MODELING CORRELATED MUTATIONS IN COMPUTATIONAL BIOLOGY
USING LOG-LINEAR ANALYSIS AND GRAPH-THEORETIC PROBABALISTIC
INERENCE METHODS

By

OMAR HAQ

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Computational Biology and Molecular Biophysics

written under the direction of

Dr. Ronald M. Levy

and approved by

________________________

________________________

________________________

________________________

New Brunswick, New Jersey

October, 2012
ABSTRACT OF THE DISSERTATION

Modeling Correlated Mutations in Computational Biology using Log-Linear Analysis and Graph-Theoretic Probabilistic Inference Methods

By OMAR HAQ

Dissertation Director:
Professor Ronald M. Levy

Point mutations are random events but selection for protein stability and function fixes specific combinations of amino acid mutations in the protein population. Many mutations are not independent but are found to be strongly correlated, the signal for which is present in multiple sequence alignment data. Using HIV Protease as a model system, this work bridges the gap between the analysis of protein sequences using statistical techniques developed by the physics and computer science communities, and the biophysical modeling of protein energetics. Using information theoretic methods together with a coarse-grained (Generalized Born) energy model we have analyzed the contribution of electrostatic interactions to protein stability among mutated residues of HIV-1 protease based on models derived from a large database of sequences which have acquired drug resistance. In the course of this work we have constructed a mean field model at the level of pair correlations (Bethe approximation) to predict the probabilities of observing mutated sequences using the HIV sequence database to parameterize the model.
Dedication

Dedicated to my loving and supportive family
Acknowledgements

I would like to take this opportunity to acknowledge the contributions of the many people in my life who supported and contributed towards my doctoral studies. I would not be at this point without these people.

First of all, I want to thank my mom and dad for their encouragement, love, and for providing me with the basis, support, and guidance to complete my PhD. I am who I am because of you. I owe you so much and I know nobody is prouder of my achievements than my parents.

I would like to thank my wife, Shafa, for standing by me through the highs and lows of graduate school life and for her love and unconditional support while I completed this long, and sometime arduous, journey. You were my rock through grad school and remain so after. I appreciate all you have done for me and continue to do for me. Every step I take in life is to try and make an even better future for us.

To my brother Taaha, you have always been closest friend growing up. You taught me how to be a leader, how to compromise, and how to achieve. These were qualities I used throughout my PhD.

I would like to thank my extended family, and especially my grandparents, for their love. Seeking a higher education was your primary goal for your grandchildren and today, I have reached the pinnacle.

I would also like to acknowledge and thank my friends. It would be impossible to withstand the ups and downs of graduate school without my Brandeis friends, my Rutgers friends, high school friends and other friends I made throughout my life. I have too many
friends to mention here because I have been so fortunate in my life thus far. I surrounded myself with so many wonderful people.

My labmates and Rutgers professors were an integral part of my PhD. I would like to thank them for their words of advice, the quick blackboard diagrams, and the quiet, quick conversations in corridors and hallways of Rutgers. All these people, made me love science, and love the scientific community. I belonged in this field because of the welcoming nature of the scientists around me. I would like to specifically mention and thank Michael Andrec and Alex Morozov. Thank you so much.

Last and certainly not least, I would like to thank my advisor, Ron for his guidance, patience and attention. I know I am a better researcher and thinker after having studied under your tutelage. The lessons I learned while in your group are going to stay with me forever.
# Contents

Abstract ii  
Dedication iii  
Acknowledgement iv  
List of Figures xxii  
List of Tables xxviii  
List of Appendices xxix  
List of Symbols 1  

1 Modeling Correlated Mutations 1  

2 Log-Linear Modeling of Correlations in HIV Protease 4  
2.1 Introduction 4  
2.2 Theory and Methods 6  
2.2.1 Pair level correlation 6  
2.2.2 Correlation at a triplet level 9  
2.2.3 The theory of pairwise and higher-order correlation 18  
2.2.4 HIV sequence databases 26  
2.3 Results 28  
2.3.1 Increased mutation frequencies under drug exposure 28
2.3.2 Evidence of pair correlations in HIV Protease 31
2.3.3 Sequence conservation and the properties of correlation measures 39
2.3.4 Correlated mutation analysis of HIV Protease 43
2.3.5 Pair correlations and the structure of HIV protease 43
2.3.6 Exhaustive analysis of residue triples 45
2.3.7 Higher-order interactions in clusters of \( n > 3 \) residues 49
2.3.8 Relationship between triplet conditional probabilities and the Kirkwood approximation 53
2.3.9 When is the Kirkwood approximation exact? 68

2.4 Appendix I: Publication attached 72

3 Introduction to Inference on Probabilistic Graphical Models 87
3.1 What are probabilistic graphical models? 87
3.2 Background to inference on graphical models 89
3.3 Examples of simple graphical models 90
3.4 Basic probabilistic inference 93
3.5 Exhaustive enumeration 93
3.6 The elimination algorithm 95
3.7 Introduction to messages 97
3.8 Decomposition of messages 99
3.9 Determination of bivariate marginals using Elimination 102
3.10 Introduction to belief propagation 103
3.11 Message update equations in matrix notation 105
3.12 Marginal probability calculations with the message update equations 107
3.13 Examples of Belief Propagation on Undirected Acyclic Graphs 109
3.13.1 3 nodes, single chain acyclic graph 110
3.13.2 5 nodes, single chain acyclic graph 111
3.13.3 Belief propagation on a simple MRF 114
4 Correspondence between Belief Propagation and the Bethe Approximation

4.1 Potts and Ising Models 120
4.2 Mean Field Free Energy 123
4.3 Bethe Mean Field Free Energy 127

5 Modeling Electrostatic Mutations in HIV Protease with a Potts Model by applying Probabilistic Inference Methods 136

5.1 Introduction 136
5.2 Methods 139
  5.2.1 HIV sequence databases 139
  5.2.2 Calculation of electrostatic folding free energies using a Generalized Born (GB) model 141
  5.2.3 Modeling Electrostatic Correlations with a Potts model 142
  5.2.4 Posing the inverse problem for the Potts model for sequence probabilities 145
  5.2.5 Brief outline of the algorithm 148
  5.2.6 Belief propagation on factor graphs 149
  5.2.7 Solving the inverse problem 151
  5.2.8 The algorithm in detail 152
5.3 Results 153
  5.3.1 Effect of electrostatic mutations on average biophysical stability 153
  5.3.2 Covariance analysis of mutation patterns 156
  5.3.3 Relationship between log-linear and Gaussian models for joint mutation probabilities 158
  5.3.4 Charge patterns explain electrostatic pair correlations 160
  5.3.5 Relationship between log-linear and discretized Gaussian models for 2 positions 162
  5.3.6 Electrostatic effects of packing charges into fixed volume 168
6 Correlated electrostatic mutations and the structure of HIV protease

6.1 Introduction

6.2 The electrostatic interaction network of wildtype HIV protease
   6.2.1 R8-D29-D30, D29-R87
   6.2.2 K14-E65-K70
   6.2.3 K45-K43-D60
   6.2.4 K20-E34
   6.2.5 E35-R57
   6.2.6 D25 (chain A) - D25 (chain B)

6.3 Energetics of short-range correlated double mutations
   6.3.1 D30N-N88D
   6.3.2 G16E-L63H

6.4 Energetics of long-range correlated double mutations
   6.4.1 G16E-N37D
   6.4.2 K20I-E35Q
   6.4.3 N37D-H69Q

6.5 Electrostatic free energy of cooperativity and sequence correlations
   6.5.1 Determining the p-value
   6.5.2 Calculating $\Delta G_{\text{coop}}$, the electrostatic free energy of cooperativity
6.5.3 $\Delta G_{coop}$ is the Born interaction energy in our coarse-grained electrostatic model 241
6.5.4 P-values for subsets of the most deviated pairs 242
6.5.5 Justification for comparing $\Delta G_{coop}$ to the observed statistical deviation 244
6.5.6 Measures of statistical deviation 247
6.5.7 Examining strongly correlated pairs of electrostatic mutations whose $\Delta G_{coop}$ sign is consistent with the sign of statistical correlation in the database 247
6.5.8 Is the sign of $\Delta G_{coop}$ consistent with like/unlike pairs of charges in a dielectric? 253

7 Understanding direct and indirect interactions in light of the Bethe approximation 255
7.1 Goal of this mini-study 255
7.2 Approach used 256
7.3 Simple 3 node complete graph 257
  7.3.1 Test 1 258
  7.3.2 Test 2 260
  7.3.3 Test 3 262
7.4 Simple 3 node acyclic graph 264
  7.4.1 Test 4 264
  7.4.2 Test 5 266
  7.4.3 Test 6 268
  7.4.4 Test 7 270

References 273

Appendices 285
List of Figures

2.1 Mutation residue frequencies for selected positions in the HIV protease as a function of the number of PIs the patient was exposed to. Data is shown for residues 10 (black, solid line), 54 (red, solid line), 90 (green), 71 (blue), 46 (orange), 77 (black, dashed line), and 35 (red, dashed line). 29

2.2 Triplet mutual information $I_3(A, B, C)$ as a function of $\alpha$ for the marginals derived from the 46-54-90 triplet. The circle represents the observed triplet distribution. 33

2.3 Triplet mutual information $I_3(A, B, C)$ as a function of $\alpha$ for the marginals derived from the 24-54-82 triplet. The circle represents the observed triplet distribution. 37

2.4 Predicted vs observed quartet probabilities for 300 residue quadruples in HIV protease. 38

2.5 An alignment of HIV protease subtype B sequences extracted from patients not exposed to any drugs. Dashes symbolize the wildtype amino acid, the sequence of which is listed at the top for reference. 43

2.6 An alignment of HIV protease subtype B sequences extracted from patients exposed to six drugs. Dashes symbolize the wildtype amino acid, the sequence of which is listed at the top for reference. 43
2.7 Histograms of the prevalence of primary drug-resistance positions among large and small $I^{(3)}_c$ triples. The red histogram corresponds to the 400 triples with the largest $I^{(3)}_c$ values, while the black histogram corresponds to the 400 triples with $I^{(3)}_c < 10^{-4}$ and the largest $I_{multi}$.

2.8 Distribution of pair correlations ($\phi$, panel A) and three-body interactions ($I^{(3)}_c$, panel B) for all pairs and triples (respectively) containing a given residue. The largest $I^{(3)}_c$ values are not distributed uniformly among all residues: residues 46, 54, 82 and 90 have the largest $I^{(3)}_c$ values ($> 0.015$), while intermediate values between 0.015 and 0.005 occur for residues 10, 20, 30, 32, 35, 36, 71, 73, 74, 77, 84, 88, and 93.

2.9 Predicted vs observed probabilities for a triple mutant for all triples of drug-associated positions in the HIV protease using the Shafer database of sequences obtained from drug-naive patients (A) and those treated with two or more inhibitors (B). Mutational states that were unobserved in the database and would have an observed probability maximum likelihood estimate of zero are not shown. The dots correspond to the best fit pair-term model, and the red lines of slope 1 correspond to the perfect agreement of the predicted probabilities with those observed, which would be obtained if the three-body interactions were included.

2.10 Probability distributions for the total number of mutations among the 41 drug-associated residues in HIV protease for sequences obtained from PI-naive patients (black), patients receiving PI monotherapy (red), and patients receiving 2 or more PIs as annotated in the Shafer database (see Methods).
2.11 Difference in the frequency of mutated residues between patients treated with 5 PIs and PI-naive patients for all positions in the HIV protease. Red bars correspond to primary drug resistance positions, black bars are accessory drug resistance positions, and gray bars are positions not associated with drug treatment.

2.12 Scatter plot of distance vs $\phi$ statistic for drug associated positions in red and non-drug associated positions in black. Pairs with cell counts of 20 or less for $A_mB_0$ and $A_0B_m$ were excluded. The value along the $y$-axis is the closest distance between any two heavy atoms of the two residues based on the crystal structure of wild type protease (PDB ID 1PRO).

2.13 Structure of the HIV protease dimer (PDB code 1HN0). The catalytic triad (residues 25, 26, and 27) is highlighted in green, while the residues from the triple with the largest $I_c^{(3)}$ (46–54–90, see Figure 2.15A) are shown in red. Residue 90 is close to the catalytic triad, whereas residues 46 and 54 are close to each other, with the smallest distance between heavy atoms of 3.11 Å.

2.14 Correlation between $I_c^{(3)}$ and $\phi$ values. Each point corresponds to one of the 45 drug-associated residue positions. The abscissa (ordinate) correspond to the average of $\phi$ ($I_c^{(3)}$) over all pairs (triples) containing that residue position.
2.15 Predicted vs observed probabilities for the residue triple in HIV protease with the largest $I^{(3)}_c$ (46–54–90) (A), as well as two three-residue mutational clusters previously identified by Wu, et al.[122]: 30–82–88 (B), and 10–46–90 (C). The black dots correspond to the best-fit independent model, while the red dots correspond to the best-fit pair-term model (Equation 2.37). The solid line of slope 1 corresponds to perfect agreement of the predicted data with the observed, which would be obtained if the three-body term were included.

2.16 Distribution of the 100 largest $I^{(3)}_c$ values observed for the Shafer and Lee databases. Panels A–C show the results for subsets of the Shafer database corresponding to patients that were PI-naive (A), PI-monotherapy treated (B), and exposed to 2 or more PIs (C). Panels D and E correspond to the complete Shafer and Lee databases, respectively.

2.17 Predicted vs observed probabilities for the $2^{10}$ mutational states of the ten-residue group 20–32–46–48–53–54–58–74–82–90 in HIV protease. Mutational states that were unobserved in the database and would have an observed probability maximum likelihood estimate of zero are not shown. The black dots correspond to the best-fit independent model, the red dots correspond to the best-fit pair-term model (Equation 2.37), and the green dots correspond to the best-fit three-body model (Equation 2.38). The solid line of slope 1 corresponds to perfect agreement of the predicted data with the observed, which would be obtained if all higher-order terms were included.
2.18 Total number of observed (solid black curve) or predicted mutations for the ten-residue group 20–32–46–48–53–54–58–74–82–90 in HIV protease. The dashed curves represent the predictions for the independent (blue), pair-term (red), and pair plus three-body model (green). 41,668 sequences from the Lee database were used to train the models (sequences for which unambiguous codons were present at all 10 positions, see Methods).

2.19 Distributions of the total number of mutations in the Lee database among the 41 drug-associated positions. The black curve corresponds to the observed data, the blue curve corresponds to the model in which mutations are assumed to occur independently of each other, and the red curve corresponds to the best fit using a two-body model under the “structural zero” approximation (see text). The independent model data were generated by a Monte Carlo procedure in which many sequences were generated using independent Bernoulli trials at each residue position.

2.20 Scatter plot of distance vs $\lambda_{ij}$ parameters estimated using the $\pm 1$ gauge. The value along the $y$-axis is the closest distance between any two heavy atoms of the two residues based on the crystal structure of wild type protease (PDB ID 1PRO).

3.1 Example of an acyclic directed graph on the left and a cyclic undirected graph on the right.

3.2 The simplest graphical model possible consists of only one node. Its graphical structure tells us everything about the class of probability functions that describe the only random variable within this graph, $A_1$. For this probabilistic graphical model, the joint probability $P(A_1) = \frac{1}{Z} \psi_1(A_1)$ where $\psi_1(A_1)$ is the field potential at node 1.
3.3 This simple graphical model consists of two nodes. The absence of an edge between $A_1$ and $A_2$ tells us that there is no probabilistic dependence between $A_1$ and $A_2$. For this probabilistic graphical model, the joint probability $P(A_1, A_2) = \frac{1}{Z} \psi_1(A_1) \psi_2(A_2)$ where $\psi_1(A_2)$ and $\psi_2(A_2)$ are the field potentials at nodes 1 and 2. 91

3.4 This simple graphical model consists of two nodes. The presence of an edge between $A_1$ and $A_2$ tells us that there is a probabilistic dependence between $A_1$ and $A_2$. For this probabilistic graphical model, the joint probability $P(A_1, A_2) \approx \psi_{12}(A_1, A_2)$ where $\psi_{12}(A_1, A_2)$ is the coupling potentials between nodes 1 and 2. In other words, the presence of an edge describes a statistical dependency between the variables. 91

3.5 This three node graph contains only two edges, and is therefore acyclic. The graph structure indicates that if we condition on the variable $A_2$, then $A_1$ and $A_3$ are conditionally independent. As a result, the joint probability for this graph will have a specific structure. For this probabilistic graphical model, the joint probability $P(A_1, A_2, A_3) \approx \psi_{12}(A_1, A_2) \psi_{23}(A_2, A_3)$ where $\psi_{12}(A_1, A_2)$ and $\psi_{23}(A_2, A_3)$ are the coupling potentials. It is this conditional independence feature that will be exploited in belief propagation. 91
This graph contains three nodes and three edges where every
genode is connected to every other node. This is known as a com-
pletely cyclic graph. There is no conditional independence struc-
ture in this graph. For this probabilistic graphical model, the
joint probability \( P(A_1, A_2, A_3) \approx \psi_{123}(A_1, A_2, A_3) \) where \( \psi_{123}(A_1, A_2, A_3) \)
is the triplet potential. It is not exactly correct to describe the
joint probability distribution of this graph in terms of the cou-
pling potentials, where \( P(A_1, A_2, A_3) \approx \psi_{12}(A_1, A_2)\psi_{23}(A_2, A_3)\psi_{13}(A_1, A_3) \).
However, in some situations, including in this work, it is a rea-
sonably accurate approximation.
5.1 Average $\Delta G_e$ of sequences as a function of the number of electrostatic mutations. Each point on a curve corresponds to $\sum_{n=1}^{N^m} P^n \Delta G^n_e$ where $N^m$ is the number of sequences with $m$ electrostatic mutations, $P^n$ is the probability of the $n$th sequence under a given model normalized with respect to the number of mutations and $\Delta G^n_e$ is its electrostatic folding energy. All points are plotted relative to $\sum_{n=1}^{N^1} P^n_{obs} \Delta G^n_{e,obs}$, the average $\Delta G_e$ of observed sequences with 1 mutation. The black curve represents the averages energies of observed sequences. The blue curve represents the average energies of sequences under an independent model in which each charge state has the observed univariate probabilities of occurrence. The red curve represents the average energies of sequences under an independent model in which each charge state has an equal probability of occurrence. The green curve represents the average energy of sequences under a pair model where the observed pair correlations have been preserved. Points on the red, blue and green curves correspond to an average of $10^6$ simulations, where in each simulation, $N^m$ sequences are randomly picked from the corresponding probability distributions. The error bars on the red, blue and green curves are the average standard errors of the mean for these $10^6$ simulations. The error bars on the black curve are the standard errors of the mean of observed sequences.

5.2 Spectrum of eigenvalues for the covariance matrix for the 18 highly correlated electrostatically active positions using a two-letter alphabet conditioned on 4, 5, or 6 electrostatic mutations. The ordinate is the eigenvalue and the abscissa is the index number sorted in order of decreasing eigenvalue.
5.3 Eigenvectors corresponding to the 3 largest eigenvalues of the covariance matrix for the 18 highly correlated electrostatically active positions using a two-letter alphabet conditioned on 4, 5, or 6 electrostatic mutations. Numbers indicate residue numbers of the larger eigenvector components.

5.4 Comparison of probabilities of individual sequences in a 2-letter wild-type/mutant alphabet over 18 positions under the log-linear model including pair terms, and the discretized Gaussian model described in section 5.3.3. The line denotes perfect agreement between the two probabilities.

5.5 Cumulative probabilities of the sequences shown in Figure 5.4. Sequences have been sorted in order of decreasing log-linear probability, and the black and red curves denote the cumulative probability under the log-linear and discretized Gaussian models. For example, the 20 sequences with the largest log-linear probabilities account for approximately 45% of the total probability, while those same 20 sequences account for 90% of the probability under the discretized Gaussian model. The red dashed curve is the result for the empirically corrected discretized Gaussian model where the factor of $1/2$ has been replaced by $0.2518$. Each curve reaches 1 for sequence number $2^{18}$.

5.6 Average electrostatic folding free energy vs total number of charges.

5.7 Average electrostatic folding free energy as a function of the number of electrostatic mutations conditional on 10 positive and 9 negative charges (black) and 11 positive and 9 negative charges (red).
5.8 Plots of $\Delta G_e.P_2$ versus $\Delta G_e.P_1$ for observed (red) and unobserved (black) sequences with 1, 2, 3, 4, 5 and 6 electrostatic mutations. These plots represent the contributions of each sequence to the average electrostatic folding free energies shown in Figure 5.1. Each point represents a unique sequence with $\Delta G_e.P_1$ and $\Delta G_e.P_2$ on the x-axis and y-axis respectively, where $P_1$ and $P_2$ are the independent and pair log-linear model probabilities of the sequence, normalized with respect to the number of mutations, and $\Delta G_e$ is the electrostatic folding energy. Sequences which contribute an extraordinary degree to the stabilization of the folded state under the pair correlation model are labeled by their mutation pattern. Mutations are represented as $aNb$ where $N$ is the residue number and $a$ and $b$ are one of the 3 charged states (+, -, n). The straight line on each diagram is a plot of $x = y$. Sequences below this line have $P_1 < P_2$, resulting in $\Delta G_e.P_1 > \Delta G_e.P_2$. For these sequences, the electrostatic stabilization is greater under the pair model than under the independent model.

5.9 Observed (red) and unobserved (green) sequences with 4 mutations sorted by their pair correlation model probabilities.

5.10 Observed (red) and unobserved (green) sequences with 5 mutations sorted by their pair correlation model probabilities.

5.11 Observed (red) and unobserved (green) sequences with 6 mutations sorted by their pair correlation model probabilities.
5.12 Observed number of mutations as a function of treatment history (number of protease inhibitors administered to the patient, panel A), and the number of electrostatic mutations (panel B) in the Shafer annotated database. In each panel, the black curve represents the total number of non-electrostatic mutations, the red curve represents the number of mutations known experimentally to be destabilizing, and the green curves represent the number of non-electrostatic primary (solid) and accessory (dashed) drug resistance mutations. In panel A, the blue line represents the total number of electrostatic mutations.

5.13 Free energy of folding for sequences with drugs versus without drugs.

5.14 Electrostatic mutation frequency for the least populated charge state for each position as a function of the number of electrostatic mutations.

6.1 Salt bridge between ARG 8 of chain A and ASP 29 of chain B

6.2 Two interactions are depicted in this figure. There is a salt bridge between R8 of chain A and D29 of chain B. There is also a close interaction between D29 and R87 of the same chain. Though not depicted in this picture, D29 is also within 6.5Å of D30. D29 is within 6.5Å of three charged residues.

6.3 Two salt bridge interactions are depicted in this figure, both of which involve E65. One is between E65 and K14 and the other is between E65 and K70.

6.4 Two electrostatic interactions are depicted in this figure, both of which involve K43. One is a destabilizing interaction between K43 and K45 and the other is a stabilizing interaction between K43 and D60.

6.5 A salt bridge between K20 and E34.
6.6 A salt bridge between E35 and R57

6.7 A destabilizing interaction between D25 of chain A and D25 of chain B

6.8 A plot of distance between pairs of positions as a function of the electrostatic free energy of cooperativity between the charges.

6.9 The percentage of pairs where statistical correlation is consistent with sign of the free energy of cooperativity, $\Delta G_{\text{coop}}$, plotted as a function of the percentage of total pairs, sorted by decreasing correlation. At the top is the plot of the negatively correlated pairs while the plot at the bottom are the positively correlated pairs. Correlation is calculated using the following measure: $\frac{(P_{\text{obs}} - P_{\text{ind}})^2}{P_{\text{ind}}}$. 244

7.1 This graph contains three nodes and three edges where every node is connected to every other node. This is known as a completely cyclic graph or simply a complete graph. There is no conditional independence structure in this graph. 257

7.2 This three node graph contains only two edges, and is therefore acyclic. The graph structure indicates that if we condition on the variable $A_2$, then $A_1$ and $A_3$ are conditionally independent. As a result, the joint probability for this graph will have a specific structure. For this probabilistic graphical model, the joint probability $P(A_1, A_2, A_3) \approx \exp [\lambda_{12}(A_1, A_2)\lambda_{23}(A_2, A_3)]$ where $\lambda_{12}(A_1, A_2)$ and $\lambda_{23}(A_2, A_3)$ are the coupling potentials. $\lambda_{13}(A_1, A_3) = 0$. 264
List of Tables

2.1 The general form of a $2 \times 2$ contingency table. .......................................................... 6
2.2 The general form of a $2 \times 2 \times 2$ contingency table. ....................................................... 10
2.3 Contingency table for the triple of residues 46, 54, and 90 of HIV protease. ......................... 31
2.4 Conditional probabilities for the observed distribution $P(A, B, C)$ for residues 46-54-90. .......... 32
2.5 Conditional probabilities for the maximum entropy distribution $\tilde{P}^{(2)}(A, B, C)$ for residues 46-54-90. 34
2.6 Contingency table for the triple of residues 24, 54, and 82 of HIV protease.. ............................ 35
2.7 Conditional probabilities for the observed distribution $P(A, B, C)$ for residues 24, 54, and 82. 35
2.8 The distribution of factorization classes for unambiguous model selections. ......................... 35
2.9 The 20 triples with the least ambiguous factorization model assignment. All cases shown here belong to Class E (Equation 2.12). .......................... 36
2.10 The 20 triples with the largest $I_C^{(3)}(A, B, C)$ values. ......................................................... 40
2.11 Triples with “large” $I_3(A, B, C)$ values. Triples with zero or very small contingency table cell counts are not shown. .......................................................... 41
2.12 Contingency table for two random variables with probabilities instead of counts

2.13 Contingency table for two completely anti-correlated random variables with equal univariate marginals of 0.50

2.14 Contingency table for two completely correlated random variables with equal univariate marginals of 0.50

2.15 Contingency table for two independent random variables with equal univariate marginals of 0.50

2.16 Correlation measures for different values of the univariate marginals

2.17 Total number of observed and predicted mutations for the 10 residue group 20–32–46–48–53–54–58–74–82–90. The total sample size is 41,668 sequences for which unambiguous codons were present at all 10 positions (see Methods).

2.18 The 10 triples in the HIV protease with the largest $I_C^{(3)}(A, B, C)$ values.

2.19 The 10 triples in the HIV protease with small $I_C^{(3)}(A, B, C)$ values but large $I_{\text{multi}}(A, B, C)$.

2.20 Actual joint probabilities versus the Kirkwood Approximation

2.21 Actual Marginal Probabilities

5.1 Univariate marginals and the resistance role for the 18 positions in HIV protease that have electrostatic mutation frequencies > 0/01%.

5.2 The 15 residue pairs in HIV protease with the largest magnitude pair correlation in the Lee database conditioned on 4, 5 or 6 electrostatic mutations, their pair correlation in a 2-letter alphabet, the predicted sense of the correlation based on a naive examination of charge patterns.
6.1 Charged amino acids within the wildtype sequence of HIV protease subtype B, the amino acid residue type and the mutation frequency for that residue.

6.2 The twenty most statistically enhanced double mutations in the Lee database, in terms of the observed number of sequences (not limited to sequences with only two mutations) that contain these mutations relative to the independent model. The measure used to test for enhancement is $\text{dev}(i, j) = \frac{(P_{ij}(M,M) - P_i(M)P_j(M))^2}{P_{ij}(M,M)}$ where $P_{ij}(M,M)$ is the joint probability of a double mutation at positions $i$ and $j$ while $P_i(M)$ is the univariate marginal of a mutation at position $i$. The wildtype and double mutant charge states, distance between charges, the free energy of folding relative to the wildtype protein ($\Delta \Delta G$) and the free energy of cooperativity ($\Delta G_{\text{coop}}$) are also listed.

6.3 The twenty most statistically suppressed double mutations in the Lee database, in terms of the observed number of sequences (not limited to sequences with only two mutations) that contain these mutations relative to the independent model. The measure used to test for enhancement is $\text{dev}(i, j) = \frac{(P_{ij}(M,M) - P_i(M)P_j(M))^2}{P_i(M)P_j(M)}$ where $P_{ij}(M,M)$ is the joint probability of a double mutation at positions $i$ and $j$ while $P_i(M)$ is the univariate marginal of a mutation at position $i$. The wildtype and double mutant charge states, distance between charges, the free energy of folding relative to the wildtype protein ($\Delta \Delta G$) and the free energy of cooperativity ($\Delta G_{\text{coop}}$) are also listed.
6.4 Observed and independent bivariate marginals for the single and double mutants of positions 30 and 88 along with their charge states, electrostatic folding energies. The positions are ~6Å apart. $\Delta G_e$ is the electrostatic folding free energy. $\Delta \Delta G_e = \Delta G_e - \Delta G_{e}^{\text{wildtype}}$ where $\Delta G_{e}^{\text{wildtype}} \approx -89$

6.5 Observed and independent bivariate marginals for the single and double mutants of positions 16 and 63 along with their charge states, electrostatic folding energies. The positions are ~8Å apart. $\Delta G_e$ is the electrostatic folding free energy. $\Delta \Delta G_e = \Delta G_e - \Delta G_{e}^{\text{wildtype}}$ where $\Delta G_{e}^{\text{wildtype}} \approx -89$

6.6 Observed and independent bivariate marginals for the single and double mutants of positions 16 and 37 along with their charge states, electrostatic folding energies. The positions are ~9Å apart. $\Delta G_e$ is the electrostatic folding free energy. $\Delta \Delta G_e = \Delta G_e - \Delta G_{e}^{\text{wildtype}}$ where $\Delta G_{e}^{\text{wildtype}} \approx -89$

6.7 Observed and independent bivariate marginals for the single and double mutants of positions 20 and 35 along with their charge states, electrostatic folding energies. The positions are ~11Å apart. $\Delta G_e$ is the electrostatic folding free energy. $\Delta \Delta G_e = \Delta G_e - \Delta G_{e}^{\text{wildtype}}$ where $\Delta G_{e}^{\text{wildtype}} \approx -89$

6.8 Observed and independent bivariate marginals for the single and double mutants of positions 37 and 69 along with their charge states, electrostatic folding energies. The positions are ~24Å apart. $\Delta G_e$ is the electrostatic folding free energy. $\Delta \Delta G_e = \Delta G_e - \Delta G_{e}^{\text{wildtype}}$ where $\Delta G_{e}^{\text{wildtype}} \approx -89$
6.9 The 60 most statistically deviated double mutations in the Lee database relative to the independent model. The measure used to test for deviation is \( \text{dev}(i, j) = \frac{(P_{ij}(M,M) - P_i(M)P_j(M))^2}{P_{ij}(M,M)} \) where \( P_{ij}(M, M) \) is the joint probability of a double mutation at positions \( i \) and \( j \) while \( P_i(M) \) is the univariate marginal of a mutation at position \( i \). The double mutant charge states, distance between charges, the free energy of cooperativity (\( \Delta G_{coop} \)) are also listed.

6.10 A sample single and double mutation

6.11 The proportion of correct predictions and the p-value for the significance of the prediction as a function of the total number of predictions after sorting by the deviation. All 612 double mutant pairs are sorted by the deviation of the observed bivariate marginals from the independent model. Predictions. The measure used to test for deviation is \( \text{dev}(i, j) = \frac{(P_{ij}(M,M) - P_i(M)P_j(M))^2}{P_{ij}(M,M)} \) where \( P_{ij}(M, M) \) is the joint probability of a double mutation at positions \( i \) and \( j \) while \( P_i(M) \) is the univariate marginal of a mutation at position \( i \). The free energy of cooperativity (\( \Delta G_{coop} \)) for each pair is determined and compared to the pairs statistical enhancement or suppression. A correct prediction for a double mutant implies that pair is enhanced and the sign of \( \Delta G_{coop} \) is negative or if the pair is suppressed and the sign of \( \Delta G_{coop} \) is positive.
6.12 The p-values for the significance of the prediction for separate subsections of the data. All 612 double mutant pairs are sorted by the deviation of the observed bivariate marginals from the independent model predictions. The measure used to test for deviation is $dev(i, j) = \frac{(P_{ij}(M, M) - P_i(M)P_j(M))^2}{P_{ij}(M, M)}$ where $P_{ij}(M, M)$ is the joint probability of a double mutation at positions $i$ and $j$ while $P_i(M)$ is the univariate marginal of a mutation at position $i$. The free energy of cooperativity ($\Delta G_{coop}$) for each pair is determined and compared to the pairs statistical enhancement or suppression. A correct prediction for a double mutant implies that pair is enhanced and the sign of $\Delta G_{coop}$ is negative or if the pair is suppressed and the sign of $\Delta G_{coop}$ is positive. The p-values for the first 60, second 60, third 60 etc predictions are listed below.

6.13 Frequency of electrostatically active positions within the top 60 most correlated double mutants.
List of Appendices

A  Operating Manual For Code  285
Chapter 1

Modeling Correlated Mutations

In protein evolution, selection for fitness coupled with structural constraints leads to the conservation of structurally and functionally important amino acids on one hand along with the non-random association of other residues on the other\[16, 121, 27, 2, 23\]. The role of conserved residues has been thoroughly documented but the correlated nature and biophysical basis behind the non-random association of amino acids has only recently begun to receive attention \[17, 18\]. In fact, most models of protein evolution treat mutations as independent events and do not consider correlated interactions between residues \[65, 41, 78\] even though they are highly prevalent and have been observed in the multiple sequence alignments (MSAs) of many protein families, especially those evolving under intense environmental selection pressure, such as viral proteins \[13, 20, 19, 44, 48, 49, 62, 73, 80, 93, 11, 122\]. Previous work has indicated that including up to pair correlations is more than sufficient in explaining observable trends in complex biological networks and that high-order terms might be unnecessary \[8, 98\].

As a result of this interest, several research groups have developed methods meant to identify pairs of coevolving mutations in multiple sequence alignments \[37, 70, 20, 19, 38, 44, 77\]. The majority of these methods are based on mutual information \[108, 37, 69, 58\], perturbation analysis \[64, 26\], or combinatorial approaches \[6\]. These methods have successfully identified pairs of coevolving residues that tend to be catalytic residues or are found to be phys-
ically interacting electrostatically or via hydrogen bonds and disulfide bridges [59, 29]. Locating physically interacting pairs through the analysis of correlated mutations has been complicated by the effects of phylogeny, because of which pairs of residues seem to be correlating, but do so only because of a historic mutation leading to divergence in the evolution of the protein. Therefore, accounting for phylogenetic effects in their correlation metrics has been a major focus of several groups and has led to better predictions of interacting pairs of residues both within a protein and between systems of interacting proteins [15, 31, 3]. Going a step further, machine learning techniques have been applied to cluster together similarly covarying pairs in an attempt to isolate groups or 'patches' of coevolving residues [122, 62, 123]. This viewpoint is interesting because a recent study has shown that semi-independent clusters of correlated amino acids may have distinct roles to play in the evolution of a protein, where one cluster of co-evolving residues is mainly responsible for protein stability while another could be mainly responsible for enzymatic function [42].

Not all residues involved in these coevolving clusters physically interact. In fact, a major complication that has existed has been the inability to distinguish directly interacting residues from those that are indirectly correlated, an effect that can arise due to the cyclic nature of the network of correlations. Weigt et al. 2009 made great progress in this regard, by determining the strength of a direct pair correlation if residues are isolated from the effects of the network [118]. They discovered that highly correlated direct interactions strongly indicated physically interacting residues that are found to be proximal to each other in the crystal structure models of the proteins.

The goal of this thesis is to connect the information gleaned from multiple sequence alignments to protein physics and energetics. Correlation mutation patterns, can give us glimpse into the interactions that govern protein folding and stability. For example, Ranganathan, et al. 1999 attempted to explain
mutational coevolution by connecting statistical free energies from multiple sequence alignments to differences in experimental folding free energies from double mutant cycle experiments [64]. Unfortunately, some of these results have been difficult to replicate [21], and are still a topic of active debate in the community [26, 33, 63]. While studies that link mutational correlations to protein structure and thermodynamics have made great progress [4, 77, 38, 95], a consensus linking protein energetics and mutational correlation patterns has not yet emerged. To help bridge this gap even further, in Chapter 2, we develop a hierarchical log-linear model that includes correlations up to the the triplet level, and show that including up to pair correlations is crucial in explaining observable trends in a multiple sequence alignment of HIV Protease. The log-linear model we developed is a two-state model. In Chapter 5, we expand upon our previous research on hierarchical log-linear models with a two-letter alphabet [44] to include three letters and apply a coarse-grained electrostatic energy model to calculate folding energies, allowing us to study the average electrostatic folding energy of HIV protease in the presence and absence of correlations. The Potts model is made up of field and coupling parameters that can describe the joint probability distribution of mutations in HIV protease. The goal is to determine the parameters of this model, such that the observed correlations in HIV protease are preserved. Correlations are function of the bivariate and univariate marginals and our problem is an inverse one: Given the univariate and bivariate marginals, we wish to determine the field and coupling parameters of a Potts model such that the marginals are preserved.
Chapter 2

Log-Linear Modeling of Correlations in HIV Protease

2.1 Introduction

The protease enzyme coded for by the pol gene of the Human Immunodeficiency Virus HIV-1 plays a critical role in the reproduction of the virus by cleaving the GAG precursor protein in a sequence-specific manner into its functional form, and as such, is a key target of several families of commonly used drugs used to control HIV infection[106]. Unfortunately, the virus has been able to evolve resistance to many of these drugs, in part due to the high mutation rates in the HIV genome[90]. In contrast to other key HIV enzymes (such as reverse transcriptase), the patterns of mutations in protease are complex, involving multiple key primary mutations that inhibit the action of the drug and a host of secondary mutations that can modulate the enzyme’s stability or activity or otherwise enhance the fitness of the virus. It is now understood that these mutations do not occur independently of each other, but instead are correlated, resulting in complex patterns of co-evolving mutations. [122, 19, 93, 114]

Previous studies have mostly focused on correlations between mutations in the HIV protease gene at the pair level. [122, 19, 93, 114]. However, recognition that the observed mutations may also be involved in higher-order interactions has led to several studies in which correlated pairs of mutations are grouped
using tools such as multidimensional scaling\cite{122, 93}, Bayesian networks\cite{25},
networks defined by patterns of conditional selection pressure\cite{19}, and clustering (both graph-theoretic\cite{62} and hierarchical\cite{105}). The underlying assumption is that understanding of higher-order interactions is critical for complete understanding of evolution in the HIV protease and in particular the emergence of drug resistance.

In this work, we investigate the impact of correlations in mutations of the HIV protease both at and beyond the pair level. We develop a hierarchy of probabilistic log-linear models that can in principle describe residue interactions of arbitrary order. We use these models in conjunction with the notion of “connected information” \cite{10, 99} to quantify inter-residue correlations at the triplet level. Unlike the Bayesian network approach\cite{82}, the information-theoretic methodology allows us to distinguish intrinsic three-body effects from the cases in which correlations between three random variables can be attributed mostly to pairwise interactions. The connected and mutual information viewpoints of higher order correlation have not been previously used in the analysis of mutational patterns in the HIV protease, although they have been employed in a much more limited analysis of the V3 loop of the HIV envelope protein\cite{36, 1}. We find that pairwise interactions are necessary to achieve even qualitative agreement with the mutational data, while including three-body and higher interactions makes predictions quantitatively accurate. In particular, higher order interactions play an important role in predicting how frequently sequences with several mutations appear in the database. The frequencies of these sequences are grossly overestimated by the pairwise model. Simultaneous appearance of multiple mutations may play an important role in the phenomenon of multiple-or cross-resistance of the viral protease.
2.2 Theory and Methods

2.2.1 Pair level correlation

Before tackling the level of triples, let us first consider the considerably simpler case of pairs of binary random variables. The probability distribution in this case can be fully specified by 3 parameters, e.g. \( P(A_0, B_0) \), \( P(A_0, B_m) \), and \( P(A_m, B_0) \), with the remaining \( P(A_m, B_m) \) being determined by normalization. Alternatively, the distribution can be parametrized by specifying the two univariate marginals \( P(A_0) \) and \( P(B_0) \) as well as one of the joint probabilities (e.g. \( P(A_0, B_0) \)). In addition, the full joint density can be factorized as \( P(A, B) = P(A)P(B|A) = P(B)P(A|B) \). This factorization is equivalent to specifying the marginals and one joint, since \( P(A|B) = P(A, B)/P(B) \).

The observed data for pairs of residues can be arranged into a 2 × 2 “contingency table” (Table 2.1). The \( N_{ij} \) values represent the number of times each binary combination was observed. Dividing each cell count by the total number of counts in the table provides the maximum likelihood estimates of the cell probabilities \( p_{ij} \).

<table>
<thead>
<tr>
<th></th>
<th>residue A mutated</th>
<th>residue A wild-type</th>
</tr>
</thead>
<tbody>
<tr>
<td>residue B mutated</td>
<td>( N_{mm} )</td>
<td>( N_{0m} )</td>
</tr>
<tr>
<td>residue B wild-type</td>
<td>( N_{m0} )</td>
<td>( N_{00} )</td>
</tr>
</tbody>
</table>

For pairs, there are only two qualitatively different situations: the random variables \( A \) and \( B \) are either independent, or they are correlated. In the former case, \( P(A, B) = P(A)P(B) \), from which it follows that \( P(A|B_0) = P(A|B_m) = P(A) \) and \( P(B|A_0) = P(B|A_m) = P(B) \). If \( A \) and \( B \) are correlated, then \( P(A|B_0) \neq P(A|B_m) \), and the magnitude of the difference is one measure of the degree of correlation, e.g. \( P(A_0|B_0) = 0.8 \) and \( P(A_0|B_m) = 0.1 \).
is strong correlation, while $P(A_0|B_0) = 0.8$ and $P(A_0|B_m) = 0.79$ is very weak correlation.

2.2.1.1 Covariance

Covariance is simply defined as the difference between the observed bivariate marginal and the product of the univariate marginals:

$$Cov(A, B) = P(A, B) - P(A)P(B)$$

2.2.1.2 Pearson correlation coefficient

The Pearson correlation coefficient is the covariance divided by the product of the standard deviations of the two variables:

$$\rho(A, B) = \frac{Cov(A, B)}{\sigma(A)\sigma(B)}$$

2.2.1.3 Log-linear representation

Yet another re-parametrization of the probability distribution can be obtained by considering a log-linear or energy function representation

$$P(A, B) = \exp(\lambda_0 + \lambda_A A + \lambda_B B + \lambda_{AB} AB), \quad (2.1)$$

where we assign numerical values to the states of $A$ and $B$, such as $A_0 = B_0 = 1$ and $A_m = B_m = -1$. In this formulation, $\lambda_A$ and $\lambda_B$ are related to the populations of $A_0$ and $B_0$, while $\lambda_{AB}$ is determined by the correlation, with $\lambda_{AB} = 0$ if $A$ and $B$ are independent [10]. The remaining parameter $\lambda_0$ For the numerical state assignment $A_0 = B_0 = 1$ and $A_m = B_m = -1$, the values of
the parameters are given by

\[
\begin{align*}
\lambda_0 &= \frac{1}{4} \ln(p_{00}p_{0m}p_{m0}p_{mm}) \\
\lambda_A &= \frac{1}{4} \ln \left( \frac{p_{00}p_{0m}}{p_{0m}p_{mm}} \right) \\
\lambda_B &= \frac{1}{4} \ln \left( \frac{p_{00}p_{m0}}{p_{0m}p_{mm}} \right) \\
\lambda_{AB} &= \frac{1}{4} \ln \left( \frac{p_{00}p_{mm}}{p_{0m}p_{m0}} \right). 
\end{align*}
\]

**2.2.1.4 Mutual information**

An alternative information-theoretic measure of correlation is the mutual information

\[
I_2(A, B) = S(A) + S(B) - S(AB) = \sum_{a,b} P(a, b) \log \frac{P(a, b)}{P(a)P(b)},
\]

where \(S(A) = -\sum_a P(a) \log P(a) \geq 0\) is the information-theoretic entropy. In general, \(0 \leq I_2(A, B) \leq \min\{S(A), S(B)\}\), with \(I_2(A, B) = 0\) if \(A\) and \(B\) are independent, and \(I_2(A, B)\) taking on its maximum value if one of the random variables is a deterministic function of the other (e.g. \(P(B_0|A_0) = 1\)) [71].

**2.2.1.5 Product moment coefficient**

Since \(\lambda_{AB} = 0\), above, implies independence, the “cross-product ratio”

\[
\rho_{AB} = \frac{p_{00}p_{mm}}{p_{0m}p_{m0}} \approx \frac{N_{00}N_{mm}}{N_{0m}N_{m0}}
\]

can be used as a measure of correlation, with \(\rho_{AB} = 1\) indicating no correlation [10].
2.2.1.6 Chi square analysis

2.2.1.7 Jaccard Coefficient

Various alternative measures of correlation between binary random variables have been proposed [40]. One which is commonly used in chemoinformatics [57] and has been applied to the analysis of mutation patterns in the HIV genome [93] is the Jaccard coefficient, which is defined as

\[ J = \frac{N_{mm}}{N_{0m} + N_{m0} + N_{mm}}. \] (2.3)

Thus, \( J \) is the number of times both positions are mutated simultaneously divided by the number of times at least one position is mutated. On average, we expect that

\[ J = \frac{P(A_m, B_m)}{1 - P(A_0, B_0)}, \]

which in the independent limit becomes

\[ J_{\text{ind}} = \frac{P(A_m)P(B_m)}{P(A_m)P(B_m) - P(A_m) - P(B_m)}. \]

Unlike \( I_2(A, B) \) and \( \lambda_{AB} \), \( J \) cannot be interpreted in the absence of knowledge of the marginal probabilities. More generally, it is necessary to compare an observed value of \( J \) compared to the distribution of \( J \) under the null hypothesis of independence [93].

2.2.2 Correlation at a triplet level

The interpretation of triplet frequencies is considerably more complex, mainly because the range of possible ways that the variables can interact with each other is much larger. No single summary statistic can capture all of the various
Table 2.2: The general form of a $2 \times 2 \times 2$ contingency table.

<table>
<thead>
<tr>
<th></th>
<th>residue A mutated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>residue B mutated</td>
</tr>
<tr>
<td>residue C mutated</td>
<td>$N_{mmm}$</td>
</tr>
<tr>
<td>residue C wild-type</td>
<td>$N_{m00}$</td>
</tr>
<tr>
<td>residue A wild-type</td>
<td></td>
</tr>
<tr>
<td>residue C mutated</td>
<td>$N_{0mm}$</td>
</tr>
<tr>
<td>residue C wild-type</td>
<td>$N_{0m0}$</td>
</tr>
</tbody>
</table>

interesting properties and characteristics of triplet distributions, requiring the use of multiple statistics that provide complimentary information and have different strengths and limitations.

For three binary variables, the data can be arranged in the form of a $2 \times 2 \times 2$ contingency table (Table 2.2), and there are 8 joint probabilities $P(A, B, C)$, 1 of which is determined by normalization. The joint probabilities can also be specified in terms of the univariate marginals $P(A_0)$, $P(B_0)$, and $P(C_0)$, and the bivariate marginals $P(A_0, B_0)$, $P(B_0, C_0)$, and $P(A_0, C_0)$. This is sufficient to constrain 6 of the 7 free parameters of the joint distribution. The remaining free degree of freedom can be fixed, for example, by specifying one of the joint probabilities. If we choose to specify $P(A_0, B_0, C_0)$ in addition to the marginals, then we can write

$$P(A_0, B_0, C_0) = \alpha \quad (2.4)$$
$$P(A_0, B_0, C_m) = D_1 - \alpha \quad (2.5)$$
$$P(A_0, B_m, C_0) = D_3 - \alpha \quad (2.6)$$
$$P(A_0, B_m, C_m) = p_a - D_1 - D_3 + \alpha \quad (2.7)$$
$$P(A_m, B_0, C_0) = D_2 - \alpha \quad (2.8)$$
$$P(A_m, B_0, C_m) = p_b - D_1 - D_2 + \alpha \quad (2.9)$$
$$P(A_m, B_m, C_0) = p_c - D_2 - D_3 + \alpha \quad (2.10)$$
$$P(A_m, B_m, C_m) = 1 - p_a - p_b - p_c + D_1 + D_2 + D_3 - \alpha, \quad (2.11)$$
where for notational convenience we have defined $D_1 = P(A_0, B_0)$, $D_2 = P(B_0, C_0)$, $D_3 = P(A_0, C_0)$, $p_a = P(A_0)$, $p_b = P(B_0)$, $p_c = P(C_0)$, and $\alpha = P(A_0, B_0, C_0)$. It should be noted that these seven parameters cannot be chosen arbitrarily, but must satisfy certain inequality constraints in order to give consistent and non-negative probabilities. In particular, the bivariate marginals must satisfy $D_1 \leq \min[p_a, p_b]$, $D_2 \leq \min[p_b, p_c]$, and $D_3 \leq \min[p_a, p_c]$. Furthermore, $\alpha$ must lie in the range $\alpha_{\min} \leq \alpha \leq \alpha_{\max}$, where

$$\alpha_{\min} = \max[0, D_1 + D_3 - p_a, D_1 + D_2 - p_b, D_2 + D_3 - p_c, D_1 + D_2 + D_3 - p_a - p_b - p_c]$$

and

$$\alpha_{\max} = \min[D_1, D_2, D_3, 1 + D_1 + D_3 - p_a, 1 + D_1 + D_2 - p_b, 1 + D_2 + D_3 - p_c, 1 - p_a - p_b - p_c + D_1 + D_2 + 1]$$

If $\alpha_{\min} > \alpha_{\max}$, then the univariate and bivariate marginals are inconsistent, and there is no valid triplet distribution that can satisfy all of them. Equations 2.4-2.11 allow us to construct all possible joint distributions consistent with a set of given marginals parametrically as a function of $\alpha$.

In general, we can rewrite the joint distribution as products of conditional probabilities:

$$P(A, B, C) = P(A)P(B|A)P(C|A, B) = P(B)P(C|B)P(A|B, C) = \text{etc.}$$

(2.12)

However, in certain special cases we can factorize the joint probability into simpler forms. There are 10 such special forms in addition to the most general form given by Equation 2.12, and are given by

$$P(A, B, C) = P(A)P(B)P(C),$$

(2.13)

$$P(A, B, C) = P(A, B)P(C),$$

(2.14)
\begin{align*}
P(A, B, C) &= P(A)P(C)P(B), \quad (2.15) \\
P(A, B, C) &= P(B)P(C | A)P(A | B), \quad (2.16) \\
P(A, B, C) &= P(A)P(B | A)P(C | A), \quad (2.17) \\
P(A, B, C) &= P(A)P(B | A)P(C | A), \quad (2.18) \\
P(A, B, C) &= P(A)P(B | A)P(C | A), \quad (2.19) \\
P(A, B, C) &= P(A)P(B | A)P(C | A), \quad (2.20) \\
P(A, B, C) &= P(C)P(A | C)P(B | C), \quad (2.21) \\
\end{align*}

and

\begin{align*}
P(A, B, C) &= P(A)P(C)P(B | A). \quad (2.22)
\end{align*}

The model of Equation 2.13 represents the case where all three variables are independent of each other (Class A), while the models of Equations 2.14-2.16 have one variable independent of the other two, which are pairwise-correlated (Class B). The remaining models represent more subtle forms of three-way independence. The models of Equations 2.20-2.22 are cases where two of the variables are independent of each other, but the third variable is simultaneously correlated with both of the others (Class D), while Equations 2.17-2.19 represent the case if we know the state of one of the three variables and hold it fixed, then the other two variables become independent of each other ("conditional independence") (Class C). The different classes of factorization models share the same number of free parameters: the Class A model requires the specification of only the 3 univariate marginal probabilities, Class B models require 4 parameters (3 univariate probabilities plus a parameter describing the bivariate correlation), Class C models require 5 parameters, while Class D models require 6 parameters. We will refer to the most general, unreduced 7 parameter model as Class E.
Given the count data, we can use a Bayesian model selection procedure to decide which of these 11 models represent the most parsimonious representation of the data. In particular, we can evaluate the marginal likelihood of the data given model $M$ with parameter vector $\theta$:

$$P(D|M) = \int d\theta \ P(D|\theta, M)P(\theta).$$

(2.23)

The marginal likelihood represents the weight of evidence in favor of model $M$, and the “Bayes factor” $P(D|M_1)/P(D|M_2)$ represents the odds in favor of model $M_1$ over model $M_2$. For a multinomial likelihood function and a uniform or Dirichlet prior, the integral in Equation 2.23 can be evaluated analytically as a ratio of Gamma functions [45].

Just as for pairs, we can re-parametrize the triple joint probability distribution in log-linear form

$$P(A, B, C) = \exp(\lambda_0 + \lambda_A A + \lambda_B B + \lambda_C C + \lambda_{AB} AB + \lambda_{BC} BC + \lambda_{AC} AC + \lambda_{ABC} ABC).$$

(2.24)

For the numerical state assignment $A_0 = B_0 = C_0 = 1$ and $A_m = B_m = C_m =$
the values of the parameters are given by

\[
\begin{align*}
\lambda_0 &= \frac{1}{8} \ln(p_{000}p_{00m}p_{0m0}p_{0mm}p_{m00}p_{m0m}p_{m0m}p_{m0m}p_{mm0}) \\
\lambda_A &= \frac{1}{8} \ln \left( \frac{p_{000}p_{00m}p_{0m0}p_{0mm}}{p_{mm0}p_{mm0}p_{m00}p_{m0m}} \right) \\
\lambda_B &= \frac{1}{8} \ln \left( \frac{p_{000}p_{0m0}p_{00m}p_{m00}}{p_{0m0}p_{m00}p_{0mm}p_{m0m}} \right) \\
\lambda_C &= \frac{1}{8} \ln \left( \frac{p_{000}p_{00m}p_{0m0}}{p_{00m}p_{0m0}p_{00m}p_{m00}} \right) \\
\lambda_{AB} &= \frac{1}{8} \ln \left( \frac{p_{000}p_{mm0}p_{00m}p_{mm0}}{p_{00m}p_{00m}p_{0mm}p_{0mm}} \right) \\
\lambda_{AC} &= \frac{1}{8} \ln \left( \frac{p_{000}p_{0m0}p_{m00}p_{mm0}}{p_{0m0}p_{0m0}p_{m00}p_{m0m}} \right) \\
\lambda_{BC} &= \frac{1}{8} \ln \left( \frac{p_{000}p_{0mm}p_{0m0}p_{mm0}}{p_{0m0}p_{0m0}p_{00m}p_{m00}} \right) \\
\lambda_{ABC} &= \frac{1}{8} \ln \left( \frac{p_{000}p_{mm0}p_{mm0}p_{mm0}}{p_{00m}p_{00m}p_{mm0}p_{mm0}} \right).
\end{align*}
\]

By analogy with pairs, where \( \lambda_{AB} = 0 \) implies lack of pair correlation (first-order interaction), we can view triplet distributions with \( \lambda_{ABC} = 0 \) as having no second-order interaction [10]. An alternative interpretation of the \( \lambda_{ABC} = 0 \) model can be obtained in terms of the cross-product ratio \( \rho_{AB} \). If \( \lambda_{ABC} = 0 \), then

\[
\frac{p_{000}p_{mm0}p_{0mm}p_{m00}}{p_{00m}p_{00m}p_{0mm}p_{0mm}} = 1
\]

or

\[
\frac{p_{000}p_{mm0}}{p_{00m}p_{000}} = \frac{p_{000}p_{mm0}}{p_{00m}p_{mm0}}.
\]

However, these two ratios are simply conditional cross-product ratios, e.g.

\[
\rho_{AB|C_0} = \frac{p_{000}p_{mm0}}{p_{00m}p_{000}}
\]

is the cross-product ratio for \( A \) and \( B \) given that \( C \) is wild-type. Therefore, \( \lambda_{ABC} = 0 \) implies that \( \rho_{AB|C_0} = \rho_{AB|C_m} \), i.e. the bivariate correlation between two of the three variables as measured by the cross-product ratio is the same
regardless of what the state of the third variable is.

There are multiple ways in which one can formulate information-theoretic characterizations of triplet distributions. One approach which has a strong connection to the log-linear model above is the concept of “connected information” [99]. It begins by defining the “multi-information” or “total correlation” [116] as the Kullback-Leibler divergence of a triplet distribution from the “independent” distribution (i.e. Equation 2.13) having the same univariate marginals:

\[
I_{\text{multi}}(A, B, C) = \sum_{abc} P(a, b, c) \log \frac{P(a, b, c)}{P(a)P(b)P(c)} = S(A) + S(B) + S(C) - S(A, B, C).
\]

We then decompose \( I_{\text{multi}} \) as a sum of second- and third-order connected information: \( I_{\text{multi}}(A, B, C) = I_C^{(2)}(A, B, C) + I_C^{(3)}(A, B, C) \) [99]. The relative magnitudes of these terms is based on maximum entropy: we define the distribution \( \tilde{P}^{(2)}(A, B, C) \) as the distribution having the maximum entropy subject to the constraint that all of its univariate and bivariate marginals are the same as that of \( P(A, B, C) \) and define \( I_C^{(2)}(A, B, C) = S[P(A)P(B)P(C)] - S[\tilde{P}^{(2)}(A, B, C)] \) [99]. One way to find \( \tilde{P}^{(2)}(A, B, C) \) is to calculate the marginals, and then maximize the entropy as a function of the parameter \( \alpha \) using the relationships in Equations 2.4-2.11.

The connection to log-linear models is provided by a theorem due to I. J. Good[39], which states that if none of the joint probabilities \( P(A, B, C) \) are zero, then \( \tilde{P}^{(2)}(A, B, C) \) can be written in the form of Equation 2.24 with \( \lambda_{ABC} = 0 \). Thus, \( \tilde{P}^{(2)}(A, B, C) \) can be thought of has having no interactions beyond the pair level. It also suggests an alternative strategy for determining \( \tilde{P}^{(2)}(A, B, C) \). Instead of maximizing \( S \) with respect to \( \alpha \), we can solve for the values of \( p_{ijk} \) which give \( \lambda_{ABC} = 0 \). Since there is only one free parameter, and we wish to preserve the marginals, this is equivalent to finding the value of \( \delta \)
such that
\[
\frac{(p_{000} + \delta)(p_{mm0} + \delta)(p_{0mm} + \delta)(p_{m0m} + \delta)}{(p_{0m0} - \delta)(p_{n00} - \delta)(p_{n0m} - \delta)(mmm - \delta)} = 1.
\]
This leads to a cubic equation in \(\delta\) which can be solved to give \(\bar{P}^{(2)}(A, B, C)\) [5, 10]. Furthermore, the maximum entropy distribution subject to the marginal constraints is equivalent to the distribution corresponding to Equation 2.24 with the parameters estimated by maximum likelihood subject to the constraint that \(\lambda_{ABC} = 0\), provided that the cell counts are sufficiently large [39].

There are potential problems with these strategies (maximization with respect to \(\alpha\) or solving for \(\delta\)) if any of the joint probabilities \(P(A, B, C)\) are zero, requiring the use of alternative methods to assess the significance of the \(\lambda_{ABC}\) term. One strategy is to make use of a Bayesian model selection procedure to choose between the models corresponding to Equation 2.24 with \(\lambda_{ABC} = 0\) or \(\lambda_{ABC} \neq 0\). Since this is a set of nested models, we can make use of a simplification that allows us to calculate the Bayes factor in favor of the \(\lambda_{ABC} = 0\) based on the marginal probability density of the posterior distribution of \(\lambda_{ABC}\) at zero compared to the corresponding probability density under the prior [112].

The relationship between the connected information and the probability factorizations of Equations 2.13-2.22 above is less straightforward, except for the Class A and B factorization models. If the joint probability factorizes fully (i.e. Equation 2.13), then obviously \(I_{\text{multi}}(A, B, C) = I^{(2)}_C(A, B, C) = I^{(3)}_C(A, B, C) = 0\). For the Class B factorizations (Equations 2.14-2.16) it can be verified that \(\lambda_{ABC} = 0\), therefore \(I^{(3)}_C(A, B, C) = 0\) and \(I_{\text{multi}}(A, B, C) = I^{(2)}_C(A, B, C) \neq 0\). The remaining classes can lead to either \(I^{(3)}_C(A, B, C) = 0\) or \(I^{(3)}_C(A, B, C) \neq 0\), depending on the values of the marginals. Thus, the joint distribution could contain triplet-level probabilistic dependencies, but still be consistent with no triplet-level connected information, if the observed triplet distribution is the maximum entropy distribution relative to its marginals (i.e. the
observed distribution contains the least amount of “additional information” beyond that contained in the marginals).

An alternative information-theoretic criterion by which the triplet distribution can be characterized is the third-order mutual information

\[
I_3(A, B, C) = S(A) + S(B) + S(C) - S(AB) - S(AC) - S(BC) + S(ABC)
\]

\[
= I_{multi}(A, B, C) - I_2(A, B) - I_2(A, C) - I_2(B, C) \quad (2.27)
\]

\[
= -\sum_{abc} P(a, b, c) \log \frac{P(a, b, c)}{\hat{P}(a, b, c)}, \quad (2.28)
\]

where

\[
\hat{P}(A, B, C) = \frac{P(A, B)P(B, C)P(A, C)}{P(A)P(B)P(C)}
\]

is the Kirkwood superposition approximation to \(P(A, B, C)\) [71]. The triplet mutual information is a measure of the degree of redundancy or synergy among the three random variables, and can be related to the difference between the marginal and conditional pairwise mutual information:

\[
I_3(A, B, C) = I_2(A, B) - I_2(A, B|C).
\]

Thus, the triplet mutual information can be positive or negative, depending on whether knowledge of the third variable decreases or increases the apparent pairwise correlation between the other two, respectively. In other words, \(I_3(A, B, C) > 0\) indicates that knowledge of one of the three variables reduces the amount of information that the second variable can provide about the third. A somewhat extreme example of such a situation would be a joint distribution that satisfies Equation 2.17. If the state of variable \(A\) is known, then the additional knowledge of the state of variable \(B\) can provide no additional predictive power with regard to variable \(C\). By contrast, if the joint distribution factorizes according to Equation 2.20, then variables \(A\) and \(B\) work synergistically to help predict \(C\), i.e. knowledge of the state of both \(A\) and \(B\) allows us to
predict $C$ more accurately than if we had $A$ or $B$ alone. Thus, joint distributions of factorization Class D have $I_3(A, B, C) < 0$, those of Class C have $I_3(A, B, C) > 0$, and those of Classes A and B have $I_3(A, B, C) = 0$, since in those cases $P(A, B, C) = \hat{P}(A, B, C)$. If the joint distribution cannot be factorized beyond the most general form of Equation 2.12, then no a priori statement about the sign of $I_3(A, B, C)$ can be made. It should be noted that $I_3(A, B, C)$ can equal zero even if $P(A, B, C) \neq \hat{P}(A, B, C)$, as will be seen below. Therefore, $I_3(A, B, C) = 0$ if at least one of the three variables are independent of the others, but the converse need not be true. Given the univariate marginals, the largest magnitude that $I_3(A, B, C)$ can attain is given by the minimum of the univariate entropies $S(A)$, $S(B)$, and $S(C)$, and is achieved only in the case of a deterministic relationship among the random variables [71].

### 2.2.3 The theory of pairwise and higher-order correlation

In this work, we will only consider the presence or absence of a non-synonymous mutation relative to a defined wild-type sequence, and not the precise base or amino acid substitution which has occurred. Thus, we will treat sequences as binary strings, and for an amino acid position $A$ we will denote the wild-type state as $A_0$ and the mutated state as $A_m$. For the sake of notational convenience, we will use a shorthand notation in the form of $p_{0m0} = P(A_0, B_m, C_0)$ for the probability of position A to have a wild-type amino acid, position B to have a mutant amino acid, and position C to have a wild-type amino acid (all other probabilities are similarly defined). We will also use $N_{0m0}$ to represent the number of times that we observe position $A$ and $C$ wild-type and $B$ mutated.

To better understand the conceptual complications that occur with higher-order correlations, let us first consider the more straightforward case of pair correlations among binary random variables. In that case there are only two
qualitatively different situations: the random variables $A$ and $B$ are either independent, or they are correlated. In the former case, $P(A, B) = P(A)P(B)$, from which it follows that $P(A|B_0) = P(A|B_m) = P(A)$ and $P(B|A_0) = P(B|A_m) = P(B)$. If $A$ and $B$ are correlated, then $P(A|B_0) \neq P(A|B_m)$, and the magnitude of the difference could be used to quantify the degree of pair correlation. Many other measures of correlation have been proposed, including the pair mutual information or Kullback-Leibler divergence,[43]

$$I_2(A, B) = \sum_{a,b} P(a,b) \log \frac{P(a,b)}{P(a)P(b)}$$ (2.29)

(with $I_2 = 0$ indicating no correlation), the binomial or “product moment”[10] correlation coefficient

$$\phi_{AB} = \frac{P(A_mB_m) - P(A_m)P(B_m)}{\sqrt{P(A_m)P(A_0)P(B_m)P(B_0)}}$$ (2.30)

(with $\phi_{AB} = 0$ indicating no correlation), the “cross-product ratio”[10]

$$\rho_{AB} = \frac{N_{00}N_{mm}}{N_{0m}N_{m0}} \approx \frac{p_{00}p_{mm}}{p_{0m}p_{m0}}$$ (2.31)

(with $\rho_{AB} = 1$ indicating no correlation), and a variety of other measures. [10, 93, 40, 57] Although many of these correlation measures are quantitatively different, and differ in their sensitivity in various regimes, they are all measuring essentially the same qualitative feature of the observed data.

The characterization of the distribution of data for three binary random variables is considerably more complex. No single summary statistic can capture all of the various characteristics of triplet distributions, leading to multiple statistics that provide complimentary information and which have different strengths and weaknesses. Perhaps the most intuitive description of the presence or absence of “three-way correlation” is in terms of factorizations of the general probability distribution. For example, if $P(A, B, C)$ can be well-
approximated by $P(A,B)P(C)$, then clearly there is no “three-way correlation”, as one of the random variables is independent of the other two. On the other hand, consider the case where the joint distribution can be written as $P(A)P(B|A)P(C|A)$. While data arising from such a distribution may appear to be consistent with correlations among all three variables, this is not the case, since the apparent pair correlation between $B$ and $C$ is due only to their common correlation with $A$. Thus, if we know the state of $A$, then $B$ and $C$ become statistically independent, and there is no “three-way correlation”.

If two of the variables are independent of each other, but the third depends jointly on the state of the other two, e.g. $P(A,B,C) = P(A)P(B|A)P(C|A,B)$, or if $P(A,B,C)$ cannot be factorized into any simpler form, then one could say that there is “three-way correlation” in the sense that the distribution of at least one of the variables depends non-trivially on the state of the other two. This strategy of looking for the simplest factorization of a joint probability distribution forms the basis of the Bayesian network method for describing the dependencies between sets of random variables [83], and statistical model selection procedures can be used to select the most parsimonious models [45]. While this type of analysis can be used to uncover conditional or unconditional independence among random variables, it has the limitation that it cannot provide insight into the nature of the correlations that are present.

Assessing how close an observed distribution is to having no three-body interaction can be used as a measure of three-way correlation that is qualitatively distinct from the Bayesian network-style factorization approach described above. One statistic that quantifies this is the “connected information” [99], which is defined as the difference in Shannon entropy $S(P) = -\sum_i p_i \log p_i$, between the distributions $P(A,B,C)$ and $\tilde{P}^{(2)}(A,B,C)$, where the latter is the maximum entropy distribution subject to the constraints that all of its univari-
ate and bivariate marginals are the same as that of $P(A, B, C)$:

$$I^{(3)}_c(A, B, C) = S[\tilde{P}^{(2)}(A, B, C)] - S[P(A, B, C)].$$  \hfill (2.32)

The connected information $I^{(3)}_c$ provides information which is complementary to a Bayesian network-style probability factorization. It can readily verified by substitution into Equation 2.35 that $I^{(3)}_c = 0$ if at least one of the random variables is independent of the other two, or if the joint distribution involves conditional independence (e.g. $P(A, B, C) = P(A)P(B|A)P(C|A)$). The other possible factorizations can lead to zero or non-zero values of $I^{(3)}_c$, depending on the values of the marginals. Thus, the joint distribution could contain triplet-level probabilistic dependencies but still be consistent with no three-body connected information if the observed triplet distribution is the maximum entropy distribution relative to its marginals (i.e. the observed distribution contains the least amount of additional information beyond that contained in the marginals).

Thus, even if a Bayesian network-style analysis shows that a given triple cannot be factorized into any simpler form, that “triplet correlation” could still be consistent with a very small or zero $I^{(3)}_c$, indicating that the observed behavior is consistent primarily with only two-body interactions. In fact, it has been shown that very complex correlation patterns among random variables can arise from large numbers of weak pairwise interactions[98]. Besides $I^{(3)}_c$, we define the “multi-information” or the Kullback-Leibler divergence between the observed distribution and the prediction based on an independent model:

$$I_{\text{multi}}(A, B, C) = \sum_{abc} P(a, b, c) \log \frac{P(a, b, c)}{P(a)P(b)P(c)} = S(A) + S(B) + S(C) - S(A, B, C).$$  \hfill (2.33)

We compute $I^{(3)}_c$ by writing triplet probability distribution in log-linear
form[10]:

\[ P(A, B, C) = \exp(\lambda_0 + \lambda_A A + \lambda_B B + \lambda_C C + \lambda_{AB} AB + \lambda_{BC} BC + \lambda_{AC} AC + \lambda_{ABC} ABC), \]  

(2.34)

where we assign numerical values to the states of A, B and C, e.g. 0 and 1. We can view triplet distributions based on the model with \( \lambda_{ABC} = 0 \) as having no three-body interactions. If in addition the pair terms are absent the model factorizes into the product of individual probabilities for A, B, and C. The model can be easily extended to include correlations beyond the third order.

Note that \( \lambda_{ABC} = 0 \) in Equation 2.34 implies that

\[ \frac{p_{000}p_{mm0}}{p_{0m0}p_{m00}} = \frac{p_{00m}p_{mmm}}{p_{0mm}p_{m0m}} \]  

(2.35)

Since the left and right hand sides of Equation 2.35 are the pair cross-product ratio (Equation 2.31) for A and B conditional on C being wild-type or mutant, respectively, the lack of three-body interaction means that the bivariate correlation between A and B conditional on C, as measured by the pair cross-product ratio is the same regardless of the state of C. This is also true by symmetry for all other permutations of the three random variables.

We can find \( \hat{P}^{(2)}(A, B, C) \) by setting \( \lambda_{ABC} = 0 \) in Equation 2.34 and fitting the six parameters \( \lambda_i \) and \( \lambda_{ij} \) to the values that maximize the likelihood of the data under a multinomial model (with appropriate regularity conditions)[39]. However, in the triplet case it is possible to avoid direct nonlinear optimization: let us represent 8 observed probabilities by the vector It is sufficient to consider only the three marginals \( P(A_m), P(B_m), \) and \( P(C_m) \), and three suitably chosen bivariate marginals, e.g. \( P(A_m, B_m), P(A_m, C_m), \) and \( P(B_m, C_m) \), since the remaining 9 bivariate marginals can be reconstructed as combinations of these: \( P(A_m, B_0) = P(A_m) - P(A_m, B_m), \) \( P(A_0, B_m) = P(B_m) - P(A_m, B_m), \)
\( P(A_0, B_0) = 1 - P(A_m) - P(B_m) + P(A_m, B_m), \) etc. The six marginals can then be written as a matrix equation involving \( p_0 \):

\[
\begin{pmatrix}
0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & 1
\end{pmatrix}
\begin{pmatrix}
P(A_m) \\
P(B_m) \\
P(C_m) \\
P(A_m, B_m) \\
P(A_m, C_m) \\
P(B_m, C_m)
\end{pmatrix}
= \begin{pmatrix}
P(A_0) \\
P(B_0) \\
P(C_0) \\
P(A_0, B_0) \\
P(A_0, C_0) \\
P(B_0, C_0)
\end{pmatrix}.
\]

Since the matrix multiplying \( p_0 \) is rectangular with dimensions \( 6 \times 8 \), it has a two-dimensional null space, with basis vectors \( \mathbf{n}_1 = (1, 0, 0, 0, 0, 0, 0, 0) \) and \( \mathbf{n}_2 = (0, -1, -1, 1, -1, 1, 1, -1) \). Then, any linear combination \( \alpha_1 \mathbf{n}_1 + \alpha_2 \mathbf{n}_2 \) added to \( p_0 \) will not change the marginals. However, \( \alpha_1 \) and \( \alpha_2 \) cannot be chosen independently without violating the normalization of \( p_0 \): we must choose \( \alpha_1 = \alpha_2 = \alpha \). Therefore, the family of all possible distributions that have the same univariate and bivariate marginals are mapped out by the parameter \( \alpha \) using the relation

\[
p_\alpha = p_0 + \alpha(1, -1, -1, 1, -1, 1, 1, -1),
\]

where the feasible values of \( \alpha \) are constrained by the non-negativity requirement for probabilities. Therefore, it suffices to find the value of \( \alpha \) which satisfies Equation 2.35 and thus leads to \( \lambda_{ABC} = 0 \):

\[
\frac{(p_{000} + \alpha)(p_{mm0} + \alpha)(p_{m0m} + \alpha)(p_{0mm} + \alpha)}{(p_{0m0} - \alpha)(p_{m00} - \alpha)(p_{00m} - \alpha)(p_{mmm} - \alpha)} = 1.
\]

(2.36)

This is a cubic equation in \( \alpha \) which can be solved to give \( \tilde{P}^{(2)}(A, B, C) \) [10].

To obtain the maximum entropy distributions for more than three binary random variables, nonlinear optimization is unavoidable. However, in-
instead of directly maximizing the entropy subject to the marginal probability constraints, we maximize the likelihood subject to the constraints that the $\lambda$’s vanish beyond a given order[39]. The number of $\lambda$ variables grows polynomially with the number of variables, while the dimensionality of the null space defined by the marginal probability constraints (which is one-dimensional for three variables) increases exponentially with the number of variables.

In general, we fit data on mutation and wild-type amino acid counts to the following hierarchy of probabilistic models: the independent model $P(A, B, C, \ldots) = P(A)P(B)P(C) \cdots$, the “two-body” model:

$$P(A, B, C, \ldots) = \exp(\lambda_0 + \sum_i \lambda_i I + \sum_{ij} \lambda_{ij} IJ),$$  \hspace{1cm} (2.37)

and the “three-body” model:

$$P(A, B, C, \ldots) = \exp(\lambda_0 + \sum_i \lambda_i I + \sum_{ij} \lambda_{ij} IJ + \sum_{ijk} \lambda_{ijk} IJK),$$  \hspace{1cm} (2.38)

where $\lambda$ is the vector of parameters, the indices $(i, j)$ and $(i, j, k)$ run over all distinct combinations of \{A, B, C, \ldots\} with $i \neq j$ and $i \neq j \neq k$, respectively, and $I$, $J$, and $K$ are numerical values of the corresponding random variable (we use 0 for wild-type and 1 for mutant). For an $n$-variate distribution $P(A, B, C, \ldots)$ there are $n(n-1)/2$ pair parameters $\lambda_{ij}$ and $n(n-1)(n-2)/6$ three-body parameters $\lambda_{ijk}$ ($\lambda_0$ is a normalization constant). The independent model was determined by forming products of the observed univariate marginals. The magnitudes of the $\lambda_i$ and $\lambda_{ij}$ parameters in the two-body model are conceptually related to the mutation frequencies at site $i$ and pair correlations between sites $i$ and $j$, respectively. In fact, the magnitudes of $\lambda_{ij}$ in the context of a two-body model have been proposed as a measure of “direct information”, i.e. the part of pair correlation resulting from direct coupling [117]. It should be
noted that the relative magnitudes of $\lambda_{ij}$ is dependent on the choice of “gauge”, i.e. the numerical values assigned to the random variables. In this work, we make use of a 0–1 gauge for computational convenience. It has been argued, however, that a more appropriate gauge is one which is symmetric about zero, e.g. ±1 [117]. While the choice of gauge will affect the values of $\lambda_{ij}$ and their interpretation as “direct information”, it will not change the best-fit two-body or three-body probabilities and consequently will have no impact on the values of $I_c^{(3)}$ and related measures of higher-order correlation.

The two-body model for $n = 3$ was fit by non-iteratively solving Equation 2.36[89]. If no feasible solution of Equation 2.36 exists, then $I_c^{(3)}$ was set to zero. For $n \geq 4$, the unknown parameters in Equations 2.37 and 2.38 were determined by maximizing the multinomial log-likelihood

\[
L(\lambda) = \sum_i N_i \log P(i),
\]

where $i$ is one of the $2^n$ states, $N_i$ is the number of times that state was observed, and $P(i|\lambda)$ is the predicted probability for state $i$ to be observed given the vector of parameters $\lambda$. Maximization was performed numerically using the “nlm” function of the R software package[28]. Statistical significance of the higher-order interaction parameters were determined using the likelihood ratio test. For $n = 3$, this was done under the null hypothesis that the data were generated by $P^{(2)}(A, B, C)$ and $p$-values were estimated by Monte Carlo sampling. It was found empirically for $n = 3$ and the sample sizes in the Lee database (see below) that $I_c^{(3)}$ values larger than 0.0002 correspond to $p$-values less than 0.0001 (data not shown).
2.2.4 HIV sequence databases

Aligned HIV-1 DNA nucleotide sequences were downloaded from Christopher Lee’s HIV Positive Selection Mutation Database\(^1\) on March 4th, 2008. The 45,161 sequences consist of primarily HIV-1 subtype B samples (Calvin Pan, personal communication) and were donated to C. Lee by Specialty Laboratories, who sequenced them between 1999 and mid-2002. Each sequence is 1443 nucleotide bases long, the first 297 bases of which correspond to protease (99 transcribing codons). The remaining 1146 nucleotides correspond to the first 382 amino acids of the p51 reverse transcriptase and were not used in this study.

The codons transcribing protease were extracted and converted to their corresponding amino acids. Codons containing any non-standard base or ambiguities were not used in subsequent analysis. The resulting amino acid sequences were then compared to the HIV-1 subtype B consensus sequence obtained from the Los Alamos National Laboratory HIV sequence database\(^2\), which was used to define the “wild-type” sequence. This resulted in strings of symbols with each amino acid position indicating whether it was “wild-type”, “mutant”, or “undefined”. For any given subset of \(n\) residue positions from among the 99, statistics were accumulated on the number of times each of the \(2^n\) possible binary combinations of wild-type and mutant were observed. All sequences of length \(n\) containing an undefined residues at any of its positions were eliminated from consideration. Only residue positions where a mutation was observed more than 6 times in the database were included, reducing the effective number of positions from 99 to 82. We will refer to this database as the “Lee database”.

\(\text{Mutations at many of these positions are associated with decreased HIV-1}\)

---

\(^1\)http://bioinfo.mbi.ucla.edu/HIV

\(^2\)http://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html
susceptibility and it is useful to classify mutations as belonging to “primary” vs “accessory” resistance classes. The terms “secondary” and “compensatory” have also been used as synonyms for “accessory”. The specific criteria for such a classification are *ad hoc* in nature, but have generally been defined as follows.

Primary mutations are usually selected first in the presence of the drug and confer resistance, even when present as single point mutations[91, 113]. They can be structurally important, such as being situated near the enzymatic active site, in which case their effect on inhibitor binding can be rationalized due to their physical proximity to the inhibitor.[22] In the case of protease, however, there are exceptions, as some mutations (such as at positions 54, 76, 88 and 90) are situated far from the active site or have no direct contacts with the substrate, but yet still reduce drug susceptibility. The mechanism of action of these mutations is not clearly understood. [73]

Accessory mutations confer resistance only when present with additional primary or accessory drug-resistance mutations and have little or no effect on inhibitor susceptibility on their own. Some of these mutations occur in the absence of drug treatment, but their frequency is observed to increase in treated patients. They may be polymorphic in nature. Accessory mutations may “rescue” possible losses of activity or stability in the enzyme that may have been caused by a destabilizing primary resistance mutation, and therefore may have a compensatory function in restoring viral fitness.[101]

We make use of a primary and accessory classification scheme based on the work of Shafer et al.[92, 103, 102, 113] We define 17 primary drug resistance positions (residues 23, 24, 30, 32, 33, 46, 47, 48, 50, 53, 54, 73, 76, 82, 84, 88, 90) and 24 accessory drug resistance positions (residues 10, 11, 13, 20, 34, 35, 36, 43, 45, 55, 58, 60, 63, 71, 74, 75, 77, 79, 83, 85, 89, 91, 93, 95). The remaining positions are polymorphic mutations not associated with drug resistance or are
invariant [46, 122, 76, 92, 53, 102].

An additional set of curated sequences were obtained from the Stanford HIV Drug Resistance Database[91]. Using their web interface\(^3\), aligned amino acid sequences for HIV-1 protease obtained from patients treated with 0 through 6 protease inhibitors were extracted. The sequences are all classified under the HIV-1 Main group, subtype B. For purposes of this paper, we are considering only the number of protease inhibitors and not the specific combinations of protease inhibitors used in the treatment. There are 13,608 sequences in this curated set, of which 8,229 sequences are drug naive, 2,677 sequence are associated with PI monotherapy, and the remaining 2,702 sequences are associated with between 2 through 6 protease inhibitors. Of the 13,608 sequences in this dataset, 2,224 sequences match sequences in the Lee database. We will refer to this database as the “Shafer database”.

2.3 Results

2.3.1 Increased mutation frequencies under drug exposure

As has been previously observed, we find that the overall number of mutations seen in the HIV protease increases significantly with the number of protease inhibitors (PIs) that the patient has been exposed to. [122] This is seen in Figure 2.10, where we show the distribution of the number of mutations at drug-associated positions for PI-naive patients, those experiencing PI monotherapy, and those exposed to 5 PIs, as annotated in the Shafer database described above. For some residue positions, the increase in the mutation frequency is nearly eightfold, while other positions show no discernible change (Figure 2.1). This observed increase in mutation frequencies at drug-associated sites is largely

\(^3\)http://hivdb.stanford.edu/cgi-bin/PI_Form.cgi
Figure 2.1: Mutation residue frequencies for selected positions in the HIV protease as a function of the number of PIs the patient was exposed to. Data is shown for residues 10 (black, solid line), 54 (red, solid line), 90 (green), 71 (blue), 46 (orange), 77 (black, dashed line), and 35 (red, dashed line).
It is of interest to ask which amino acid positions exhibit elevated mutation frequencies under drug treatment. As can be seen in Figure 2.11, these positions are for the most part primary drug resistance positions. Accessory positions 10 and 71 also exhibit increased mutation frequency upon treatment with multiple PIs. It is interesting to note that sequences bearing mutations at any one of the 6 residues with the most elevated mutation frequencies (10, 46, 54, 71, 82, and 90) are resistant to most of the current PIs, including Amprenavir (APV), Indinavir (IDV), Nelfinavir (NFV), Ritonavir (RTV), Saquinavir (SQV), Atazanavir (ATV), and lopinavir (LPV). For instance, according to the Stanford HIV database[91], the common mutation at these residues (L10I, M46I, I54V, A71V, V82A, and L90M) are associated with six or all seven of the drugs mentioned above. On the other hand, primary drug resistance positions that show the least amount of change in their mutation frequency (23, 30, 76, and 88) are generally inhibitor specific to inhibitors. For example, D30N is associated only with NFV treatment, L23I and L76V are specific to two drugs each (NFV and SQV, and IDV and LPV, respectively), while N88D is specific to IDV, SQV and NFV. Thus, mutations at positions that provide drug resistance to multiple inhibitors are more frequently mutated in sequences that are treated by multiple drugs than positions that are specific to a small number of protease inhibitors.

Furthermore, most of the positions exhibiting large changes in mutation frequency upon multi-drug therapy are relatively distant from the active site catalytic triad (D25, T26, G27): the closest distance to any heavy atom of the catalytic triad is 8.0 Å for residue 10, 13.0 Å for residue 46, 11.7 Å for residue 54, 13.2 Å for residue 71, 6.4 Å for residue 82, and 3.7 Å for residue 90 (based on the PDB structure 1PRO). Additionally, examination of 278 ligand-bound crystal structures of HIV-1 protease, no heavy atom of residues 10, 46, 54, 71,
or 90 is ever within 3.4 Å of any ligand bound at the cleft (data not shown). The atoms of residue 82, however, do contact ligands in 54% of the crystal structures examined.

2.3.2 Evidence of pair correlations in HIV Protease

Before providing an overview of all of the results for HIV protease mutation patterns at the triplet level, let us first consider in detail two illustrative (but not particularly representative) examples. The first example is that of the triple of residues 46, 54 and 90. Residues 54 and 90 are known to be sites of drug-resistance mutations [20], so this triple is of some biological interest. The data for this triple are shown in Table 2.3. We can use these data to estimate joint, marginal and/or conditional probabilities for the occurrence of given mutation patterns. In Table 2.4 we show the probabilities corresponding to a factorization of the form of Equation 2.12, from which we can see evidence of substantial probabilistic dependencies. In particular, we see that the probability of a mutation at residue 90 is substantially increased if residue 46 is mutated, while the probability of a mutation at residue 54 is greatly reduced if both 46 and 90 are wild-type. This observation is confirmed by a Bayesian statistical analysis based on the marginal likelihoods of the data (Equation 2.23), which indicates that of the 11 models corresponding to Equations 2.12-2.22, the model with the greatest support from the data is the general Class E model of Equation 2.12, with the next-best model having a smaller odds by a factor of $\approx 10^{690}$. This constitutes overwhelming statistical support for the general model.

Table 2.3: Contingency table for the triple of residues 46, 54, and 90 of HIV protease.

<table>
<thead>
<tr>
<th></th>
<th>residue 46 mutated</th>
<th>residue 46 wild-type</th>
</tr>
</thead>
<tbody>
<tr>
<td>residue 54 mutated</td>
<td>1310</td>
<td>2427</td>
</tr>
<tr>
<td>residue 54 wild-type</td>
<td>1036</td>
<td>1393</td>
</tr>
<tr>
<td>residue 90 mutated</td>
<td>1758</td>
<td>4143</td>
</tr>
<tr>
<td>residue 90 wild-type</td>
<td>1038</td>
<td>30814</td>
</tr>
</tbody>
</table>
Table 2.4: Conditional probabilities for the observed distribution $P(A, B, C)$ for residues 46-54-90.

| $P(A_m)$ = 0.1404 | $P(C_m|A_m)$ = 0.6061 | $P(C_m|A_0)$ = 0.1563 | $P(B_m|A_mC_m)$ = 0.3505 | $P(B_m|A_mC_0)$ = 0.4265 | $P(B_m|A_0C_m)$ = 0.2979 | $P(B_m|A_0C_0)$ = 0.0326 |

We now look at the same data from the mutual and connected information perspectives. First, we ask whether the extremely significant probabilistic dependencies can be explained using only pairwise energies in a log-linear model without the need to invoke second-order interactions. In Figure 2.2 we show the values of $I_3(A, B, C)$ for the family of all distributions having the same univariate and bivariate marginals as is observed. The family was constructed by calculating the marginals, and varying the parameter $\alpha = P(A_0, B_0, C_0)$ using Equations 2.4-2.11. Since the entropies of the univariate and bivariate distributions are constant by construction, it follows from Equation 2.26 that the distribution corresponding to the extremum of the curve in Figure 2.2 is the maximum entropy distribution. The observed distribution does not match the maximum entropy distribution, as indicated by the position of the filled circle in relation to the extremum, and in fact it is extremely unlikely that the maximum entropy distribution could have generated the observed data [we still need to confirmed this directly in some way, either using p-values or z-scores, or the Bayesian method based on density ratios mentioned above]. Therefore, there is evidence for a second-order interaction beyond the pairwise level for these residues. However, this interaction is not essential to generate the probabilistic dependencies observed above, since a factorization of the maximum entropy distribution along the lines of Table 2.4 shows qualitatively very similar features as the observed distribution (Table 2.5). This is in agreement with recent work by Bialek et al. which emphasizes how purely pairwise interactions can give rise to strong higher-level probabilistic dependencies [98].

While there is evidence for both probabilistic dependence and second-
Figure 2.2: Triplet mutual information $I_3(A, B, C)$ as a function of $\alpha$ for the marginals derived from the 46-54-90 triplet. The circle represents the observed triplet distribution.
Table 2.5: Conditional probabilities for the maximum entropy distribution $\tilde{P}^{(2)}(A, B, C)$ for residues 46-54-90.

| $P(A_m) = 0.1404$ | $P(C_m|A_m) = 0.6061$ | $P(C_m|A_0) = 0.1563$ |
|-------------------|---------------------|---------------------|
| $P(B_m|A_mC_m) = 0.5083$ | $P(B_m|A_mC_0) = 0.1839$ | $P(B_m|A_0C_m) = 0.1980$ | $P(B_m|A_0C_0) = \ldots$ |

order interactions for the 46, 54, 90 triple, the fact that the observed value of $I_3(A, B, C)$ is close to zero indicates that there is neither strong redundancy nor synergy among these three residues. [I have some ideas for how to visualize this, but it’s not ready yet]

As another example displaying qualitatively different features, we next consider the triple of residues 24, 54, and 82. All three of these residues are sites of known drug-resistance mutations [20]. The data for this triple are shown in Table 2.6, and a general factorization into conditionals is shown in Table 2.7. The results are similar to those seen for the 46-54-90 triple: the probability of a mutation at residue 82 is larger if residue 24 is mutated, and the probability of a mutation at residue 54 is smaller if both 24 and 82 are wild-type. The Bayesian model selection calculation also indicates that the Class E model has the largest support, with the next-best model less likely by a factor of $\approx 10^{34}$. However, from the plot of $I_3(A, B, C)$ for the constant-marginal family (Figure), we see that there is a much smaller difference between the observed and maximum entropy distributions, indicating that there is little evidence for second-order interactions. The observed value of $I_3(A, B, C) = 0.0277$ is somewhat larger compared to the 46-54-90 triple, and is a larger percentage of the maximal possible value given the univariate marginals (0.1473). This indicates that while there are no second order interactions, there is some degree of redundancy among the three variables. [Again, this needs to be made more concrete]

Overall, we examined a set of 88,560 residue triples from HIV protease from all of these perspectives. Of these, 4910 triples showed very strong statisti-
Table 2.6: Contingency table for the triple of residues 24, 54, and 82 of HIV protease.

<table>
<thead>
<tr>
<th>residue 24 mutated</th>
<th>residue 24 wild-type</th>
<th>residue 54 mutated</th>
<th>residue 54 wild-type</th>
<th>residue 54 mutated</th>
<th>residue 54 wild-type</th>
<th>residue 82 mutated</th>
<th>residue 82 wild-type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>633</td>
<td>168</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3039</td>
<td>1972</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>28</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1467</td>
<td>36667</td>
</tr>
</tbody>
</table>

Table 2.7: Conditional probabilities for the observed distribution $P(A, B, C)$ for residues 24, 54, and 82.

\[
P(A_m) = 0.0211 \quad P(C_m|A_0) = 0.1161 \quad P(C_m|A_m) = 0.8631 \quad P(B_m|A_0C_m) = 0.2207 \quad P(B_m|A_mC_0) = 0.6065 \quad P(B_m|A_0C_0) = 0.7902
\]

Table 2.8: The distribution of factorization classes for unambiguous model selections.

<table>
<thead>
<tr>
<th>odds margin</th>
<th>factorization class</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^3-10^{10}$</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>$&gt;10^{10}$</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 2.9: The 20 triples with the least ambiguous factorization model assignment. All cases shown here belong to Class E (Equation 2.12).

<table>
<thead>
<tr>
<th>residue triple</th>
<th>$\log_{10}(\text{odds margin})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>46_54_90</td>
<td>696.3</td>
</tr>
<tr>
<td>10_46_90</td>
<td>477.2</td>
</tr>
<tr>
<td>10_54_71</td>
<td>457.0</td>
</tr>
<tr>
<td>46_82_90</td>
<td>423.4</td>
</tr>
<tr>
<td>10_71_90</td>
<td>423.1</td>
</tr>
<tr>
<td>10_46_82</td>
<td>402.4</td>
</tr>
<tr>
<td>46_71_84</td>
<td>402.3</td>
</tr>
<tr>
<td>46_54_71</td>
<td>381.8</td>
</tr>
<tr>
<td>10_84_90</td>
<td>347.4</td>
</tr>
<tr>
<td>20_46_54</td>
<td>331.5</td>
</tr>
<tr>
<td>46_71_73</td>
<td>328.3</td>
</tr>
<tr>
<td>46_84_90</td>
<td>327.1</td>
</tr>
<tr>
<td>46_54_84</td>
<td>324.6</td>
</tr>
<tr>
<td>46_71_82</td>
<td>322.6</td>
</tr>
<tr>
<td>46_73_84</td>
<td>322.4</td>
</tr>
<tr>
<td>10_71_82</td>
<td>304.8</td>
</tr>
<tr>
<td>20_46_82</td>
<td>303.4</td>
</tr>
<tr>
<td>46_71_90</td>
<td>284.6</td>
</tr>
<tr>
<td>10_71_73</td>
<td>273.2</td>
</tr>
<tr>
<td>71_73_84</td>
<td>267.7</td>
</tr>
</tbody>
</table>
Figure 2.3: Triplet mutual information \( I_3(A, B, C) \) as a function of \( \alpha \) for the marginals derived from the 24-54-82 triplet. The circle represents the observed triplet distribution.

We explore this question by examining the 20 triples with the largest \( I_3^{(3)} \) values (Table 2.10). We see that, in general, \( I_3^{(3)} \) is less than \( I_3^{(2)} \), sometimes substantially so (e.g. 54-82-90 and 54-82-84). This suggests that the role of second-order effects is relatively small, and that most of the strong observed probabilistic dependency can be explained by “pairwise energy” effects. [we should continue to discuss this to make sure the argument makes sense]. Interestingly, all of the residue triples in Table 2.10 have overwhelming evidence in favor of the Class E factorization, with the exception of 10-30-71, for which the model selection is ambiguous with respect to Class D and Class E. [we also need to determine for which, if any, of these the \( \lambda_{ABC} \neq 0 \) is statistically significant].

We have also examined the triples having the largest triplet mutual information relative to the maximum possible value given the univariate marginals (Table 2.11). The vast majority of these have positive \( I_3 \) value, meaning that
Figure 2.4: Predicted vs observed quartet probabilities for 300 residue quadruplets in HIV protease.

despite the probabilistic dependencies (most are Class E) there is some degree of informational redundancy inherent in the distribution. However, the most interesting are those with the negative $I_3$ values, particularly 48-82-94 and 10-48-51, which have strong evidence in favor of a Class D factorization, indicating significant synergy. Furthermore, all of the 616 Class D triples from Table 2.8 also have statistically significant synergistic effects.

To gain further insight into the importance of three-body terms, we constructed 4-way $2 \times 2 \times 2 \times 2$ contingency tables for protease. For 300 of these with the largest overall cell counts, we constructed three different approximations to the joint probability, one under the assumption of independence, the second using log-linear models with terms up to pairwise, and the third using terms up to three-body. The results are shown in Figure
2.3.3 Sequence conservation and the properties of correlation measures

There are various methods of measuring correlation between covarying residues in multiple sequence alignments. Several of these have been highlighted in Subsection 2.2.1. These properties of these measures vary and each of them has its own strengths and weakness. In order to understand the measures better, we investigated their properties. The following is a contingency
Table 2.10: The 20 triples with the largest $I_C^{(3)}(A,B,C)$ values.

<table>
<thead>
<tr>
<th>residue triple</th>
<th>$I_{multi}(A,B,C)$</th>
<th>$I_C^{(3)}(A,B,C)$</th>
<th>$I_C^{(2)}(A,B,C)$</th>
<th>$I_2(A,B)$</th>
<th>$I_2(A,C)$</th>
<th>$I_2(B,C)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>46_54_90</td>
<td>0.20436</td>
<td>0.02837</td>
<td>0.17599</td>
<td>0.05893</td>
<td>0.08576</td>
<td>0.06571</td>
</tr>
<tr>
<td>46_82_90</td>
<td>0.19778</td>
<td>0.02712</td>
<td>0.17066</td>
<td>0.07983</td>
<td>0.08540</td>
<td>0.03219</td>
</tr>
<tr>
<td>82_84_90</td>
<td>0.12002</td>
<td>0.01834</td>
<td>0.10167</td>
<td>0.00356</td>
<td>0.03196</td>
<td>0.06971</td>
</tr>
<tr>
<td>54_82_90</td>
<td>0.26960</td>
<td>0.01468</td>
<td>0.25492</td>
<td>0.18942</td>
<td>0.06505</td>
<td>0.03187</td>
</tr>
<tr>
<td>35_36_90</td>
<td>0.10383</td>
<td>0.01436</td>
<td>0.08946</td>
<td>0.07484</td>
<td>0.00415</td>
<td>0.01396</td>
</tr>
<tr>
<td>20_46_82</td>
<td>0.13403</td>
<td>0.01428</td>
<td>0.11974</td>
<td>0.03013</td>
<td>0.02498</td>
<td>0.08042</td>
</tr>
<tr>
<td>46_54_84</td>
<td>0.12776</td>
<td>0.01382</td>
<td>0.11394</td>
<td>0.05885</td>
<td>0.04430</td>
<td>0.02981</td>
</tr>
<tr>
<td>20_46_54</td>
<td>0.12622</td>
<td>0.01360</td>
<td>0.11262</td>
<td>0.03001</td>
<td>0.04108</td>
<td>0.05952</td>
</tr>
<tr>
<td>54_82_84</td>
<td>0.23624</td>
<td>0.01200</td>
<td>0.22424</td>
<td>0.18935</td>
<td>0.02977</td>
<td>0.00349</td>
</tr>
<tr>
<td>30_35_36</td>
<td>0.09905</td>
<td>0.01172</td>
<td>0.08733</td>
<td>0.00548</td>
<td>0.01073</td>
<td>0.07498</td>
</tr>
<tr>
<td>46_84_90</td>
<td>0.17939</td>
<td>0.01131</td>
<td>0.16808</td>
<td>0.04406</td>
<td>0.08528</td>
<td>0.06920</td>
</tr>
<tr>
<td>10_71_88</td>
<td>0.12843</td>
<td>0.01075</td>
<td>0.11768</td>
<td>0.10190</td>
<td>0.00091</td>
<td>0.01532</td>
</tr>
<tr>
<td>54_84_90</td>
<td>0.15248</td>
<td>0.01067</td>
<td>0.14181</td>
<td>0.02951</td>
<td>0.06508</td>
<td>0.06907</td>
</tr>
<tr>
<td>46_54_73</td>
<td>0.11034</td>
<td>0.01065</td>
<td>0.09969</td>
<td>0.05954</td>
<td>0.03563</td>
<td>0.01730</td>
</tr>
<tr>
<td>35_36_88</td>
<td>0.09279</td>
<td>0.01051</td>
<td>0.08229</td>
<td>0.07458</td>
<td>0.00327</td>
<td>0.00678</td>
</tr>
<tr>
<td>10_30_71</td>
<td>0.12474</td>
<td>0.01033</td>
<td>0.11441</td>
<td>0.00027</td>
<td>0.10198</td>
<td>0.00899</td>
</tr>
<tr>
<td>46_71_90</td>
<td>0.24616</td>
<td>0.01030</td>
<td>0.23585</td>
<td>0.05373</td>
<td>0.08574</td>
<td>0.13841</td>
</tr>
<tr>
<td>10_71_93</td>
<td>0.16304</td>
<td>0.00977</td>
<td>0.15327</td>
<td>0.10266</td>
<td>0.02702</td>
<td>0.04203</td>
</tr>
<tr>
<td>10_46_90</td>
<td>0.27272</td>
<td>0.00943</td>
<td>0.26329</td>
<td>0.10588</td>
<td>0.12817</td>
<td>0.08631</td>
</tr>
<tr>
<td>46_71_84</td>
<td>0.13909</td>
<td>0.00920</td>
<td>0.12989</td>
<td>0.05315</td>
<td>0.04420</td>
<td>0.05515</td>
</tr>
</tbody>
</table>
Table 2.11: Triples with “large” $I_3(A, B, C)$ values. Triples with zero or very small contingency table cell counts are not shown.

| residue triple | $I_3(A, B, C)$ | max $|I_3(A, B, C)|$ | ratio |
|----------------|----------------|----------------------|-------|
| 24_54_82       | 0.0277         | 0.1473               | 0.1881|
| 48_54_82       | 0.0241         | 0.1349               | 0.1787|
| 10_54_82       | 0.0887         | 0.5179               | 0.1713|
| 24_46_82       | 0.0227         | 0.1466               | 0.1547|
| 10_24_82       | 0.0216         | 0.1464               | 0.1476|
| 10_24_54       | 0.0203         | 0.1464               | 0.1384|
| 10_73_90       | 0.0477         | 0.3463               | 0.1378|
| 32_46_47       | 0.0121         | 0.0895               | 0.1347|
| 10_24_46       | 0.0191         | 0.1459               | 0.1308|
| 10_48_82       | 0.0171         | 0.1359               | 0.1258|
| 10_54_71       | 0.0626         | 0.5194               | 0.1206|
| 48_82_94       | -0.0010        | 0.0088               | 0.1197|
| 54_71_82       | 0.0590         | 0.5209               | 0.1132|
| 24_46_90       | -0.0166        | 0.1468               | 0.1132|
| 10_84_90       | 0.0412         | 0.3643               | 0.1130|
| 24_46_54       | 0.0165         | 0.1467               | 0.1126|
| 10_48_54       | 0.0147         | 0.1356               | 0.1082|
| 10_48_51       | -0.0013        | 0.0122               | 0.1080|
| 24_54_90       | -0.0154        | 0.1474               | 0.1042|
| 32_46_82       | 0.0153         | 0.1471               | 0.1038|

Table 2.12: Contingency table for two random variables with probabilities instead of counts

<table>
<thead>
<tr>
<th></th>
<th>$Y = W$</th>
<th>$Y = M$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = W$</td>
<td>$P_{X,Y}(W,W)$</td>
<td>$P_{X,Y}(W,M)$</td>
<td>$P_X(W)$</td>
</tr>
<tr>
<td>$X = M$</td>
<td>$P_{X,Y}(M,W)$</td>
<td>$P_{X,Y}(M,M)$</td>
<td>$P_X(M)$</td>
</tr>
<tr>
<td>Total</td>
<td>$P_Y(W)$</td>
<td>$P_Y(M)$</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 2.13: Contingency table for two completely anti-correlated random variables with equal univariate marginals of 0.50

<table>
<thead>
<tr>
<th></th>
<th>$Y = W$</th>
<th>$Y = M$</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = W$</td>
<td>0.0</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>$X = M$</td>
<td>0.50</td>
<td>0.0</td>
<td>0.50</td>
</tr>
<tr>
<td>total</td>
<td>0.50</td>
<td>0.50</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 2.14: Contingency table for two completely correlated random variables with equal univariate marginals of 0.50

<table>
<thead>
<tr>
<th></th>
<th>$Y = W$</th>
<th>$Y = M$</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = W$</td>
<td>0.50</td>
<td>0.0</td>
<td>0.50</td>
</tr>
<tr>
<td>$X = M$</td>
<td>0.0</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>total</td>
<td>0.50</td>
<td>0.50</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Table 2.15: Contingency table for two independent random variables with equal univariate marginals of 0.50

<table>
<thead>
<tr>
<th></th>
<th>$Y = W$</th>
<th>$Y = M$</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = W$</td>
<td>0.25</td>
<td>0.25</td>
<td>0.50</td>
</tr>
<tr>
<td>$X = M$</td>
<td>0.25</td>
<td>0.25</td>
<td>0.50</td>
</tr>
<tr>
<td>total</td>
<td>0.50</td>
<td>0.50</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 2.16: Correlation measures for different values of the univariate marginals

<table>
<thead>
<tr>
<th>$P_X(M)$</th>
<th>$P_Y(M)$</th>
<th>$P_{X,Y}(M, M)$</th>
<th>$MI(X, Y)$</th>
<th>$Cov(X, Y)$</th>
<th>$\phi(X, Y)$</th>
<th>$\chi(X, Y)$</th>
<th>$\rho(X, Y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
<td>0.25</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.97</td>
<td>0.24</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.88</td>
<td>0.21</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.72</td>
<td>0.16</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.47</td>
<td>0.09</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 2.16: Correlation measures for different values of the univariate marginals
2.3.4 Correlated mutation analysis of HIV Protease

Analysis of drug annotated mutations in HIV Protease indicates that the frequency of mutations increases as drugs are used.

![HomologieAlign.png](attachment:HomologieAlign.png)

**Figure 2.5:** An alignment of HIV protease subtype B sequences extracted from patients not exposed to any drugs. Dashes symbolize the wildtype amino acid, the sequence of which is listed at the top for reference.

![HomologieAlign2.png](attachment:HomologieAlign2.png)

**Figure 2.6:** An alignment of HIV protease subtype B sequences extracted from patients exposed to six drugs. Dashes symbolize the wildtype amino acid, the sequence of which is listed at the top for reference.

2.3.5 Pair correlations and the structure of HIV protease

We begin by investigating pair correlations in the larger Lee database and their association with the structure of the HIV protease. Several groups have studied pair correlations as a means of identifying functionally dependent residues in the HIV protease and other systems. [93, 19, 49, 122, 114] Since our dataset
of 45,161 protease sequences is unannotated, it was important to qualitatively match prior studies that employed smaller datasets with known drug treatment profiles. For each of the \( \binom{82}{2} = 3,321 \) pairs of positions, a \( 2 \times 2 \) contingency table was constructed and the \( \phi \) correlation coefficient (Equation 2.30) was calculated. Of the 3,321 \( \phi \) values, 246 were considered to be statistically significant (\(|\phi| > 0.1, |Z| > 5.0\), where \( Z \) is the number of standard deviations from the mean under the null hypothesis of no correlation). Our correlations match well with those observed previously, with a Spearman rank order correlation coefficient of 0.75 between the \( \phi \) values for our top 22 correlations and those of Wu, et al.[122] (data not shown). Furthermore, we observe larger \( \phi \) values for pairs of drug-associated positions compared to non-drug associated pairs (Figure 2.12), which is consistent with previous observations.[122]

Previous studies of pair correlations in protein families have indicated that coevolving pairs of residues tend to lie closer to each other in structure than random residue pairs.[87, 109, 37, 32, 84] In the case of the HIV protease the distribution of \( \phi \) values for drug-associated positions with distance shows a characteristic triangular shape (Figure 2.12)[34]. In particular, distances associated with the most correlated residue pairs (30–88, 54–82, 32–47, and 37–77) are all within a few Å of each other. Furthermore, the difference in the distribution of \( \phi \) values for pairs with distance > 10 Å and < 10 Å is highly statistically significant (\( p < 10^{-9} \) using the non-parametric Mann-Whitney test[24]). The triangular shape in Figure 2.12 is not surprising, given that most of the drug-associated positions are on the substrate cleft and thus tend to be relatively close in space. As a result, 89 pairs (36% of the statistically significant pairs) are within 8 Å of each other. As far as long-range interactions go, we do not observe substantial correlation between residues further than 22 Å apart. Non-drug associated positions, by contrast, display little correlation at any distance (Figure 2.12, black dots).
It is interesting to note that of the 8 significantly negatively correlated pairs, none involve a primary position. The positions involved in these negatively correlated pairs are the polymorphic positions 13, 15, 35, 36, 63, 64, 71, 77, and 93. Of these, 15 and 64 are not associated with drug treatment, while the rest play accessory roles in drug resistance[74, 85, 86]. There are no statistically significant negative correlations between primary and accessory mutations.

In contrast to the negatively correlated set, the 8 pairs with the largest positive correlations consist of 4 primary-accessory and 4 primary-primary pairs. Further breaking down the set of primary positions into residues within the cleft or away from the cleft (non-active site residues are defined as having no interaction with drugs) does not lead to any observable trends. We also see no tendency for exposed residues to preferentially coevolve, in contrast to previous studies on mutation covariation in other protein families. [87, 37]

### 2.3.6 Exhaustive analysis of residue triples

#### 2.3.6.1 Unannotated Lee database

The connected information $I_c^{(3)}$ was calculated for all $82 \choose 3 = 88,560$ residue triples in the HIV protease by solving Equation 2.36 as described above using counts of mutated residues from the Lee database. We find that 2246 triples have $I_c^{(3)} > 0.0002$, and that these are significantly enriched in drug resistance associated positions: 77% of the set of 2246 triples with high $I_c^{(3)}$ consist of only drug resistance associated positions, compared to only 16% of the full set of 88,560. The 10 triples with the largest $I_c^{(3)}$ values are shown in Table 2.18, along with the value of the “multi-information” or the Kullback-Leibler divergence between the observed distribution and the prediction based on an independent model. Residues from the largest $I_c^{(3)}$ triple (46–54–90) are also displayed on the structure of the HIV protease (Figure 2.13). In Table 2.19 we show triples with
very little three-body interaction: we selected those triples with $I^{(3)}_c < 10^{-4}$ and then chose the 10 from that subset that had the largest $I_{multi}$ in order to eliminate the trivial cases where $I^{(3)}_c$ is small simply because the independent model already fits the data very well. Strikingly, virtually all of the residues that appear in Tables 2.18 and 2.19 are associated with drug resistance. The 10 large $I^{(3)}_c$ triples contain only 9 of the 45 drug resistance associated residues (20, 30, 35, 36, 46, 54, 82, 84, and 90), of which 6 are primary and 3 are accessory positions. The 10 small-$I^{(3)}_c$ triples contain 13 residues (10, 13, 30, 35, 43, 54, 62, 63, 71, 73, 77, 82, and 88), of which only 5 are primary positions, and one of which (residue 62) is non-drug associated (though it was initially identified as treatment-associated by Wu, et al.[122]). A similar pattern can be seen in the analysis of 400 largest $I^{(3)}_c$ triples and 400 triples with small $I^{(3)}_c$ but large $I_{multi}$ – we find that primary positions are somewhat more common in the large $I^{(3)}_c$ set (Figure 2.7).

Residue positions which on average have larger pair correlations tend to have larger average $I^{(3)}_c$ values, and vice versa (Figure 2.14). This is understandable, as we would not expect significant three-body interactions in the absence of pair correlations. The average trend does not, however, imply that all of the $I^{(3)}_c$ values associated with correlated residue pairs are large. This can be seen more clearly if we plot $\phi$ and $I^{(3)}_c$ values for all triples associated with a given residue position (Figure 2.8).

In order to obtain a feeling for the magnitude of the three-body interactions, let us look at the triple with the largest $I^{(3)}_c$ value (46–54–90). In Figure 2.15A we show the correlation between the predicted and observed probabilities for each of the 8 binary states for the independent model (red) and the two-body fit (black) for this triple. It is clear that the independent model fits the data poorly: the probability that all three residues are mutated is underestimated by an order of magnitude, and some of the others are over- or underestimated by
Figure 2.7: Histograms of the prevalence of primary drug-resistance positions among large and small $I_c^{(3)}$ triples. The red histogram corresponds to the 400 triples with the largest $I_c^{(3)}$ values, while the black histogram corresponds to the 400 triples with $I_c^{(3)} < 10^{-4}$ and the largest $I_{multi}$.

Figure 2.8: Distribution of pair correlations ($\phi$, panel A) and three-body interactions ($I_c^{(3)}$, panel B) for all pairs and triples (respectively) containing a given residue. The largest $I_c^{(3)}$ values are not distributed uniformly among all residues: residues 46, 54, 82 and 90 have the largest $I_c^{(3)}$ values (> 0.015), while intermediate values between 0.015 and 0.005 occur for residues 10, 20, 30, 32, 35, 36, 71, 73, 74, 77, 84, 88, and 93.
factors of 2 or 3. Although the two-body fit does a better job of reproducing the observed probabilities, it still leads to noticeable deviations from the straight line for the 46–54–90 triple. In contrast, the two-body fit makes very accurate predictions in the case of the two mutational clusters previously identified by Wu, et al.[122] (Figure 2.15B,C).

One way in which the magnitude of the effect of the three-body interactions on mutational patterns can be visualized is by considering conditional probabilities for the 46–54–90 triple. We find clear evidence of “triplet correlation” in the data: \( P(54_m|46_m90_m) = 0.3505 \), \( P(54_m|46_m90_0) = 0.4265 \), \( P(54_m|46_090_m) = 0.2979 \), and \( P(54_m|46_090_0) = 0.0326 \), as compared to the independent estimate \( P(54_m) = 0.1171 \). Thus the probability that residue 54 is mutated depends very strongly on the state of both of the other two residues.

Much of this probabilistic dependency, however, can be accounted for by pair interactions. The corresponding probabilities for the best-fit two-body model are \( \hat{P}^{(2)}(54_m|46_m90_m) = 0.5083 \), \( \hat{P}^{(2)}(54_m|46_m90_0) = 0.1839 \), \( \hat{P}^{(2)}(54_m|46_090_m) = 0.1980 \), and \( \hat{P}^{(2)}(54_m|46_090_0) = 0.0511 \). Even though the model contains no three-body interactions, a qualitative “triplet correlation” remains, indicating that non-trivial three-way probabilistic dependencies can arise purely from pair correlations[98]. This can also be seen in the numerical magnitude of \( I^{(3)}_c \) and \( I_{multi} \): if we imagine \( \hat{P}^{(2)}(A, B, C) \) as lying “in between” the independent model and the observed distribution in terms of information entropy, and we define \( I^{(2)}_c(A, B, C) \) as \( S[P(A)P(B)P(C)] - S[\hat{P}^{(2)}(A, B, C)] \), then \( I^{(2)}_c + I^{(3)}_c = I_{multi} \), and \( I^{(3)}_c \) can be interpreted as the part of \( I_{multi} \) that is explained by three-body interaction[99]. Since \( I^{(3)}_c \) is an order of magnitude smaller than \( I_{multi} \) even for the triples in Table 2.18, the three-body term plays a secondary role to the pair interactions.
2.3.6.2 Annotated Shafer database

A similar analysis was performed using amino acid counts from the Shafer database. We divided the Shafer database into three classes based on how many drugs the patients had been exposed to: PI-naive, PI monotherapy, and 2-6 PIs. For each class, we found the 100 residue triples with the largest $I_c^{(3)}$ (Figure 2.16A–C). It is clear that treating patients with PIs not only increases the overall number of mutations (Figure 2.10), but also increases the importance of three-body interactions. This is also evident from the comparison of the observed frequencies of triple mutants in the Shafer database with predictions of the pair-term model (Figure 2.9). The pair model is much less accurate with the sequences obtained from patients who had been treated with two or more inhibitors. Reassuringly, comparison of the $I_c^{(3)}$ distributions for the complete Lee and Shafer databases shows that the two databases have quantitatively similar behavior when considered in aggregate (Figure 2.16D,E). In addition, many of the top-100 residue triples found for the Lee and the complete Shafer database overlap significantly: 67 of the top-100 residue triples found for the Lee database are also found among the top-100 triples from the complete Shafer database. Moreover, of the top-100 Lee database triples, 12, 23, and 40 triplets overlap with the top-100 triples from the PI-naive, PI monotherapy, and 2-6 PIs data respectively. This is strikingly similar to the overlap between the top-100 triplets from the complete Shafer dataset and the PI-naive, PI monotherapy, and 2-6 PIs data, which have an overlap of 10, 20 and 39 triplets respectively.

2.3.7 Higher-order interactions in clusters of $n > 3$ residues

Exposure to PIs may induce a large-scale mutational response in protease that goes well beyond the triplet level, making it necessary to study higher-order interactions in larger residue groups. The largest cluster of residues for which the
Figure 2.9: Predicted vs observed probabilities for a triple mutant for all triples of drug-associated positions in the HIV protease using the Shafer database of sequences obtained from drug-naive patients (A) and those treated with two or more inhibitors (B). Mutational states that were unobserved in the database and would have an observed probability maximum likelihood estimate of zero are not shown. The dots correspond to the best fit pair-term model, and the red lines of slope 1 correspond to the perfect agreement of the predicted probabilities with those observed, which would be obtained if the three-body interactions were included.
exact calculation is computationally feasible via the exact methods used here is $n = 10$. However, it is not practical to enumerate all $\binom{41}{10} \approx 10^9$ 10-residue groups. Instead, we restrict our attention to the three key primary drug resistance positions 30, 82 and 90, and the known accessory positions associated with them. This reduced the search space to approximately 9,000 10-residue groups. Using the Lee database, we searched for the 10-residue group for which the pair model had the largest deviation from the observed probabilities. We found that the group 20–32–46–48–53–54–58–74–82–90 has the most prominent contribution from higher-order interactions. We observe the same general pattern as for the 46–54–90 triple: there are strong pair interactions which bring observed and predicted probabilities into qualitative, order-of-magnitude agreement, weaker three-body interactions which further improve the agreement, and very weak four-body and higher interactions which have quantitative impact on a small number of state probabilities (Figure 2.17).

We also studied the distribution of the total number of mutations in the 10-residue group 20–32–46–48–53–54–58–74–82–90. Appropriate subsets of the predicted state probabilities were summed to obtain the total number of mutated residues for the independent, two-body, and three-body models. The results are shown in Figure 2.18 and Table 2.17. As expected, the distribution for the independent model is very different from the observed distribution: the probability of having no mutations is considerably underestimated, and the upper tail is much too thin (predicting no more than 4 or so mutations). Adding pair terms greatly improves the “no mutation” probability and considerably extends the length of the tail. The tail length is further modulated by the addition of the three-body interactions, bringing the distribution very close to the observed probabilities.

In order to investigate the impact of higher-order interactions on the full set of 41 drug-associated positions, we made use of an approximate fitting
Table 2.17: Total number of observed and predicted mutations for the 10 residue group 20–32–46–48–53–54–58–74–82–90. The total sample size is 41,668 sequences for which unambiguous codons were present at all 10 positions (see Methods).

<table>
<thead>
<tr>
<th>number of mutations</th>
<th>independent model</th>
<th>pair-term model</th>
<th>pair+three-body model</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>16,503</td>
<td>24,933</td>
<td>26,674</td>
</tr>
<tr>
<td>1</td>
<td>16,508</td>
<td>8,150</td>
<td>5,617</td>
</tr>
<tr>
<td>2</td>
<td>6,873</td>
<td>3,771</td>
<td>3,115</td>
</tr>
<tr>
<td>3</td>
<td>1,557</td>
<td>1,949</td>
<td>2,789</td>
</tr>
<tr>
<td>4</td>
<td>210</td>
<td>1,162</td>
<td>2,100</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>860</td>
<td>1,049</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>580</td>
<td>278</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>216</td>
<td>42</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>42</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

procedure in which we assumed that binary combinations (from among the $2^{41}$ possible) that were not observed in the Lee database are in fact exactly zero and will never be observed no matter how large a future database may become. In other words, we make the probabilities of unobserved states “structural zeros” in the model[10]. This greatly reduces the computational burden, since the normalization calculation needed to determine $\lambda_0$ in Equation 2.37 now requires a summation over at most $\approx 45,000$ terms (the number of sequences in the Lee database), instead of the $2^{41}$ terms needed for the exact model. The actual number of terms is somewhat smaller, since some of the $2^{41}$ potential binary states occur more than once, resulting in a total of 8331 populated states.

The fit of one-body and two-body parameters $\lambda_i$ and $\lambda_{ij}$ while enforcing the structural zeros was performed by nonlinear optimization as before, and the resulting distribution of the total mutation counts is shown in Figure 2.19. As for the 10-residue case, the qualitative distribution is well accounted for by the pair interaction model, with the three-body and higher interactions modulating the details of the shape of the distribution, such as the increase in frequency of 3–5 and 13–15 mutations, at the expense of a decrease in the frequency of 6–10
2.3.8 Relationship between triplet conditional probabilities and the Kirkwood approximation

Consider the joint distribution of three random variables $X$, $Y$, and $Z$ (that may take on a discrete number of values, e.g. $X \in \{x_1, x_2, \ldots\}$, though the derivation below applies equally well to continuous state spaces), with a joint probability distribution $P(X, Y, Z)$. In general, we can rewrite the joint probability as

$$P(X, Y, Z) = P(X)P(Y|X)P(Z|X, Y).$$  \hspace{1cm} (2.39)

Analogous expressions can be written for any permutation of $X$, $Y$, and $Z$ as well. This most general form contains “three-body interactions” via the last factor, in which the probability of $Z$ depends on both $X$ and $Y$. A joint distribution which has only “two-body interactions” can be written by eliminating
Table 2.18: The 10 triples in the HIV protease with the largest $I_C^{(3)}(A, B, C)$ values.

<table>
<thead>
<tr>
<th>residue triple</th>
<th>$I_{multi}(A, B, C)$</th>
<th>$I_C^{(3)}(A, B, C)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>46–54–90</td>
<td>0.20436</td>
<td>0.02837</td>
</tr>
<tr>
<td>46–82–90</td>
<td>0.19778</td>
<td>0.02712</td>
</tr>
<tr>
<td>82–84–90</td>
<td>0.12002</td>
<td>0.01834</td>
</tr>
<tr>
<td>54–82–90</td>
<td>0.26960</td>
<td>0.01468</td>
</tr>
<tr>
<td>35–36–90</td>
<td>0.10383</td>
<td>0.01436</td>
</tr>
<tr>
<td>20–46–82</td>
<td>0.13403</td>
<td>0.01428</td>
</tr>
<tr>
<td>46–54–84</td>
<td>0.12776</td>
<td>0.01382</td>
</tr>
<tr>
<td>20–46–54</td>
<td>0.12622</td>
<td>0.01360</td>
</tr>
<tr>
<td>54–82–84</td>
<td>0.23624</td>
<td>0.01200</td>
</tr>
<tr>
<td>30–35–36</td>
<td>0.09905</td>
<td>0.01172</td>
</tr>
</tbody>
</table>

Table 2.19: The 10 triples in the HIV protease with small $I_C^{(3)}(A, B, C)$ values but large $I_{multi}(A, B, C)$.

<table>
<thead>
<tr>
<th>residue triple</th>
<th>$I_{multi}(A, B, C)$</th>
<th>$I_C^{(3)}(A, B, C)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>30–71–88</td>
<td>0.1423</td>
<td>$1.89 \times 10^{-1}$</td>
</tr>
<tr>
<td>43–54–82</td>
<td>0.1416</td>
<td>$3.23 \times 10^{-6}$</td>
</tr>
<tr>
<td>13–30–88</td>
<td>0.1458</td>
<td>$2.06 \times 10^{-5}$</td>
</tr>
<tr>
<td>30–63–88</td>
<td>0.1380</td>
<td>$2.33 \times 10^{-5}$</td>
</tr>
<tr>
<td>10–71–82</td>
<td>0.1625</td>
<td>$4.25 \times 10^{-5}$</td>
</tr>
<tr>
<td>30–73–88</td>
<td>0.1373</td>
<td>$4.92 \times 10^{-5}$</td>
</tr>
<tr>
<td>30–82–88</td>
<td>0.1389</td>
<td>$6.15 \times 10^{-5}$</td>
</tr>
<tr>
<td>30–35–88</td>
<td>0.1364</td>
<td>$7.52 \times 10^{-5}$</td>
</tr>
<tr>
<td>54–62–82</td>
<td>0.1452</td>
<td>$7.57 \times 10^{-5}$</td>
</tr>
<tr>
<td>30–77–88</td>
<td>0.1365</td>
<td>$9.64 \times 10^{-5}$</td>
</tr>
</tbody>
</table>
Figure 2.11: Difference in the frequency of mutated residues between patients treated with 5 PIs and PI-naive patients for all positions in the HIV protease. Red bars correspond to primary drug resistance positions, black bars are accessory drug resistance positions, and gray bars are positions not associated with drug treatment.
Figure 2.12: Scatter plot of distance vs $\phi$ statistic for drug associated positions in red and non-drug associated positions in black. Pairs with cell counts of 20 or less for $A_m B_0$ and $A_0 B_m$ were excluded. The value along the $y$-axis is the closest distance between any two heavy atoms of the two residues based on the crystal structure of wild type protease (PDB ID 1PRO).
Figure 2.13: Structure of the HIV protease dimer (PDB code 1HN0). The catalytic triad (residues 25, 26, and 27) is highlighted in green, while the residues from the triple with the largest $I_3^{(R)}$ (46–54–90, see Figure 2.15A) are shown in red. Residue 90 is close to the catalytic triad, whereas residues 46 and 54 are close to each other, with the smallest distance between heavy atoms of 3.11 Å.)
Figure 2.14: Correlation between $I_c^{(3)}$ and $\phi$ values. Each point corresponds to one of the 45 drug-associated residue positions. The abscissa (ordinate) correspond to the average of $\phi (I_c^{(3)})$ over all pairs (triples) containing that residue position.
Figure 2.15: Predicted vs observed probabilities for the residue triple in HIV protease with the largest $I_{c}^{(3)}$ (46–54–90) (A), as well as two three-residue mutational clusters previously identified by Wu, et al.[122]: 30–82–88 (B), and 10–46–90 (C). The black dots correspond to the best-fit independent model, while the red dots correspond to the best-fit pair-term model (Equation 2.37). The solid line of slope 1 corresponds to perfect agreement of the predicted data with the observed, which would be obtained if the three-body term were included.
Figure 2.16: Distribution of the 100 largest $I_{c}^{(3)}$ values observed for the Shafer and Lee databases. Panels A–C show the results for subsets of the Shafer database corresponding to patients that were PI-naive (A), PI-monotherapy treated (B), and exposed to 2 or more PIs (C). Panels D and E correspond to the complete Shafer and Lee databases, respectively.
Figure 2.17: Predicted vs observed probabilities for the $2^{10}$ mutational states of the ten-residue group 20–32–46–48–53–54–58–74–82–90 in HIV protease. Mutational states that were unobserved in the database and would have an observed probability maximum likelihood estimate of zero are not shown. The black dots correspond to the best-fit independent model, the red dots correspond to the best-fit pair-term model (Equation 2.37), and the green dots correspond to the best-fit three-body model (Equation 2.38). The solid line of slope 1 corresponds to perfect agreement of the predicted data with the observed, which would be obtained if all higher-order terms were included.
Figure 2.18: Total number of observed (solid black curve) or predicted mutations for the ten-residue group 20–32–46–48–53–54–58–74–82–90 in HIV protease. The dashed curves represent the predictions for the independent (blue), pair-term (red), and pair plus three-body model (green). 41,668 sequences from the Lee database were used to train the models (sequences for which unambiguous codons were present at all 10 positions, see Methods).
Figure 2.19: Distributions of the total number of mutations in the Lee database among the 41 drug-associated positions. The black curve corresponds to the observed data, the blue curve corresponds to the model in which mutations are assume to occur independently of each other, and the red curve corresponds to the best fit using a two-body model under the “structural zero” approximation (see text). The independent model data were generated by a Monte Carlo procedure in which many sequences were generated using independent Bernoulli trials at each residue position.
Figure 2.20: Scatter plot of distance vs $\lambda_{ij}$ parameters estimated using the $\pm 1$ gauge. The value along the $y$-axis is the closest distance between any two heavy atoms of the two residues based on the crystal structure of wild type protease (PDB ID 1PRO).
this dependence. For example, if the probability of $Z$ is independent of $Y$, then we can replace the last factor with $P(Z|X)$.

The Kirkwood approximation for the joint probability is given by

$$K(X, Y, Z) = \frac{P(X, Y)P(Y, Z)P(X, Z)}{P(X)P(Y)P(Z)}.$$  \hfill (2.40)

If we make use of the fact that $P(A, B) = P(A)P(B|A)$, then we can rewrite $K(X, Y, Z)$ as

$$\frac{P(X)P(Y|X)P(Z|Y)P(Z)P(X|Z)}{P(X)P(Y)P(Z)} = P(Y|X)P(Z|Y)P(X|Z).$$  \hfill (2.41)

We wish to know under what conditions $P(X, Y, Z) = K(X, Y, Z)$. To do this, we set Equation 2.39 equal to Equation 2.41:

$$P(X)P(Y|X)P(Z|X, Y) = P(Y|X)P(Z|Y)P(X|Z)$$

or

$$P(Z|X, Y) = \frac{P(Z|Y)P(X|Z)}{P(X)}.$$  \hfill (2.42)

However, from Bayes’ Theorem we know that

$$P(X|Z) = \frac{P(Z|X)P(X)}{P(Z)},$$

so

$$P(Z|X, Y) = \frac{P(Z|Y)P(Z|X)}{P(Z)}.$$  \hfill (2.42)

Therefore, the Kirkwood approximation reproduces the joint probability if and only if Equation 2.42 is satisfied, a necessary condition of which is that the “three-body” probability $P(Z|X, Y)$ be proportional to the product of two “two-body” probabilities. One somewhat trivial case in which this is true is
if there is only one “two-body interaction”, e.g. the joint distribution can be written as $P(X,Y,Z) = P(X)P(Y,Z) = P(X)P(Y)P(Z|Y)$. Then $P(Z|X,Y)$ and $P(Z|X)$ reduce to $P(Z|Y)$ and $P(Z)$, respectively, and Equation 2.42 is satisfied.

One case in which the Kirkwood approximation may not reproduce the joint probability even there is no explicit “three-body interaction” is when, for example, both $Y$ and $Z$ are dependent on $X$, but $Y$ and $Z$ are not directly dependent on each other (i.e. $Y$ and $Z$ are conditionally independent). In that case, $P(X,Y,Z) = P(X)P(Y|X)P(Z|X)$. Repeating the above argument starting with this special case, we find that the only way that $P(X,Y,Z)$ can equal $K(X,Y,Z)$ is if $P(Z|Y) = P(Z)$. This may never be true, since the dependence of $Y$ and $Z$ on $X$ induces a correlation between them, while $P(Z|Y) = P(Z)$ implies that $Y$ and $Z$ are independent (marginally with respect to $X$).

Let us consider a numerical example. Suppose that $P(X,Y,Z) = P(X)P(Y|X)P(Z|X)$ and that the random variables are binary, e.g. $X \in \{x_0, x_m\}$ representing a wild-type and mutant at amino acid sequence position $X$. Let $P(x_m) = 0.7$, $P(y_m|x_m) = 0.8$, $P(y_m|x_0) = 0.3$, $P(z_m|x_m) = 0.9$, and $P(z_m|x_0) = 0.1$. The corresponding probabilities of $y_0$ and $z_0$ can be obtained from the normalization condition, e.g. $P(y_0|x_m) = 1 - P(y_m|x_m) = 0.2$. From these we can obtain all of the possible three-way joint probabilities, for example $P(x_0,y_m,z_0) = (1 - 0.7) \times 0.3 \times (1 - 0.1) = 0.081$. The resulting probabilities $P(X,Y,Z)$ are shown in Table 2.20.

From those probabilities, we can calculate the remaining marginal and joint distributions, e.g. $P(y_m) = 0.081 + 0.009 + 0.056 + 0.504 = 0.65$. In particular, we can obtain the bivariate marginal distributions needed to calculate $K(X,Y,Z)$ using Equation 2.40. For example, $P(x_m,y_0) = 0.014 +
Table 2.20: Actual joint probabilities versus the Kirkwood Approximation

<table>
<thead>
<tr>
<th></th>
<th>( P(X, Y, Z) )</th>
<th>( K(X, Y, Z) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_0, y_0, z_0 )</td>
<td>0.189</td>
<td>0.322</td>
</tr>
<tr>
<td>( x_0, y_0, z_m )</td>
<td>0.021</td>
<td>0.013</td>
</tr>
<tr>
<td>( x_0, y_m, z_0 )</td>
<td>0.081</td>
<td>0.050</td>
</tr>
<tr>
<td>( x_0, y_m, z_m )</td>
<td>0.009</td>
<td>0.011</td>
</tr>
<tr>
<td>( x_m, y_0, z_0 )</td>
<td>0.014</td>
<td>0.024</td>
</tr>
<tr>
<td>( x_m, y_0, z_m )</td>
<td>0.126</td>
<td>0.080</td>
</tr>
<tr>
<td>( x_m, y_m, z_0 )</td>
<td>0.056</td>
<td>0.035</td>
</tr>
<tr>
<td>( x_m, y_m, z_m )</td>
<td>0.504</td>
<td>0.603</td>
</tr>
</tbody>
</table>

Table 2.21: Actual Marginal Probabilities

| \( P(x_m) \) | \( P(y_m) \) | \( P(z_m) \) | \( P(x_0, y_0) \) | \( P(y_0, z_0) \) | \( P(x_0, z_0) \) | \( P(x_m, y_0) \) | \( P(y_m, z_0) \) | \( P(x_m, z_0) \) | \( P(x_0, y_m) \) | \( P(y_0, z_m) \) | \( P(x_0, z_m) \) | \( P(x_m, y_m) \) | \( P(y_m, z_m) \) | \( P(x_m, z_m) \) |
|--------------|--------------|--------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| 0.7          | 0.65         | 0.66         | 0.21             | 0.203            | 0.27             | 0.14             | 0.137            | 0.07             | 0.09             | 0.147            | 0.03             | 0.56             | 0.513            | 0.63             |

0.126 = 0.140. The quantities needed to calculate the Kirkwood approximation are given in Table 2.21, and the resulting values of \( K(X, Y, Z) \) are shown in Table 2.20. As can be seen in Table 2.20, the Kirkwood approximation does not reproduce the joint density, with some occasional large discrepancies, such as 0.322 vs 0.189 for \( (x_0, y_0, z_0) \). These deviations arise because neither \( P(z_m|y_0) = (0.021 + 0.126)/(0.189 + 0.021 + 0.014 + 0.126) = 0.420 \) nor \( P(z_m|y_m) = (0.009 + 0.504)/(0.0.081 + 0.009 + 0.056 + 0.504) = 0.789 \) equal \( P(z_m) = 0.66 \).
2.3.9 When is the Kirkwood approximation exact?

2.3.9.1 Theorem

The Kirkwood approximation for the joint probability of binary random variables $A$, $B$, or $C$ is exact, i.e.

$$P(A, B, C) = \frac{P(A, B)P(B, C)P(A, C)}{P(A)P(B)P(C)}$$  \hspace{1cm} (2.43)

if and only if at least one of the random variables is independent of the others.

2.3.9.2 Proof

Let us assume that there exists some joint distribution for which the Kirkwood approximation is exact. This means that Equation 2.43 must hold for all eight possible binary combinations of the variables $A$, $B$, or $C$. The most general specification of the joint distributions $P(A, B, C)$ is given by seven independent probabilities, e.g.

- $P(A_0, B_0, C_0)$,
- $P(A_0, B_0, C_m)$,
- $P(A_0, B_m, C_0)$,
- $P(A_0, B_m, C_m)$,
- $P(A_m, B_0, C_0)$,
- $P(A_m, B_0, C_m)$,
- $P(A_m, B_m, C_0)$.

Equivalently, a general joint distribution can be specified by three independent univariate marginals $P(A_0)$, $P(B_0)$, $P(C_0)$, three independent bivariate marginals, e.g.

- $P(A_0, B_0)$,
- $P(B_0, C_0)$,
- $P(A_0, C_0)$,

and one of the joint probabilities, e.g. $P(A_0, B_0, C_m)$. The remainder of the joint probabilities can be written in terms of these. For notational convenience, let $D_1 = P(A_0, B_0)$, $D_2 = P(B_0, C_0)$, $D_3 = P(A_0, C_0)$, $p_a = P(A_0)$,
\( p_b = P(B_0), p_c = P(C_0), \) and \( \alpha = P(A_0, B_0, C_m). \) Then

\[
\begin{align*}
P(A_0, B_0, C_0) &= D_1 - \alpha \\
P(A_0, B_0, C_m) &= \alpha \\
P(A_0, B_m, C_0) &= D_3 - D_1 + \alpha \\
P(A_0, B_m, C_m) &= p_a - D_3 - \alpha \\
P(A_m, B_0, C_0) &= D_2 - D_1 + \alpha \\
P(A_m, B_0, C_m) &= p_b - D_2 - \alpha \\
P(A_m, B_m, C_0) &= p_c + D_1 - D_2 - D_3 - \alpha \\
P(A_m, B_m, C_m) &= 1 - p_a - p_b + D_2 + D_3 + \alpha.
\end{align*}
\]

We can also express all of the remaining bivariate marginals in terms of the other marginals:

\[
\begin{align*}
P(A_m, B_0) &= p_b - D_1 \\
P(A_0, B_m) &= p_a - D_1 \\
P(A_m, B_m) &= 1 - p_a - p_b + D_1 \\
P(B_m, C_0) &= p_c - D_2 \\
P(B_0, C_m) &= p_b - D_2 \\
P(A_m, C_0) &= p_c - D_3 \\
P(A_0, C_m) &= p_a - D_3 \\
P(A_m, C_m) &= 1 - p_c - p_a + D_3.
\end{align*}
\]

We proceed by finding the values of \( D_1, D_2, D_3, p_a, p_b, p_c \) and \( \alpha \) for which Equation 2.45 is satisfied. We begin with the probability \( P(A_0, B_0, C_0) = \)
\( D_1 - \alpha \), which, if the Kirkwood approximation is exact, must equal

\[
D_1 - \alpha = \frac{P(A_0, B_0)P(B_0, C_0)P(A_0, C_0)}{P(A_0)P(B_0)P(C_0)} = \frac{D_1 D_2 D_3}{p_a p_b p_c}.
\]

This places a constraint on the possible values that the marginal probabilities can take on. We will choose to express this constraint by writing \( D_1 \) in terms of the other parameters:

\[
D_1 = \frac{\alpha p_a p_b p_c}{p_a p_b p_c - D_2 D_3}. \tag{2.44}
\]

Next, consider \( P(A_0, B_0, C_m) = \alpha \). Again, the requirement that Equation 2.43 be satisfied means that

\[
\alpha = \frac{P(A_0, B_0)P(B_0, C_m)P(A_0, C_m)}{P(A_0)P(B_0)[1 - P(C_0)]} = \frac{D_1 D_2 (p_a - D_3)}{p_a p_b (1 - p_c)}.
\]

Substituting Equation 2.44 and solving for \( D_2 \) we find that all \( D_3 \)-containing terms vanish and

\[
D_2 = p_b p_c. \tag{2.45}
\]

For the moment, let us skip \( P(A_0, B_m, C_0) \) and \( P(A_0, B_m, C_m) \) and move on to \( P(A_m, B_0, C_0) \). If Equation 2.43 holds, then

\[
D_2 - D_1 + \alpha = \frac{P(A_m, B_0)P(B_0, C_m)P(A_m, C_0)}{[1 - P(A_0)]P(B_0)P(C_0)} = \frac{(p_b - D_1) D_2 (p_c - D_3)}{(1 - p_a) p_b p_c}.
\]

Substituting expressions 2.44 and 2.45, we find that this leads to the constraint equation

\[
\frac{[\alpha + p_b (D_3 - p_a)](D_3 - p_a p_c)}{(D_3 - p_a)(p_a - 1)} = 0. \tag{2.46}
\]

There are two possible value of \( D_3 \) which satisfy this equation, namely \( p_a p_c \) and \( p_b^{-1} (p_a p_b - \alpha) \).
Let’s consider the $D_3 = p_a p_c$ case first. This implies that

$$D_1 = \frac{\alpha}{1 - p_c}. \quad (2.47)$$

Substituting Equations 2.45 and 2.47 and $D_3 = p_a p_c$ into the analogous constraints for $P(A_0, B_m, C_0)$, $P(A_0, B_m, C_m)$, $P(A_m, B_0, C_m)$, $P(A_m, B_m, C_0)$, and $P(A_m, B_m, C_m)$ shows that they are automatically satisfied and thus provide no further constraints on the parameters of the triplet joint distributions. Thus, the set of all triplet joint distributions subject to the constraint that Equation 2.43 is satisfied and that $D_3 = p_a p_c$ can be specified using only four parameters: $p_a$, $p_b$, $p_c$, and $\alpha$. Since the case where the random variables $A$, $B$, and $C$ are completely independent requires the three parameters $p_a$, $p_b$, and $p_c$ to describe its joint distribution, the fact that the joint distribution for the Kirkwood-consistent case requires only four parameters implies that there can be correlation between at most one pair of the random variables (since correlation between more than one pair would require additional parameters).

We can see this explicitly by noting that in the $D_3 = p_a p_c$ case the triplet probabilities are equal to

$$P(A_0, B_0, C_0) = p_a p_b - \alpha$$
$$P(A_0, B_0, C_m) = \alpha$$
$$P(A_0, B_m, C_0) = p_a (1 - p_b) - \alpha (p_b - 1)/p_b$$
$$P(A_0, B_m, C_m) = \alpha (1 - p_b)/p_b$$
$$P(A_m, B_0, C_0) = p_b (p_a - p_c) + \alpha$$
$$P(A_m, B_0, C_m) = p_b (1 - p_c) - \alpha$$
$$P(A_m, B_m, C_0) = p_c + D_1 - D_2 - D_3 - \alpha$$
$$P(A_m, B_m, C_m) = 1 - p_a - p_b + D_2 + D_3 + \alpha.$$
2.4 Appendix I: Publication attached

Parts of the contents of the previous chapter were published in BMC Bioinformatics (2009), S10[44]. The publication is attached.
Abstract

Background: The reaction of HIV protease to inhibitor therapy is characterized by the emergence of complex mutational patterns which confer drug resistance. The response of HIV protease to drugs often involves both primary mutations that directly inhibit the action of the drug, and a host of accessory resistance mutations that may occur far from the active site but may contribute to restoring the fitness or stability of the enzyme. Here we develop a probabilistic approach based on connected information that allows us to study residue, pair level and higher-order correlations within the same framework.

Results: We apply our methodology to a database of approximately 13,000 sequences which have been annotated by the treatment history of the patients from which the samples were obtained. We show that including pair interactions is essential for agreement with the mutational data, since neglect of these interactions results in order-of-magnitude errors in the probabilities of the simultaneous occurrence of many mutations. The magnitude of these pair correlations changes dramatically between sequences obtained from patients that were or were not exposed to drugs. Higher-order effects make a contribution of as much as 10% for residues taken three at a time, but increase to more than twice that for 10 to 15-residue groups. The sequence data is insufficient to determine the higher-order effects for larger groups. We find that higher-order interactions have a significant effect on the predicted frequencies of sequences with large numbers of mutations. While relatively rare, such sequences are more prevalent after multi-drug therapy. The relative importance of these higher-order interactions increases with the number of drugs the patient had been exposed to.

Conclusion: Correlations are critical for the understanding of mutation patterns in HIV protease. Pair interactions have substantial qualitative effects, while higher-order interactions are individually smaller but may have a collective effect. Together they lead to correlations which could have an
important impact on the dynamics of the evolution of cross-resistance, by allowing the virus to pass through otherwise unlikely mutational states. These findings also indicate that pairwise and possibly higher-order effects should be included in the models of protein evolution, instead of assuming that all residues mutate independently of one another.

### Background

The protease enzyme coded for by the pol gene of the Human Immunodeficiency Virus HIV-1 plays a critical role in the reproduction of the virus by cleaving the GAG precursor protein in a sequence-specific manner into its functional form, and as such, is a key target of several families of commonly used drugs used to control HIV infection [1]. Unfortunately, the virus has been able to evolve resistance to many of these drugs, in part due to the high mutation rates in the HIV genome [2]. The patterns of mutations in protease are complex, involving multiple primary mutations that inhibit the action of drugs and a host of accessory mutations that can modulate the enzyme’s stability or activity or otherwise enhance the fitness of the virus. It is now understood that these mutations do not occur independently of each other, but instead are correlated, resulting in complex patterns of co-evolving mutations [3-7].

Previous studies have mostly focused on correlations between mutations in the HIV protease gene at the pair level [3,5-7]. However, recognition that the observed mutations may also be involved in higher-order interactions has led to a few studies in which correlated pairs of mutations are grouped using tools such as multidimensional scaling [3,6], Bayesian networks [8], networks defined by patterns of conditional selection pressure [5], and clustering [9,10]. The underlying assumption is that understanding higher-order interactions is important for a complete understanding of the evolution of resistance in HIV protease.

In this paper, we investigate correlations among HIV protease mutations at and beyond the pair level, and the impact of drug treatment on the nature of those correlations. We only consider the presence or absence of a non-synonymous mutation relative to a defined wild-type sequence, and not the precise base or amino acid substitution which has occurred. We develop a hierarchy of probabilistic log-linear models [11] that can in principle describe residue interactions of arbitrary order, and use those to analyze HIV protease sequence data obtained from patient cohorts with varying protease inhibitor (PI) treatment histories.

We use “connected information” [12] to quantify inter-residue interactions at the triplet and higher level. Unlike the Bayesian network approach [13], the information-theoretic methodology allows us to distinguish intrinsic three-body effects from the cases in which correlations between three random variables can be attributed mostly to pairwise interactions. The connected information viewpoint of higher order correlation has not been previously used in the analysis of mutational patterns in HIV protease, although it has been employed in a much more limited analysis of the V3 loop of the HIV envelope protein [14], and log-linear models have been used to study protein-protein interactions [15]. We find that pairwise interactions are necessary to achieve even qualitative agreement with the mutational data, while higher order interactions play an important role in predicting how frequently sequences with several mutations appear in the database. Simultaneous appearance of multiple mutations may play an important role in the phenomenon of multiple- or cross-resistance of the viral protease.

### Results

**Increased mutation frequencies under drug exposure**

As has been previously observed [3], we find that the overall number of mutations seen in HIV protease increases significantly with the number of PIs that the patient has been exposed to. This is seen in Figure 1, where we show the distribution of the number of mutations at drug-associated positions for sequences isolated from the drug-dominated samples.

![Probability distributions for the total number of mutations among the 41 drug-associated residues in HIV protease for sequences obtained from the PI0 (blue), PI1 (red), and PI2+ (green) cohorts.](image-url)
naive cohort (PI0), from a PI monotherapy cohort (PI1), and from a cohort treated with 2 or more PIs (PI2+), as estimated from the database described in the Methods. For some residue positions, the increase in the mutation frequency between the PI0 and PI2+ cohorts is nearly eightfold, while other positions show no discernible change (Additional File 1). This observed increase in mutation frequencies at drug-associated sites is largely responsible for the shift in the distribution shown in Figure 1.

It is of interest to ask which amino acid positions exhibit elevated mutation frequencies under drug treatment. Mutations at many of these positions are associated with decreased HIV-1 inhibitor susceptibility, and it is useful to classify mutations as belonging to "primary" vs "accessory" resistance classes. The terms "secondary" and "compensatory" have also been used as synonyms for "accessory". The specific criteria for such a classification are ad hoc in nature, but have generally been defined as follows.

Primary mutations are usually selected first in the presence of the drug and confer resistance, even when present as single point mutations [16,17]. They can be structurally important, e.g. situated near the enzymatic active site, in which case their effect on inhibitor binding can be rationalized due to their physical proximity to the inhibitor [18]. In the case of protease, however, there are exceptions, as some mutations (such as positions 54, 76, 88 and 90) are situated far from the active site or have no direct contacts with the substrate, yet still reduce drug susceptibility [16]. The mechanism of action of these mutations is not clearly understood [19].

Accessory mutations confer resistance only when present with additional primary or accessory drug-resistance mutations and have little or no effect on inhibitor susceptibility on their own. Some of these mutations occur in the absence of drug treatment, but their frequency of occurrence is observed to increase in treated patients. Accessory mutations may "rescue" possible losses of activity or stability in the enzyme that may have been caused by a destabilizing primary resistance mutation, and therefore may have a compensatory function in restoring viral fitness [20].

We make use of a primary and accessory classification scheme based on the work of Shafer et al. [21-23,17]. We define 17 primary drug resistance positions (residues 23, 24, 30, 32, 33, 46, 47, 48, 50, 53, 54, 73, 76, 82, 84, 88, 90) and 24 accessory drug resistance positions (residues 10, 11, 13, 20, 34, 35, 36, 43, 45, 55, 58, 60, 63, 71, 74, 75, 77, 79, 83, 85, 89, 91, 93, 95). The remaining positions are polymorphic mutations not associated with drug resistance or are conserved sites [24,3,25,21,26,23].

As seen in Figure 2, the positions which exhibit the most elevated mutation frequencies in PI2+ relative to PI0 sequences are for the most part primary and accessory drug resistance positions. It is interesting to note that sequences bearing mutations at any one of the 6 residues with the most elevated mutation frequencies (10, 46, 54, 71, 82 and 90) are resistant to most of the current PIs, including amprenavir (APV), indinavir (IDV), nelfinavir (NFV), ritonavir (RIV), saquinavir (SQV), atazanavir (ATV), and lopinavir (LPV). For instance, according to the Stanford HIV database [16], the common mutations at these residues (L10I, M46I, I54V, A71V, V82A and L90M) are associated with six or all seven of the drugs mentioned above. On the other hand, primary drug resistance positions that show the least amount of change in their mutation frequency (23, 30, 76 and 88) are generally inhibitor-specific. For example, D30N is associated only with NFV treatment, L23I and L76V are specific to two drugs each (NFV and SQV, and IDV and LPV, respectively), while N88D is specific to IDV, SQV and NFV treatment. Thus, positions that provide drug resistance to multiple inhibitors are more frequently mutated in the PI2+ cohort than positions that are specific to a small number of protease inhibitors. Upon examination of 278 ligand-bound crystal structures of HIV-1 protease, no heavy atom of residues 10, 46, 54, 71, 90 is ever within 3.4 Å of any ligand bound at the cleft (data not shown). The atoms of residue

![Figure 2](http://www.biomedcentral.com/1471-2105/10/S8/S10)

**Figure 2**

**Difference in the frequency of mutated residues between patients treated with 5Pis and PI-naive patients for all positions in HIV protease.** Red bars correspond to primary drug resistance positions, black bars are accessory drug resistance positions, and grey bars are positions not associated with drug resistance.
82, however, do contact ligands in 54% of the crystal structures examined.

**Exhaustive analysis of residue pairs**

We begin by investigating pair correlations and their association with the structure of HIV protease. Several groups have studied pair correlations as a means of identifying functionally dependent residues in the HIV protease and other systems [6,5,4,3,7]. For each of the \( \binom{82}{2} = 3,321 \) pairs of positions, a 2 × 2 contingency table was constructed for both the drug-naive sequences and the sequences treated by two or more drugs. The binomial or "product moment" [11] correlation coefficient

\[
\phi_{AB} = \frac{P(A_m B_m) - P(A_m) P(B_m)}{\sqrt{P(A_m) P(A_\text{m}) P(B_m) P(B_\text{m})}}
\]

was calculated for both datasets, where for amino acid position \( A \) we denote the wild-type state as \( A_w \) and the mutated state as \( A_m \). Of the 3, 321 \( \phi \) values, 98 from the drug-naive and 223 from the treated set were considered to be statistically significant with substantial correlations (\( |\phi| > 0.1, p < 0.001 \)). These correlations match qualitatively with a prior study which used the same database but had fewer sequences: the \( \phi \) values of the top 15 positively correlated pairs for the PI2+ cohort and those of Wu, et al. [3] have a Spearman rank order correlation coefficient of 0.80 (data not shown). Furthermore, we observe larger \( \phi \) values for pairs of drug-associated positions compared to non-drug associated pairs (Figure 3), which is consistent with previous observations [3].

Previous studies of pair correlations in protein families have indicated that coevolving pairs of residues tend to lie closer to each other in structure than random residue pairs [27-31]. In the case of the HIV protease, the distribution of \( \phi \) values for drug-associated positions with distance shows a characteristic triangular shape, particularly for the PI2+ cohort (Figure 3B) [32]. In particular, distances associated with the most correlated residue pairs (30–88, 54–82, 32–47, and 37–77) are all within a few Ångstroms of each other. The triangular shape in Figure 3B is not surprising, given that most of the drug-associated positions are on the substrate cleft and thus tend to be relatively close to each other in space. As a result, 78 pairs (36% of the statistically significant pairs) are within 8 Å of each other. We also see no tendency for exposed residues to preferentially coevolve, in contrast to previous studies on mutation covariation in other protein families [27,29,15].

Of the 26 statistically significant negatively correlated pairs in the PI2+ cohort, 10 pairs involve either residues 30 or 88. Residue 30 is negatively correlated with positions 82, 10, 46, 90, 54, 73 and 84 (in decreasing order of the magnitude of the correlation), while position 88 is negatively correlated with positions 82, 73 and 54. It is interesting to note that position 63, which has a high mutation rate in both the PI0 and PI2+ cohorts, is negatively correlated with positions 80, 52, 5, 83, 64, and 61, of which only position 83 is associated with drug resistance [3]. In fact, of the 23 unique positions involved in the 26 negatively correlated pairs, most, but not all, are positions of drug resistance. It is possible that the 7 non-drug associated positions, 5, 15, 20, 52, 61, 64 and 80 play a role in the stability or function of the protein, even if they do not interfere with inhibitors [33-35]. Residue 80, in particular is negatively correlated with three residues, 63, 71 and 90, all of which play either primary or accessory roles in drug resistance.

Drug treatment has a significant impact on the pair correlations, as can be seen in Figure 4. Of the \( \binom{82}{2} = 3,321 \) pairs of positions, only 100 have statistically significant (\( p < 0.001 \)) \( \phi \) values and are common to both the PI0 and PI2+ cohorts, and of these, most of the positively correlated pairs in the PI0 cohort become more strongly correlated in the PI2+ cohort. However, some pairs of residues which are weakly positively correlated in the PI0 cohort become negatively correlated in the PI2+ cohort. There are 12 such pairs, almost all of which have at least one primary drug resistance position, and 8 of them involve...
Exhaustive analysis of residue triples

To study interactions among mutations beyond the pair level, we quantify the amount of information in the observed distribution that cannot be explained by pair correlation. This is done using the three-body “connected information” [12], which is defined as the difference in Shannon entropy $S(P) = -\sum p_i \log p_i$ between the distributions $P(A, B, C)$ and $\tilde{P}^{(2)}(A, B, C)$, where the latter is the maximum entropy distribution subject to the constraints that all of its univariate and bivariate marginals are the same as that of $P(A, B, C)$:

$$I_c^{(3)}(A, B, C) = S[\tilde{P}^{(2)}(A, B, C)] - S[P(A, B, C)].$$

(2)

Similarly, we define the two-body connected information to be

$$I_c^{(2)}(A, B, C) = S[P(A)P(B)P(C)] - S[\tilde{P}^{(2)}(A, B, C)].$$

(3)

The total information arising from correlation at any level is given by the “multi-information” or the Kullback-Leibler divergence between the observed distribution and the prediction based on an independent model:

$$I_{\text{multi}}(A, B, C) = \sum_{a,b,c} P(a,b,c) \log \frac{P(a,b,c)}{P(a)P(b)P(c)} = S(A) + S(B) + S(C) - S[A,B,C].$$

(4)

The maximum entropy distribution $\tilde{P}^{(2)}(A, B, C)$ can be thought of as being “in between” the independent model and the observed distribution, since it is more constrained than the independent model but does not have the full correlation structure of the observed data (Figure 5). Since $I_c^{(2)} + I_c^{(3)} = I_{\text{multi}}$, we can interpret the two- and three-body connected information as the part of the correlation that can be explained by pair interactions alone, and that which arises from three-body interactions, respectively [12].

We examined the degree of connected information in both the PI0 and PI2+ cohort sequences. The connected information $I_c^{(3)}$ was calculated for all $\binom{82}{3} = 88,560$ residue triples in the HIV protease as described in the Methods section below. In the PI0 cohort sequences, only 175 residue triples have statistically significant three-body interaction at the $p = 0.001$ level, while the PI2+ cohort sequences have 6,300 significant triples. Furthermore, the significant triples from the PI2+ cohort are enriched in drug-resistance associated positions: 32% of them consist of either residue 30 or 88. It is interesting to note that in PI0 cohort sequences, residue 30 is positively correlated with positions 24, 46, 54, 84 and 90, but becomes strongly negatively correlated in the PI2+ cohort. The anticorrelation of residue 30 with the other primary positions after drug treatment has been previously observed experimentally [36] and in a prior statistical study [3], but it is not clear why this anticorrelation exists only in the presence of drugs.

Additionally, the types of residues involved in pair correlations changes upon treatment, with accessory positions becoming more prominent: 10% of the correlated pairs in the PI0 cohort consist of a primary and an accessory position, and that this combination increases to 31% in the PI2+ cohort. Furthermore, 45% of the pairs in the PI0 cohort involve at least one accessory position, which increases to 61% in the PI2+ cohort. As expected, pairs of non-resistance-associated positions decrease from 52% to 35%. It is interesting to note that the same trend is not observed for primary positions: 57% of pairs in the PI0 cohort contain at least one primary position, and this is essentially unchanged in the PI2+ cohort (59%). Therefore, drug treatment causes correlated pairs involving primary and non-resistance associated positions to be replaced by pairs involving primary and accessory positions.

Figure 4
Scatterplot showing the change in $\varphi$ statistic upon drug treatment, with drug associated positions in red and non-drug associated positions in black. Only pairs with statistically significant ($p < 0.001$) pair correlations are shown. The solid line corresponds to no change upon drug treatment.
The entropy of the distribution denoted by \((A, B, C)\) in figure by a star. the same as the observed distribution, and is denoted in the constraints that all of the univariate and bivariate marginals are the text which has the maximum entropy subject to the con-

\[ I_C^{(2)} = \text{Shannon entropy for the family of all triplet distributions having the same univariate and bivariate marginals as the data for the triplet 46-54-90 (PI2+ cohort) plotted as a function of the parameter } \alpha \text{ defined in Equation 8. The observed distribution is indicated by the filled circle, and the entropies of the independent model (} S_{\text{indep}} \text{), the model with only pair interactions (} S_{\text{pair}} \text{) and the observed data (} S_{\text{obs}} \text{) are indicated, as well as } I_{\text{mult}} \text{ and the connected information measures } I_C^{(2)} \text{ and } I_C^{(3)}. S_{\text{obs}} \text{ is the entropy of the distribution denoted by } \tilde{P}^{(2)}(A, B, C) \text{ in the text which has the maximum entropy subject to the con-

of only drug associated positions, compared to 12% of the full set of 88,560. In contrast, the significant triples from the PI0 cohort show no such enrichment: only 9% consist solely of drug resistance associated positions.

The 10 triples with the largest \(I_C^{(3)}\) values from the PI2+ cohort sequences are shown in Table 1, along with the corresponding values of \(I_{\text{mult}}\). Residues from the largest \(I_C^{(3)}\) triple (46-54-90) are also displayed on the structure of the HIV protease (Figure 6). As can be seen in Table 1, the total contribution of three-body interactions to the information content of the observed data is at most 10%. Although this is a relatively small effect, there is a clear association with drug therapy, with the largest \(I_C^{(3)}\) values increasing substantially with the number of PIs the patient was exposed to (Figure 7). There is relatively little overlap between the largest 100 \(I_C^{(3)}\) triples in the PI0 and PI2+ cohorts, with only 11 triples in common. The impact of these higher-order interactions is particularly manifest in the probability of the occurrence of three simultaneous mutations in a given residue triple (Figure 8), which is significantly increased compared to what would be predicted based on a pair model for sequences from the PI2+ cohort (but not in sequences from the PI0 cohort).

While all of the triples in Table 1 have three-body interactions that are highly statistically significant, it is also important to obtain a practical feeling for the magnitude of these interactions. Let us consider the 46-54-90 triple. In Figure 9 we show the correlation between the predicted and observed probabilities for each of the 8 binary states for the independent model (red) and the two-body fit (black) for this triple in the PI2+ cohort. It is clear that the independent model fits the data poorly; the probability that all three residues are mutated is underestimated by an order of magnitude; and some of the others are over- or underestimated by factors of 2 or 3. Although the two-body fit does a better job of reproducing the observed probabilities, it still leads to noticeable deviations from the straight line for the 46-54-90 triple.

Another way in which the magnitude of the effect of the three-body interactions can be visualized is by considering conditional probabilities for the 46-54-90: \(P(46|46,90) = 0.4186, P(54|46,90) = 0.5508, P(54|46,90) = 0.4359,\) and \(P(54|46,90) = 0.0280.\) All of these probabilities differ significantly from the independent estimate of \(P(46) = 0.1178.\) We also see clear evidence of “triplet correlation” in the data, in the sense that the probability of a mutation depends very strongly on the state of both of the other residues, e.g. residue 54 is much less likely to be mutated if both 46 and 90 are wild-type than if only one of them is wild-type.

<table>
<thead>
<tr>
<th>residue triple</th>
<th>(I_C^{(3)} (A, B, C))</th>
<th>(I_{\text{mult}} (A, B, C))</th>
</tr>
</thead>
<tbody>
<tr>
<td>46-54-90</td>
<td>0.03219</td>
<td>0.20170</td>
</tr>
<tr>
<td>46-71-90</td>
<td>0.03101</td>
<td>0.29041</td>
</tr>
<tr>
<td>46-82-90</td>
<td>0.03085</td>
<td>0.20640</td>
</tr>
<tr>
<td>82-84-90</td>
<td>0.02570</td>
<td>0.13237</td>
</tr>
<tr>
<td>10-46-90</td>
<td>0.02445</td>
<td>0.31475</td>
</tr>
<tr>
<td>46-71-73</td>
<td>0.02195</td>
<td>0.16610</td>
</tr>
<tr>
<td>36-46-90</td>
<td>0.02191</td>
<td>0.15949</td>
</tr>
<tr>
<td>46-54-71</td>
<td>0.02072</td>
<td>0.23595</td>
</tr>
<tr>
<td>46-77-82</td>
<td>0.02068</td>
<td>0.14573</td>
</tr>
<tr>
<td>20-82-90</td>
<td>0.02030</td>
<td>0.14661</td>
</tr>
</tbody>
</table>
Much of this probabilistic dependency, however, can be accounted for by pair interactions. The corresponding probabilities for the best-fit two-body model are 
\[
\begin{align*}
\hat{p}^{(2)}(54_m|46_m90_m) &= 0.6286, \\
\hat{p}^{(2)}(54_m|46_090_m) &= 0.2236, \\
\hat{p}^{(2)}(54_m|46_0900) &= 0.2316, \\
\hat{p}^{(2)}(54_0|46_090) &= 0.0488.
\end{align*}
\]
Even though this model contains no three-body interactions, a qualitative “triplet correlation” of the type seen above remains (i.e. 
\[
\hat{p}^{(2)}(54_m|46_m90_0) \approx \hat{p}^{(2)}(54_m|46_090_m) > \hat{p}^{(2)}(54_m|46_0900)
\]
indicating that non-trivial three-way probabilistic dependencies can arise purely from pair correlations [37]. Overall, three-body interactions do quantitatively modulate the probabilities, but only to a small degree, since \(I_C^{(3)}\) is an order of magnitude smaller than \(I_{\text{null}}\) even for the triples in Table 1. In other words, the contribution of three-body interactions to any of the triplet distributions that describe the mutational patterns of protease taken three at a time is roughly 10% or less of the effect induced by the pairwise interactions acting on these positions. Nonetheless, there can be substantial effects on a “micro level”, such as the more than twofold difference in \(P(54_m|46_m90_0)\) seen here.

**Increased higher-order interactions in larger residue groups**

The small but consistent systematic deviations seen in Figure 8B raise the possibility that these interactions could combine synergistically to produce more substantial effects over larger clusters of residues. Ideally, this would be studied by fitting log-linear models to increasingly larger clusters using the data from the PI2+ cohort. How-
ever, the size of cohort limits our ability to do this to clusters of no more than \( \approx 15 \) residues. We begin by examining the 10-residue group 20-32-46-53-54-58-74-82-90, which was chosen to have the largest higher-order interactions from among a limited set of residues defined by the three key primary drug resistance positions 30, 82 and 90, and the known accessory positions associated with them. For this 10-residue group in the PI2+ cohort, we observe strong pair interactions which bring observed and predicted probabilities into qualitative, order-of-magnitude agreement, weaker three-body interactions which further improve the agreement, and very weak four-body and higher interactions which have quantitative impact on a small number of state probabilities (Additional File 2). This can be quantified in terms of connected information: from the entropies of the of the observed, three-body model, pair model, and independent distributions (Figure 10), we find that \( I_{\text{pair}}^{(2)} = S(\text{ind}) - S(\text{pair}) = 0.4785, I_{\text{trip}}^{(3)} = S(\text{pair}) - S(\text{trip}) = 0.1187 \), and the sum of the remaining connected information measures of fourth order and higher is \( S(\text{trip}) - S(\text{obs}) = 0.0703 \) (these information theoretic measures were found to be robust with respect to sampling error as determined by bootstrap). When compared to \( I_{\text{multi}} = S(\text{ind}) - S(\text{obs}) = 0.6676, we see that three-body and higher-order interactions make up 28% of the total correlation information. This is substantial increase over the (at most) 10% contribution from higher-order interactions to the observed triplet distributions. By contrast, the same 10-residue group for the P10 cohort displays a substantially smaller overall degree of correlation, as seen by the small \( I_{\text{multi}} \) in Figure 10.

We also studied the distribution of the total number of mutations in the same 10-residue group (20-32-46-48-53-54-58-74-82-90). Appropriate subsets of the state probabilities determined above were summed to obtain the distribution of the total number of mutated residues for the independent, two-body, and three-body models, and the results are shown in Figure 11A. The distribution for the independent model is very different from the observed distribution: the probability of having no mutations is considerably underestimated, and the upper tail is much too thin. Adding pair terms greatly improves the "no mutation" probability and considerably extends the length of the tail. The tail length is further modulated by the addition of the three-body interactions, bringing the distribution very close to the observed probabilities.

It should be noted that since the univariate marginals are preserved by all of the models, the mean total number of mutations is the same for all 4 curves in Figure 11A. Therefore, under- or overestimation of the total number of mutations in one part of the distribution must be compensated by over- or under-estimation (respectively) in
To see if the synergistic effects seen for the 10-residue group become even stronger for 15 residues, we repeated this analysis for the 15-residue group 10-20-33-36-46-54-55-63-71-73-74-82-84-90-93 (Figure 10), which was chosen by selecting the residues with the largest change in mutation frequencies upon PI treatment. For the PI2+ cohort, we find that \( I_2^{(1)} = S(\text{ind}) - S(\text{pair}) = 1.1810 \), the sum of the remaining connected information measures of third order and higher is \( S(\text{pair}) - S(\text{obs}) = 0.8249 \) and \( I_{\text{multi}} = S(\text{ind}) - S(\text{obs}) = 2.0059 \). The contribution of three-body and higher-order interactions now make up 41% of the total correlation information. However, this result may somewhat overestimate the true amount of higher-order correlation. A fit with a three-body model (Equation 12) gives an estimate of \( I_3^{(3)} \) of 0.2638, and the ratio \( I_3^{(3)}/(I_2^{(2)} + I_3^{(3)}) = 18\% \) represents a lower bound on the contribution from correlations beyond the pair level. A more complete account of the many issues involved in estimating the amount of higher order correlation and its precision and accuracy for finite data sets will be the subject of a future communication. The comparison of the predicted and observed distributions of total number of mutations bears out this result (Figure 11B), showing more pronounced differences between the observed (black) and pair-model (red) distributions. In addition to the overestimation of the upper tail similar to that seen for the 10-residue group, we now also see that the pair model cannot reproduce the bimodal shape seen in the observed data. Again, the same 15-residue group for the PI0 cohort shows considerably weaker overall correlation (Figure 10).

**Discussion and conclusion**

Treating HIV protease with drugs results in the appearance of complex mutational patterns: observed mutations are not limited to the active site and often occur in groups that involve two or more residues. Furthermore, some mutations occur even in the absence of drugs, presumably following neutral rather than adaptive evolution. To study correlations between different residue positions in HIV protease, we have developed a hierarchy of models that allows us to include inter-residue correlations of arbitrary order within a consistent framework. Using only HIV protease sequences as input, we find that pair interactions become common and quite strong after PI treatment. In fact, it is often impossible to achieve even qualitative agreement with the data without including the two-body terms (Figures 9, 11, and Additional File 2). This finding calls into question a common assumption employed in another, implying that the curves must cross one other. To quantify this effect, we can compare the predicted probabilities for seeing 5 or more mutations under each model and comparing to the observed probability. Those probabilities are 0.0372, 0.1197, and 0.1030 for the independent, pair-term model, and pair+three-body model, respectively, compared to the observed probability of 0.1007. The deviations of the first two from the observed are highly statistically significant, while the latter has a p-value of \( \approx 0.06 \). The qualitative distribution is well accounted for by the pair interaction model, with the three-body and higher interactions modulating the details of the shape of the distribution, such as the increase in frequency of 3–5 and 13–15 mutations, at the expense of a decrease in the frequency of 6–10 mutations.

**Figure 11**

Total number of observed (solid black curve) or predicted mutations in the 2,702 sequences of the PI2+ cohort for the 10-residue group 20-32-46-48-53-54-58-74-82-90 (A) and the 15-residue group 10-20-33-36-46-54-55-63-71-73-74-82-84-90-93 (B). The dashed curves represent the predictions for the independent (blue), pair-term (red), and pair+three-body model (green, not fit for the 15-residue group).
current probabilistic approaches to phylogeny [38] that most residues evolve independently.

We have developed an information-theoretic method to study interactions between mutations beyond the pairwise level. Our approach is based on the notion of the connected information $I_c^{(3)}$ (Equation 2) [12]. While there are a variety of quantitatively different measures of pair correlation [11,39] that may differ in their sensitivity in various regimes, they all measure essentially the same qualitative feature of the observed data. On the other hand, no single summary statistic can capture all of the various characteristics of higher-order behavior, leading to multiple descriptions that provide complimentary information. Connected information is one intuitive statistic that provides insight into the degree of structure in the data beyond the pair correlation level.

Connected information provides information which is complementary to Bayesian network analyses based on factorizations of the joint probability. It can readily verified that $I_c^{(3)} = 0$ if at least one of the random variables is independent of the other two, or if the joint distribution involves conditional independence (e.g. $P(A, B, C) = P(A)P(B|A)P(C|A)$). However, a joint distribution with triplet-level probabilistic dependencies (in the sense that two of the variables are independent of each other, but the third depends jointly on the state of the other two, e.g. $P(A, B, C) = P(A)P(B|A)P(C|A, B)$, or if $P(A, B, C)$ cannot be factorized into any simpler form) could still be consistent with no three-body connected information if the observed triplet distribution is the maximum entropy distribution relative to its marginals. Thus, even if a Bayesian network-style analysis shows that a given triple cannot be factorized into any simpler form, that "triplet correlation" could still be consistent with a very small or zero $I_c^{(3)}$, indicating that the observed behavior is dominated by two-body interactions. In fact, it has been shown that very complex correlation patterns among random variables can arise from large numbers of weak pairwise interactions [37].

Other information-theoretic measures of "higher-order correlation" have also been proposed, including higher-order mutual information, which measures "frustration" or the degree of synergy vs redundancy among several random variables [40]. While this measure has been used in the analysis of HIV envelope protein sequence data [41], its interpretation is considerably less intuitive. Similarly, ad hoc methods for finding putative clusters of mutually correlated residues [3,9,8,10] cannot reliably uncover sets that have intrinsic higher order interactions, as defined by large $I_c^{(3)}$ (data not shown).

Plotting $\varphi$ values for pairs of residues as a function of the distance between them (Figure 3B) reveals that while some large pair correlations arise from direct contacts between residues (e.g. $\varphi \approx 0.5$, $d < 5 \text{ Å}$), there are also strong correlations ($\varphi \approx 0.5$) between amino acids separated by 15 Å or more, making physical coupling between them very unlikely. To provide an example of the former, we consider mutations involving residues 30 and 88. The closest distance between heavy atoms of residues 30 and 88 is just 3.66 Å, making likely some sort of physical interaction between them. Mutations at residue 30 are strongly and uniquely associated with resistance to the protease inhibitor nelfinavir, and there exists a strong correlation between mutations at positions 30 and 88 [42] which may be due in part to a compensation of the loss of a surface negative charge from the D30N mutation being restored by N88D [9].

It is possible that chains of intermediate interactions result in long-range coupling between two coevolving yet physically distant residues [43]. However, because non-zero values of $\Delta_{ij}$ indicate a presence of direct interactions between residues $i$ and $j$ in our model [15], we can decompose such "energetically connected pathways" into contributions from separate pairs. In contrast to a previous study [15], we find that non-zero values of $\Delta_{ij}$ are only weakly correlated with distance (Additional File 3). This lack of correlation is not entirely surprising, since even direct interactions between a pair of residues need not have a purely physical origin. Indeed, if protein fitness is a non-linear function of its stability or enzymatic activity [44], two mutations can be correlated because they compensate each other by making independent and opposite contributions to the overall fitness, even if there is no direct or indirect physical interaction between them [45]. By the same argument, the three-body terms also result from a mixture of physical and epistatic (compensatory) origins.

We have shown that three-body and higher-order correlations have the largest effect on the probabilities of the simultaneous occurrence of multiple mutations in the HIV protease (Figure 11). Since both this and previous studies have found that the total number of mutated positions is correlated with treatment by multiple protease inhibitors (Figure 1) [3], the presence of higher-order interactions may influence how protease reacts to multiple drugs, and could have an important impact on the evolution of cross-resistance, for example, by providing
the virus with an "escape hatch" of large numbers of mutations. Higher-order interactions could also impact the time evolution of mutations by allowing the virus to pass through otherwise unlikely mutational states. We have seen that the impact of higher-order interactions in 10 to 15 residue clusters is at least a factor of two larger than the largest $I_{C}^{(3)}$ values for residue triples (approximately 20% or more of the total entropy change). One of the outstanding questions raised by this work is whether the impact of higher-order interactions for HIV evolving under the pressure of multiple drugs continues to become stronger for larger residue groups (ultimately in the study of all 41 drug-associated positions). Unfortunately, there is not enough sequence data to perform such an analysis. Short of obtaining additional data, it may also be possible to explore this question by constructing synthetic data sets using $\lambda_{A}$ and $\lambda_{B}$ values consistent with an observed $I_{C}^{(3)}$ distribution at the level of residue triples (i.e. Figure 7C).

The sequence-based approach presented here is not limited to the HIV protease and its response to drug treatment, and should be equally useful in studies of the evolution of drug resistant in other systems. Moreover, it will be of interest to extend our techniques to other examples of short-term neutral and adaptive evolution, including controlled evolution in the lab accompanied by protein sequencing at different timepoints. Recent work has suggested that evolutionary pathways of proteins are relatively restricted and may be predictable in general [46], and specific methods for predicting the mutational dynamics of HIV protease have been proposed, based on Bayesian network models [47] or pairwise conditional selection pressure [5]. A better understanding of the nature of the probabilistic dependencies underlying the network models should lead to improved prediction strategies. However, our model cannot distinguish between physical and epistatic origins of the observed co-evolution. To do this, we need a different approach which would explicitly introduce protein fitness as a function of residue energies (including interactions across protein-protein and protein-ligand interface). These energies would be fit against the sequence data, resulting in a prediction that decomposes observed inter-residue correlations into the physical and epistatic parts. This approach is currently being pursued in our laboratories.

**Methods**

**HIV sequence database**

Aligned and annotated HIV-1 protease amino acid sequences were obtained using the web interface of the Stanford HIV Drug Resistance Database [http://hivdb.stanford.edu/cgi-bin/PI_Form.cgi][16]. The sequences are all classified under the HIV-1 Main group, subtype B. For the purposes of this paper, we are considering only the number of protease inhibitors and not the specific combinations of protease inhibitors used in the treatment. There are 13,608 sequences in this curated set, of which 8,229 sequences are drug naive, 2,677 sequence are associated with PI monotherapy, and the remaining 2,702 sequences are associated with between 2 through 6 protease inhibitors.

**Calculation of higher-order interactions**

For residue triples, we compute $I_{C}^{(3)}$ by writing the triplet probability distribution in log-linear form [11]:

$$P(A, B, C) = \exp(\lambda_{A}A + \lambda_{B}B + \lambda_{C}C + \lambda_{AB}AB + \lambda_{BC}BC + \lambda_{AC}AC + \lambda_{ABC}ABC),$$

(5)

where we assign numerical values to the states of $A$, $B$ and $C$, e.g. 0 and 1. We find $\hat{I}_{C}^{(2)}(A, B, C)$ by setting $\lambda_{ABC} = 0$ in Equation 5 and fitting the six parameters $\lambda_{A}$, $\lambda_{B}$, and $\lambda_{C}$ to the values that maximize the likelihood of the data under a multinomial model [48]. However, in the triplet case it is possible to avoid direct nonlinear optimization: let us represent the 8 observed probabilities by the vector

$$p_{0} = (p_{000}, p_{00m}, p_{0m0}, p_{0mm}, p_{m00}, p_{m0m}, p_{mm0}, p_{mmm}),$$

(6)

where, e.g. $p_{00m} = P(A_{0}B_{m}C_{0})$. It is sufficient to consider only the three marginals $P(A_{0})$, $P(B_{m})$, and $P(C_{0})$, and three suitably chosen bivariate marginals, e.g. $P(A_{m}B_{0})$, $P(A_{m}C_{0})$, and $P(B_{m}C_{0})$, since the remaining 9 bivariate marginals can be reconstructed as combinations of these: $P(A_{m}B_{m}) = P(A_{m}) \cdot P(B_{m})$, $P(A_{m}B_{m}C_{0}) = P(A_{m}) \cdot P(B_{m}) \cdot P(C_{0})$. The six marginals can then be written as a matrix equation involving $p_{0}$:

$$\begin{pmatrix}
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
P(A_{0}) \\
P(B_{m}) \\
P(C_{0}) \\
P(A_{m}B_{m}) \\
P(A_{m}C_{0}) \\
P(B_{m}C_{0})
\end{pmatrix} = \lambda - I_{C}^{(3)}$$

(7)

Since the matrix multiplying $p_{0}$ is rectangular with dimensions $6 \times 8$, it has a two-dimensional null space, with basis vectors $n_{1} = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ and $n_{2} = (0, -1, -1, 1, -1, 1, -1, 1, -1)$. Then, any linear combination $\alpha_{1}n_{1} + \alpha_{2}n_{2}$ added to $p_{0}$ will not change the marginals. However, $\alpha_{1}$ and $\alpha_{2}$ cannot be chosen independently without violating the normalization of $p_{0}$: we must choose $\alpha_{1} = \alpha_{2} = \alpha$. There-


fore, the family of all possible distributions that have the same univariate and bivariate marginals are mapped out by the parameter \( \alpha \) using the relation

\[
P_\alpha = p_0 + \alpha (1,-1,-1,1,1,-1,1,1,-1),
\]

where the feasible values of \( \alpha \) are constrained by the non-negativity requirement for probabilities. Furthermore, \( \lambda_{abc} = 0 \) in Equation 5 implies that

\[
P_{000}P_{mm0}P_{00m}P_{0m0}P_{m00}P_{0mm}P_{00m}P_{m00} = 1.
\]

Therefore, to find \( \hat{P}^{(2)} (A, B, C) \) it suffices to find the value of \( \alpha \) which satisfies Equation 9:

\[
\frac{(p_{000}+\alpha)(p_{mm0}+\alpha)(p_{00m}+\alpha)(p_{m00}+\alpha)}{(p_{0m0}+\alpha)(p_{m00}+\alpha)(p_{0mm}+\alpha)} = 1,
\]

leading to a cubic equation in \( \alpha \) [11].

To obtain the maximum entropy distributions for more than three binary random variables, nonlinear optimization is unavoidable. However, instead of directly maximizing the entropy subject to the marginal probability constraints, we maximize the likelihood subject to the constraints that the \( \lambda \)'s vanish beyond a given order [48]. In most cases, the latter will be more computationally efficient since the number of \( \lambda \) variables grows polynomially with the number of variables, while the dimensionality of the null space defined by the marginal probability constraints (which is one-dimensional for three variables) increases exponentially with the number of variables.

In general, we fit data on mutation and wild-type amino acid counts to the following hierarchy of probabilistic models: the independent model \( P(A, B, C,...) = P(A)P(B)P(C) \), the "two-body" model:

\[
P(A, B, C,...) = \exp(\lambda_0 + \sum_i \lambda_i I + \sum_j \lambda_j J),
\]

and the "three-body" model:

\[
P(A, B, C,...) = \exp(\lambda_0 + \sum_i \lambda_i I + \sum_j \lambda_j J + \sum_{jk} \lambda_{ijk} JK),
\]

where \( \lambda \) is the vector of parameters, the indices \( i, j, k \) run over all distinct combinations of \( \{A, B, C,...\} \) with \( i \neq j \) and \( i \neq j \neq k \), respectively, and \( I, J, \) and \( K \) are numerical values of the corresponding random variable (we use 0 for wild-type and 1 for mutant). For an \( n \)-variate distribution \( P(A, B, C,...) \) there are \( n(n-1)/2 \) pair parameters \( \lambda_{ij} \) and \( n(n-1)(n-2)/6 \) three-body parameters \( \lambda_{ijk} \) (\( \lambda_0 \) is a normalization constant). The independent model was determined by forming products of the observed univariate marginals. The magnitudes of the \( \lambda_{ij} \) and \( \lambda_{ijk} \) parameters in the two-body model are related to the mutation frequencies at site \( i \) and pair correlations between sites \( i \) and \( j \), respectively. In fact, the magnitudes of \( \lambda_{ij} \) in the context of a two-body model have been proposed as a measure of "direct information", i.e. the part of pair correlation resulting from direct coupling [15]. It should be noted that the relative magnitudes of \( \lambda_{ij} \) is dependent on the choice of the numerical values assigned to the random variables \( I, J \), and \( K \) in Equations 11 and 12. In this work, we assign values of 0 and 1 for wild-type and mutant, respectively, for computational convenience. It has been argued, however, that a more appropriate choice of numerical values is one which is symmetric about zero, e.g. ± 1, which allows "gauge constraints" to be introduced [15]. While this choice will affect the the values of \( \lambda_{ij} \) and their interpretation as "direct information", it will not change the best-fit two-body or three-body probabilities and consequently will have no impact on the values of \( I_c^{(3)} \) and related measures of higher-order interactions.

The two-body model for \( n = 3 \) was fit by solving Equation 10 exactly [49]. If no feasible solution of Equation 10 exists, then \( I_c^{(3)} \) was set to zero. For \( n \geq 4 \), the unknown parameters in Equations 11 and 12 were determined by maximizing the multinomial log-likelihood

\[
L(\lambda) = \sum_i N_i \ln P(i),
\]

where \( i \) is one of the \( 2^n \) states, \( N_i \) is the number of times that state was observed, and \( P(i|\lambda) \) is the predicted probability for state \( i \) to be observed given the vector of parameters \( \lambda \). Maximization was performed numerically using the "nlm" function of the R software package [50]. All entropy and connected information values are given in natural log units. Statistical significance of the three-body interactions was estimated using the likelihood ratio test under the null hypothesis that the data were generated by \( \hat{P}^{(2)} (A, B, C) \) by Monte Carlo sampling. For all of the residue triples in Table 1, the \( p \)-values for the observed likelihood ratio were too small to be estimated (\( p \ll 10^{-6} \)).
indicating very strong statistical significance for the three-body interaction.

**Competing interests**

The authors declare that they have no competing interests.

**Authors' contributions**

OH and MA wrote software and performed the analyses, and OH, RML, AVM and MA wrote the paper.

**Additional material**

Additional File 1  
Mutation frequencies for selected positions in the HIV protease as a function of the number of PIs the patient was exposed to. Data is shown for residues 10 (black, solid line), 54 (red, solid line), 90 (green), 71 (blue), 46 (orange), 77 (black, dashed line), and 35 (red, dashed line). Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-S8-S10-S1.pdf]

Additional File 2  
Predicted vs observed probabilities for the $2_{\lambda}$ mutational states of the ten-residue group 20-32-46-48-53-54-74-82-90 in HIV protease for the PI2+ cohort. Mutational states that were unobserved in the database and would have an observed probability maximum likelihood estimate of zero are not shown. The black dots correspond to the best-fit independent model, the red dots correspond to the best-fit pair-term model (Equation 11), and the green dots correspond to the best-fit three-body model (Equation 12). The solid line of slope 1 corresponds to perfect agreement of the predicted data with the observed, which would be obtained if all higher-order terms were included. Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-S8-S10-S2.pdf]

Additional File 3  
Scatterplot of distance vs $\lambda_2$ parameters estimated using $L_1$ and $L_2$ values of $\pm 1$ as described in the Methods for the 15-residue group 10-20-33-42-45-46-75-82-90. The value along the y-axis is the closest distance between any two heavy atoms of the two residues based on the crystal structure of wild type protease (PDB ID 1PR0). Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-S8-S10-S3.pdf]

**Acknowledgements**

This work has been supported by a grant from the National Institutes of Health, GM-30580 (to RML) and by an Alfred P Sloan Research Fellowship (to AVM).

This article has been published as part of BMC Bioinformatics Volume 10 Supplement 8, 2009: Proceedings of the European Conference on Computational Biology (ECCB) 2008 Workshop: Annotation, interpretation and management of mutations. The full contents of the supplement are available online at  
http://www.biomedcentral.com/1471-2105/10/S8/S10

**References**

23. Shafer RW, Rhee SY, Pillay D, Miller V, Sandstrom, P, Schapiro JM, Kuritzkes DR, Bennett D: HIV-1 protease and reverse tran-
Chapter 3

Introduction to Inference on Probabilistic Graphical Models

3.1 What are probabilistic graphical models?

Probabilistic graphical models provide a general framework to approach problems that involve complex networks of interacting and/or independent random variables. The models not only simplify the way we visualize and understand complicated networks, but also allow for the efficient inference of marginal and joint probabilities [9]. Graphical models have their roots in a number of academic communities, such as artificial intelligence[82], statistics and neural networks. They provide a simple, yet elegant mathematical formalism, allowing researchers to understand complex probabilistic relationships between elements of networks. Probabilistic graphical models, specifically, are a union between graph theory and probability theory. They use graphs to model or describe the joint probability distributions for a class of distributions that share a common structure. A graph has nodes, drawn as circles, that represent the variables of the joint probability. Edges connect nodes to each other and they are drawn as lines and represent probabilistic dependencies or relationships between variables. There are two main kinds of graphs used in graphical models, directed and undirected. Directed graphical models include Bayesian networks and are depicted on the left hand side of Figure 3.1. An undirected graph consists of edges which have no directionality (right side of Figure 3.1). This form of graph
Figure 3.1: Example of an acyclic directed graph on the left and a cyclic undirected graph on the right.

is known as a Markov Random Field (MRF). Another widely used undirected graph is the factor graph.

Regardless of directionality, all graphical models have a structural component, represented by the network of edges between nodes. They may be completed cyclic (also known as a completed connected graph or completely loopy), with every node connected to all other nodes or completely acyclic, which is a network devoid of any cycles (or loops), or some mixture of the two. The undirected graph in Figure 3.1 contains loops but is not completely cyclic; every node is not connected to every other node. However, the absence of an edge between two nodes is indicative of a conditional independence structure. Graphical models also have a parametric component represented by the coupling potentials associated with edges and field potentials associated with nodes. The network structure of the graph and its underlying potentials define the particular graphical model being studied. Representing the data in a graphical model allows for inference algorithms to efficiently calculate marginal and joint probabilities, conditional probabilities, and likelihoods. Undirected graphs such as the one in Figure 3.1 are used in many applications including
protein folding, protein evolution and structural bioinformatics where causality of interactions is not the primary concern and where the interactions between nodes can be described by non-directional potential functions. This work focuses on the determination of marginal and joint probabilities for undirected graphs, mainly the Markov random field and the factor graph.

3.2 Background to inference on graphical models

This chapter deals with functions of many variables. Let $A_1, A_2, A_3, ..., A_N$ be the set of $N$ random variables representing a sequence of $N$ amino acids for example. Each random variable, $A_i$, takes on values in some finite or discrete domain. The domain, or alphabet as it sometime referred to, represents the $q$ possible states or realizations of $A_i$. Let the probability mass function, $P(A_1, ..., A_N)$, be the joint probability for some specific sequence of random variables. Therefore $\sum_s P(A_1, ..., A_N) = 1$ where $s$ is the universe of all unique sequences.

For a Bayesian network like the one depicted on the left hand side of Figure 3.1, the probability mass function is defined as

$$ P(A_1, ..., A_N) = \frac{1}{Z} \prod_i P(A_i | par(A_i)) $$

(3.1)

where $P(A_i | par(A_i))$ is known as the local conditional probability associated with node $i$. $A_i$ is the random variable at node $i$, $par(A_i)$ are the parents of node $i$ and $Z$ is the partition function. For nodes with no parents, the empty set are its parents.

On the other hand, in a MRF, the probability mass function can be written out as a product of potentials.

$$ P(A_1, ..., A_N) = \frac{1}{Z} \prod_i \psi_i $$

(3.2)
where the index \(i\), labels the set of potential functions, \(\psi_i\), which describe probabilistic relationships between variables and can act on any sized subset of the random variables, \(A_1, ..., A_N\). If the potential acts on only one node, it is called a field potential, \(\psi_i(A_i)\), and it describes the probabilistic tendencies for the states on the particular node on which it is acting. If the potential acts on two nodes, it is known as a coupling potential, \(\psi_{ij}(A_i, A_j)\), and it describes the probabilistic tendencies for the pairs of states belonging to the two nodes.

The next few sections of this chapter describe the basics of probabilistic inference on simple undirected graphs like the Markov Random Field described above. From the basics, we will then derive the belief propagation equations.

### 3.3 Examples of simple graphical models

Below are some examples and descriptions of several simple graphical models

![Figure 3.2: The simplest graphical model possible consists of only one node. Its graphical structure tells us everything about the class of probability functions that describe the only random variable within this graph, \(A_1\). For this probabilistic graphical model, the joint probability \(P(A_1) = \frac{1}{Z} \psi_1(A_1)\) where \(\psi_1(A_1)\) is the field potential at node 1.](image-url)
Figure 3.3: This simple graphical model consists of two nodes. The absence of an edge between $A_1$ and $A_2$ tells us that there is no probabilistic dependence between $A_1$ and $A_2$. For this probabilistic graphical model, the joint probability $P(A_1, A_2) = \frac{1}{2} \psi_1(A_1) \psi_2(A_2)$ where $\psi_1(A_2)$ and $\psi_2(A_2)$ are the field potentials at nodes 1 and 2.

Figure 3.4: This simple graphical model consists of two nodes. The presence of an edge between $A_1$ and $A_2$ tells us that there is a probabilistic dependence between $A_1$ and $A_2$. For this probabilistic graphical model, the joint probability $P(A_1, A_2) \approx \psi_{12}(A_1, A_2)$ where $\psi_{12}(A_1, A_2)$ is the coupling potentials between nodes 1 and 2. In other words, the presence of an edge describes a statistical dependency between the variables.

Figure 3.5: This three node graph contains only two edges, and is therefore acyclic. The graph structure indicates that if we condition on the variable $A_2$, then $A_1$ and $A_3$ are conditionally independent. As a result, the joint probability for this graph will have a specific structure. For this probabilistic graphical model, the joint probability $P(A_1, A_2, A_3) \approx \psi_{12}(A_1, A_2) \psi_{23}(A_2, A_3)$ where $\psi_{12}(A_1, A_2)$ and $\psi_{23}(A_2, A_3)$ are the coupling potentials. It is this conditional independence feature that will be exploited in belief propagation.
Figure 3.6: This graph contains three nodes and three edges where every node is connected to every other node. This is known as a completely cyclic graph. There is no conditional independence structure in this graph. For this probabilistic graphical model, the joint probability $P(A_1, A_2, A_3) \approx \psi_{123}(A_1, A_2, A_3)$ where $\psi_{123}(A_1, A_2, A_3)$ is the triplet potential. It is not exactly correct to describe the joint probability distribution of this graph in terms of the coupling potentials, where $P(A_1, A_2, A_3) \approx \psi_{12}(A_1, A_2)\psi_{23}(A_2, A_3)\psi_{13}(A_1, A_3)$. However, in some situations, including in this work, it is a reasonably accurate approximation.
3.4 Basic probabilistic inference

Consider the representation of a three residue peptide in Figure 3.7, with each node representing a single amino acid. This single-chained undirected graph contains no loops or cycles, indicating that we are not trying to model any cyclic interactions except for the interactions between adjacent nodes. Only neighboring nodes are linked to each other and in the absence of field potentials, the joint probability distribution over this MRF using Equation 3.2 can be described by the product of the coupling potentials.

\[
P(A) = \frac{1}{Z}\psi_{1,2}(A_1, A_2)\psi_{2,3}(A_2, A_3)
\]  

(3.3)

Exhaustive enumeration of this equation can be used to determine the partition function, and the marginal probability of any particular amino acid, \(A_i\), at position \(i\).

3.5 Exhaustive enumeration

We begin by considering the inference problem of determining the exact marginal distribution of a single mutation at a particular site for the graphical in Figure 3.7. For example, in order to find the marginal of the state \(a_1\) at position 1, we need to take the sum of the probabilities of all sequences that contain \(a_1\) at position 1.

\[
P(a_1) = \sum_{A_2} \sum_{A_3} P(A)
\]  

(3.4)
In this particular example, let the number of states be equal to $q = 2$. That is, the specific amino acid $A_i$ at position $i$ can be either $a_1^i$ or $a_2^i$. For two states and three positions, there are a total of eight unique sequences: $a_1^1a_2^1a_3^1$, $a_1^1a_2^1a_3^2$, $a_1^2a_2^1a_3^1$, $a_1^2a_2^1a_3^2$, $a_1^1a_2^2a_3^1$, $a_1^1a_2^2a_3^2$, $a_1^2a_2^2a_3^1$, and $a_1^2a_2^2a_3^2$. In general for $N$ positions and $q$ states, there are a total of $q^N$ unique sequences, a number that scales exponentially with $N$. Of these eight sequences, four contain the state $a_1^1$ at position 1, therefore the marginal probability of $a_1^1$ is equal to the sum of four probabilities.

$$P(a_1^1) = P(a_1^1a_2^1a_3^1) + P(a_1^1a_2^1a_3^2) + P(a_1^1a_2^2a_3^1) + P(a_1^1a_2^2a_3^2)$$

With potential functions, the partition function needs to be determined in order to normalize the data so it is not $q^{N-1}$ calculations. In general, the marginal probability at a single site using exhaustive enumeration costs $q^N$ computations. The nodes in the example above have no field potentials and as a result, the marginal probability can be written in terms of pair potentials.

$$P(a_1^1) = \frac{1}{Z} \sum_{A_2} \sum_{A_3} \psi_{1,2}(a_1^1, A_2)\psi_{2,3}(A_2, A_3)$$

(3.5)

The summation in the example of this particular exhaustive enumeration involves the probabilities of the four unique sequences containing the state $a_1^1$ at position 1.

$$P(a_1^1) = P(a_1^1a_2^1a_3^1) + P(a_1^1a_2^1a_3^2) + P(a_1^1a_2^2a_3^1) + P(a_1^1a_2^2a_3^2)$$

(3.6)

Each of these sequences has probability associated with it, which can be expressed in terms of the potential functions. For example, $P(a_1^1a_2^1a_3^1) \approx \psi_{1,2}(a_1^1, a_2^1)\psi_{2,3}(a_2^1, a_3^1)$.

As a result, the marginal probability for $P(a_1^1)$ becomes four individual sums.
of the products of potentials.

\[ P(a_1^1) = \frac{1}{Z} \sum_{a_2^1, a_3^1} \psi_{1,2}(a_1^1, a_2^1) \psi_{2,3}(a_2^1, a_3^1) + \psi_{1,2}(a_1^2, a_2^1) \psi_{2,3}(a_2^1, a_3^1) + \psi_{1,2}(a_1^1, a_2^2) \psi_{2,3}(a_2^2, a_3^1)\]  

(3.7)

where the partition function, \( Z \), is equal to

\[ Z = P(a_1^1 a_2^1 a_3^1) + P(a_1^1 a_2^3 a_3^1) + P(a_1^3 a_2^1 a_3^1) + P(a_1^3 a_2^3 a_3^1) \]

\[ + P(a_1^1 a_2^1 a_3^2) + P(a_1^1 a_2^3 a_3^2) + P(a_1^3 a_2^1 a_3^2) + P(a_1^3 a_2^3 a_3^2) \]

(3.8)

\[ Z = \psi_{1,2}(a_1^1, a_2^1) \psi_{2,3}(a_2^1, a_3^1) + \psi_{1,2}(a_1^2, a_2^1) \psi_{2,3}(a_2^1, a_3^1) + \psi_{1,2}(a_1^1, a_2^2) \psi_{2,3}(a_2^2, a_3^1)\]

(3.9)

Subsequently we will explicitly show that this summation above is equivalent to the Bethe approximation in belief propagation for tree-like acyclic graphs.

### 3.6 The elimination algorithm

In this subsection, we will introduce a basic inference algorithm called elimination. The crux of the idea behind elimination is that one can exploit the independent nature of the specific potential functions in a graphical model, thus allowing for efficient summation. For example, in the probability mass function of Equation 3.3 which describes the probability distribution of the graphical model in Figure 3.7, summation over \( A_3 \) applies only to the potential function \( \psi_{2,3}(A_2, A_3) \). Therefore we can perform the summation of \( A_3 \) over only \( \psi_{2,3}(A_2, A_3) \), and not over both \( \psi_{1,2}(A_1, A_2) \) and \( \psi_{2,3}(A_2, A_3) \).

\[ P(a_1^1) = \frac{1}{Z} \sum_{A_2} \sum_{A_3} \psi_{1,2}(a_1^1, A_2) \psi_{2,3}(A_2, A_3) \]

(3.10)
\[ P(a_1) = \frac{1}{Z} \left[ \sum_{A_2} \psi_{1,2}(a_1, A_2) \right] \left[ \sum_{A_3} \psi_{2,3}(A_2, A_3) \right] \]  
(3.11)

This method is referred to in the literature as the elimination algorithm since the first step of the factorized summation effectively removes \( A_3 \) from the overall summation and subsequently the graph. \( \sum_{A_3} \psi_{2,3}(A_2, A_3) \) results in a marginalization over the random variable \( A_3 \), which can then be multiplied by \( \psi_{1,2}(A_1, A_2) \) in a summation over \( A_2 \). In larger graphs, factorization of similar independent terms can be introduced, such that every time a summation over the states of a node is performed, that node is eliminated and is effectively removed from the graph. For example, the marginal probability of \( P(a_1) \) in the exhaustive enumeration above in Equation 3.7 can be further factorized as follows.

\[
P(a_1) = \frac{1}{Z} [\psi_{1,2}(a_1, a_2^1)\psi_{2,3}(a_2^1, a_3^1) + \psi_{1,2}(a_1, a_2^2)\psi_{2,3}(a_2^2, a_3^1) + \psi_{1,2}(a_1, a_2^3)\psi_{2,3}(a_2^3, a_3^2)]
\]  
(3.12)

\[
P(a_1) = \frac{1}{Z} [\psi_{1,2}(a_1^1, a_2^1)[\psi_{2,3}(a_2^1, a_3^1) + \psi_{2,3}(a_2^1, a_3^2)] + \psi_{1,2}(a_1^1, a_2^2)[\psi_{2,3}(a_2^2, a_3^1) + \psi_{2,3}(a_2^2, a_3^2)]]
\]  
(3.13)

which simplifies to

\[
P(a_1) = \frac{1}{Z} [\psi_{1,2}(a_1^1, a_2^1)[\sum_{A_3} \psi_{2,3}(a_2^1, A_3)] + \psi_{1,2}(a_1^1, a_2^2)[\sum_{A_3} \psi_{2,3}(a_2^2, A_3)]]
\]  
(3.14)

This equation above has the flavor of matrix multiplication

\[
P(a_1) = \frac{1}{Z} \begin{bmatrix} \psi_{1,2}(a_1^1, a_2^1) & \psi_{1,2}(a_1^1, a_2^2) \end{bmatrix} \begin{bmatrix} \sum_{A_3} \psi_{2,3}(a_2^1, A_3) \\ \sum_{A_3} \psi_{2,3}(a_2^2, A_3) \end{bmatrix}
\]  
(3.15)

and since

\[
\begin{bmatrix} \sum_{A_3} \psi_{2,3}(a_2^1, A_3) \\ \sum_{A_3} \psi_{2,3}(a_2^2, A_3) \end{bmatrix} = \sum_{A_3} \psi_{2,3}(A_2, A_3)
\]
and

\[
\begin{bmatrix}
\psi_{1,2}(a_1^1, a_2^1) & \psi_{1,2}(a_1^1, a_2^2)
\end{bmatrix} = \sum_{A_2} \psi_{1,2}(a_1^1, A_2)
\]

\[
P(a_1) = \frac{1}{Z} \left[ \sum_{A_2} \psi_{1,2}(a_1^1, A_2) \left[ \sum_{A_3} \psi_{2,3}(A_2, A_3) \right] \right]
\]

which of course is identical to Equation 3.11. This factorization leads to a substantial reduction in computations. There are four summations in total, each of which is a summation over a $2 \times 2$ matrix. In general, for $N$ positions, the total cost of evaluating a marginal is $O(NQ^2)$, linear with respect to the number of positions, much more efficient than the exponential cost of the previous approach. The reason factorization is so useful can be illustrated by the simple example below. Multiplication is distributive over addition, so something like

\[
ab + ac = a(b + c)
\]

(3.17)

can be rearranged such that there is one fewer computation on the right-hand side than on the left-hand side. This is the key concept that is used repeatedly for inference in graphs.

### 3.7 Introduction to messages

A second point to note is that factorization can lead to products of incoming marginalizations for nodes with more than one incoming edge. This can be highlighted by factorizing the exhaustive enumeration for the marginal probability of $P(a_2^1)$.

\[
P(a_2^1) = \frac{1}{Z} \sum_{A_1} \sum_{A_3} \psi_{1,2}(A_1, a_2^1) \psi_{2,3}(a_2^1, A_3)
\]

(3.18)

Node 2 has edges that connect to both node 1 and node 3. Factorization of the marginal probability at node 2 results in two incoming marginalizations over
the variables in the leaf nodes, node 1 and 3.

\[ P(a_1^2) = \frac{1}{Z} \left[ \psi_{1,2}(a_1^1, a_2^1) \psi_{2,3}(a_2^1, a_3^2) + \psi_{1,2}(a_1^1, a_2^1) \psi_{2,3}(a_2^2, a_3^1) + \psi_{2,3}(a_2^1, a_3^1) \right] \]  
(3.19)

\[ P(a_2^1) = \frac{1}{Z} \left[ \psi_{1,2}(a_1^1, a_2^1) \psi_{2,3}(a_2^1, a_3^1) + \psi_{1,2}(a_1^1, a_2^1) \psi_{2,3}(a_2^2, a_3^1) \right] \]  
(3.20)

\[ P(a_2^2) = \frac{1}{Z} \left[ \psi_{1,2}(a_1^1, a_2^1) \psi_{2,3}(a_2^1, a_3^1) + \psi_{2,3}(a_2^1, a_3^1) \right] \]  
(3.21)

\[ P(a_2^3) = \frac{1}{Z} \left[ \sum_{A_1} \psi_{1,2}(A_1, a_1^1) \right] \left[ \sum_{A_3} \psi_{2,3}(a_2^1, A_3) \right] \]  
(3.22)

or we can simply rewrite this equation in terms of the random variable \( A_2 \) itself.

\[ P(A_2) = \frac{1}{Z} \left[ \sum_{A_1} \psi_{1,2}(A_1, A_2) \right] \left[ \sum_{A_3} \psi_{2,3}(A_2, A_3) \right] \]  
(3.23)

Let \( \mu_{1 \to 2}(A_2) \) and \( \mu_{3 \to 2}(A_2) \) refer to the intermediate terms that arise while performing these sums. These incoming marginalizations over the variables of the incoming nodes are called \textit{messages}.

\[ P(A_2) = \frac{1}{Z} \left[ \sum_{A_1} \psi_{1,2}(A_1, A_2) \right] \left[ \sum_{A_3} \psi_{2,3}(A_2, A_3) \right] \]  
(3.24)

For example, the marginal probability of one specific state of random variable \( A_2 \) is proportional to the product of the message from node 1 to node 2, \( \mu_{1 \to 2}(A_2) \), and the message from node 3 to node 2, \( \mu_{3 \to 2}(A_2) \).

\[ P(A_2) = \frac{1}{Z} \mu_{1 \to 2}(A_2) \mu_{3 \to 2}(A_2) \]  
(3.25)
And in general, the marginal probability of any state at position \( i \), which is described by the random variable \( A_i \), is proportional to the product of all incoming messages.

\[
P(A_i) \approx \prod_{j \in N(i)} \mu_{j \rightarrow i}(A_i)
\]  

(3.26)

where \( j \) is a neighboring node of \( i \). If the network is larger than the simple one depicted above, each incoming message can be further decomposed into its incoming messages. Further examples of this message decomposition can be found in the next subsection.

### 3.8 Decomposition of messages

Figure 3.8 below shows an acyclic network of five adjacent nodes. Only neighboring nodes are linked to each other and the joint distribution over this MRF using Equation 3.2 can be described a product of four potential functions, one for each pair of adjacent nodes.

\[
P(A) = \frac{1}{Z} \psi_{1,2}(A_1, A_2) \psi_{2,3}(A_2, A_3) \psi_{3,4}(A_3, A_4) \psi_{4,5}(A_4, A_5)
\]

(3.27)

We wish to find the marginal probability of a state at position 3, which is described by the random variable \( A_3 \). In this example, we will begin with the exact and exhaustive enumeration, factorize the enumeration into messages of messages.

![Undirected acyclic graph with 5 nodes](image)

Figure 3.8: Undirected acyclic graph with 5 nodes

We begin by considering the inference problem of finding the marginal distribution of a single mutation. For example, in order to find the marginal of
a particular state of the random variable $A_3$, one needs to take the sum of the joint distribution over all states at each node, except for the state of interest at $A_3$.

$$P(A_3) = \sum_{A_1} \sum_{A_2} \sum_{A_4} \sum_{A_5} P(A)$$  \hspace{1cm} (3.28)

For 3 possible states per site, there are $3^5 = 243$ values for $P(A)$, therefore explicit summation is quite cumbersome. In general for $N$ positions and $Q$ states, there are $Q^N$ terms in the joint distribution, a number that scales exponentially with $N$. However, one can exploit the independent nature of the potentials in the graph, which allows for a much more efficient summation. For example, the summation over $A_5$ applies only to the potential function $\psi_{4,5}(A_4, A_5)$, therefore we can perform the summation over $A_5$ only over $\sum_{A_5} \psi_{4,5}(A_4, A_5)$ and not over all the four potentials. As described in the section above, this method is referred to in the literature as the elimination algorithm since the factorized summation effectively removes $A_5$ from the overall summation, and subsequently the graph. The result is a function over the variable $A_4$, which can then be further marginalized in a summation over $A_4$. Similarly, factorization of other terms can be introduced, such that every time a summation over a variable is performed, that variable is eliminated and the node associated with that variable is removed from the graph. If the potential functions which describe the joint distribution $P(A)$ are grouped together with their corresponding summations, the marginal calculation for $P(A_3)$ can be factorized as follows

$$P(A_3) = \frac{1}{Z} \left[ \sum_{A_2} \psi_{2,3}(A_2, A_3) \left[ \sum_{A_1} \psi_{1,2}(A_1, A_2) \right] \right] \left[ \sum_{A_4} \psi_{3,4}(A_3, A_4) \left[ \sum_{A_5} \psi_{4,5}(A_4, A_5) \right] \right] \mu_{2\rightarrow3}(A_3) \mu_{4\rightarrow3}(A_3)$$  \hspace{1cm} (3.29)

A message is essentially a vector of beliefs or probabilities, marginalized over an eliminated node, passed from the eliminated node to the receiving node. In this example, the marginal probability of $A_3$ is proportional to the product
of the message from node 2 to node 3 and the message from node 4 to node 3.

\[ P(A_3) = \frac{1}{Z} \mu_{2\rightarrow3}(A_3)\mu_{4\rightarrow3}(A_3) \]  

(3.30)

In fact each incoming message can be further decomposed into its incoming messages. For example, the isolated message from node 2 to node 3 is a marginal summation over the product of the potential function at node 3 and the message from node 1 to node 2.

\[ \mu_{2\rightarrow3}(A_3) = \sum A_2 \psi_{2,3}(A_2, A_3)[\sum A_1 \psi_{1,2}(A_1, A_2)] \]

(3.31)

where the message from node 1 to node 2 is the summation over \( A_1 \)

\[ \mu_{1\rightarrow2}(A_2) = \sum_{A_1} \psi_{1,2}(A_1, A_2) \]  

(3.32)

Similarly, the message from node 4 to node 3 can be factorized further into a product of its incoming messages.

\[ \mu_{4\rightarrow3}(A_3) = \sum A_4 \psi_{3,4}(A_3, A_4)[\sum A_5 \psi_{4,5}(A_4, A_5)] \]

(3.33)

where the message from node 5 to node 4 is

\[ \mu_{5\rightarrow4}(A_4) = \sum_{A_5} \psi_{4,5}(A_4, A_5) \]  

(3.34)

In general messages can be written as

\[ \mu_{j\rightarrow i}(A_i) = \sum_{A_j} \psi_{i,j}(A_i, A_j) \prod_{k \in N(j) \setminus i} \mu_{k\rightarrow j}(A_j) \]  

(3.35)

This message passing protocol is summarized in Figure 3.9.
3.9 Determination of bivariate marginals using Elimination

These message passing factorizations that are described can also be applied to infer bivariate marginals. For example, the joint probability distribution of mutations $A_2$ and $A_3$ can be written as a product of the message from node 1 to node 2, the message from node 4 to node 3, the potential between node 2 and 3 and the normalization constant.

$$P(A_2, A_3) = \frac{1}{Z} \sum_{A_1} \sum_{A_4} \sum_{A_5} P(A)$$

$$= \frac{1}{Z} \sum_{A_1} \sum_{A_4} \sum_{A_5} \psi_{1,2}(A_1, A_2) \psi_{2,3}(A_2, A_3) \psi_{3,4}(A_3, A_4) \psi_{4,5}(A_4, A_5)$$

$$= \frac{1}{Z} \psi_{2,3}(A_2, A_3) \sum_{A_1} \psi_{1,2}(A_1, A_2) \sum_{A_4} \psi_{3,4}(A_3, A_4) \sum_{A_5} \psi_{4,5}(A_4, A_5)$$

$$= \frac{1}{Z} \psi_{2,3}(A_2, A_3) \mu_{1\rightarrow2}(A_2) \mu_{4\rightarrow3}(A_3)$$

(3.36)

And in general, for a pair of mutations in a tree-like graph with $N-1$ nodes, the bivariate marginal is written as the product of the incoming messages to each node, the potential function between the nodes and a normalizing factor

$$P(A_{n-1}, A_n) = \frac{1}{Z} \mu_{n-2\rightarrow n-1}(A_{n-1}) \psi_{n-1,n}(A_{n-1}, A_n) \mu_{n+1\rightarrow n}(A_n)$$

(3.37)

For tree-like graphs such as the one in Figure 3.9, this marginalization is simple to compute. But in complete graphs that are fully connected, this marginalization requires further approximations.
3.10 Introduction to belief propagation

The elimination inference method described in Section 3.4 yields only a single marginal. In order to calculate marginals for multiple sites, we could run the elimination algorithm several times, but that would be highly inefficient as several factors would need to be recalculated repeatedly. Instead of determining each marginal using Elimination, one can pass local messages throughout all the nodes in the tree and repeat this process till the messages converge. The converged messages can then yield all the marginals. This process is called Belief Propagation (BP) or the Sum-Product algorithm [125, 56]. It is an iterative procedure applied to efficiently calculate all the marginals in a tree-like graph[82]. It consists of leaf nodes passing messages to their parents, which in turn process these messages and pass them onwards towards the root. The root then sends messages back to its children nodes and so on until the messages

Figure 3.10: Example of belief propagation messages in an undirected acyclic graph with 5 nodes
eventually reach the leaf nodes. At this point, for acyclic graphs, the messages will converge\[126, 56\]. For cyclic graphs, this method is approximate but may converge after several cycles of message passing\[72, 47, 50\]. The messages update rules are as follows. To compute the message from node $j$ to node $i$:

1. Multiple together all the messages coming into node $j$, except for the message coming from node $i$ to node $j$ (For a vector of messages, take the scalar product).

2. Multiply by the pair potential.

3. Marginalize over $A_i$ (matrix multiplication will handle the marginalization).

The message update equation is

$$
\mu_{j\rightarrow i}(A_i) = \sum_{A_j} \psi_{i,j}(A_i, A_j) \prod_{k \in N(j) \setminus i} \mu_{k\rightarrow j}(A_j) \tag{3.38}
$$

where $k \in N(j) \setminus i$ indicates that choose any node $k$ which is a neighbor of node $j$, except node $i$. To compute the marginal probability at node $i$, we take the product of all incoming messages into node $i$.

$$
P(A_i) \approx \prod_{j \in N(i)} \mu_{j\rightarrow i}(A_i) \tag{3.39}
$$

If a field potential exists at node $i$, it is also multiplied to the messages.

$$
P(A_i) \approx \psi_i(A_i) \prod_{j \in N(i)} \mu_{j\rightarrow i}(A_i) \tag{3.40}
$$
3.11 Message update equations in matrix notation

For the simple tree-like graph in Figure 3.7, we can determine the marginal
probability of one of possible realizations of the random variable $A_1$, using the
message update equation:

$$
\mu_{j\rightarrow i}(A_i) = \sum_{A_j} \psi_{ij}(A_i, A_j) \prod_{k \in N(j) \setminus i} \mu_{k\rightarrow j}(A_j)
$$

and belief equation:

$$
P(A_i) \approx \prod_{j \in N(i)} \mu_{j\rightarrow i}(A_i)
$$

where $k \in N(j) \setminus i$ indicates that choose any node $k$ which is a neighbor of node
j, except node $i$. In this work, implementation of belief and update equations
is done with the help of matrices. Therefore, in matrix notation, the message
update equations can be written in terms of vectors of messages, $\overline{\mu}_{j\rightarrow i}(A_i)$, of
size $q \times 1$, where $q$ is the number of possible states, $a_1^i, ..., a_q^i$, that the random
variable $A_i$ can take at position $i$. The belief propagation update equations in
matrix notation are:

$$
\overline{\mu}_{j\rightarrow i}(A_i) = \sum_{A_j} \psi_{ij}(A_i, A_j) \prod_{k \in N(j) \setminus i} \overline{\mu}_{k\rightarrow j}(A_j)
$$

and belief equation:

$$
P(A_i) \approx \prod_{j \in N(i)} \overline{\mu}_{j\rightarrow i}(A_i)
$$

where the products of incoming messages, $\prod_{k \in N(j) \setminus i} \overline{\mu}_{k\rightarrow j}(A_j)$ or $\prod_{j \in N(i)} \overline{\mu}_{j\rightarrow i}(A_i)$,
in terms of matrix algebra, is a Hadamard multiplication of $q \times 1$ dimensional
vectors resulting in a $q \times 1$ dimensional incoming message vector.

The Hadamard matrix product, named after the French mathematician
Jacques Hadamard, is a matrix multiplication between two $m \times n$ dimensional
matrices, whose product also has the same dimensions. In a Hadamard multiplication, each element, $e_{ij}$, of the product matrix is a product of the same elements, $e_{ij}$, of the parent matrices. For example the Hadamard product of the following $3 \times 1$ matrices is:

$$
\begin{bmatrix}
a_1 \\
a_2 \\
a_3
\end{bmatrix} \circ 
\begin{bmatrix}
b_1 \\
b_2 \\
b_3
\end{bmatrix} = 
\begin{bmatrix}
a_1b_1 \\
a_2b_2 \\
a_3b_3
\end{bmatrix}
$$

where the $\circ$ operator represents a Hadamard matrix multiplication.

On the other hand, the marginal sum over the variable $A_j$, in terms of matrices, is a matrix multiplication of the $q \times q$ matrix of coupling potentials, $\psi_{ij}(A_i, A_j)$ with a $q \times 1$ vector of incoming messages from the output of the Hadamard multiplication. For example, let node $i$ be connected to node $j$, and we wish to find the message from node $j$ to $i$. Imagine there only two incoming edges into node $j$, besides the edge from $i$ itself. These edges come from the nodes $k$ and $l$. For simplicity, let nodes $k$ and $l$ be leaf nodes. Therefore, in this example, the message from $j$ to $i$ can be written in terms of the Hadamard product of the incoming messages into node $j$ from nodes $k$ and $l$ and the potential between $i$ and $j$, $\psi_{ij}(A_i, A_j)$. The update equation for the message $\mu_{j \rightarrow i}(A_i)$ thus has the form:

$$
\mu_{j \rightarrow i}(A_i) = \psi_{i,j}(A_i, A_j) \left[ \mu_{k \rightarrow j}(A_j) \circ \mu_{k \rightarrow j}(A_j) \right]
$$

If $q = 3$, the random variable $A_i$ can exist in the states $a^1_i$, $a^2_i$ and $a^3_i$. In matrix format, $\mu_{k \rightarrow j}(A_j) \circ \mu_{k \rightarrow j}(A_j)$ looks like:

$$
\mu_{k \rightarrow j}(A_j) \circ \mu_{k \rightarrow j}(A_j) = 
\begin{bmatrix}
\sum_{q=1}^{3} \psi_{jk}(a^1_j, a^q_k) \\
\sum_{q=1}^{3} \psi_{jk}(a^2_j, a^q_k) \\
\sum_{q=1}^{3} \psi_{jk}(a^3_j, a^q_k)
\end{bmatrix} \circ 
\begin{bmatrix}
\sum_{q=1}^{3} \psi_{jl}(a^1_j, a^q_l) \\
\sum_{q=1}^{3} \psi_{jl}(a^2_j, a^q_l) \\
\sum_{q=1}^{3} \psi_{jl}(a^3_j, a^q_l)
\end{bmatrix}
$$
where each of the marginal sums, $\sum_{q=1}^{3} \psi_{jk}(a_{1j}^{1}, a_{k}^{q})$ are equal to $\psi_{jk}(a_{1j}^{1}, a_{1k}^{q}) + \psi_{jk}(a_{1j}^{1}, a_{2k}^{q}) + \psi_{jk}(a_{1j}^{1}, a_{3k}^{q})$. The result of the Hadamard multiplication is simply:

$$\mu_{k \rightarrow j}(A_{j}) = \frac{1}{\mu_{k \rightarrow j}(A_{j})} = \left[ \begin{array}{c}
\sum_{q=1}^{3} \psi_{jk}(a_{1j}^{1}, a_{k}^{q}) \\
\sum_{q=1}^{3} \psi_{jk}(a_{1j}^{2}, a_{k}^{q}) \\
\sum_{q=1}^{3} \psi_{jk}(a_{1j}^{3}, a_{k}^{q})
\end{array} \right] \left[ \begin{array}{c}
\sum_{q=1}^{3} \psi_{jl}(a_{1j}^{1}, a_{l}^{q}) \\
\sum_{q=1}^{3} \psi_{jl}(a_{1j}^{2}, a_{l}^{q}) \\
\sum_{q=1}^{3} \psi_{jl}(a_{1j}^{3}, a_{l}^{q})
\end{array} \right]$$

and the message update equation, $\mu_{j \rightarrow i}(A_{i})$ simplifies to the following matrix multiplication

$$
\begin{bmatrix}
\mu_{j \rightarrow i}(a_{1}^{1}) \\
\mu_{j \rightarrow i}(a_{1}^{2}) \\
\mu_{j \rightarrow i}(a_{1}^{3})
\end{bmatrix} =
\begin{bmatrix}
\psi_{ij}(a_{1}^{1}, a_{j}^{1}) & \psi_{ij}(a_{1}^{1}, a_{j}^{2}) & \psi_{ij}(a_{1}^{1}, a_{j}^{3}) \\
\psi_{ij}(a_{1}^{2}, a_{j}^{1}) & \psi_{ij}(a_{1}^{2}, a_{j}^{2}) & \psi_{ij}(a_{1}^{2}, a_{j}^{3}) \\
\psi_{ij}(a_{1}^{3}, a_{j}^{1}) & \psi_{ij}(a_{1}^{3}, a_{j}^{2}) & \psi_{ij}(a_{1}^{3}, a_{j}^{3})
\end{bmatrix}
\begin{bmatrix}
\sum_{q=1}^{3} \psi_{jk}(a_{1}^{1}, a_{k}^{q}) \\
\sum_{q=1}^{3} \psi_{jk}(a_{1}^{2}, a_{k}^{q}) \\
\sum_{q=1}^{3} \psi_{jk}(a_{1}^{3}, a_{k}^{q})
\end{bmatrix}
\begin{bmatrix}
\sum_{q=1}^{3} \psi_{jl}(a_{1}^{1}, a_{l}^{q}) \\
\sum_{q=1}^{3} \psi_{jl}(a_{1}^{2}, a_{l}^{q}) \\
\sum_{q=1}^{3} \psi_{jl}(a_{1}^{3}, a_{l}^{q})
\end{bmatrix}
$$

In the next subsection, we will apply these BP update and belief equations, in matrix notation, to calculate the marginal probability of a particular state at node 1 from the graph Figure 3.7.

### 3.12 Marginal probability calculations with the message update equations

Using the BP belief and update equations, we can determine the marginal probability of the specific state $a_{1}^{1}$ at node 1 in terms of its incoming messages:

$$P(a_{1}^{1}) \approx \prod_{j \in N(i)} \mu_{2 \rightarrow 1}(a_{1}^{1})$$  \hspace{1cm} (3.45)$$

As Figure 3.7 indicates, there is only one incoming message into node 1, and that is from node 2. Therefore the belief at node 1 is proportional to message from node 2 to node 1.

$$P(a_{1}^{1}) \approx \mu_{2 \rightarrow 1}(a_{1}^{1})$$  \hspace{1cm} (3.46)$$
Using the message update rules, the message from node 2 to node 1 can be expanded out as the product of the marginal sum over the variables at node 2 and the incoming messages into node 2, which in this example, is only one message; the message from node 3 to node 2.

$$\mu_{2\to1}(a_1^1) = \sum_{A_2} \psi_{1,2}(a_1^1, A_2) \mu_{3\to2}(A_2)$$

(3.47)

Therefore the marginal probability at node 1 can be rewritten in terms of the messages coming from the leaf node, node 3, into node 1.

$$P(a_1^1) \approx \sum_{A_2} \psi_{1,2}(a_1^1, A_2) \mu_{3\to2}(A_2)$$

(3.48)

According to the update rules, the message from node 3 to node 2 is the product of all incoming messages into node 3, except for the one coming from node 2. Therefore since node 3 is a leaf node, the message from node 3 to node 2 is just a marginal sum.

$$\mu_{3\to2}(A_2) = \sum_{A_3} \psi_{2,3}(A_2, A_3)$$

(3.49)

giving

$$P(a_1^1) \approx \left[ \sum_{A_2} \psi_{1,2}(a_1^1, A_2) \left[ \sum_{A_3} \psi_{2,3}(A_2, A_3) \right] \right]$$

(3.50)

If the number of states, $q = 2$, all the coupling potentials are $2 \times 2$ matrices and the message from node 3 to node 2, $\mu_{3\to2}(A_2)$, can be written in terms of these matrices.

$$\sum_{A_3} \psi_{2,3}(A_2, A_3) = \sum_{A_3} \begin{bmatrix} \psi_{2,3}(a_2, a_3) & \psi_{2,3}(a_2, b_3) \\ \psi_{2,3}(b_2, a_3) & \psi_{2,3}(b_2, b_3) \end{bmatrix}$$

(3.51)

$$\mu_{3\to2}(A_2) = \begin{bmatrix} \psi_{2,3}(a_2, a_3) + \psi_{2,3}(a_2, b_3) \\ \psi_{2,3}(b_2, a_3) + \psi_{2,3}(b_2, b_3) \end{bmatrix}$$

(3.52)
\[ \mu_{3\rightarrow2}(A_2) = \left[ \sum_{A_3} \psi_{2,3}(a_2, A_3) \right] \]

This is the message from node 3 to node. Equation 3.48 shows that this message needs to be multiplied by the potential between node 1 and node 2 in order to get the marginal probability at node 1.

\[ P(a_1) \approx \sum_{A_2} \psi_{1,2}(a_1, A_2) \mu_{3\rightarrow2}(A_2) \]  

(3.54)

\[ P(a_1) \approx \left[ \psi_{1,2}(a_1, a_2) \psi_{1,2}(a_1, b_2) \right] \left[ \sum_{A_3} \psi_{2,3}(a_2, A_3) \right] \left[ \sum_{A_3} \psi_{2,3}(b_2, A_3) \right] \]  

(3.55)

\[ P(a_1) \approx \left[ \psi_{1,2}(a_1, a_2) \sum_{A_3} \psi_{2,3}(a_2, A_3) + \psi_{1,2}(a_1, b_2) \sum_{A_3} \psi_{2,3}(b_2, A_3) \right] \]  

(3.56)

This result is equivalent to the output of exhaustive enumeration presented earlier in Equation 3.7 for the marginal probability of \( P(a_i) \).

### 3.13 Examples of Belief Propagation on Undirected Acyclic Graphs

The following section will provide worked out examples of marginal probability calculations using exhaustive enumeration, simple mean field and the Bethe mean field (belief propagation).

The belief propagation update and belief rules are:

\[ \overrightarrow{\mu_{j\rightarrow i}}(A_i) = \sum_{A_j} \psi_{ij}(A_i, A_j) \prod_{k \in N(j) \setminus i} \overrightarrow{\mu_{k\rightarrow j}}(A_j) \]  

(3.57)
\[ P(A_i) \approx \prod_{j \in N(i)} \mu_{j \rightarrow i}(A_i) \quad (3.58) \]

### 3.13.1 3 nodes, single chain acyclic graph

This simple graph below has no fields and only coupling potentials between adjacent nodes. Moreover, the graph is acyclic and has only two states per node, \(a_i\) and \(b_i\). As a result, the coupling potentials are \(2 \times 2\) matrices.

\[
\psi_{1,2} = \begin{bmatrix}
\psi_{1,2}(a_1, a_2) & \psi_{1,2}(a_1, b_2) \\
\psi_{1,2}(b_1, a_2) & \psi_{1,2}(b_1, b_2)
\end{bmatrix} \quad (3.59)
\]

\[
\psi_{2,3} = \begin{bmatrix}
\psi_{2,3}(a_2, a_3) & \psi_{2,3}(a_2, b_3) \\
\psi_{2,3}(b_2, a_3) & \psi_{2,3}(b_2, b_3)
\end{bmatrix} \quad (3.60)
\]

![Figure 3.11: Undirected acyclic graph with 3 nodes](image)

### 3.13.1.1 Exhaustive Enumeration

\[
P(a_1) = \frac{1}{Z} \sum_{A_2} \sum_{A_3} \psi_{1,2}(a_1, A_2) \psi_{2,3}(A_2, A_3) \quad (3.61)
\]

\[
P(a_1) \approx [\psi_{1,2}(a_1, a_2) \psi_{2,3}(a_2, a_3) + \psi_{1,2}(a_1, a_2) \psi_{2,3}(a_2, b_3) + \psi_{1,2}(a_1, b_2) \psi_{2,3}(b_2, a_3) + \psi_{1,2}(a_1, b_2) \psi_{2,3}(b_2, b_3)] \quad (3.62)
\]
\[ P(a_1) \approx [\psi_{1,2}(a_1, a_2)[\psi_{2,3}(a_2, a_3) + \psi_{2,3}(a_2, b_3)] + \psi_{1,2}(a_1, b_2)[\psi_{2,3}(b_2, a_3) + \psi_{2,3}(b_2, b_3)]] \]

(3.63)

\[ P(a_1) \approx [ \psi_{1,2}(a_1, a_2) \sum_{A_3} \psi_{2,3}(a_2, A_3) + \psi_{1,2}(a_1, b_2) \sum_{A_3} \psi_{2,3}(b_2, A_3) ] \]

(3.64)

### 3.13.1.2 Belief Propagation

\[ P(a_1) \approx \mu_{2\rightarrow 1}(a_1) \]  

(3.65)

\[ P(a_1) \approx \sum_{A_2} \psi_{1,2}(a_1, A_2) \mu_{3\rightarrow 2}(A_2) \]  

(3.66)

\[ P(a_1) \approx \sum_{A_2} \psi_{1,2}(a_1, A_2) \sum_{A_3} \psi_{2,3}(A_2, A_3) \]  

(3.67)

\[ P(a_1) \approx \sum_{A_2} \psi_{1,2}(a_1, A_2) \begin{bmatrix} \sum_{A_3} \psi_{2,3}(a_2, A_3) \\ \sum_{A_3} \psi_{2,3}(b_2, A_3) \end{bmatrix} \]  

(3.68)

\[ P(a_1) \approx \begin{bmatrix} \psi_{1,2}(a_1, a_2) & \psi_{1,2}(a_1, b_2) \end{bmatrix} \begin{bmatrix} \sum_{A_3} \psi_{2,3}(a_2, A_3) \\ \sum_{A_3} \psi_{2,3}(b_2, A_3) \end{bmatrix} \]  

(3.69)

\[ P(a_1) \approx \begin{bmatrix} \psi_{1,2}(a_1, a_2) \sum_{A_3} \psi_{2,3}(a_2, A_3) + \psi_{1,2}(a_1, b_2) \sum_{A_3} \psi_{2,3}(b_2, A_3) \end{bmatrix} \]  

(3.70)

### 3.13.2 5 nodes, single chain acyclic graph

This simple graph below has no fields and only coupling potentials between adjacent nodes. Moreover, the graph is acyclic and has only two states per node, \( a_i \) and \( b_i \). As a result, the four coupling potentials are \( 2 \times 2 \) matrices.

\[ \psi_{1,2} = \begin{bmatrix} \psi_{1,2}(a_1, a_2) & \psi_{1,2}(a_1, b_2) \\ \psi_{1,2}(b_1, a_2) & \psi_{1,2}(b_1, b_2) \end{bmatrix} \]  

(3.71)
\[ \psi_{2,3} = \begin{bmatrix} \psi_{2,3}(a_2, a_3) & \psi_{2,3}(a_2, b_3) \\ \psi_{2,3}(b_2, a_3) & \psi_{2,3}(b_2, b_3) \end{bmatrix} \]  
\[ (3.72) \]

\[ \psi_{3,4} = \begin{bmatrix} \psi_{3,4}(a_3, a_4) & \psi_{3,4}(a_3, b_4) \\ \psi_{3,4}(b_3, a_4) & \psi_{3,4}(b_3, b_4) \end{bmatrix} \]  
\[ (3.73) \]

\[ \psi_{4,5} = \begin{bmatrix} \psi_{4,5}(a_4, a_5) & \psi_{4,5}(a_4, b_5) \\ \psi_{4,5}(b_4, a_5) & \psi_{4,5}(b_4, b_5) \end{bmatrix} \]  
\[ (3.74) \]

Figure 3.12: Undirected acyclic graph with 5 nodes

### 3.13.2.1 Exhaustive Enumeration

\[ P(a_3) = \frac{1}{Z} \sum_{A_1} \sum_{A_2} \sum_{A_4} \sum_{A_5} \psi_{1,2}(A_1, A_2) \psi_{2,3}(A_2, a_3) \psi_{3,4}(a_3, A_4) \psi_{4,5}(A_4, A_5) \]  
\[ (3.75) \]
\[ P(a_1) \approx \psi_{1,2}(a_1, a_2)\psi_{2,3}(a_2, a_3)\psi_{3,4}(a_3, a_4)\psi_{4,5}(a_4, a_5) + \psi_{1,2}(a_1, a_2)\psi_{2,3}(a_2, a_3)\psi_{3,4}(a_3, a_4)\psi_{4,5}(a_4, b_5) \\
+ \psi_{1,2}(a_1, a_2)\psi_{2,3}(a_2, a_3)\psi_{3,4}(a_3, b_4)\psi_{4,5}(b_4, a_5) + \psi_{1,2}(a_1, a_2)\psi_{2,3}(a_2, a_3)\psi_{3,4}(a_3, b_4)\psi_{4,5}(b_4, b_5) \\
+ \psi_{1,2}(a_1, b_2)\psi_{2,3}(b_2, a_3)\psi_{3,4}(a_3, a_4)\psi_{4,5}(a_4, a_5) + \psi_{1,2}(a_1, b_2)\psi_{2,3}(b_2, a_3)\psi_{3,4}(a_3, a_4)\psi_{4,5}(a_4, b_5) \\
+ \psi_{1,2}(a_1, b_2)\psi_{2,3}(b_2, a_3)\psi_{3,4}(a_3, b_4)\psi_{4,5}(b_4, a_5) + \psi_{1,2}(a_1, b_2)\psi_{2,3}(b_2, a_3)\psi_{3,4}(a_3, b_4)\psi_{4,5}(b_4, b_5) \\
+ \psi_{1,2}(a_1, b_2)\psi_{2,3}(b_2, b_3)\psi_{3,4}(b_3, a_4)\psi_{4,5}(a_4, a_5) + \psi_{1,2}(a_1, b_2)\psi_{2,3}(b_2, b_3)\psi_{3,4}(b_3, a_4)\psi_{4,5}(a_4, b_5) \\
+ \psi_{1,2}(a_1, b_2)\psi_{2,3}(b_2, b_3)\psi_{3,4}(b_3, b_4)\psi_{4,5}(b_4, a_5) + \psi_{1,2}(a_1, b_2)\psi_{2,3}(b_2, b_3)\psi_{3,4}(b_3, b_4)\psi_{4,5}(b_4, b_5) \]  

\[ P(a_1) \approx [\psi_{1,2}(a_1, a_2)[\psi_{2,3}(a_2, a_3) + \psi_{2,3}(a_2, b_3)] + \psi_{1,2}(a_1, b_2)[\psi_{2,3}(b_2, a_3) + \psi_{2,3}(b_2, b_3)]] \]  

\[ P(a_1) \approx [\psi_{1,2}(a_1, a_2)\sum_{A_3}\psi_{2,3}(a_2, A_3) + \psi_{1,2}(a_1, b_2)\sum_{A_3}\psi_{2,3}(b_2, A_3)] \]  

3.13.2.2 Belief Propagation

\[ P(a_1) \approx \mu_{2\rightarrow 1}(a_1) \]  

\[ P(a_1) \approx \sum_{A_2}\psi_{1,2}(a_1, A_2)\mu_{3\rightarrow 2}(A_2) \]  

\[ P(a_1) \approx \sum_{A_2}\psi_{1,2}(a_1, A_2)\sum_{A_3}\psi_{2,3}(A_2, A_3) \]
3.13.3 Belief propagation on a simple MRF

The following will be a numerical example of belief propagation. We will attempt to solve for the marginal probabilities in a simple graphical model of two-state random variables. A cartoon of the sample graph is show below in Figure 3.13. Again, for simplicity this graph contains no field parameters and consists of only coupling potentials between adjacent nodes. Since the ran-

$$P(a_1) \approx \sum_{A_2} \psi_{1,2}(a_1, A_2) \left[ \frac{\sum_{A_3} \psi_{2,3}(a_2, A_3)}{\sum_{A_3} \psi_{2,3}(b_2, A_3)} \right]$$  \hspace{1cm} (3.82)$$

$$P(a_1) \approx \left[ \psi_{1,2}(a_1, a_2) \psi_{1,2}(a_1, b_2) \right] \left[ \frac{\sum_{A_3} \psi_{2,3}(a_2, A_3)}{\sum_{A_3} \psi_{2,3}(b_2, A_3)} \right]$$  \hspace{1cm} (3.83)$$

$$P(a_1) \approx \left[ \psi_{1,2}(a_1, a_2) \sum_{A_3} \psi_{2,3}(a_2, A_3) + \psi_{1,2}(a_1, b_2) \sum_{A_3} \psi_{2,3}(b_2, A_3) \right]$$  \hspace{1cm} (3.84)$$

Figure 3.13: Undirected acyclic graph with four nodes and three pairwise potentials

random variables are two-state ($a_i$ and $b_i$), the three coupling potentials are $2 \times 2$
matrices.

\[
\psi_{1,2} = \begin{pmatrix}
\psi_{1,2}(a_1, a_2) & \psi_{1,2}(a_1, b_2) \\
\psi_{1,2}(b_1, a_2) & \psi_{1,2}(b_1, b_2)
\end{pmatrix}
\]

\[
\psi_{1,2}(a_1, a_2) = \begin{pmatrix}
1.0 \\
0.9
\end{pmatrix}
\]

\[
\psi_{2,3} = \begin{pmatrix}
\psi_{2,3}(a_2, a_3) & \psi_{2,3}(a_2, b_3) \\
\psi_{2,3}(b_2, a_3) & \psi_{2,3}(b_2, b_3)
\end{pmatrix}
\]

\[
\psi_{2,3}(a_2, a_3) = \begin{pmatrix}
0.1 \\
1.0
\end{pmatrix}
\]

\[
\psi_{2,4} = \begin{pmatrix}
\psi_{2,4}(a_2, a_4) & \psi_{2,4}(a_2, b_4) \\
\psi_{2,4}(b_2, a_4) & \psi_{2,4}(b_2, b_4)
\end{pmatrix}
\]

\[
\psi_{2,4}(a_2, a_4) = \begin{pmatrix}
1.0 \\
0.1
\end{pmatrix}
\]

### 3.13.3.1 Calculating all univariate marginals using belief propagation

In the first step of belief propagation, following the protocol of a synchronous parallel update scheme, leaf nodes pass messages inwards along their edges (Figure 3.14). Therefore, we need to determine the following three messages, \( \mu_{1\rightarrow 2}(A_2) \), \( \mu_{3\rightarrow 2}(A_2) \) and \( \mu_{4\rightarrow 2}(A_2) \). For the first message the message update equation is:

\[
\mu_{1\rightarrow 2}(A_2) = \sum_{A_1} \psi_{1,2}(A_1, A_2) \prod_{k \in N(1) \setminus 2} \mu_k(A_1)
\]
However, since node 1 is a leaf node, it has no incoming messages and the update equation reduces to

\[ \mu_{1 \rightarrow 2}(A_2) = \sum_{A_1} \psi_{1,2}(A_1, A_2) \]  

which is a simple marginalization over the states of variable \( A_1 \).

\[ \mu_{1 \rightarrow 2}(A_2) = \sum_{A_1} \begin{pmatrix} 1.0 & 0.9 \\ 0.9 & 1.0 \end{pmatrix} \]  

\[ \mu_{1 \rightarrow 2}(A_2) = \begin{pmatrix} 1.9 \\ 1.9 \end{pmatrix} = k \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} \]

Messages are normalized so that their vector sum adds up to 1 for numerical stability. Eventually all the marginal probabilities have to sum up to 1, so normalizing the messages at each step is good practice.

The message from node 3 to node 2 is

\[ \mu_{3 \rightarrow 2}(A_2) = \sum_{A_3} \psi_{3,2}(A_3, A_2) \prod_{k \in N(3) \setminus 2} \mu_{k \rightarrow 3}(A_3) \]  

\[ \mu_{3 \rightarrow 2}(A_2) = \sum_{A_3} \begin{pmatrix} 0.1 & 1.0 \\ 1.0 & 0.1 \end{pmatrix} \]  

\[ \mu_{3 \rightarrow 2}(A_2) = \begin{pmatrix} 1.1 \\ 1.1 \end{pmatrix} = k \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} \]

The message from node 4 to node 2 is

\[ \mu_{4 \rightarrow 2}(A_2) = \sum_{A_4} \begin{pmatrix} 1.0 & 0 \\ 0.1 & 0 \end{pmatrix} \]
Now that all the leaf node to inner node messages have been completed, we need to calculate the outgoing messages to the leaf nodes. There are three such messages, \( \mu_{2\rightarrow1}(A_1) \), \( \mu_{2\rightarrow3}(A_2) \) and \( \mu_{2\rightarrow4}(A_4) \) (Figure 3.15). The message from node 2 to node 1 uses the Hadamard product of the message from node 4 to node 2 with the message from node 3 to node 2.

\[
\mu_{2\rightarrow1}(A_1) = \sum_{A_1} \psi_{1,2}(A_1, A_2) \prod_{k \in N(2) \setminus 1} \mu_{k\rightarrow2}(A_2) 
\]

(3.98)

\[
\mu_{2\rightarrow1}(A_1) = \sum_{A_1} \psi_{1,2}(A_1, A_2) \mu_{4\rightarrow2}(A_2) \mu_{3\rightarrow2}(A_2) 
\]

(3.99)

\[
\mu_{2\rightarrow1}(A_1) = \begin{bmatrix} 1.0 & 0.9 \\ 0.9 & 1.0 \end{bmatrix} \begin{bmatrix} 1.0 \\ 0.1 \end{bmatrix} \odot \begin{bmatrix} 1 \\ 1 \end{bmatrix} 
\]

(3.100)

\[
\mu_{2\rightarrow1}(A_1) = \begin{bmatrix} 1.0 & 0.9 \\ 0.9 & 1.0 \end{bmatrix} \begin{bmatrix} 1.0 \\ 0.1 \end{bmatrix} 
\]

(3.101)
\[
\mu_{2\rightarrow 1}(A_1) = \begin{pmatrix} 1.09 \\ 1.0 \end{pmatrix} 
\] (3.102)

The message from node 2 to node 3 uses the Hadamard product of the message from node 4 to node 2 with the message from node 1 to node 2.

\[
\mu_{2\rightarrow 3}(A_3) = \sum_{A_2} \psi_{2,3}(A_2, A_3) \prod_{k \in N(2) \setminus 3} \mu_{k\rightarrow 2}(A_2) 
\] (3.103)

\[
\mu_{2\rightarrow 3}(A_3) = \sum_{A_2} \psi_{2,3}(A_2, A_3) \mu_{4\rightarrow 2}(A_2) \mu_{1\rightarrow 2}(A_2) 
\] (3.104)

\[
\mu_{2\rightarrow 3}(A_3) = \begin{pmatrix} 0.1 & 1.0 \\ 1.0 & 0.1 \end{pmatrix} \left[ \begin{pmatrix} 1.0 \\ 0.1 \end{pmatrix} \circ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right] = \begin{pmatrix} 0.2 \\ 1.01 \end{pmatrix} 
\] (3.105)

The message from node 2 to node 4 uses the Hadamard product of the message from node 1 to node 2 with the message from node 3 to node 2.

\[
\mu_{2\rightarrow 4}(A_4) = \sum_{A_2} \psi_{2,4}(A_2, A_4) \prod_{k \in N(2) \setminus 4} \mu_{k\rightarrow 2}(A_2) 
\] (3.106)

\[
\mu_{2\rightarrow 4}(A_4) = \sum_{A_2} \psi_{2,4}(A_2, A_4) \mu_{3\rightarrow 2}(A_2) \mu_{1\rightarrow 2}(A_2) 
\] (3.107)

\[
\mu_{2\rightarrow 4}(A_4) = \begin{pmatrix} 1.0 & 0 \\ 0.1 & 0 \end{pmatrix} \left[ \begin{pmatrix} 1.0 \\ 0.1 \end{pmatrix} \circ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right] = \begin{pmatrix} 0.2 \\ 1.01 \end{pmatrix} 
\] (3.108)

Now that every single message has been computed, we can determine the node marginals from the Hadamard product of incoming messages. The only message arriving into node 1 is the message from node 2, \( \mu_{2\rightarrow 1}(A_1) \). As a result, the marginal at \( A_1 \) is just the normalized message \( \mu_{2\rightarrow 1}(A_1) \).

\[
P_1(A_1) = \mu_{2\rightarrow 1}(A_1) = \begin{pmatrix} 1.09 \\ 1.0 \end{pmatrix} 
\] (3.109)

The marginal at node 2 however, is the Hadamard product of 3 incoming messages.
\[
P_2(A_2) = \mu_{1\rightarrow 2}(A_2)\mu_{4\rightarrow 2}(A_2)\mu_{3\rightarrow 2}(A_2)
\]  
(3.110)

\[
P_2(A_2) = \begin{pmatrix}
1 \\
1 \\
1 \\
0.1 \\
1
\end{pmatrix}
\]  
(3.111)

\[
P_2(A_2) = \begin{pmatrix}
1 \\
0.1
\end{pmatrix}
\]  
(3.112)

The marginal probability at node 3, similar to node 1, is simply the normalized message from node 3 to node 1.

\[
P_3(A_3) = \mu_{2\rightarrow 3}(A_3) = \begin{pmatrix}
0.20 \\
1.01
\end{pmatrix}
\]  
(3.113)

Now we have determined all the marginal probabilities using belief propagation. The power of belief propagation lies in the fact that once the messages converge, we can calculate all the marginal probabilities, not just the ones we require, by reusing the intermediate messages between node. By exploiting the structure of the graph, we are basically marginalizing the joint probabilities across different edges in order to perform the final summation more efficiently. The cost of this computation linear with the number of nodes and quadratic with the number of states per node variable.
Chapter 4

Correspondence between Belief Propagation and the Bethe Approximation

As described in Chapter 3, belief propagation is an algorithm for probabilistic inference on graphs. That is, given a particular graph model and the functions that describe the probabilistic dependencies between the variables of the graphs, belief propagation is can be applied to efficiently compute the marginals for the variables. In this chapter, the correspondence between the Bethe mean field approximation from statistical mechanics and belief propagation is explicitly worked out. This chapter is a regurgitation from my understanding of Dr. Alex Morozov’s notes on the subject. Credit for this work should primarily go to Dr. Morozov.

4.1 Potts and Ising Models

Let \( \{A_1, ..., A_N\} \) be a set of \( N \) random variables. Each random variable, \( A_i \), can take on a discrete set of values which represent the \( q \) possible states or discrete realizations of \( A_i \). In a Markov Random Field, the joint probability mass function of the random variables, \( P(A_1, ..., A_N) \), can be written out as a product of the field and coupling potentials.

\[
P(A_1, ..., A_N) = \frac{1}{Z} \prod_{i<j} \psi_{ij}(A_i, A_j) \prod_i \psi_i(A_i)
\]  

(4.1)
where the indexes $i$ and $j$ label the set of potential functions $\psi_i(A_i)$ and $\psi_{ij}(A_i, A_j)$, which describe the probabilistic dependencies between the states of the random variables and act on some subset of the random variables, $\{A_1, \ldots, A_N\}$. If the potential acts on only one node, it is called a field potential, $\psi_i(A_i)$. If the potential acts on two nodes, it is known as a coupling potential, $\psi_{ij}(A_i, A_j)$. $Z$ is the partition function or the normalization constant. Moreover, there is no causality present in this model. Our goal is to calculate beliefs, or the inferred probabilities, $b_i$ and $b_{ij}$, for all nodes, $i$, and pairs of nodes, $ij$. The Markov Random Field, or MRF, described above is equivalent to the Potts Model recognizable in the statistical physics community.

Define $\psi_{ij}(A_i, A_j) = e^{J_{ij}}, \psi_i(A_i) = e^{h_i}$ assuming $T = 1$. $J_{ij}$ can be defined as the interaction between the variables $i$ and $j$. $h_i$ can be defined as the field acting on $i$. Then using the Boltzmann’s law from statistical mechanics:

$$P(A_1, \ldots, A_N) = \frac{1}{Z} e^{-E(A_1, \ldots, A_N)}$$

where the Potts model energy is

$$E(A_1, \ldots, A_N) = - \sum_{i<j} J_{ij}(A_i, A_j) - \sum_i h_i(A_i)$$

Our pairwise MRF corresponds exactly to a Potts model at a temperature of $T = 1$ and $Z$, the normalization constant corresponds to the partition function. If $q = 2$, the model is known as an Ising model and if $q > 2$, the model is known as the Potts model.

Define $F_{\text{exact}} = U - S$, assuming $T = 1$ where the internal energy or the average energy of the ensemble is

$$U = \sum_{A_1, \ldots, A_N} P(A_1, \ldots, A_N)E(A_1, \ldots, A_N)$$
and the average entropy of the ensemble is

\[ S = - \sum_{A_1, \ldots, A_N} P(A_1, \ldots, A_N) \log P(A_1, \ldots, A_N) \]

Note that \( F_{\text{exact}} = \langle E + \log P \rangle \), where

\[ \langle ... \rangle \equiv \sum_{A_1, \ldots, A_N} P(A_1, \ldots, A_N) \]

Define

\[ \tilde{F}_{\text{exact}} = F_{\text{exact}} + \lambda \left( \sum_{A_1, \ldots, A_N} P(A_1, \ldots, A_N) - 1 \right) \]
\[ \tilde{F}_{\text{exact}} = U - S + \lambda \sum_{A_1, \ldots, A_N} P(A_1, \ldots, A_N) - \lambda \]
\[ \tilde{F}_{\text{exact}} = \sum_{A_1, \ldots, A_N} P(A_1, \ldots, A_N) E(A_1, \ldots, A_N) + \]
\[ + \sum_{A_1, \ldots, A_N} P(A_1, \ldots, A_N) \log P(A_1, \ldots, A_N) + \lambda \sum_{A_1, \ldots, A_N} P(A_1, \ldots, A_N) - \lambda \]

then

\[ \frac{\partial \tilde{F}_{\text{exact}}}{\partial P(A_1, \ldots, A_N)} = 0 \]
\[ \frac{\partial \tilde{F}_{\text{exact}}}{\partial P(A_1, \ldots, A_N)} = E(A_1, \ldots, A_N) + [\log P(A_1, \ldots, A_N) + 1] + \lambda = 0 \]

\[ \log P = -E - \lambda - 1 \]

\[ P = e^{-E} e^{-(\lambda+1)} \]

\[ P(A_1, \ldots, A_N) = e^{-E(A_1, \ldots, A_N)} e^{-(\lambda+1)} \]

Using \( \sum_{A_1, \ldots, A_N} P(A_1, \ldots, A_N) = 1 \), we obtain:

\[ 1 = e^{-E(A_1, \ldots, A_N)} e^{-(\lambda+1)} \]

\[ \frac{1}{Z} = e^{-(\lambda+1)} \]
where $Z = \sum_{A_1,\ldots,A_N} e^{-E(A_1,\ldots,A_N)}$. Thus

$$P(A_1,\ldots,A_N) = \frac{1}{Z} e^{-E(A_1,\ldots,A_N)} \quad (4.2)$$

Note that if Equation 4.2 is true,

$$F_{\text{exact}} = U - S$$

$$F_{\text{exact}} = \sum_{A_1,\ldots,A_N} P(A_1,\ldots,A_N) E(A_1,\ldots,A_N) + \sum_{A_1,\ldots,A_N} P(A_1,\ldots,A_N) \log P(A_1,\ldots,A_N)$$

$$F_{\text{exact}} = \sum_{A_1,\ldots,A_N} \frac{1}{Z} e^{-E(A_1,\ldots,A_N)} E(A_1,\ldots,A_N) + \sum_{A_1,\ldots,A_N} \frac{1}{Z} e^{-E(A_1,\ldots,A_N)} \log \frac{1}{Z} e^{-E(A_1,\ldots,A_N)}$$

$$F_{\text{exact}} = \sum_{A_1,\ldots,A_N} \left[ E(A_1,\ldots,A_N) + \sum_{A_1,\ldots,A_N} \log \frac{1}{Z} e^{-E(A_1,\ldots,A_N)} \right] \frac{1}{Z} e^{-E(A_1,\ldots,A_N)}$$

$$F_{\text{exact}} = \sum_{A_1,\ldots,A_N} \left[ E(A_1,\ldots,A_N) + \log \frac{1}{Z} + \log e^{-E(A_1,\ldots,A_N)} \right] \frac{1}{Z} e^{-E(A_1,\ldots,A_N)}$$

$$F_{\text{exact}} = \sum_{A_1,\ldots,A_N} \left[ E(A_1,\ldots,A_N) - \log Z - E(A_1,\ldots,A_N) \right] \frac{1}{Z} e^{-E(A_1,\ldots,A_N)}$$

$$F_{\text{exact}} = \sum_{A_1,\ldots,A_N} \left[ - \log Z \right] \frac{1}{Z} e^{-E(A_1,\ldots,A_N)}$$

$$F_{\text{exact}} = \left[ - \log Z \right] \frac{1}{Z} Z$$

$$F_{\text{exact}} = - \log Z$$

### 4.2 Mean Field Free Energy

Now consider the mean-field ansatz:

$$P(A_1,\ldots,A_N) = \prod_i b_i(A_i) \quad (4.3)$$
One can show that the corresponding free energy $F_{MF} \geq F_{exact}$, according to the variational approach. Thus we need to minimize $F_{MF}$ (this is true of all $F_{var}$). Now, with Equation 4.3:

$$
\sum_{A_1,\ldots,A_N} \left( \prod_i b_i \right) \left( \sum_j \log b_j \right) = \sum_{A_1,\ldots,A_N} b_1 \ldots b_N (\log b_1 + \ldots + \log b_N)
$$

$$
= \sum_{A_1} b_1 \log b_1 + \sum_{A_2} b_2 \log b_2 + \ldots + \sum_{A_N} b_N \log b_N
$$

$$
= \sum_i \sum_A b_i \log b_i
$$

where $\sum_A b_i(A_i) = 1, \forall i$. Likewise,

$$
\sum_{A_1,\ldots,A_N} \left( \prod_i b_i \right) \left[ -\sum_{i<j} J_{ij}(A_i, A_j) - \sum_i h_i(A_i) \right]
$$

$$
= -\sum_{i<j} \sum_{A_i,A_j} b_i b_j J_{ij}(A_i, A_j) - \sum_i \sum_A h_i(A_i)
$$

Thus

$$
F_{MF} \{ \{ b_i \} \} = -\sum_{i<j} \sum_{A_i,A_j} b_i(A_i) b_j(A_j) J_{ij}(A_i, A_j) - \sum_i \sum_{A_i} b_i(A_i) h_i(A_i) + \sum_i \sum_{A_i} b_i(A_i) \log b_i(A_i)
$$

Then

$$
\tilde{F}_{MF} = F_{MF} + \sum_i \lambda_i \left( \sum_{A_i} b_i(A_i) - 1 \right)
$$

Indeed

$$
\frac{\delta \tilde{F}_{MF}}{\delta \lambda_i} = \sum_{A_i} b_i - 1 = 0, \forall i
$$

recovering the constraints. Now,

$$
\frac{\delta \tilde{F}_{MF}}{\delta b_i(A_i)} = - \sum_{j \in N(i)} b_j J_{ij} - h_i + \lambda_i + [1 + \log b_i] = 0
$$

giving

$$
\log b_i = h_i + \sum_{j \in N(i)} \sum_{x_j} b_j J_{ij} - \lambda_i - 1
$$
Just as before, $\sum A_i b_i(A_i) = 1$ gives $e^{-(\lambda_i+1)} = \frac{1}{Z_i}$, where

$$Z_i = \sum_{A_i} e^{-E_{MF}^{i}(A_i)}$$

and

$$E_{MF}^{i}(A_i) = -h_i(A_i) - \sum_{j \in N(i)} \sum_{x_j} b_j(A_j) J_{ij}(A_i, A_j)$$

Thus

$$b_i(A_i) = \frac{1}{Z_i} e^{-E_{MF}^{i}(A_i)}$$

This is a system of self-consistent equations, to be solved iteratively.

For Ising spins, $A_i = \pm 1, \forall i$. Hence

$$\sum \sum b_i(A_i)b_j(A_j)J_{ij}(A_i, A_j) =$$

Introduce

$$J_{ij}(A_i) = \begin{pmatrix} J_{ij} & -J_{ij} \\ -J_{ij} & J_{ij} \end{pmatrix}$$

$$= \sum_{i<j} [J_{ij}b_i(1)b_j(1) + J_{ij}b_i(-1)b_j(-1) - J_{ij}b_i(1)b_j(-1) - J_{ij}b_i(-1)b_j(1)]$$

$$= \sum_{i<j} J_{ij} [b_i(1) - b_i(-1)] [b_j(1) - b_j(-1)]$$

Introduce $m_i \equiv b_i(1) - b_i(-1) = < A_i >$

$$= \sum_{i<j} J_{ij} m_i m_j$$

Likewise,

$$\sum_{i} \sum_{A_i} h_i(A_i)b_i(A_i) =$$

Introduce

$$h_i(A_i) = \begin{pmatrix} h_i \\ -h_i \end{pmatrix}$$
\[= \sum_i h_i (b_i(1) - b_i(-1)) = \sum_i h_i m_i\]

Finally,
\[
\begin{cases}
  b_i(1) + b_i(-1) = 1 & \Rightarrow \left\{ \begin{array}{l}
b_i(1) = \frac{1 + m_i}{2} \\
b_i(-1) = \frac{1 - m_i}{2}
\end{array} \right., \\
  b_i(1) - b_i(-1) = m_i
\end{cases}
\]

Then
\[
\sum_i \sum_{A_i} b_i(A_i) \log b_i(A_i) = \\
\sum_i \left[ \frac{1 + m_i}{2} \log \left( \frac{1 + m_i}{2} \right) + \frac{1 - m_i}{2} \log \left( \frac{1 - m_i}{2} \right) \right]
\]

Thus
\[
F_{MF} = - \sum_{i<j} J_{ij} m_i m_j - \sum_i h_i m_i + \sum_i \left[ \frac{1 + m_i}{2} \log \left( \frac{1 + m_i}{2} \right) + \frac{1 - m_i}{2} \log \left( \frac{1 - m_i}{2} \right) \right] = 0
\]

Note that we already explicitly used the normalization constraint, so there is no need to use it again. Thus, in the absence of the external field,
\[
\frac{\delta F_{MF}}{\delta m_i} = 0
\]

\[
\Rightarrow - \sum_{j \in N(i)} J_{ij} m_j + \left[ \frac{1}{2} + \frac{1}{2} \log \left( \frac{1 + m_i}{2} \right) - \frac{1}{2} - \frac{1}{2} \log \left( \frac{1 - m_i}{2} \right) \right] = 0
\]

\[
\frac{1}{2} \log \left( \frac{1 + m_i}{1 - m_i} \right) = \sum_{j \in N(i)} J_{ij} m_j
\]

This is equivalent to
\[
m_i = \tanh \left( \sum_{j \in N(i)} J_{ij} m_j \right)
\]

This MF approach is the first two terms in the \(\beta\) expansion of \(F_{exact}\) for the Ising spin glass with zero external field: \((\beta = \frac{1}{T})\)
\[- \beta F_{exact} = - \sum_i \left[ \frac{1 + m_i}{2} \log \left( \frac{1 + m_i}{2} \right) + \frac{1 - m_i}{2} \log \left( \frac{1 - m_i}{2} \right) \right] + \]
\[ + \beta \sum_{i<j} J_{ij} m_i m_j + \frac{\beta^2}{2} \sum_{i<j} J_{ij}^2 \left(1 - m_i^2\right) \left(1 - m_j^2\right) + ..., \]

where the last term is the TAP correction. Further discussion of the MF approach and its extensions is beyond the scope of this work.

### 4.3 Bethe Mean Field Free Energy

Consider

\[ P(A_1, ..., A_N) = \prod_{i<j} b_{ij}(A_i, A_j) \prod_i b_i(A_i)^{(1-q_i)} \tag{4.4} \]

where \( q_i \) is the number of nodes in the neighborhood of \( i, N_i \). Indeed, this is equivalent to the product of conditional probabilities on a tree. For example, for

\[ P(A_1, A_2, A_3, A_4) = b_{12} b_{23} b_{24} \]

\[ = b_{12} b_{23} b_{24} \]

since \( b_{12} = b_1(A_1)^{b_{21}^{cond}(A_2|A_1)} \), \( b_{23} = b_{32}^{cond}(A_3|A_2) \) and \( b_{24} = b_{42}^{cond}(A_4|A_2) \).

Therefore

\[ P(A_1, A_2, A_3, A_4) = b_1(A_1) b_{21}^{cond}(A_2|A_1) b_{32}^{cond}(A_3|A_2) b_{42}^{cond}(A_4|A_2) \]

which is exact for this tree. Not that this prescription automatically chooses the root node to be 1. Thus the Bethe ansatz (unlike MF) is guaranteed to be exact on trees. In any case, with this decomposition we obtain:

\[ \sum_{A_1, ..., A_N} P(A_1, ..., A_N) \log P(A_1, ..., A_N) = \sum_{A_1, ..., A_N} \prod_{i<j} b_{ij} \prod_i b_i^{1-q_i} \left[ \sum_{i<j} \log b_{ij} + \sum_i (1-q_i) \log b_i \right] \]
For the tree above, we have

\[
\sum_{A_1,A_2,A_3,A_4} b_{12} b_{23} b_{24} \left[ \log b_{12} + \log b_{23} + \log b_{24} + (1-q_1) \log b_1 \\
+ (1-q_2) \log b_2 + (1-q_3) \log b_3 + (1-q_4) \log b_4 \right] = \]

Note that \( q_1 = q_3 = q_4 = 1 \), but we keep the corresponding one-body terms for generality. Using

\[
\begin{align*}
\sum_{A_3,A_4} \frac{b_{23} b_{24}}{b_2^2} &= 1, \\
\sum_{A_1,A_4} \frac{b_{12} b_{24}}{b_2^2} &= 1, \\
\sum_{A_1,A_3} \frac{b_{12} b_{23}}{b_2^2} &= 1,
\end{align*}
\]

we obtain:

\[
\sum_{i<j} \sum_{A_i,A_j} b_{ij} \log b_{ij}
\]

for the two-body term. Using

\[
\sum_{A_2,A_3,A_4} \frac{b_{12} b_{23} b_{24}}{b_2^2} = b_1
\]

, etc., we obtain:

\[
\sum_i (1-q_i) \sum_{A_i} b_i \log b_i
\]

for the one-body term. Thus

\[
\sum_{A_1,...,A_N} P(A_1,...,A_N) \log P(A_1,...,A_N) = \sum_i \sum_{A_i} \sum_{i<j} b_{ij} \log b_{ij} + \sum_i (1-q_i) \sum_{A_i} b_i \log b_i
\]

This expression is exact on a tree.

Finally

\[
\sum_{A_1,...,A_N} P(A_1,...,A_N) \left[ - \sum_{i<j} J_{ij}(A_i,A_j) - \sum_i h_i(A_i) \right] = - \sum_{i<j} \sum_{A_i,A_j} b_{ij} J_{ij} - \sum_i \sum_{A_i} b_i h_i
\]
This is true for any $E$ which only has one- and two-body terms. Thus

$$U = -\sum_{i<j} \sum_{A_i,A_j} b_{ij} J_{ij} - \sum_{i} \sum_{A_i} b_i h_i$$

and

$$F_\beta = -\sum_{i<j} \sum_{A_i,A_j} b_{ij} J_{ij} - \sum_{i} \sum_{A_i} b_i h_i + \sum_{i} \sum_{A_i,A_j} b_{ij} \log b_{ij} + \sum_{i} (1 - q_i) \sum_{A_i} b_i \log b_i$$

$U$ can be rewritten in a slightly different form:

$$U = \sum_{i<j} \sum_{A_i,A_j} b_{ij} [J_{ij} + h_i + h_j] + \sum_{i} (q_i - 1) \sum_{A_i} b_i h_i$$

Indeed,

$$-\sum_{i<j} \sum_{A_i,A_j} b_{ij} [h_i + h_j] = -\sum_{i<j} \left[ \sum_{A_i} b_i h_i + \sum_{A_j} b_j h_j \right] = -\sum_{i} q_i \sum_{A_i} b_i h_i$$

Indeed, for our expression earlier,

$$-\sum_{i<j} \left[ \sum_{A_i} b_i h_i + \sum_{A_j} b_j h_j \right] =$$

$$= - \left[ \sum_{A_1} b_1 h_1 + \sum_{A_2} b_2 h_2 + \sum_{A_3} b_3 h_2 + \sum_{A_2} b_2 h_3 + \sum_{A_4} b_4 h_4 \right] =$$

$$= - \left[ \sum_{A_1} b_1 h_1 + 3 \sum_{A_2} b_2 h_2 + \sum_{A_3} b_3 h_2 + \sum_{A_4} b_4 h_4 \right]$$

which is consistent with $q_i$'s for this graph.

So, an alternative form of $F_\beta$ is:

$$F_\beta = -\sum_{i<j} \sum_{A_i,A_j} b_{ij} [J_{ij} + h_i(A_i) + h_j(A_j)] + \sum_{i} (q_i - 1) \sum_{A_i} b_i(A_i) h_i(A_i) +$$

$$+ \sum_{i<j} \sum_{A_i,A_j} b_{ij}(A_i, A_j) \log b_{ij}(A_i, A_j) + \sum_{i} (1 - q_i) \sum_{A_i} b_i(A_i) \log b_i(A_i)$$
Now,

\[ \tilde{F}_\beta = F_\beta + \sum_i \gamma_i \left[ 1 - \sum_{A_i} b_i(A_i) \right] + \sum_{i,j} \gamma_{ij} \left[ 1 - \sum_{A_i, A_j} b_{ij}(A_i, A_j) \right] + \]

\[ \sum_{i<j} \lambda_{ij}(A_j) \left[ b_j(A_j) - \sum_{A_i} b_{ij}(A_i, A_j) \right] + \sum_{i<j} \sum_{A_i} \lambda_{ji}(A_i) \left[ b_i(A_i) - \sum_{A_j} b_{ij}(A_i, A_j) \right] \]

And the following constraints

\[
\left\{ \begin{align*}
\sum_{A_i} b_i(A_i) &= 1, \\
\sum_{A_i, A_j} b_{ij}(A_i, A_j) &= 1, \\
\sum_{A_j} b_{ij}(A_i, A_j) &= b_i(A_i), \\
\sum_{A_i} b_{ij}(A_i, A_j) &= b_j(A_j),
\end{align*} \right.
\]

are enforced through Lagrange multipliers.

\[
\frac{\delta \tilde{F}_\beta}{\delta b_{ij}} = -\tilde{J}_{ij} + [1 + \log b_{ij}] - \gamma_{ij} - \lambda_{ij}(A_j) - \lambda_{ji}(A_i) = 0
\]

or

\[ \log b_{ij} = \tilde{J}_{ij} + \gamma_{ij} + \lambda_{ij}(A_j) + \lambda_{ji}(A_i) - 1 \]

\[ b_{ij} = e^{\tilde{J}_{ij} + \gamma_{ij}} e^{\lambda_{ij}(A_j) + \lambda_{ji}(A_i)} \]

\[ \sum_{A_i, A_j} b_{ij} = 1 \Rightarrow e^{\gamma_{ij} - 1} \sum_{A_i, A_j} e^{\tilde{J}_{ij} + \lambda_{ij}(A_j) + \lambda_{ji}(A_i)} = 1 \]

\[ Z_{ij} = \sum_{A_i, A_j} e^{\tilde{J}_{ij} + \lambda_{ij}(A_j) + \lambda_{ji}(A_i)} \]

\[ b_{ij} = \frac{1}{Z_{ij}} e^{\tilde{J}_{ij} + \lambda_{ij}(A_j) + \lambda_{ji}(A_i)} \]

Furthermore,

\[ \frac{\delta \tilde{F}_\beta}{\delta b_i} = (q_i - 1) h_i - (q_i - 1) [1 + \log b_i] - \gamma_i + \sum_{j \in N(i)} \lambda_{ji}(A_i) = 0 \]
or

\[
\log b_i = h_i + \frac{1}{(q_i - 1)} \sum_{j \in N(i)} \lambda_{ji}(A_i) - \frac{\gamma_i}{q_i - 1} - 1
\]

\[
\sum_{A_i} b_i = 1 \Rightarrow \sum_{A_i} e^{-\left(\frac{\gamma_i}{q_i - 1}+1\right)} \sum_{A_i} e^{h_i + \frac{1}{q_i - 1} \sum_{j \in N(i)} \lambda_{ji}(A_i)} = 1
\]

\[
Z_i = \sum_{A_i} e^{h_i + \frac{1}{q_i - 1} \sum_{j \in N(i)} \lambda_{ji}(A_i)}
\]

\[
b_i = \frac{1}{Z_i} e^{h_i + \frac{1}{q_i - 1} \sum_{j \in N(i)} \lambda_{ji}(A_i)}
\]

Finally, \( \sum_{A_i} b_{ij}(A_i, A_j) = b_j(A_j) \) gives:

\[
\frac{1}{Z_{ij}} \sum_{A_i} e^{\tilde{J}_{ij} + \lambda_{ij}(A_j) + \lambda_{ji}(A_i)} = \frac{1}{Z_j} e^{h_j + \frac{1}{q_j - 1} \sum_{k \in N(j)} \lambda_{kj}(A_j)}
\]

These are self-consistent equations for the \( \lambda \)'s. Assume

\[
e^{\lambda_{ij}(A_j)} = \prod_{k \in N(j)/i} m_{kj}(A_j)
\]

which gives

\[
e^{\lambda_{ji}(A_i)} = \prod_{k \in N(i)/j} m_{ki}(A_i)
\]

Then

\[
b_{ij} = \frac{1}{Z_{ij}} e^{\tilde{J}_{ij}} \prod_{k \in N(i)/j} m_{ki}(A_i) \prod_{l \in N(j)/i} m_{lj}(A_j)
\]

\[
b_i = \frac{1}{Z_i} e^{h_i} \prod_{j \in N(i)} \left( \prod_{k \in N(i)/j} m_{ki}(A_i) \right) \frac{1}{q_i - 1} = \frac{1}{Z_i} e^{h_i} \prod_{k \in N(i)} m_{ki}(A_i)
\]

Indeed, consider example

Then

\[
[m_{24}m_{34}] [m_{14}m_{34}] [m_{14}m_{24}] = (m_{14}m_{24}m_{34})^{q_i-1}
\]
which is consistent with the expression above. Finally

\[
\frac{1}{Z_j} e^{h_j} \prod_{k \in N(j)} m_{kj}(A_j) = \frac{1}{Z_{ij}} e^{h_j} e^{\lambda_{ij}(A_j)} \sum_{A_i} e^{J_{ij}+h_i} \prod_{k \in N(i)/j} m_{ki}(A_i)
\]
or

\[
\frac{\prod_{k \in N(j)} m_{kj}(A_j)}{\prod_{k \in N(j)/i} m_{kj}(A_j)} = m_{ij}(A_j) = \frac{Z_j}{Z_{ij}} \sum_{A_i} e^{J_{ij}+h_i} \prod_{k \in N(i)/j} m_{ki}(A_i)
\]

where

\[
e^{\lambda_{ij}(A_j)} = \prod_{k \in N(j)/i} m_{kj}(A_j)
\]

Thus we obtain the belief propagation equations:

\[
m_{ij}(A_j) = \frac{Z_j}{Z_{ij}} \sum_{A_i} e^{J_{ij}+h_i} \prod_{k \in N(i)/j} m_{ki}(A_i)
\]

\[
b_i = \frac{1}{Z_i} e^{h_i} \prod_{j \in N(i)} m_{ji}(A_i)
\]

\[
b_{ij} = \frac{1}{Z_{ij}} e^{J_{ij}+h_i+h_j} \prod_{k \in N(i)/j} m_{ki}(A_i) \prod_{l \in N(j)/i} m_{lj}(A_j)
\]

\[m_{ki}(A_i)\] can be interpreted as the probability of \(A_i\) if all the links except to \(k\) have been removed, and \(h_i\) has been turned off. Indeed, \(b_i = \frac{1}{Z_i} m_{ki}(x_i)\) in this case.

Introducing \(\tilde{m}_{ji}(A_i) = \log m_{ji}(A_i)\), we obtain:

\[
b_i = \frac{1}{Z_i} e^{h_i+\sum_{j \in N(i)} \tilde{m}_{ji}(A_i)}
\]

\[
b_{ij} = \frac{1}{Z_{ij}} e^{J_{ij}+h_i+h_j+\sum_{k \in N(i)/j} \tilde{m}_{ki}(A_i)+\sum_{l \in N(j)/i} \tilde{m}_{lj}(A_j)}
\]

We can think of \(\tilde{m}\)'s as auxiliary fields which capture the influence of neighboring nodes (“external environment”). Not also that \(b_{ij}\) is “almost” given by \(b_ib_j\), except that \(\tilde{m}_{ji}(A_i) + \tilde{m}_{ij}(A_j)\) is replaced by the “true” coupling, \(J_{ij}(A_i, A_j)\).
\(\lambda\)'s and \(m\)'s are interchangeable:

\[ e^{\lambda_{ij}(A_j)} = \prod_{k \in N(j)/i} m_{kj}(A_j) \tag{4.5} \]

and

\[ b_j(A_j) = \frac{1}{Z_j} e^{h_j} \prod_{k \in N(j)/i} m_{kj}(A_j) = \frac{1}{Z_j} e^{h_j} m_{ij}(A_j) \]

Thus

\[ m_{ij}(A_j) = Z_j e^{-h_j} e^{-\lambda_{ij}(A_j)} \frac{1}{Z_j} e^{h_j} e^{\frac{1}{q_j-1} \sum_{k \in N(j)} \lambda_{kj}(A_j)} \]

or

\[ m_{ij}(A_j) = e^{\frac{1}{q_j-1} \sum_{k \in N(j)} \lambda_{kj}(A_j) - \lambda_{ij}(A_j)} \]

\(e^{\lambda_{ij}(A_j)}\) can be interpreted as the probability of \(A_j\) in a system where a link to \(i\) has been removed as can be seen in Equation 4.5. Also, \(h_j\) is set to \(\emptyset\). Indeed, in such a system:

\[ b_j = \frac{1}{Z_j} \prod_{k \in N(j)} m_{kj}(A_j) = \frac{1}{Z_j} e^{\lambda_{ij}(A_j)} \]

Thus \(\lambda\)'s and \(m\)'s are complementary, as can be seen from this example:

\[ e^{\frac{1}{2}[-\lambda_{ij} + \lambda_{kj} + \lambda_{lj}]} = e^{\frac{1}{2}[\lambda_{ij} + \lambda_{kj} + \lambda_{lj}]} = e^{\frac{1}{2}[\lambda_{ij} + \lambda_{kj} + \lambda_{lj}]} = \left[ \frac{m_{mj}m_{mj}m_{ij}}{m_{kj}m_{ij}} \right]^{\frac{1}{2}} = m_{ij} \]

as expected.

Interestingly, if the link is removed,

\[ b_j = \frac{1}{Z_j} \prod_{k \in N(j)} m_{kj}(A_j) = \frac{1}{Z_j} e^{\lambda_{ij}(A_j)} \]

on one hand, and

\[ b_j = \frac{1}{Z_j} e^{\frac{1}{q_j-1} \sum_{k \in N(j)/i} \lambda_{kj}(A_j)} \]
on the other, where $q_i$ is the number of links without $i$. For example, in the graph below, $q_i = 2$ and

$$\frac{1}{2-1}(\lambda_{kj} + \lambda_{lj}) = \bar{m}_{kj} + \bar{m}_{lj}$$

is indeed equivalent to $\lambda_{ij} = \bar{m}_{kj} + \bar{m}_{lj}$.

Finally, Weigt et al. define messages by:

$$m_{ki}(A_i) = \sum_{A_k} e^{J_{ki}(A_k, A_i)} P_{k\to i}(A_k)$$

which gives

$$b_i = \frac{1}{Z_i} e^{h_i} \prod_{k \in N(i)} \left[ \sum_{A_k} e^{J_{ki}(A_k, A_i)} P_{k\to i}(A_k) \right]$$

$$m_{ij}(A_j) = \frac{Z_i}{Z_{ij}} \sum_{A_i} e^{J_{ij}(A_i, A_j) + h_i} \prod_{k \in N(i)/j} m_{ki}(A_i)$$

which gives

$$P_{i\to j}(A_i) = \frac{Z_i}{Z_{ij}} e^{h_i + \lambda_{ij}(A_i)}$$

Thus

$$P_{i\to j}(A_i) = \frac{Z_i}{Z_{ij}} e^{h_i + \lambda_{ij}(A_i)}$$

On the other hand,

$$P_{j\to i}(A_j) = \frac{Z_j}{Z_{ij}} e^{h_j + \lambda_{ji}(A_j)}$$

With these definitions,

$$b_{ij} = \frac{1}{Z_{ij}} e^{J_{ij} + h_i + h_j} \prod_{k \in N(i)/j} \left[ \sum_{A_k} e^{J_{ki}(A_k)} P_{k\to i}(A_k) \right] \prod_{l \in N(j)/i} \left[ \sum_{A_l} e^{h_l} P_{l\to i}(A_l) \right]$$

$$b_{ij} = \frac{1}{Z_{ij}} e^{J_{ij}} P_{i\to j}(A_i) P_{j\to i}(A_j)$$

Remarkably, this is called $P_{ij}^{dir}(A_i, A_j)$ in Weigt et al. Here we see that it is
the same as $b_{ij}$ under the Bethe approximation. Since $P_{i \rightarrow j}(A_i) \sim e^{h_i + \lambda_{ji}(A_i)}$, it gives the probability of $A_i$ in a system where a link to $j$ has been removed. Indeed, in such a system,

$$b_i = \frac{1}{Z_i} \prod_{k \in N(i)/j} m_{ki}(A_i)e^{h_i}$$

$$\sim e^{h_i + \lambda_{ji}(A_i)}$$

Unlike the formulation with $\lambda$’s, $h_i \neq 0$ is allowed, which makes these messages more interpretable in terms of marginal probabilities.
Chapter 5

Modeling Electrostatic Mutations in HIV Protease
with a Potts Model by applying Probabilistic
Inference Methods

5.1 Introduction

Proteins evolve through random mutagenesis and their evolutionary selection is constrained by structural, functional and environmental limitations[16, 121]. Thermodynamic stability is by far the most important structural factor, as most proteins need to be folded in order to function. The stability range of proteins is narrow, and is estimated experimentally to be around 10 kcal/mol, which is of the order of three hydrogen bonds. As a result of this marginal stability, proteins operate “on a knife’s edge”[79], whereby a single highly deleterious mutation could potentially lead to an unfolded protein[27]. By the same token, a single stabilizing mutation could be advantageous from an evolutionary point of view. Arnold and co-workers have found that more stable forms of cytochrome P450 allowed for greater exploration of mutational space in directed evolution experiments than sequences without stabilizing mutations[12]. It is now widely recognized that thermodynamic stability is intimately linked with the evolvability of a protein[12, 30, 111].

Another constraint on protein evolution is sign epistasis, which places
limits on mutational pathways. This occurs when a new mutation is improbable, or even impossible, under one genetic background, but possible under another genetic background[120]. Sign epistasis could arise from protein stability: a mutation deleterious to stability can lead to a viable or non-viable protein depending on the overall stability of the protein in which the mutation is occurring[27, 119]. This results in pathway restriction and reduces the complexity of evolutionary sequence space.

Even though the process of mutagenesis is random, the genetic and structural constraints mentioned above coupled with functional selection ensure that certain mutations in evolving proteins are associated with each other in a highly non-random fashion[23]. These correlated mutations are an inherent property of evolving amino acid sequences and are a signature of viable, fit proteins. A multitude of methods have been developed to identify such pairs and groups of mutations[37, 70], some of which have been applied to HIV protease sequences to locate pairs [19, 93, 49] or groups[122, 62] of coevolving residues. Our own previous work on higher-order correlations showed that for HIV-1 protease, including pair correlations are essential to reproduce the patterns observed in multiple sequence alignments of protease sequences of HIV from patients undergoing anti-retroviral therapy[44].

It is very tempting to attribute sequence correlations to the effects of epistasis arising from protein stability. Theoretical work has shown that typical protein folding energetics is consistent with biological observations related to genome size and mutation rate[127], and we believe that it should be possible to connect sequence correlations with protein biophysics on a more detailed molecular level. For example, previous work by Ranganathan, et al.[104] has attempted to explain mutational coevolution by connecting statistical free energies from multiple sequence alignments to differences in experimental free energies from double mutant cycle experiments. Unfortunately, some of
these results have been difficult to replicate[21], and are still a topic of active debate in the community[26, 33, 63]. Others have attempted to correlate coevolution with protein interaction data[81, 14, 117]. However, interactions involved in protein binding domains tend to be local, and clearly not all coevolving mutations are close in space. Therefore, it would be difficult to explain the biophysics of allostery and long-range interactions using this strategy. Thus, while studies of these kinds have made great progress, a consensus linking protein biophysics and evolutionary mutational correlations has not yet emerged. This has motivated us to focus on how correlated electrostatic mutations in HIV protease could be connected with the biophysics of long-range interactions via their impact on protein stability.

Methods for predicting the free energy of folding and its change upon mutation based upon structural and physical considerations are an area of active research. However, most current methods are not very accurate at calculating free energies differences upon point mutation[51, 88], and are for the most part unable to deal with multiple mutations. Therefore, we have chosen to focus instead on the electrostatic part of the total energy, for which we have much more confidence. We apply a coarse-grained model of electrostatics, by placing whole charges on the terminal carbons of amino acids, and score the electrostatic energy using the high resolution Analytical Generalized Born (AGB) scoring function[35].

Focusing on the electrostatic part of the folding energy captures many of the effects of sequence mutation on the stability of HIV protease. Residues which change their charge state upon mutation are particularly common in protease, and it has been long believed that such electrostatic mutations can influence enzymatic activity through long-range interactions[115]. Furthermore, it has been shown experimentally that mutations distal to the binding site seem to a large effect on resistance to drugs[73]. It is therefore reasonable to assume
that long-range interactions of non-random evolutionary selected charged amino acids at very specific sites also influence protease stability and contribute to drug resistance.

In this paper, we present results that suggest that a substantial part of the sequence correlations observed in HIV-1 protease can be attributed to effects on the folding free energy of the protein. We show that the average electrostatic stabilization of HIV protease increases with the number of electrostatic mutations, consistent with the hypothesis that electrostatic mutations provide a means by which destabilizing effects of drug resistance mutations can be mitigated. Furthermore, we show that the particular patterns of correlation observed in the multiple sequence alignment are consistent with electrostatic stabilization and destabilization (i.e. like charges close in space are depleted, while opposite charges are enhanced).

5.2 Methods

5.2.1 HIV sequence databases

Aligned HIV-1 DNA nucleotide sequences were downloaded from Christopher Lee’s HIV Positive Selection Mutation Database\textsuperscript{1} on March 4th, 2008. The 45,161 sequences consist of primarily HIV-1 subtype B samples (Calvin Pan, personal communication) and were donated to C. Lee by Specialty Laboratories, who sequenced them between 1999 and mid-2002. Each sequence is 1443 nucleotide bases long, the first 297 bases of which correspond to protease (99 transcribing codons). The remaining 1146 nucleotides correspond to the first 382 amino acids of the p51 reverse transcriptase and were not used in this study.

\textsuperscript{1}http://bioinfo.mbi.ucla.edu/HIV
The codons (other than those containing non-standard bases or sequencing ambiguities) corresponding to the protease portion of the pol gene were extracted and converted to their corresponding amino acids and then into strings of the characters “0”, “-” and “+”, indicating whether the given residue is neutral, negatively or positively charged at pH 6 (i.e. His, Arg, and Lys are positively charged, while Asp and Glu are negatively charged). The resulting “charge signatures” were also compared to the HIV-1 subtype B consensus signature obtained from the Los Alamos National Laboratory HIV sequence database\(^2\), which was used to define the “wild-type”. We define an electrostatic mutation to be a mutation which changes the charge of the amino acid relative to the wild-type sequence. For example, a Ser to Cys mutation involving polar, uncharged amino acids is not an electrostatic mutation, and neither is an Arg to Lys mutation, even though both amino acids are charged.

Examination of the charge signatures reveals that 55 of the 99 amino acid positions in HIV protease are observed to have at least one electrostatic mutation. However, some positions mutate more than others, and different residue pairs show greater or lesser degrees of correlation. In order to facilitate the log-linear analysis (below), a smaller, more manageable subset of the 55 positions was chosen. To do this, we performed a pair correlation analysis on all of the electrostatically active positions. We converted the database of charge signatures to the 2 letter alphabet \(W\) and \(M\), representing wild-type and mutant charge states, respectively. We then calculated product-moment correlation coefficients\(^4\)

\[
\phi_{ij} = \frac{p_{i,j} - p_i p_j}{\sqrt{p_i (1 - p_i) p_j (1 - p_j)}} \tag{5.1}
\]

for each residue pair (where \(p_i\) is the mutation probability at position \(i\), and \(p_{i,j}\) is the probability of the double mutant for positions \(i\) and \(j\)). This allowed us to

\(^2\)http://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html
identify 18 positions (Q7, T12, G16, Q18, K20, D30, E34, E35, N37, K43, Q58, Q61, L63, H69, K70, I72, N88, and Q92) that have high electrostatic mutation frequencies and/or are highly coupled with each other at the pair level.

In order to explore correlation patterns involving multiple residues, we also constructed a subset of electrostatic signatures described above consisting of only those signatures with 4, 5 or 6 electrostatic mutations. A similar analysis of pair correlation was performed on this reduced set, and chose a different set of 18 positions that have high electrostatic mutation frequency and/or correlation (G16, G17, Q18, K20, D30, E34, E35, N37, K43, Q58, K45, Q58, D60, L63, H69, K70, T74, N88, and Q92).

5.2.2 Calculation of electrostatic folding free energies using a Generalized Born (GB) model

The electrostatic energy of protein folding $\Delta G_e$ was estimated as

$$\Delta G_e = G_e^{(f)} - G_e^{(u)},$$

(5.2)

where $G_e^{(f)}$ and $G_e^{(u)}$ are electrostatic free energies of the folded and unfolded states, respectively. The folded state electrostatic free energy was calculated by placing charges corresponding to a particular charge signature onto the most-distal sidechain carbon atom of the corresponding amino acid for a wild-type dimer crystal structure (PDB ID 1NH0). All of the other sidechain atoms were neutral, and we placed a partial charge dipole of $\pm 0.4e$ on every backbone amide and carbonyl group. The electrostatic free energy of this system was calculated using AGB, an implementation of the pairwise descreening Generalized Born model that makes use of a parameter-free algorithm to take into account atomic overlaps[35]. The calculation for the unfolded state was similar, except that a
maximally extended structural representation of one monomer chain was used: backbone dihedral angles were set to 180° (except for proline) and sidechain rotamer states were chosen all-trans. To account for the fact that only one chain of the dimer is used in the unfolded state calculation, the resulting AGB free energy was doubled to obtain \( G_{\text{e}}^{(u)} \). This procedure was then repeated for every electrostatic signature corresponding to observed and synthetic charge signatures.

5.2.3 Modeling Electrostatic Correlations with a Potts model

For sequences of amino acids, \( A_1, \cdots, A_N \), of length \( N \) with \( q \) degrees of freedom at each site (types of amino acids for example), there exist \( q^N \) unique sequences, a number that grows exponentially with the sequence length. The global joint probability distribution of this system, \( P \), can be written as a sum over the probability of every unique sequence

\[
P = \sum_{n=1}^{q^N} P_n(A_1, \cdots, A_N)
\]  

(5.3)

Modeling such a system and getting estimates of the probabilities of each state at every position is computationally difficult. One solution to modeling such systems is to apply the principle of maximum entropy, which assumes a model that is maximally random but is still constrained enough to explain certain average features of the training data[39]. For instance, we can preserve the univariate marginals of amino acids. We call this model the independent model as the frequencies of amino acids at site \( i \) are independent of the frequencies of mutations at site \( j \). The constraints on the independent model are

\[
P_i(A_i) \equiv F_i(A_i) = \sum_{A_i} P(A_1, \cdots, A_N)
\]  

(5.4)
where $P_i(A_i)$ is the estimated independent model probability of amino acid $A$ at site $i$ and $F_i(A_i)$ is the observed probability of amino acid $A$ at site $i$. We can also preserve the univariate and bivariate marginals. We call this model the pair model as the joint probabilities of all pairs of mutations are preserved. The constraints on the pair model are

\begin{align}
  P_i(A_i) & \equiv F_i(A_i) = \sum_{A_1} P(A_1, \cdots, A_N) \\
  P_{ij}(A_i, A_j) & \equiv F_{ij}(A_i, A_j) = \sum_{A_i, A_j} P(A_1, \cdots, A_N)
\end{align}

(5.5)

where $P_{ij}(A_i, A_j)$ is the estimated pair model probability for the joint probability of a pair of amino acids, $A_i$ and $A_j$ at position $i$ and $j$ and $F_{ij}(A_i, A_j)$ is the observed joint probability for a pair of amino acids, $A_i$ and $A_j$. The entropy $S$, from a probabilistic perspective, can be written in terms of the global probability distribution.

\[
S = - \sum \mathbf{P} \ln \mathbf{P}
\]

(5.6)

where $\mathbf{P}$ is the global probability distribution from Equation 5.3. Maximizing Equation 5.6 and introducing Lagrange multipliers with the constraints posed in Equation 5.4 and Equation 5.5 leads to the following Boltzmann distribution for the independent model

\[
P_1(A_1, \ldots, A_i, \ldots, A_N) = \frac{1}{Z} \exp \left[ \sum_i \lambda_i(A_i) \right]
\]

(5.7)

and the following Boltzmann distribution for pair model

\[
P_2(A_1, \ldots, A_i, \ldots, A_N) = \frac{1}{Z} \exp \left[ \sum_i \lambda_i(A_i) - \sum_{i<j} \lambda_{ij}(A_i, A_j) \right]
\]

(5.8)

where $\lambda_i(A_i)$ is the unknown Lagrange multiplier for site $i$ with amino acid $A_i$, $\lambda_{ij}(A_i, A_j)$ is the unknown Lagrange multiplier for a pair of sites $i$ and $j$ with amino acids $A_i$ and $A_j$ associated with the constraints of the pair model and $Z$ is the partition function. The independent model is constrained by only the
univariate marginal, for which $\lambda_i(A_i)$ is a fitting parameter that represents the field of amino acid $A_i$ at position $i$. Similarly, adding additional constraints to the independent model by preserving the bivariate marginals alongside the univariate marginals resulting in additional fitting parameters, $\lambda_{ij}(A_i, A_j)$, which represents the coupling between amino acids $A_i$ and $A_j$ at positions $i$ and $j$. In statistical mechanics, this pair model is referred to as the $Q$-state disordered Potts model[52, 98, 118]. In previous work we used a two-letter Potts model, commonly known as the Ising model, for modeling wild-type and mutated residues in HIV protease sequences[44]. For that previous study, we used iterative proportional fitting (IPF), an algorithm applied to hierarchical log-linear models to fit observables and interactions such as such univariate and bivariate marginals[28].

Though IPF fits the desired marginals, it converges slowly and it does not explicitly calculate or output the fitting parameters. These parameters are important, because in a subsequent study, we used a 3-letter Potts model to study the energetic role of correlated electrostatic mutations in HIV protease, for which we wished to separate the direct coupling between pairs of positions from the indirect effect of the network. In order to do so, we needed to determine the sign and magnitude of the field and coupling fitting parameters that preserve the observed univariate and bivariate marginals. This problem of inferring the fitting parameters from the probability distribution is known in the literature as the ‘inverse Ising problem’ and several approaches have been proposed to solve this problem[54].

In contrast, the approach we took used efficient inference of probabilities over a graphical representation of the data to iteratively determine the fitting parameters; a method that was first prescribed by Weigt et al. 2009[118, 66]. Graphical models are often used in computer science to describe joint probability distributions. They allow the user to reason with the data, even in highly
complex graphs with many variables and states. Inference techniques can then be applied to determine or estimate (depending on the graph structure) the joint distributions efficiently. After we model the data as a graph, we then used a mean field approach to determine the direct contribution of the total correlation between a pair of mutations. Further discussion of iterative proportional fitting, the inverse Ising problem, and the graph-theoretic solution to the inverse Ising problem, follows in subsequent chapters.

5.2.4 Posing the inverse problem for the Potts model for sequence probabilities

Our Potts model for correlated electrostatic mutations deals with a reduced sequence length of 18 (from 99 amino acids) and a reduced amino acid alphabet size of 3. For length $N = 18$ and a $q = 3$ letter alphabet ($0$, $+$, $-$), there are $q^N = 3^{18} = 387,420,489$ possible sequences or unique charge signatures. For every signature, we wish to calculate the probability of that sequence under two models; an independent model which preserves the database derived univariate marginals and a mean field model (Bethe approximation) which preserves both the database derived univariate and bivariate marginals. We call this the pair correlation model and we refer to the probabilities of a sequence under the independent and pair correlation model as $P_1$ and $P_2$ respectively. $P_1$ probabilities of individual sequences can be obtained by simply taking the product of the observed univariate marginals at each position.

$$P_1(A_1, ..., A_N) = \prod_{i=1}^{N} P_{i}^{obs}(A_i)$$

(5.9)

where $A_i = \{0, +, -\}$ is one of the three possible charges at position $i$ and $P_{i}^{obs}(A_i)$ is the observed univariate marginal of charge $A_i$ at position $i$. All observed univariate and bivariate frequencies are derived from the Lee database[20].
The probability of a sequence under the pair correlation model $P_2$, on the other hand, is determined by a model which generates sequences that preserve both the observed univariate and bivariate marginals. In order to do so, we fit the sequence signature probabilities to a 3-state Potts model where the Hamiltonian, $\mathcal{H}$ is described by field and coupling parameters, which reflect the mutation frequency at a site and the strength of the statistical coupling between two sites respectively.

$$P_2(A_1, ..., A_N) = \frac{1}{Z} \exp\{\mathcal{H}(A_1, ..., A_N)\} \quad (5.10)$$

where $P_2(A_1, ..., A_N)$ is the probability of a sequence of length $N$ consisting of charges $A_i$ at position $i$ which preserves the univariate and bivariate marginals. The Hamiltonian can be defined as

$$\mathcal{H}(A_1, ..., A_N) = \sum_i \lambda_i(A_i) - \sum_{i<j} \lambda_{ij}(A_i, A_j) \quad (5.11)$$

where $\lambda_i(A_i)$ is the field at position $i$ for charge $A_i$, $\lambda_{ij}(A_i, A_j)$ is the coupling between charges $A_i$ and $A_j$ at positions $i$ and $j$ and $Z$ is the partition function.

$$Z = \sum_{A_i} \exp\{\mathcal{H}(A_1, ..., A_N)\} \quad (5.12)$$

This model is the maximum entropy solution to the probability distribution that matches the single-point and double-point correlations [39]. Note that if there were two possible states at each site instead of three, this Potts model would be equivalent to the famous Ising model, which is widely applied in the study of spin glass systems in statistical physics.

For a system of 18 positions and 3 states, $18 \times 3 = 54 \lambda_i(A_i)$ and $\binom{18}{2} \times 3^2 = 1377 \lambda_{ij}(A_i, A_j)$ parameters need to be determined to accurately describe the Hamiltonian which preserves the observed marginals. But the parameters are not independent as we can apply conditions known as gauge constraints which
connect the parameters[118]. For each position $\sum_{A_i} \lambda_i(A_i) = 0$ and for each pair of positions $\sum_{A_i} \lambda_{ij}(A_i, A_j) = 0$. This results in two free field parameters per position and four free coupling parameters for every pair of positions. In this work, following Weigt et al.[118], we have chosen the free parameters so as to maximize the fields and minimize the couplings, on average. In a future communication, we will investigate how the choice of the free parameters affects the information carried by the couplings about spatial proximity.

This inverse problem of determining the fields and couplings, given the univariate and bivariate marginals, is computationally challenging[118]. This problem has been described in the literature as the inverse pairwise Ising problem (for two states) and it is computationally expensive because exact methods to determine the marginals from an initial set of Hamiltonian parameters are slow and therefore iterative procedures to search for the field and coupling parameters for many positions and more than two states is unfeasible. Our own previously described method of fitting pair marginals using iterative proportional fitting (IPF) of log-linear model parameters is slow and may not converge within a desired time frame for the problem at hand[44]. Other proposed methods such as Monte Carlo sampling have been applied on Ising models with a few sites, but may require exponential computational time to converge[97]. The approach we have taken involves iterative inference on a probabilistic graphical model using belief propagation described by Weigt et al. 2009[118, 66]. The difference between our approach and the approach taken by Weigt and coworkers is that while converging the bivariate marginals, we use a mean field model which includes pair correlations in the Bethe approximation[7]. Applying the Bethe mean field approximation consistently is just as accurate as using the fluctuation dissipation approach taken by Weigt and coworkers[118, 66]. We will compare the two approaches in a future communication.
5.2.5 Brief outline of the algorithm

The algorithm iteratively converges upon the field and coupling parameters using gradient descent. The outline of the algorithm is as follows.

1. For a given set of field and coupling parameters, determine the corresponding univariate and bivariate marginals using Belief Propagation and the Bethe mean field approximation.

2. Compare these computed marginals to the observed marginals and update the field and coupling parameters.

3. Repeat steps 1 and 2 until the Bethe mean field approximated univariate and bivariate marginals determined from the updated fields and couplings converge to their observed values from the sequence alignment.

For a large system such as the one we are trying to model, Step 1 is the bottleneck and can be extremely expensive. However, graph-theoretic inference methods can reduce the computational time substantially and following the prescription of Weigt et al, we apply graph-theoretic inference methods to iteratively determine field and coupling parameters that describe the Hamiltonian in Equation 5.10 [118]. We model the amino acid positions as nodes connected by edges on an undirected graph, a Markov Random Field (MRF) and exploit its graphical nature by applying computationally efficient inference methods to estimate the univariate and bivariate marginals.

The following sections will introduce and describe basic inference methods on simple acyclic graphs, thus providing a foundation through which inference on complex cyclic graphs can be applied and understood. We will then describe how these graph-theoretic inference methods can be adjusted to solve the inverse problem of determining the fields and couplings given an observed set of
univariate and bivariate marginals.

5.2.6 Belief propagation on factor graphs

We model the variables representing the residues and their interactions on a factor graph. A factor graph is similar to a Markov Random Field as both are models of undirected graphs. However a factor graph is a bipartite graph which is useful if we wish to model and parametrize not just variable nodes and the univariate marginals of each state, but also the edges, and the bivariate marginals for every pair of residues. Since our goal in this inverse problem is to find the field and coupling parameters, given the univariate and bivariate marginals, we use a factor graph representation of the system.

Let \( \{A_1, ..., A_N\} \) be a set of \( N \) random variables representing \( N \) amino acids for example. Each random variable, \( A_i \), has a discrete set of values \( a_i \), which represent the \( q \) possible states or realizations of \( A_i \). This set can be written as \( \{a_1^i, ..., a_q^i\} \). The probability mass function \( P(A_1 = a_1, ..., A_N = a_N) \) is written as \( P(A) \) for short where \( A = \{a_1, ..., a_N\} \), which is a specific realization of a sequence. Similar to a MRF, \( P(A) \) can be factored into a product of factors.

\[
P(A) = \frac{1}{Z} \prod_{\alpha}^{M} f_\alpha(A_\alpha) \tag{5.13}
\]

where the index \( \alpha \), labels the \( M \) factors \( f_\alpha,...,f_M \). Each factor is a function, \( f_\alpha(A_\alpha) \), that takes an argument \( A_\alpha \) which is some subset of \( \{a_1, ..., a_N\} \), each of which can take on \( q \) possible states.

To define the belief propagation equations on a factor graph, we must introduce messages between variable nodes and their corresponding factor nodes and vice versa. The message \( m_{\alpha \rightarrow i}(A_i) \) from the factor node \( \alpha \) to the variable node \( i \) is a vector over the possible states of \( A_i \). This vector of messages can
be interpreted as the beliefs from the factor node to the variable node about
the relative distribution of probabilities at the variable node, based on the
potential at the factor node. The message \( n_{i \rightarrow \alpha}(A_i) \) from variable \( i \) to factor \( \alpha \)
can be interpreted as the beliefs at the variable node regarding its distribution
of probabilities. The messages are updated according to the following rules:

\[
n_{i \rightarrow \alpha}(A_i) \sim \psi_i(A_i) \prod_{c \in N(i) \setminus \alpha} n_{c \rightarrow i}(A_i)
\]

\[
m_{\alpha \rightarrow i}(A_i) \sim \sum_{A_j \setminus A_i} \psi(A_j, A_i) \prod_{j \in N(\alpha) \setminus i} n_{j \rightarrow \alpha}(A_j)
\]

These update rule apply to nodes and factors. The self-consistent belief prop-
agation equations we have implemented in this paper follow the approach of
Weigt et al, 2009[118]. Here the update rules are defines only in terms of vari-
able nodes. As as result, the update rules are rewritten in terms of the message
from one node to another. The messages from the factors to the nodes are
implicit in the update equation. The corresponding BP message update rule
is

\[
P_{i \rightarrow j}(A_i) \sim e^{\lambda_i(A_i)} \prod_{k \neq i, j} \left[ \sum_{A_k} e^{-\lambda_{ik}(A_i, A_k)} P_{k \rightarrow i}(A_k) \right] \quad (5.14)
\]

\[
e^{\lambda_i(A_i)} \sim \frac{P_{i \rightarrow j}(A_i)}{\prod_{k \neq i, j} \left[ \sum_{A_k} e^{-\lambda_{ik}(A_i, A_k)} P_{k \rightarrow i}(A_k) \right]} \quad (5.15)
\]

where \( P_{i \rightarrow j}(A_i) \) is the local message passed from node \( i \) to node \( j \). This message
is a function of the field at \( i \) and the product of all incoming messages from the
neighbors of \( i \), not including \( j \). The BP propagation messages are passed locally
between nodes with random initial values for the messages. Updates are made
and the process is repeated until the messages converge. The proportionality
constant is applied so that the messages at a site sum to 1. Once the messages
have converged, marginals are evaluated by taking the product of the field at
a site with all the incoming messages to that site

\[ P_i(A_i) \sim e^{\lambda_i(A_i)} \prod_{k \neq i} \left[ \sum_{A_k} e^{-\lambda_k(A_i,A_k)} P_{k \to i}(A_k) \right] \]  

(5.16)

Since our implementation of the network is a completely connected undirected graph, with all nodes interconnected to one another, belief propagation is not guaranteed to converge[124, 126, 50]. However belief propagation on cyclic graphs, called loopy belief propagation, may closely approximate the solutions after several iterations[72, 47, 50].

5.2.7 Solving the inverse problem

For our problem, the marginals are known quantities and it is the fields and couplings that we wish to find. Therefore, we actually have an inverse problem; we need to find the fields and couplings given the marginals. This can be achieved by taking the ratios of Equations 6 and 7, a trick described by Weigt et al. 2009[118, 66], thus allowing us to write the message from \( i \) to \( j \) in terms of the known marginal at \( i \).

\[ \frac{P_{i \to j}(A_j)}{P_i(A_i)} = \frac{e^{\lambda_i(A_i)} \prod_{k \neq i,j} \left[ \sum_{A_k} e^{-\lambda_k(A_i,A_k)} P_{k \to i}(A_k) \right]}{e^{\lambda_j(A_j)} \prod_{k \neq i,j} \left[ \sum_{A_k} e^{-\lambda_k(A_i,A_k)} P_{k \to i}(A_k) \right]} \]

(5.17)

Equation 8 can be used to force the univariate marginals estimated by BP to be the observed marginals. As a result, the field parameters never require updating; once the messages converge, the fields can be explicitly calculated using Equation 7. In other words, the univariate marginals are always conserved.

On the other hand, the predicted bivariate marginals need to match the observed bivariate marginals. This can be approximated by the following equa-
\[
P_{ij}^{\text{bethe}}(A_i, A_j) = \frac{\exp[\lambda_{ij}(A_i, A_j)]P_{i\rightarrow j}(A_i)P_{j\rightarrow i}(A_j)}{Z}
\]  

(5.18)

where \( A_i \) and \( A_j \) are the mutations at positions \( i \) and \( j \), \( \lambda_{ij}(A_i, A_j) \) is the statistical coupling parameter between \( i \) and \( j \), \( P_{i\rightarrow j}(A_i) \) is the message passed from \( i \) to \( j \), \( P_{j\rightarrow i}(A_j) \) is the message passed from \( j \) to \( i \) and \( Z \) is the partition function. This equation has been proven by Yedidia and coworkers to be mathematically equivalent to the Bethe approximation, a mean field model, and is what we apply in our code to approximate the bivariate marginals in our system[124, 47, 126].

**5.2.8 The algorithm in detail**

1. Initialization: Set all \( \lambda_{ij}(A_i, A_j) = 0 \) and all \( \lambda_i(A_i) = c_i + \ln P_i(A_i) \), where \( c_i \) is a normalization constant for the gauge constraints which were described earlier.

2. Update messages using Equation 8 for all pairs of residues iteratively until the belief propagation messages converge.

3. Update bivariate marginals \( P_{ij}^{\text{bethe}}(A_i, A_j) \) using the Bethe approximation (Equation 9).

4. Compare \( P_{ij}^{\text{bethe}}(A_i, A_j) \) to \( P_{ij}^{\text{obs}}(A_i, A_j) \), which is the database derived frequency of a double mutation. If the couplings have converged, then stop. If the couplings have not converged by a desired amount, update \( \lambda_{ij}(A_i, A_j) \) as follows

\[
\Delta \lambda_{ij}(A_i, A_j) = -\epsilon[(P_{ij}^{\text{obs}}(A_i, A_j) - P_{ij}^{\text{bethe}}(A_i, A_j)]
\]  

(5.19)

where \( \epsilon \) is the gradient descent step size, set to 0.0001, and repeat steps
2, 3 and 4 until the pair probabilities converge.

5.3 Results

5.3.1 Effect of electrostatic mutations on average biophysical stability

If we divide the Lee database into subsets that have 1, 2, 3… electrostatic mutations and calculate the average AGB folding free energy for each subset, we find that on average the folded state becomes more stable with more electrostatic mutations relative to an independent model which includes only the frequencies of occurrence of each charge at each position (Figure 5.1, solid black curve), as well as in absolute terms (data not shown). It is possible that this increase in electrostatic stability is due to an energetic compensation: if mutations are occurring under selective pressure from drug treatment, and drug-associated mutations which confer resistance are destabilizing to the protein, then the electrostatically active residues could be providing a “reservoir of stabilization”.

To bolster this hypothesis, we examined the Stanford HIV database\(^3\), which has DNA sequences annotated with patient drug treatment history. We find that as the number of protease inhibitors a patient receives increases, the total number of mutations also increases (Figure 5.12A). Some of these mutations have been classified as primary drug resistance mutations, which directly help protease evade drugs at a cost to stability and activity. We reviewed the recent literature and found 12 point mutations, most of which are primary drug resistance positions, that are experimentally destabilizing to protease based on urea denaturation studies\([107, 68, 67, 96, 60, 61, 73, 110]\). Even though 12 positions is a small subset of the 41 positions associated with drug resistance

\(^3\)http://hivdb.stanford.edu/
Figure 5.1: Average $\Delta G_e$ of sequences as a function of the number of electrostatic mutations. Each point on a curve corresponds to $\sum_{n=1}^{N^m} P^n \Delta G^m_n$ where $N^m$ is the number of sequences with $m$ electrostatic mutations, $P^n$ is the probability of the $n$th sequence under a given model normalized with respect to the number of mutations and $\Delta G^m_e$ is its electrostatic folding energy. All points are plotted relative to $\sum_{n=1}^{N^1} P^m \Delta G^m_1$, the average $\Delta G_e$ of observed sequences with 1 mutation. The black curve represents the averages energies of observed sequences. The blue curve represents the average energies of sequences under an independent model in which each charge state has the observed univariate probabilities of occurrence. The red curve represents the average energies of sequences under an independent model in which each charge state has an equal probability of occurrence. The green curve represents the average energy of sequences under a pair model where the observed pair correlations have been preserved. Points on the red, blue and green curves correspond to an average of $10^6$ simulations, where in each simulation, $N^m$ sequences are randomly picked from the corresponding probability distributions. The error bars on the red, blue and green curves are the average standard errors of the mean for these $10^6$ simulations. The error bars on the black curve are the standard errors of the mean of observed sequences.
(17 primary and 24 accessory), the average number of these destabilizing mutations per isolate increases from essentially zero within isolates exposed to no drugs, to 2 mutations within isolates exposed to 6 drugs. Similarly, the average number of electrostatic mutations per sequence increases when going from 0 to 6 drugs. A modest increase can also be seen in Figure 5.12B, where we plot the average number of destabilizing mutations against the number of electrostatic mutations.

The stabilization seen in the observed sequences has a non-trivial origin that requires not only the correct charge frequencies, but also the presence of correlations. Generating random signatures with either equal probabilities for the various charge states results in a substantial destabilization of the protein (Figure 5.1, red curve). However, if we include the correct charge frequencies and introduce pair correlations at the level of the 2-letter W/M alphabet as described above, we see a stabilization comparable to that for the observed
sequences (Figure 5.1, solid green curve).

It is conceivable that the average stabilization observed between the independent and pair-correlated signatures could be distributed among many signatures with each contributing slightly to the average. In fact, this is not the case. In each case, the stabilization can be attributed to a relatively small number of low-energy signatures which are highly unlikely under the independent model, but become very probable when we consider pair correlation. For example, in the case of 2 electrostatic mutations, the \( \approx 7 \) kcal/mol stabilization can be almost entirely attributed to the well-known 30–88 double mutant associated with NFV treatment (Figure 5.8A). For 3 electrostatic mutations, there are three such signatures, all of which involve 30 and 88 with the addition of residues 37, 20, and 69 (Figure 5.8B). The 30–88 theme continues in the 4 and 5 electrostatic mutation cases, with the signature 20, 30, 37, 88 making the largest contribution to the former (Figure 5.8C), and the latter being dominated by the single set of mutations at positions 20, 30, 35, 37, and 88 (Figure 5.8D). All these patterns are associated with NFV\(^4\) treatment in general, and include treatment combinations such as IDV-NFV, NFV-SQV and IDV-NFV-RTV.

5.3.2 Covariance analysis of mutation patterns

We have seen above that groups of mutations play an important role in the average electrostatic energetics of protease. Therefore, it is of interest to find ways of visualizing groups and patterns of mutations that are positively and/or negatively correlated with each other. We can pursue this by making use of the eigendecomposition of the covariance matrix. To focus on the role of correlation in the limit of relatively large numbers of electrostatic mutations, we condition the Lee database on only those signatures with 4, 5 or 6 electrostatic mutations, and select a (different) set of 18 positions which are strongly correlated within

\(^4\)NFV: Nelfinavir, IDV: Indinavir, SQV: Saquinavir, RTV: Ritonavir
this conditioned set of signatures as described above. We then calculate the covariance matrix

$$\Sigma = \begin{bmatrix}
p_1(1 - p_1) & p_{1,2} - p_1 p_2 & \cdots & p_{1,18} - p_1 p_{18} \\
p_{1,2} - p_1 p_2 & p_2(1 - p_2) & \cdots & p_{2,18} - p_2 p_{18} \\
\vdots & \vdots & \ddots & \vdots \\
p_{1,18} - p_1 p_{18} & p_{2,18} - p_2 p_{18} & \cdots & p_{18}(1 - p_{18})
\end{bmatrix}, \quad (5.20)$$

where, as above, $p_i$ is the mutation probability at position $i$, and $p_{i,j}$ is the probability of the double mutant for positions $i$ and $j$. The largest eigenvalues and corresponding eigenvectors of $\Sigma$ represent the largest modes of variability and the relative contributions of the 18 positions to those modes.

In Figure 5.2 we show the eigenvalue spectrum of $\Sigma$ for the conditional database. We see that there are three eigenvalues which are “large”, followed by a gap, and the remainder forming a continuum. The eigenvectors corresponding to these three eigenvalues are shown in corresponding to the three largest eigenvalues are shown in Figure 5.3. We see that the components of these eigenvectors are not uniformly distributed, but rather certain positions “stick out”. For example, the eigenvector corresponding to the largest eigenvalue (Figure 5.3A) is dominated by mutation at residues 30 and 88 (which are positively correlated with each other, since they have the same sign contribution to the eigenvector) with additional contributions from 20 and 92 (which are positively correlated with each other, but negatively correlated with 30 and 88). This pattern is consistent with the mutation pattern associated with NFV, and includes drug combinations such as IDV-NFV, IDV-NFV-RTV, NFV-SQV. The second “eigenmode” (Figure 5.3B), by contrast, is dominated by the single position 37 (which has a high mutation probability), with weak negative correlations of 37 to positions 16 and 20. This mutation pattern is associated with patients receiving APV-IDV-SQV and IDV-NFV. The third eigenmode (Figure 5.3C) is considerably more complex, with a prominent negative correlation of 20 and
69 and smaller contributions from an additional 6 positions. The negative correlation between 20 and 69 is associated with the following combinations of drug therapies: NFV, IDV-NFV-RTV-SQV, IDV-NFV-RTV-SQV, IDV-NFV, IDV-SQV, RTV-SQV.

5.3.3 Relationship between log-linear and Gaussian models for joint mutation probabilities

We made use of a log-linear model (Equation 5.10) to capture the contribution of pair correlations to a joint distribution of mutation probabilities over many positions in HIV protease in Figure 5.1 above, as well as in our previous work[44]. However, the covariance matrix (Equation 5.20) also contains information on pair correlations and their interactions, as we see in Figure 5.3. It is therefore of interest to further examine the relationship between the two methods of capturing pair correlation information.

The log-linear model with parameters estimated by maximum likelihood represents the distribution with maximum entropy consistent with a given set of univariate and bivariate marginals (in our case, the marginal probability of mutation at each position and the probability of a double mutant at each pair of positions, respectively)[39]. The continuous distribution having a mean vector \( \mu \) and covariance matrix \( \Sigma \) with the maximum entropy is the multivariate Gaussian

\[
P(x) \propto \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right].
\] (5.21)

In our case, a natural choice for the mean \( \mu \) is the vector of univariate marginals \((p_1, p_2, \ldots, p_{18})\). The mutation data, however, are discrete and not continuous. We can discretize the Gaussian probability density by restricting \( x \) to be a 0–1 sequence (i.e. evaluating Equation 5.21 at the \( 2^{18} \) vertices of the 18-dimensional unit cube). The constant of proportionality in Equation 5.21 can then be chosen
so that the probabilities over all sequences sum to 1. Note that although the continuous Gaussian distribution is of maximum entropy among all continuous distributions with the same means and covariances, the discretized version need not have any special properties with respect to entropy.

The relationship between the log-linear and discretized Gaussian models can be made explicit if we only consider a small number of positions. In the Appendix, we derive the result for the simplest case of two positions. We show that the log-linear model can be recast as an exponential of a quadratic form just as Equation 5.21, but with a matrix that is not identical to $\Sigma^{-1}$. Therefore, the two models are not mathematically equivalent, and the discretized Gaussian model does not maximize the entropy. Nonetheless, the two models can give similar predictions for the probabilities of individual mutation patterns. Fitting both discretized Gaussian and log-linear models using the Lee database conditioned on 4, 5 or 6 electrostatic mutations, we obtain the joint probabilities shown in Figure 5.4. There is very strong correlation between the two sets of predicted probabilities (Pearson correlation = 0.94 for probabilities $> 10^{-6}$), though the Gaussian probabilities are more strongly peaked than the log-linear, i.e. more of the total probability is concentrated in fewer sequences (Figure 5.5). This latter effect arises because of a systematic underestimation of the probability of low-probability sequences, and overestimation of high-probability sequences. We can introduce an empirical correction for this effect by replacing the factor of $1/2$ in Equation 5.21 with 0.2518 to maximize the agreement between the two sets of probabilities, which greatly reduces the systematic error (Figure 5.5, red dashed curve). Thus, although the log-linear and discretized Gaussian models are different in a mathematical sense (in that the quadratic forms involve different matrices), they both nonetheless capture sequence probabilities in a similar way for real data sets, thereby providing a basis for the interpretation of the eigendecomposition of $\Sigma$. 
5.3.4 Charge patterns explain electrostatic pair correlations

The pair correlations essential for reproducing the observed dependence of stability on the number of mutations can often be understood using a qualitative examination of electrostatic interactions. In particular, we expect that if two sites are interacting electrostatically, then placing like charges at those sites will be destabilizing, while unlike charges will be stabilizing. This is in fact seen in the observed data (Figure ??). For example, for the residue pair 20–35, there is positive correlation in that the wild-type–wild-type combination is enhanced, and in this case this corresponds to the interaction of unlike charges. Similarly, the positive correlation in the pair 16–63 can be understood based on the unlike charges of the most populated mutant forms. On the other hand, the positive correlation in the 18–20 pair arises in part from the unfavorable like-charge combination that arises if 20 remains wild-type and 18 mutates to its most populated state.

The residue pair 63–70 is interesting, in that there is an enhancement involving the two different mutant forms. This suggests that the use of a three-letter alphabet will ultimately be necessary to fully capture the energetics of electrostatic mutations, and that reduction to a two-letter that collapses $M_1$ and $M_2$ may miss some interesting signals. However, the statistical correlations between the mutant forms cannot be extremely strong, as eliminating them in the way that we did above (by introducing correlations at the two-letter level, and adding the information about the mutant forms only as an independent contribution) still gets us close to the observed patterns (e.g. compare the green and black curves of Figure 5.1).

To further investigate the extent to which the qualitative examination of like/unlike charge patterns could be used to explain observed pair correlations, we again made use of the subset of the Lee database conditioned on 4, 5, or 6
electrostatic mutations, for which we calculated $\phi$ using a two-letter alphabet. The sign of $\phi$ (as can be seen from Equation 5.1) indicates which of the 4 mutational states ($WW$, $WM$, $MW$, and $MM$) are enhanced and depleted relative to the independent model: $\phi > 0$ means that $WW$ and $MM$ are enhanced and $WM$ and $MW$ are depleted, while $\phi < 0$ means that $WM$ and $MW$ are enhanced and $WW$ and $MM$ are depleted. Based on the charges of the $W$ and $M = M_1$ states for each position, we can make a qualitative prediction of the sign of the correlation based on an examination of charge patterns, i.e., mutations which lead to like charges are predicted to be disfavored, while mutations which lead to unlike charges are favored. For example, the 30–88 pair would be predicted to have $\phi > 0$, to avoid the juxtaposition of two negative charges that would be caused by the $WM$ state. On the other hand, we expect that 30–92 will have $\phi < 0$, since the juxtaposition of opposite charges that occurs in the $WM$ state is favorable.

If we examine the residue pairs with the largest magnitude pair correlations (Table 5.2), then 11 of the top 15 can be rationalized in this qualitative manner. The 4 that are predicted incorrectly (88–92, 18–88, 20–30, and 34–88) can be rationalized by “frustration” involving a third residue: the most favorable mutation pattern at positions $A$ and $B$ cannot be achieved because it would lead to stronger unfavorable interactions with a third mutually interacting position $C$. For example, the pair 34–88, if it were isolated, would like to have a positive correlation to reduce the probability of the like charge interaction in the $WM$ state. However, both of these residues also have strong correlations with residue 30 which prevents 34–88 from realizing this optimal charge configuration. In particular, if 88 is negative, then 30 will prefer to be neutral due to the 30–88 correlation. If 30 is neutral, then 34 will prefer to be negative due to the 30–34 correlation. However, this leads to 34 and 88 both being negative. On the other hand, if we start with 88 neutral, then 30 is preferred to be negative, and 34 neutral. This also corresponds to the lo-
cally unfavorable state for 34–88, since if the neutral-neutral state is enhanced relative to independent, then so is the negative-negative state.

To confirm this frustration interpretation, we estimated the parameters of the log-linear model for the distribution of electrostatic signatures in the reduced 4–6 mutation Lee database. As has been previously proposed by Weight, et al. [117], each $\lambda_{ij}$ pair parameter of the log-linear model can be interpreted as a statistical interaction energy, and can be thought of as measuring the “direct” pair correlation. In other words, it measures the part of the total correlation given by $\phi_{ij}$ that comes from the direct interaction of the two positions that would occur if the pair was in isolation, as opposed to the “indirect” part that comes from the remainder of the interaction network. If the total correlation between two positions is being influenced by a frustration effect, then we would expect to see $\lambda_{ij}$ and $\phi_{ij}$ to have opposite signs: the effect of the remainder of the network is qualitatively outweighing the direct interaction. In fact, we find this to be the case for 3 of these 4 putative “frustrated” pairs (88–92, 18–88, and 34–88): for each of these, the sign of $\lambda_{ij}$ that is the same as that predicted by the qualitative examination of like/unlike charge patterns, and opposite that of $\phi_{ij}$.

5.3.5 Relationship between log-linear and discretized Gaussian models for 2 positions

For the case of two positions, we can derive an explicit relationship between the log-linear and discretized Gaussian models, since we can write each as an exponential of a quadratic form

$$P(x, y) \propto \exp[-f(x, y)],$$

(5.22)
where
\[
f(x, y) = \begin{pmatrix} x - a \\ y - b \end{pmatrix}^T M \begin{pmatrix} x - a \\ y - b \end{pmatrix},
\]
(5.23)
x and y are discrete 0/1 variables, and \( M \) is a symmetric \( 2 \times 2 \) matrix. The discretized Gaussian model corresponds to \( M = \frac{1}{2} \Sigma^{-1} \), and \((a, b) = (p_1, p_2)\). To transform the log-linear model into a quadratic form, we begin by writing
\[
M = \begin{pmatrix} m_1 & m_{12} \\ m_{12} & m_2 \end{pmatrix}
\]
and expanding out the quadratic form as
\[
f(x, y) = m_1 x^2 + m_2 y^2 + 2m_{12}xy - (2am_1 + 2bm_{12})x - (2bm_2 + 2am_{12})y + c,
\]
where \( c \) is a constant independent of \( x \) and \( y \) that sets the zero of energy. We will neglect the constant \( c \) in the remainder of the derivation. For 0/1 variables, \( x^2 = x \) and \( y^2 = y \). Therefore, we can rewrite the quadratic form (neglecting the constant term) as
\[
f(x, y) = (m_1 - 2am_1 - 2bm_{12})x + (m_2 - 2bm_2 - 2am_{12})y + 2m_{12}xy.
\]
Since the argument of the exponential in the log-linear model is \( \lambda_1 x + \lambda_2 y + \lambda_{12}xy \), we can equate coefficients:
\[
m_{12} = \lambda_{12}/2,
\]
\[
m_1 = \frac{\lambda_1 + b\lambda_{12}}{1 - 2a},
\]
and
\[
m_2 = \frac{\lambda_2 + a\lambda_{12}}{1 - 2b}.
\]
Note that the “mean vector” \((a, b)\) can be chosen arbitrarily.
Table 5.2: The 15 residue pairs in HIV protease with the largest magnitude pair correlation in the Lee database conditioned on 4, 5 or 6 electrostatic mutations, their pair correlation in a 2-letter alphabet, the predicted sense of the correlation based on a naive examination of charge patterns.

<table>
<thead>
<tr>
<th>residue pair</th>
<th>obs. $\phi$</th>
<th>wild type</th>
<th>double mutant</th>
<th>pred. $\phi$ sign</th>
<th>pred. correct?</th>
<th>$\lambda_{ij}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>30–88</td>
<td>0.884</td>
<td>-, 0</td>
<td>0, -</td>
<td>+</td>
<td>yes</td>
<td>11.3</td>
</tr>
<tr>
<td>16–63</td>
<td>0.532</td>
<td>0, 0</td>
<td>-, +</td>
<td>+</td>
<td>yes</td>
<td>0.7</td>
</tr>
<tr>
<td>30–92</td>
<td>-0.345</td>
<td>-, 0</td>
<td>0, +</td>
<td>-</td>
<td>yes</td>
<td>-3.5</td>
</tr>
<tr>
<td>18–30</td>
<td>-0.312</td>
<td>0, -</td>
<td>+, 0</td>
<td>-</td>
<td>yes</td>
<td>-2.6</td>
</tr>
<tr>
<td>88–92</td>
<td>-0.305</td>
<td>0, 0</td>
<td>-, +</td>
<td>+</td>
<td>frustration with 30</td>
<td>2.5</td>
</tr>
<tr>
<td>18–88</td>
<td>-0.282</td>
<td>0, 0</td>
<td>+, -</td>
<td>+</td>
<td>frustration with 30</td>
<td>1.5</td>
</tr>
<tr>
<td>20–30</td>
<td>-0.279</td>
<td>+, -</td>
<td>0, 0</td>
<td>+</td>
<td>frustration with 88</td>
<td>-0.6</td>
</tr>
<tr>
<td>20–88</td>
<td>-0.277</td>
<td>+, 0</td>
<td>0, -</td>
<td>-</td>
<td>yes</td>
<td>-0.2</td>
</tr>
<tr>
<td>17–34</td>
<td>0.268</td>
<td>0, -</td>
<td>-, 0</td>
<td>+</td>
<td>yes</td>
<td>0.4</td>
</tr>
<tr>
<td>63–70</td>
<td>0.264</td>
<td>0, +</td>
<td>+, -</td>
<td>+</td>
<td>yes</td>
<td>0.6</td>
</tr>
<tr>
<td>30–34</td>
<td>-0.249</td>
<td>-, -</td>
<td>0, 0</td>
<td>-</td>
<td>yes</td>
<td>-3.7</td>
</tr>
<tr>
<td>16–37</td>
<td>-0.244</td>
<td>0, 0</td>
<td>-, -</td>
<td>-</td>
<td>yes</td>
<td>-0.9</td>
</tr>
<tr>
<td>20–35</td>
<td>0.240</td>
<td>+, -</td>
<td>0, 0</td>
<td>+</td>
<td>yes</td>
<td>0.2</td>
</tr>
<tr>
<td>16–88</td>
<td>-0.224</td>
<td>0, 0</td>
<td>-, -</td>
<td>-</td>
<td>yes</td>
<td>-0.5</td>
</tr>
<tr>
<td>34–88</td>
<td>-0.201</td>
<td>-, 0</td>
<td>0, -</td>
<td>+</td>
<td>frustration with 30</td>
<td>2.4</td>
</tr>
</tbody>
</table>

Figure 5.2: Spectrum of eigenvalues for the covariance matrix for the 18 highly correlated electrostatically active positions using a two-letter alphabet conditioned on 4, 5, or 6 electrostatic mutations. The ordinate is the eigenvalue and the abscissa is the index number sorted in order of decreasing eigenvalue.
Figure 5.3: Eigenvectors corresponding to the 3 largest eigenvalues of the covariance matrix for the 18 highly correlated electrostatically active positions using a two-letter alphabet conditioned on 4, 5, or 6 electrostatic mutations. Numbers indicate residue numbers of the larger eigenvector components.
Figure 5.4: Comparison of probabilities of individual sequences in a 2-letter wild-type/mutant alphabet over 18 positions under the log-linear model including pair terms, and the discretized Gaussian model described in section 5.3.3. The line denotes perfect agreement between the two probabilities.
Figure 5.5: Cumulative probabilities of the sequences shown in Figure 5.4. Sequences have been sorted in order of decreasing log-linear probability, and the black and red curves denote the cumulative probability under the log-linear and discretized Gaussian models. For example, the 20 sequences with the largest log-linear probabilities account for approximately 45% of the total probability, while those same 20 sequences account for 90% of the probability under the discretized Gaussian model. The red dashed curve is the result for the empirically corrected discretized Gaussian model where the factor of $1/2$ has been replaced by $0.2518$. Each curve reaches 1 for sequence number 218.
Let us choose these parameters to be the univariate marginals as in the Gaussian model: \((a, b) = (p_1, p_2)\). Then the matrix \(M\) is given by

\[
M = -\begin{pmatrix}
\lambda_1 + p_2 \lambda_{12} & \lambda_{12}/2 \\
\lambda_{12}/2 & \lambda_1 + p_1 \lambda_{12} + p_1 \lambda_{12}
\end{pmatrix},
\]

where the log-linear parameters are related to the univariate and bivariate marginals via

\[
\lambda_1 = \ln \left( \frac{p_1 - p_{12}}{1 - p_1 - p_2 + p_{12}} \right),
\]

\[
\lambda_2 = \ln \left( \frac{p_2 - p_{12}}{1 - p_1 - p_2 + p_{12}} \right),
\]

and

\[
\lambda_{12} = \ln \left( \frac{p_{12}(1 - p_1 - p_2 + p_{12})}{(p_1 - p_{12})(p_2 - p_{12})} \right).
\]

Note that the matrices \(M\) for the Gaussian model and the quadratic form representation of the log-linear model are never identical, even if \(x\) and \(y\) are independent. Similar relationships can be derived for more than 2 positions, although then there is no simple relationship between the marginal probabilities and the \(\lambda_i\) and \(\lambda_{ij}\) parameters.

### 5.3.6 Electrostatic effects of packing charges into fixed volume

The interpretation of the averages in Figure 5.1 in terms of a “reservoir of stabilization” compensating for destabilizing drug resistance mutations is not a foregone conclusion, since the folding free energy is dependent on the total number of charges in the protein. [discussion of the physics of packing charges into a fixed volume could go here] We see in Figure 5.6 that the degree of
stabilization of the folded state decreases with the number of charges. Thus, if there is an inverse correlation between the number of electrostatic mutations and total number of charges in the observed sequences, then the stabilization seen in the black curve in Figure 5.1 could be due entirely to the physics of the electrostatic interactions rather than to effects of drug pressure.

That the increasing stabilization of the observed sequences is not due to total charge can be seen if we partition the observed sequences into groups with the same number of positive and negative charges. Since the wild-type state at each position can correspond to any of the three electrostatic states, the number of electrostatic mutations will vary even if we hold the total number of positive and negative charges fixed. If we examine the average electrostatic folding free energy as a function of the number of electrostatic mutations for a given total number of positive and negative charges, we see the same trend as in the total set of observed sequences unpartitioned by total charge: the
Figure 5.7: Average electrostatic folding free energy as a function of the number of electrostatic mutations conditional on 10 positive and 9 negative charges (black) and 11 positive and 9 negative charges (red).

The protein becomes more stable with respect to electrostatics as the number of electrostatic mutations increases (Figure 5.7).

In addition, partitioning the signatures by total charge gives us clues as to the range of validity for our reduced electrostatic model: observed signatures with extreme numbers of charges can lead to predicted folding free energies that are positive, due to the extra stabilization of the unfolded state and increased repulsion within the protein globule. Since the observed sequences presumably came from viable viruses that had a folded and functional protease enzyme, this clearly indicates that the reduced electrostatic model is failing to correctly predict the stability of the protein. This arises in part from the limitations of the model arising from too many charges being placed within a constant volume, and the neglect of entropy. Therefore, in the subsequent analysis, we
excluded all signatures with fewer than 9 or more than 12 positive charges, or fewer than 7 or more than 10 negative charges (i.e. limiting ourselves to the central boxes of the “checkerboard” of Figure 5.7). This is not a substantial limitation, as this eliminates fewer than 0.5% of the observed sequences.

5.3.7 Covariance matrix for a 3-letter charge alphabet

5.3.7.1 Review of 2-letter alphabet

In the 2-letter wild-type/mutant (0/1) alphabet, we assume that each position can take on values of 0 or 1 with probability \((1 - p)\) and \(p\), respectively. Let us consider 2 random variables \(x\) and \(y\), corresponding to 2 different residue positions, and that the probability of mutation at \(x\) and \(y\) is \(p_x\) and \(p_y\), respectively. Let \(p_{xy}\) be the probability of the double mutant at these positions. The variance of \(x\) is defined to be \(\langle x^2 \rangle - \langle x \rangle^2\). Since \(\langle x \rangle = p_x\) and \(x^2 = x\), then the variance is \(p_x - p_x^2 = p_x(1 - p_x)\). The covariance of \(x\) and \(y\) is defined to be \(\langle xy \rangle - \langle x \rangle \langle y \rangle\). Since \(xy\) is non-zero only for the double-mutant, \(\langle xy \rangle = p_{xy}\), and the covariance is \(p_{xy} - p_x p_y\). Thus, the diagonal elements for any position \(x\) is \(p_x(1 - p_x)\), and the off-diagonal element for the pair \(x\) and \(y\) is \(p_{xy} - p_x p_y\). The complete covariance matrix for an arbitrary number of positions can be constructed by “looping over” all possible positions and pairs.

5.3.7.2 Extension to 3-letter alphabet

Suppose we now have 3 possible letters at each position, corresponding to the three possible charge states +/0/-. One way in which we can proceed is to define the random variable \(x\) which takes on the values 1, -1, or 0 with probabilities \(p_{+,x}\), \(p_{-,x}\), and \((1 - p_{+,x} - p_{-,x})\), respectively, corresponding to the probabilities
of the various charges at position $x$. In addition, for 2 arbitrary positions $x$ and $y$, let $p_{++}$ be the probability of 2 positive charges at those positions, $p_{+-}$ be the probability of a positive charge at $x$ and a negative charge at $y$, and similarly for the other pairwise charge combinations. Then $\langle x \rangle = p_{+x} - p_{-x}$, which can be understood as the “average charge” at position $x$. To calculate the variance of $x$, we need to calculate $\langle x^2 \rangle$. The 1 and $-1$ charge states both contribute a value of 1 to this average with probabilities of $p_{+x}$ and $p_{-x}$. Therefore, the variance is $\langle x^2 \rangle - \langle x \rangle^2 = p_{+x} + p_{-x} = (p_{+x} - p_{-x})^2$. The quantity $xy$ which must be averaged to obtain the covariance is non-zero for only 4 charge combinations, namely $+1$ for $++$ and $--$, and $-1$ for $+-$ and $-+$. Therefore, $\langle xy \rangle = p_{++} + p_{--} - p_{+-} - p_{-+}$, and the covariance is

$$
\langle xy \rangle - \langle x \rangle \langle y \rangle = p_{++} + p_{--} - p_{+-} - p_{-+} - (p_{+x} - p_{-x})(p_{+y} - p_{-y}).
$$

As for the 2-letter case, the complete covariance matrix for an arbitrary number of positions can be constructed simply by “looping over” all possible positions $x$ and pairs $xy$.

### 5.3.8 Change of basis for log-linear models

In a 2-letter alphabet, the log-linear model is given by

$$P(x_1 \ldots x_n) \propto \exp(E), \quad (5.24)$$

where the energy $E$ is given by

$$E = \sum_i \lambda_i x_i + \sum_{i,j} \lambda_{ij} x_i x_j, \quad (5.25)$$
where (suppose) $x_i$ is 0 if position $i$ is wild-type, and 1 if it is mutant. Suppose we have found the values of $\lambda_i$ and $\lambda_{ij}$ in this basis, and we now want to change the basis to one where $x_i$ is −1 if position $i$ is wild-type, and 1 if it is mutant. We can do this without refitting the $\lambda$’s. This basis change is equivalent to changing variables from $x_i = 0$ if position $i$ is wild-type, and 1 if it is mutant to an new variable $x'_i = 2x_i - 1$. From this, it follows that $x_i = (x'_i + 1)/2$.

Substituting into Equation 5.25, we find
\[
E = \frac{1}{2} \sum_i \lambda_i (x'_i + 1) + \frac{1}{4} \sum_i \sum_{j>i} \lambda_{ij} (x'_i + 1)(x'_j + 1). \tag{5.26}
\]

We expand out Equation 5.27 leaving out any terms that are independent of $x'$, since such terms will only have an effect on the normalization constant in Equation 5.24. Doing this, we find that
\[
E = \sum_i \left[ x'_i \left( \frac{1}{2} \lambda_i + \frac{1}{4} \sum_{j>i} \lambda_{ij} \right) \right] + \frac{1}{4} \sum_i \sum_{j>i} \lambda_{ij} x'_i x'_j \tag{5.27}
\]
\[
= \sum_i \lambda'_i x'_i + \sum_i \sum_{j>i} \lambda'_{ij} x'_i x'_j. \tag{5.28}
\]

Thus, we end up with an energy of exactly the same form as Equation 5.25, except that the $\lambda$’s have been transformed as $\lambda'_i = \lambda_i/2 + \frac{1}{4} \sum_{j>i} \lambda_{ij}$ and $\lambda'_{ij} = \lambda_{ij}/4$. A similar calculation can be done for any basis change by writing it in the form $x'_i = ax_i + b$.

### 5.3.9 Predicted probabilities for observed and unobserved sequences
Figure 5.8: Plots of $\Delta G_e.P_2$ versus $\Delta G_e.P_1$ for observed (red) and unobserved (black) sequences with 1, 2, 3, 4, 5 and 6 electrostatic mutations. These plots represent the contributions of each sequence to the average electrostatic folding free energies shown in Figure 5.1. Each point represents a unique sequence with $\Delta G_e.P_1$ and $\Delta G_e.P_2$ on the x-axis and y-axis respectively, where $P_1$ and $P_2$ are the independent and pair log-linear model probabilities of the sequence, normalized with respect to the number of mutations, and $\Delta G_e$ is the electrostatic folding energy. Sequences which contribute an extraordinary degree to the stabilization of the folded state under the pair correlation model are labeled by their mutation pattern. Mutations are represented as $aNb$ where $N$ is the residue number and $a$ and $b$ are one of the 3 charged states (+, -, n). The straight line on each diagram is a plot of $x = y$. Sequences below this line have $P_1 < P_2$, resulting in $\Delta G_e.P_1 > \Delta G_e.P_2$. For these sequences, the electrostatic stabilization is greater under the pair model than under the independent model.
Figure 5.9: Observed (red) and unobserved (green) sequences with 4 mutations sorted by their pair correlation model probabilities.

Figure 5.10: Observed (red) and unobserved (green) sequences with 5 mutations sorted by their pair correlation model probabilities.
Figure 5.11: Observed (red) and unobserved (green) sequences with 6 mutations sorted by their pair correlation model probabilities.
5.4 Discussion

The above results show the evidence of a significant impact of electrostatic interactions on the coevolution of mutations in HIV protease to a degree not previously noted in the literature. In fact, at least one experimental study of protease has minimized the role of electrostatics in favor of the impact of compensatory mutations on protein flexibility[86], despite the fact that long-range electrostatic interactions can in general have a substantial effect on protein stability[94]. Almost all previous studies of electrostatics in HIV protease have focused on the short-range effects of the mutations D30N and N88D. D30N is an almost exclusively an NFV resistance mutation, as it results in a loss of a stabilizing hydrogen bond leading to weaker ligand binding[48, 67, 75]. N88D, though considered to be a primary drug resistance mutation for some drugs, is an accessory compensatory mutation for NFV evasion and coevolves very strongly with D30N[75, 48, 93]. The N88D mutation, which is less than 3Å away from the sidechain of position 30, either compensates for the loss of negative charge at the active site, or pulls the position 30 asparagine away from the inhibitor [48, 67, 75, 55].

Electrostatics, however, is only one part of the total energy of the protein, and the remaining contributions to the stability such as nonbonded and torsional energies are not completely negligible. Nonetheless, electrostatic interactions are extremely important[115, 94], and there are reliable models that can be used to calculate their energetic contribution[35]. The importance of the role of electrostatic mutations in protein function has been underscored recently by a study that showed that sequence correlations among charged residues are more prevalent in proteins associated with disease[29]. Though their focus was on diseases related to protein aggregation, the authors observe that even HIV proteins have a high degree of mutational correlation among charged residues,
Figure 5.12: Observed number of mutations as a function of treatment history (number of protease inhibitors administered to the patient, panel A), and the number of electrostatic mutations (panel B) in the Shafer annotated database. In each panel, the black curve represents the total number of non-electrostatic mutations, the red curve represents the number of mutations known experimentally to be destabilizing, and the green curves represent the number of non-electrostatic primary (solid) and accessory (dashed) drug resistance mutations. In panel A, the blue line represents the total number of electrostatic mutations.
and they further speculate that blocking these charges may incapacitate the virus. Our results suggest that the presence of mutational correlations among electrostatic residues may be intimately connected to protein stability. HIV protease is constantly under strong selection pressure due to clinical therapy using protease inhibitors. As a result of the initial build up of resistance to these drugs in the form of primary drug resistance mutations, protease becomes less stable. In order to compensate for this loss of stability, it requires stabilizing accessory mutations. These mutations not only bring the protein back to a more viable state, but may give the protein more “breathing room” on the evolutionary fitness landscape [12, 30, 111]. These mutations increase the evolvability of the protein by helping them to escape traps on their pathway towards higher fitness.

We observe that the number of electrostatic mutations increase with both the number of drugs, and the number of experimentally verified destabilizing mutations. This represents further circumstantial evidence in favor of the hypothesis that compensation for these destabilizing mutations comes from a reservoir of electrostatic mutations. In fact, even the electrostatic stabilization of sequences with increasing number of drugs reflects this trend (Figure 5.13). We recognize that in grouping together destabilizing point mutations we have implicitly assumed that their destabilizing effects are additive, and, as a result, have ignored the possible effects of epistasis. Even so, Figure 5.12 shows many mutations whose frequency increases substantially with both drug exposure and the number of electrostatic mutations, but whose stabilities have not yet been classified experimentally. These mutations could be destabilizing or neutral, but we conjecture that the majority of them cannot be stabilizing. A protein can only survive in a certain stability range, and can never be too stable. We therefore claim that the evolvability of HIV protease is substantially, if not completely, dependent on stabilizing electrostatic mutations which compensate for the loss of stability due to drug resistant mutations and serve
Figure 5.13: Free energy of folding for sequences with drugs versus without drugs.
to keep HIV protease within a narrow stability range. This evidence is, admittedly, equivocal, as the sequence database information contains confounding factors. For example, it may initially be puzzling as to why the number of electrostatic mutations (Figure 5.12A, blue curve) does not increase more strongly as a function of protease inhibitor exposure. This is due in part to the fact that the Lee database is dominated by HIV subtype B, whereas the Shafer database has a substantial number of other HIV subtypes which differ in their content of charged residues. Furthermore, the number of annotated protease inhibitors may be misleading, as some of the treatment durations can be very short.

Clearly, a more definitive exploration of these questions requires the ability to directly calculate changes in protein stability due to mutations beyond the electrostatic component studied here. This is a difficult problem[51, 88], and we are currently developing computational methods based on modern molecular mechanics energy functions and sophisticated sampling methods that will provide estimates of free energy changes due to single and multiple mutations. This will allow us to more directly test our electrostatic stabilization hypothesis, and help to further understand the link between observed sequence correlations and protein structure and energetics. In addition, a more comprehensive statistical treatment of pair correlation using log-linear models is desirable, both in terms of expanding the number of positions that we can simultaneously consider (from 18 to 55, in principle), as well as treating the problem in the natural 3-letter alphabet (without the extra step of reduction to the 2-letter wild-type/mutant alphabet). Figure 5.14 shows that not only is the 3rd state significantly mutated, but that its mutation frequency increases with the number of electrostatic mutations. Both of these goals could in principle be accomplished by advanced methods such as message passing[117], which we have implemented. Nonetheless, our results suggest that the 2-letter approximation is reasonably good, and that most of the important correlation effects are being captured in the 18 residues subsets that we have chosen.
Figure 5.14: Electrostatic mutation frequency for the least populated charge state for each position as a function of the number of electrostatic mutations.
5.5 Appendix I: Response letter attached

The response letter to the reviews from initial submission to PNAS is attached.
April 3, 2012
Response to reviews of PNAS MS# 2012-011171
Title: Correlated electrostatic mutations provide a reservoir of stability in HIV protease

Dear PNAS Editorial Board,

I am writing to request that our manuscript “Correlated electrostatic mutations provide a reservoir of stability in HIV protease” be sent out for further review. We submitted the manuscript for publication in PNAS (track II), the Biophysics section on January 20, 2012 and were notified on March 1 that the Editorial Board rejected our manuscript. Although the two PNAS reviewers both stated that our work was of high quality with conclusions that were well justified, it was rejected because it was deemed to not be of sufficient general interest for the PNAS audience. For the reasons stated below, we strongly disagree with their opinion. Our work will be of interest to a very broad range of scientists working in such disciplines as biophysics (electrostatic energy calculations), structural biology (drug design), evolutionary biology (correlated mutations), computational biology (sequence analysis methods), virology (HIV), and clinical pharmacology (drug resistance). In any case, we found several of the reviewers’ suggestions to be helpful in improving aspects of our paper and have revised the text to improve the clarity of the presentation. The revisions are shown in red. We include the reviewers’ critique of the manuscript and our response below:

Reviewer #1 writes: “The authors’ work supports the well established importance of electrostatics in protein stability using HIV protease as an example. This is a small incremental advance in the field of protein stability. While the general concept of electrostatics and stability is of common interest, this paper would be better suited in a specialty journal perhaps in the field of biostatistics or biophysics as the article is more about correlating the authors’ pair wise correlation model for predicting mutations with observed drug resistant mutations.”

Response: Our work bridges the gap between the analysis of protein sequences using statistical techniques developed by the physics and computer science communities, and the biophysical modeling of protein energetics. As such, it is a significant advance in our understanding of how electrostatic interactions among mutating residues contribute to protein stability. Using information theoretic methods together with a a coarse-grained (Generalized Born) energy model we have analyzed the contribution of electrostatic interactions to protein stability among mutated residues of HIV-1 protease based on models derived from a large database of sequences which have acquired drug resistance. We find that the increased stability, which results from the electrostatic interactions among the mutated residues, is intimately related to the correlations between the mutations. Uncorrelated mutations, even if the mutation frequencies at individual positions are preserved, would strongly destabilize the enzyme. In the course of this work we have constructed a mean field model at the level of pair correlations (Bethe approximation) to predict the probabilities of observing mutated sequences using the HIV sequence database to parameterize the model. The model has remarkable predictive power. We demonstrate that it can be used to predict complex mutation patterns involving many simultaneous mutations whose probabilities lie in the tails of the distribution, yet
the accuracy of the model is much better than corresponding estimates based on sequence counts from the database itself.

We go beyond a purely statistical approach to modeling patterns of electrostatic mutations in HIV protease, and show that the statistical results are consistent when viewed in the context of a structure based energy model. We build coarse grained structural models of the charge distributions corresponding to mutated sequences and show that the placement of charges according to the pair correlation model leads to stabilizing electrostatic contributions to the folding energy, while if a primitive mean field model is used which preserves only the frequencies of the charged mutations at individual residue positions but not the pair correlations, the resulting charge patterns are destabilizing. There is no other work in the literature that we are aware of that integrates statistical inference and biophysical structure-energy viewpoints and shows how important pair correlations are for protein stability.

The novelty of our approach to analyzing how correlated electrostatic mutations can provide a reservoir of stability for an enzyme, the power of our mean field pair correlation model to predict complex mutation patterns, and the consistency of the statistical model with the structure based energy model, together warrant publication in PNAS.

Reviewer #1 writes: “Also, the authors focused on only 18 amino acid positions without addressing the effects of polymorphic electrostatic mutations on protein stability. How will the model described here be able to distinguish between polymorphic mutations and drug resistant ones?”

Response: In our pair correlation model, we include all mutations which change the charge at any position above a threshold frequency, including all primary, accessory, and polymorphic drug resistance mutations (as designated by the Stanford HIV database). Positions not included in our analysis had electrostatic mutation frequencies below 0.01%. We have revised the text on p. 2 to make this clear.

Reviewer #1 writes: “This is an interesting approach to predict mutations, however, it fails to correlate well when more than 3 mutations in a particular molecule are present which is often the case in multi-drug resistant sequences.”

Response: This is a very unfortunate misunderstanding as a major point of our paper is that although our model for the joint distribution of mutations at eighteen positions involving changes in the charge state of HIV Protease is based on pairwise correlations, the model correctly captures the probabilities of sequences in the tail of the distribution with many more than three mutations. The model has remarkable predictive power; the probabilities of observing sequences in the tail of the distribution when queries are made from finite size samples of the order of tens of thousands of sequences are predicted accurately. This can be seen by looking at Figure 2 (reproduced below) and Supplementary Figure S4.
Figure 2 plots the pair correlation probabilities of sequences as a function of the sequence probabilities from the Lee database (the training database). Each dot represents a unique sequence. Additionally, the dots are shaded according to a color gradient which corresponds to the probability of observing the sequence in a database that was not used to parameterize the model (the Stanford database), relative to its probability in the Lee database. The plot clearly shows that within the tail of the Lee distribution (which include sequences with up to six mutations), sequences that are frequently observed in the Stanford database have higher pair correlation model probabilities than sequences which are absent or not frequently observed in Lee. For example, sequences to the left of the abscissa break are not observed in our training set. Within this subset, however, sequences that are observed in the Stanford database (red dots) have higher pair correlation model probabilities than sequences that are unobserved in both databases (green dots). Similar trends can be observed for other subsets of sequences seen once, twice three times etc. in the Lee database, indicating that the pair correlation model is a very good predictor of the probabilities of sequences in the tail. In other words, the model can be parameterized on one of the databases and predict the properties in the tail of the other,
including the probabilities of seeing sequences in the test database that were not in the training database.

We have revised the manuscript on p. 3 to make this important point clearer and added a new figure in the supplementary material (Supplementary Figure 4) that shows the distribution of pair correlation model probabilities for sequences in the tail that are observed at least once in the Stanford database, is significantly different from the distribution of predicted probabilities for sequences that are unobserved in the Stanford database (p-value that the means of the two distributions are the same < 10^{-4}).

**Reviewer #2 writes:** “1) I don't think there has been much of an attempt to make this article accessible to a general audience with a lot of detailed mathematical modeling. Simple things for instance rather than using the nomenclature 30n or 88-8, using D30N(n) and N88D(-) would help extremely in making it more accessible. A description early on about frustration and how it fits in also would help.”

**Response:** We have changed the nomenclature to make it consistent and accessible throughout. At the suggestion of the referee, we have added a description about frustration early on (p. 2). Reviewer #2 also seems to be concerned that the manuscript will not be accessible to a general audience because it contains a lot of mathematical modeling. In fact there is very little reference to the details of the mathematical modeling in the main text. The mathematical framework for our pair correlation model is summarized in the supplementary material, but even in the supplementary material very few details are presented. The pair correlation model we use to fit the distribution of HIV protease sequences is well known in condensed matter physics and in computer science and is used to solve pattern recognition problems when posed as an inverse problem. In the last few years it has been taken up by biophysicists. Papers have appeared in the Biophysics section of PNAS very recently using the same model – also called the maximum entropy model – to relate the coevolution of sequences across protein families with native contacts of the corresponding structures in 3D space, and also to model the sequence ensemble that gives rise to the diversity of antibodies. Our manuscript treats the mathematical details of the pair correlation model at the same level of detail as those papers which recently appeared in the Biophysics section of PNAS. [PNAS Papers in 2009 and 2011 by Onuchic, Hwa and colleagues, and a paper by Callan and Bialek et al. in PNAS 2010]

**Reviewer #2 writes:** “2) The analysis is limited to 18 sites of 99 sites where charge changes occur in HIV protease. From the supplementary data it seems that the vast majority include changes at 30 and 88. In fact if these two sites were eliminated from the analysis how robust would the conclusions be? While charged changes occur at the other 16 sites these are likely very low probability events when querying the Stanford database. Thus how general are the conclusions. To make a general statement like the title infers, several independent sets of charged changes should be observed.”

**Response:** The analysis includes all positions in HIV protease where there is a change in the charge with frequency greater than 0.01%, of which there are 18 positions. With three possible charge states at each of the 18 positions, we are modeling the probability distribution of more than 380 million states. While it is true that D30N-N88D are strongly coupled double mutations in both the Lee and Stanford HIV protease databases, only about 11% of the mutated sequence signatures involve D30N and N88D. The remaining 89% do not involve D30N and N88D.
Moreover, the univariate marginals indicate that a third of the positions have mutation frequencies higher than the rate for D30N and N88D. In other words, charged changes involving the other 16 sites are not very low probability events, as the reviewer suggests. We believe that our conclusions are highly robust because we observe approximately 840 independent sets of charged changes in HIV protease, of which ~600 do not involve D30N-N88D. All of these patterns are well described by the pair correlation model, albeit the patterns with three or more mutations with the highest probabilities do involve D30N-N88D.

Reviewer #2 writes: “3) If this is a general mechanism by which the enzyme stabilizes resistance mutations, why are so few charged changes occurring. This is an important question to be addressed.”

Response: First, we would like to reiterate that we are not suggesting that electrostatic mutations are the only mechanism by which the mutated enzyme is stabilized, but it is an important one. Analysis of the drug annotated Stanford HIV database shows that while 10% of the sequences from drug-naïve patients contain two or more electrostatic mutations, this percentage quadruples to approximately 40% for drug exposed sequences. Conversely, for those (two hundred unique) sequences with four or more electrostatic mutations, all have acquired resistance through exposure to HIV protease drugs.

Reviewer #2 writes: “4) Stronger explanation and justification for the extensive use of the Lee database is needed, while the Stanford database is a well established, well maintained and well known database, the Lee database is not.

Response: We agree that the Stanford database is well established and better known than the Lee database, although the Lee database is larger (45,000+ sequences) than the Stanford database (13,000+ sequences). The reason we used the Lee database as a training set is mostly historical, as we worked with it for several years and we published a paper in 2009 using the Lee database as a basis for parameterization. Having said that, the differences between the univariate and bivariate marginals of the two databases is negligible. The correlation coefficient is ~0.99 which means that we could have parameterized our pair correlation model on the Stanford database and we would have reached the same conclusions.

There was also included with the reviews, an attachment which appears to have been included with review #2.

Additional concerns raised by a reviewer:

First, it is unlikely that the wild-type HIV-1 protease charge signature was determined based on a sufficiently large sample size. There are several HIV-1 protease sequences considered being “wild-type”.

Second, only 18 mutation positions were considered in this study. Mutations in other positions may also contribute to protein stability, such as compensatory mutations of the electrostatic effects caused by mutations in the 18 selected positions.

Third, the study does not explain the multidrug resistant HIV-1 protease variants having only aliphatic amino acid mutations. There are numerous multidrug resistant HIV-1 protease variants only containing neutral drug resistance mutations.
Fourth, the statement that the predicted “mutation patterns...have not yet been observed due to finite sample size effects” is probably inaccurate considering that HIV-1 protease is an enzyme that recognizing multiple substrates and most of the predicted mutation patterns based on stabilization may cause enzyme malfunction.

**Response:** The definition of the HIV protease wild type we adopted is the HIV-1 subtype B consensus sequence that is defined by the Los Alamos National Laboratory HIV sequence database (http://www.hiv.lanl.gov). This consensus sequence was verified by comparing it to the Stanford HIV database consensus sequence (http://sierra2.stanford.edu/sierra/html/asi/releaseNotes/index.shtml#consensussequences) and also by independently determining the consensus sequence from the ~60,000 HIV protease subtype B sequences from Lee and Stanford databases. We have already addressed the second and third concerns, namely that electrostatic mutations are not the only source of mutations which contribute to the stability among drug resistant HIV protease sequences, but they are an important source. The story we tell about the coupling among the mutations at all the electrostatically active positions and the importance of their correlations for protein stability is a compelling one, even if there are other mutating positions which undoubtedly play a role in the acquisition of drug resistance.

The fourth concern suggests that because HIV-1 protease is an enzyme that must recognize many substrates, most of the predicted mutation patterns based on stabilization which we claim have not yet been observed due to finite sample size effects, may instead be because they cause enzyme malfunction. This is an interesting point; we have indirect evidence that the as yet unobserved mutation patterns among the set of eighteen electrostatically active positions included in our pair correlation model are indeed not observed because of finite sample size effects. We have performed sampling tests on the pair correlation model to ask how many unique sequences would be observed in datasets constructed from this distribution of sample size ~13,000 (corresponding to Stanford) and ~45,000 (corresponding to Lee). When we do this test many times and construct the distribution of results, we find that the number of unique sequences actually observed in Lee and Stanford are entirely consistent with this test (see Fig. S6). If there were additional strong constraints on the mutation patterns for the set of eighteen positions we are modeling beyond stability, then the number of unique sequences observed in Lee and Stanford would not be consistent with draws from the pair correlation model.

In summary, we have in this paper constructed a pair correlation model for the probabilities of observing mutations which change the charge state at any one of the eighteen residue positions in HIV protease observed in the Lee or Stanford databases which mutate at frequencies greater than 0.01%. We have demonstrated that this model has remarkable predictive power; the probabilities of observing sequences in the tail of the distribution when queries are made from finite size samples of the order of tens of thousands of sequences are predicted accurately. We have gone beyond a purely statistical approach to modeling and have shown that our statistical results are entirely consistent when viewed in the context of a structure based energy model. Using a coarse grained structure-energy model for the charge distributions corresponding to the mutated sequences, we show that the placement of charges according to the Bethe mean field distribution which preserves the pair correlations, leads to stabilizing contributions to the folding energy on average, while if a primitive mean field model is used, the
resulting charge patterns are destabilizing on average. The combination of state-of-the-art statistical inference modeling with biophysical energy modeling is highly novel. We apply our methods to obtain insights concerning a problem of great importance, the molecular basis for the acquisition of drug resistance in HIV protease through correlated electrostatic mutations which affect the stability of the enzyme. We strongly believe that our manuscript is of the caliber that meets the high standards of PNAS.

Thank you for considering our request to have our revised manuscript sent out for further review.

Sincerely,

Ronald M. Levy

Board of Governors Professor of Chemistry and Chemical Biology
Rutgers University
5.6 Appendix II: Publication attached

Parts of the previous chapter were submitted for publication. The submitted publication is attached.
Correlated electrostatic mutations provide a reservoir of stability in HIV protease

Omar Haq\textsuperscript{1}, Michael Andrec\textsuperscript{1,2}, Alexandre V. Morozov\textsuperscript{3,1} Ronald M. Levy\textsuperscript{1,2,*}

\textbf{1} BioMaPS Institute for Quantitative Biology, Rutgers University, Piscataway, New Jersey, USA
\textbf{2} Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, New Jersey, USA
\textbf{3} Department of Physics & Astronomy, Rutgers University, Piscataway, New Jersey, USA

\* E-mail: Corresponding ronlevy@lutece.rutgers.edu

Abstract

HIV protease, an aspartyl protease crucial to the life cycle of HIV, is the target of many drug development programs. Though many protease inhibitors are on the market, protease eventually evades these drugs by mutating at a rapid pace and building drug resistance. The drug resistance mutations, called primary mutations, are often destabilizing to the enzyme and this loss of stability has to be compensated for. Using a coarse-grained biophysical energy model together with information-theoretic methods, we observe that accessory mutations of charged residues increase protein stability, playing a key role in compensating for destabilizing primary drug resistance mutations. Increased stability is intimately related to correlations between electrostatic mutations – uncorrelated mutations would strongly destabilize the enzyme. Additionally, statistical modeling indicates that the network of correlated electrostatic mutations has a simple topology and has evolved to minimize frustrated interactions. The model's statistical coupling parameters reflect this lack of frustration and strongly distinguish like-charge electrostatic interactions from unlike-charge interactions for \( \approx 90\% \) of the most significantly correlated double mutants. Finally, we demonstrate that our model has considerable predictive power and can be used to predict complex mutation patterns, that have not yet been observed due to finite sample size effects, and which are likely to exist within the larger patient population whose virus has not yet been sequenced.

Author Summary

HIV is incurable because its enzymes evolve rapidly by developing resistance mutations to retroviral inhibitors. Most of these mutations work synergistically, but the biophysical basis behind their cooperation is not well understood. Moreover, it is not yet possible to accurately predict patterns of mutations that are not yet observed in the patient population. As a result, the ability to model and predict patterns of drug resistance mutations is of significant therapeutic relevance. Our work addresses these important issues by bridging the gap between the statistical modeling of HIV protease subtype B sequences with the energetics of mutations involving charged amino acids by showing that electrostatic stability is intimately related to correlations. Moreover, we demonstrate that our statistical model has considerable predictive power and can be used to predict complex mutation patterns that have not yet been observed due to the finite sizes of the current sequence databases. In other words, as the database size increases, our model has the ability to predict the identities of the high probability mutations patterns, which are more likely to be observed. Knowing which currently unobserved mutations are more likely to be observed can be very advantageous in combating the disease.

Introduction

Proteins evolve through random mutagenesis and their evolutionary selection is constrained by structural, functional and environmental factors [1]. Thermodynamic stability is by far the most important structural
factor, as most proteins need to be folded in order to function. The stability range for each protein, however, is narrow and is estimated experimentally to be around 10 kcal/mol, which is of the order of three hydrogen bonds [2]. As a result of this marginal stability, proteins operate “on a knife’s edge” [3], whereby a single highly deleterious mutation could potentially lead to decreased stability and loss of activity [4]. By the same token, a single stabilizing mutation could be advantageous from an evolutionary point of view. For example, more stable forms of cytochrome P450 allowed for greater exploration of mutational space in directed evolution experiments than sequences without stabilizing mutations [5]. This increased “evolvability” is not just limited to directed evolution experiments, but may be a general property of proteins evolving under selective pressure [6]. In fact, recent experimental work on HIV protease has shown that accessory mutations compensate for the loss of stability due to destabilizing primary drug resistance mutations, helping the virus evade drugs [7]. This stabilizing effect can have an external source as well: Hsp90, a molecular chaperone, buffers deleterious mutations, allowing for polymorphisms to appear and new traits to evolve [8]. As a result of this work and prior research by other groups, it is now widely recognized that thermodynamic stability is intimately linked with the evolvability of a protein [9–11].

Another constraint on protein evolution is sign epistasis, which put limits on mutational pathways. This occurs when a new mutation is improbable, or even impossible, under one genetic background, but possible under a different genetic background [12]. Sign epistasis could be due to a change in protein stability: a particular mutation can be stabilizing or deleterious depending on the existing amino acids in the protein sequence, leading to a viable or nonviable protein [1,13]. This phenomenon results in pathway restriction, further reducing the complexity of evolutionary sequence space.

Even though the process of mutagenesis is random, the genetic and structural constraints mentioned above, coupled with functional selection, ensure that certain mutations in evolving proteins are associated with each other in a highly non-random fashion [14]. These correlated mutations are an inherent property of evolving amino acid sequences, and an evolutionary signature of viable, natural proteins. A multitude of methods have been developed to identify such pairs and groups of mutations [15], some of which have been applied to HIV protease sequences to locate pairs [16–18] or groups [19,20] of coevolving residues [16,17,20]. Our previous work on higher-order correlations showed that for HIV-1 protease, including at least pair correlations is essential for reproducing statistical patterns of primary and accessory mutations observed in protease sequences from patients undergoing antiretroviral therapy [21].

It is tempting to attribute sequence correlations to the effects of epistasis arising from protein stability constraints [13], and several groups have tried to connect sequence correlations with protein energetics on a detailed atomic level. For example, Ranganathan et al. have attempted to explain mutational coevolution by connecting statistical free energies from multiple sequence alignments to differences in experimental folding free energies [22]. However, some of these results have been difficult to replicate [23] and are still a topic of active debate in the community [24,25]. Thus while studies that link mutational correlations to thermodynamic constraints have made great progress [14,26–28], a consensus linking protein energetics and mutational correlation patterns has not yet emerged. These observations have motivated us to explore how correlated mutations in HIV protease are connected to energetics via their impact on protein stability.

Since current methods for predicting stability changes upon mutation based on detailed atomic models are not sufficiently accurate [29], we have chosen to focus instead on the electrostatic part of the total energy for which a coarse-grained model of electrostatics is appropriate. We find that this model captures many important effects of mutations on energetics and stability of HIV protease. We show that the average electrostatic stabilization of HIV protease increases with the number of electrostatic mutations (an electrostatic mutation changes the charge of that mutating residue relative to the wild-type residue), consistent with the hypothesis that accessory electrostatic mutations buffer the destabilizing effects of primary drug-resistance mutations, most of which are non-electrostatic mutations and are therefore not modelled here. We demonstrate that correlations among electrostatic mutations are critical for stabiliza-
tion; uncorrelated mutations would strongly destabilize the protein. We show that our method, which employs both electrostatic calculations and sequence analysis based on statistical inference techniques, can be used as a predictive tool for novel mutational patterns that have not yet been observed. Finally, we comment on the structure of the electrostatic mutation network of HIV protease. Energy landscape theory, which provided the framework for understanding protein folding through funnels, introduced the concept of a smooth, minimally frustrated landscape for foldable, natural proteins [30, 31]. Our results indicate that the electrostatic interaction network is minimally frustrated as is evident in the derived statistical coupling parameters which strongly predict the underlying charge patterns, providing additional evidence that proteins have evolved to minimize frustrated interactions.

Results

Effect of electrostatic mutations on protein stability.

Our analysis of electrostatic mutation patterns is based on the alignment of $\sim 45,000$ HIV protease sequences from Christopher Lee’s HIV Positive Selection Mutation Database (http://bioinfo.mbi.ucla.edu/HIV) [32]. Each amino acid sequence in the Lee database is converted into a charge signature, which is a three letter alphabet representation of that sequence (+, -, n) corresponding to positively-charged, negatively-charged, and neutral residues. These charge signatures are compared to the wild-type charge signature to determine electrostatic mutations. We examined all primary, accessory and polymorphic drug resistance mutation positions (as designated by the Stanford HIV database [33]) and limited our analysis to a subset of 18 positions whose charged state mutates above a threshold frequency of 0.01%. Our model therefore includes more than 380 million states or unique charge signatures involving these 18 positions. Of the 18, 9 are sites which have been characterized as primary or accessory drug resistance mutations while the rest are sites labeled as polymorphic mutations. Mutations are labelled “polymorphic” if they are observed to mutate in the absence of drugs and whose compensatory effect has not yet been experimentally verified, even though drugs may have a significant affect on their correlations with other mutating residues. [33]

If we divide this database of charge signatures into subsets with 1, 2, 3... electrostatic mutations and calculate the electrostatic contribution to the average folding free energy $\Delta G_e$, for each subset, we find that on average the stability of the folded state increases by $\approx 5$ kcal/mol from 1 to 3 mutations and maintains this level of stabilization beyond 3 mutations (Figure 1, black curve). Since selective pressure in the presence of inhibitors often leads to destabilizing primary drug resistance mutations [34, 35], the observed increase in electrostatic stability is due to energetic compensation: destabilizing mutations occur due to selective pressure and electrostatically active residues provide a “reservoir of stability”.

The observed stabilization requires not only the correct frequencies of occurrence of each of the three possible charge states at each position, but also the presence of correlations. Generating random sequences with equal mutation frequencies for the three charge states results in a substantial destabilization of the protein (Figure 1, red curve). Introducing observed frequencies of occurrence of each charge at every position improves the stabilization relative to the previous model with equal mutation frequencies, but still results in substantial destabilization (Figure 1, blue curve). We refer to this latter model as the independent model as it generates an alignment in which mutations at each position occur independently with correct frequencies.

If pair correlations are introduced by preserving the observed joint mutation frequencies (see Methods), substantial protein stabilization occurs, and the energies predicted by this pair correlation model (Figure 1, green curve) become comparable to the energies of the observed sequences. The magnitude of the difference between the observed and pair correlation model average energies is less than 2 kcal/mol for sequences with $\leq 5$ mutations, suggesting that introducing pair correlations is sufficient for explaining the observed energetic stabilization trends. Overall, the difference between the independent model and the pair correlation model is statistically significant (e.g. $p < 10^{-4}$ for sequences with 4 or fewer
mutations), although the sample size dependent error bars grow larger with the number of mutations as fewer sequences are observed in the database.

**Contribution of specific sequences to the average protein stability and significant drug associations.**

We find that the observed electrostatic stabilization can be attributed to a relatively small number of low-energy signatures which are highly unlikely under the independent model but become very probable once pair correlations are introduced (Supporting Information (SI) Figure 1). For example, the well-studied pair of primary drug resistance mutations, D30N-N88D [17, 34], which occurs 2220 times in the Lee database, contributes \( \approx 64\% \) to the \( \approx 5 \) kcal/mol stabilization of the pair correlation model relative to the independent model shown in Figure 1. Together, the top 10 double mutants account for \( \approx 83\% \) of the stabilization of the pair correlation model relative to the independent model.

With increasing numbers of mutations, the stabilization spreads among multiple patterns (SI Figure 1). For 3 electrostatic mutations, the top contributor D30N-N37D-N88D is responsible for 20\%, while the top 10 signatures account for 64\% of the \( \approx 8 \) kcal/mol stabilization of the pair correlation model relative to the independent model. For 4 mutations, K20I-D30N-N37D-N88D accounts for 10\%, while the top 10 signatures account for 33\% of the stabilization. For 5 mutations, K20I-D30N-E35Q-N37D-N88D accounts for 17\% and the top 10 signatures account for 36\% of the stabilization.

These highly stabilizing charge patterns are also strongly associated with protease inhibitor therapies, as determined by our drug association analysis (see Supplementary Material). Most protease drug association studies focus on point mutations or pairs of mutations [19,32]. Our drug association analysis allows us to examine the significance of drug association for patterns of more than two mutations. SI Table 1 lists the most significant associations between drugs and charge patterns of 2, 3 and 4 electrostatic mutations with the highest probabilities as predicted by the pair correlation model. Most of the patterns listed are strongly associated with at least one drug and several are associated with many drugs. For example the D30N-N88D double mutant and the D30N-N37D-N88D triple mutant are both strongly associated with Nelfinavir monotherapy and Indinavir-Nelfinavir combination therapy with \( p < 10^{-7} \). We also find strong association between drugs and patterns predicted by the pair correlation model with more than three mutations. For example K20I-D30N-N37D-N88D and K20I-D30N-E35Q-N88D are both associated with Indinavir-Nelfinavir combination therapy while K20I-D30N-H69Q-N88D is associated with Ritonavir-Nelfinavir therapy with \( p < 10^{-7} \).

**Predicting novel mutational patterns.**

The pair correlation model allows us to predict the probabilities of arbitrary charge signatures, many of which have not yet been experimentally observed. SI Figure 2 shows that most of the sequences with less than 5 mutations, whose probabilities are significantly enhanced by pair correlations, are observed in the Lee database, indicating that these mutational patterns are routinely utilized by the virus. However, for 6 mutations the most stabilizing pattern, K20I-D30N-E35Q-N37D-Q58E-N88D, was not observed (SI Figure 1). The probability of this pattern under the pair correlation model is \( 1.2 \times 10^{-5} \), too small to appear frequently in a database of \( \sim 45,000 \) sequences due to finite sample size effects. If the size of the database were to increase five-fold, the probability of observing at least one copy of this pattern would be \( > 0.90 \).

The proportion of sequences not observed in the Lee database with significantly enhanced pair correlation model probabilities increases greatly with the number of mutations, of which it is likely that many are not observed because of finite sample size effects (SI Figure 2). In order to test our ability to predict novel patterns of favorable electrostatic mutations unobserved in the Lee database due to finite sample size effects, we examined the contents of a separate database, the drug-annotated Stanford database...
which contains HIV protease subtype B sequences from various sources [33]. Figure 2 plots the probabilities of sequences using the pair correlation model, $P_2$, as a function of the observed probabilities in the Lee database. Sequences are also shaded according to a gradient that represents how often the sequence occurs in the Stanford database, relative to the Lee database. The plot shows that sequences with the highest predicted $P_2$ probabilities that are unobserved in the Lee database are overwhelmingly shaded red, indicating that most are observed in the Stanford database. In fact, of the top 25 most probable sequences predicted by the pair correlation model that are not found in the Lee database, 19 are present in the Stanford database (SI Table 2). Most of these sequences are also significantly associated with drug therapies (e.g. $p < 0.05$). As the predicted $P_2$ probability decreases, the shading of the dots gradually changes to green, indicating that sequences with the lowest predicted $P_2$ probabilities are unobserved in both the Lee and the Stanford database. Thus our approach exhibits considerable predictive power.

If we examine other sequences in the tail of the Lee probability distribution that are observed once, twice, three times etc in the Lee database, we notice a similar trend; sequences with higher predicted $P_2$ probabilities are present in the Stanford database at much higher frequencies than sequences with lower predicted $P_2$ probabilities, even though the pair correlation model was parameterized on the Lee database. SI Figure 3 highlights this effect and shows that the distribution of pair correlation model probabilities for sequences in the tail that are observed at least once in the Stanford database is very different from the distribution of predicted probabilities for sequences that are unobserved in Stanford. The p-value for the null hypothesis, which states that the means of these two distributions are equal, is $< 10^{-4}$, indicating that the difference between the two means is statistically significant (SI Figure 3).

Though the Stanford database is smaller than the Lee database, the differences between the univariate and bivariate marginals of the two databases is negligible. The correlation coefficient is 0.99 which means we could have parameterized our pair correlation model on the Stanford database and reached the same conclusions. In fact, both databases are simply finite size draws from the distribution accurately described by the pair correlation model. Sampling tests on the pair correlation model highlight this effect. We asked how many unique signatures would be observed in datasets constructed from the distribution described by the pair correlation model of sample size 13,000 (corresponding to Stanford) and 45,000 (corresponding to Lee). When we do this test many times and construct the distribution of results, we find that the number of unique sequences actually observed in Lee and Stanford are entirely consistent with this test (SI Figure 4 and 5).

These results suggests that the pair correlation model is a much better predictor of actual sequence probabilities than using the sequence counts from the databases themselves, because of finite sample size effects. In other words, the tail of the distribution is very well represented by the pair correlation model.

Structure of the electrostatic mutation network.

Determining the statistical field and coupling parameters (written as $\lambda_i$ and $\lambda_{ij}$ for simplicity) that best fit the pair correlation model given a set of observed univariate and bivariate marginals ($P_{ij}^{obs}$ and $P_{ij}^{obs}$), is known in the literature as the inverse Ising problem. As described in the Methods, we iteratively determine these parameters using a graph-theoretic inference algorithm, called belief propagation (BP), [36–38] a method which has recently been applied by other research groups to study protein conformational entropy, ligand binding and protein-protein interactions [39,40]. Within the BP framework, we apply a mean field model which includes pair correlations in the Bethe approximation to estimate the bivariate marginals, $P_{ij}^{bethe}(A_i, A_j)$ [37,41,42]. However, it is well known that while $P_{ij}^{bethe}$ is exact and converges to $P_{ij}^{obs}$ on acyclic networks, $P_{ij}^{bethe}$ only approximates $P_{ij}^{obs}$ and can become unstable on cyclic networks [43]. For the electrostatic correlation network of HIV protease, we observe that the belief propagation algorithm converges quickly to the observed bivariate marginals (Figure 3). The convergence towards the observed probabilities for triplets and larger multiplets is also well approximated, a result that is non-trivial since the Bethe approximation is a pair-level approximation and does not guarantee the convergence for marginals beyond pairs even for acyclic networks [44]. For trivariate marginals, the correlation coefficient
between \(P_{bethe}^{ijkl}\) and \(P_{obs}^{ijkl}\) is 0.98 while for four mutations, the correlation coefficient between \(P_{bethe}^{ijkl}\) and \(P_{obs}^{ijkl}\) is 0.90 (Figure 3). This close correlation between observed and predicted marginals argues for a simple network structure and suggests that for this system, the Bethe approximation is a good approximation and by implication the electrostatic mutation network is minimally frustrated.

Another strong indicator of the lack of frustration in the electrostatic mutation network of HIV protease is that the statistical coupling parameters of the (Bethe) mean field model are able to distinguish like-charge patterns from unlike-charge patterns with high accuracy (Figure 4). We find that the sign of \(\lambda_{ij}\), a quantity derived from the sequence analysis alone, is able to correctly predict the charge patterns for \(\approx 90\%\) of the top 35 most significantly correlated charge pairs (inset, Figure 4), reflecting the evolutionary optimization of the protein electrostatic interaction network. Moreover, Figure 4 also indicates that the magnitude of \(\lambda_{ij}\) correlates with the spatial distance between residues. Of the top ten pairs of residues with the largest statistical coupling parameters, nine are situated close to one another (< 10 Å) in the folded structure of the HIV protease homodimer. In this context, we note that Morcos et al. [45] used a similar approach to infer spatial contacts between residues of many proteins, through an analysis of the coupling parameters of a corresponding mean field model.

**Discussion**

Our results suggest that electrostatic interactions play an important role in the coevolution of mutations in HIV protease. The extent to which electrostatics influences protein stability has been the subject of debate in the literature [46–52]. One experimental study of protease has minimized the role of electrostatics in favor of the impact of compensatory mutations on protein flexibility [53]. Others have suggested that buried charges play a more important role in protein function than stability [54, 55]. We recognize that electrostatics is only part of the total energy, and that contributions to stability from van der Waals interactions, hydrogen bonding and hydrophobic effects are significant. Nonetheless, long-range electrostatics is likely to have a substantial effect on protein stability [49, 51, 56–58]. Moreover, compensatory electrostatic mutations have been shown to be significantly correlated in some systems [59], and have been suggested to provide viable evolutionary pathways [60].

Almost all previous studies of electrostatics in protease have focused on the short-range effects of the mutations D30N and N88D [34]. D30N is believed to be a Nelfinavir resistance mutation, as it results in the loss of a stabilizing hydrogen bond leading to weaker ligand binding [61]. It is speculated that N88D, which coevolves strongly with D30N [17], either compensates for the loss of the negative charge at the active site, or helps pull 30N away from the inhibitor [61]. While it is true that D30N-N88D is a statistically prominent double mutation in both the Lee and Stanford databases, only about 11% of the mutated sequences involve D30N and N88D. The remaining 89% of sequences do not involve D30N and N88D. We believe our conclusions are highly robust because approximately 840 independent sets of charged changes are observed in HIV protease, of which 600 do not involve D30N-N88D. All of these mutation patterns are well described by the pair correlation model, albeit the patterns with three or more mutations with the highest probabilities do involve D30N-N88D. Complementing previous studies on HIV protease, our drug association analysis finds multiple statistically significant links between several electrostatic mutation patterns and drug treatment therapies.

Our results support the proposal that the presence of correlations among electrostatic mutations arises from the constraints imposed by the need to maintain the stability of the folded protein. HIV protease is under strong selective pressure from drugs. As a result of the initial build up of drug resistance, protease becomes less stable [34]. We hypothesize that electrostatic mutations play a role in bringing the protein back to a more viable state, by giving the protein more “breathing room” on the evolutionary fitness landscape. Analysis of the Stanford drug annotated sequence database shows that while 10% of the sequences from drug-naive patients contain two or more electrostatic mutations, this percentage quadruples to approximately 40% for drug exposed sequences [33]. Conversely, for those (two hundred)
sequences with four or more electrostatic mutations, all have acquired resistance through exposure to HIV protease drugs. These trends clearly show that electrostatic mutations are playing a major role in the evolution of drug resistance in HIV protease. However, random, uncorrelated patterns of electrostatic mutations are impossible. Manipulating the charge distribution of HIV protease is complex and we find that uncorrelated mutations would tend to strongly destabilize the enzyme, contrary to the stability gain observed in the database. Therefore, we propose that sets of electrostatic mutations occur together, increasing the “evolvability” of a protein by providing a “reservoir of stability” which allows it to escape epistatic traps along evolutionary pathways towards higher fitness [13].

The absence of frustration could reflect evolutionary optimization of the electrostatic interaction network in HIV protease under selection pressure from drugs, or it could be a general property of protein electrostatic interaction networks. Indeed, natural proteins tend not to be frustrated systems [31,62] – they are fine-tuned biological machines with restricted evolutionary pathways [13]. Within these pathways, proteins are highly robust and the physics underlying their folding display a kind of simplicity [31,63–65]. Our conclusions based on a coarse-grained electrostatics model combined with information-theoretic techniques reflect this lack of frustration.

Our information-theoretic analysis of HIV sequences captures biophysical constraints in the form of a statistical network of correlated mutations. Even though the model is based on pairwise correlations, it captures the higher-order effects and correctly predicts the probabilities of sequences with six or more mutations found in the tail of the distribution. The fact that many of these patterns are also strongly associated with protease inhibitors from patients undergoing antiretroviral therapy, highlights the clinical relevance of our method. Other mutation patterns that we predict are likely to exist within patients whose virus has not yet been sequenced. Having knowledge of these unique, but as yet unobserved patterns, can be important for the design of future inhibitors to combat drug resistance.

In this work, we go beyond a purely statistical approach to modeling patterns of electrostatic mutations, and show that the statistical results are entirely consistent when viewed in the context of a structure based energy model. Though electrostatics is only part of the total energy, our work has highlighted its importance and provided support for the proposal that correlated electrostatic mutations provide a reservoir of stability for HIV protease as it builds resistance to drugs.

Methods

HIV sequence databases.

45,161 aligned HIV-1 DNA nucleotide sequences were downloaded from Christopher Lee’s HIV Positive Selection Mutation Database (http://bioinfo.mbi.ucla.edu/HIV) on March 4th, 2008 [32]. This database of sequences, which we call the Lee database, consists primarily of HIV-1 subtype B samples [32]. The amino acid sequences were converted into strings of characters “n”, “-” and “+”, indicating whether a given residue is neutral, negatively or positively charged at pH=6 (i.e. His, Arg, and Lys are positively charged, while Asp and Glu are negatively charged). A second database of subtype B sequences, which we call the Stanford database, was downloaded from the Stanford HIV database on April 7th, 2010 [33]. This drug-annotated dataset was used to associate correlated mutation patterns with specific anti-retroviral therapies (see SI). The univariate and bivariate marginals extracted from the Lee database and from the Stanford database are effectively the same (correlation coefficient of 0.999), indicating that our results would be unchanged if we used the Stanford database to parameterize the model.

To locate electrostatic mutations, the resulting charge signatures were compared to the HIV-1 subtype B consensus sequence from the Los Alamos National Laboratory HIV sequence database (http://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html). This consensus sequence was used to define the wild-type charge signature. We define an electrostatic mutation as an amino acid mutation which changes the charge at a certain position along the protein sequence, relative to the wild-type amino acid at that position (e.g. D30N and N88D). In contrast, the L90M and R8K are not considered to be electrostatic mutations.
We examined all the primary, accessory and polymorphic drug resistance mutations positions (as designated by the Stanford database [33]) and included all electrostatic mutations above a threshold frequency of 0.01%. The 18 positions included are the primary drug resistance mutation sites D30 and N88, accessory mutation sites K20, E34, E35, K43, Q58, L63, and Q92, and polymorphic mutation sites Q7, T12, G16, Q18, N37, Q61, H69, K70, and I72.

**Calculation of electrostatic folding free energies.**

The electrostatic energy of protein folding $\Delta G_e$ was estimated as

$$\Delta G_e = G_e^{(f)} - G_e^{(u)}$$

where $G_e^{(f)}$ and $G_e^{(u)}$ are electrostatic free energies of the folded and unfolded states, computed using an Analytical Generalized Born (AGB) model, an implementation of the pairwise descreening Generalized Born (GB) model that makes use of a parameter-free algorithm to take into account atomic overlaps [66]. Both $G_e^{(f)}$ and $G_e^{(u)}$ were obtained using

$$G_e \geq G_{GB} = u_e \sum_i q_i^2 / B_i + 2u_e \sum_{i<j} q_i q_j / r_{ij} + \sum_{i<j} \frac{q_i q_j}{\epsilon_{in} r_{ij}}$$

where $q_i$ is the charge of atom $i$, $B_i$ is its Born radius, $f_{ij} = \sqrt{r_{ij}^2 + B_i B_j \exp(-r_{ij}^2/AB_i B_j)}$ is the GB distance between atoms $i$ and $j$, $r_{ij}$ is the distance between the charges, and $u_e = 1/(1/(\epsilon_{in} - 1/\epsilon_{in})$ ($\epsilon_{in}$ is the solute dielectric constant and $\epsilon_{in}$ is the solvent dielectric constant) [66].

The folded state electrostatic free energy $G_e^{(f)}$ was calculated by placing unit charges corresponding to a particular charge signature onto the most-distal sidechain carbon atom of the corresponding wild-type amino acid within a dimer crystal structure (PDB ID 1NH0 [67]). All other sidechain atoms remain neutral, although a partial charge dipole of ±0.4e is placed on every backbone amide and carbonyl group to retain the helix dipole effects [68]. Our approximation of the denatured state is a maximally extended structural representation of chain A from 1NH0, with backbone dihedral angles set to 180° (except for prolines) and sidechain rotamer states set to all-trans. Similarly to the folded state, charges on the unfolded state are placed on the most-distal sidechain carbon atom and backbone dipoles are switched on.

**Statistical modeling of sequence probabilities.**

As in our previous work [21], we make use of a Potts model to capture the effects of pair interactions between residues. Since our electrostatic data consists of sequences with three possible charge states at each site, we use a 3-letter alphabet (+, -, n), for positively-charged, negatively charged, and neutral residues. Including all three charge states in our study leads to $3^{18} = 387,420,489$ possible charge signatures for 18 positions.

For each signature we calculate $P_1$, the independent model probability, and $P_2$, the pair correlation model probability. Specifically, we fit the frequencies of charge states at each position and the joint frequencies of charge states at pairs of positions to the 3-state Potts model:

$$P_2(A_1 A_2 \ldots A_N) = \frac{1}{Z} \exp[\sum_{i=1}^N \lambda_i(A_i) - \sum_{i<j} \lambda_{ij}(A_i, A_j)]$$

where $A_1 A_2 \ldots A_N$ is a sequence of +’s, −’s or n’s of length $N$, $i, j$ are position indices, $\lambda_i$ and $\lambda_{ij}$ are the fitting parameters for the fields and couplings, and $Z$ is the partition function. The independent model, obtained by setting all $\lambda_{ij} = 0$, corresponds to $P_1(A_1 A_2 \ldots A_N) = P(A_1)P(A_2)\ldots P(A_N)$. For the equal frequency model in Figure 1, we set $P(A_i) = 1/3, \forall i$.

The joint probability distribution given by the Potts model has the largest entropy constrained by the univariate (independent model) or both univariate ($P^{obs}_i$) and bivariate ($P^{obs}_{ij}$) marginals from the
data [69]. To solve the inverse Ising problem, we implemented an efficient graph-theoretic inference algorithm called belief propagation (BP) [36–38]. Our algorithm employs a two-step procedure: first, all the univariate and bivariate marginals are determined for a given set of $\lambda_i$ and $\lambda_{ij}$ in the Bethe approximation within BP [37, 41, 42]. Second, the predicted marginals are compared to the observed marginals to determine updated $\lambda_i$ and $\lambda_{ij}$ via gradient descent [38]. See the supporting information for further information about the algorithm.

Acknowledgments

We thank Emilio Gallicchio for many discussions and assistance with implementation of the coarse-grained Generalized Born electrostatic model. This work was supported by National Institutes of Health (GM30580 to R.M.L. and HG004708 to A.V.M.). A.V.M. is a recipient of an Alfred P. Sloan Research Fellowship.

References


Figure 1. Average electrostatic free energy of folding as a function of the number of electrostatic mutations. Each point on a curve corresponds to $\sum_{n=1}^{N^m} P^m_n \Delta G_e^n$, where $N^m$ is the number of sequences with $m$ electrostatic mutations, $P^m_n$ is the probability of the $n$th sequence under a given model conditional upon the number of mutations, and $\Delta G_e^n$ is its electrostatic folding energy (Equation (1); see Methods). All points are plotted relative to $\sum_{n=1}^{N^1} P^m_{obs} \Delta G_e^n$, the average $\Delta G_e$ of observed sequences with one electrostatic mutation. The black curve shows the average energies of observed sequences ($P^m_{obs} = 1/N^m$), the red curve represents the average energies of sequences under a model in which each charge state occurs with equal frequency, the blue curve shows the average energies of sequences under a model in which each charge state occurs with frequencies observed in the data, and the green curve represents the average energy of sequences under a pair correlation model which preserves observed pair frequencies. The error bars on the black curve are the standard errors of the mean of observed sequences. Note that $N^m = \{13470, 4798, 1515, 337, 50, 11\}$ for $m = 1 \ldots 6$. 


Figure 2. Comparison of the sequence probabilities in the tail of the Lee database and the pair correlation model with the sequence probabilities in the Stanford database. The probabilities of sequences under the pair correlation model, $P_2$, predicted using the Bethe approximation, are plotted as a function of the sequence probabilities from the Lee database, $P_{LEE}$. Sequences with a probability of 0 in the Lee database, i.e. unobserved sequences, are plotted to the left of the abscissa break. Every sequence is shaded using a color gradient corresponding to $P_{ST}/P_{LEE}$, which represents the number of times the sequence occurs in the Stanford database, relative to its probability in the Lee database. Sequences that occur frequently in the Stanford database as compared to the Lee database have a higher ratio and are shaded red, while the sequences that do not occur as frequently in the Stanford database as compared to the Lee database have a lower ratio and are shaded blue. Sequences that are shaded green have equal probabilities in both databases. The plot legend further explains the correspondence between the color gradient and the ratio of probabilities each color in the gradient represents. Sequences unobserved in the Lee database, but observed in the Stanford database have a ratio that is artificially set to equal 4, which corresponds to the color red. Unobserved Lee sequences that are shaded green are also unobserved in the Stanford database. Sequences with probabilities $<10^{-7}$ or $>10^{-3}$ are not shown. The indices (0), (1), (2), etc mark the locations of sequences observed zero, once, twice (etc) in the Lee database. Each dot corresponds to a unique sequence.
Figure 3. Comparison between the observed and predicted mutivariate marginals for 2, 3 and 4 mutations. Predicted marginals determined using belief propagation in the Bethe approximation are plotted against the observed marginals for sets of 2, 3, and 4 mutations. The correlation between $P_{ij}^{bethe}$ and $P_{ij}^{obs}$ is 1.00. The correlation between $P_{ijk}^{bethe}$ and $P_{ijk}^{obs}$ is 0.98. The correlation between $P_{ijkl}^{bethe}$ and $P_{ijkl}^{obs}$ is 0.90.
Figure 4. Distance between like and unlike-charge pairs as a function of the statistical coupling parameter, $\lambda_{ij}$. The statistical coupling parameter $\lambda_{ij}$ is a fitting parameter that describes the statistical interaction energy between pairs of mutations. A strongly negative $\lambda_{ij}$ indicates that a pair of mutations is enhanced, while a strongly positive $\lambda_{ij}$ indicates that a pair of mutations is suppressed. Using simple electrostatics, we observe that like-charge patterns (blue) are mostly suppressed (negatively correlated) while unlike-charge patterns (red) are enhanced (positively correlated). Inset: Fraction of pairs, whose electrostatic interaction is correctly predicted by the sign of $\lambda_{ij}$, arranged in order of decreasing $|\lambda_{ij}|$. 
Supplementary Information: Correlated electrostatic mutations provide a reservoir of stability in HIV protease

Omar Haq\(^1\), Michael Andrec\(^{1,2}\), Alexandre V. Morozov\(^{1,3}\), Ronald M. Levy\(^{1,2,*}\)

\(^1\) BioMaPS Institute for Quantitative Biology, Rutgers University, Piscataway, New Jersey, USA
\(^2\) Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, New Jersey, USA
\(^3\) Department of Physics & Astronomy, Rutgers University, Piscataway, New Jersey, USA

* E-mail: Corresponding ronlevy@lutece.rutgers.edu

Materials and Methods

Posing the inverse problem for the Potts model for sequence probabilities

Our Potts model for correlated electrostatic mutations deals with a reduced sequence length of 18 (from 99 amino acids) and a reduced amino acid alphabet size of 3. For length \(N = 18\) and a \(Q = 3\) letter alphabet (0, +, -), there are \(3^{18} = 3^{18} = 387,420\) possible sequences or unique charge signatures. For every signature, we wish to calculate the probability of that sequence under two models: an independent model which preserves the database derived univariate marginals and a mean field model (Bethe approximation) which preserves both the database derived univariate and bivariate marginals. We call this the pair correlation model and we refer to the probabilities of a sequence under the independent and pair correlation model as \(P_1\) and \(P_2\) respectively. \(P_1\) probabilities of individual sequences can be obtained by simply taking the product of the observed univariate marginals at each position.

\[
P_1(A_1, ..., A_N) = \prod_{i=1}^{N} P_{\text{obs}}(A_i) \tag{1}
\]

where \(A_i = \{0, +, -\}\) is one of the three possible charges at position \(i\) and \(P_{\text{obs}}(A_i)\) is the observed univariate marginal of charge \(A_i\) at position \(i\). All observed univariate and bivariate frequencies are derived from the Lee database [1].

The probability of a sequence under the pair correlation model \(P_2\), on the other hand, is determined by a model which generates sequences that preserve both the observed univariate and bivariate marginals. In order to do so, we fit the sequence signature probabilities to a 3-state Potts model where the Hamiltonian, \(\mathcal{H}\) is described by field and coupling parameters, which reflect the mutation frequency at a site and the strength of the statistical coupling between two sites respectively.

\[
P_2(A_1, ..., A_N) = \frac{1}{Z} \exp[\mathcal{H}(A_1, ..., A_N)] \tag{2}
\]

where \(P_2(A_1, ..., A_N)\) is the probability of a sequence of length \(N\) consisting of charges \(A_i\) at position \(i\) which preserves the univariate and bivariate marginals. The Hamiltonian can be defined as

\[
\mathcal{H}(A_1, ..., A_N) = \sum_i \lambda_i(A_i) - \sum_{i<j} \lambda_{ij}(A_i, A_j) \tag{3}
\]

where \(\lambda_i(A_i)\) is the field at position \(i\) for charge \(A_i\), \(\lambda_{ij}(A_i, A_j)\) is the coupling between charges \(A_i\) and \(A_j\) at positions \(i\) and \(j\) and \(Z\) is the partition function.

\[
Z = \sum_{A_i} \exp[\mathcal{H}(A_1, ..., A_N)] \tag{4}
\]

This model is the maximum entropy solution to the probability distribution that matches the single-point and double-point correlations [2]. Note that if there were two possible states at each site instead of three,
this Potts model would be equivalent to the famous Ising model, which is widely applied in the study of spin glass systems in statistical physics.

For a system of 18 positions and 3 states, \( 18 \times 3 = 54 \) \( \lambda_i(A_i) \) and \( \binom{18}{2} \times 3^2 = 1377 \) \( \lambda_{ij}(A_i, A_j) \) parameters need to be determined to accurately describe the Hamiltonian which preserves the observed marginals. But the parameters are not independent as we can apply conditions known as gauge constraints which connect the parameters [3]. For each position \( \sum_{A_i} \lambda_i(A_i) = 0 \) and for each pair of positions \( \sum_{A_i} \lambda_{ij}(A_i, A_j) = 0 \). This results in two free field parameters per position and four free coupling parameters for every pair of positions. In this work, following Weigt et al. [3], we have chosen the free parameters so as to maximize the fields and minimize the couplings, on average. In a future communication, we will investigate how the choice of the free parameters affects the information carried by the couplings about spatial proximity.

This inverse problem of determining the fields and couplings, given the univariate and bivariate marginals, is computationally challenging [3]. This problem has been described in the literature as the inverse pairwise Ising problem (for two states) and it is computationally expensive because exact methods to determine the marginals from an initial set of Hamiltonian parameters are slow and therefore iterative procedures to search for the field and coupling parameters for many positions and more than two states is unfeasible. Our own previously described method of fitting pair marginals using iterative proportional fitting (IPF) of log-linear model parameters is slow and may not converge within a desired time frame for the problem at hand [4]. Other proposed methods such as Monte Carlo sampling have been applied on Ising models with a few sites, but may require exponential computational time to converge [5]. The approach we have taken involves iterative inference on a probabilistic graphical model using belief propagation described by Weigt et al. 2009 [3,6]. The difference between our approach and the approach taken by Weigt and coworkers is that while converging the bivariate marginals, we use a mean field model which includes pair correlations in the Bethe approximation [7]. Applying the Bethe mean field approximation consistently is just as accurate as using the fluctuation dissipation approach taken by Weigt and coworkers [3,6]. We will compare the two approaches in a future communication.

**Outline of the algorithm**

The algorithm iteratively converges upon the field and coupling parameters using gradient descent. The outline of the algorithm is as follows.

1. For a given set of field and coupling parameters, determine the corresponding univariate and bivariate marginals using Belief Propagation and the Bethe mean field approximation.
2. Compare these computed marginals to the observed marginals and update the field and coupling parameters.
3. Repeat steps 1 and 2 until the Bethe mean field approximated univariate and bivariate marginals determined from the updated fields and couplings converge to their observed values from the sequence alignment.

**Belief propagation**

Belief Propagation (BP) or the Sum-Product algorithm is an iterative procedure applied to efficiently calculate all the marginals in a tree-like graph [8]. It consists of leaf nodes passing messages to their parents, which in turn process these messages and pass them onwards towards the root. The root then sends messages back to its children nodes and so on until the messages eventually reach the leaf nodes. At this point, for acyclic graphs, the messages will converge [9,10]. For cyclic graphs, this method is approximate but may converge after several cycles of message passing [11–13].

The self-consistent belief propagation equations we have implemented in this paper follow the approach of Weigt et al. 2009 [3]. The probability distribution of the pair correlation model \( P_2 \) is given by Equation 3. For this distribution, the corresponding BP message update rule is

\[
P_{i \rightarrow j}(A_i) \sim e^{\lambda_i(A_i)} \prod_{k \neq i,j} \left[ e^{-\lambda_{ik}(A_i, A_k)} P_{k \rightarrow i}(A_k) \right] \tag{5}
\]
where $P_{i \to j}(A_i)$ is the local message passed from node $i$ to node $j$. This message is a function of the field at $i$ and the product of all incoming messages from the neighbors of $i$, not including $j$. The BP propagation messages are passed locally between nodes with random initial values for the messages. Updates are made and the process is repeated until the messages converge. The proportionality constant is such that the messages at a site sum to 1. Once the messages have converged, marginals are evaluated by taking the product of the field at a site with all the incoming messages to that site

$$P_i(A_i) \propto e^{\lambda_i(A_i)} \prod_{k \neq i} e^{-\lambda_{ik}(A_i, A_k)} P_{k \to i}(A_k)$$

(6)

Since our implementation of the network is a completely connected undirected graph, with all nodes interconnected to one another, belief propagation is not guaranteed to converge [9, 13, 14]. However belief propagation on cyclic graphs, called loopy belief propagation, may closely approximate the solutions after several iterations [11–13].

For our problem, the marginals are known quantities and it is the fields and couplings that we wish to find. Therefore, we actually have an inverse problem; we need to find the fields and couplings given the marginals. This can be achieved by taking the ratios of Equations 6 and 7, a trick described by Weigt et al. 2009 [3, 6], thus allowing us to write the message from $i$ to $j$ in terms of the known marginal at $i$.

$$\frac{P_{i \to j}(A_i)}{P_i(A_i)} = \frac{e^{\lambda_i(A_i)} \prod_{k \neq i} e^{-\lambda_{ik}(A_i, A_k)} P_{k \to i}(A_k)}{\prod_{k \neq i} e^{-\lambda_{ik}(A_i, A_k)} P_{k \to i}(A_k)}$$

$$P_i(A_i) = \frac{\sum_{A_j} e^{-\lambda_{ij}(A_i, A_j)} P_{j \to i}(A_j)}{\sum_{A_j} e^{-\lambda_{ij}(A_i, A_j)} P_{j \to i}(A_j)}$$

(7)

Equation 8 can be used to force the univariate marginals estimated by BP to be the observed marginals. As a result, the field parameters never require updating; once the messages converge, the fields can be explicitly calculated using Equation 7. In other words, the univariate marginals are always conserved.

On the other hand, the predicted bivariate marginals need to match the observed bivariate marginals. This can be approximated by the following equation:

$$\mathcal{P}^{\text{bethe}}_{ij}(A_i, A_j) = \frac{\exp[\lambda_{ij}(A_i, A_j)] P_{i \to j}(A_i) P_{j \to i}(A_j)}{Z}$$

(8)

where $A_i$ and $A_j$ are the mutations at positions $i$ and $j$, $\lambda_{ij}(A_i, A_j)$ is the statistical coupling parameter between $i$ and $j$, $P_{i \to j}(A_i)$ is the message passed from $i$ to $j$, $P_{j \to i}(A_j)$ is the message passed from $j$ to $i$ and $Z$ is the partition function. This equation has been proven by Yedidia and coworkers to be mathematically equivalent to the Bethe approximation, a mean field model, and is what we apply in our code to approximate the bivariate marginals in our system [9, 12, 14].

**Algorithm in detail.**

1. Initialization: Set all $\lambda_{ij}(A_i, A_j) = 0$ and all $\lambda_i(A_i) = c_i + \ln P_i(A_i)$, where $c_i$ is a normalization constant for the gauge constraints which were described earlier.

2. Update messages using Equation 8 for all pairs of residues iteratively until the belief propagation messages converge.

3. Update bivariate marginals $\mathcal{P}^{\text{bethe}}_{ij}(A_i, A_j)$ using the Bethe approximation (Equation 9).

4. Compare $\mathcal{P}^{\text{bethe}}_{ij}(A_i, A_j)$ to $\mathcal{P}^{\text{obs}}_{ij}(A_i, A_j)$, which is the database derived frequency of a double mutation. If the couplings have converged, then stop. If the couplings have not converged by a desired amount, update $\lambda_{ij}(A_i, A_j)$ as follows

$$\triangle \lambda_{ij}(A_i, A_j) = -\epsilon [(\mathcal{P}^{\text{obs}}_{ij}(A_i, A_j) - \mathcal{P}^{\text{bethe}}_{ij}(A_i, A_j))]$$

(9)

where $\epsilon$ is the gradient descent step size, set to 0.0001, and repeat steps 2, 3 and 4 until the pair probabilities converge.
Associating mutation patterns with protease inhibitors.

A drug-annotated sequence alignment consisting of 38,420 HIV protease isolates of multiple subtypes was downloaded from the Stanford HIV database on April 7th, 2010 [15]. This dataset was used to determine which inhibitors are significantly associated with specific electrostatic mutations patterns. Since multiple isolates in this database are associated with a single patient, only the most recent subtype B isolate for each patient was extracted, leaving 13,286 protease sequences. Upon examination, many sequences come from patients undergoing antiretroviral therapy with one or more protease inhibitors. However, the majority of sequences come from patients who have not been exposed to any drugs. The difference in the proportion of sequences with a particular mutation pattern in the drug-naive cohort as compared to the proportion of sequences with the same mutation pattern but exposed to a specific drug cocktail, can be used as a measure of association between that drug cocktail and the mutation pattern.

To find significant associations, the sequence alignment was converted into strings of charge states “n”, “-” and “+” as described above, and used to calculate \( p_{drug} \), the proportion of sequences with a unique mutation pattern and exposed to a particular set of drugs, and \( p_{naive} \), the proportion of sequences with the same mutation pattern but exposed to no drugs. A pooled sample proportion t-test was then performed to determine the significance of association for a drug combination with a group of mutations, with the z-score for the null hypothesis of \( p_{drug} - p_{naive} = 0 \) given by

\[
z = \frac{p_{drug} - p_{naive}}{SE}.
\]

The standard error, SE is

\[
SE = \sqrt{p(1-p)(1/n_1 + 1/n_2)},
\]

where \( p = (p_{drug}n_{drug} + p_{naive}n_{naive})/(n_{drug} + n_{naive}) \) is the pooled sample proportion, \( n_{drug} \) is the number of sequences exposed to the particular combination of drugs, and \( n_{naive} \) is the number of sequences not exposed to drugs.
References


Figure S1. Contribution of individual sequences to the average electrostatic folding energy. Each contribution is given by $\Delta G_e P$, where $\Delta G_e$ is the electrostatic folding energy of a given sequence (see Methods) and $P$ is its probability under the independent or pair correlation model conditional upon the number of mutations. Red: mutation patterns observed in the Lee database [1], black: mutation patterns not observed in the Lee database. Several outliers are labeled explicitly by their mutation pattern. Mutations are represented as $aNb$, where $N$ is the residue number and $a$ and $b$ are one of the 3 charged states (+, -, n). The straight line on each diagram is a plot of $x = y$. Sequences below this line have $P_1 < P_2$, resulting in $\Delta G_e P_1 > \Delta G_e P_2$ ($\Delta G_e < 0$). For these sequences, the electrostatic stabilization is greater under the pair correlation model than under the independent model.
Figure S2. Comparison of sequence probabilities under the independent and pair correlation model. The probability of a given sequence under the pair correlation model, $P_2$, is plotted against the probability of the same sequence under the independent model, $P_1$, for all sequences with 1 through 6 electrostatic mutations. Both independent and pair correlation model probabilities are renormalized and are conditional upon the number of mutations. Red: mutation patterns observed in the Lee database [1], black: mutation patterns not observed in the Lee database. Several outliers are labeled explicitly by their mutation pattern. Mutations are represented as $aNb$, where $N$ is the residue number and $a$ and $b$ are one of the 3 charged states (+, - , n). The straight line on each diagram is a plot of $x = y$. Sequences below this line have $P_1 < P_2$. 
Figure S3. Distribution of pair correlation model probabilities for sequences in the tail of the Lee distribution that are observed (red) or unobserved (blue) in the Stanford database. The histogram in red is the distribution of pair correlation model probabilities for sequences found in the tail of the Lee database that also exist in the Stanford database. The histogram in blue is the distribution of pair correlation model probabilities for sequences that are not observed in the Stanford database. The null hypothesis which states that the means of these two distributions are equal, has a low p-value of $< 10^{-4}$, indicating that the null hypothesis must be rejected. Therefore, the difference between the means of these two distributions is statistically significant.
Figure S4. Distribution of the number of unique sequences for sample sizes equal to the size of Stanford and Lee databases. 13,286 and 45,161 sequences, corresponding to the sizes of the Stanford and Lee databases, were each randomly sampled from the probability distribution described by the pair correlation model. The distribution of the number of unique sequences for 1,000 simulations for both sampling distributions is plotted as a histogram. The sample average for the Stanford-sized sample distribution is 452.9 and the standard deviation is 14.3. The sample average for the Lee-sized sample distribution is 862.1 and the standard deviation is 18.7. The number of unique sequences in the Stanford database is 431 while the number of unique sequences in the Lee database is 828, both of which lie within 1.4 standard deviations of their respective sample means.
Figure S5. Distribution of the number of sampled sequences not observed in the Lee database. 13,286 sequences, corresponding to the size of the Stanford database, were randomly sampled from the probability distribution described by the pair correlation model. The distribution of the number of sequences not observed in the Lee database for each of the 1,000 simulations, is plotted as a frequency distribution. The sample average for this distribution is 124.2 and the standard deviation is 10.6. The actual number of sequences in the Stanford database that are not observed in the Lee database is 128 (plotted as a straight red line), a number which lies well within 1 standard deviation of the sample mean.
Table S1. Electrostatic mutation patterns with the highest probabilities under the pair correlation model and the drug combinations they are most strongly associated with.

Shown are the top 5 patterns with 2, 3 and 4 electrostatic mutations for which the pair correlation model predicted probability, $P_2$, is the highest, together with the drug combination they are most significantly associated with. Drug combinations are listed in order of treatment. The test of statistical association between drugs and electrostatic mutation patterns is based on the the Stanford database [15] (SI Methods). The proportion of sequences with the mutation pattern and exposed to a specific drug was compared to the proportion of sequences with the same mutation pattern but exposed to no drugs. The null hypothesis is that that the two proportions are equal, and the p-value to test the significance of this hypothesis is listed alongside the drug combination. NFV: Nelfinavir, IDV: Indinavir, SQV: Saquinavir, RTV: Ritonavir, APV: Amprenavir. The acronym PI, protease inhibitor, is used in the Stanford database when the drug was unknown. The 30n,37−,61−,88− pattern is not significantly associated with any drug combination.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>$P_2$</th>
<th>Drugs</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 electrostatic mutations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D30N, N88D</td>
<td>$2.7 \times 10^{-4}$</td>
<td>NFV</td>
<td>&lt; $10^{-7}$</td>
</tr>
<tr>
<td>K20I, N37D</td>
<td>$6.1 \times 10^{-3}$</td>
<td>IDV,NFV</td>
<td>&lt; $10^{-4}$</td>
</tr>
<tr>
<td>N37D, H69Q</td>
<td>$1.6 \times 10^{-4}$</td>
<td>PI</td>
<td>&lt; $10^{-4}$</td>
</tr>
<tr>
<td>N37D, Q61E</td>
<td>$2.9 \times 10^{-4}$</td>
<td>RTV,SQV,PI</td>
<td>&lt; $10^{-4}$</td>
</tr>
<tr>
<td>Q7E, N37D</td>
<td>$2.1 \times 10^{-3}$</td>
<td>RTV,PI</td>
<td>&lt; $10^{-4}$</td>
</tr>
<tr>
<td>3 electrostatic mutations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D30N, N37D, N88D</td>
<td>$4.7 \times 10^{-3}$</td>
<td>IDV,NFV,RTV</td>
<td>&lt; $10^{-4}$</td>
</tr>
<tr>
<td>K20I, D30N, N88D</td>
<td>$3.1 \times 10^{-3}$</td>
<td>IDV,NFV,PI</td>
<td>&lt; $10^{-4}$</td>
</tr>
<tr>
<td>D30N, H69Q, N88D</td>
<td>$2.7 \times 10^{-3}$</td>
<td>IDV,NFV,RTV,SQV</td>
<td>&lt; $10^{-4}$</td>
</tr>
<tr>
<td>D30N, Q61E, N88D</td>
<td>$8.1 \times 10^{-4}$</td>
<td>NFV</td>
<td>&lt; $10^{-7}$</td>
</tr>
<tr>
<td>Q7E, D30N, N88D</td>
<td>$7.4 \times 10^{-3}$</td>
<td>NFV</td>
<td>&lt; $10^{-4}$</td>
</tr>
<tr>
<td>4 electrostatic mutations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K20I, D30N, N37D, N88D</td>
<td>$5.5 \times 10^{-4}$</td>
<td>IDV,NFV</td>
<td>&lt; $10^{-4}$</td>
</tr>
<tr>
<td>D30N, N37D, H69Q, N88D</td>
<td>$3.0 \times 10^{-4}$</td>
<td>APV,IDV,NFV,RTV,SQV</td>
<td>&lt; $10^{-7}$</td>
</tr>
<tr>
<td>K20I, D30N, H69Q, N88D</td>
<td>$2.4 \times 10^{-4}$</td>
<td>NFV,RTV,PI</td>
<td>&lt; $10^{-7}$</td>
</tr>
<tr>
<td>K20I, D30N, E35Q, N88D</td>
<td>$2.2 \times 10^{-4}$</td>
<td>IDV,NFV</td>
<td>&lt; $10^{-7}$</td>
</tr>
<tr>
<td>D30N, N37D, Q61E, N88D</td>
<td>$1.5 \times 10^{-4}$</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
Table S2. Prediction of novel electrostatic mutation patterns. Shown are 25 electrostatic mutation patterns with the highest probabilities under the pair correlation model that are not observed in the Lee database [1]. \( P_2 \) is the probability of the sequence under the pair correlation model, \( N_{LEE} \) is the number of times the mutation pattern was found in the Lee database [1], \( N_{SH} \) is the number of times the mutation pattern was found in the Stanford database [15]. If the sequence is found in the Stanford database, it may be significantly associated with specific drugs combinations. The drug combinations listed are in order of treatment and have strong p-values of association with the mutation pattern. The test of statistical association between drugs and electrostatic mutation patterns is described in SI Methods. NFV: Nelfinavir, IDV: Indinavir, SQV: Saquinavir, RTV: Ritonavir, APV: Amprenavir. The acronym PI, protease inhibitor, is used in the Stanford database when the drug was unknown.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>( P_2 )</th>
<th>N_{LEE}</th>
<th>N_{SH}</th>
<th>Drugs</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>H69Q,I72R</td>
<td>1.8 \times 10^{-4}</td>
<td>0</td>
<td>11</td>
<td>APV-IDV-NFV-RTV</td>
<td>&lt; 10^{-7}</td>
</tr>
<tr>
<td>K20I,N37D,Q58E,Q92K</td>
<td>8.3 \times 10^{-5}</td>
<td>0</td>
<td>5</td>
<td>PI</td>
<td>&lt; 10^{-5}</td>
</tr>
<tr>
<td>K20I,E34Q,Q58E</td>
<td>7.4 \times 10^{-5}</td>
<td>0</td>
<td>16</td>
<td>PI</td>
<td>&lt; 10^{-7}</td>
</tr>
<tr>
<td>K20I,L63H,K70E</td>
<td>6.0 \times 10^{-5}</td>
<td>0</td>
<td>4</td>
<td>ATV</td>
<td>&lt; 10^{-7}</td>
</tr>
<tr>
<td>D30N,H69Q,I72E,N88D</td>
<td>5.3 \times 10^{-5}</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>K20I,D30N,K70E,N88D</td>
<td>5.0 \times 10^{-5}</td>
<td>0</td>
<td>1</td>
<td>PI</td>
<td>&lt; 4.3 \times 10^{-2}</td>
</tr>
<tr>
<td>Q7E,N37D,Q58E</td>
<td>4.6 \times 10^{-5}</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>D30N,I72R,N88D</td>
<td>4.4 \times 10^{-5}</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q18H,K43T</td>
<td>4.4 \times 10^{-5}</td>
<td>0</td>
<td>25</td>
<td>LPV-NFV-SQV</td>
<td>&lt; 10^{-7}</td>
</tr>
<tr>
<td>D30N,L63H,H69Q,N88D</td>
<td>4.2 \times 10^{-5}</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>G16E,I72R</td>
<td>4.1 \times 10^{-5}</td>
<td>0</td>
<td>3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>E34Q,K70E</td>
<td>4.1 \times 10^{-5}</td>
<td>0</td>
<td>4</td>
<td>PI</td>
<td>&lt; 5.0 \times 10^{-5}</td>
</tr>
<tr>
<td>G16E,K20I,N37K</td>
<td>3.8 \times 10^{-5}</td>
<td>0</td>
<td>2</td>
<td>PI</td>
<td>&lt; 4.1 \times 10^{-5}</td>
</tr>
<tr>
<td>Q18H,D30N,N37D,N88D</td>
<td>3.6 \times 10^{-5}</td>
<td>0</td>
<td>2</td>
<td>NFV-PI</td>
<td>&lt; 10^{-5}</td>
</tr>
<tr>
<td>K20I,D30N,E35Q</td>
<td>3.6 \times 10^{-5}</td>
<td>0</td>
<td>6</td>
<td>NFV-RTV</td>
<td>&lt; 10^{-7}</td>
</tr>
<tr>
<td>K20I,K70E,Q92K</td>
<td>3.6 \times 10^{-5}</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>T12K,K70T</td>
<td>3.5 \times 10^{-5}</td>
<td>0</td>
<td>18</td>
<td>NFV-PI</td>
<td>&lt; 7.2 \times 10^{-4}</td>
</tr>
<tr>
<td>N37K,K43T,Q61E</td>
<td>3.4 \times 10^{-5}</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q18H,K20I,N88D</td>
<td>3.4 \times 10^{-5}</td>
<td>0</td>
<td>3</td>
<td>LPV-RTV-SQV</td>
<td>&lt; 10^{-5}</td>
</tr>
<tr>
<td>E35Q,Q58E</td>
<td>3.4 \times 10^{-5}</td>
<td>0</td>
<td>21</td>
<td>IDV-LPV-NFV-PI</td>
<td>&lt; 10^{-5}</td>
</tr>
<tr>
<td>N37D,Q58E,I72E</td>
<td>3.4 \times 10^{-5}</td>
<td>0</td>
<td>3</td>
<td>APV-IDV-NFV-RTV-SQV</td>
<td>&lt; 10^{-7}</td>
</tr>
<tr>
<td>T12K,N37D,H69Q</td>
<td>3.4 \times 10^{-5}</td>
<td>0</td>
<td>5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>D30N,N37D,L63H,N88D</td>
<td>3.4 \times 10^{-5}</td>
<td>0</td>
<td>1</td>
<td>PI</td>
<td>0.04</td>
</tr>
<tr>
<td>G16E,K70E</td>
<td>3.4 \times 10^{-5}</td>
<td>0</td>
<td>6</td>
<td>NFV</td>
<td>&lt; 4.8 \times 10^{-4}</td>
</tr>
<tr>
<td>K20I,D30N,N37D,N88D,Q92K</td>
<td>3.4 \times 10^{-5}</td>
<td>0</td>
<td>1</td>
<td>PI</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Chapter 6

Correlated electrostatic mutations and the structure of HIV protease

6.1 Introduction

One of the major goals of this project is to connect the realm of protein biophysics and energetics with the realm of sequence statistics extracted from multiple sequences in an alignment. In previous chapters we discussed in detail our results that highlight the average electrostatic properties of many sequences. For example we showed that on average, the electrostatic contribution towards the total folding energy gets more stabilizing as the number of electrostatic mutations increases. Additionally we also show that the pair correlation model more closely reflects these observed trends in electrostatic stabilization and that the independent model is not a good approximation. Though links between sequence correlations and energetics within ensembles of sequences are important and highlight important features of the observed distribution of sequences, we wish to examine individual sequences of mutations, calculate their energies, and correlate these to the structure of HIV protease. In order to focus on specific mutations, we need to first understand the structure of HIV protease and its wildtype network of electrostatic interactions. Table 6.1 lists the charged amino acids of wildtype protease. Of the 19 charged residues, 11 are positively charged.
Table 6.1: Charged amino acids within the wildtype sequence of HIV protease subtype B, the amino acid residue type and the mutation frequency for that residue.

<table>
<thead>
<tr>
<th>residue #</th>
<th>wildtype residue</th>
<th>% conserved in the Lee database</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>ARG</td>
<td>100</td>
</tr>
<tr>
<td>14</td>
<td>LYS</td>
<td>100</td>
</tr>
<tr>
<td>20</td>
<td>LYS</td>
<td>93</td>
</tr>
<tr>
<td>21</td>
<td>GLU</td>
<td>100</td>
</tr>
<tr>
<td>25</td>
<td>ASP</td>
<td>100</td>
</tr>
<tr>
<td>29</td>
<td>ASP</td>
<td>100</td>
</tr>
<tr>
<td>30</td>
<td>ASP</td>
<td>93</td>
</tr>
<tr>
<td>34</td>
<td>GLU</td>
<td>99</td>
</tr>
<tr>
<td>35</td>
<td>GLU</td>
<td>99</td>
</tr>
<tr>
<td>41</td>
<td>ARG</td>
<td>100</td>
</tr>
<tr>
<td>43</td>
<td>LYS</td>
<td>98</td>
</tr>
<tr>
<td>45</td>
<td>LYS</td>
<td>100</td>
</tr>
<tr>
<td>55</td>
<td>LYS</td>
<td>100</td>
</tr>
<tr>
<td>57</td>
<td>ARG</td>
<td>100</td>
</tr>
<tr>
<td>60</td>
<td>ASP</td>
<td>100</td>
</tr>
<tr>
<td>65</td>
<td>GLU</td>
<td>100</td>
</tr>
<tr>
<td>69</td>
<td>HIS</td>
<td>93</td>
</tr>
<tr>
<td>70</td>
<td>LYS</td>
<td>99</td>
</tr>
<tr>
<td>87</td>
<td>ARG</td>
<td>100</td>
</tr>
</tbody>
</table>
Table 6.2: The twenty most statistically enhanced double mutations in the Lee database, in terms of the observed number of sequences (not limited to sequences with only two mutations) that contain these mutations relative to the independent model. The measure used to test for enhancement is $dev(i,j) = \frac{(P_{ij}(M,M)-P_i(M)P_j(M))^2}{P_{ij}(M,M)}$ where $P_{ij}(M,M)$ is the joint probability of a double mutation at positions $i$ and $j$ while $P_i(M)$ is the univariate marginal of a mutation at position $i$. The wildtype and double mutant charge states, distance between charges, the free energy of folding relative to the wildtype protein ($\Delta \Delta G$) and the free energy of cooperativity ($\Delta G_{coop}$) are also listed.

<table>
<thead>
<tr>
<th>residues</th>
<th>wildtype</th>
<th>mutations</th>
<th>distance</th>
<th>$\Delta \Delta G$</th>
<th>$\Delta G_{coop}$</th>
<th>$dev(i,j)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>30–88</td>
<td>-, 0</td>
<td>0, -</td>
<td>6</td>
<td>-10</td>
<td>-14.6</td>
<td>1910</td>
</tr>
<tr>
<td>20–35</td>
<td>+, -</td>
<td>0, 0</td>
<td>11</td>
<td>-7</td>
<td>-0.8</td>
<td>126</td>
</tr>
<tr>
<td>16–63</td>
<td>0, 0</td>
<td>-, +</td>
<td>8</td>
<td>-31</td>
<td>-1.4</td>
<td>118</td>
</tr>
<tr>
<td>18–20</td>
<td>0, +</td>
<td>+, 0</td>
<td>5</td>
<td>-4</td>
<td>-2.9</td>
<td>91</td>
</tr>
<tr>
<td>20–92</td>
<td>+, 0</td>
<td>0, +</td>
<td>21</td>
<td>5</td>
<td>-0.6</td>
<td>75</td>
</tr>
<tr>
<td>63–70</td>
<td>0, +</td>
<td>+, -</td>
<td>7</td>
<td>-20</td>
<td>-5.5</td>
<td>56</td>
</tr>
<tr>
<td>20–88</td>
<td>+, 0</td>
<td>0, -</td>
<td>18</td>
<td>41</td>
<td>0.7</td>
<td>53</td>
</tr>
<tr>
<td>20–58</td>
<td>+, 0</td>
<td>0, -</td>
<td>20</td>
<td>18</td>
<td>0.6</td>
<td>53</td>
</tr>
<tr>
<td>63–70</td>
<td>0, +</td>
<td>+, 0</td>
<td>7</td>
<td>-32</td>
<td>-2.8</td>
<td>46</td>
</tr>
<tr>
<td>16–72</td>
<td>0, 0</td>
<td>-, +</td>
<td>13</td>
<td>1</td>
<td>-0.8</td>
<td>46</td>
</tr>
<tr>
<td>35–37</td>
<td>-, 0</td>
<td>0, -</td>
<td>5</td>
<td>3</td>
<td>-1.0</td>
<td>38</td>
</tr>
<tr>
<td>58–92</td>
<td>0, 0</td>
<td>-, +</td>
<td>15</td>
<td>33</td>
<td>-0.7</td>
<td>37</td>
</tr>
<tr>
<td>16–37</td>
<td>0, 0</td>
<td>-, +</td>
<td>9</td>
<td>-2</td>
<td>-1.4</td>
<td>32</td>
</tr>
<tr>
<td>20–30</td>
<td>+, -</td>
<td>0, 0</td>
<td>19</td>
<td>-47</td>
<td>-0.5</td>
<td>30</td>
</tr>
<tr>
<td>20–72</td>
<td>+, 0</td>
<td>0, +</td>
<td>18</td>
<td>-5</td>
<td>-0.6</td>
<td>28</td>
</tr>
<tr>
<td>61–72</td>
<td>0, 0</td>
<td>-, +</td>
<td>8</td>
<td>2</td>
<td>-3.0</td>
<td>27</td>
</tr>
<tr>
<td>37–92</td>
<td>0, 0</td>
<td>-, +</td>
<td>23</td>
<td>16</td>
<td>-0.5</td>
<td>22</td>
</tr>
<tr>
<td>43–61</td>
<td>0, 0</td>
<td>0, -</td>
<td>13</td>
<td>-3</td>
<td>0.7</td>
<td>19</td>
</tr>
<tr>
<td>63–69</td>
<td>+, 0</td>
<td>0, -</td>
<td>14</td>
<td>-7</td>
<td>-0.5</td>
<td>18</td>
</tr>
<tr>
<td>20–43</td>
<td>0, +</td>
<td>+, 0</td>
<td>24</td>
<td>-13</td>
<td>0.5</td>
<td>17</td>
</tr>
</tbody>
</table>
Table 6.3: The twenty most statistically suppressed double mutations in the Lee database, in terms of the observed number of sequences (not limited to sequences with only two mutations) that contain these mutations relative to the independent model. The measure used to test for enhancement is $\text{dev}(i, j) = \frac{(P_{ij}(M, M) - P_i(M)P_j(M))^2}{P_i(M)P_j(M)}$ where $P_{ij}(M, M)$ is the joint probability of a double mutation at positions $i$ and $j$ while $P_i(M)$ is the univariate marginal of a mutation at position $i$. The wildtype and double mutant charge states, distance between charges, the free energy of folding relative to the wildtype protein ($\Delta \Delta G$) and the free energy of cooperativity ($\Delta G_{\text{coop}}$) are also listed.

<table>
<thead>
<tr>
<th>residues</th>
<th>wildtype</th>
<th>mutations</th>
<th>distance</th>
<th>$\Delta \Delta G$</th>
<th>$\Delta G_{\text{coop}}$</th>
<th>$\text{dev}(i, j)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>16–37</td>
<td>0, 0</td>
<td>-, -</td>
<td>9</td>
<td>8</td>
<td>1.4</td>
<td>48</td>
</tr>
<tr>
<td>37–69</td>
<td>0, +</td>
<td>-, 0</td>
<td>24</td>
<td>31</td>
<td>0.5</td>
<td>39</td>
</tr>
<tr>
<td>16–69</td>
<td>0, +</td>
<td>-, 0</td>
<td>17</td>
<td>28</td>
<td>0.6</td>
<td>27</td>
</tr>
<tr>
<td>43–58</td>
<td>+, 0</td>
<td>0, -</td>
<td>5</td>
<td>22</td>
<td>7.7</td>
<td>19</td>
</tr>
<tr>
<td>12–37</td>
<td>0, 0</td>
<td>-, -</td>
<td>19</td>
<td>6</td>
<td>0.5</td>
<td>18</td>
</tr>
<tr>
<td>58–69</td>
<td>0, +</td>
<td>-, 0</td>
<td>23</td>
<td>48</td>
<td>0.4</td>
<td>17</td>
</tr>
<tr>
<td>61–63</td>
<td>0, 0</td>
<td>-, +</td>
<td>8</td>
<td>-27</td>
<td>0.6</td>
<td>16</td>
</tr>
<tr>
<td>16–88</td>
<td>0, 0</td>
<td>-, -</td>
<td>18</td>
<td>48</td>
<td>0.7</td>
<td>13</td>
</tr>
<tr>
<td>37–70</td>
<td>0, +</td>
<td>-, -</td>
<td>19</td>
<td>23</td>
<td>1.0</td>
<td>13</td>
</tr>
<tr>
<td>61–72</td>
<td>0, 0</td>
<td>-, -</td>
<td>8</td>
<td>20</td>
<td>3.0</td>
<td>11</td>
</tr>
<tr>
<td>18–20</td>
<td>0, +</td>
<td>-, -</td>
<td>18</td>
<td>48</td>
<td>0.7</td>
<td>9</td>
</tr>
<tr>
<td>20–69</td>
<td>+, +</td>
<td>0, 0</td>
<td>18</td>
<td>21</td>
<td>0.6</td>
<td>9</td>
</tr>
<tr>
<td>37–69</td>
<td>0, +</td>
<td>+, 0</td>
<td>24</td>
<td>23</td>
<td>-0.5</td>
<td>9</td>
</tr>
<tr>
<td>16–72</td>
<td>0, 0</td>
<td>-, -</td>
<td>13</td>
<td>15</td>
<td>0.8</td>
<td>8</td>
</tr>
<tr>
<td>18–37</td>
<td>0, 0</td>
<td>-, -</td>
<td>7</td>
<td>23</td>
<td>1.7</td>
<td>8</td>
</tr>
<tr>
<td>63–72</td>
<td>0, 0</td>
<td>+, -</td>
<td>5</td>
<td>-25</td>
<td>-6.5</td>
<td>8</td>
</tr>
<tr>
<td>61–69</td>
<td>0, +</td>
<td>-, 0</td>
<td>20</td>
<td>30</td>
<td>0.4</td>
<td>7</td>
</tr>
<tr>
<td>35–69</td>
<td>-, +</td>
<td>0, 0</td>
<td>25</td>
<td>24</td>
<td>-0.4</td>
<td>6</td>
</tr>
<tr>
<td>37–63</td>
<td>0, 0</td>
<td>-, +</td>
<td>14</td>
<td>-27</td>
<td>-0.6</td>
<td>6</td>
</tr>
<tr>
<td>20–72</td>
<td>+, 0</td>
<td>0, -</td>
<td>18</td>
<td>8</td>
<td>0.6</td>
<td>6</td>
</tr>
</tbody>
</table>
After examining the wildtype electrostatic network, we will focus on correlated electrostatic mutations that are statistically enhanced or suppressed. There are \( \binom{N}{2} (Q - 1)^2 = \binom{18}{2} 2^2 = 612 \) pairs of unique double mutants in our database where the number of positions, \( N = 18 \), and \( Q = 3 \) possible mutants per site. Our goal is to understand the energetic basis behind why these pairs mutations are statistically enhanced or suppressed in the sequence database. Moreover, it is curious to see that some of these mutations are distal from one another on the protein, yet have a strong correlation signal. Another goal of ours is to understand long-range correlations. To start, Tables 6.2 and 6.3 list the sequence counts for the top 20 most enhanced and suppressed double mutations relative to the independent model in the Lee database, our of the 612 total double mutations. We observed that several of these pairs of mutations are more than 15\( \text{Å} \) apart, yet the electrostatic folding free energies of only the most enhanced or suppressed double mutants are consistent with the sequence statistics; the top 4 enhanced double mutants have negative \( \Delta \Delta G \)'s which lead to more stable proteins while the top 6 suppressed double mutations have positive \( \Delta \Delta G \)'s and destabilized. This is a remarkable yet highly non-trivial result given the coarse-grained nature of our electrostatic energy model. However this trend is limited to only the topmost positively or negatively correlated double mutants. If we examine the top 20 positively correlated residues, only 12 have negative \( \Delta \Delta G \)'s. In contrast, 17 of the top 20 suppressed double mutants have positive \( \Delta \Delta G \)'s suggesting that there is perhaps a weak signal. Moreover, many of these highly correlated pairs are more than 10 Angstroms apart and we wish to understand this observation in more detail by studying the structure of HIV protease and examining how local interactions influence long-range correlations between distal mutations.

In order to do so, it is therefore first necessary to map out the electrostatic network of wildtype protease and highlight the important local electrostatic
interactions between the 19 charged wildtype amino acids of HIV protease. We will especially focus on interactions involving conserved residues since these are presumably highly stabilizing to the protein. To do this analysis we use the wildtype structure of HIV protease from the PDB file 1HN0. This PDB structure was previously used in our electrostatic correlation analysis of Chapter 5.

6.2 The electrostatic interaction network of wildtype HIV protease

Wildtype HIV protease subtype B is a dimer where each chain consists of 99 amino acids. Of these 99 residues, 19 are charged residues, 11 of which are positively charged (arginine, lysine or histidine). Table 6.1 lists the charged amino acids for wildtype HIV protease with their mutation frequency rate within the Lee database. The conservation percentages show that of the 19 charged residues, 12 are completely conserved in the Lee database, i.e., these charged residues never mutate. The remaining 7 residues have low mutation frequencies ranging from 1% to 7%.

Our goal in this section is to map out the wildtype electrostatic interactions of HIV protease. For that we require a certain threshold distance to define an electrostatic interaction. This is not a totally trivial task because in solution, modeled as a continuum dielectric medium, the decay of the Coulomb interaction remains \( \frac{1}{r} \) but is scaled by a factor of 80, the dielectric constant of water. In ionic solutions according to the Debye-Huckel theory, the decay of electrostatic interactions with distance is faster than in pure water. In a practical sense it is better to ask, at what distance does a charge-charge interaction become negligible compared to the local interactions around the charges? In this analysis distances of less than 6.5Å are considered to be close contacts.
Several of the 19 charged amino acids of wildtype HIV protease lie within 6.5Å of each other forming stabilizing salt bridges. As a result, these 19 residues can be clustered into 6 distinct clusters of interacting residues. Three clusters involve two residues, two clusters involve three residues, while one cluster involves 4 charged residues. 4 of the 19 residues do not lie within 6.5Å of any other charged residue. The specific patterns of interactions among these groups of charged residues are described below.

### 6.2.1 R8-D29-D30, D29-R87

Using 6.5Å as the distance threshold, there is a inter-chain interaction between R8 of chain A and D29 of chain B and vice versa (Figure 6.1). This interaction is probably an important dimer stabilizing salt bridge. D29 on either chain is also within 6.5Å of D30 and R87 (Figure 6.2). D29-D30 is a destabilizing interaction. The presence of this destabilizing like-charge interaction explains why D30N is a stabilizing single mutation (ΔΔG ~ -41 kcal/mol).
Figure 6.2: Two interactions are depicted in this figure. There is a salt bridges between R8 of chain A and D29 of chain B. There is also a close interaction between D29 and R87 of the same chain. Though not depicted in this picture, D29 is also within 6.5Å of D30. D29 is within 6.5Å of three charged residues.

6.2.2 K14-E65-K70

E65 is involved with two favorable electrostatic interactions within 6.5Å, with K14 and with K70 (Figure 6.3). K14 and E65 are conserved residues, but K70 mutates from the positively charged wildtype lysine to other (negative) polar or non-polar residues in 1% of the sequences in the Lee database.

6.2.3 K45-K43-D60

E43 is involved in one favorable and one unfavorable electrostatic interaction within 6.5Å, with K45 and with D60 (Figure 6.4). K45 and D60 are conserved residues, but K43 mutates from the positively charged wildtype lysine to other (negative) polar or non-polar residues in 2% of the sequences in the Lee database.
Figure 6.3: Two salt bridge interactions are depicted in this figure, both of which involve E65. One is between E65 and K14 and the other is between E65 and K70.

Figure 6.4: Two electrostatic interactions are depicted in this figure, both of which involve K43. One is a destabilizing interaction between K43 and K45 and the other is a stabilizing interaction between K43 and D60.
6.2.4 K20-E34

K20 has a favorable electrostatic interaction with E34 (Figure 6.5). Both residues are not completely conserved in the Lee database suggesting that gain or loss of these charges plays a role in protein stability.

6.2.5 E35-R57

E35 has a favorable electrostatic interaction with R57. The arginine on residue 57 is completely conserved but the glutamate on residue 35 mutates in the Lee database.

6.2.6 D25 (chain A) - D25 (chain B)

In the active site, the aspartate on position 25 in chain A interacts with its corresponding aspartate on chain B position 25 (Figure 6.7). The interaction is unfavorable. However this is a binding site residue and both residues are
completely conserved, indicating that the presence of these residues is crucial for enzymatic function.

6.3 Energetics of short-range correlated double mutations

Table 6.2 and Table 6.3 show that several highly correlated double mutations are within a few angstroms of each other.

6.3.1 D30N-N88D

D30N-N88D is the most statistically enhanced double mutation in the Lee database relative to the independent model. The positions are ~6Å apart. Moreover, the electrostatic folding energy of the double mutant is consistent with the statistics since it is also a stabilizing double mutation. If we examine the energetics in more detail (Table 6.4), we note that the single mutants D30N and N88D cause very different effects. The D30N single mutant is very stabilizing but the N88D single mutant is highly destabilizing. One must look
Figure 6.7: A destabilizing interaction between D25 of chain A and D25 of chain B at the structure of HIV protease, especially the local interactions around these mutations to rationalize this results. D30N is stabilizing because of the local network of electrostatic interactions around residue 30. In wildtype protease, D30 is very close to D29 (<6Å, See Figure 6.2). As a result, once the aspartate on residue 30 becomes a neutral residue, this unfavorable interaction gets removed. As a result, D30N becomes a highly favorable mutation. In contrast, the reason why N88D is unfavorable is because of the presence of D30 so close to D88. This is a highly unfavorable like-charge interaction which leads to poor stability. Along the same line of reasoning, the double mutant relieves this unfavorable interaction.

6.3.2 G16E-L63H

Table 6.2 shows that G16E-L63H is the third most statistically enhanced double mutant in the Lee database. The wildtype state for both positions 16 and 63 is electrostatically neutral. The double mutant consists of like charges at a distance of ~8Å. The electrostatic folding free energy of this double mutation is
-120 kcal/mol indicating that the double mutant is stabilized compared to the wildtype. This folding free energy is consistent with sequence statistics as the double mutant is enhanced relative to the independent model. To further understand why the double mutant is stabilizing, we must examine the energies of the single mutants and the local electrostatic structure around 16 and 63. Table 6.5 shows that the L63H single mutant is highly stabilizing. This stabilization is probably due to the presence of the negatively E65 at a distance of 5.2Å. G16E, on the other hand is only slightly unstable. This suggests that there are very few local effects around positions 16. Indeed, within ~6.5Å of residue 16, there are no charged residues. The stabilization of the double mutant is almost entirely due to the L63H single mutant, with some contribution from the favorable energetic coupling between the mutations E16 and H63 themselves.

### 6.4 Energetics of long-range correlated double mutations

#### 6.4.1 G16E-N37D

Table 6.3 shows that G16E-N37D is the most statistically suppressed double mutant relative to the independent model in the Lee database. The wildtype states for both position 16 and 37 are electrostatically neutral. This double mutant consists of like charges which are ~9Å apart. The electrostatic folding free energy of this double mutation is -81 kcal/mol which means the double mutant is less stable than wildtype protease and that this energy is consistent with the idea that this destabilizing mutation should be suppressed relative to the independent model predictions. Table 6.6 shows that both single mutants are slightly less stable than the wildtype, though N37D is the more unstable. This suggests that there are very few local effects around positions 16 in particular. Indeed, within ~6.5Å of residue 16, there are no charged residues. On the other hand, E35 is 5.5Å away from the negative charge of D37. As a result, the
Table 6.4: Observed and independent bivariate marginals for the single and double mutants of positions 30 and 88 along with their charge states, electrostatic folding energies. The positions are ~6Å apart. $\Delta G_e$ is the electrostatic folding free energy. $\Delta \Delta G_e = \Delta G_e - \Delta G_{e^{\text{wildtype}}}$ where $\Delta G_{e^{\text{wildtype}}} \approx -89$

<table>
<thead>
<tr>
<th>mutations</th>
<th>observed counts</th>
<th>independent counts</th>
<th>charges</th>
<th>$\Delta G_e$</th>
<th>$\Delta \Delta G_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D30N-N88D</td>
<td>2220</td>
<td>161</td>
<td>0, -</td>
<td>-99</td>
<td>-10</td>
</tr>
<tr>
<td>N88D</td>
<td>248</td>
<td>2307</td>
<td>-, -</td>
<td>-43</td>
<td>+46</td>
</tr>
<tr>
<td>D30N</td>
<td>726</td>
<td>2785</td>
<td>0, 0</td>
<td>-130</td>
<td>-41</td>
</tr>
<tr>
<td>-</td>
<td>41965</td>
<td>39906</td>
<td>-, 0</td>
<td>-89</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6.5: Observed and independent bivariate marginals for the single and double mutants of positions 16 and 63 along with their charge states, electrostatic folding energies. The positions are ~8Å apart. $\Delta G_e$ is the electrostatic folding free energy. $\Delta \Delta G_e = \Delta G_e - \Delta G_{e^{\text{wildtype}}}$ where $\Delta G_{e^{\text{wildtype}}} \approx -89$

<table>
<thead>
<tr>
<th>mutations</th>
<th>observed counts</th>
<th>independent counts</th>
<th>charges</th>
<th>$\Delta G_e$</th>
<th>$\Delta \Delta G_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>G16E-L63H</td>
<td>159</td>
<td>22</td>
<td>-, +</td>
<td>-120</td>
<td>-31</td>
</tr>
<tr>
<td>G16E</td>
<td>1153</td>
<td>1289</td>
<td>-, 0</td>
<td>-87</td>
<td>+2</td>
</tr>
<tr>
<td>L63H</td>
<td>601</td>
<td>738</td>
<td>0, +</td>
<td>-120</td>
<td>-32</td>
</tr>
<tr>
<td>-</td>
<td>43149</td>
<td>43013</td>
<td>0, 0</td>
<td>-89</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6.6: Observed and independent bivariate marginals for the single and double mutants of positions 16 and 37 along with their charge states, electrostatic folding energies. The positions are ~9Å apart. $\Delta G_e$ is the electrostatic folding free energy. $\Delta \Delta G_e = \Delta G_e - \Delta G_{e^{\text{wildtype}}}$ where $\Delta G_{e^{\text{wildtype}}} \approx -89$

<table>
<thead>
<tr>
<th>mutations</th>
<th>observed counts</th>
<th>independent counts</th>
<th>charges</th>
<th>$\Delta G_e$</th>
<th>$\Delta \Delta G_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>G16E-N37D</td>
<td>92</td>
<td>187</td>
<td>-, -</td>
<td>-81</td>
<td>+8</td>
</tr>
<tr>
<td>G16E</td>
<td>1167</td>
<td>1114</td>
<td>-, 0</td>
<td>-87</td>
<td>+2</td>
</tr>
<tr>
<td>N37D</td>
<td>6342</td>
<td>6246</td>
<td>0, -</td>
<td>-84</td>
<td>+5</td>
</tr>
<tr>
<td>-</td>
<td>37102</td>
<td>37156</td>
<td>0, 0</td>
<td>-89</td>
<td>0</td>
</tr>
</tbody>
</table>
the N37D mutation is destabilizing mutation due to the presence of the nearby glutamate residue. It is interesting to note the $\Delta G_e$ of the G16E-N37D double mutant is almost equal to the sums of the individual single mutant $\Delta G_e$’s. The remaining difference is probably due to the interaction between E16 and D16. Since this biophysical interaction energy is quite small, the statistical correlation seen in the database cannot entirely be due to direct interaction, and as a result, there is probably something else more complicated happening.

6.4.2 K20I-E35Q

Table 6.2 shows that K20I-E35Q is the second most statistically enhanced double mutant in the Lee database. The wildtype state for position 20 is electrostatically positive whereas position 35 is negatively charged. The double mutant consists of two neutral charges at a distance of ~11Å. The electrostatic folding free energy of this double mutation is -96 kcal/mol which means the double mutant is stable compared to the wildtype folding energy. This energy is consistent with sequence statistics as the double mutant is enhanced relative to the independent model. However it is not clear from the charge patterns themselves as to why this energy is so destabilizing. Moreover, this pattern is very interesting as it mutates from unlike charges to neutral charges. To understand why the double mutant is stabilizing, we must examine the energies of the single mutants and the local electrostatic structure around 20 and 35. Table 6.7 shows that both single mutants and the double mutant are more stable than the wildtype, though the K20I single mutation is slightly more stable than the E35Q single mutation. The energies suggest that the loss of a positive charge at position 20 is energetically favorable. However, within 6.5Å of position 20, the only charged residue is the negatively charged E34. The breakage of this favorable interaction will not lead to lower energies. However E34 is also <9Å away from position 35. When the aspartate at 35 mutates to a neutral
residue, the folding energy should become slightly more favorable given that
the like-charge interaction between 34 and 35 has been removed.

6.4.3 N37D-H69Q

Table 6.2 shows that N37D-H69Q is the second most statistically suppressed
double mutant in the Lee database. The wildtype state for position 37 is electro-
statically neutral whereas position 69 is positively charged. The double mutant
consists of a negative charge at position 37 and a neutral charge at position 69
at a distance of ~24Å. The electrostatic folding free energy of this double mu-
tation is -58 kcal/mol which means the double mutant is very unstable. Once
again, this energy is consistent with sequence statistics as the double mutant
is suppressed relative to the independent model. However it is not clear from
the charge patterns themselves as to why this energy is so destabilizing. For
that we must examine the energies of the single mutants and the local electro-
static structure around 37 and 69. Table 6.8 shows that both single mutants
and the double mutant are less stable than the wildtype, though N37D is the
less unstable than H69Q. We have already discussed N37D above. There is a
negatively charged residue 5.5Å away from the negative charge of position 37.
H69Q is little more interesting. It is highly destabilizing with a $\Delta \Delta G_e$ of +26
kcal/mol. If we examine the local electrostatic network around position 69,
we find no charged residues within 6.5Å. However ASP 65 is 7.4Å away, which
could explain the stability decrease. Regardless of the exact mechanism, it is
interesting to note the $\Delta \Delta G_e$ of the G16E-N37D double mutant is equal to the
sums of the individual single mutant $\Delta \Delta G_e$’s. The mutations are 24Å apart so
they essentially do not interact explaining why the single mutant $\Delta \Delta G_e$’s sum
up to the double mutant $\Delta \Delta G_e$. Again, the fact that $\Delta \Delta \Delta G_e \sim 0$ kcal/mol
indicates that the reason behind the statistical suppression of N37D-H69Q is
not due to the energetic coupling between the residues.
6.5 Electrostatic free energy of cooperativity and sequence correlations

The folding energies (relative to the wildtype state, $\Delta\Delta G$) for each of the top 40 positively and negatively correlated double mutants are not entirely consistent with the statistics of enhancement and suppression (Tables 6.2 and 6.3). Apart from the most correlated pairs, the majority of pairs with strong deviations from the independent model have $\Delta\Delta G$’s with signs that are inconsistent with the statistics. However, if we calculate the $\Delta\Delta\Delta G$’s (known as the free energy of cooperativity, $\Delta G_{coop}$, from now on) for each pair, we observe highly significant correlation between the database-derived statistics of coevolution and the sign of the $\Delta G_{coop}$. For example, of the top 20 most positively correlated double mutants, 17 have favorable $\Delta G_{coop}$’s ($\Delta G_{coop} < 0$) while 16 out of the top 20 most negatively correlated double mutants have unfavorable $\Delta G_{coop}$’s ($\Delta G_{coop} > 0$, Tables 6.2 and 6.3). If we expand this list to the top 60 double mutants regardless of whether the pair is enhanced or suppressed, we observed that 50 out the 60 have $\Delta G_{coop}$’s whose sign is consistent with the sign of the observed deviation (Table 6.9). This observation is highly statistically significant (p-value $\sim 10^{-8}$, please see the next section on how this p-value was calculated) and we believe it has a convincing physical interpretation and we will justify its usage. We should keep in mind that we are only examining the top $\sim 10\%$ of all 612 double mutants. The following subsections discuss this result in further detail and provide justification for comparing the sign of the $\Delta G_{coop}$ for a pair of positions to the observed trends in the statistics of the bivariate marginals for the most deviated double mutations.
6.5.1 Determining the p-value

The p-value was calculated using a cumulative binomial probability distribution. Having no prior information, it is fair to assume that the sign of $\Delta G_{coop}$ can be positive or negative with equal probability, that is the choice is a binomial decision of equal probability. Therefore the chance of success of one trial, $p = 0.5$. In our data, however, we observe that the sign of $\Delta G_{coop}$ is consistent with the sign of the correlation, 50 out of 60 times. Therefore the cumulative probability of this event occurring at least 50 or more times is the sum of binomial probabilities from 50 till 60:

$$
\sum_{k=50}^{n=60} \binom{n}{k} p^k (1-p)^{n-k} = 8.08 \times 10^{-8}
$$

This is the probability of seeing 50 or more correct predictions through random chance from a sample size of 60 is $< 10^{-7}$.

6.5.2 Calculating $\Delta G_{coop}$, the electrostatic free energy of cooperativity

Double mutant cycles involve a wildtype protein, two single mutants and the double mutant and are used experimentally and computationally to understand the energetic coupling between residues. If the change in electrostatic free energy of the double mutant differs from the sum of the changes in free energy due to the single mutations, then the residues are energetically coupled. $\Delta G_{coop}$ is the measure of energetic coupling. This coupling free energy, $\Delta G_{coop} = \Delta G_{DM}^{\text{elec}} - \Delta G_{SM_1}^{\text{elec}} - \Delta G_{SM_2}^{\text{elec}} + \Delta G_{WT}^{\text{elec}}$, where $\Delta G_{DM}^{\text{elec}}$, $\Delta G_{SM_1}^{\text{elec}}$ and $\Delta G_{WT}^{\text{elec}}$ are the electrostatic free energies of folding for the double mutant, the single mutant and the wildtype protein respectively. $\Delta G_{coop}$ is negative if the coupling between the residues is favorable and positive if the coupling is unfavorable. In our coarse-grained electrostatics model, mutations at residues that are located
Table 6.7: Observed and independent bivariate marginals for the single and double mutants of positions 20 and 35 along with their charge states, electrostatic folding energies. The positions are ~11Å apart. \( \Delta G_e \) is the electrostatic folding free energy. \( \Delta \Delta G_e = \Delta G_e - \Delta G_{e\text{wildtype}} \)
where \( \Delta G_{e\text{wildtype}} \approx -89 \)

<table>
<thead>
<tr>
<th>mutations</th>
<th>observed counts</th>
<th>independent counts</th>
<th>charges</th>
<th>( \Delta G_e )</th>
<th>( \Delta \Delta G_e )</th>
</tr>
</thead>
<tbody>
<tr>
<td>K20I-E35Q</td>
<td>162</td>
<td>19</td>
<td>0, 0</td>
<td>-96</td>
<td>-7</td>
</tr>
<tr>
<td>K20I</td>
<td>3155</td>
<td>3299</td>
<td>0, -</td>
<td>-94</td>
<td>-5</td>
</tr>
<tr>
<td>E35Q</td>
<td>100</td>
<td>243</td>
<td>+, 0</td>
<td>-90</td>
<td>-1</td>
</tr>
<tr>
<td>-</td>
<td>41736</td>
<td>41592</td>
<td>+, -</td>
<td>-89</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6.8: Observed and independent bivariate marginals for the single and double mutants of positions 37 and 69 along with their charge states, electrostatic folding energies. The positions are ~24Å apart. \( \Delta G_e \) is the electrostatic folding free energy. \( \Delta \Delta G_e = \Delta G_e - \Delta G_{e\text{wildtype}} \)
where \( \Delta G_{e\text{wildtype}} \approx -89 \)

<table>
<thead>
<tr>
<th>mutations</th>
<th>observed counts</th>
<th>independent counts</th>
<th>charges</th>
<th>( \Delta G_e )</th>
<th>( \Delta \Delta G_e )</th>
</tr>
</thead>
<tbody>
<tr>
<td>N37D-H69Q</td>
<td>304</td>
<td>434</td>
<td>- , 0</td>
<td>-58</td>
<td>+31</td>
</tr>
<tr>
<td>N37D</td>
<td>6130</td>
<td>6000</td>
<td>- , +</td>
<td>-84</td>
<td>+5</td>
</tr>
<tr>
<td>H69Q</td>
<td>2726</td>
<td>2580</td>
<td>0, 0</td>
<td>-63</td>
<td>+26</td>
</tr>
<tr>
<td>-</td>
<td>35549</td>
<td>35695</td>
<td>0, +</td>
<td>-89</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 6.8: A plot of distance between pairs of positions as a function of the electrostatic free energy of cooperativity between the charges.
Table 6.9: The 60 most statistically deviated double mutations in the Lee database relative to the independent model. The measure used to test for deviation is \( \text{dev}(i,j) = \frac{(P_{ij}(M,M) - P_i(M)P_j(M))^2}{P_{ij}(M,M)} \) where \( P_{ij}(M,M) \) is the joint probability of a double mutation at positions \( i \) and \( j \) while \( P_i(M) \) is the univariate marginal of a mutation at position \( i \). The double mutant charge states, distance between charges, the free energy of cooperativity (\( \Delta G_{coop} \)) are also listed.

<table>
<thead>
<tr>
<th>#</th>
<th>residues</th>
<th>charges</th>
<th>distance</th>
<th>enhanced or suppressed</th>
<th>( \Delta G_{coop} )</th>
<th>( \text{dev}(i,j) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30–88</td>
<td>0, -</td>
<td>6</td>
<td>enhanced</td>
<td>-14.6</td>
<td>1910</td>
</tr>
<tr>
<td>2</td>
<td>20–35</td>
<td>0, 0</td>
<td>11</td>
<td>enhanced</td>
<td>-0.8</td>
<td>126</td>
</tr>
<tr>
<td>3</td>
<td>16–63</td>
<td>-, +</td>
<td>8</td>
<td>enhanced</td>
<td>-1.4</td>
<td>118</td>
</tr>
<tr>
<td>4</td>
<td>18–20</td>
<td>+, 0</td>
<td>5</td>
<td>enhanced</td>
<td>-2.9</td>
<td>91</td>
</tr>
<tr>
<td>5</td>
<td>20–92</td>
<td>0, +</td>
<td>21</td>
<td>enhanced</td>
<td>-0.6</td>
<td>75</td>
</tr>
<tr>
<td>6</td>
<td>63–70</td>
<td>+, -</td>
<td>7</td>
<td>enhanced</td>
<td>-5.5</td>
<td>56</td>
</tr>
<tr>
<td>7</td>
<td>20–88</td>
<td>0, -</td>
<td>18</td>
<td>enhanced</td>
<td>0.7</td>
<td>53</td>
</tr>
<tr>
<td>8</td>
<td>20–58</td>
<td>0, -</td>
<td>20</td>
<td>enhanced</td>
<td>0.6</td>
<td>53</td>
</tr>
<tr>
<td>9</td>
<td>16–37</td>
<td>-, -</td>
<td>9</td>
<td>suppressed</td>
<td>1.4</td>
<td>48</td>
</tr>
<tr>
<td>10</td>
<td>63–70</td>
<td>+, 0</td>
<td>7</td>
<td>enhanced</td>
<td>-2.8</td>
<td>46</td>
</tr>
<tr>
<td>11</td>
<td>16–72</td>
<td>-, +</td>
<td>13</td>
<td>enhanced</td>
<td>-0.8</td>
<td>46</td>
</tr>
<tr>
<td>12</td>
<td>37–69</td>
<td>-, 0</td>
<td>24</td>
<td>suppressed</td>
<td>0.5</td>
<td>39</td>
</tr>
<tr>
<td>13</td>
<td>35–37</td>
<td>0, -</td>
<td>5</td>
<td>enhanced</td>
<td>-1.0</td>
<td>38</td>
</tr>
<tr>
<td>14</td>
<td>58–92</td>
<td>-, +</td>
<td>15</td>
<td>enhanced</td>
<td>-0.7</td>
<td>37</td>
</tr>
<tr>
<td>15</td>
<td>16–37</td>
<td>-, +</td>
<td>9</td>
<td>enhanced</td>
<td>-1.4</td>
<td>32</td>
</tr>
<tr>
<td>16</td>
<td>20–30</td>
<td>0, 0</td>
<td>19</td>
<td>enhanced</td>
<td>-0.5</td>
<td>30</td>
</tr>
<tr>
<td>17</td>
<td>20–72</td>
<td>0, +</td>
<td>18</td>
<td>enhanced</td>
<td>-0.6</td>
<td>28</td>
</tr>
<tr>
<td>18</td>
<td>16–69</td>
<td>-, 0</td>
<td>17</td>
<td>suppressed</td>
<td>0.6</td>
<td>27</td>
</tr>
<tr>
<td>19</td>
<td>61–72</td>
<td>-, +</td>
<td>8</td>
<td>enhanced</td>
<td>-3.0</td>
<td>27</td>
</tr>
<tr>
<td>20</td>
<td>37–92</td>
<td>-, +</td>
<td>23</td>
<td>enhanced</td>
<td>-0.5</td>
<td>22</td>
</tr>
<tr>
<td>21</td>
<td>43–58</td>
<td>0, -</td>
<td>5</td>
<td>suppressed</td>
<td>7.7</td>
<td>19</td>
</tr>
<tr>
<td>22</td>
<td>43–61</td>
<td>0, -</td>
<td>13</td>
<td>enhanced</td>
<td>0.7</td>
<td>19</td>
</tr>
<tr>
<td>23</td>
<td>12–37</td>
<td>-, -</td>
<td>19</td>
<td>suppressed</td>
<td>0.5</td>
<td>18</td>
</tr>
<tr>
<td>24</td>
<td>63–69</td>
<td>0, -</td>
<td>14</td>
<td>enhanced</td>
<td>-0.5</td>
<td>18</td>
</tr>
<tr>
<td>25</td>
<td>20–43</td>
<td>+, 0</td>
<td>24</td>
<td>enhanced</td>
<td>0.5</td>
<td>17</td>
</tr>
<tr>
<td>26</td>
<td>20–34</td>
<td>0, 0</td>
<td>4</td>
<td>enhanced</td>
<td>-11</td>
<td>17</td>
</tr>
<tr>
<td>27</td>
<td>20–70</td>
<td>0, -</td>
<td>17</td>
<td>enhanced</td>
<td>1.2</td>
<td>17</td>
</tr>
<tr>
<td>28</td>
<td>58–69</td>
<td>-, 0</td>
<td>23</td>
<td>suppressed</td>
<td>0.4</td>
<td>17</td>
</tr>
<tr>
<td>29</td>
<td>61–63</td>
<td>-, +</td>
<td>8</td>
<td>suppressed</td>
<td>0.6</td>
<td>16</td>
</tr>
<tr>
<td>30</td>
<td>18–37</td>
<td>+, -</td>
<td>6</td>
<td>enhanced</td>
<td>-1.7</td>
<td>16</td>
</tr>
<tr>
<td>#</td>
<td>residues</td>
<td>charges</td>
<td>distance</td>
<td>enhanced or suppressed</td>
<td>$\Delta G_{\text{coop}}$</td>
<td>$dev(i,j)$</td>
</tr>
<tr>
<td>----</td>
<td>----------</td>
<td>---------</td>
<td>----------</td>
<td>------------------------</td>
<td>-------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>31</td>
<td>30–35</td>
<td>0, 0</td>
<td>17</td>
<td>enhanced</td>
<td>0.3</td>
<td>16</td>
</tr>
<tr>
<td>32</td>
<td>34–58</td>
<td>0, -</td>
<td>19</td>
<td>enhanced</td>
<td>-0.6</td>
<td>15</td>
</tr>
<tr>
<td>33</td>
<td>63–72</td>
<td>-, +</td>
<td>5</td>
<td>enhanced</td>
<td>-6.5</td>
<td>14</td>
</tr>
<tr>
<td>34</td>
<td>35–88</td>
<td>0, -</td>
<td>18</td>
<td>enhanced</td>
<td>-0.7</td>
<td>14</td>
</tr>
<tr>
<td>35</td>
<td>69–70</td>
<td>0, -</td>
<td>9</td>
<td>enhanced</td>
<td>-0.4</td>
<td>14</td>
</tr>
<tr>
<td>36</td>
<td>16–88</td>
<td>-, -</td>
<td>19</td>
<td>suppressed</td>
<td>0.7</td>
<td>13</td>
</tr>
<tr>
<td>37</td>
<td>37–70</td>
<td>-, -</td>
<td>18</td>
<td>suppressed</td>
<td>1.0</td>
<td>13</td>
</tr>
<tr>
<td>38</td>
<td>61–72</td>
<td>+, -</td>
<td>8</td>
<td>enhanced</td>
<td>-3.0</td>
<td>13</td>
</tr>
<tr>
<td>39</td>
<td>20–34</td>
<td>0, +</td>
<td>4</td>
<td>enhanced</td>
<td>-21.0</td>
<td>13</td>
</tr>
<tr>
<td>40</td>
<td>12–20</td>
<td>+, 0</td>
<td>10</td>
<td>enhanced</td>
<td>-0.9</td>
<td>11</td>
</tr>
<tr>
<td>41</td>
<td>43–92</td>
<td>0, +</td>
<td>19</td>
<td>enhanced</td>
<td>-0.6</td>
<td>11</td>
</tr>
<tr>
<td>42</td>
<td>69–72</td>
<td>0, +</td>
<td>11</td>
<td>enhanced</td>
<td>-1.0</td>
<td>11</td>
</tr>
<tr>
<td>43</td>
<td>12–70</td>
<td>+, -</td>
<td>13</td>
<td>enhanced</td>
<td>-1.5</td>
<td>11</td>
</tr>
<tr>
<td>44</td>
<td>61–72</td>
<td>-, -</td>
<td>8</td>
<td>suppressed</td>
<td>3.0</td>
<td>11</td>
</tr>
<tr>
<td>45</td>
<td>43–72</td>
<td>0, +</td>
<td>18</td>
<td>enhanced</td>
<td>-0.6</td>
<td>10</td>
</tr>
<tr>
<td>46</td>
<td>58–63</td>
<td>-, -</td>
<td>16</td>
<td>enhanced</td>
<td>0.3</td>
<td>10</td>
</tr>
<tr>
<td>47</td>
<td>20–37</td>
<td>0, -</td>
<td>11</td>
<td>enhanced</td>
<td>0.9</td>
<td>10</td>
</tr>
<tr>
<td>48</td>
<td>18–20</td>
<td>-, 0</td>
<td>5</td>
<td>suppressed</td>
<td>2.9</td>
<td>9</td>
</tr>
<tr>
<td>49</td>
<td>70–92</td>
<td>-, +</td>
<td>12</td>
<td>enhanced</td>
<td>-1.9</td>
<td>9</td>
</tr>
<tr>
<td>50</td>
<td>20–69</td>
<td>0, 0</td>
<td>18</td>
<td>suppressed</td>
<td>0.6</td>
<td>9</td>
</tr>
<tr>
<td>51</td>
<td>37–69</td>
<td>+, 0</td>
<td>24</td>
<td>suppressed</td>
<td>-0.5</td>
<td>9</td>
</tr>
<tr>
<td>52</td>
<td>61–70</td>
<td>+, -</td>
<td>15</td>
<td>enhanced</td>
<td>-1.1</td>
<td>9</td>
</tr>
<tr>
<td>53</td>
<td>30–69</td>
<td>0, 0</td>
<td>20</td>
<td>enhanced</td>
<td>-0.6</td>
<td>9</td>
</tr>
<tr>
<td>54</td>
<td>35–58</td>
<td>0, -</td>
<td>15</td>
<td>enhanced</td>
<td>-0.6</td>
<td>8</td>
</tr>
<tr>
<td>55</td>
<td>16–72</td>
<td>-, -</td>
<td>13</td>
<td>suppressed</td>
<td>0.8</td>
<td>8</td>
</tr>
<tr>
<td>56</td>
<td>18–37</td>
<td>-, -</td>
<td>7</td>
<td>suppressed</td>
<td>1.7</td>
<td>8</td>
</tr>
<tr>
<td>57</td>
<td>63–72</td>
<td>+, -</td>
<td>5</td>
<td>suppressed</td>
<td>-6.5</td>
<td>8</td>
</tr>
<tr>
<td>58</td>
<td>61–69</td>
<td>-, 0</td>
<td>20</td>
<td>suppressed</td>
<td>0.4</td>
<td>7</td>
</tr>
<tr>
<td>59</td>
<td>20–61</td>
<td>0, +</td>
<td>22</td>
<td>enhanced</td>
<td>-0.5</td>
<td>7</td>
</tr>
<tr>
<td>60</td>
<td>37–72</td>
<td>-, +</td>
<td>18</td>
<td>enhanced</td>
<td>-0.6</td>
<td>7</td>
</tr>
</tbody>
</table>
close to each other on the protein are expected to have stronger couplings than residues that are further apart. That is, we expect to see very strong association of $\Delta G_{\text{coop}}$ with distance. Indeed, if we examine the pairs with the highest magnitude $\Delta G_{\text{coop}}$'s, they are close to each other on the protein (Figure 6.8).

6.5.3 $\Delta G_{\text{coop}}$ is the Born interaction energy in our coarse-grained electrostatic model

The free energy of cooperativity is written as $\Delta G_{\text{coop}} = \Delta G_{\text{DM}}^{\text{elec}} - \Delta G_{\text{SM1}}^{\text{elec}} - \Delta G_{\text{SM2}}^{\text{elec}} + \Delta G_{\text{WT}}^{\text{elec}}$. It is the difference between the stability change of the double mutant with the single mutants. This form is derived from:

$$\Delta G_{\text{coop}} = (\Delta G_{\text{DM}}^{\text{elec}} - \Delta G_{\text{WT}}^{\text{elec}}) - (\Delta G_{\text{SM1}}^{\text{elec}} - \Delta G_{\text{WT}}^{\text{elec}}) - (\Delta G_{\text{SM2}}^{\text{elec}} - \Delta G_{\text{WT}}^{\text{elec}})$$

$$\Delta G_{\text{coop}} = \Delta G_{\text{DM}}^{\text{elec}} - \Delta G_{\text{WT}}^{\text{elec}} - \Delta G_{\text{SM1}}^{\text{elec}} + \Delta G_{\text{WT}}^{\text{elec}} - \Delta G_{\text{SM2}}^{\text{elec}} + \Delta G_{\text{WT}}^{\text{elec}}$$

Let us now examine a specific wildtype double mutant to determine what exactly this free energy of cooperativity really is in our coarse-grained model which allows no backbone flexibility or relaxation. Consider Table 6.10 which describes the charge states for a pair of hypothetical positions. The wildtype states are both neutral and the double mutant has a pair of negative charges. For this pair, the free energy of cooperativity will be

$$\Delta G_{\text{coop}} = \Delta G_{\text{WT}}^{\text{elec}} - \Delta G_{\text{SM1}}^{\text{elec}} - \Delta G_{\text{SM2}}^{\text{elec}} + \Delta G_{\text{WT}}^{\text{elec}}$$
In this example, the free energy of cooperativity is only due to the interaction between the negative charges of the double mutant. All other interactions get canceled out. For example, the interaction between the negative charge on position 1 and the background in the double mutant is canceled out by the same interaction in the single mutant which contains a negative charge on position 1 ($\Delta G_{-, -} - \Delta G_{-, 0}$). Similar cancellations occur for the negative charge on position 2 ($\Delta G_{-, -} - \Delta G_{0, -}$). Moreover, the interactions of the background with itself is also canceled out. The only interaction that is not canceled out is the interaction between the negative charges on the double mutant. This is the Born energy between the charges and its magnitude decreases with the distance between the charges. The free energy of cooperativity, therefore, is simply the Born energy between non-zero charges on the pair of positions. This interaction does not necessarily have to occur on the double mutant, but can occur on the single mutant itself. However, calculating the free energy of cooperativity in this way by including all 4 states, and not just the double mutant, allows for the possible interactions on the single mutant or the wildtype configuration to be counted or included in our analysis.

6.5.4 P-values for subsets of the most deviated pairs

We selected the top 60 most deviated pairs as the subset to report the p-value for (Table 6.9). If we had selected the top 20, then our p-value might have been lower because of the sample size effect in the cumulative binomial distribution calculation. Table 6.11 lists the correct predictions for the top 10, top 20, top 30, etc pairs as ranked by $\text{dev}(i, j)$. Table 6.12 lists the p-values for the correct predictions within the 1st 60 pairs of double mutations, 2nd 60, 3rd 60, etc to highlight the fact that the signal is strong for only the top 10% of the sorted data.
Table 6.11: The proportion of correct predictions and the p-value for the significance of the prediction as a function of the total number of predictions after sorting by the deviation. All 612 double mutant pairs are sorted by the deviation of the observed bivariate marginals from the independent model. The measure used to test for deviation is $$dev(i,j) = \frac{(P_{ij}(M,M) - P_i(M)P_j(M))^2}{P_{ij}(M,M)}$$ where $$P_{ij}(M,M)$$ is the joint probability of a double mutation at positions $$i$$ and $$j$$ while $$P_i(M)$$ is the univariate marginal of a mutation at position $$i$$. The free energy of cooperativity ($$\Delta G_{coop}$$) for each pair is determined and compared to the pairs statistical enhancement or suppression. A correct prediction for a double mutant implies that pair is enhanced and the sign of $$\Delta G_{coop}$$ is negative or if the pair is suppressed and the sign of $$\Delta G_{coop}$$ is positive.

<table>
<thead>
<tr>
<th>Correct predictions</th>
<th>Total number of predictions</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>10</td>
<td>$$5.5 \times 10^{-2}$$</td>
</tr>
<tr>
<td>18</td>
<td>20</td>
<td>$$2.0 \times 10^{-4}$$</td>
</tr>
<tr>
<td>25</td>
<td>30</td>
<td>$$1.6 \times 10^{-4}$$</td>
</tr>
<tr>
<td>34</td>
<td>40</td>
<td>$$4.2 \times 10^{-6}$$</td>
</tr>
<tr>
<td>42</td>
<td>50</td>
<td>$$5.8 \times 10^{-7}$$</td>
</tr>
<tr>
<td>50</td>
<td>60</td>
<td>$$8.1 \times 10^{-8}$$</td>
</tr>
<tr>
<td>55</td>
<td>70</td>
<td>$$8.3 \times 10^{-9}$$</td>
</tr>
<tr>
<td>60</td>
<td>80</td>
<td>$$4.3 \times 10^{-6}$$</td>
</tr>
</tbody>
</table>

Table 6.12: The p-values for the significance of the prediction for separate subsections of the data. All 612 double mutant pairs are sorted by the deviation of the observed bivariate marginals from the independent model. The measure used to test for deviation is $$dev(i,j) = \frac{(P_{ij}(M,M) - P_i(M)P_j(M))^2}{P_{ij}(M,M)}$$ where $$P_{ij}(M,M)$$ is the joint probability of a double mutation at positions $$i$$ and $$j$$ while $$P_i(M)$$ is the univariate marginal of a mutation at position $$i$$. The free energy of cooperativity ($$\Delta G_{coop}$$) for each pair is determined and compared to the pairs statistical enhancement or suppression. A correct prediction for a double mutant implies that pair is enhanced and the sign of $$\Delta G_{coop}$$ is negative or if the pair is suppressed and the sign of $$\Delta G_{coop}$$ is positive. The p-values for the first 60, second 60, third 60 etc predictions are listed below.

<table>
<thead>
<tr>
<th>Pair subsets</th>
<th>Correct predictions/Total Predictions</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st set of 60 pairs</td>
<td>50/60</td>
<td>$$8.1 \times 10^{-8}$$</td>
</tr>
<tr>
<td>2nd set of 60 pairs</td>
<td>31/60</td>
<td>0.45</td>
</tr>
<tr>
<td>3rd set of 60 pairs</td>
<td>35/60</td>
<td>0.12</td>
</tr>
<tr>
<td>4th set of 60 pairs</td>
<td>34/60</td>
<td>0.18</td>
</tr>
<tr>
<td>5th set of 60 pairs</td>
<td>36/60</td>
<td>0.08</td>
</tr>
</tbody>
</table>
6.5.5 Justification for comparing $\Delta G_{\text{coop}}$ to the observed statistical deviation

50 out of the top 60 most positively and negatively correlated pairs of mutations have a $\Delta G_{\text{coop}}$ whose sign is consistent with the statistical deviation (Table 6.9). The p-value is $\sim 10^{-8}$ indicating that statistical significance of this observation. But does it make sense to compare the observed deviation from the independent model of the double mutant state to a free energy of cooperativity that is a function of the four states associated with the double mutant (wildtype, single mutant, single mutant and double mutant)? To justify its usage, let us examine the underlying probabilities that make up a measure that describes the statistical deviation from the independent model. Let $P(M_1, M_2)$ be the

Figure 6.9: The percentage of pairs where statistical correlation is consistent with sign of the free energy of cooperativity, $\Delta G_{\text{coop}}$, plotted as a function of the percentage of total pairs, sorted by decreasing correlation. At the top is the plot of the negatively correlated pairs while the plot at the bottom are the positively correlated pairs. Correlation is calculated using the following measure: $\frac{(P_{\text{obs}}-P_{\text{ind}})^2}{P_{\text{ind}}}$. 
bivariate marginal of the double mutant state. If this marginal can be explained by the independent model then

\[ P(M_1, M_2) = P(M_1)P(M_2) \]

But

\[ P(M_1) = P(M_1, M_2) + P(M_1, W_2) \]

and

\[ P(M_2) = P(M_1, M_2) + P(W_1, M_2) \]

Therefore

\[ P(M_1)P(M_2) = (P(M_1, M_2) + P(M_1, W_2))(P(M_1, M_2) + P(W_1, M_2)) \]

\[ P(M_1)P(M_2) = P(M_1, M_2)P(M_1, M_2) + P(M_1, M_2)P(W_1, M_2) \]

\[ + P(M_1, W_2)P(M_1, M_2) + P(M_1, W_2)P(W_1, M_2) \]

\[ P(M_1)P(M_2) = P(M_1, M_2)(P(M_1, M_2) + P(W_1, M_2) + P(M_1, W_2)) + P(M_1, W_2)P(W_1, M_2) \]

and finally

\[ P(M_1)P(M_2) = P(M_1, M_2)(1 - P(W_1, W_2)) + P(M_1, W_2)P(W_1, M_2) \]

since

\[ P(M_1, M_2) + P(W_1, M_2) + P(M_1, W_2) + P(W_1, W_2) = 1 \]

Therefore if the bivariate marginal can be explained as a product of the univariate marginals

\[ P(M_1, M_2) = P(M_1)P(M_2) \]

and using our previously derived form of the univariate marginals

\[ P(M_1, M_2) = P(M_1, M_2)(1 - P(W_1, W_2)) + P(M_1, W_2)P(W_1, M_2) \]
and

\[ P(M_1, M_2)P(W_1, W_2) = P(M_1, W_2)P(W_1, M_2) \]

or

\[ P(M_1, M_2)P(W_1, W_2) - P(M_1, W_2)P(W_1, M_2) = 0 \]

where \( P(M_1, M_2)P(W_1, W_2) - P(M_1, W_2)P(W_1, M_2) \) is the determinant of the 2x2 matrix:

\[
\begin{vmatrix}
P(M_1, M_2) & P(M_1, W_2) \\
P(W_1, M_2) & P(W_1, W_2)
\end{vmatrix}
\]

In other words, if the determinant of this matrix is equal to 0, then the bivariate marginals positions 1 and 2 are independent of each other. Converting the probabilities to Boltzmann factors we observe that if the independent model holds then

\[ P(M_1, M_2)P(W_1, W_2) - P(M_1, W_2)P(W_1, M_2) = e^{-\Delta G_{M_1} - \Delta G_{W_1} - \Delta G_{M_2} - \Delta G_{W_2}} = 0 \]

Further rearrangement leads to

\[ e^{-\Delta G_{M_1} - \Delta G_{W_1} - \Delta G_{M_2}} - e^{-\Delta G_{M_1} - \Delta G_{W_2}} = 0 \]

\[ e^{-\Delta G_{M_1} - \Delta G_{W_1} - \Delta G_{M_2}} = e^{-\Delta G_{M_1} - \Delta G_{W_1}} \]

After removing the exponents, we get

\[ -\Delta G_{M_1} - \Delta G_{W_1} - \Delta G_{M_2} = -\Delta G_{M_1} - \Delta G_{W_1} \]

\[ \Delta G_{M_1} + \Delta G_{W_2} - \Delta G_{M_1} - \Delta G_{W_1} = \Delta G_{coop} = 0 \]

That is, if the mutation probabilities are independent, then \( \Delta G_{coop} = 0 \), illustrating that qualitatively, the direction of the observed statistical deviation from the independent model results directly from the free energy of \( \Delta G_{coop} \). In other words, the sign of the free energy of cooperativity can be used to predict
the direction of statistical deviation and which is why 50 of the 60 of the most
correlated double mutants have $\Delta G_{\text{coop}}$’s that are energetically consistent with
enhancement or suppression of the double mutant.

6.5.6 Measures of statistical deviation

Though the previous section advocates the use of the determinant as our pri-
mary measure of statistical deviation, we need a measure that is a bit more
robust to scaling and sorting since our goal is to be able to study the most
statistically deviated pairs. We therefore make use a measure which we define
as

$$dev(i, j) = \frac{[P_{ij}(A_i, A_j) - P_i(A_i)P_j(A_j)]^2}{\max(P_{ij}(A_i, A_j), P_i(A_i)P_j(A_j))}$$

This measure scales from 0 to infinity, where a score closer to 0 indicates sta-
tistical independence.

6.5.7 Examining strongly correlated pairs of electrostatic mutations

whose $\Delta G_{\text{coop}}$ sign is consistent with the sign of statistical
correlation in the database

There are 18 electrostatically active residues in our analysis. Therefore there
are $\binom{18}{2} = 153$ possible unique pairs of positions. For each pair, there are
9 possible charge states in a 3 letter charge alphabet. 4 of these 9 states are
double mutants, 4 are single mutants while 1 is the wildtype charge state. As
a result, there are 153 $\times$ 4 = 612 unique pairs of charged mutations. If we
examine the sequence database and calculate the observed deviation from the
independent model for these 612 double mutants, we find that the majority of
them are either unobserved in the database, or uncorrelated. That is, the inde-
dependent model predictions match the observed marginals. However, there are
many pairs that are either statistically enhanced or suppressed in the database. Of the top 60 pairs of double mutants listed in Table 6.9, whose free energy of cooperativity is consistent with the pair being suppressed or enhanced, not all 18 positions are present evenly. 306 positions are listed for the 153 pairs. Each of the 18 positions is listed 17 times each in this list of 306. If the positions were distributed evenly, then within a subset of 60 pairs, then each position would be listed 6.67 times each within Table 6.9. However, Table 6.13 shows that position 20 (accessory), 37 (polymorphic), 69 (polymorphic) and 72 (polymorphic) are overrepresented while positions 7 (polymorphic), 12 (polymorphic), 30 (primary), 34 (accessory) and 88 (primary) are underrepresented. Some of these positions are very interesting. The next few subsubsections discuss these unevenly distributed positions in some detail by focusing on their pharmaceutical relevance.

6.5.7.1 K20 (overrepresented accessory mutation position)

Position 20 is an accessory mutation position as designatated by the Stanford HIV database (author?) [101, 100, 102, 103]. The wildtype state for this position in HIV protease subtype B is K (lysine). The most common mutations at this site according to the Stanford HIV database, in decreasing order of mutation frequency, are: K20I, K20R, K20T, K20M, K20V, K20L, K20E, K20Q (author?) [101, 100, 102, 103]. Several of these mutations are actually consensus amino acids for other viral subtypes of HIV protease. Moreover, according to work by Christopher Lee et al., the selection pressure at this site is highly indicative of treatment related drug-pressure (author?) [20], providing further statistical evidence of its clinically defined role as an accessory drug resistance mutation.

According the the Stanford HIV database, the K20R/M/I/T/V mutations
are associated with resistance to multiple PIs when present with other mutations (author?) [101, 100, 102, 103]. K20R/M/I are polymorphic; K20T/V are nonpolymorphic. Compared with other PIs, K20I/T/V are more likely to be selected by NFV and to reduce its susceptibility. K20ITV are common relatively nonpolymorphic accessory mutations. They contribute to reduced susceptibility to NFV and possibly other PIs. K20I is the consensus residue for subtype G. K20I and T74S occur commonly in other subtypes. K20R is a common highly polymorphic compensatory mutation, compensating for the decreased fitness associated with major PI-resistance mutations.

In terms of statistical correlations, mutations at K20 are highly correlated with other mutations in the Lee and Shafer sequence databases. Of the top 60 highly correlated pairs of double mutations, 16 involve position 20 (Table 6.9). The majority of these pairs are positively correlated. The same trends hold for statistical covariation studies on non-electrostatic mutations. For example, the mutation K20R has strong, statistically significant positive correlations with M36I, I54V and E35D (author?) [122]. All three mutations are non-electrostatic, drug resistance mutations. 35 and 36 are considered to be accessory mutation sites while 54 is a site for primary drug resistance mutations. K20R is also strongly negatively correlated with V77I, which is an accessory drug resistance mutation (author?) [122].

**6.5.7.2 N37 (overrepresented)**

N37D - IND

N37E - APV, NFV

N37K - LPV

N37H - IND
N37R- SQV

Plausible structural explanations for covariation demonstrated by a subset of these pairs (35:37, 12:19, 71:93, 15:77, and 36:77) are possible. For example, the E35D N37D combination occurred around twice as often as expected in both the notx and the tx sequence sets (Table 3). N37D appears less often than E35D, suggesting that N37D is stabilized in the presence of E35D to a greater extent than the reverse. This is also reflected in the values of $U(35|37)$ and $U(37|35)$ for these two positions which are 0.087 and 0.047, respectively, in the tx set (i.e., knowing the composition of position 37 reduces the uncertainty of position 35 more than vice versa). Positions 35 and 37 are near one another in the protease structure in the hinge region of the flap, but the side chains are 5.5 Å apart. E35 forms an ionic bond with R57. In the context of an E35D mutation, R57 might extend to maintain an interaction with the shorter aspartic acid side chain; this extension could also bring R57 within bonding distance of the side chain of position 37. Thus N37D would be stabilized by interacting with R57. *(author?)* [48]

6.5.7.3 H69 *(overrepresented)*

The A and C HIV-1 subtype proteases used in these studies carry eight and five mutations that have been incorporated within the framework of the B subtype protease [I13V/E35D/M36I/ S37N/R41K/R57K/H69K/L89M for the A subtype and M36I/S37N/R41K/H69K/L89M for the C subtype]

In protease (Fig. 1b), the majority of subtype A isolates were polymorphic at codons I13, E35, M36, R41, G57, H69 and L89. More than 80% of isolates were polymorphic at M36, R41, G57, H69 and L89 for subtype C isolates, and also at codons I15, L19 and I93.
6.5.7.4 I72 (overrepresented)

I72E

I72R - IDV

I72K - SQV

6.5.7.5 Q7 (underrepresented)

Q7H, Q7E, Q7R - APV, Q7K - NFV, Q7D - NFV

Autoproteolysis of the retroviral aspartyl proteases is a major obstacle to purification and analysis of these enzymes. A mutagenic approach to rendering autolytic cleavage sites less labile was applied to the primary cleavage site between Leu5 and Trp6 in human immunodeficiency virus-1 (HIV-1) protease. From predictions based on known substrates it was concluded that amino acids Lys or Ser in place of Gln at position 7 would prevent cleavage at the Leu5-Trp6 peptide bond, therefore stabilizing the protein. Autoproteolytic stability was enhanced at least 100-fold by these mutations. At longer time points the protease was degraded at secondary sites which contained adequate substrate sequences but were conformationally restricted. Conversely, a mutation in HIV-2 protease which changed Lys7 to Gln rendered the protein 3-fold less stable and shifted the position of the initial autoproteolytic cleavage from Phe3-Ser4 to Leu5-Trp6. The effects of these mutations demonstrate that small changes in protein sequence can have a major impact on their autoproteolytic stability. The work described here suggests a general method for stabilizing proteases and perhaps other recombinantly produced proteins to autolysis.

The mutant protein, HIV-1 protease Q7K, has a similar inhibition profile, similar kinetic parameters, and a more than 100-fold greater stability toward
proteolytic degradation than the wild-type enzyme (24

6.5.7.6 T12 (underrepresented)

For example, at protease codon 12 several amino acid changes were positively
selected (T12K, T12P, and T12S), resulting in an overall Ka/Ks value of 6.92
for the codon. (author?) [20]

6.5.7.7 D30 (underrepresented)

30 D N D30N causes high-level resistance to NFV. Positions 30, 32, 46, 47, 48,
50, 54, 82, and 84 are in or near (positions 46, 47, and 54) the substrate cleft.

6.5.7.8 E34 (underrepresented)

NON-POLYMORPHIC PI-SELECTED MUTATIONS AT POSITIONS NOT
GENERALLY RECOGNIZED AS PI-RESISTANCE POSITIONS: Nonpoly-
morphic mutations that occur significantly more frequently in PI-experienced
compared with PI-naive patients: A22V, E34Q, E35G, K55RN, I66FVL, C67FL,
V75I, P79ASD, I85V, T91S, Q92K, and C95F (Carrillo 98; Parkin 03; Wu 03;
Ceccherini-Silberstein 04; Svicher 05; Kagan 06; Margerison 08; Shahriar 09).

6.5.7.9 N88 (underrepresented)

RARE VARIANTS AT KNOWN DRUG-RESISTANCE POSITIONS: L10RY,
V11L, M46V, G48ASTLQ, F53Y, I54S, V82MC, I84AC, N88TG, and V89T
(Camacho 05; Baxter 06; Mo 07; Vermeiren 07; Shahriar 09; Rhee 10);

N88S increases FPV susceptibility (Ziermann 00).
INTER-SUBTYPE DIFFERENCES: In several non-B subtypes including C, 01_AE, and G, D30N is less likely to cause NFV resistance than other mutations such as L90M or N88S.

88 N D N88D is selected by NFV and, in combination with D30N, increases NFV resistance. It causes low-level cross-resistance to ATV. 88 N GT N88S causes high-level resistance to NFV and ATV and low-level resistance to IDV; it increases susceptibility to FPV. N88T/G are rare PI-selected mutations that may have an effect similar to but less pronounced than N88S. 88 N S N88S causes high-level resistance to NFV and ATV and low-level resistance to IDV; it increases susceptibility to FPV.

Positions 76, 88, and 90 interfere with PI susceptibility indirectly (Patrick 98; Atkinson 00; Mahalingam 01; Malan 08; Rhee 10; Louis 11).

6.5.8 Is the sign of $\Delta G_{coop}$ consistent with like/unlike pairs of charges in a dielectric?
Table 6.13: Frequency of electrostatically active positions within the top 60 most correlated double mutants.

<table>
<thead>
<tr>
<th>#</th>
<th>position</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>34</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>35</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>37</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>43</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>58</td>
<td>7</td>
</tr>
<tr>
<td>12</td>
<td>61</td>
<td>8</td>
</tr>
<tr>
<td>13</td>
<td>63</td>
<td>8</td>
</tr>
<tr>
<td>14</td>
<td>69</td>
<td>10</td>
</tr>
<tr>
<td>15</td>
<td>70</td>
<td>8</td>
</tr>
<tr>
<td>16</td>
<td>72</td>
<td>10</td>
</tr>
<tr>
<td>17</td>
<td>88</td>
<td>4</td>
</tr>
<tr>
<td>18</td>
<td>92</td>
<td>6</td>
</tr>
</tbody>
</table>
Chapter 7

Understanding direct and indirect interactions in light of the Bethe approximation

7.1 Goal of this mini-study

Study model problems such as a 3 node acyclic graph. Select fields and couplings. Run the Bethe approximation in the inverse direction.

Does the Bethe approximation used inversely, get the original fields and couplings back?

In a 3 node acyclic/cyclic graph. Put in fields/take out fields. Increase the strength of the the 3 edge, while keeping 2 constant. See if the inverse Bethe approximation gets the original fields and couplings back.

Run the Bethe approximation in a forward direction. That is, given fields and couplings, determine the univariate and bivariate marginals. How well do these Bethe approximated marginals match the exact marginals? How well does the Bethe approximation approximate the marginals if the strength of the fields is increased? How well does it approximate the marginals if the couplings are increased? What about the cyclic nature of the graph? If the graph becomes more increasingly cyclic, does the approximation break down?

\[ p_{ij}^{\text{Direct}} \text{ vs } p_{ij}^{\text{Exact}} \]
If the Bethe approximation works, does that imply that $P_{ij}^{\text{Direct}} = P_{ij}^{\text{Exact}}$? What is the distinction between $P_{ij}^{\text{Direct}}$ and $P_{ij}^{\text{Exact}}$? Study this on a 3 node model system.

7.2 Approach used

In order to study the properties of the Bethe approximation, we need to understand it on a simple system. The simplest system possible is a 3 node graph. This can be cyclic or acyclic. In this mini-study, we create model systems with both kinds of graphs.

1. Choose appropriate field and coupling parameters. This will be the original set of parameters: $\lambda_i^{\text{exact}}(A_i)$ and $\lambda_{ij}^{\text{exact}}(A_i, A_j)$.

2. Determine the univariate and bivariate marginals for the graph by calculating exhaustively, the probabilities of all possible states using $P(A_1, A_2, A_3) = \frac{1}{Z} \exp[\sum_i \lambda_i^{\text{exact}}(A_i) - \sum_{i<j} \lambda_{ij}^{\text{exact}}(A_i, A_j)]$. These will be the original univariate and bivariate marginals: $P_i^{\text{exact}}(A_i)$ and $P_{ij}^{\text{exact}}(A_i, A_j)$.

3. From the exact univariate and bivariate marginals, run the inverse incrementally updated Bethe approximation algorithm to determine approximate fields and coupling parameters: $\lambda_i^{\text{Bethe}}(A_i)$ and $\lambda_{ij}^{\text{Bethe}}(A_i, A_j)$.

4. Use the approximate field and coupling parameters in this Hamiltonian $P(A_1, A_2, A_3) = \frac{1}{Z} \exp[\sum_i \lambda_i^{\text{Bethe}}(A_i) - \sum_{i<j} \lambda_{ij}^{\text{Bethe}}(A_i, A_j)]$, to exhaustively calculate the univariate and bivariate marginals that result from these parameters: $P_i^{\text{Bethe}}(A_i)$ and $P_{ij}^{\text{Bethe}}(A_i, A_j)$.
7.3 Simple 3 node complete graph

In the following model experiment, I wanted to investigate the effect of fields on the accuracy of the inverse Bethe approximation algorithm. I started with a complete (cyclic) 3 node graph. One edge, however, is a magnitude weaker than the other two. In subsequent experiments, I will change this around. I start with strong fields at each node, then subsequently reduce the magnitude of the fields, while holding all the couplings at the same magnitude. I then determine the Bethe fields and couplings and their corresponding univariate and bivariate marginals. Test 1, Test 2, and Test 3 are the three mini-experiments that I ran.

I started with fields for the Ising spins around a magnitude of 2 (Test 1). I subtracted 1 from the magnitude of the fields in Test 2. In test 3, I removed the fields all-together.

From the results, it seems that the presence of fields is important for the inverse Bethe approximation to work accurately. That is, my observation is that if the univariate marginals are around 0.90, then the system is well approximated by the field and coupling parameters that fit the Bethe distribution.
for this system. This is an interesting observation considering that the HIV protease maintains this property.

### 7.3.1 Test 1

<table>
<thead>
<tr>
<th>$\lambda_1^{exact}(A_1)$</th>
<th>$\lambda_1^{exact}(A_2)$</th>
<th>1.999</th>
<th>−1.999</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_2^{exact}(A_1)$</td>
<td>$\lambda_2^{exact}(A_2)$</td>
<td>2.088</td>
<td>−2.088</td>
</tr>
<tr>
<td>$\lambda_3^{exact}(A_1)$</td>
<td>$\lambda_3^{exact}(A_2)$</td>
<td>1.987</td>
<td>−1.987</td>
</tr>
</tbody>
</table>

Test 1: Original Fields and Couplings

| $\lambda_{12}^{exact}(A_1, A_1)$ | $\lambda_{12}^{exact}(A_1, A_2)$ | 0.300 | −0.300 |
| $\lambda_{12}^{exact}(A_2, A_1)$ | $\lambda_{12}^{exact}(A_2, A_2)$ | −0.300 | 0.300 |
| $\lambda_{23}^{exact}(A_1, A_1)$ | $\lambda_{23}^{exact}(A_1, A_2)$ | 0.499 | −0.499 |
| $\lambda_{23}^{exact}(A_2, A_1)$ | $\lambda_{23}^{exact}(A_2, A_2)$ | −0.499 | 0.499 |
| $\lambda_{13}^{exact}(A_1, A_1)$ | $\lambda_{13}^{exact}(A_1, A_2)$ | −0.013 | 0.013 |
| $\lambda_{13}^{exact}(A_2, A_1)$ | $\lambda_{13}^{exact}(A_2, A_2)$ | −0.013 | −0.013 |

Test 1: Original Marginals

| $P_1^{exact}(A_1)$ | $P_1^{exact}(A_2)$ | 0.921 | 0.079 |
| $P_2^{exact}(A_1)$ | $P_2^{exact}(A_2)$ | 0.975 | 0.025 |
| $P_3^{exact}(A_1)$ | $P_3^{exact}(A_2)$ | 0.956 | 0.044 |
| $P_{12}^{exact}(A_1, A_1)$ | $P_{12}^{exact}(A_1, A_2)$ | 0.897 | 0.025 |
| $P_{12}^{exact}(A_2, A_1)$ | $P_{12}^{exact}(A_2, A_2)$ | 0.078 | 0.001 |
| $P_{23}^{exact}(A_1, A_1)$ | $P_{23}^{exact}(A_1, A_2)$ | 0.878 | 0.044 |
| $P_{23}^{exact}(A_2, A_1)$ | $P_{23}^{exact}(A_2, A_2)$ | 0.078 | 0.001 |
| $P_{13}^{exact}(A_1, A_1)$ | $P_{13}^{exact}(A_1, A_2)$ | 0.932 | 0.043 |
| $P_{13}^{exact}(A_2, A_1)$ | $P_{13}^{exact}(A_2, A_2)$ | 0.024 | 0.001 |
### Test 1: Bethe Predicted Fields and Couplings

<table>
<thead>
<tr>
<th>$\lambda_{Bethe}^1(A_1)$</th>
<th>$\lambda_{Bethe}^1(A_2)$</th>
<th>$\lambda_{Bethe}^2(A_1)$</th>
<th>$\lambda_{Bethe}^2(A_2)$</th>
<th>$\lambda_{Bethe}^3(A_1)$</th>
<th>$\lambda_{Bethe}^3(A_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.999</td>
<td>-1.999</td>
<td>2.077</td>
<td>-2.077</td>
<td>1.974</td>
<td>-1.974</td>
</tr>
</tbody>
</table>

### Test 1: Bethe Predicted Marginals

<table>
<thead>
<tr>
<th>$P_{Bethe}^1(A_1)$</th>
<th>$P_{Bethe}^1(A_2)$</th>
<th>$P_{Bethe}^2(A_1)$</th>
<th>$P_{Bethe}^2(A_2)$</th>
<th>$P_{Bethe}^3(A_1)$</th>
<th>$P_{Bethe}^3(A_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.921</td>
<td>0.079</td>
<td>0.975</td>
<td>0.025</td>
<td>0.956</td>
<td>0.044</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$P_{Bethe}^1(A_1, A_1)$</th>
<th>$P_{Bethe}^1(A_1, A_2)$</th>
<th>$P_{Bethe}^2(A_1, A_1)$</th>
<th>$P_{Bethe}^2(A_1, A_2)$</th>
<th>$P_{Bethe}^3(A_1, A_1)$</th>
<th>$P_{Bethe}^3(A_1, A_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.897</td>
<td>0.025</td>
<td>0.078</td>
<td>0.001</td>
<td>0.878</td>
<td>0.044</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$P_{Bethe}^1(A_2, A_1)$</th>
<th>$P_{Bethe}^1(A_2, A_2)$</th>
<th>$P_{Bethe}^2(A_2, A_1)$</th>
<th>$P_{Bethe}^2(A_2, A_2)$</th>
<th>$P_{Bethe}^3(A_2, A_1)$</th>
<th>$P_{Bethe}^3(A_2, A_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.932</td>
<td>0.043</td>
<td>0.024</td>
<td>0.001</td>
<td>0.932</td>
<td>0.043</td>
</tr>
</tbody>
</table>
### Test 2

#### Original Fields and Couplings

| $\lambda_{1}^{\text{exact}}(A_1)$ | $\lambda_{1}^{\text{exact}}(A_2)$ | 0.999 | -0.999 |
| $\lambda_{2}^{\text{exact}}(A_1)$ | $\lambda_{2}^{\text{exact}}(A_2)$ | 1.088 | -1.088 |
| $\lambda_{3}^{\text{exact}}(A_1)$ | $\lambda_{3}^{\text{exact}}(A_2)$ | 0.987 | -0.987 |

#### Original Marginals

| $P_{1}^{\text{exact}}(A_1)$ | $P_{1}^{\text{exact}}(A_2)$ | 0.688 | 0.312 |
| $P_{2}^{\text{exact}}(A_1)$ | $P_{2}^{\text{exact}}(A_2)$ | 0.865 | 0.135 |
| $P_{3}^{\text{exact}}(A_1)$ | $P_{3}^{\text{exact}}(A_2)$ | 0.799 | 0.201 |

Test 2: Original Fields and Couplings

Test 2: Original Marginals
\[
\begin{array}{|c|c|}
\hline
\lambda_{1}^{\text{Bethe}}(A_1) & \lambda_{1}^{\text{Bethe}}(A_2) \\
\hline
1.002 & -1.002 \\
\hline
\lambda_{2}^{\text{Bethe}}(A_1) & \lambda_{2}^{\text{Bethe}}(A_2) \\
\hline
1.052 & -1.052 \\
\hline
\lambda_{3}^{\text{Bethe}}(A_1) & \lambda_{3}^{\text{Bethe}}(A_2) \\
\hline
0.940 & -0.940 \\
\hline
\end{array}
\]

Test 2: Bethe Predicted Fields and Couplings

\[
\begin{array}{|c|c|}
\hline
\lambda_{12}^{\text{Bethe}}(A_1, A_1) & \lambda_{12}^{\text{Bethe}}(A_1, A_2) \\
\hline
-0.303 & 0.303 \\
\hline
\lambda_{12}^{\text{Bethe}}(A_2, A_1) & \lambda_{12}^{\text{Bethe}}(A_2, A_2) \\
\hline
0.303 & -0.303 \\
\hline
\lambda_{23}^{\text{Bethe}}(A_1, A_1) & \lambda_{23}^{\text{Bethe}}(A_1, A_2) \\
\hline
-0.501 & 0.501 \\
\hline
\lambda_{23}^{\text{Bethe}}(A_2, A_1) & \lambda_{23}^{\text{Bethe}}(A_2, A_2) \\
\hline
0.501 & -0.501 \\
\hline
\lambda_{13}^{\text{Bethe}}(A_1, A_1) & \lambda_{13}^{\text{Bethe}}(A_1, A_2) \\
\hline
0.080 & -0.080 \\
\hline
\lambda_{13}^{\text{Bethe}}(A_2, A_1) & \lambda_{13}^{\text{Bethe}}(A_2, A_2) \\
\hline
-0.080 & 0.080 \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|}
\hline
P_{1}^{\text{Bethe}}(A_1) & P_{1}^{\text{Bethe}}(A_2) \\
\hline
0.687 & 0.313 \\
\hline
P_{2}^{\text{Bethe}}(A_1) & P_{2}^{\text{Bethe}}(A_2) \\
\hline
0.864 & 0.136 \\
\hline
P_{3}^{\text{Bethe}}(A_1) & P_{3}^{\text{Bethe}}(A_2) \\
\hline
0.798 & 0.202 \\
\hline
\end{array}
\]

Test 2: Bethe Predicted Marginals

\[
\begin{array}{|c|c|}
\hline
P_{12}^{\text{Bethe}}(A_1, A_1) & P_{12}^{\text{Bethe}}(A_1, A_2) \\
\hline
0.568 & 0.119 \\
\hline
P_{12}^{\text{Bethe}}(A_2, A_1) & P_{12}^{\text{Bethe}}(A_2, A_2) \\
\hline
0.296 & 0.017 \\
\hline
P_{23}^{\text{Bethe}}(A_1, A_1) & P_{23}^{\text{Bethe}}(A_1, A_2) \\
\hline
0.500 & 0.187 \\
\hline
P_{23}^{\text{Bethe}}(A_2, A_1) & P_{23}^{\text{Bethe}}(A_2, A_2) \\
\hline
0.298 & 0.015 \\
\hline
P_{13}^{\text{Bethe}}(A_1, A_1) & P_{13}^{\text{Bethe}}(A_1, A_2) \\
\hline
0.702 & 0.162 \\
\hline
P_{13}^{\text{Bethe}}(A_2, A_1) & P_{13}^{\text{Bethe}}(A_2, A_2) \\
\hline
0.096 & 0.040 \\
\hline
\end{array}
\]
### 7.3.3 Test 3

#### Original Fields and Couplings

<table>
<thead>
<tr>
<th>$\lambda_{\text{exact}}^1(A_1)$</th>
<th>$\lambda_{\text{exact}}^1(A_2)$</th>
<th>0.000</th>
<th>0.000</th>
<th>Test 3: Original Fields and Couplings</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_{\text{exact}}^2(A_1)$</td>
<td>$\lambda_{\text{exact}}^2(A_2)$</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>$\lambda_{\text{exact}}^3(A_1)$</td>
<td>$\lambda_{\text{exact}}^3(A_2)$</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\lambda_{\text{exact}}^{12}(A_1, A_1)$</th>
<th>$\lambda_{\text{exact}}^{12}(A_1, A_2)$</th>
<th>$\lambda_{\text{exact}}^{12}(A_2, A_1)$</th>
<th>$\lambda_{\text{exact}}^{12}(A_2, A_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_{\text{exact}}^{12}(A_1, A_1)$</td>
<td>$\lambda_{\text{exact}}^{12}(A_1, A_2)$</td>
<td>$\lambda_{\text{exact}}^{12}(A_2, A_1)$</td>
<td>$\lambda_{\text{exact}}^{12}(A_2, A_2)$</td>
</tr>
<tr>
<td>$\lambda_{\text{exact}}^{13}(A_1, A_1)$</td>
<td>$\lambda_{\text{exact}}^{13}(A_1, A_2)$</td>
<td>$\lambda_{\text{exact}}^{13}(A_2, A_1)$</td>
<td>$\lambda_{\text{exact}}^{13}(A_2, A_2)$</td>
</tr>
</tbody>
</table>

#### Original Marginals

<table>
<thead>
<tr>
<th>$P_{\text{exact}}^1(A_1)$</th>
<th>$P_{\text{exact}}^1(A_2)$</th>
<th>0.500</th>
<th>0.500</th>
<th>Test 3: Original Marginals</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{\text{exact}}^2(A_1)$</td>
<td>$P_{\text{exact}}^2(A_2)$</td>
<td>0.500</td>
<td>0.500</td>
<td></td>
</tr>
<tr>
<td>$P_{\text{exact}}^3(A_1)$</td>
<td>$P_{\text{exact}}^3(A_2)$</td>
<td>0.500</td>
<td>0.500</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$P_{\text{exact}}^{12}(A_1, A_1)$</th>
<th>$P_{\text{exact}}^{12}(A_1, A_2)$</th>
<th>$P_{\text{exact}}^{12}(A_2, A_1)$</th>
<th>$P_{\text{exact}}^{12}(A_2, A_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{\text{exact}}^{12}(A_1, A_1)$</td>
<td>$P_{\text{exact}}^{12}(A_1, A_2)$</td>
<td>$P_{\text{exact}}^{12}(A_2, A_1)$</td>
<td>$P_{\text{exact}}^{12}(A_2, A_2)$</td>
</tr>
<tr>
<td>$P_{\text{exact}}^{13}(A_1, A_1)$</td>
<td>$P_{\text{exact}}^{13}(A_1, A_2)$</td>
<td>$P_{\text{exact}}^{13}(A_2, A_1)$</td>
<td>$P_{\text{exact}}^{13}(A_2, A_2)$</td>
</tr>
<tr>
<td>$P_{\text{exact}}^{13}(A_1, A_1)$</td>
<td>$P_{\text{exact}}^{13}(A_1, A_2)$</td>
<td>$P_{\text{exact}}^{13}(A_2, A_1)$</td>
<td>$P_{\text{exact}}^{13}(A_2, A_2)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$P_{\text{exact}}^{23}(A_1, A_1)$</th>
<th>$P_{\text{exact}}^{23}(A_1, A_2)$</th>
<th>$P_{\text{exact}}^{23}(A_2, A_1)$</th>
<th>$P_{\text{exact}}^{23}(A_2, A_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{\text{exact}}^{23}(A_1, A_1)$</td>
<td>$P_{\text{exact}}^{23}(A_1, A_2)$</td>
<td>$P_{\text{exact}}^{23}(A_2, A_1)$</td>
<td>$P_{\text{exact}}^{23}(A_2, A_2)$</td>
</tr>
<tr>
<td>$P_{\text{exact}}^{23}(A_1, A_1)$</td>
<td>$P_{\text{exact}}^{23}(A_1, A_2)$</td>
<td>$P_{\text{exact}}^{23}(A_2, A_1)$</td>
<td>$P_{\text{exact}}^{23}(A_2, A_2)$</td>
</tr>
</tbody>
</table>
Test 3: Bethe Predicted Fields and Couplings

<table>
<thead>
<tr>
<th>( \lambda_{1}^{\text{Bethe}}(A_1) )</th>
<th>( \lambda_{1}^{\text{Bethe}}(A_2) )</th>
<th>0.000</th>
<th>0.000</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda_{2}^{\text{Bethe}}(A_1) )</td>
<td>( \lambda_{2}^{\text{Bethe}}(A_2) )</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>( \lambda_{3}^{\text{Bethe}}(A_1) )</td>
<td>( \lambda_{3}^{\text{Bethe}}(A_2) )</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Test 3: Bethe Predicted Marginals

<table>
<thead>
<tr>
<th>( P_{12}^{\text{Bethe}}(A_1, A_1) )</th>
<th>( P_{12}^{\text{Bethe}}(A_1, A_2) )</th>
<th>0.500</th>
<th>0.500</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_{12}^{\text{Bethe}}(A_2, A_1) )</td>
<td>( P_{12}^{\text{Bethe}}(A_2, A_2) )</td>
<td>0.500</td>
<td>0.500</td>
</tr>
<tr>
<td>( P_{12}^{\text{Bethe}}(A_1, A_1) )</td>
<td>( P_{12}^{\text{Bethe}}(A_1, A_2) )</td>
<td>0.161</td>
<td>0.339</td>
</tr>
<tr>
<td>( P_{12}^{\text{Bethe}}(A_2, A_1) )</td>
<td>( P_{12}^{\text{Bethe}}(A_2, A_2) )</td>
<td>0.339</td>
<td>0.161</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( P_{23}^{\text{Bethe}}(A_1, A_1) )</th>
<th>( P_{23}^{\text{Bethe}}(A_1, A_2) )</th>
<th>0.126</th>
<th>0.374</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_{23}^{\text{Bethe}}(A_2, A_1) )</td>
<td>( P_{23}^{\text{Bethe}}(A_2, A_2) )</td>
<td>0.374</td>
<td>0.126</td>
</tr>
<tr>
<td>( P_{13}^{\text{Bethe}}(A_1, A_1) )</td>
<td>( P_{13}^{\text{Bethe}}(A_1, A_2) )</td>
<td>0.320</td>
<td>0.180</td>
</tr>
<tr>
<td>( P_{13}^{\text{Bethe}}(A_2, A_1) )</td>
<td>( P_{13}^{\text{Bethe}}(A_2, A_2) )</td>
<td>0.180</td>
<td>0.320</td>
</tr>
</tbody>
</table>
Figure 7.2: This three node graph contains only two edges, and is therefore acyclic. The graph structure indicates that if we condition on the variable $A_2$, then $A_1$ and $A_3$ are conditionally independent. As a result, the joint probability for this graph will have a specific structure. For this probabilistic graphical model, the joint probability $P(A_1, A_2, A_3) \approx \exp \left[ \lambda_{12}(A_1, A_2)\lambda_{23}(A_2, A_3) \right]$ where $\lambda_{12}(A_1, A_2)$ and $\lambda_{23}(A_2, A_3)$ are the coupling potentials. $\lambda_{13}(A_1, A_3) = 0$.

### 7.4 Simple 3 node acyclic graph

#### 7.4.1 Test 4

<table>
<thead>
<tr>
<th>$\lambda_{12}^{exact}(A_1, A_1)$</th>
<th>$\lambda_{12}^{exact}(A_1, A_2)$</th>
<th>$\lambda_{13}^{exact}(A_1, A_1)$</th>
<th>$\lambda_{13}^{exact}(A_1, A_2)$</th>
<th>$\lambda_{23}^{exact}(A_1, A_1)$</th>
<th>$\lambda_{23}^{exact}(A_1, A_2)$</th>
<th>$\lambda_{23}^{exact}(A_2, A_1)$</th>
<th>$\lambda_{23}^{exact}(A_2, A_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>-0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>-0.300</td>
</tr>
<tr>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>-0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>-0.500</td>
</tr>
<tr>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
\[
\begin{array}{c|c|c}
P_{1}^\text{exact}(A_1) & P_{1}^\text{exact}(A_2) & 0.500 \\
\hline
P_{2}^\text{exact}(A_1) & P_{2}^\text{exact}(A_2) & 0.500 \\
\hline
P_{3}^\text{exact}(A_1) & P_{3}^\text{exact}(A_2) & 0.500 \\
\end{array}
\]

Test 4: Original Marginals

\[
\begin{array}{c|c|c}
P_{12}^\text{exact}(A_1, A_1) & P_{12}^\text{exact}(A_1, A_2) & 0.177 \\
\hline
P_{12}^\text{exact}(A_2, A_1) & P_{12}^\text{exact}(A_2, A_2) & 0.323 \\
\hline
P_{23}^\text{exact}(A_1, A_1) & P_{23}^\text{exact}(A_1, A_2) & 0.134 \\
\hline
P_{23}^\text{exact}(A_2, A_1) & P_{23}^\text{exact}(A_2, A_2) & 0.366 \\
\end{array}
\]

\[
\begin{array}{c|c|c}
P_{13}^\text{exact}(A_1, A_1) & P_{13}^\text{exact}(A_1, A_2) & 0.284 \\
\hline
P_{13}^\text{exact}(A_2, A_1) & P_{13}^\text{exact}(A_2, A_2) & 0.216 \\
\end{array}
\]

\[
\begin{array}{c|c|c}
\lambda_{12}^\text{Bethe}(A_1) & \lambda_{12}^\text{Bethe}(A_2) & 0.000 \\
\hline
\lambda_{12}^\text{Bethe}(A_2, A_1) & \lambda_{12}^\text{Bethe}(A_2, A_2) & 0.000 \\
\hline
\lambda_{13}^\text{Bethe}(A_1) & \lambda_{13}^\text{Bethe}(A_2) & 0.000 \\
\hline
\lambda_{13}^\text{Bethe}(A_2, A_1) & \lambda_{13}^\text{Bethe}(A_2, A_2) & 0.000 \\
\end{array}
\]

Test 4: Bethe Predicted Fields and Couplings

\[
\begin{array}{c|c|c}
\lambda_{12}^\text{Bethe}(A_1, A_1) & \lambda_{12}^\text{Bethe}(A_1, A_2) & -0.300 \\
\hline
\lambda_{12}^\text{Bethe}(A_2, A_1) & \lambda_{12}^\text{Bethe}(A_2, A_2) & 0.300 \\
\hline
\lambda_{23}^\text{Bethe}(A_1, A_1) & \lambda_{23}^\text{Bethe}(A_1, A_2) & -0.500 \\
\hline
\lambda_{23}^\text{Bethe}(A_2, A_1) & \lambda_{23}^\text{Bethe}(A_2, A_2) & 0.500 \\
\hline
\lambda_{13}^\text{Bethe}(A_1, A_1) & \lambda_{13}^\text{Bethe}(A_1, A_2) & 0.135 \\
\hline
\lambda_{13}^\text{Bethe}(A_2, A_1) & \lambda_{13}^\text{Bethe}(A_2, A_2) & -0.135 \\
\end{array}
\]
### Test 4: Bethe Predicted Marginals

<table>
<thead>
<tr>
<th>$P_{\text{Bethe}}(A_1)$</th>
<th>$P_{\text{Bethe}}(A_2)$</th>
<th>$P_{\text{Bethe}}(A_1)$</th>
<th>$P_{\text{Bethe}}(A_2)$</th>
<th>$P_{\text{Bethe}}(A_1)$</th>
<th>$P_{\text{Bethe}}(A_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{\text{Bethe}}(A_1)$</td>
<td>$P_{\text{Bethe}}(A_2)$</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
</tr>
<tr>
<td>$P_{\text{Bethe}}(A_2)$</td>
<td>$P_{\text{Bethe}}(A_2)$</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
</tr>
</tbody>
</table>

### Test 5: Original Fields and Couplings

<table>
<thead>
<tr>
<th>$\lambda_{\text{exact}}(A_1)$</th>
<th>$\lambda_{\text{exact}}(A_2)$</th>
<th>$\lambda_{\text{exact}}(A_1)$</th>
<th>$\lambda_{\text{exact}}(A_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_{\text{exact}}(A_1)$</td>
<td>$\lambda_{\text{exact}}(A_2)$</td>
<td>0.100</td>
<td>-0.100</td>
</tr>
<tr>
<td>$\lambda_{\text{exact}}(A_1)$</td>
<td>$\lambda_{\text{exact}}(A_2)$</td>
<td>0.100</td>
<td>-0.100</td>
</tr>
<tr>
<td>$\lambda_{\text{exact}}(A_1)$</td>
<td>$\lambda_{\text{exact}}(A_2)$</td>
<td>0.100</td>
<td>-0.100</td>
</tr>
</tbody>
</table>

Test 4: Bethe Predicted Marginals

Test 5: Original Fields and Couplings
\[
\begin{array}{c|c|c}
\text{Test 5: Original Marginals} & P_{1}^{\text{exact}}(A_1) & P_{1}^{\text{exact}}(A_2) \\
\hline
P_{2}^{\text{exact}}(A_1) & 0.542 & 0.458 \\
\hline
P_{3}^{\text{exact}}(A_1) & 0.542 & 0.466 \\
\hline
\end{array}
\]

\[
\begin{array}{c|c|c}
\text{Test 5: Bethe Predicted Fields and} & \lambda_{12}^{\text{Bethe}}(A_1, A_1) & \lambda_{12}^{\text{Bethe}}(A_1, A_2) \\
\hline
\lambda_{12}^{\text{Bethe}}(A_1, A_2) & 0.134 & -0.134 \\
\lambda_{12}^{\text{Bethe}}(A_2, A_1) & 0.134 & -0.134 \\
\lambda_{12}^{\text{Bethe}}(A_2, A_2) & -0.300 & 0.300 \\
\lambda_{23}^{\text{Bethe}}(A_1, A_1) & -0.500 & 0.500 \\
\lambda_{23}^{\text{Bethe}}(A_1, A_2) & 0.500 & -0.500 \\
\lambda_{23}^{\text{Bethe}}(A_2, A_1) & 0.500 & -0.500 \\
\end{array}
\]
<table>
<thead>
<tr>
<th>$P_{\text{Bethe}}(A_1)$</th>
<th>$P_{\text{Bethe}}(A_2)$</th>
<th>0.512</th>
<th>0.488</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{\text{Bethe}}^2(A_1)$</td>
<td>$P_{\text{Bethe}}^2(A_2)$</td>
<td>0.541</td>
<td>0.459</td>
</tr>
<tr>
<td>$P_{\text{Bethe}}^3(A_1)$</td>
<td>$P_{\text{Bethe}}^3(A_2)$</td>
<td>0.532</td>
<td>0.468</td>
</tr>
<tr>
<td>$P_{\text{Bethe}}^{12}(A_1, A_1)$</td>
<td>$P_{\text{Bethe}}^{12}(A_1, A_2)$</td>
<td>0.191</td>
<td>0.321</td>
</tr>
<tr>
<td>$P_{\text{Bethe}}^{12}(A_2, A_1)$</td>
<td>$P_{\text{Bethe}}^{12}(A_2, A_2)$</td>
<td>0.350</td>
<td>0.138</td>
</tr>
<tr>
<td>$P_{\text{Bethe}}^{23}(A_1, A_1)$</td>
<td>$P_{\text{Bethe}}^{23}(A_1, A_2)$</td>
<td>0.151</td>
<td>0.361</td>
</tr>
<tr>
<td>$P_{\text{Bethe}}^{23}(A_2, A_1)$</td>
<td>$P_{\text{Bethe}}^{23}(A_2, A_2)$</td>
<td>0.382</td>
<td>0.106</td>
</tr>
<tr>
<td>$P_{\text{Bethe}}^{13}(A_1, A_1)$</td>
<td>$P_{\text{Bethe}}^{13}(A_1, A_2)$</td>
<td>0.353</td>
<td>0.188</td>
</tr>
<tr>
<td>$P_{\text{Bethe}}^{13}(A_2, A_1)$</td>
<td>$P_{\text{Bethe}}^{13}(A_2, A_2)$</td>
<td>0.180</td>
<td>0.280</td>
</tr>
</tbody>
</table>

| $\lambda_{\text{exact}}^1(A_1)$ | $\lambda_{\text{exact}}^1(A_2)$ | 1.100 | −1.100 |
| $\lambda_{\text{exact}}^2(A_1)$ | $\lambda_{\text{exact}}^2(A_2)$ | 1.100 | −1.100 |
| $\lambda_{\text{exact}}^3(A_1)$ | $\lambda_{\text{exact}}^3(A_2)$ | 1.100 | −1.100 |
| $\lambda_{\text{exact}}^{12}(A_1, A_1)$ | $\lambda_{\text{exact}}^{12}(A_1, A_2)$ | −0.300 | 0.300 |
| $\lambda_{\text{exact}}^{12}(A_2, A_1)$ | $\lambda_{\text{exact}}^{12}(A_2, A_2)$ | 0.300 | −0.300 |
| $\lambda_{\text{exact}}^{23}(A_1, A_1)$ | $\lambda_{\text{exact}}^{23}(A_1, A_2)$ | −0.500 | 0.500 |
| $\lambda_{\text{exact}}^{23}(A_2, A_1)$ | $\lambda_{\text{exact}}^{23}(A_2, A_2)$ | 0.500 | −0.500 |
| $\lambda_{\text{exact}}^{13}(A_1, A_1)$ | $\lambda_{\text{exact}}^{13}(A_1, A_2)$ | 0.000 | 0.000 |
| $\lambda_{\text{exact}}^{13}(A_2, A_1)$ | $\lambda_{\text{exact}}^{13}(A_2, A_2)$ | 0.000 | 0.000 |

Test 5: Bethe Predicted Marginals

Test 6: Original Fields and Couplings
\[
\begin{array}{c|c|c}
\text{Test 6: Original Marginals} & & \\
\hline
P_{exact}^1(A_1) & P_{exact}^1(A_2) & 0.721 \\
P_{exact}^2(A_1) & P_{exact}^2(A_2) & 0.863 \\
P_{exact}^3(A_1) & P_{exact}^3(A_2) & 0.822 \\
\end{array}
\]

\[
\begin{array}{c|c|c|c|c}
\text{Test 6: Bethe Predicted Fields} & & & & \\
\hline
\lambda_{Bethe}^{12}(A_1, A_1) & \lambda_{Bethe}^{12}(A_1, A_2) & -0.300 & 0.300 \\
\lambda_{Bethe}^{12}(A_2, A_1) & \lambda_{Bethe}^{12}(A_2, A_2) & 0.300 & -0.300 \\
\lambda_{Bethe}^{23}(A_1, A_1) & \lambda_{Bethe}^{23}(A_1, A_2) & -0.500 & 0.500 \\
\lambda_{Bethe}^{23}(A_2, A_1) & \lambda_{Bethe}^{23}(A_2, A_2) & 0.500 & -0.500 \\
\lambda_{Bethe}^{13}(A_1, A_1) & \lambda_{Bethe}^{13}(A_1, A_2) & 0.059 & -0.059 \\
\lambda_{Bethe}^{13}(A_2, A_1) & \lambda_{Bethe}^{13}(A_2, A_2) & -0.059 & 0.059 \\
\end{array}
\]
Test 6: Bethe Predicted Marginals

<table>
<thead>
<tr>
<th>( P_{\text{Bethe}}^1(A_1) )</th>
<th>( P_{\text{Bethe}}^1(A_2) )</th>
<th>( P_{\text{Bethe}}^2(A_1) )</th>
<th>( P_{\text{Bethe}}^2(A_2) )</th>
<th>( P_{\text{Bethe}}^3(A_1) )</th>
<th>( P_{\text{Bethe}}^3(A_2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.720</td>
<td>0.280</td>
<td>0.862</td>
<td>0.138</td>
<td>0.822</td>
<td>0.178</td>
</tr>
</tbody>
</table>

Test 7: Original Fields and Couplings

<table>
<thead>
<tr>
<th>( \lambda_1^{\text{exact}}(A_1) )</th>
<th>( \lambda_1^{\text{exact}}(A_2) )</th>
<th>( \lambda_2^{\text{exact}}(A_1) )</th>
<th>( \lambda_2^{\text{exact}}(A_2) )</th>
<th>( \lambda_3^{\text{exact}}(A_1) )</th>
<th>( \lambda_3^{\text{exact}}(A_2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.100</td>
<td>-2.100</td>
<td>2.100</td>
<td>-2.100</td>
<td>2.100</td>
<td>-2.100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \lambda_{12}^{\text{exact}}(A_1, A_1) )</th>
<th>( \lambda_{12}^{\text{exact}}(A_1, A_2) )</th>
<th>( \lambda_{12}^{\text{exact}}(A_2, A_1) )</th>
<th>( \lambda_{12}^{\text{exact}}(A_2, A_2) )</th>
<th>( \lambda_{23}^{\text{exact}}(A_1, A_1) )</th>
<th>( \lambda_{23}^{\text{exact}}(A_1, A_2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>-0.300</td>
<td>-0.500</td>
<td>0.500</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \lambda_{13}^{\text{exact}}(A_1, A_1) )</th>
<th>( \lambda_{13}^{\text{exact}}(A_1, A_2) )</th>
<th>( \lambda_{13}^{\text{exact}}(A_2, A_1) )</th>
<th>( \lambda_{13}^{\text{exact}}(A_2, A_2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>( P_{1 \text{exact}}(A_1) )</td>
<td>( P_{1 \text{exact}}(A_2) )</td>
<td>0.934</td>
<td>0.066</td>
</tr>
<tr>
<td>( P_{2 \text{exact}}(A_1) )</td>
<td>( P_{2 \text{exact}}(A_2) )</td>
<td>0.975</td>
<td>0.025</td>
</tr>
<tr>
<td>( P_{3 \text{exact}}(A_3) )</td>
<td>( P_{3 \text{exact}}(A_2) )</td>
<td>0.963</td>
<td>0.037</td>
</tr>
</tbody>
</table>

| \( P_{12 \text{exact}}(A_1, A_1) \) | \( P_{12 \text{exact}}(A_1, A_2) \) | 0.909 | 0.025 |
| \( P_{12 \text{exact}}(A_2, A_1) \) | \( P_{12 \text{exact}}(A_2, A_2) \) | 0.065 | 0.001 |
| \( P_{23 \text{exact}}(A_1, A_1) \) | \( P_{23 \text{exact}}(A_1, A_2) \) | 0.898 | 0.037 |
| \( P_{23 \text{exact}}(A_2, A_1) \) | \( P_{23 \text{exact}}(A_2, A_2) \) | 0.065 | 0.000 |
| \( P_{13 \text{exact}}(A_1, A_1) \) | \( P_{13 \text{exact}}(A_1, A_2) \) | 0.939 | 0.036 |
| \( P_{13 \text{exact}}(A_2, A_1) \) | \( P_{13 \text{exact}}(A_2, A_2) \) | 0.024 | 0.001 |

| \( \lambda_{1 \text{Bethe}}^\text{Bethe}(A_1) \) | \( \lambda_{1 \text{Bethe}}^\text{Bethe}(A_2) \) | 2.098 | -2.098 |
| \( \lambda_{2 \text{Bethe}}^\text{Bethe}(A_1) \) | \( \lambda_{2 \text{Bethe}}^\text{Bethe}(A_2) \) | 2.090 | -2.090 |
| \( \lambda_{3 \text{Bethe}}^\text{Bethe}(A_1) \) | \( \lambda_{3 \text{Bethe}}^\text{Bethe}(A_2) \) | 2.089 | -2.089 |

and Couplings

<p>| ( \lambda_{12 \text{Bethe}}(A_1, A_1) ) | ( \lambda_{12 \text{Bethe}}(A_1, A_2) ) | -0.300 | 0.300 |
| ( \lambda_{12 \text{Bethe}}(A_2, A_1) ) | ( \lambda_{12 \text{Bethe}}(A_2, A_2) ) | 0.300 | -0.300 |
| ( \lambda_{23 \text{Bethe}}(A_1, A_1) ) | ( \lambda_{23 \text{Bethe}}(A_1, A_2) ) | -0.499 | 0.499 |
| ( \lambda_{23 \text{Bethe}}(A_2, A_1) ) | ( \lambda_{23 \text{Bethe}}(A_2, A_2) ) | 0.499 | -0.499 |
| ( \lambda_{13 \text{Bethe}}(A_1, A_1) ) | ( \lambda_{13 \text{Bethe}}(A_1, A_2) ) | 0.011 | -0.011 |
| ( \lambda_{13 \text{Bethe}}(A_2, A_1) ) | ( \lambda_{13 \text{Bethe}}(A_2, A_2) ) | -0.011 | 0.011 |</p>
<table>
<thead>
<tr>
<th>$P_{\text{Bethe}}(A_1)$</th>
<th>$P_{\text{Bethe}}(A_2)$</th>
<th>(0.934)</th>
<th>(0.066)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{\text{Bethe}}(A_1)$</td>
<td>$P_{\text{Bethe}}(A_2)$</td>
<td>(0.975)</td>
<td>(0.025)</td>
</tr>
<tr>
<td>$P_{\text{Bethe}}(A_1)$</td>
<td>$P_{\text{Bethe}}(A_2)$</td>
<td>(0.963)</td>
<td>(0.037)</td>
</tr>
</tbody>
</table>

Test 7: Bethe Predicted Marginals

<table>
<thead>
<tr>
<th>$P_{\text{Bethe}}(A_1, A_1)$</th>
<th>$P_{\text{Bethe}}(A_1, A_2)$</th>
<th>(0.909)</th>
<th>(0.025)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{\text{Bethe}}(A_1, A_1)$</td>
<td>$P_{\text{Bethe}}(A_1, A_2)$</td>
<td>(0.065)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>$P_{\text{Bethe}}(A_1, A_1)$</td>
<td>$P_{\text{Bethe}}(A_1, A_2)$</td>
<td>(0.898)</td>
<td>(0.037)</td>
</tr>
<tr>
<td>$P_{\text{Bethe}}(A_1, A_1)$</td>
<td>$P_{\text{Bethe}}(A_1, A_2)$</td>
<td>(0.065)</td>
<td>(0.000)</td>
</tr>
<tr>
<td>$P_{\text{Bethe}}(A_1, A_1)$</td>
<td>$P_{\text{Bethe}}(A_1, A_2)$</td>
<td>(0.939)</td>
<td>(0.036)</td>
</tr>
<tr>
<td>$P_{\text{Bethe}}(A_1, A_1)$</td>
<td>$P_{\text{Bethe}}(A_1, A_2)$</td>
<td>(0.024)</td>
<td>(0.001)</td>
</tr>
</tbody>
</table>
References


[59] Daniel Y. Little and Lu Chen. Identification of coevolving residues and


[81] Florencio Pazos and Alfonso Valencia. In silico two-hybrid system for


[102] Robert W. Shafer, Son-Yon Rhee, Deenan Pillay, Veronica Miller, Paul Sandstrom, Jonathan M. Schapiro, Daniel R. Kuritzkes, and Diane Ben-


[113] Françoise Brun-Vézinet MD PhD Bonaventura Clotet MD PhD Huldrych


Appendix A

Operating Manual For Code

A.1 The Forward Problem

Problem Statement: Given field and coupling parameters of a completely con- 
nected graph, determine the univariate and bivariate marginals.

File names

mpforward_bethe.c

vars.c

Compiling

gcc -lm -Wall params.c mpforward_bethe.c -o mpforward_bethe

Usage

mp_forward_bethe npos nletters < parametersFile

where npos is the number of positions $N$, nletters is the number of letters $k$ and parametersFile is the file containing the field and coupling parameters.
**Input file format**

The input consists of field and coupling parameters in the input file called `parametersFile`. The file can take any name. The file consists of the numerical values of the field and coupling parameters, \( \lambda_i \)'s and \( \lambda_{ij} \)'s. For \( N \) positions and \( k \) letters per position, there will be \( k \) field parameters per position and \( k^2 \) coupling parameters for each pair of positions. The format for this file has to be the field parameters followed by the coupling. The field parameters should take this format:

\[
\begin{array}{cccc}
\lambda^1_1 & \lambda^2_1 & \cdots & \lambda^k_1 \\
\lambda^1_2 & \lambda^2_2 & \cdots & \lambda^k_2 \\
\vdots & \vdots & \ddots & \vdots \\
\lambda^1_i & \lambda^2_i & \cdots & \lambda^k_i \\
\vdots & \vdots & \ddots & \vdots \\
\lambda^1_N & \lambda^2_N & \cdots & \lambda^k_N \\
\end{array}
\]

where the symbol \( \lambda^c_i \) represents the field parameter of the \( i \)th position taking on the \( c \)th state. The coupling parameters have to be written into the input file as
well and should follow right after the field parameters and take this format.

\[
\begin{align*}
\lambda_{12}^{11} & \quad \lambda_{12}^{12} & \cdots & \quad \lambda_{12}^{1k} \\
\lambda_{12}^{21} & \quad \lambda_{12}^{22} & \cdots & \quad \lambda_{12}^{2k} \\
\vdots & \quad \vdots & \ddots & \quad \vdots \\
\lambda_{12}^{k1} & \quad \lambda_{12}^{k2} & \cdots & \quad \lambda_{12}^{kk} \\
\lambda_{13}^{11} & \quad \lambda_{13}^{12} & \cdots & \quad \lambda_{13}^{1k} \\
\lambda_{13}^{21} & \quad \lambda_{13}^{22} & \cdots & \quad \lambda_{13}^{2k} \\
\vdots & \quad \vdots & \ddots & \quad \vdots \\
\lambda_{13}^{k1} & \quad \lambda_{13}^{k2} & \cdots & \quad \lambda_{13}^{kk} \\
\vdots & \quad \vdots & \ddots & \quad \vdots \\
\lambda_{1N}^{11} & \quad \lambda_{1N}^{12} & \cdots & \quad \lambda_{1N}^{1k} \\
\lambda_{1N}^{21} & \quad \lambda_{1N}^{22} & \cdots & \quad \lambda_{1N}^{2k} \\
\vdots & \quad \vdots & \ddots & \quad \vdots \\
\lambda_{1N}^{k1} & \quad \lambda_{1N}^{k2} & \cdots & \quad \lambda_{1N}^{kk} \\
\vdots & \quad \vdots & \ddots & \quad \vdots 
\end{align*}
\]

where the symbol \(\lambda_{ij}^{cd}\) represents the coupling parameter between positions \(i\) and \(j\), with states \(c\) and \(d\) respectively. Since this is an undirected graph, \(\lambda_{ij}^{cd} = \lambda_{ji}^{dc}\).

**Output**

The output is the Bethe approximated set of univariate and bivariate marginals following by some additional information about the convergence properties of this particular algorithm. The output format is similar to the input format. It consists of the univariate marginals followed by the bivariate
marginals. The following is the output format for the univariate marginals.

\[
\begin{array}{cccc}
P_1^0 & P_2^0 & \cdots & P_k^0 \\
P_1^1 & P_2^1 & \cdots & P_k^1 \\
\vdots & \vdots & \ddots & \vdots \\
P_1^i & P_2^i & \cdots & P_k^i \\
\vdots & \vdots & \ddots & \vdots \\
P_1^N & P_2^N & \cdots & P_k^N \\
\end{array}
\]

where the symbol \( P_{ci} \) represents the univariate marginals of the \( c \)th state of position \( i \).

\[
\begin{array}{cccc}
P_{11}^{11} & P_{12}^{12} & \cdots & P_{12}^{1k} \\
P_{12}^{21} & P_{12}^{22} & \cdots & P_{12}^{2k} \\
\vdots & \vdots & \ddots & \vdots \\
P_{12}^{k1} & P_{12}^{k2} & \cdots & P_{12}^{kk} \\
P_{13}^{11} & P_{13}^{12} & \cdots & P_{13}^{1k} \\
P_{13}^{21} & P_{13}^{22} & \cdots & P_{13}^{2k} \\
\vdots & \vdots & \ddots & \vdots \\
P_{13}^{k1} & P_{13}^{k2} & \cdots & P_{13}^{kk} \\
P_{1N}^{11} & P_{1N}^{12} & \cdots & P_{1N}^{1k} \\
P_{1N}^{21} & P_{1N}^{22} & \cdots & P_{1N}^{2k} \\
\vdots & \vdots & \ddots & \vdots \\
P_{1N}^{k1} & P_{1N}^{k2} & \cdots & P_{1N}^{kk} \\
\end{array}
\]

where the symbol \( P_{cd}^{ij} \) represents the bivariate marginals of the states \( cd \) at positions \( i \) and \( j \).
A.2 The Inverse Problem

Problem Statement: Given the univariate and bivariate marginals, determine the field and coupling parameters that preserve the pair correlations within the alignment.

File names

mpinverse_bethe.c

vars.c

Compiling

gcc -lm -Wall params.c mpinverse_bethe.c -o mpinverse_bethe

Usage

mp_foward_bethe npos nletters epsilon iterations < marginalsFile

where npos is the number of positions \( N \), nletters is the number of letters \( k \), epsilon is the gradient descent parameter, iterations are the maximum iterations of inverse belief propagation in the Bethe approximation that one wants to do, and marginalsFile is the file containing the field and coupling parameters.

Input file format

The input consists of univariate and bivariate marginals in the input file called marginals. The file can take any name of course. The file consists of the numerical values of the univariate and bivariate marginals, \( P_i \)'s and \( P_{ij} \)'s. For \( N \) positions and \( k \) letters per position, there will be \( k \) univariate marginals per position and \( k^2 \) bivariate marginals for every pair of positions. The format for this file has to be the univariate marginals followed by the bivariate marginals.
The univariate marginals should take this format.

\[
\begin{array}{cccc}
P_0^1 & P_0^2 & \cdots & P_0^k \\
P_1^1 & P_1^2 & \cdots & P_1^k \\
\vdots & \vdots & \ddots & \vdots \\
P_i^1 & P_i^2 & \cdots & P_i^k \\
\vdots & \vdots & \ddots & \vdots \\
P_N^1 & P_N^2 & \cdots & P_N^k \\
\end{array}
\]

where the symbol \( P_{i}^{c} \) represents the univariate marginal of the \( i \)th position taking on the \( c \)th state. The bivariate marginals have to be written into the input file as well and should follow right after the univariate marginals and take this format.

\[
\begin{array}{cccc}
P_{12}^{11} & P_{12}^{12} & \cdots & P_{12}^{1k} \\
P_{12}^{21} & P_{12}^{22} & \cdots & P_{12}^{2k} \\
\vdots & \vdots & \ddots & \vdots \\
P_{12}^{k1} & P_{12}^{k2} & \cdots & P_{12}^{kk} \\
P_{13}^{11} & P_{13}^{12} & \cdots & P_{13}^{1k} \\
P_{13}^{21} & P_{13}^{22} & \cdots & P_{13}^{2k} \\
\vdots & \vdots & \ddots & \vdots \\
P_{13}^{k1} & P_{13}^{k2} & \cdots & P_{13}^{kk} \\
P_{1N}^{11} & P_{1N}^{12} & \cdots & P_{1N}^{1k} \\
P_{1N}^{21} & P_{1N}^{22} & \cdots & P_{1N}^{2k} \\
\vdots & \vdots & \ddots & \vdots \\
P_{1N}^{k1} & P_{1N}^{k2} & \cdots & P_{1N}^{kk} \\
\end{array}
\]

where the symbol \( P_{ij}^{cd} \) represents the bivariate marginal between positions \( i \) and \( j \), with states \( c \) and \( d \) respectively. Since this is an undirected graph, \( P_{ij}^{cd} = P_{ji}^{dc} \).
Output

The output is the Bethe approximated set of field and coupling parameters followed by some additional information about the convergence properties of this particular algorithm. The output format is similar to the input format. It consists of the fields followed by the couplings. The following is the output format for the fields.

\[
\begin{align*}
\lambda_1^1 & \quad \lambda_1^2 & \quad \cdots & \quad \lambda_1^k \\
\lambda_2^1 & \quad \lambda_2^2 & \quad \cdots & \quad \lambda_2^k \\
& \vdots & & \vdots \\
\lambda_i^1 & \quad \lambda_i^2 & \quad \cdots & \quad \lambda_i^k \\
& \vdots & & \vdots \\
\lambda_N^1 & \quad \lambda_N^2 & \quad \cdots & \quad \lambda_N^k \\
\end{align*}
\]

where the symbol \( \lambda_i^c \) represents the field parameter for \( c \)th state of position \( i \).
where the symbol $\lambda_{ij}^{cd}$ represents the coupling between the states $cd$ at positions $i$ and $j$. 