A SNAPSHOT OF THE *ARTEMIA* GENOME – TO CODE OR NOT TO CODE

By

STACEY LYNN WITTIG

A thesis submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

And

The Graduate School of Biomedical Sciences

University of Medicine and Dentistry of New Jersey

In partial fulfillment of the requirements

For the degree of

Master of Science

Graduate Program in Microbiology and Molecular Genetics

Written under the direction of

Dr. Andrew K. Vershon

And approved by

____

____

____

New Brunswick, New Jersey

October, 2009

ABSTRACT OF THE THESIS

A Snapshot of the *Artemia* Genome – To Code or Not to Code

By

STACEY LYNN WITTIG

Thesis Director

Dr. Andrew K. Vershon

The Waksman Student Scholars Program, along with the Introduction to Molecular Biology and Biochemical Research class, were responsible for the publication of 628 *Artemia* sequences. Surprisingly, 361 of these sequences (58%) did not contain an open reading frame larger than 80 residues. It was originally presumed that this was due to a high level of genomic DNA contamination. While it is possible that some of our *Artemia* sequences are genomic contamination, I believe a large majority of our non-coding sequences are long non-coding RNA (ncRNA), newly recognized players in transcriptional regulation. This high percentage of non-coding sequences is reasonable, as other genomic studies indicate about 50% of an organism's RNA is non-coding. Our average non-coding sequence length was 600nt, significantly longer than our average *Artemia* 3'UTR length of 175nt, which can easily be explained if we acknowledge these

sequences as long non-coding RNAs. Many of our non-coding RNAs also contain polyA tails, as well as polyadenylation signals. Considering many ncRNAs are polyadenylated, this data supports my hypothesis. Fifty-two percent of our non-coding sequences match other *Artemia* sequences in NCBI, and of these matches, 33% are in the reverse direction. Transcription in the reverse direction is a method used by ncRNA to inhibit gene transcription.

In addition to my analysis of the 628 analyzed *Artemia* sequences, I used DNASTAR software to analyze all 5,947 *Artemia* sequences generated from 2005 through 2008. This software validated sequence quality and assembled similar sequences into 2,848 contiguous sequences. These contiguous sequences were further processed using Blast2GO, a gene ontology tool, where only 268 contiguous sequences were of high enough quality to be considered annotated genes. These genes were further characterized according to their Gene Ontology.

ACKNOWLEGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor Drew, for his support throughout the last five years. I am indebted to him for his flexibility, considering I earned this degree while still working a full-time job. Despite the time constraints, he would patiently encourage me to continue with my research, even when I felt the end was so far away. His insight and enthusiasm aided the writing of this thesis in innumerable ways. He is a positive, warm-hearted mentor, and I was fortunate to have him as my advisor.

I am also grateful to my committee members, Todd Michael and Samuel Gunderson, for agreeing to support my continued education. Even though some of my meetings with Todd were less than efficient, working with him was truly a pleasure and I consider him to be an excellent mentor and teacher.

I owe my deepest gratitude to my family for their unwavering support and encouragement in everything I do, but especially for their support during my graduate years. They have always had faith in my ability to reach any goal I set for myself. That kind of unfaltering love has molded me into the woman I am today, and for that, I can never thank them enough.

Finally, I would like to thank my husband Bob. He has been a source of incredible love and support throughout the past three years. He was my sounding board for this thesis, when I thought no one would want to listen. I am especially appreciative of his patience, considering how frequently this degree kept me away from times we could be spending together. His faith in me is unsurpassed and I will always love him dearly for that. He is my true companion.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. General Introduction

I.1. All About Artemia

Artemia franciscana, commonly known as brine shrimp or sea monkeys, is one of the most primitive crustaceans alive today, evolving very little from their ancestors of 250 million years ago [1]. They exist in extreme environments of high salinity (30ppt) and abnormal pH (8-9), which effectively allows them to avoid their predators [5]. *Artemia* are filter feeders, and enjoy diets of algae, bacteria and detritus, which in turn determine their color. They are incapable of active dispersion, and subsequently rely on wind, waves and waterfowl to carry them to new surroundings.

Artemia can breed through sexual reproduction or via parthenogenesis if little or no males are present. Fertilized eggs can either develop into free-swimming nauplius larvae, or if living conditions are poor, they can be surrounded by a protective shell (cyst) and they hatch when living conditions improve (Figure 1). Females can reproduce at a rate of 300 nauplii/cysts every 4 days [3]. *Artemia* can grow from nauplius to adult in eight days [3] and grow to +/- 1cm in length, depending on the sex of the organism [6]. When in favorable environments, *Artemia* can live for several months [3].

During periods of stress, such as higher salinity, higher temperature, desiccation, food shortage or O_2 fluctuations, cysts are formed instead of living nauplii. These cysts, comprised of about 4000 cells, will show little signs of metabolism and energy consumption and can stay dormant for decades until favorable conditions present themselves [7]. This discontinuous development, known as diapause, ensures their survival. These cysts are incredibly resilient and can withstand ionizing radiation,

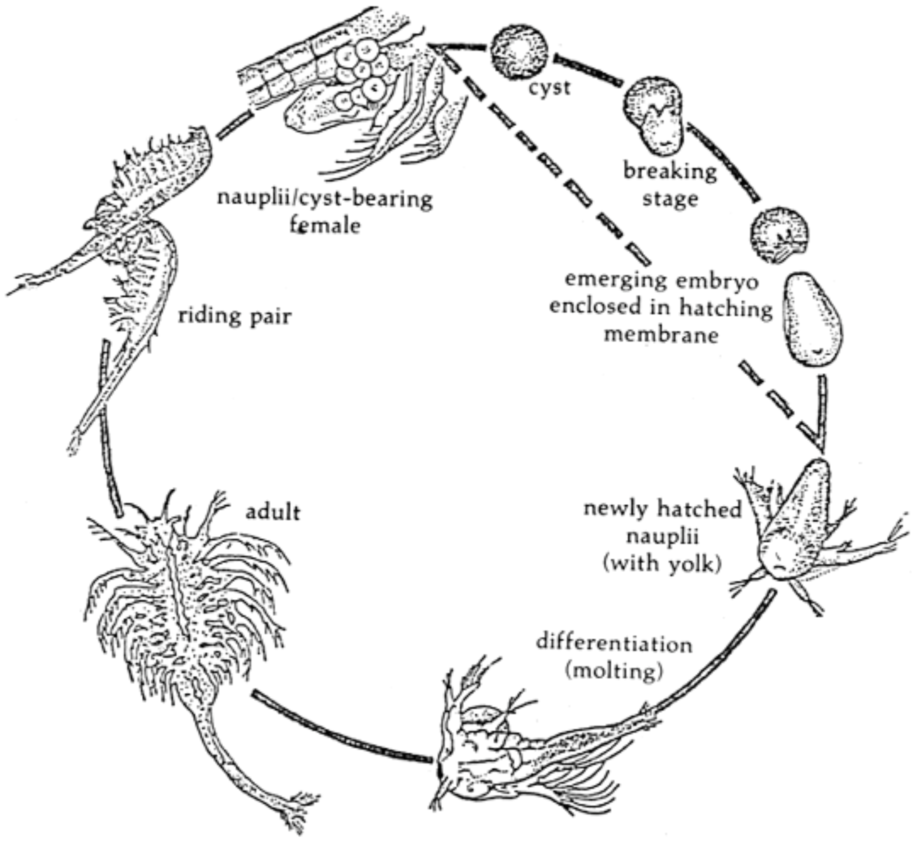

nauplii/cyst-bearing female
cyst
breaking stage
emerging embryo enclosed in hatching membrane
riding pair
newly hatched nauplii (with yolk)
adult
differentiation (molting)

Figure 1. **Artemia lifecycle**.
Picture taken from **Manual on the Production and Use of Live Food for Aquaculture** [3]

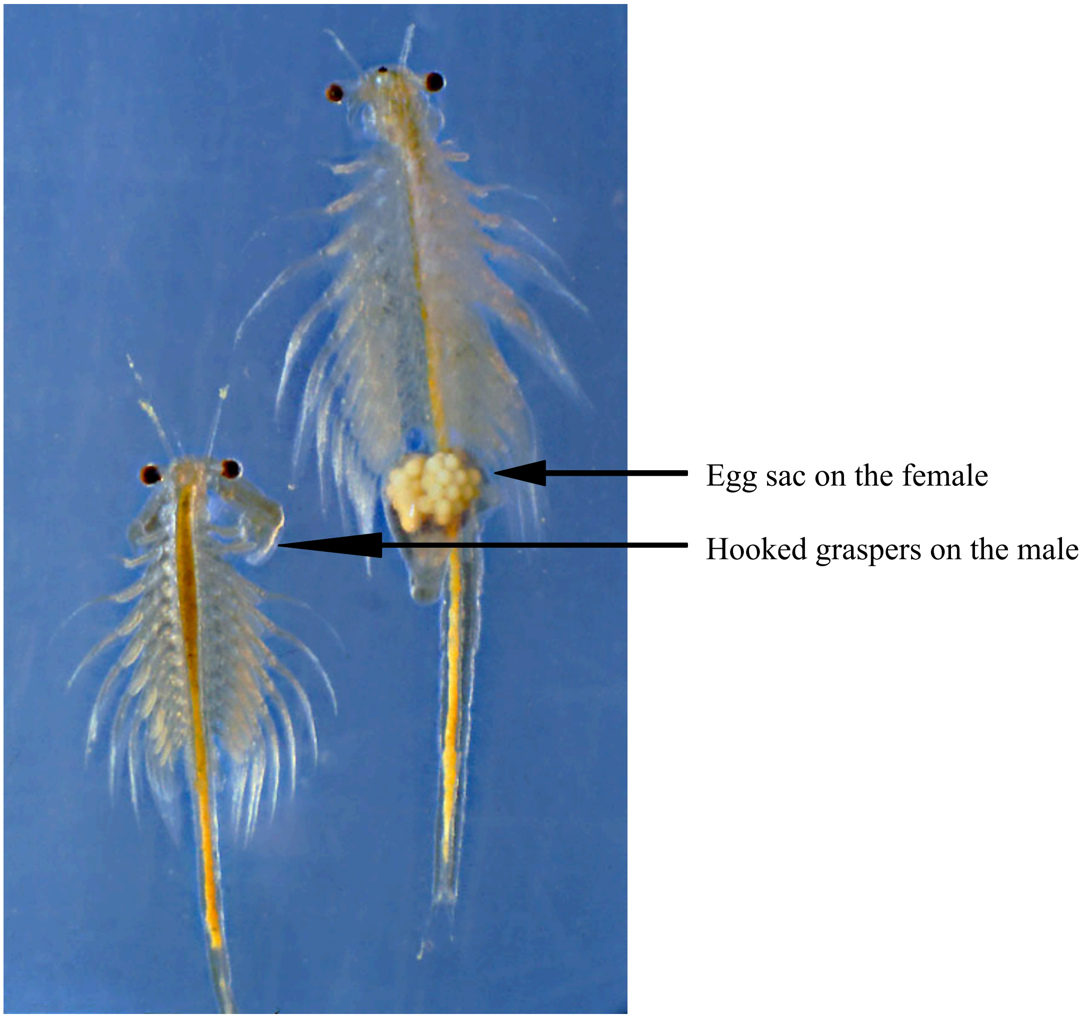


Figure 2. Male and female *Artemia*.
Note the claspers identifying the male, and the egg sac identifying the female. [4]

extreme drying, vacuum, and temperature fluctuations between -273°C and +90°C [3]. Dehydrated cysts are often stored in a vacuum or in nitrogen gas, as oxygen exposure results in detrimental free radicals that decrease the cysts' viability.

These calorie-rich organisms are an excellent food source for industrial fish and shrimp, as they are high in lipids, unsaturated fatty acids, and protein. They have become one of the more favorable forms of aquarium fodder, as *Artemia* cysts can be purchased and stored until needed, at which point they can easily be processed in order to continue development into edible nauplii. Currently, over 2000 metric tons of dry *Artemia* cysts are sold every year [3], constituting the most widely used live aquaculture diets. While *Artemia* can be found in salty waters throughout the Americas, Asia, Europe and Australia, about 90% of marketed brine shrimp come from the Great Salt Lake in Utah [3]. Due to their commercial significance, it is important to develop a thorough understanding of this organism's genome, metabolism, and development.

I.2. Artemia in Research

Artemia are a suitable organism for laboratory analysis due to their commercial availability, low cost and small size. The ability to synchronize larvae development in the lab is also advantageous and allows for separation of organisms based on particular life cycle stages. However, what makes these organisms most unique is their ability to exist long periods of dormancy (diapause), as well as their ability to survive prolonged periods without oxygen (anoxia). In 2005, only 246 *Artemia* sequences were published in the EST database in NCBI, making it a favorable organism for novel genetic analysis.

For this reason, they were used as the model eukaryotic organism for the Waksman Student Scholars Program (WSSP) as well as the Rutgers University undergraduate course Introduction to Molecular Biology and Biochemical Research. The Waksman Student Scholars Program is a year-long program designed to help high school students learn modern molecular genetics by engaging them in genuine scientific research projects. From 2005 through 2008, these students isolated and sequenced cDNA clones from *Artemia* to further analyze this unique organism's life cycle. The students were also interested in completing expression studies to identify novel genes associated with *Artemia's* ability to undergo diapause, as well as its ability to exist in such extreme environments.

Over 3000 clones were generated throughout these four years, but only about 20% of them were individually analyzed by the students. No comparison had been made across the entire dataset. This thesis describes the analysis of the data from this project.

II. Construction of cDNA library

II.1. Extraction of mRNA

A mixed age population of male and female *Artemia* was obtained from Petland Discounts on Stelton Road in Piscataway, New Jersey. Dr. Marty Nemeroff harvested 0.5 grams of the nauplii by filtering them on a 0.22 μm filter. They were then homogenized by freezing with liquid N_2 in a sterile mortar and then lysed with a sterile pestle. RA1 and β-mercaptoethanol were added and the cells were pulled through a 16 gauge and then a 21 gauge syringe needle to further lyse the cells and shear the DNA. This solution was transferred to a NucleoSpin filter where it was subsequently spun to filter the lysate. The flow-through was transferred to microfuge tubes where it was washed with ethanol and briefly vortexed to adjust for RNA binding conditions. The RNA was then loaded onto another NucleoSpin column and the flow-through was discarded. MDB was added to the column to desalt the membrane, and DNase was added to digest any remaining DNA. The column was then washed and dried using RA2 buffer, as well as RA3 buffer. All flowthrough was discarded. The sample was eluted into a clean tube by the addition of RNase-free H_2O.

II.2. Synthesis of cDNA Library

The cDNA was made using the Creator SMART cDNA Library Construction Kit by Clontech. 1 μl of *Artemia* mRNA was combined with Smart IV Oligonucleotides containing an *Sfi*IA site and CDS III 3'PCR primer containing an *Sfi*IB site. This was then mixed before adding the First-strand buffer, DTT, dNTP mix and PowerScript Reverse Transcriptase used to create the first strand of the cDNA. After one hour of incubation, 2 μl of the mix was removed and added to a PCR tube with Advantge 2 PCR buffer, dNTP mix, 5’PCR primers, CDS III 3’PCR primers and Advantage 2 Polimerase mix containing Taq (Figure 3) and amplified for 20 cycles. 50 μl of the sample was transferred to a microfuge tube, where Proteinase K was added in order to inactivate DNA polymerase activity and phenolchloroform was added to separate the cDNA from any remaining proteins, primers, or dNTPs. The cDNA was precipated with ethanol, dried and suspended in dedeionized water. The cDNA was then digested with *Sfi*I and cloned into an *SfiI* digested pTriplEX2 vector (Figure 4). The vectors were transformed into DH5α competent *E. coli* cells and plate amplified.

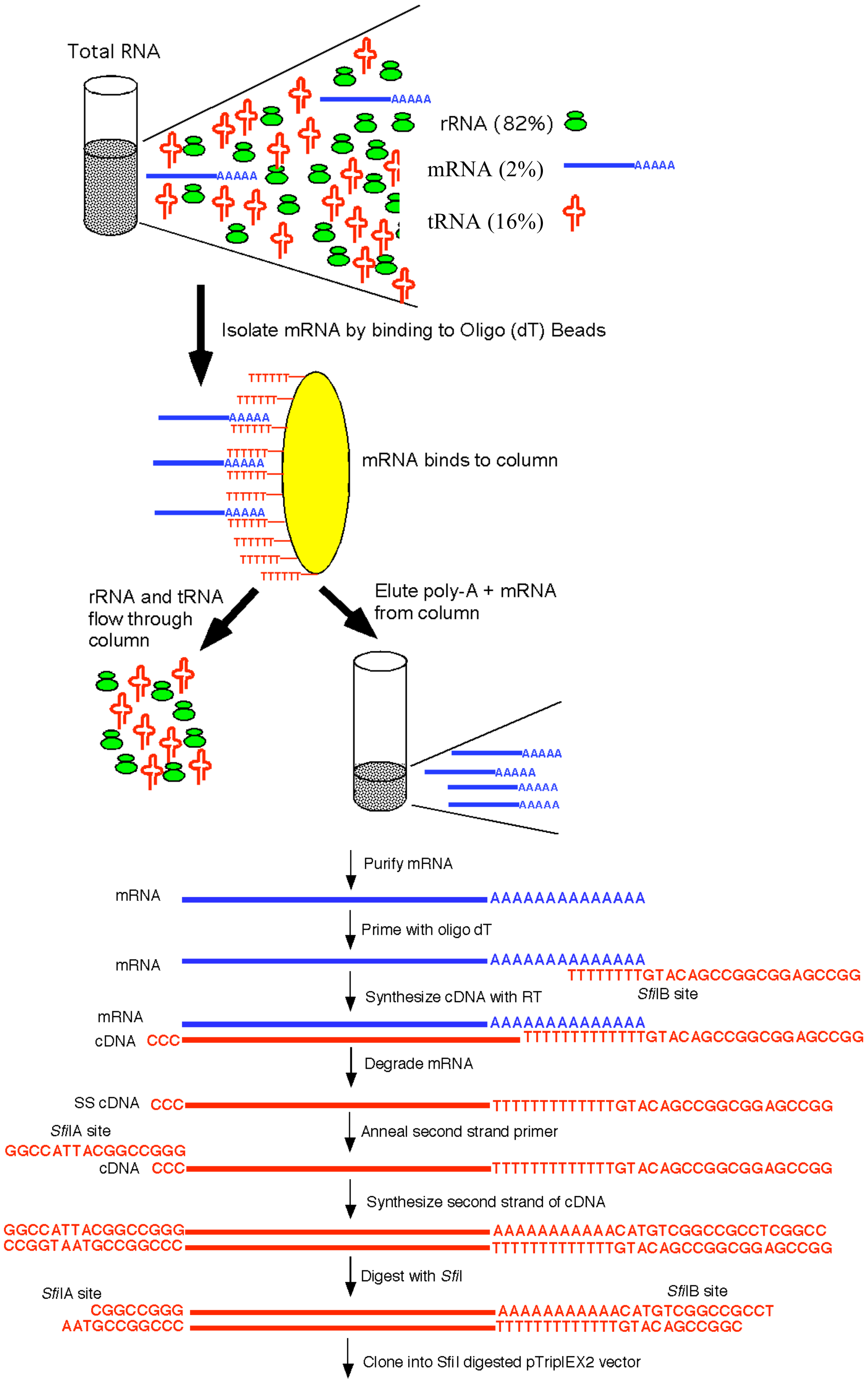

Total RNA
AAAAA
rRNA (82%)
mRNA (2%)
AAAAA
tRNA (16%)
Isolate mRNA by binding to Oligo (dT) Beads
TTTTTT
AAAAA
mRNA binds to column
rRNA and tRNA flow through column
Elute poly-A + mRNA from column
AAAAA
Purify mRNA
mRNA
AAAAAAAAAAAAAA
Prime with oligo dT
mRNA
AAAAAAAAAAAAAA
TTTTTTTTGTACAGCCGGCGGAGCCGG
Synthesize cDNA with RT
SfiIB site
mRNA
AAAAAAAAAAAAAA
cDNA CCC
TTTTTTTTTTTTTGTACAGCCGGCGGAGCCGG
Degrade mRNA
SS cDNA CCC
TTTTTTTTTTTTTGTACAGCCGGCGGAGCCGG
SfiIA site
GGCCATTACGGCCGGG
Anneal second strand primer
cDNA CCC
TTTTTTTTTTTTTGTACAGCCGGCGGAGCCGG
Synthesize second strand of cDNA
GGCCATTACGGCCGGG
AAAAAAAAAAACATGTCGGCCGCCTCGGCC
CCGGTAATGCCGGCCC
TTTTTTTTTTTTTGTACAGCCGGCGGAGCCGG
Digest with SfiI
SfiIA site
SfiIB site
CGGCCGGG
AAAAAAAAAAACATGTCGGCCGCCT
AATGCCGGCCC
TTTTTTTTTTTTTGTACAGCCGGC
Clone into SfiI digested pTriplEX2 vector

Figure 3. Purification of polyA containing mRNA and synthesis of cDNA from mRNA. The total RNA was passed over an oligo-dT column in order to select for the mRNA. Primers containing either an *SfiIA* site or an *SfiIB* site with a polyT run were used to create the cDNA and ensured directional cloning into the pTriplEX2 vector. Figure from Vershon, 2008 [2].

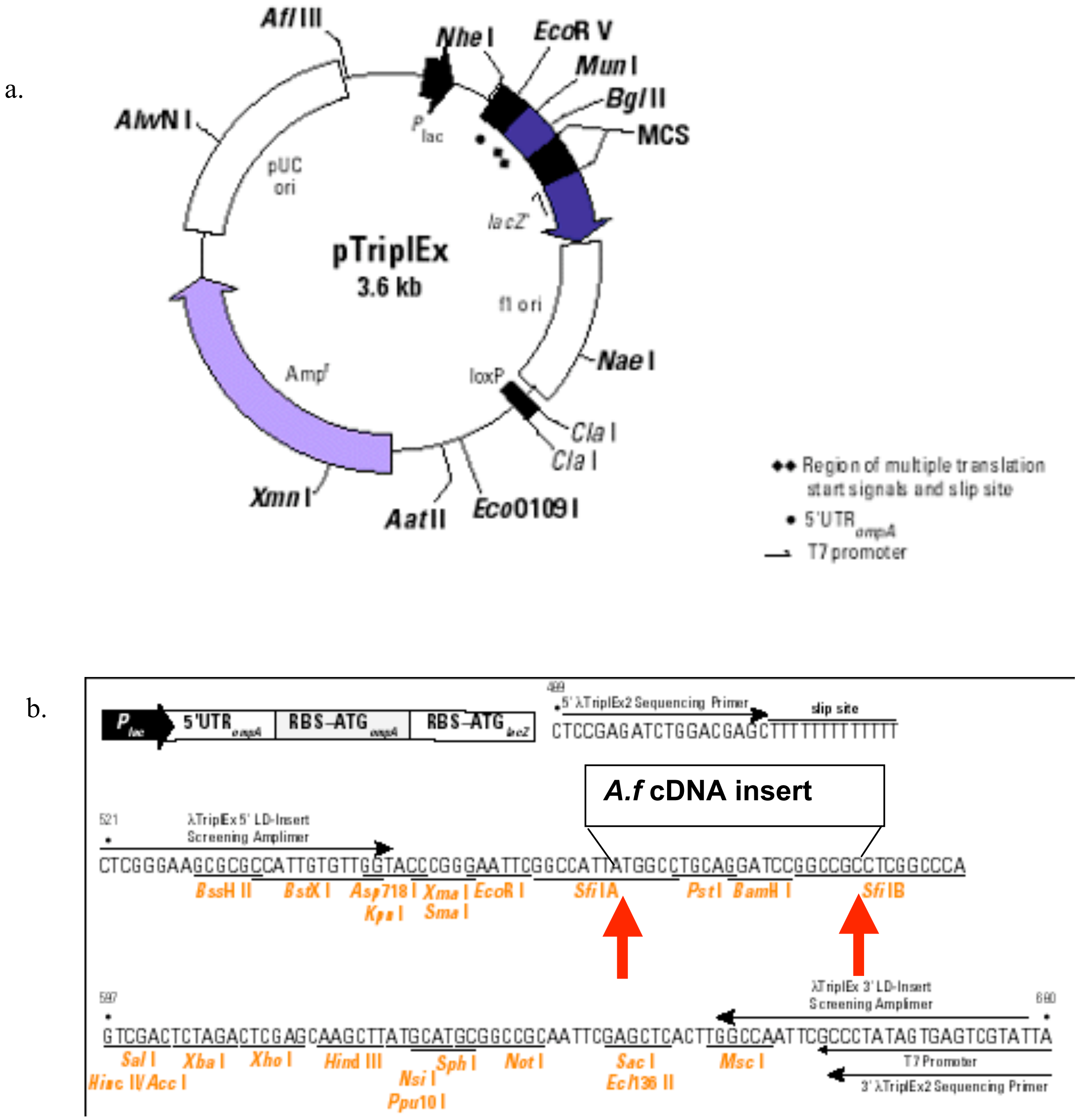

a.
Afl III
AlwN I
Nhe I
EcoR V
Mun I
Bgl II
MCS
P lac
pUC ori
lacZ'
pTriplEx
3.6 kb
f1 ori
Nae I
loxP
Cla I
Cla I
Amp r
Xmn I
Aat II
Eco0109 I
◆◆ Region of multiple translation start signals and slip site
● 5'UTR ompA
T7 promoter
b.
P lac
5'UTR ompA
RBS–ATG ompA
RBS–ATG lacZ
499
5' λTriplEx2 Sequencing Primer
slip site
CTCCGAGATCTGGACGAGCTTTTTTTTTTTT
A.f cDNA insert
521
λTriplEx 5' LD-Insert Screening Amplimer
CTCGGGAAGCGCGCCATTGTGTTGGTACCCGGGAATTCGGCCATTATGGCCTGCAGGATCCGGCCGCCTCGGCCCA
BssH II
BstX I
Asp718 I
Kpn I
Xma I
Sma I
EcoR I
Sfi IA
Pst I
BamH I
Sfi IB
597
λTriplEx 3' LD-Insert Screening Amplimer
680
GTCGACTCTAGACTCGAGCAAGCTTATGCATGCGGCCGCAATTCGAGCTCACTTGGCCAATTCGCCCTATAGTGAGTCGTATTA
Sal I
Hinc II/Acc I
Xba I
Xho I
Hind III
Nsi I
Ppu10 I
Sph I
Not I
Sac I
Ecl136 II
Msc I
T7 Promoter
3' λTriplEx2 Sequencing Primer

Figure 4. pTriplEx vector and insert site.
a. Map of the pTriplEx vector used to amplify the *Artemia* clones.
b. The sequence of the polylinker (MCS) is shown with restriction sites. The *Artemia* cDNA inserts were cloned into the *SfiI* sites. Figure from Vershon, 2008 [2].

II.3. Selection of Clones

A *lacZ* blue-white screen using β-galactosidase was used on the plated *E. coli* to determine if the bacteria contained a vector with an insert. If a clone was inserted into the pTriplEx vector, it would interrupt the plasmid's *lacZ* gene, which is responsible for producing β-galactosidase. β-galactosidase is a protein that cleaves X-gal, resulting in a blue bacterial colony. If an insert interrupted the *lacZ* gene, only a partial β-galactosidase protein would be produced, resulting in a white bacterial colony. To screen for *E. coli* that contained inserts, students selected colonies that were white in color.

The students grew colonies overnight, and performed minipreps on them to isolate the plasmid DNA. Part of the miniprep DNA was digested with *Pvu*II, while another part of the miniprep DNA was amplified using the polymerase chain reaction (PCR). Agarose gel electrophoresis was used to determine the size of each insert. Samples of the uncut, *Pvu*II digested and PCR amplified DNA from each clone were loaded onto an agarose gel and run for 40 minutes at 100 volts, or until the blue tracking dye had migrated through about 75% of the gel. The gels were photographed under UV light and the resulting data (Figure 5) allowed the students to identify the size of each insert. Only inserts larger than 500 nt were further analyzed, as they were more likely to contain a large portion the protein-coding region of the gene, and not simply the 3'UTR and polyA tail.

The plasmids containing inserts larger than 500 nt were sent to GE Healthcare in Piscataway, NJ for sequencing from both the 5' and 3' ends of the insert. 5,947 total sequences were generated by the WSSP and Introduction to Research courses over the four year period. 5’ and 3’ sequences from the same clone were checked for overlap in order to create larger contiguous sequences. 647 of these sequences were published in NCBI. 98 were published by the Rutgers Molecular Biology and Biochemical (MBB) research class, and 549 were published by WSSP (Figure 6).

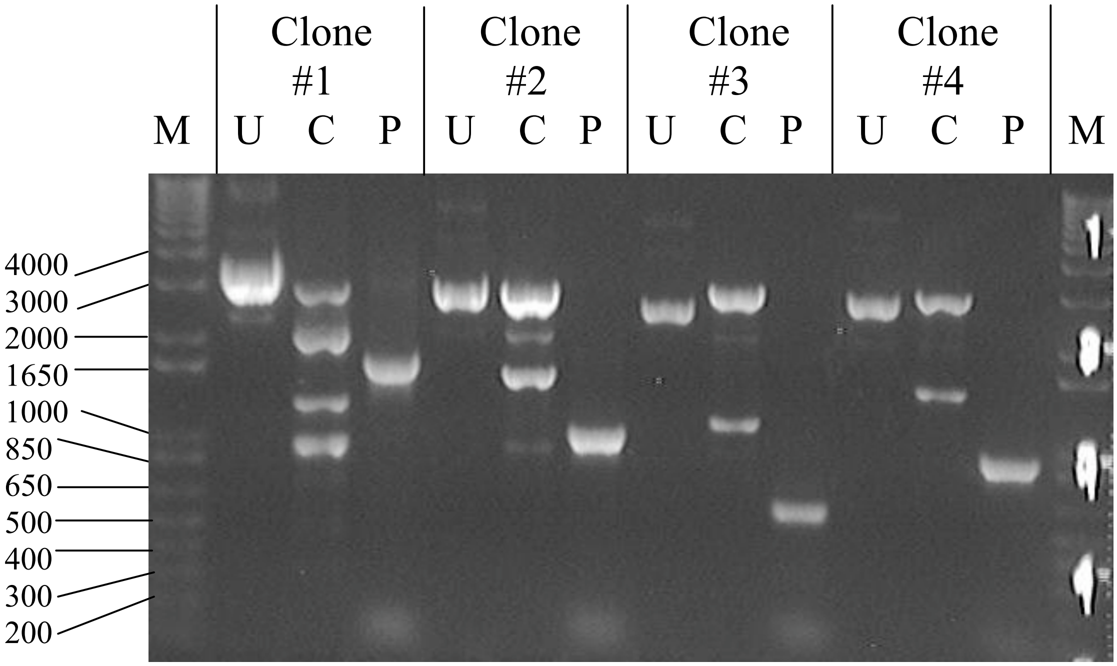


Figure 5. A selection of my clones (SW 1-4) from the cDNA library.
Uncut (U), digested with *Pvu*II (C) or amplified by PCR (P). A 1 kb DNA marker (M) was used to determine the size of the inserts. The 2.9kb vector backbone can be seen in all four digested lanes. Each insert is 700 bp less than the digested insert size, and 200 bp less than the PCR product size.
The insert from clone #1 seems to be about 1300 bp, containing a *Pvu*II site within the insert. The insert from clone #2 seems to be about 650 bp. The insert from clone #3 seems to be about 300 bp. The insert from clone #4 seems to be about 700 bp.

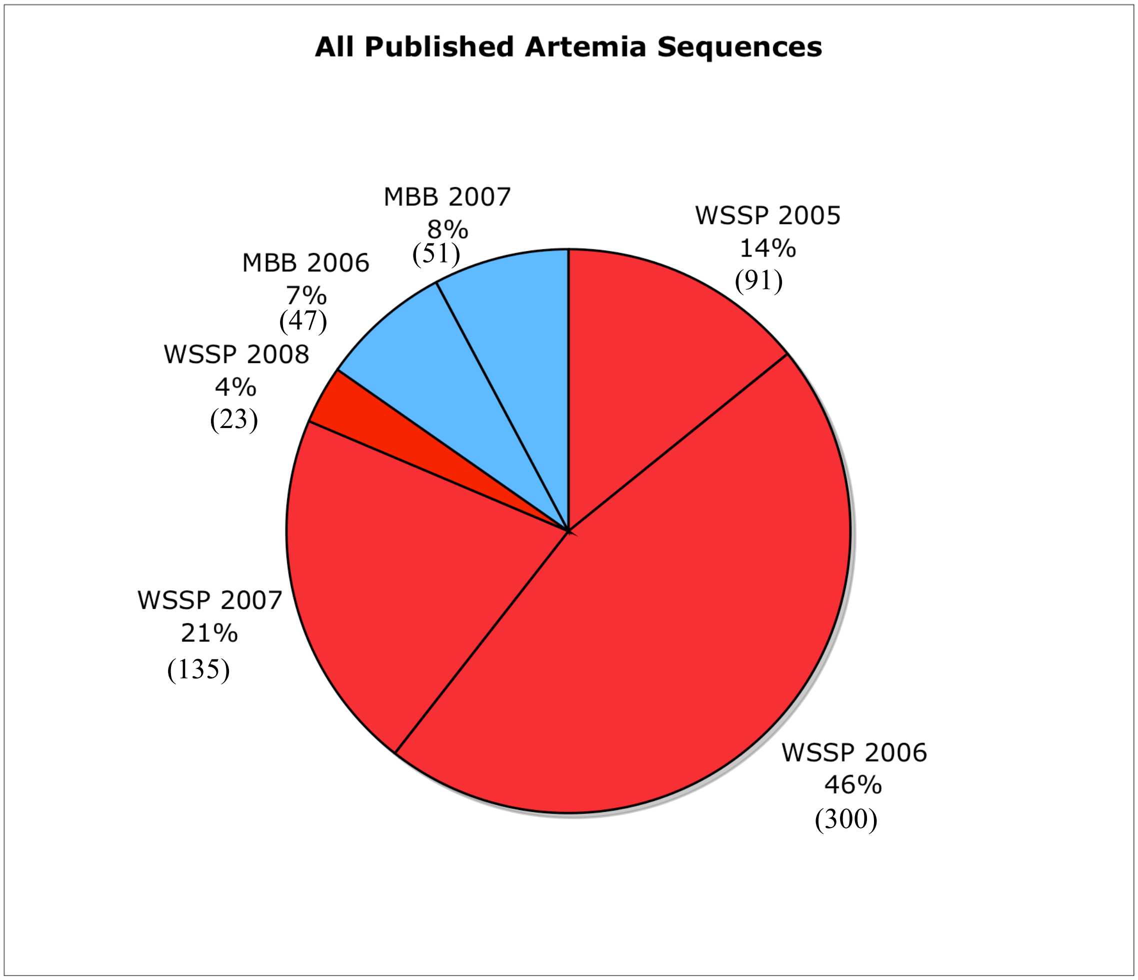


Figure 6. Breakdown of 647 published sequences in NCBI database as of 5/12/09. The 98 published MBB published sequences are in blue. The 549 published WSSP sequences are in red.

III. Analysis of WSSP Published *Artemia* Sequences

III.1. Evolution of this Project

This project began as a means to analyze the *Artemia* genome, with the intention to perform expression studies identifying the novel genes associated with *Artema's* unique lifestyle involving anoxia and dormancy. However, in January 2009, a consortium from the Chinese Academy of Sciences published a paper in BMC Genomics on *Artemia franciscana*. [8]. They published 28,039 sequences on NCBI and identified 324 differentially-expressed genes based on pairwise comparisons of the cDNA libraries from four different time points. These time points included dehydration, as well as 5, 10, and 15 hours after rehydration. While their focus was on protein expression, they failed to acknowledge any information regarding the non-coding regions of the *Artemia* genome.

Investigation into the WSSP sequences identified an abundance of non-coding RNAs (ncRNA), many of which were longer than anticipated based on current dogma. Our focus therefore shifted, and we decided instead to compare the results from our coding sequences to Chen et al. [8], while additionally analyzing our abundance of non-coding sequences. The non-coding regions of the genome are recently receiving recognition for their roles in gene expression and subsequently warrant analysis. We wanted to analyze our long ncRNAs, and see if they were long 3'UTRs, or possibly something more. We expected that these potential long ncRNAs held additional information overlooked by previous research, and as such, fueled our focus to further explore these regions of the genome.

III.2. Materials and Methods

Nucleotide sequences from the nr/nt and EST databases generated from the Waksman Student Scholars Program (WSSP) and the Rutgers University undergraduate course were downloaded from NCBI on 9/14/08 using the search words 'artemia nemeroff', as all sequences were published with Dr. Nemeroff's name. 418 EST sequences and 210 nucleotide sequences were downloaded. Each sequence was analyzed individually using NCBI's BLAST programs [9]. BlastN searches were performed on the EST and nr databases [10]. BlastX analysis was performed on the 'non-redundant protein sequences' database [11]. "Toolbox" software, provided by the WSSP, was also employed for each sequence to translate each sequence into six reading frames. Internal open reading frames (ORF) larger than 80 residues, containing a methionine (Met) were then analyzed using BlastP of the non-redundant protein database. ORFs longer than 50 residues at the beginning or end of the sequence were also analyzed using BlastP, in the event that the sequence contained only a part of the full-length protein. These ORFs were recognized as part of the coding region if the BlastX alignment score was larger than 80. The 5' and 3' untranslated regions, ORF and polyA tail were identified and noted for all sequences. Sequences that did not contain an ORF larger than 50 residues and did not match any genes in NCBI were considered non-coding RNA (ncRNA). If the ncRNA contained a stretch of 10 or more adenines at the end of the sequence, these adenines were labeled as the "polyA tail" and the remainder of the sequence was labeled as "polyA ncRNA". If the ncRNA did not contain a stretch of 10 or more adenines, it was labeled as "ncRNA".

III.3. Overall Annotation

Of these 628 sequences, 207 sequences (33%) contained a region that coded for a known gene, while 23 sequences (4%) coded for a ‘hypothetical protein’, based on Blast results from NCBI. Some sequences originally classified as ESTs in the NCBI database coded for a protein as evidenced by a BlastX alignment score greater than 80. OrfPredictor [12] was run to identify open reading frames longer than 100 nucleotides, containing a Met. OrfPredictor identified 28 additional ‘hypothetical proteins’ (4%) that were not characterized by Blast results from NCBI. 361 sequences (58%) did not code for any known protein and did not have an open reading frame larger than 100 amino acids. They were therefore identified as ncRNA. 219 of these ncRNAs contained at least 10 adenines at the end of their sequence, and were subsequently labeled as polyA ncRNA. (Figure 7).

While 58% may sound like a higher percentage of non-coding sequences than expected, EST analyses from *Daphnia magna*, a close relative of *Artemia*, resulted in a similar percentage of 59% [13]. The FANTOM (Functional Annotation of the Mouse) consortium analyzed mouse EST libraries and concluded that 48% of mouse RNA is non-coding [14]. Our data support these values. Chen et al. [8] did not determine the abundance of non-coding *Artemia* sequences in their study, so we are therefore, unable to compare our values to theirs.

Approximately 1% of the ESTs (8 out of 628) were identified as the large mitochondrial 16S rRNA gene. Other systematic sequencing projects find about 2% ribosomal RNA

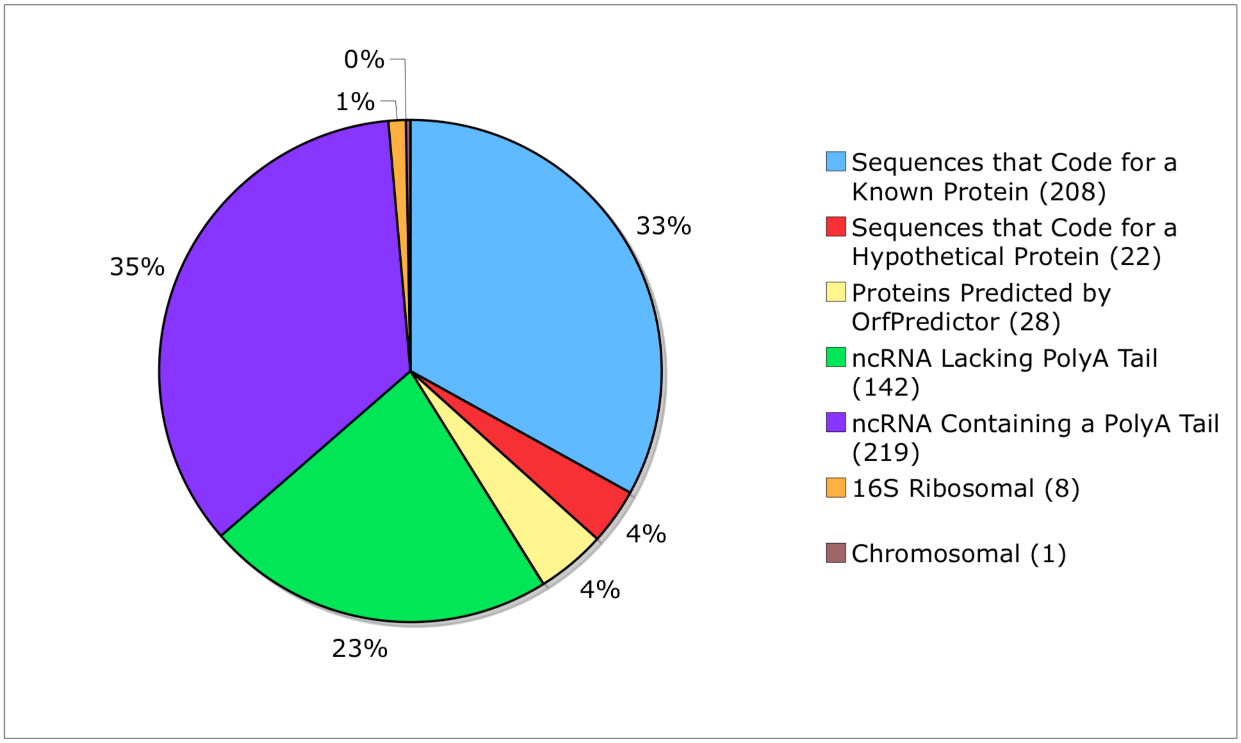


Figure 7. A majority of our published sequences are non-coding RNA. This pie chart illustrates the large percentage of non-coding RNA extracted from the library, relative to coding sequences.

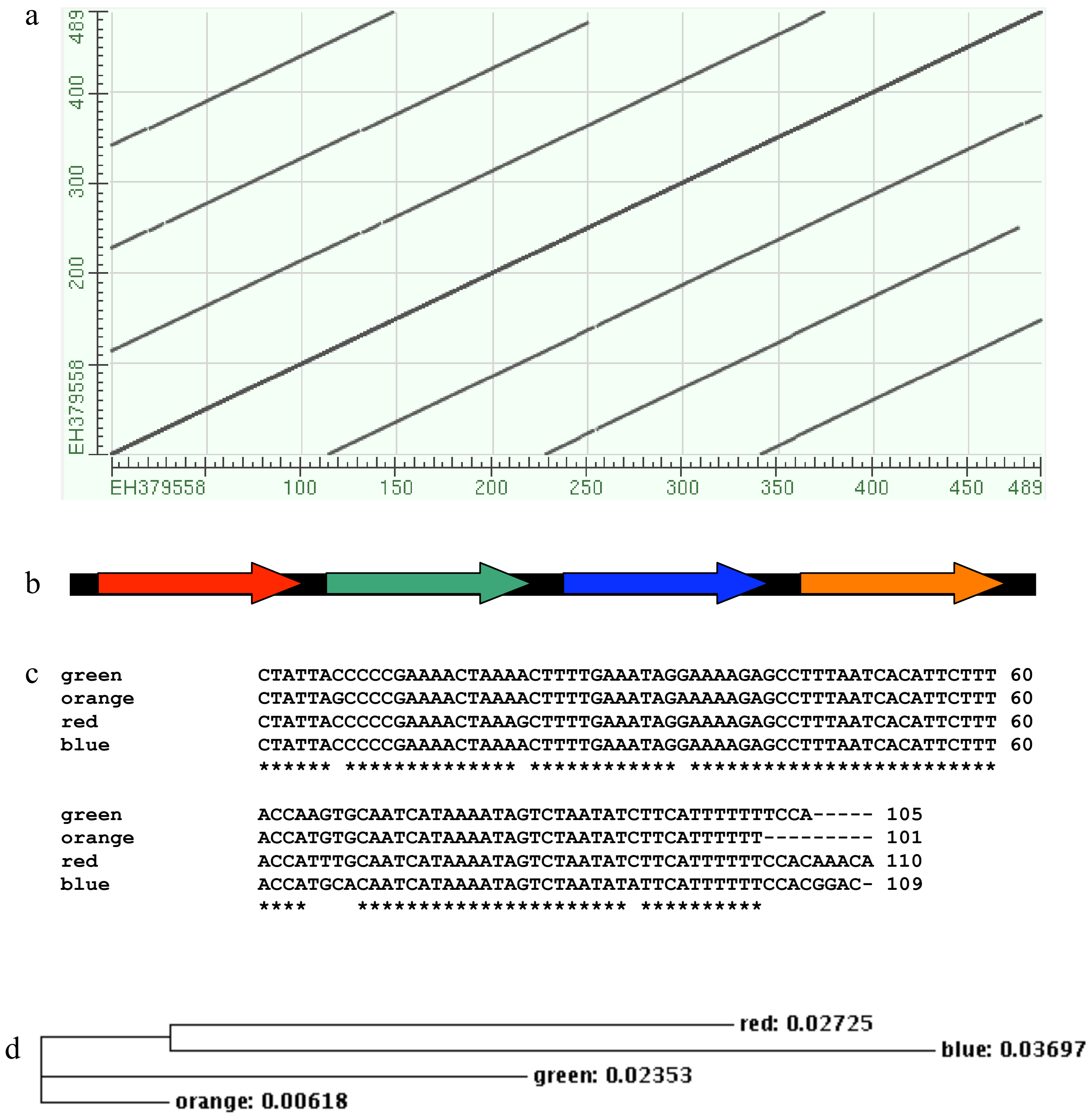
contamination [15]. Eleven additional mitochondrial sequences, including cytochrome b and c, were also recognized and are included in the 207 sequences that code for a gene.

III.4. Identification of Microsatellites

Microsatellites, also known as Single Sequence Tandem Repeats (SSR) are useful genetic markers for parental identification, linkage mapping, and population genetics. They can be found in the 5' and 3'UTR, as well as the coding and non-coding regions of a gene. They have recently been recognized for their functional role of affecting gene transcription and regulation, as well as mRNA splicing [16]. Among the 628 WSSP sequences examined in this study, SSRs were identified in 18 sequences (2.9%) by using MISA software [17]. The minimum number of repeats for dinucleotides was six, while the minimum number of repeats for tri-, tetra- and penta-nucleotides was five.

The most frequent motifs were di- (50%) and tri- (39%) nucleotides with predominance of TA (6 out of 9) and AAC (3 out of 7). Sequence conservation analysis was performed on these short motifs through BlastN. Three of the sequences containing the TA repeat matched *Gasterosteus aculeatus* sequences with e-values of 0.0, however, the TA repeat was interrupted in all three of these sequences. Two of the sequences containing the remaining TA repeats did not match any other sequences in NCBI, while one of the sequences containing the TA repeat was conserved in *Nematostella* (FC318433). Of the seven trinucleotide repeats, 3 of them were not conserved in any published sequences in NCBI. Two of these seven trinucleotide repeats were highly conserved, and two were located within sequences published by Chen et al. [8] These comparisons imply that the location of microsatellite sequences, particularly the short ones, frequently changes from one organism to the next.

A stretch of bases between 101-110 nt in length was also repeated four times within one of our sequences (EH379558), with each repeat slightly different from the next (Figure 8). This repeat was also identified in *Artemia salina* clones and could be used as a genetic marker to identify different species of *Artemia*.



```
green           CTATTACCCCCGAAAACTAAAACTTTTGAAATAGGAAAAGAGCCTTTAATCACATTCTTT 60
orange          CTATTAGCCCCGAAAACTAAAACTTTTGAAATAGAAAAAGAGCCTTTAATCACATTCTTT 60
red             CTATTACCCCCGAAAACTAAAGCTTTTGAAATAGGAAAAGAGCCTTTAATCACATTCTTT 60
blue            CTATTACCCCCGAAAACTAAAACTTTTGAAATAGGAAAAGAGCCTTTAATCACATTCTTT 60
                ****** ************** ************ *************************

green           ACCAAGTGCAATCATAAAATAGTCTAATATCTTCATTTTTTTCCA----- 105
orange          ACCATGTGCAATCATAAAATAGTCTAATATCTTCATTTTTTT--------- 101
red             ACCATTTGCAATCATAAAATAGTCTAATATCTTCATTTTTTTCCACAAACA 110
blue            ACCATGCACAATCATAAAATAGTCTAATATATTCATTTTTTTCCACGGAC- 109
                ****     ********************** **********
```

Figure 8. Repeat sequence within an *Artemia* clone.
a. Illustration of the alignment when the sequence is blasted against itself.
b. Showing the order and direction of the four repeats.
c. A clustalW alignment of all four repeats. Notice that each repeat is slightly different from the each of the other repeats.
d. A phylogram of the four repeats.

III.5. GC % Content

The average GC% content of the 5'UTR, coding region and 3'UTR of the 197 published *Artemia* proteins (11 mitochondrial proteins were removed from the original 208 total) was calculated to be 30.5%, 42.8%, and 30.2%, with standard deviations of 7.6, 3.7, and 6.4, respectively (Figure 9). The GC% content of the 5'UTR, coding region and 3'UTR of the hypothetical proteins, as well as the sequences predicted to be proteins by OrfPredictor were also calculated. These GC% values were subsequently used to further support the likelihood that these hypothetical proteins were authentic proteins. Observations were made to identify if each region of the hypothetical gene fell within the calculated accepted range for that region. The GC% content of the coding region relative to the 5'UTR and 3'UTR was also noted, as the GC% content of the coding region should be higher than that of the non-coding region. This is due to the three hydrogen bonds found within the GC pair that subsequently makes GC pairs more stable than AT pairs, which only contain two hydrogen bonds. Of the 50 sequences predicted to be proteins, only 21 (43%) of them contained a higher GC content in the coding region when compared to the non-coding region and also fell within the accepted range in each region of the gene. These sequences likely code for proteins native to *Artemia* and subsequently deserve further investigation.

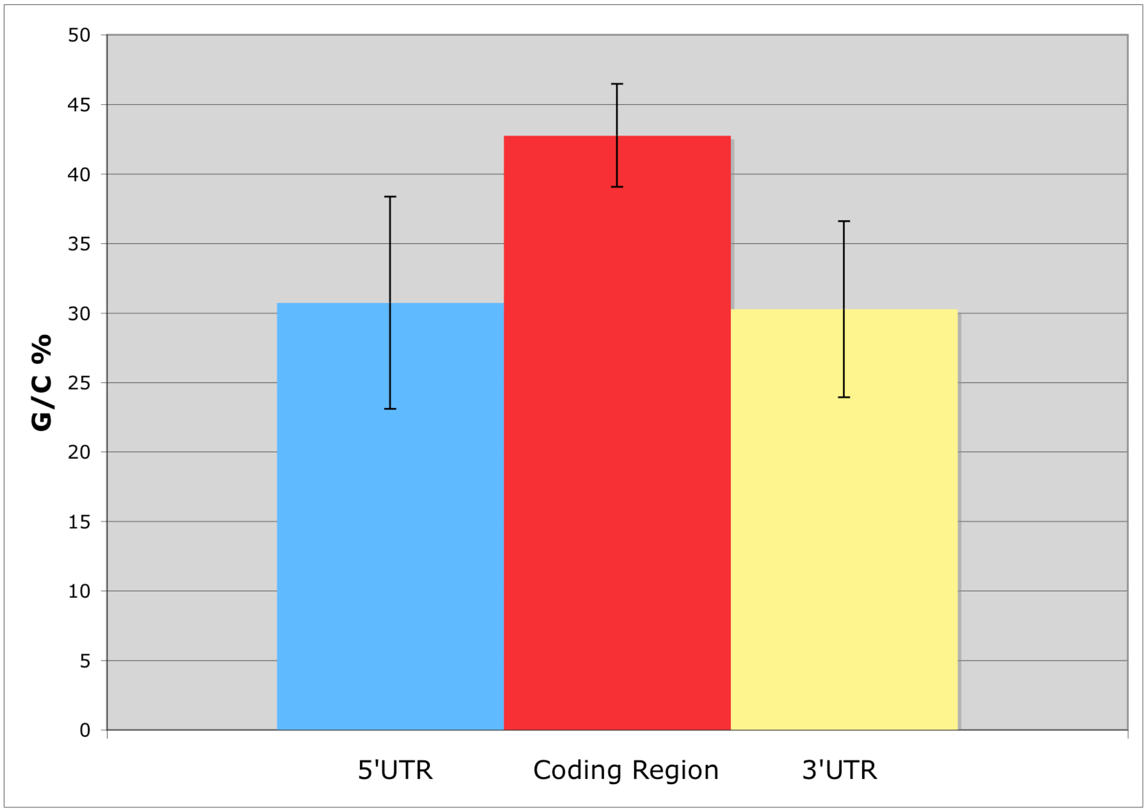


Figure 9. GC % content of coding vs. non-coding regions of sequences containing known genes.
This graph illustrates that the average GC% content of the coding region of the known genes is larger than the average GC% content of the non-coding regions of the known genes.

III.6. UTR Analysis

Of the 418 EST sequences downloaded from NCBI, only 361 of them were considered to be ESTs, as 57 of these 418 EST sequences coded for proteins. All EST sequences were compared to all other sequences in the NCBI database. Duplicate sequences were noted, and any sequence with a significant similarity was documented. 'Significant' similarity was identified with a BlastN alignment score larger than 80. Special attention was paid to these 'significant similarities', noting if the match was to an *Artemia* sequence, or to a non-*Artemia* sequence. In addition, if our sequence matched an *Artemia* sequence, strand direction was identified as +/+ or +/-. If less than 60% of our sequence was matched by other *Artemia* sequences, it was identified as a 'short' match. When our sequence contained a "polyA tail" it was compared to sequences with significant similarities to help determine if the clone was generated from an mRNA or was isolated due to an adenine-rich region of the genome that passed through the purification steps in the construction of the cDNA library.

Only 189 (52%) of these 361 EST sequences analyzed significantly matched other *Artemia* sequences published in NCBI, with alignment scores greater than 80. 78 ESTs (22%) significantly matched sequences belonging to other organisms such as *Gasterosteus aculeatus* (stickleback), and *Hippoglossus hippoglossus* (halibut). 70 of these 78 sequences also significantly matched to a published *Artemia* sequence. 164 sequences (46%) were unique (Figure 10).

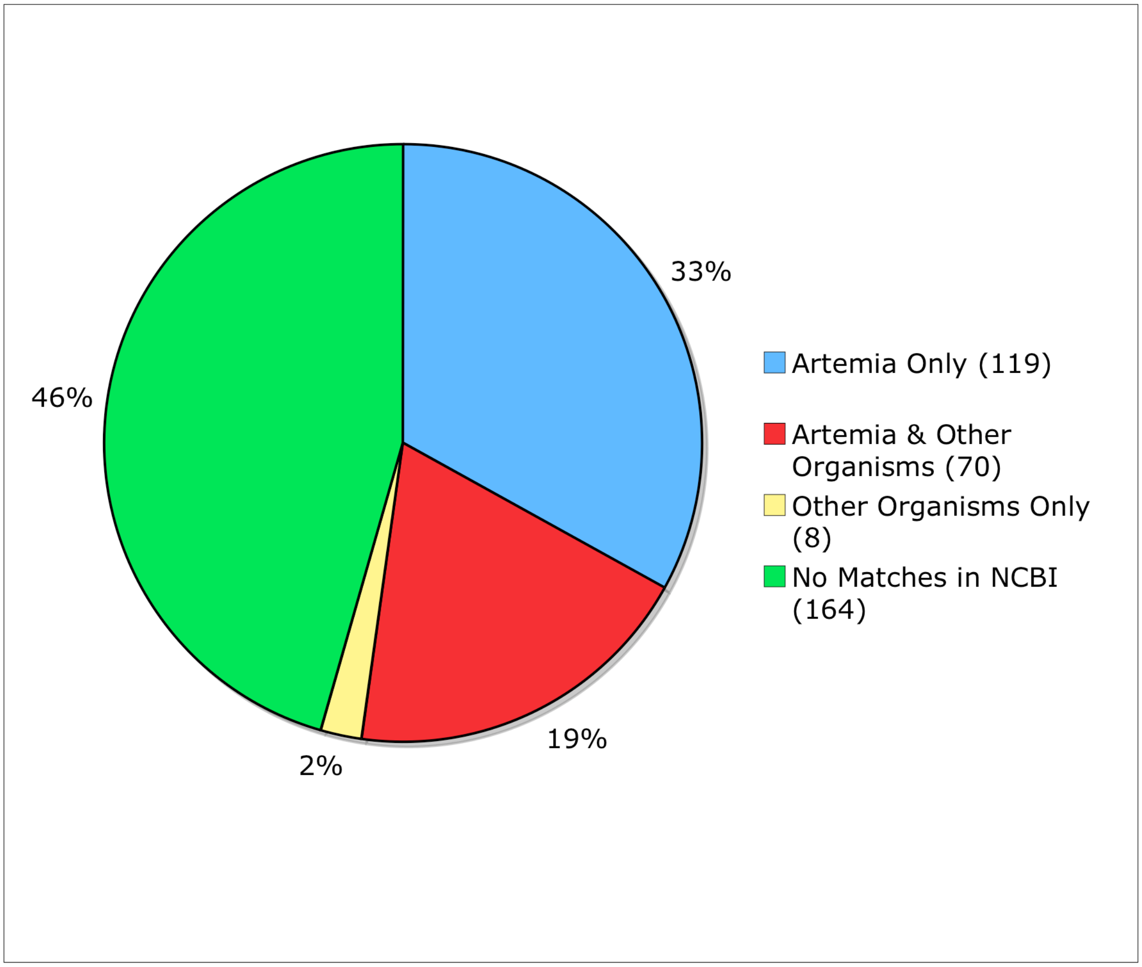


Figure 10. Significant EST matches of our *Artemia* sequences to other sequences in NCBI.
Note the large number of sequences that do not have any significant matches.

Among the 189 EST sequences that matched other *Artemia* sequences in NCBI, 62 sequences (33%) matched the *Artemia* sequences in the antisense direction. Antisense transcription is a mechanism for gene regulation in the transcriptome of almost all organisms, as it can result in the degradation of the sense transcript [14]. Simply the act of transcribing the gene in the reverse direction, as well as the transcript itself, can result in positively or negatively affecting the expression of nearby genes [18]. These long non-coding RNAs are a newly recognized means of gene regulation, and may be represented among these 62 sequences.

Upon analyzing the 189 sequences that matched other *Artemia* sequences in NCBI, 87 of them (46%) were considered 'short' matches, as only part of our sequence matched the published sequence. This usually occurred when the published sequence was shorter than ours, but it also occurred when we still had overlapping sequences (see ES525195 and FL685674 in Figure 11). This evidence of only part of two sequences matching, even when they are from the same organism, supports that non-lethal recombination, mutations and duplicate sequences occur within the non-coding regions of the *Artemia* genome. This kind of alignment also suggests that this part of the EST may be important and could be used in gene regulation. 21 of these short matches matched in the reverse direction, and may be involved in gene regulation. These short matching regions deserve further investigation.

```
ES525195        GGATTATATATTTGTTCTGTACTGAG-CACTACATCATTAATGCAAAATGTATTCGTCAA 119
FL685674        ---------------ACTACAGTAAGTCCCTGGAACAC--ACCCTCCATGTAGCTCCTAA 43
                                **  * * ** * **  * **   *  *   *****      **

ES525195        ATAGTTCACGAC-GCGAAATATCAAAAAGACGAATGTGTTGATGACGT-TCTGCAAAATT 177
FL685674        TCACTTTGCTACTACTGCCAATTAAAGTACTAGACATTCTTTTGCTCTGTCTGATGGAGG 103
                  * **  * **  *     ** ***       *  *  *  **   * ****    *

ES525195        TTGACAGGATTGTCTTGGAGCAAGCTCAGTTTGAAAGACGGAAATTTGTTGCATTTGAGA 237
FL685674        AGCAATTTTTTCTCTTACAGATAATGCATGCACAATTGTCAAGACTT----CGATTTTGG 159
                   *     ** ****  **  *   **     **      * * **       **  *

ES525195        AGAAAAAAGGCTTAGAATCGCGTATATCAGAATTGACAAATGAAATTGAGCAGACCTCAA 297
FL685674        ATAGCCTAGGCTTAGAATCGCGTTGATCAGAATTGACAAATCAAATTGAGCAGACTTCAA 219
                * *    ****************  **************** ************* ****

ES525195        AGTTGTTGGAGGTTACACTAGCACACGTGCAAGACGAAGAAAATGCATTGAAAGAAGCCC 357
FL685674        AGCTGTTGGAGGTTACACTGGCACACGTGCAAGACGAAGAAAATGCATTGAAAGAAGTCC 279
                ** **************** ************************************* **

ES525195        AATCAACTTACAAAAG----------CTTACGTCGGGCTAT--------TCAGAATAAAA 399
FL685674        AAACAACTTACAAAAAGTACTTTTGTTTTGCCTAAAGGTATAGCCGATATTTTAATCTTG 339
                ** ************            ** * *   * ***        *   ***

ES525195        TGGAGCTGA-ATAAACTGAAGCAACAGCAAGATGC---TGTCAAAAAGTTCCACGCTGAT 455
FL685674        TGATTTTGGCAAGAATTTGGCCAACAGCACTGCACATTTGACAGGAATAACTGAGCAAAA 399
                **    **  *  ** *    ********     *   ** **  **   *   **  *

ES525195        TCTTTGC--GATCTCAAATTCAAGCTAACTT--------CTATCTAAACATGATTGTT-- 503
FL685674        AGTATGCATGACTGCCAACTCACCGTGGCCTTTTATTCACTAATTAAAGATAAATGCTTG 459
                  * ***  **   * ** ***   *  * *        ***  **** ** * ** *

ES525195        -GAAAGAAGCAAGG--AACGTTTTCACAAGTGTTATACTT---TCAACACATTTTTTCTC 557
FL685674        AAAAAAAAGCCAGCTTGAGGCTAGCTAGGCCACTTCGCTTGATTGAAGATATGCCAATTC 519
                  *** **** **    * * *  *        *   ***   * ** * **      **

ES525195        ACGAAACTCCTTGAAAGTGG--TCACGCCGG-ACTAAAGAGATGGACGAGAAAGATAGAT 614
FL685674        TTTACATTCTTTTAAAACGAATTTATTCCGTCACTATACAAAGCAAAATGCGTGCTTGAT 579
                   * * ** ** ***  *   * *  ***  **** * * *   *   *   * * ***

ES525195        ATAT-TTGAGAACGAGATAATTCTCGTCC--CTGTGCATCTAGCTGTT------------ 659
FL685674        TCATATTTCAAGTTATTTGATTTGCATATAACAGTATTGTCTGCAATTGATAACTATGAA 639
                  ** **   *   *  * ***  * *    * **       **  **
```

Figure 11. Illustration of the partial match between a WSSP sequence (FL685674) and another *Artemia* sequence in NCBI (ES525195).
The e-value for this matching segment highlighted in red is $3e^{-45}$. Note how well the red segments match one another, and how poorly the black segments match.

ESTs with a ‘polyA tail’ (219) were analyzed to see if this polyA region was observed in other published sequences. This would help identify if the stretch of adenines at the end of our sequence was a polyA tail, or if it was simply an adenine rich region of the genome. When BlastN analysis was performed on these 219 ESTs, 159 sequences did not match their polyA tail region to anything in NCBI, while 60 ESTs matched their polyA tail region to sequences in the database. Of these 60 matches, 27 supported the ‘polyA tail’, through a significant polyA match at the end of the subject sequence (Figure 12). 47 sequences however, rejected the validity of our ‘polyA tail’. This was recognized when the subject sequence matched our query sequences well and contained an adenine-rich region within the sequence at the same position where we identified our ‘polyA tail’ to be (Figure 13). There were also alignments where we had a ‘polyA tail’ and the subject sequence did not contain any adenines at all. This could be due to alternate polyadenylation. 14 sequences showed alignments to sequences that both supported and rejected the authenticity of our ‘polyA tails’ (Figure 14). These numbers indicate that some of our ‘polyA tails’ may, in fact, not be polyA tails, but may have arose from adenine rich regions of the genome that were expanded during oligo-dT priming during construction of our cDNA library. Other unmatched polyA tails may be the result of alternate adenylation.

```
EH093845        GCATTCGCAATCTTTTCTGCATTGATGTTTGTCTGCGCAGCATTGGCTAAGCCATCACCT 60
ES494148        GCATTCGCAATCTTT-CTGCATTGATGTTTGTCTGCGCAGCATTGGCTAAGCCATCACCT 59
                *************** ********************************************

EH093845        GACTCTGGTTGGTGGGATGAAAAATTCGAAATACCAGAGCATTTAATAATTGGGTAAACA 120
ES494148        GACTCTGGTTGGTGGGATGAAAAATTCGAAATACCAGAGCATTTAATAATTGGGTAAACA 119
                ************************************************************

EH093845        GAGTTCCGAAAAAGGACCCTGCATACTGGAAATATCATCGGTATTTGTAACAAATCTTTT 180
ES494148        GAGTTCCGAAAAAGGACCCTGCATACTGGAAATATCATCGGTATTTGTAGCAAATCTTTT 179
                ************************************************* **********

EH093845        TTAAAGTTTTCAAACTCTACTGGCATGGATTATTTCATTAATTTTCTGATGTGCTGTTTA 240
ES494148        TTAAAGTTTTCAAACTCTACTGGCATGGATTATTTCATTAATTTTCTGATGTGCTGTTTA 239
                ************************************************************

EH093845        GCTTATTCTATTCTTGTAATTTCAATAAAGGAACTTTTGGCTTAAAAAAAAAAAAAAAAA 300
ES494148        GCTTATTCTATTCTTGTAATTTCAATAAAGGAACTTTTGGCTTAAAAAAAAAAAAAAAAA 299
                ************************************************************

EH093845        AAAAAAAAAAAAAAAAAAA-------------- 319
ES494148        AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA 332
                *******************
```

Figure 12. Some polyA stretches at the end of a WSSP sequence are polyA tails. This match illustrates an example when the polyA tail of our WSSP sequence (EH093845) matched the polyA tail of another *Artemia* sequence (ES494148).

```
ES584503        TATTAATATTTTTATTAATTTATATTTTTAATTATTTTTTATTTAGTTATTGTGAATTGG 300
ES500865        -------------------------------------------------------TGCAG 5
                                                                           *

ES584503        AAAATAGTTATTTTTATTTCCTTTCATAATAAATGCTCAACAATAGCCCAGCAGAACTTT 360
ES500865        GAATTCGAATTTTTTATTTGTAATAACAATAA-TGCTCCATAATAGCCCAGCAAAACGTC 64
                 ** * *   *********    * * ***** ***** * ************ *** *

ES584503        TAGTGCTCCATAATAGCATAGCAAAATCGCCTGAAAATTTTAGCTATAAAAAAA----GA 416
ES500865        TAATGCTCCATAATAATATAGCAAAATCGACTAAAAATTATAGCTGCAAAAAAAAAAAGA 124
                ** ************  ************ ** ****** *****  *******    **

ES584503        AGAAAACAAAAATGTTGCGAATGAAGCAACATTTTGATAATTTTGCAGCCAGCCAGGGAT 476
ES500865        AGGAAAGGGAAATTTTGCGAGTGAACCAAGAGTTGGATAATGTTGCAGCCAGACCGGGAT 184
                ** ***   **** ****** **** *** * ** ****** ********** * *****

ES584503        GTTTAGTTCTCTTTATTTGAGGTATCGCAACTTTAATCTGAAGTACTAATGCAAGTTCGA 536
ES500865        ATTTAGTTGTCTTTATTTGAGGTATCGCAACTTTAATCTGAAGTATTAATGCAAGTTCGA 244
                 ******* ************************************ **************

ES584503        AGTGTACTGGATTATTTTTTTTGTTTATGCAAGTTGTTTCTTTTGTGAATAAGCATTCTT 596
ES500865        AGTGTACTGGATTATTTTTG--GATCATGCAATTTGTTTCTTTTGTGAATAAGCATTCTT 302
                *******************   * * ****** ***************************

ES584503        GAATAAAAGCTAGAGTTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA------------ 644
ES500865        GAATAAAAACTAGAGTTTAGAAAAAAAAAAACCGATGCTTATTATTAACATTGAGCAGTC 362
                ******** ******** * ***********   *     *  *  **

ES584503        ------------------------------------------------------------
ES500865        ATAATGATAAAATTAGTATATGTATAAAAATGCAATAGAATCTATTCCAAAAATTTGGGA 422

ES584503        ------------------------------------------------------------
ES500865        AGGAATCCGGCAAAACTTGCCTCCGCCTGTTTAACAAAAACATCGCCTCCTAACTTTAGG 482
```

Figure 13. Not all polyA stretches at the end of a WSSP sequence are polyA tails. Above is an illustration of the match between a WSSP sequence (ES584503) and another *Artemia* sequence in NCBI (ES500865). The e-value for this match is $9e^{-95}$. This sequence was most likely found in the WSSP library due to the 11 adenines within the sequence. However, the additional adenines are most likely an artifact that arose during oligo-dT priming during construction of our cDNA library. The boxed region indicates the potential oligo-dT priming site.

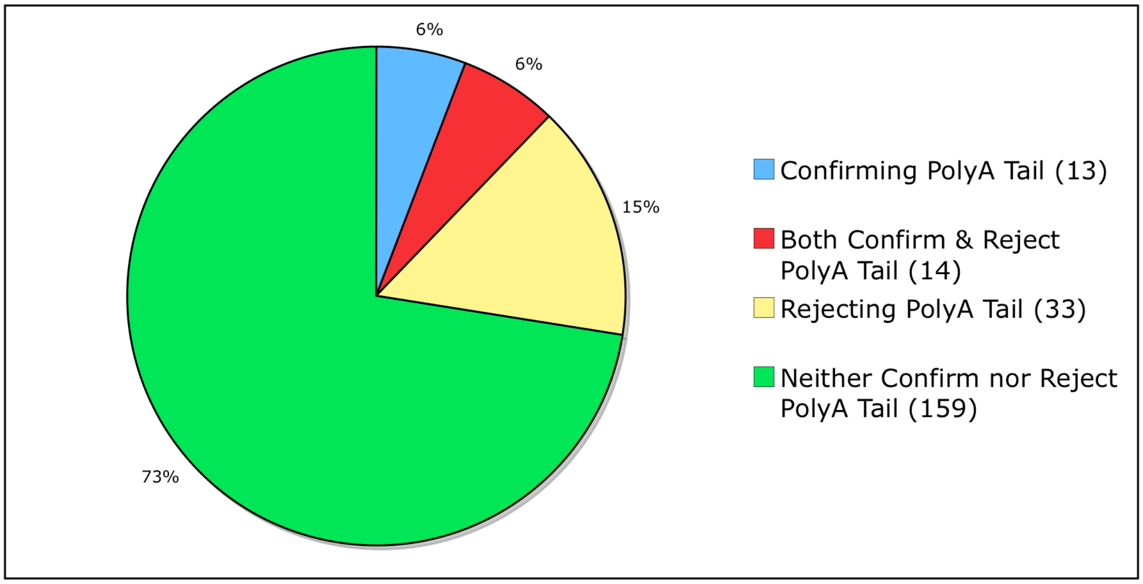


Figure 14. Many polyA tails cannot be verified.
The figure illustrates how many polyA tails from sequences identified as "3'UTR" were supported and rejected by sequences in NCBI. A large number of polyA tails cannot be verified based on the lack of sufficient sequences in NCBI.

III.7. 3'UTR Analysis

Gene expression can be controlled through transcriptional and translational regulation. Many methods for translational regulation involve the 3' and 5' UTR. Quantitative analysis of the 3'UTR was performed on *Artemia* sequences that contained a known coding region, as well as a polyA tail. These parameters ensured the untranslated region was in fact 3'UTR, and not 5'UTR, and it guaranteed the analysis of the entire 3'UTR, and not simply a partial sequence. These criteria fit 119 sequences. The average 3'UTR length was 175 nucleotides long, with a maximum length of 857 nucleotides, a minimum length of 2 nucleotides, and a standard deviation of 189 nucleotides (Figure 15). Almost half of these sequences (44%) contained a 3'UTR with a length between 30-90 nucleotides. A recent study found the average length of the 3'UTR in invertebrates is about 300 nucleotides [19]. Our data does not support these values. Nevertheless, further analysis into sequences with the longest 3'UTR lengths found relative length similarities among other invertebrates. BlastN searches were performed on the top 10 sequences with the longest 3'UTR lengths (480+). The top Blast returns were to other invertebrates with UTR lengths well over 300 nucleotides as well, including, but not limited to *Aedes, Apis, Bombyx,* & *Nasonia*. BlastN searches were also performed on the 10 sequences with the shortest 3'UTR lengths (< 40), five of which were ribosomal proteins. The top Blast returns were to other invertebrates with UTR lengths under 300 nucleotides as well. These invertebrates included *Spodoptera* and *Acyrthosiphon*. When 2 or more of our sequences coded for the same protein, the 3'UTR lengths of each protein were typically within 15 nucleotides of one another. These results support that our 3'UTR lengths are correct, despite their deviation from the average.

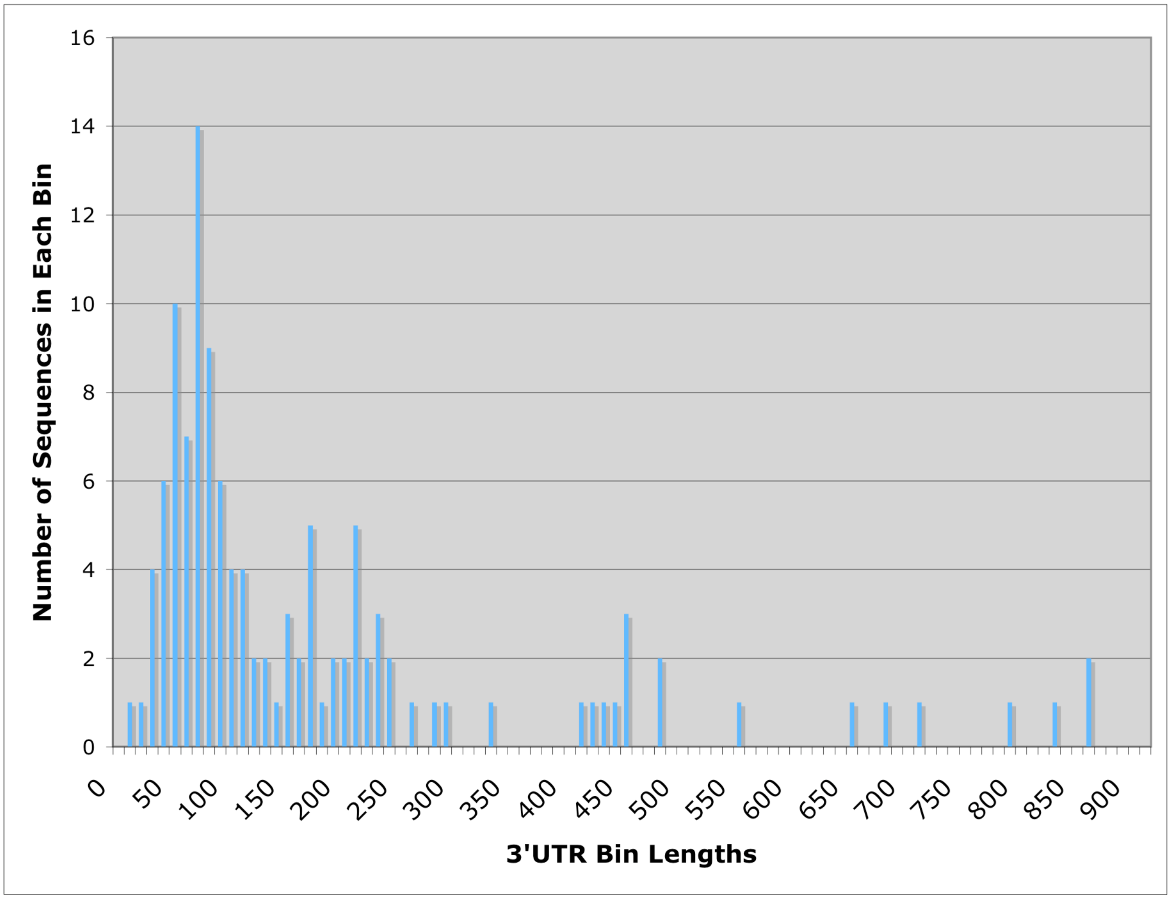


Figure 15. Average 3'UTR length is 175 nucleotides.
Illustration of 3'UTR lengths of all sequences containing both a coding region and a polyA tail of 10 or more adenines. While most sequences have a 3'UTR length less than 150 nucleotides, many sequences have 3'UTR lengths between 150 and 900 nucleotides. The average length is 175 nucleotides. Each bin is 10 nucleotides.

The average length of the 219 sequences identified as “polyA ncRNA” was 680 nucleotides long, with a standard deviation of 278. These sequences that do not contain the coding region for a gene seem unnaturally long, considering our calculated average of 175 for 3’UTRs of sequences that do contain a coding region. This suggests that they are not long 3’UTRs, but instead, could be ncRNAs. PolyA signals were used to help confirm the authenticity of the known 3’UTRs. The 3’UTRs of the confirmed 119 genes were run through a program written to analyze the nucleotide profile at specific positions in relation to the end of the transcript. This program is therefore able to reveal conserved motifs within a sequence. The polyadenylation signal (AAUAAA, and the 11 single base variants) was searched for and revealed, 10-25 nucleotides upstream of the polyA tail, in the 119 sequences with confirmed genes (Figure 16). The frequency of adenines was about 60% in this region, identifying it as the polyA signal site. When the same program was run with the 219 polyA ncRNA seqeunces that did not contain a coding region, this trend was recognized again, albeit not so obviously. The frequency of adenines was about 45% within the polyA signal site (Figure 17). It is therefore highly probable that a majority the stretches of adenines at the end of these 219 polyA ncRNAs are polyA tails, and not from adenine-rich regions of the genome. Any sequences containing mistaken ‘polyA tails’ will not have the polyadenylation signal, and are subsequently responsible for decreasing the frequency of adenines within the signal site.

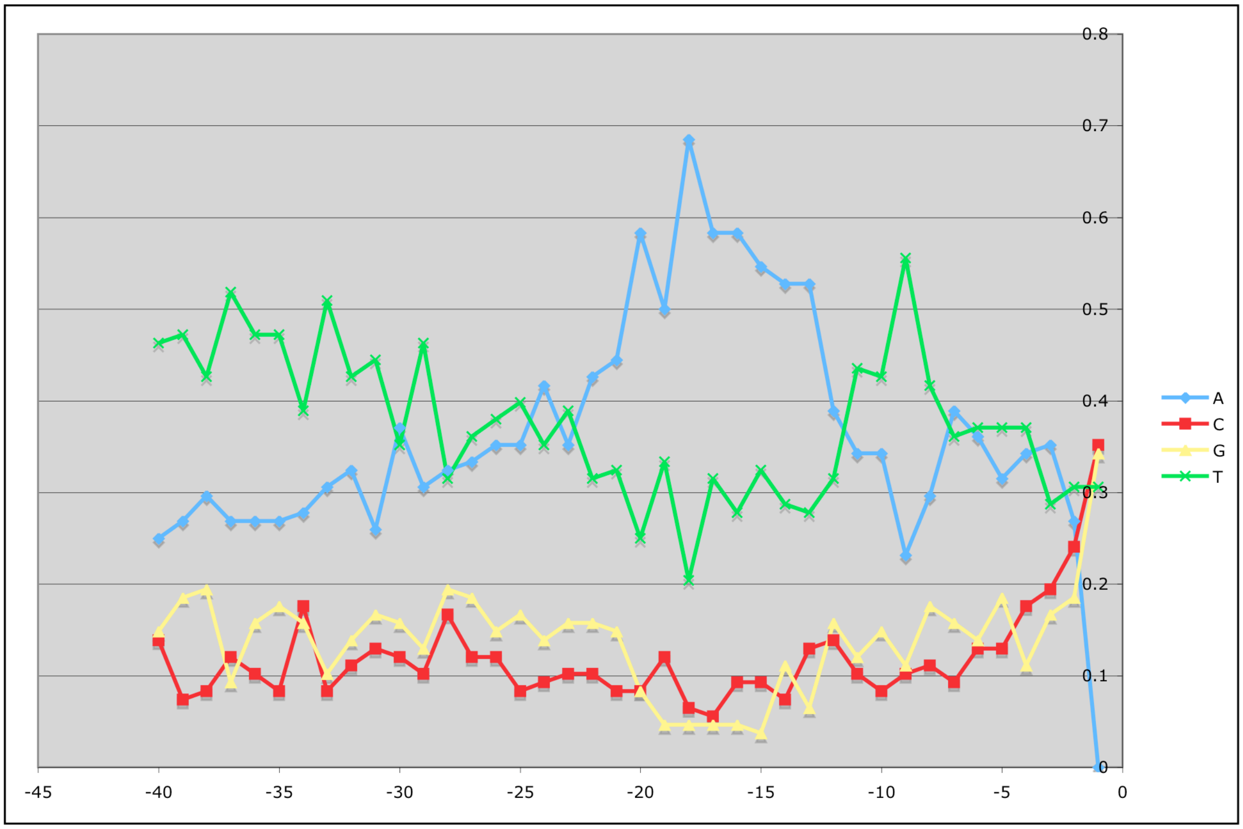


Figure 16. PolyA signal confirms 3'UTR in ESTs with transcripts.
Note the high percentage of adenines in the region 10-25 bases upstream. This high percentage correlates to the polyadenylation signal, AAUAAA.

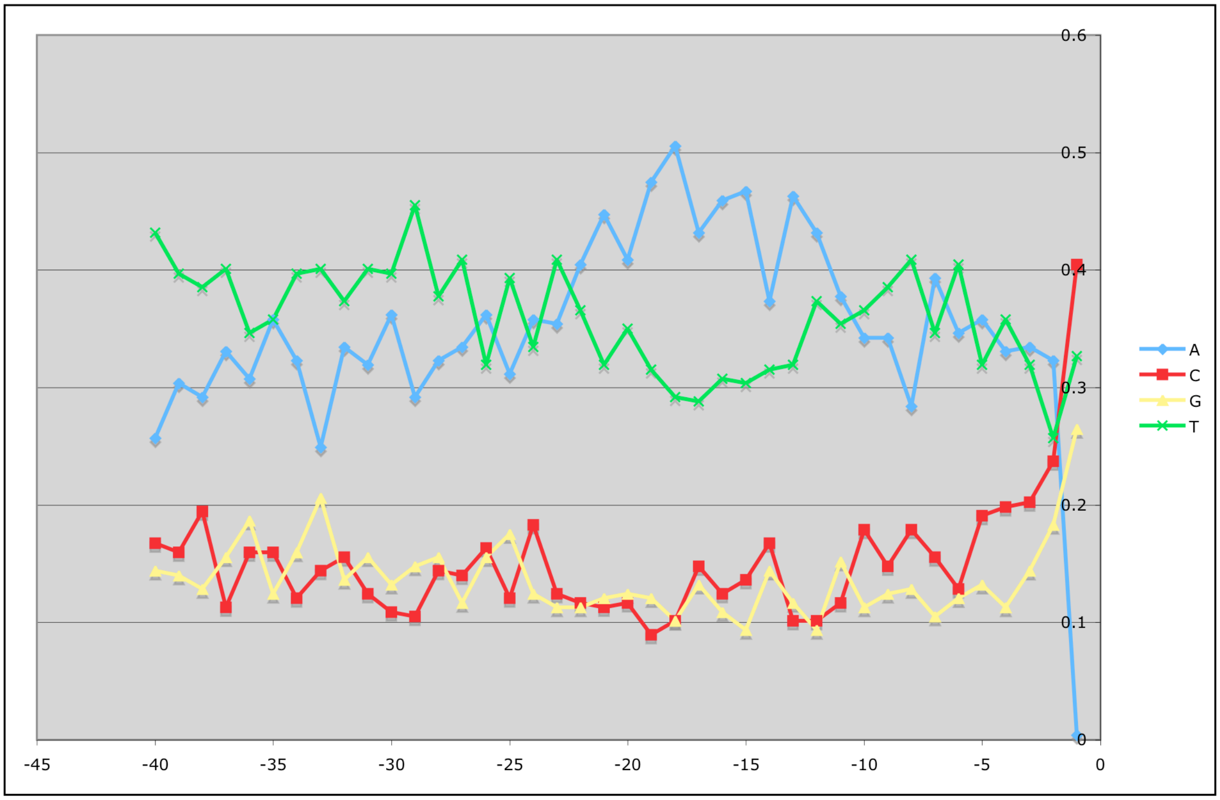


Figure 17. PolyA signal suggests polyA tail is real in many of the ncRNA. Note the high percentage of adenines in the region 10-25 bases upstream. This high percentage most likely correlates to the polyadenylation signal, AAUAAA, although it is not as strong as the signal found in sequences containing coding regions. This suggests that while most of the polyA tails are legitimate, some may not be, and these sequences will subsequently be missing the polyA signal.

IV. Analysis of All *Artemia* Contigs

IV.1. Construction of Artemia Contig Catalog

The Waksman Student Scholars Program and the Molecular Biology and Biochemical Research class at the Waksman Instutute at Rutgers University generated a total of 5,947 sequences from 2005 through 2008. Only about 15% of these sequences had been individually analyzed by the students and subsequently published on NCBI. Unfortunately, about 85% of the sequences generated were never analyzed individually. One of the objectives of my research was to analyze this large body of data to extract useful information regarding the *Artemia* genome, such as which proteins were expressed, and to what degree. I felt that this data was a more accurate representation of *Artemia's* protein expression profile.

These 5,947 sequences were initially analyzed to validate sequence quality. Vector and poor sequences were removed using the SeqMan application within the DNASTAR software. 'Poor sequences' were identified as sequences containing 3 or more Ns, or with a phred quality value of 12. Phred is a base-calling algorithm based on DNA sequencer trace data that indicates base-calling reliability. A phred value of 12 indicates about a one in ten chance of a miscalled base. Two-hundred-forty-nine sequences were removed due to low quality or sequence length less than 100nt, while 5,698 sequences (95.8%) were recognized as high quality. These high quality sequences were subsequently cropped of their vector sequence, as well as any poor sequence at the end. The Classic Assembler module of DNASTAR was then used to assemble these high quality sequences into 1,019 clusters and 1,829 single sequences based on sequence similarity, resulting in a total of 2,848 non-redundant sequences. These will be referred to as ‘contigs’.

The number of sequences that make up each contig varies from 2 (584 contigs) to 149 (1 contig) (Figure 18). The average length of the original sequences was 1,156 while the average length of the sequences once the vector was cropped was 732. The average phred quality score of the cropped high quality sequences was 41, indicating that the average base call was 99.99% accurate. All clusters made up of 4 or more sequences were then further analyzed individually to confirm vector sequence was removed, and also to confirm contig construction validity. All clusters containing 3 or less sequences that returned ‘alpha peptide’ from a BlastX search were also individually analyzed and vector sequences were cropped off by hand. A total of 41 contigs were created solely on identical vector sequence. 25 contigs were incorrectly created due to a minimal overlap. The longest assembled sequence is 2,955bp while the shortest assembled sequence is 101bp. The minimum sequence length allowed in the DNASTAR program was 100bp. The average contig length is 780bp. Creation of a contig required a minimum of 12 base pair overlap and at least 80% sequence similarity.

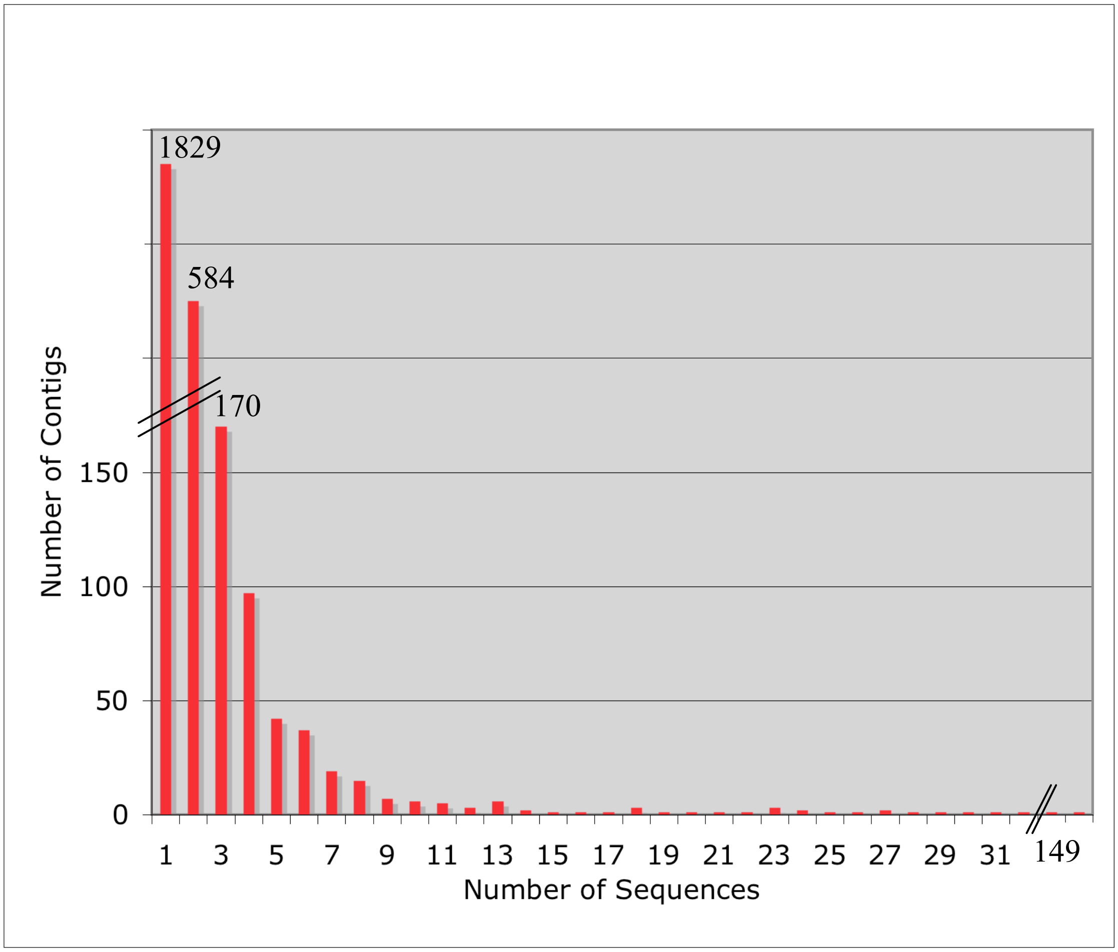


Figure 18 – Many contigs were made up of 4 or fewer sequences. 1829 contigs were made up of one sequence. 584 contigs were made up of two sequences. 170 contigs were made up of three sequences. One contig is made up of 149 sequences and codes for the 16S ribosomal RNA.

IV.2. GO Categorization

All non-redundant sequences were imported into the software program Blast2GO, a gene ontology, visualization and analysis tool [20]. A BlastX analysis was performed through Blast2GO on all 2,848 contigs, with an e-value cutoff of e^{-3}. Gene ontology (GO) identifications were obtained for all sequences that were recognized as genes. Only 380 contigs were classified as having GO identifications. Of these 380 sequences, further annotations were run in order to validate the gene matches. A BlastX annotation cutoff value of e^{-6} and a sequence similarity of at least 30% was used in order to strictly confirm the existence of a gene. Annotation was based on the highest BlastX hit similarity. Subsequently, only 268 contigs were of high enough quality to be considered annotated. The annotations of these 268 contigs were then further streamlined using GO Slim, a reduced version of the Gene Ontologies containing fewer nodes [21].

Given that the library used was not normalized, the number of clones returning a specific gene can be used to indicate gene expression to some degree. A number of genes were highly expressed, and these are indicated in Table 1. The top three genes returned, cytochrome c oxidase, NADH subunit 2, and ATP synthase f0 subunit 6, were mitochondrial. Genes related to ATP synthesis were identified in 250 of the 416 clones listed in the table, suggesting that *Artemia* have an active energy metabolism.

Gene Description	Number of clones that make up those contigs
Cytochrome c oxidase subunit 1	87
NADH subunit 2	44
ATP synthase f0 subunit 6	40
Cofilin actin-depolymerizing factor homolog	35
Cytochrome b	34
Cuticle protein	24
Cytochrome c oxidase subunit 3	22
NADH subunit 4	16
Myosin light chain 2	15
NADH subunit 6	13
mpv17 protein	12
Sodium/potassium transporting ATPase subunit B-2	12
Ferritin	11
Elongation factor 1 alpha	9
ADP ATP translocase	8
Cytochrome c oxidase subunit 2	6
NADH subunit 1	6
B-cell translocation gene 2	6
Eukaryotic translation elongtion factor 1 gamma	5
Eukaryotic translation initiation factors (4-6)	5
Ubiquitin-conjugating enzyme	4
Cytochrome c oxidase subunit 4	2

Table 1. Many of the sequenced genes are involved in ATP synthesis.
Note the large number of genes involved in ATP synthesis, designated by red font.

In order to carry out functional genomic studies in *Artemia*, the 268 contigs showing similarity with known genes or proteins were grouped into 10 categories according to Gene Ontology. Gene Ontology is widely used to classify genes according to molecular function, biological process and cellular component, as it provides the vocabulary necessary to accurately compare and contrast different organisms and their genes [21]. GO categories were assigned to 268 *Artemia* sequences with BlastX hits better than e^{-6} using the generic GO Slim feature of Blast2GO.

369 molecular functions were identified for the 268 contigs. These molecular functions identify the action characteristic of the gene product. A majority of the contigs (51.9%) contained genes with “binding function”, closely followed by 41.8% of the contigs whose genes displayed “catalytic activity” (Figure 19). The most common binding functions identified within our library of sequences were protein binding, nucleotide binding and nucleic acid binding. Hydrolase activity and transferase activity were the only two subcategories identified within the catalytic activity division. Other than the binding and catalytic functions, 20.9% of the genes contained some kind of “structural molecule activity”, while 10.4% contained genes with “transporter activity”, including, but not limited to, ion channel activity. The remainder of the molecular functions categories was identified in less than 5% of the contigs. Chen et al. [8] reported similar values when analyzing their *Artemia* sequences, with binding and catalytic activity classifications well exceeding all other molecular functions.

All 268 contigs were also categorized according to their biological processes, resulting in 593 identified biological processes. The most dominant biological process identified within the *Artemia* contigs was “cellular processes”, which was associated with 57.5% of the identified genes. This category included genes involved with gene expression, cell cycle and biogenesis. “Cellular processes” was followed closely by “metabolic processes”, which was assigned to 54.9% of the genes within the contigs and contained primary metabolic processes, biosynthetic processes, and cellular metabolic processes, which is cross listed as a subcategory under cellular processes as well (Figure 20). 23.1% of the genes were responsible for localization or the establishment of localization, while the remainder of the biological processes were identified in less than 20% of the contigs. Chen et al. [8] reported similar values when analyzing their sequences, as their classification resulted in a substantially large number of genes falling into the category of either cellular processes or metabolic processes as well.

All contigs were also categorized based on their cellular component; the location at which the gene products can be found in the cell. 691 categorizations were identified, many of which were localized in the “cell”, or “cell part”, 77.6 % 74.3% respectively. The “cell part” category is a subsection of the “cell” category (Figure 21). The next category most frequently identified (59.0%) was “organelle”, which contains membrane and non-membrane organelles, as well as intracellular organelles. “Macromolecular complexes”, identified in 35.0% of the sequences, followed this and contained the subcategories protein complexes and ribonucleoprotein complexes. Chen et al. [8] also reported that the largest majority of their contigs could be found in the cell or organelle.

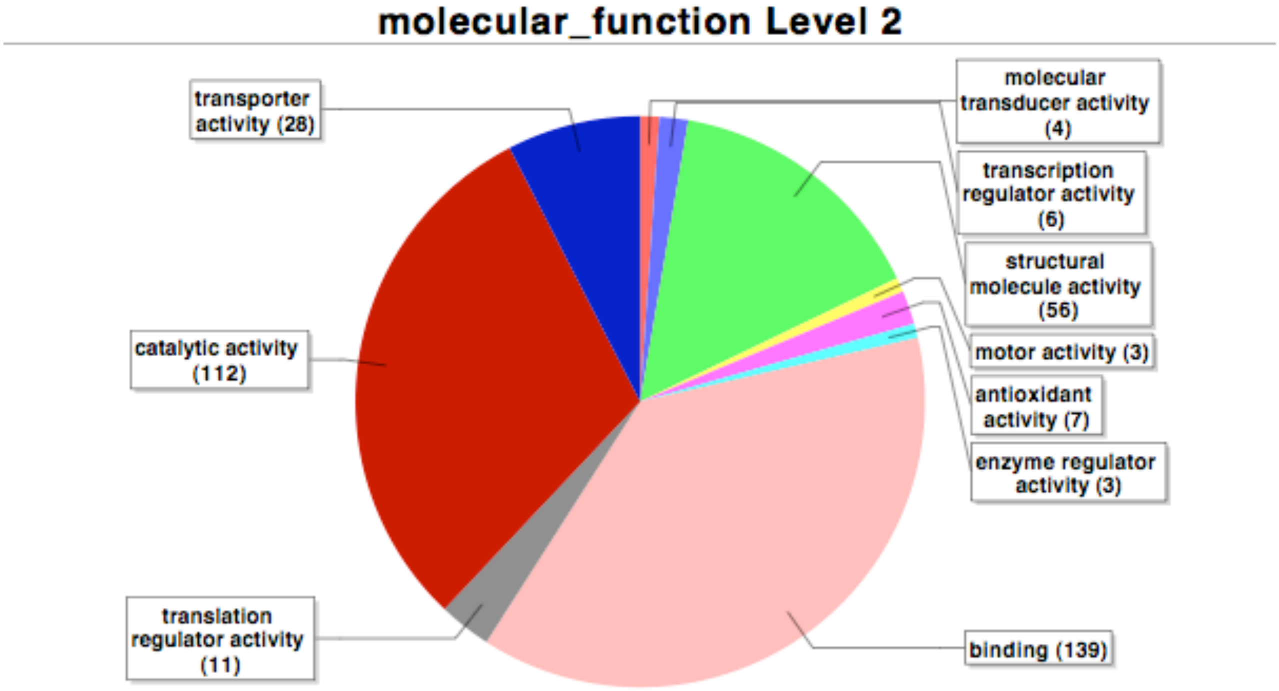


Figure 19. Molecular functions of the 268 annotated genes.
A large percentage of the genes possess binding (51.9%) or catalytic (41.8%) activity.

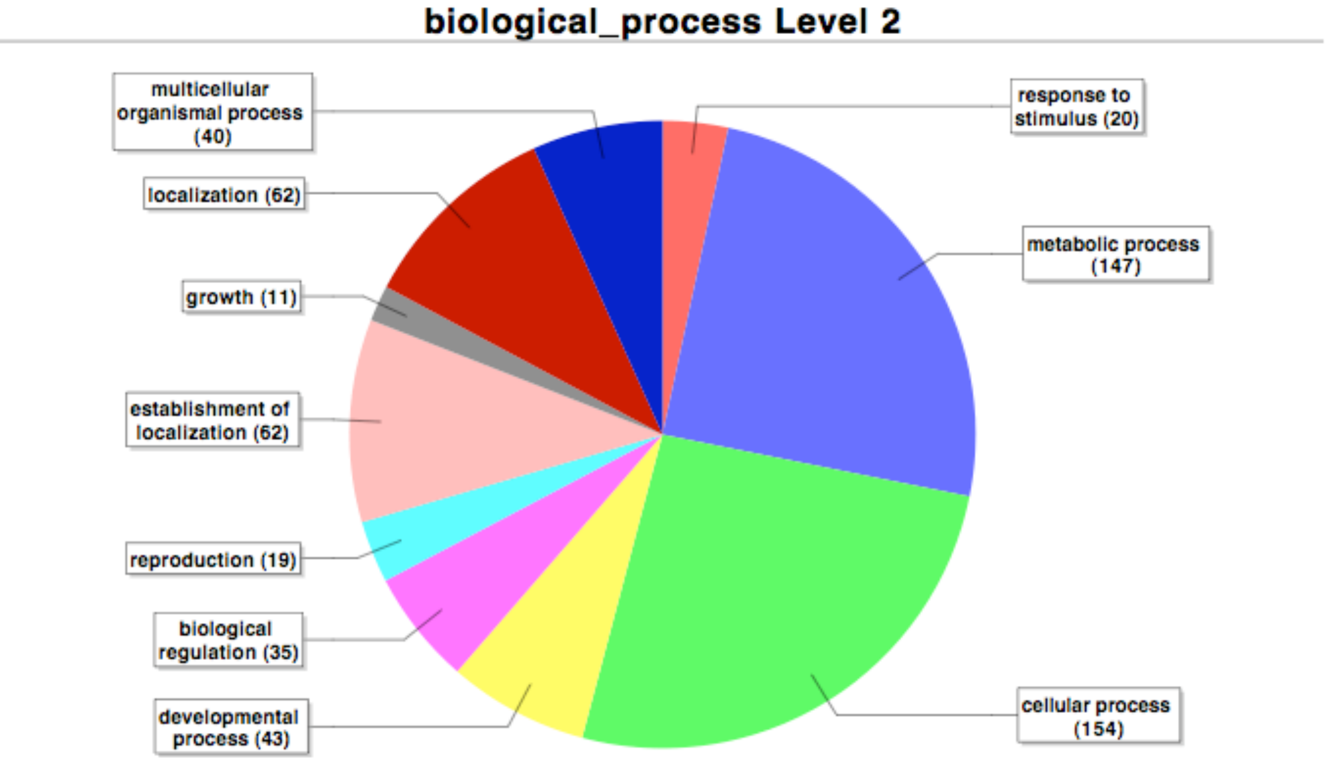


Figure 20. Biological processes of the 268 annotated genes.
A large percentage of the genes are responsible for cellular (57.5%) or metabolic (54.9%) processes.

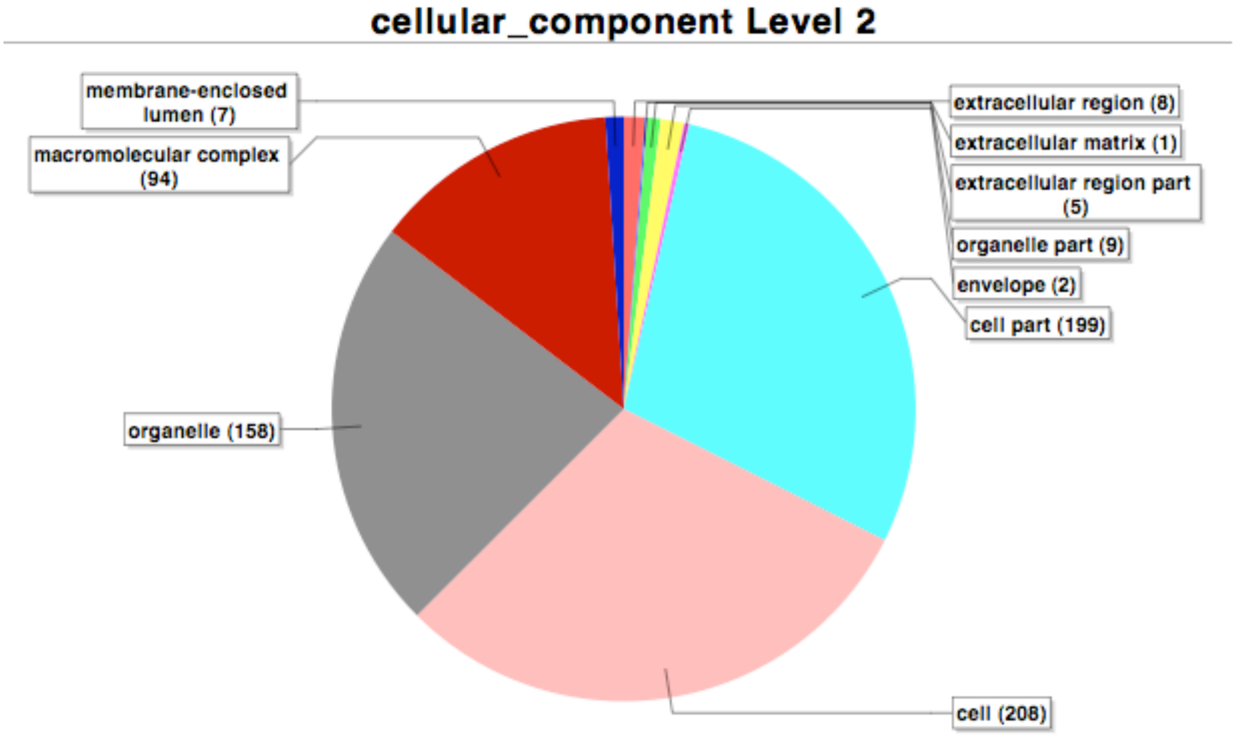


Figure 21. Cellular component of the 268 annotated genes.
A large percentage of the genes are located in the cell (77.6%) or cell part (74.3%) or organelle (59.0%).

IV.3. BlastX Analysis

Overall, only about 10% of the sequences imported into Blast2GO (268 of 2,848) were annotated and analyzed. With reference to the BlastX results, the organism with the most hits (216) was *Drosophila melanogaster*, followed by *Artemia franciscana* with 162 hits, and *Mus musculus* with 159 hits. A significantly large number of hits were made to other *Drosophila* species as well (Figure 22). This unexpected distribution of hits to fly instead of brine shrimp can be attributed to the large number of *Drosophila* sequences (2,835,902 nucleotide and 587,260 protein) in the public database as of June 2009 with respect to *Artemia franciscana* (38,186 nucleotide and 444 protein).

A similarity distribution chart was created to illustrate the number of BlastX hits that correlated to the percentage of similarity between the contigs and their BlastX matches. Figure 23a represents the distribution when all 2,848 sequences were graphed. A majority of the sequences have between 50-97% similarity to other sequences in the public database. Figure 23b reveals the distribution when only the 268 annotated sequences were graphed. The number of hits correlating to the original 50-97% similarity seen in Figure 23a drops significantly, revealing that a considerable number of contigs between 50-70% sequence similarity were not identified as containing genes. This kind of correlation could be used to create stricter guidelines for identifying likely genes from unknown sequences.

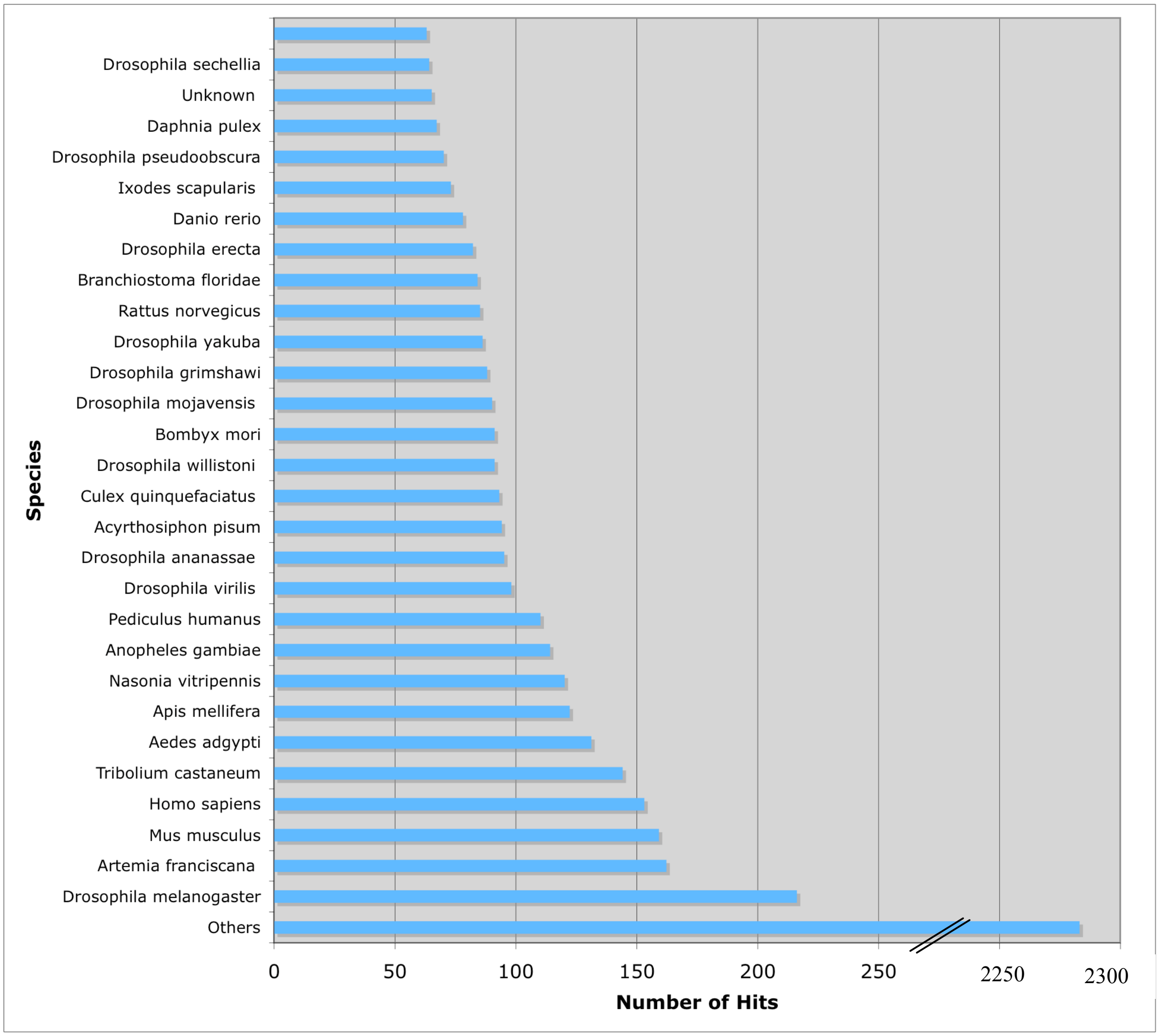


Figure 22. Most *Artemia* sequences hit *drosophila*.
This figure identifies the species distribution from the top 20 BlastX hits for each of the 268 annotated sequences. Note the high number of *Drosophila* hits.

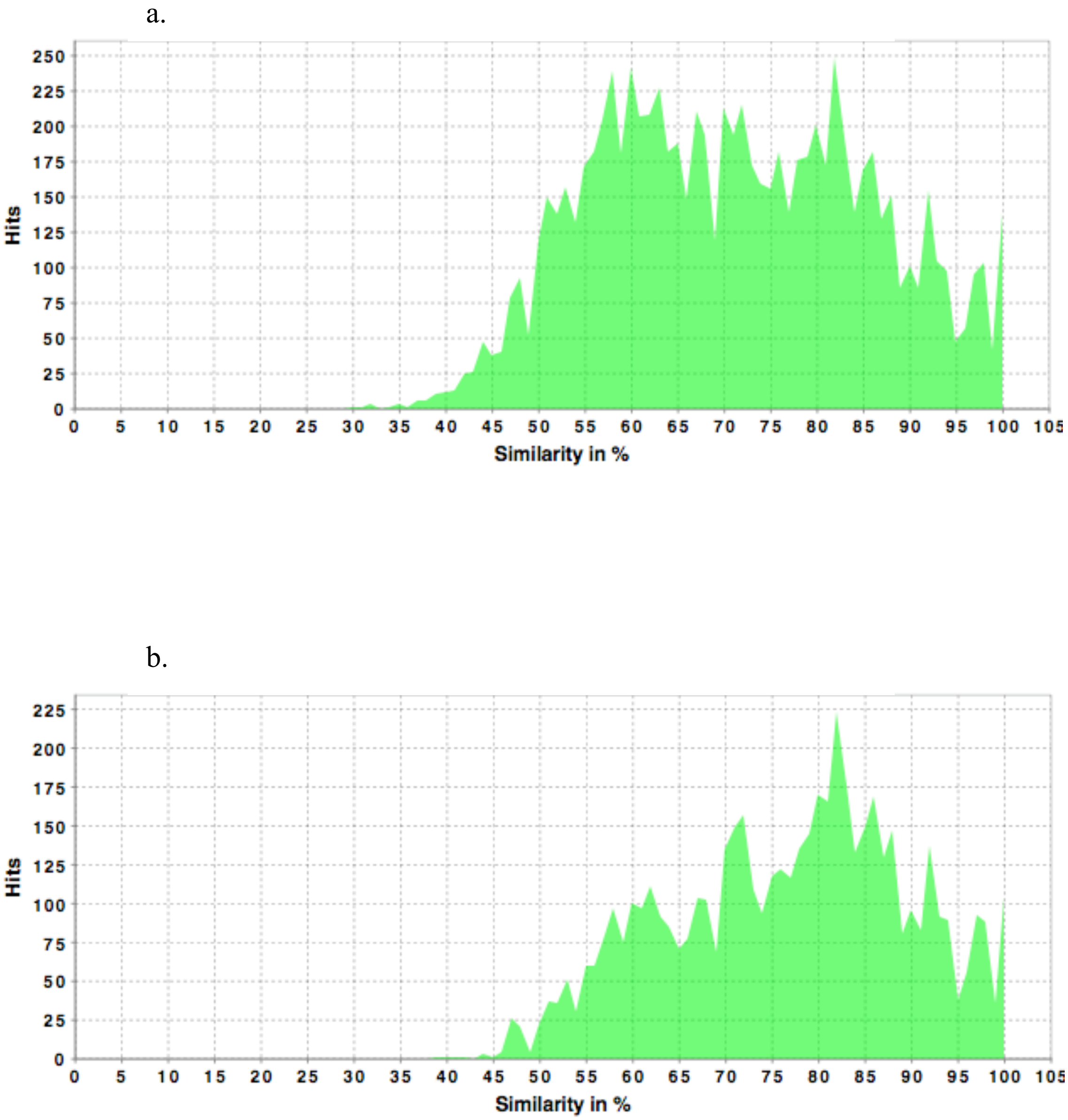
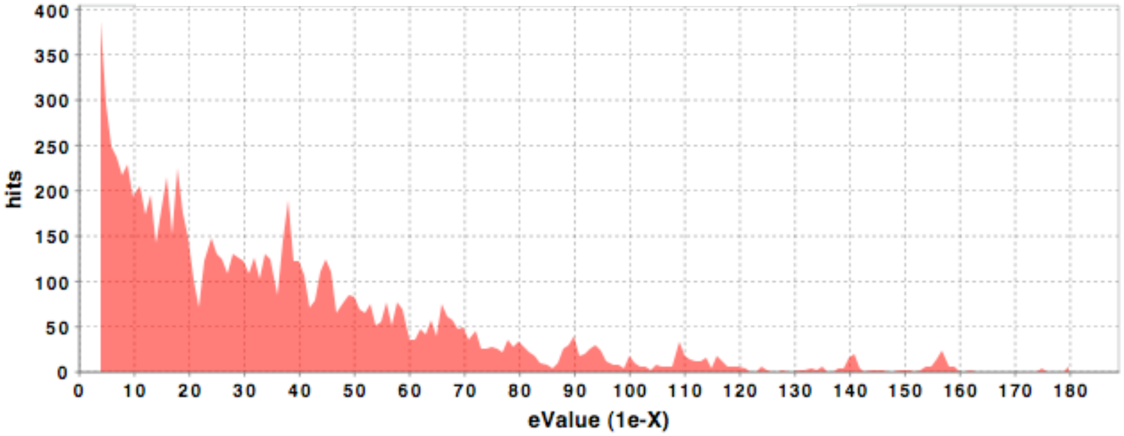


Figure 23. The 268 annotated Artemia ESTs have higher BlastX similarities with published sequences.
a. This illustrates a BlastX similarity distribution of all 2,848 sequences to other sequences posted in NCBI.
b. This illustrates a BlastX similarity distribution of the 268 annotated

An e-value distribution chart was generated in order to display the distribution of e-values returned from the BlastX analysis. Figure 24a confirms that a large distribution of hits returned an e-value ($1e^{-x}$) of less than 30. The number of hits returned with each e-value gradually decreases as the e-value increases. This should not be surprising, as only significant matches will return high e-values. Figure 24b illustrates the e-value distribution of the BlastX hits from the 268 annotated sequences. It is evident from this chart that many of the hits returning e-values of 30 or less have been removed. However, the number of hits returning e-values higher than 30 seem to have remained, indicating that the e-values of 30 or higher were returned from the 268 annotated sequences. This information could be useful when setting cut-offs for identification of a gene product from an unknown sequence, as BlastX hits returning e-values of higher than 30 were typically identified as gene products that were able to be annotated, indicating a high level of confidence.

a.



b.

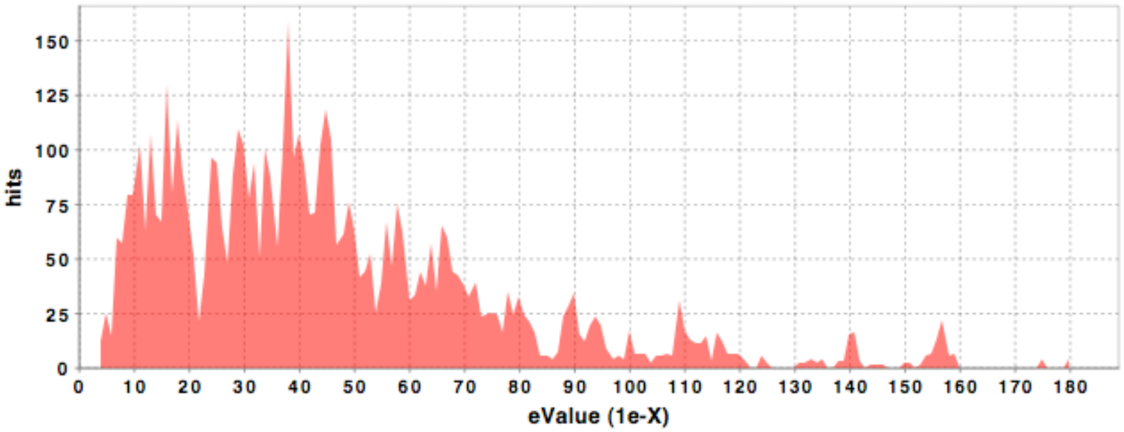


Figure 24. E-value distribution
a. E-value distribution of all 2,848 sequences to other sequences posted in NCBI
b. E-value distribution of the 268 annotated sequences to other sequences posted in NCBI

V. Discussion

The Waksman Student Scholars Program, along with the Introduction to Molecular Biology and Biochemical Research class, were responsible for the publication of 628 *Artemia* sequences. Surprisingly, 361 of these sequences (58%) were non-coding. It was originally presumed that this was due to a high level of genomic DNA contamination. While it is possible that some of our *Artemia* sequences are genomic contamination, I hypothesize a large majority of our non-coding sequences are long non-coding RNA (ncRNA). The high percentage of non-coding sequences is reasonable, as is the length of these sequences. Furthermore, some of these non-coding sequences contain polyA tails, similar to other ncRNAs, as well as the polyA signal.

Long (>200) non-coding RNAs have recently been identified as key regulatory molecules in the cell. The very act of transcribing the ncRNA has been linked to transcriptional regulation. For example, when the ncRNA is transcribed across the promoter region of the gene, it may directly interfere with transcription initiation [22]. Transcription can also be repressed through the cooperative efforts of ncRNA and histone modification. This is seen in *PHO84* where the accumulation of antisense RNAs leads to targeted histone deacetylation and the silencing of *PHO84* sense transcription [23]. The ncRNA molecule itself can also be responsible for transcriptional regulation. One particular ncRNA, *HSR1*, is required for heat-shock transcription factor 1 activation [24]. In addition to binding to proteins, some ncRNAs are known to bind to miRNAs, rendering them incapable to interfere with translation. Long ncRNAs, like their shorter cousins miRNA and siRNA, are also believed to be misregulated in many diseases, including cancer [25].

Since the FANTOM consortium indicated that 48% of mouse RNA is non-coding, and *daphnia* contains 59% ncRNA, it is reasonable to believe that our non-coding sequences (58%) are long ncRNA, and not contamination. Of the 361 non-coding sequences, 189 sequences (52%) matched other *Artemia* sequences in NCBI. Most of these matches were to the *Artemia* sequences published by Chen et al. [8] Their protocol also selected for mRNAs using the polyA tail. It is possible, albeit unlikely, that both groups had contaminated cDNA libraries. However, it is more likely that both groups extracted long ncRNA transcripts, in addition to their mRNA transcripts, considering their apparent prevalence within the cell.

62 of the 189 sequences (33%) matched other *Artemia* sequences in the reverse direction. 3 of these sequences had ORFs larger than 100AA in the reverse direction. While there are many methods in which ncRNA functions, one of them is to transcribe a message in the reverse direction, subsequently inhibiting the gene to be transcribed in the correct direction. Identifying these sequences as ncRNA provides a reasonable explanation for their reverse direction. These 3 sequences in particular deserve further investigation.

While it is possible that some of these long UTRs are in fact, 5' and 3'UTRs, it is more likely that they are ncRNAs. The average length of the 361 sequences that did not contain a coding region is 600nt, while the average length of the 3'UTRs following known genes is 175nt. While some *Artemia* 3'UTRs are known to be longer than 175nt, ncRNAs are identified as being significantly longer than 175nt. *H19,* the first imprinted ncRNA locus to be discovered, produces a 2.3 kb transcript [26]. The *roX* genes, responsible for binding to the X chromosome in male *Drosophila*, are 3.7 kb and 1 kb [27]. It is more likely that our lengthy UTRs are long ncRNAs rather than long *Artemia* 5' or 3'UTRs, considering how infrequently these lengthy UTRs occur.

Many of our ncRNAs are polyadenylated. This suggests that they could be the 3'UTR of *Artemia* mRNA transcripts. However, many long ncRNAs, including, but not limited to, *H19*, *roX*, *Xist*, and *Air* (*antisense Igf2r*), are also polyadenylated [28]. Considering many long ncRNAs are known to contain a polyA tail, it is reasonable to believe that our long polyadenylated non coding sequences are ncRNAs, and not simply long 3'UTRs. The polyA signal present in many of these transcripts also disproves the idea that these long non coding sequences are genomic contamination.

Research has yet to be conducted concerning the prevalence of non-coding RNAs in *Artemia*. I believe our quantity and quality of data provides an excellent starting point. What was once considered genomic sequence, unable to provide us with any useful information, is now novel data, enabling us to continue expanding our understanding of one of the newest players in transcriptional regulation – long noncoding RNAs.

References

1. http://www.reference.com/browse/wiki/Artemia.

2. Vershon A: **Chapter 1 - Vectors and Libraries**. *WSSP-08 Lecture and Lab Files* 2008:8.

3. Lavens P, Sorgeloos P: **Manual on the Production and Use of Live Food for Aquaculture**. 1996:295.

4. Albert: **Natura Mediterraneo**. *http://wwwnaturamediterraneocom/Public/data3/albert2006/Artemia%20salina%20m%20e%20%20fjpg_200648174214_Artemia%20salina%20m%20e%20%20fjpg* 2006.

5. Browne RA, Sorgeloos P, Trotman CNA: **Artemia Biology**. 1991:374.

6. Abatzopoulos TJ, Beardmore JA, Clegg JS, Sorgeloos P: **Artemia: Basic and Applied Biology (Biology of Aquatic Organisms)**. 2002.

7. Warner AH, Clegg JS: **Diguanosine nucleotide metabolism and the survival of artemia embryos during years of continuous anoxia**. *Eur J Biochem* 2001, **268**(6):1568-1576.

8. Chen WH, Ge X, Wang W, Yu J, Hu S: **A gene catalogue for post-diapause development of an anhydrobiotic arthropod Artemia franciscana**. *BMC Genomics* 2009, **10**:52.

9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

10. Zhang Z, Schwartz S, Wagner L, Miller W: **A greedy algorithm for aligning DNA sequences**. *J Comput Biol* 2000, **7**(1-2):203-214.

11. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

12. Min XJ, Butler G, Storms R, Tsang A: **OrfPredictor: predicting protein-coding regions in EST-derived sequences**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W677-680.

13. Watanabe H, Tatarazako N, Oda S, Nishide H, Uchiyama I, Morita M, Iguchi T: **Analysis of expressed sequence tags of the water flea Daphnia magna**. *Genome* 2005, **48**(4):606-609.

14. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J *et al*: **Antisense transcription in the mammalian transcriptome**. *Science* 2005, **309**(5740):1564-1566.

15. De Pitta C, Bertolucci C, Mazzotta GM, Bernante F, Rizzo G, De Nardi B, Pallavicini A, Lanfranchi G, Costa R: **Systematic sequencing of mRNA from the Antarctic krill (Euphausia superba) and first tissue specific transcriptional signature**. *BMC Genomics* 2008, **9**:45.

16. Li YC, Korol AB, Fahima T, Nevo E: **Microsatellites within genes: structure, function, and evolution**. *Mol Biol Evol* 2004, **21**(6):991-1007.

17. MISA s: **MISA: MIcroSAtellite Identification Tool**. 2002.

18. Louro R, Smirnova AS, Verjovski-Almeida S: **Long intronic noncoding RNA transcription: expression noise or expression choice?** *Genomics* 2009, **93**(4):291-298.

19. Mazumder B, Seshadri V, Fox PL: **Translational control by the 3'-UTR: the ends specify the means**. *Trends Biochem Sci* 2003, **28**(2):91-98.

20. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics* 2005, **21**(18):3674-3676.

21. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.

22. Martens JA, Laprade L, Winston F: **Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene**. *Nature* 2004, **429**(6991):571-574.

23. Camblong J, Iglesias N, Fickentscher C, Dieppois G, Stutz F: **Antisense RNA stabilization induces transcriptional gene silencing via histone deacetylation in S. cerevisiae**. *Cell* 2007, **131**(4):706-717.

24. Shamovsky I, Ivannikov M, Kandel ES, Gershon D, Nudler E: **RNA-mediated response to heat shock in mammalian cells**. *Nature* 2006, **440**(7083):556-560.

25. Perez DS, Hoage TR, Pritchett JR, Ducharme-Smith AL, Halling ML, Ganapathiraju SC, Streng PS, Smith DI: **Long, abundantly expressed non-coding transcripts are altered in cancer**. *Hum Mol Genet* 2008, **17**(5):642-655.

26. Brannan CI, Dees EC, Ingram RS, Tilghman SM: **The product of the H19 gene may function as an RNA**. *Mol Cell Biol* 1990, **10**(1):28-36.

27. Meller VH, Rattner BP: **The roX genes encode redundant male-specific lethal transcripts required for targeting of the MSL complex**. *Embo J* 2002, **21**(5):1084-1091.

28. Prasanth KV, Spector DL: **Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum**. *Genes Dev* 2007, **21**(1):11-42.