

Patterns and Signals of Biology: An Emphasis On The Role of Post Translational Modifications in Proteomes for Function and Evolutionary Progression

by
Oliver Bonham-Carter

A DISSERTATION

Presented to the Faculty of
The Graduate College at the University of Nebraska
In Partial Fulfillment of Requirements
For the Degree of Doctor of Philosophy

Major: Information Technology

Under the Supervision of Dhundy Bastola, PhD and Hesham Ali, PhD
Omaha, Nebraska

May, 2016

SUPERVISORY COMMITTEE

Chair: Dhundy (Kiran) Bastola, PhD
Co-Chair: Hesham Ali, PhD
Lotfollah Najjar, PhD
Steven From, PhD

Abstract

Patterns and Signals of Biology: An Emphasis On The Role of Post Translational Modifications in Proteomes for Function and Evolutionary Progression

Oliver Bonham-Carter, MS, PhD

University of Nebraska, 2016

Advisors: Dhundy (Kiran) Bastola, PhD and Hesham Ali, PhD

After synthesis, a protein is still immature until it has been customized for a specific task. Post-translational modifications (PTMs) are steps in biosynthesis to perform this customization of protein for unique functionalities. PTMs are also important to protein survival because they rapidly enable protein adaptation to environmental stress factors by conformation change. The overarching contribution of this thesis is the construction of a computational profiling framework for the study of biological signals stemming from PTMs associated with stressed proteins. In particular, this work has been developed to predict and detect the biological mechanisms involved in types of stress response with PTMs in mitochondrial (Mt) and non-Mt protein.

Before any mechanism can be studied, there must first be some evidence of its existence. This evidence takes the form of signals such as biases of

biological actors and types of protein interaction. Our framework has been developed to locate these signals, distilled from “Big Data” resources such as public databases and the the entire PubMed literature corpus. We apply this framework to study the signals to learn about protein stress responses involving PTMs, modification sites (MSs). We developed of this framework, and its approach to analysis, according to three main facets: (1) by statistical evaluation to determine patterns of signal dominance throughout large volumes of data, (2) by signal location to track down the regions where the mechanisms must be found according to the types and numbers of associated actors at relevant regions in protein, and (3) by text mining to determine how these signals have been previously investigated by researchers. The results gained from our framework enable us to uncover the PTM actors, MSs and protein domains which are the major components of particular stress response mechanisms and may play roles in protein malfunction and disease.

Copyright

Copyright 2016, Oliver Bonham-Carter.

We never work alone in this
business.

Oliver Bonham-Carter

Dedication

I dedicate this work to my wife; Janyl and son; Vincent, my sisters; Kate and Daisy, my brother; Alexander, and my mother and father; Jane and Nicholas. These are the people who I, likely, subjected to many long-winded and unwarranted updates about the progression of this document and that of my degree, in general.

This work is also dedicated to my advisors Dr. Dhundy (Kiran) Bastola, and Dr. Hesham Ali who served on my doctorate committee with Dr. Lotfollah Najjar and Dr. Steven From. By your constant and tireless efforts, mentoring, and coaching, I have been inspired to create an exciting work of scientific investigation.

I would also like to thank the following wonderful people for their constant warmth, support and friendship: Ishwor Thapa, Dr. Jasjit Banwait-Kaur, Sean West and Scott McGrath. It is very likely that I also subjected these good people to far too many general updates on this work and I thank them earnestly for their counsel.

This work is also dedicated to my friends from the bioinformatics lab at the University of Nebraska at Omaha: Dr. Sanjukta Bhowmick, Dr.

Kate Cooper, Julia Warnke, Vladimir Ufimtsev, Jay Pedersen, Kritika Karri, Kaitlin Goettsch, Asuda Sharma, Sunandini Sharma, Suyeon Kim, Sahil Sethi and many others too!

Without the encouragement from my family and friends, none of this work would have been possible to return to our scientific community, from whom I have taken so much.

Contents

1	Introduction	1
1.1	Signals After Mechanisms	1
1.2	A Focus On PTMs, Proteins And Their Typical Stress Responses	2
1.2.1	Protein Regulation And Modification	4
1.2.2	Sudden Stresses	7
1.2.3	PTM Localization And Study	8
1.2.4	Protein Domains	9
1.3	Mitochondrial And Non-Mitochondrial Protein	10
1.3.1	Mitochondria: Energy Production And Stress Response	12
1.3.1.1	An Energy Crisis	13
1.3.2	Mechanisms From Signal Detection	16
2	The Ubiquitous Nature Of Signals In Biology And Living Systems	18
2.1	Traffic Signals	18
2.2	Signals Of The Heart	21
2.3	Natural Signals Of Organismal Biology	23
2.3.1	Mobile Organisms	23
2.3.2	Immobile Organisms	25
2.4	Signals, Semantics And Comprehension	27

3 Objectives	29
3.1 Motivation	29
3.2 General Organization	30
3.3 Thesis Contributions	31
3.3.1 Contribution 1	31
3.3.2 Contribution 2	31
3.3.3 Contribution 3	32
3.3.4 Contribution 4	32
3.3.5 Contribution 5	33
3.3.6 Contribution 6	33
3.3.7 Contribution 7	34
3.3.8 Contribution 8	34
3.3.9 Contribution 9	35
3.3.10 Contribution 10	35
4 Alignment-Free Genetic Sequence Comparisons: A Review of Recent Approaches By Word Analysis	37
4.1 Abstract	37
4.2 Introduction	38
4.3 Background	41
4.4 Factor Frequencies	41
4.4.1 BBC By Analysis Of Mutual Information	42
4.4.2 Feature Frequency Profiles (FFPs)	44
4.4.2.1 A Comparison By Frequencies And The Jensen-Shannon Divergence Test	46
4.4.3 Suffix Trees By k -mer Frequencies	49
4.4.4 Composition Vectors Based On k -mer Frequencies	50
4.4.4.1 Creation Of Composition Vectors, (CVs, CCVs)	51

4.4.4.2	Creation Of Improved CCV's	52
4.4.4.3	Distance Measurement	53
4.4.5	A Revised String Composition Method	54
4.4.6	D ₂ Statistic	56
4.5	Data Compression And Dictionaries	59
4.5.1	Text Compression Algorithms	61
4.5.2	Average Common Substring (ACS)	62
4.6	Applications Of Alignment-Free Methods	64
4.6.1	Biological Data And Sequence Assembly	64
4.6.2	Chromosomal Data And Phylogeny	65
4.6.3	Horizontal Gene Transfer	66
4.7	Advantages And Disadvantages Of Methods	68
4.8	Conclusion	70
4.9	Article Details	71
5	An Analysis Of Palindromic Content In The Coding And Non-coding DNA Regions Of Bacteria	73
5.1	Abstract	73
5.2	Introduction	74
5.3	Methods	74
5.4	Results And Discussion	77
5.4.1	Lengths-{4,6,8,10} Palindromes	78
5.5	Conclusions	79
5.6	Article Details	80
6	A Base Composition Analysis Of Natural Patterns For The Pre-processing Of Metagenome Sequences	81
6.1	Abstract	81

6.2	Introduction And Related Work	82
6.3	Methods	85
6.3.1	Genome Sequences	85
6.3.2	Read And Contig Sequences	86
6.3.3	Motifs	87
6.3.3.1	Base Compositions And Spectrum Sets	87
6.3.4	Proportions	90
6.4	Results And Discussion	91
6.4.1	Sequence Data	92
6.4.2	<i>Clostridium</i> And <i>Staphylococcus</i>	93
6.4.3	Proportional Differences In Contigs By Spectrum Sets	94
6.4.3.1	Tests To Determine Pliable Spectrum Sets	96
6.4.3.2	Removal Of The Contrasting Contig Group	97
6.5	Phylogeny From Full Chromosomes	100
6.6	Conclusion	105
6.7	Article Details	107
7	sEncrypt - An Encryption Algorithm Inspired From Biological Processes	108
7.1	Abstract	108
7.2	Introduction And Related Work	109
7.2.1	Background On DNA To Protein Translation	112
7.2.2	Lossy Biological Functions	113
7.3	Methods	113
7.3.1	Phase One	115
7.3.1.1	Latin Squares	115
7.3.1.2	Sequence Encoding	115
7.3.1.3	Choosing The Encryption Key	116

7.3.1.4	Encryption	117
7.3.2	Phase Two: Translation Of DNA To Protein Code	118
7.3.2.1	Huffman Codes From Triplet Frequencies	119
7.3.2.2	Encoding Triplet Codes By Protein Amino Acid Codes	120
7.3.3	Decryption	121
7.4	Results And Discussion	122
7.4.1	Entropy	123
7.4.2	Autocorrelation Evaluation	124
7.4.3	General File Reductions	125
7.5	Conclusion And Future Work	126
7.6	Article Details	127
8	Evidence Of A Pathway Of Reduction In Bacteria: Reduced Quantities Of Restriction Sites Impact tRNA Activity In A Trial Set	128
8.1	Abstract	128
8.2	Introduction	129
8.2.1	Palindromes And Restriction Enzymes	129
8.2.2	Methylation And Damage Control	130
8.2.3	The Pathway Of Reduction	131
8.3	Methods	132
8.3.1	Data Collection	132
8.3.2	Regression Models	133
8.3.3	An Analysis Of tRNAs	134
8.3.4	Biological Importance	134
8.3.5	Model Building By Stepwise Regression	136
8.4	Results	138
8.4.1	Predictor Significance	138

8.4.2	An Analysis By Transfer RNA Composition	140
8.4.3	Available tRNAs	143
8.5	Discussion	143
8.6	Conclusions	145
8.7	Article Details	146
9	Evidence Of Post-translational Modification Bias Extracted From The tRNA And Corresponding Amino Acid Interplay Across A Set Of Diverse Organisms	147
9.1	Abstract	147
9.2	Introduction	148
9.2.1	PTM Bias	148
9.2.2	tRNA Bias	151
9.2.3	Amino Acid Bias	152
9.3	Methods	153
9.3.1	Diverse Organisms	153
9.3.2	Computing Frequencies	155
9.3.3	Network Models	156
9.4	Results and Discussion	157
9.4.1	Notable PTMs	163
9.5	Conclusions	166
9.6	Article Details	169
10	A Content And Structural Assessment Of Oxidative Motifs Across A Diverse Set Of Life Forms	170
10.1	Abstract	170
10.2	Introduction	171
10.2.1	The Effects Of Weightlessness On Mitochondrial Function . .	171

10.2.2 Carbonylation And PEST Protein Regulation Mechanics	173
10.2.3 Mitochondria	175
10.3 Methods	177
10.3.1 Protein Sequence Data From Organisms	177
10.3.2 RKPT Motifs - Attractors Of Oxidative Carbonylation	179
10.3.3 A Sequence Analysis By <i>l</i> -Word Proportions	181
10.4 Cluster Analysis	182
10.4.1 A Comparison Across Organism Sequence Data	184
10.4.2 Statistical Analysis	185
10.5 Structural Analysis	188
10.5.1 Grouping Structural Elements	192
10.5.2 Structural Results	193
10.6 Discussion	196
10.7 Conclusions	198
10.8 Article Details	203
11 A Study of Bias And Increasing Organismal Complexity From Their Post-Translational Modifications And Modification Site Interplays	204
11.1 Abstract	204
11.2 Introduction	205
11.2.1 PTMs	205
11.2.2 Biases	206
11.2.2.1 Research Statement	207
11.3 Methods	209
11.3.1 Organismal Protein Samples	210
11.3.2 Computing Frequencies	212
11.3.3 Building Heatmaps And Networks	216
11.4 Results And Discussion	217

11.4.1 Heatmaps	217
11.4.2 Domains And PTMs	220
11.4.3 Networks	222
11.4.4 Pearson's And Kendall-Tau Correlation	228
11.4.4.1 Mt And Non-Mt Networks	228
11.4.5 Bias Analysis Using Gene Ontologies	232
11.4.6 Poisson Approximation By The Chen-Stein Method	233
11.4.7 Protein Isoforms In Organisms	237
11.4.8 Notable PTMs	237
11.5 Conclusions	239
11.6 Article Details	241
12 A Text Mining Application For Linking Functionally Stressed-Proteins To Their Post-Translational Modifications	242
12.1 Abstract	242
12.2 Introduction	243
12.3 Methods	246
12.4 Results And Discussion	249
12.5 Conclusion	251
12.6 Article Details	252
13 PTM Tracker: A System For Determining Trends Of PTM Modification Sites Relative To Protein Domains	253
13.1 Abstract	253
13.2 Introduction	254
13.3 Methods	256
13.3.0.1 PTMs	257
13.3.0.2 Modification Sites	258

13.3.0.3 Domains	258
13.3.0.4 Plots	259
13.3.0.5 Heatmaps	260
13.3.1 Organism-Centric Study Of PTM Distances	260
13.3.1.1 Three Criteria Of Proximity	260
13.3.1.2 MSs Found Before A Domain	261
13.3.1.3 MSs Found Inside A Domain	261
13.3.1.4 MSs Found After A Domain	262
13.3.2 Domain-Centric Study Of PTM Distances	262
13.3.2.1 MSs Found Relative To Domains	263
13.4 Results And Discussion	263
13.4.0.1 Graphical Interpretation - Organism-Centric Study Of PTM Distances	265
13.4.0.2 Graphical Interpretation - Domain-Centric Study Of PTM Distances	271
13.4.0.3 Analyses By Heatmaps	274
13.5 Conclusions	276
13.6 Article Details	279
14 Conclusions	285
14.1 Gains From The Contributions	287
14.2 Concluding Thoughts	292
14.3 Future Works	295
References	323

List of Tables

4.1 Positions 1 through 16 of the table of vectors for V_{S_1} from $S_1 = \text{ACGTGCTATG}$ and V_{S_2} from $S_2 = \text{ACGCGCTA}$, aligned with position. The elements of combined vector \mathbf{M} by index is also shown. Frequencies of each 2-mer are made by normalizing the occurrences of each 2-mer in S_1 and S_2 , respectively, by the total number of 2-mer occurrences in each sequence.	47
4.2 The similar and different chunks, taken in order from each sequence.	63
4.3 The sequences to compare by the <i>alfy</i> method. To compare sequences, we find the shortest sequence in the query (S_Q) which is absent from a subject.	67
4.4 S_1 is compared to S_Q to determine the shortest substring in S_Q which is absent from S_1 . The matching numbers indicate the shortest unique substring starting at this position that is absent from the subject.	67
4.5 Sequences S_1 and S_2 are compared to S_Q . The matching numbers indicate the shortest unique substring starting at this position that is absent from the subject. The HGT is described by a string of S_1 and S_2 characters to indicate where the subsequences likely originated.	67
4.6 We wish to determine the sequence relations based on common sequence material. The query sequence is S_Q and subjects are S_1 through S_3	67

4.7 Summary of the discussed methods in this contribution. The column “Alignment” contains the best suggested use of the method.	70
5.1 The percentage of the exhaustive lists of all possible palindromes (lengths 4, 6, 8 and 10) which are found in higher proportions in the non-coding regions than the coding regions, according to their significant p-values (Mann-Whitney tests). The row “ $p < 0.05$ only” excludes the set from $p < 0.01$ and indicates that these palindromes were not as significant as the $\alpha = 0.01$ group. Each column of this table correlate to our listed Hypothesis 1 through 4.	77
6.1 This table displays the genera used in our study. The <i>Read Originator</i> column displays the sequences which we processed via MetaSim for its reads. To determine their general associative behaviors, we studied ten trials of five freshly drawn reads. We chose two organisms from each of the three divisions from which to draw our contigs.	86
6.2 The numbers of available motifs belonging to each spectrum. The motifs in the spectrum set are non-palindromic and are permutations of the set seeds. The set created from the permutation of AATTAG is called, the <i>AATTAG</i> -spectrum, for example.	90
6.3 The organisms used in the base composition experiment. We note that rabbit, dog, mouse and rat are seemingly more closely related than the bacteria, fruit fly and the worm. This observation is used as a <i>first-glance</i> assessment of the heatmaps below.	102

6.4 Ranking of Spectrum Sets over Chromosomal Data: We note the best to worse resemblance of spectrum set trees to actual taxonomy data. For this data set, the <i>CCGGAT</i> -spectrum set created a tree which most closely resembled the one based on the classification in NCBI taxonomy database in Figure 6.14.	105
7.1 A Summary of Steps. The plain text moves through phases one and two before becoming the cipher text.	114
7.2 The quasigroup table used.	115
8.1 The organisms, their abbreviations and the type of data used in our study. This selection of organisms is the from Gelfand and Koonin's published results ^[101] . We note that "Mito" and "Chloro" indicate "mitochondria" and "chloroplasts," respectively.	133
8.2 Our SPSS code for stepwise regression. We did nine experiments where each organism was a <i>Dependent</i> variable to be regressed over all the others of the pool (the <i>Predictors</i> variables).	137
8.3 A complete listing all codons for amino acids (AAs) that were extracted from the DNA of the avoided palindromes (APs). The columns contain the counts of codons correlating to each extracted amino acid. The gray cells indicate that a triplet from the AP code was also missing a corresponding tRNA (listed in Table 8.4) according to our analysis using BLAST. These cells are evidence for the pathways of reduction of our study.	141
8.4 The tRNAs which were absent across all organisms of our study according to a BLAST analysis. The above tRNAs, by extension from reduced codon content according to avoided palindromic DNA, are the end-points of the pathway of reduction.	142

11.2 The table to show the number of protein records by organism available for our work. The second and third columns display the number of Mt and non-Mt UniProt protein records, respectively. The forth column describes the exhaustive number of protein records where PTM are discussed in some of the articles. The fifth column provides an estimation of the number of scientific articles from the literature that may have been sources of PTM information for protein records. This data was furnished by text mining the NCBI body of literature.	211
11.3 The Pearson and Kendall-tau correlation values of implied organism complexity, and PTM and MS magnitudes. In this table, we assigned complexity values based on inspection of the general number of connections between PTMs and MSs in the networks. We computed a parametric (Pearson) and a non-parametric (Kendall-tau) correlation coefficient between the set of organism ranks and the set of PTM counts for each protein type (i.e., Mt and non-Mt). These two different correlation tests were employed to provide a wider view of correlation between the limited number of data points of each set. Three out of the four Pearson and Kendall-tau rank correlation tests were found to be significant. However, in both tests, the Mt and non-Mt PTMs were consistently correlated with the organism rankings. The Mt and non-Mt MSs were correlated, but not consistently across both tests.	229
11.4 A ranking of proteomes in terms of number of unique PTMs observed (Mt and non-Mt). The gray fields indicate that the ranking is not the same for both the Mt and non-Mt sets. The majority of proteomes have the same ranking in both sets.	230

11.5 The number of PTMs and MSs, associated with each organism. Organisms without PTM or MS data are absent from this list. These results are displayed as a scatter plot in Figure 11.18.	231
11.6 The <i>p</i> -values of from our Poisson approximation by the Chen-Stein Method over (PTM, MS) pairs in Mt networks. Significant <i>p</i> -values (i.e., values less than $\frac{\alpha}{66}$) are denoted by stars (*) to suggest that these pairs of organisms differ in complexity according to their networks.	236
11.7 The <i>p</i> -values of from our Poisson approximation by the Chen-Stein Method over (PTM, MS) pairs in non-Mt networks. Significant <i>p</i> -values (i.e., values less than $\frac{\alpha}{66}$) are denoted by stars (*) to suggest that these pairs of organisms differ in complexity according to their networks.	236
13.1 Diverse organisms of the study. The number of MS encountered in Mt and non-Mt are shown in the organism's row.	257

List of Figures

1.1	A protein is still immature until it has been customized for its specific function after translation.	3
1.2	A stressed protein undergoes a conformational adaptation by one or more PTMs to allow it to continue its function unabatedly while still in stress. Once the stress has elapsed, PTMs are again able to restore the protein to its original conformation.	4
1.3	During protein synthesis, codons of mRNA's attract tRNAs which place a unique amino acid in a string. PTMs target specific AAs for modification.	5
1.4	The short term solution implies that protein is temporally re-purposed to cope with stress where as, the long term solution implies that the stress is constant and protein evolution may be involved.	7
1.5	Two mitochondrial organelles located within a mammal lung tissue. Here we note their matrix and membranes as shown by electron microscopy. Image: https://en.wikipedia.org/wiki/Mitochondrion (21 March 2016).	10

1.6 Neurons require more ATP to drive their signals and associated processes and mechanisms. We note that all of these mechanisms require energy to function and an absence of ATP is likely to cause malfunctions and the onset of ailments. Graphic taken from Knott <i>et al.</i> ^[151]	14
1.7 In addition to all the communication between the Mt and the nucleus, the Mt also provides the energy to support these mechanisms. Here we note that even the basic mechanism of synthesizing protein requires much communication with the Mt to gain the energy for the process.	15
2.1 One may see signals all over and if one is not familiar with their mechanisms, then their meanings may not be obvious.	19
2.2 The waveforms from PPG and ECG containing noise. The ECG waveform is mostly noisy over a 40-second interval and the PPG waveform is partially noisy. The poor-quality waveform regions have been detected by SVMs and are outlined by thick horizontal bars. This image was taken from Yu <i>et al.</i> ^[321] .	21
6.1 Sequence fragments are separated into groups (called, “bins”) of relatedness by a quick pre-processing step. This graphic taken from our previous work in ^[28] .	85
6.2 The spectrum set taken from the four restriction sites of the <i>Clostridium</i> genera. There are ten unique recognition sites covering all four spectrum sets (shown in Figure 6.4). This graphic taken from our previous work in ^[28] .	88
6.3 The spectrum set taken from the four restriction sites of the <i>Staphylococcus</i> genera. The motif <i>ATGCAT</i> is common to <i>Clostridium</i> . This graphic taken from our previous work in ^[28] .	88

6.4 From its base composition, each bacterial restriction site fits into only one of the four spectrum sets, featured by unique color patterns. The motifs of each set are made up by the permutations of one of the following words; <i>AAATT</i> , <i>AATTG</i> , <i>CCGGAT</i> or <i>CCCGGG</i> . This result taken from our previous work in [28].	89
6.5 The flowchart that we applied to the clustering operation using heatmaps.	92
6.6 Separation by the <i>AAATT</i> -Spectrum set. There is a clear distinction between each bin; <i>Closteridium</i> and <i>Staphylococcus</i> of the Firmicute division. The data is segregated except for the two middle sequences forming a separate group. We had similar results from the <i>AATTG</i> -Spectrum set. This result from our previous work in [28].	95
6.7 Separation by the motifs of the <i>CCCGGG</i> -Spectrum set. Note a clear distinction between each bin. In addition, we note that there is no longer a color pattern showing that <i>Clostridium botulinum</i> are closely related, as we saw in Figure 6.6. This result taken from our previous work in [28].	95
6.8 The <i>AAATT</i> -Spectrum set test. The sequence data is applied to our base composition analysis to determine its relatedness.	96
6.9 The <i>CCCGGG</i> -Spectrum set test. The sequence data is analyzed by base composition to determine relatedness.	96
6.10 The <i>AAATT</i> -spectrum set analysis taken across all sequence data in a pool. The <i>Burkholderia pseudomallei</i> sequence data, having elevated proportions of the motifs of this spectrum set, create a contrast from those of <i>Clostridium tetani</i> and <i>Staphylococcus aureus</i> and have mixed proportions.	97

6.11 The <i>CCCGGG</i> -spectrum set analysis taken across all the contigs in the pool. We note that the <i>Burkholderia pseudomallei</i> sequence data, having low proportions of the motifs of this spectrum set, create a contrast from those of <i>Clostridium tetani</i> and <i>Staphylococcus aureus</i> . These organisms are observed to have mixed proportions by this heatmap.	98
6.12 The <i>AATTCG</i> -Spectrum set test: The genomes or chromosomes are analyzed by base composition to determine the expected clustering behavior of their contigs. 99
6.13 Separation of contigs of <i>Clostridium tetani</i> and <i>Staphylococcus aureus</i> by the <i>AATTCG</i> -spectrum set. We found that this spectrum set worked well to separate the contigs. The <i>AAATT</i> -spectrum set did not perform as well as we had expected from our work in Figure 6.6. We suggest that the contigs of these particular organisms followed trends shown in Figure 6.12. 99
6.14 NCBI's Taxonomy Tree used for validation and comparison. This phylogenetic tree was used to compare the results of the spectrum set analysis of the organisms listed in Table 6.3. We ranked the results on a scale of highest to lowest resemblance in Table 6.4. 100
6.15 The <i>CCGGAT</i> -spectrum set. This tree perfectly resembles the taxonomy tree of Figure 6.14 and shows the great evolutionary distances between the organisms. The rat and mouse are found to be closely related. We note tree distinct subtrees: one containing the bacterium, one for the mammals and one containing the worm and fruit fly. The location of these subtrees conforms to taxonomy tree.	. . 101

6.16 The <i>AAATT</i> -spectrum set. This tree also resembles the taxonomy tree but there is a slight distance between mouse and rat which is not found in Figure 6.15. We note tree distinct subtrees conforming to the taxonomy tree.	101
6.17 The <i>AATTCG</i> -spectrum set. We note that mouse and rat are not closely related. The bacterium is also evolutionarily located between the mammals and the subtree containing the worm and fruit fly. . . .	103
6.18 The <i>CCCGGG</i> -spectrum set. This tree is inaccurate because it indicates that the rabbit and the fruit fly are closely related.	103
6.19 Length-6 palindromic spectrum set. Here we note that this tree does not conform well to the validation tree in Figure 6.14. Rat and mouse are shown to be closely related but inaccurately, the tree shows that the bacterium and the worm are also closely related.	104
7.1 The flow chart of the encryption and decryption phases.	114
7.2 The process to convert PT into PtDNA. The plain text is encoded to binary words of length eight. This binary sequence is read a pair at a time and each is encoded by the correlating DNA bases.	116
7.3 The application of the Latin square to the PtDNA and the KeyDNA. This step serves to encrypt the information by recombination in function of two input-sequences.	118
7.4 Encryption using the Latin square. Here the KeyDNA character is ‘A’ from the left column and the plain text character is ‘C’ of the top row. At the intersection of these two characters is the cipher text character ‘t’.	119

7.5 The transformation of DNA code to Huffman binary codes. Here we coded these triplets and their amino acids according to the codon usage frequencies of <i>Bacillus phage PBS2</i> . We note that the final product of phase two contains two levels of code: the amino acid from the translation and its exact RNA triplet. Since there are redundant triplets encoding the same amino acid, coding the triplets all together make lengthy codes. To maintain shorter codes, we made sets of triplet codes, corresponding to each unique protein amino acid.	120
7.6 Decryption using the Latin square. Here the KeyDNA character is ‘A’ from the left column and the cipher text character is ‘t’. At the top of the column is the plain text character is ‘C.	122
7.7 Raising entropy in CT from PT forms. We note an incline in entropy from sEncrypt. The Gene code was the only information type that already had high entropy when PT form. There was an entropy increase of all other forms of information we tested.	124
7.8 Autocorrelation of the poetry CT sequence was tested by <i>dottup</i> from Emboss. We note some tiny regions where the cipher text has repetition but these are too small to be significant.	125
7.9 There is a significant reduction in the file size containing the CT when saved in binary.	126
8.1 The taxonomy tree of the organisms of this study. This information was obtained from NCBI taxonomy.	135
8.2 The size of each set of palindromes as arranged by word length. We note that there also are some odd-length relevant palindromes. This data taken from REBASE.	135

8.3 Palindrome Avoidance Data, Length-4. Our notation, * and ** specifies a significance at the $\alpha = 0.01$ and $\alpha = 0.05$ levels, respectively. The listed value is the unstandardized b coefficient. The variable names are given in Table 8.1	138
8.4 Palindrome Avoidance Data, Length-5 (upper) and Length-6 (lower). Our notation, * and ** specifies a significance at the $\alpha = 0.01$ and $\alpha = 0.05$ levels, respectively. The listed value is the unstandardized b coefficient. The variable names are given in Table 8.1	139
9.1 A comparison between the number of PTMs between our Mt and non-Mt sequence data.	150
9.2 The taxonomy tree of our study's data.	154
9.3 A heatmap of PTM frequencies in Mt protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.	159
9.4 A heatmap of PTM frequencies in non Mt protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.	159
9.5 Mt amino acid frequency heatmaps prepared using Equation 11.3. Only the active site frequencies greater than 0.1 are displayed.	160
9.6 Non-Mt amino acid site frequency heatmaps prepared using Equation 11.3. Only the active site frequencies greater than 0.1 are displayed.	161
9.7 The <i>Arabidopsis thaliana</i> Mt Network: PTM frequencies (left) and associated active sites by tRNA frequencies (right). These models tend to be <i>less dense</i> than those of non-Mt proteins.	164
9.8 The <i>Arabidopsis thaliana</i> non-Mt Network: PTM frequencies (left) and associated active sites by tRNA frequencies (right). These models tend to be <i>more dense</i> than those of non-Mt proteins.	165
9.9 The <i>Caenorhabditis elegans</i> Mt Network: PTM frequencies (left) and associated active sites by tRNA frequencies (right).	165

9.10 The <i>Caenorhabditis elegans</i> non-Mt Network: PTM frequencies (left) and associated active sites by tRNA frequencies (right).	166
10.1 The four protein classes from <i>Zea mays</i> (maize) sequence data describe a typical heatmap from our data. Each bar is made up of colored cells which represent the amount of motif coverage in a particular protein class. The darker the shade of cell (blue) then the closer to zero is the coverage of motif which is thought to attract oxidative activity. We noted that the darker regions were usually more pronounced in the mitochondrial proteins than the non-mitochondrial proteins and were often more pronounced in the enzymatic data sets. The other heatmaps from this work are included in the supplemental data.	183
10.2 The four protein classes from <i>Sophophora melanogaster</i> (fruitfly) sequence data in a heatmap. The heatmaps from the other organisms discussed in this work are included in the supplemental data.	184
10.3 The four protein classes from <i>Homo sapiens</i> (human) sequence data in a heatmap.	184
10.4 Rankings of R, K, T, P, E and S residues across the protein classes of all organisms. We note that P and T, common to both the PEST and RKTP motif sets, have general upward trends but several residues do not.	188
10.5 The flowchart for the method (Part 1). Here we describe our method using an sample protein sequence which is processed by EMBOSS' protein prediction tool to ascertain the protein's functional structure (C = Coils, E = Sheets, H = Helices, T = Turns). The individual locations where oxidation sites were detected are extracted from the structural sequence to uncover the structure-feature words at these locations. These words are then used to build a new sequence containing all oxidation site structural information.	191

10.12 Test Number Six: The structural words made up of coils (C), sheets (E), helices (H) and turns (T) are along the x-axis. The y-axis represents the mean coverage of the structural words of all proteins in the set. The <i>Mt</i> and <i>NonMt</i> refer to mitochondrial and non-mitochondrial content, respectively.	200
10.13 Test Number Seven: The structural words made up of coils (C), sheets (E), helices (H) and turns (T) are along the x-axis. The y-axis represents the mean coverage of the structural words of all proteins in the set. The <i>RKPT</i> and <i>PEST</i> refer to these motif contents, respectively.	200
10.14 Test Number Eight: The structural words made up of coils (C), sheets (E), helices (H) and turns (T) are along the x-axis. The y-axis represents the mean coverage of the structural words of all proteins in the set. The <i>RKPT</i> and <i>PEST</i> legends refer to these motif contents, respectively.	201
11.1 A comparison between the number of PTMs in our Mt and non-Mt sequence data. Here we exclude all PTMs that are labeled by UniProt as <i>InterChain</i> due to a lack of information available for our study.	208
11.2 All Mt and non-Mt proteins were examined in each organism of our study. We recorded the protein type (Mt or non-Mt), the PTMs of the protein and their associated MSs. This information was used to assemble relative frequency data.	212
11.3 An example of how relative frequency information was extracted from protein data. For each organism, all Mt and non-Mt protein records were queried to ascertain their observed PTMs that have been curated by UniProt. The type and count of each PTM, including its associated MS was recorded to calculate frequencies by Equations 11.1 and 11.2. Not shown, the occurrence magnitudes of all amino acids (non-MSs) were also obtained and applied to Equation 11.3 to determine the general amino acid compositions of each proteome.	214

11.4 Mt and non-Mt PTM compositions prepared using Equation 11.3. High magnitudes of frequency are described by warmer colors. We note that phosphorylation and acetylation were common PTMs across the organisms. We note that all frequency values greater than 0.18 (threshold) are included here.	217
11.5 Mt and non-Mt MS compositions prepared using Equation 11.2. High magnitudes of frequency are described by warmer colors. Unlike the non-Mt heatmap where nearly all amino acids played a roles as MSs, there were many AAs in the Mt proteomes that were never involved with the PTMs.	219
11.6 Mt and non-Mt amino acid compositions prepared using Equation 11.3. High magnitudes of frequency are described by warmer colors. Although all organisms display a common theme of color bands indicating that their amino acid composition is similar, we note that related organisms have especially similar patterns of color indicating that the amino acid distributions are similar.	220
11.7 <i>Arabidopsis thaliana</i> : A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.	222
11.8 <i>Aspergillus nidulans</i> : A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.	223
11.9 <i>Caenorhabditis elegans</i> : A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.	223

11.10 <i>Canis familiaris</i> : A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.	224
11.11 <i>Danio rerio</i> : A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.	224
11.12 <i>Homo sapiens</i> : A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.	225
11.13 <i>Mus musculus</i> : A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.	225
11.14 <i>Oryctolagus cuniculus</i> : A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.	226
11.15 <i>Rattus norvegicus</i> : A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.	226
11.16 <i>Saccharomyces cerevisiae</i> : A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.	227
11.17 <i>Xenopus laevis</i> : A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.	227
11.18 The number of PTMs and MSs, associated with each organism. This data is also shown in Table 11.5.	232

11.19 The number of isoforms of the organisms in our study. These counts were prepared by querying all organismal proteins in UniProt and then determining how many isoform proteins were present. The increasing number of isoforms may help to explain the increasing number of PTMs in higher organisms. Note that <i>Aspergillus nidulans</i> has been omitted due to the lack of isoform information.	238
12.1 <i>ALN</i> : The stressed Mt proteins associated by Alzheimer's disease. The three types of nodes featured the networks are: green pentagons (<i>PTMs</i>), red circles (<i>protein types</i>), and blue squares (<i>stresses</i>).	251
13.1 The description of where the distances of PTMs are found: <i>before</i> , <i>inside</i> and <i>after</i> a protein domain. The amino acid residues for each region were collected and analyzed for composition. The MSs are shown inside the diamonds and the arrows-above describe the collected distances.	261
13.2 We note that a protein may exist having multiple domains as influenced by the same MS. In this case, each domain is processed separately by PTM-Tracker with the same MS.	264
13.3 We note that a protein may exist having multiple MSs that influence the same domain. In this case, the domain is processed for each MS. .	264

13.4 How to read plots containing organismal MS information occurring <i>before</i> domains. The domain start is imagined to be at the far right and the green bars describe the MS neighbourhoods which are located upstream of the domain. The magnitude is understood to be the number of other MSs for <i>acetylation</i> (or any single PTM of interest) found at the same locations. The plots for showing MS neighbourhoods <i>inside</i> and <i>after</i> the domains of a particular organism are similar. The plot in this example describes that MSs are found in three general locations (i.e., neighbourhoods) <i>before</i> all domains of a particular organism.	265
13.5 The non-Mt protein plot of all locations of <i>acetylation</i> MSs found <i>before</i> the domains in <i>H. sapiens</i> (human). The <i>x</i> -axis represents the location of the MS neighbourhoods (green). The <i>y</i> -axis describes the number of times that this same location was observed for the element across the samples. This is a typical plot for the mammal protein data. . . .	266
13.6 The non-Mt protein plot of all locations of <i>acetylation</i> MSs found <i>before</i> the domains in <i>C. familiaris</i> (dog). This plot resembles that of Figure 13.5 which also has the same axes.	267
13.7 The non-Mt protein plot of all locations of <i>acetylation</i> MSs found <i>before</i> the domains in <i>M. musculus</i> (mouse).	268
13.8 The non-Mt protein plot of all locations of <i>acetylation</i> MSs found <i>before</i> the domains in <i>O. cuniculus</i> (rabbit).	269
13.9 The Mt protein plot of all locations of <i>acetylation</i> MSs found <i>inside</i> the domains in <i>H. sapiens</i> (human).	270
13.10 The Mt protein plot of all locations of <i>acetylation</i> MSs found <i>inside</i> the domains in <i>M. musculus</i> (mouse).	271

13.11The non-Mt protein plot of all locations of <i>acetylation</i> MSs found <i>before</i> the domains in <i>O. cuniculus</i> (rabbit).	272
13.12The non-Mt protein plot of all locations of <i>acetylation</i> MSs found <i>inside</i> the domains in <i>H. sapiens</i> (human).	273
13.13The non-Mt protein plot of all locations of <i>acetylation</i> MSs found <i>inside</i> the domains in <i>C. familiaris</i> (dog).	274
13.14The non-Mt protein plot of all locations of <i>acetylation</i> MSs found <i>inside</i> the domains in <i>M. musculus</i> (mouse).	275
13.15The non-Mt protein plot of all locations of <i>acetylation</i> MSs found <i>inside</i> the domains in <i>O. cuniculus</i> (rabbit).	276
13.16The Mt protein plot of all locations of <i>acetylation</i> MSs found <i>inside</i> the domains in <i>M. musculus</i> (mouse).	277
13.17The Mt protein plot of all locations of <i>acetylation</i> MSs found <i>before</i> the domains of <i>O. cuniculus</i> (rabbit).	278
13.18How to read plots containing domain and MS information. Red and blue bars represent the proportional distances ventured across the protein to reach the beginning and end of a domain. Their magnitudes indicate the number of other domains having the same proportional values. The green bar describes the proportional locations of the MS neighbourhoods for <i>acetylation</i> (or any single PTM of interest). The magnitude describes the number of encountered MSs at these same locations. The plot in this example describes that the MS neighbourhood is found <i>inside</i> domains.	279

13.19 The proportional distribution of <i>acetylation</i> MSs, encountered in all proteins in UniProt for the 11 organisms of our study, before the <i>atp-grasp2</i> domain. We note how this plot is simple to that of Figure 13.20.	280
13.20 The non-Mt plot of <i>acetylation</i> MS encountered before the <i>atp-grasp2</i> domain. We note how this plot is similar to that of Figure 13.20.	281
13.21 A heatmap of amino acid composition of Mt and Non-Mt protein in MS neighbourhoods <i>before</i> the domains. We note that the amino acid compositions are very similar across the samples.	282
13.22 A heatmap of amino acid composition of Mt and Non-Mt protein in MS neighbourhoods <i>inside</i> the domains. We note that the amino acid compositions are very similar across the samples.	283
13.23 A heatmap of amino acid composition of Mt and Non-Mt protein in MS neighbourhoods <i>after</i> the domains. We note that the amino acid compositions are very similar across the samples.	284
14.1 Lenses for different depths of field and levels of granularity.	295

If beauty isn't genius, it usually
signals at least a high level of
animal cunning.

Peter York

Chapter 1

Introduction

1.1 Signals After Mechanisms

Biology is not made up of magical systems; there is always a mechanism of logic to explain a phenomenon. Naturally, this mechanism may remain in the dark until it is illuminated by an investigator whose goal is set on its discovery. To confirm its existence, it is by the meticulous collection of biological data surrounding this mechanism, in conjunction with its acute analysis, that the form of the mechanism emerges. Once the signals originating from their natural environment have been identified, the mechanism may be isolated for study and eventually presented to the community as new knowledge.

Signals may be detected by their effects on members of their environments. For example, cellular stresses (i.e., originating from heat, cold, salt and others) may damage proteins in short amounts of time requiring stressed proteins to quickly cope to continue their functions. Failure to tolerate environmental stresses has

unfortunate consequences. For example, muscular tissues experience atrophy when exposed to the stress of weightlessness (microgravity) during space travel after only a few weeks^[4;209]. To combat the effects of stress, it generally takes less time to recondition an existing protein that has already completed its transcription and associated biological processing, than to re-synthesize steady-state proteins^[115]. Although we may not directly observe this transformation, we can, in fact, note that the alteration has occurred by the nature of the signals which have been emitted as a result of this mechanism.

In this dissertation, we explore the signals of different types of biological data (i.e., sequences of DNA, RNA, protein and curated data from public databases) to locate biological mechanisms which we believe to exist due to their tell-tail signals emerging from the distillation from large amounts of the data. This work is primarily an investigation into the detection and isolation of signals that may be used to describe the existence of their biological mechanisms. Each contribution that we provide reinforces two main concepts: (1) that signals may be detected and, (2) that these signals are not randomly distributed in the data. This non-random distribution signifies that there are elements in the data that must have deeper meanings and may likely originate from conserved biological mechanisms. The data for this work is informational-based and has been curated by researchers in wet-labs and placed on public databases where it can be obtained for studies such as our own.

1.2 A Focus On PTMs, Proteins And Their Typical Stress Responses

Before a building material may be applied to some specific task, it must be crafted into the some shape or form which will be useful to its intended usage. In much the

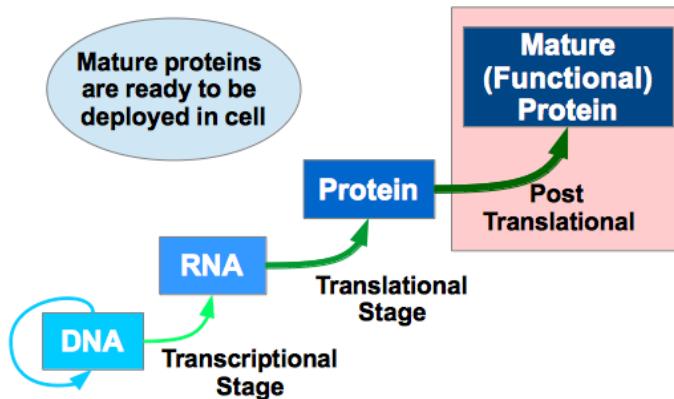


Figure 1.1: A protein is still immature until it has been customized for its specific function after translation.

same way, after the synthesis of protein material, the protein is still immature and unfit for a specific duty until it has been customized for its intended function. During the process of maturing the protein for a special duty, it must be tailored and uniquely folded to give it special qualities to enable its behavior and function for its specialized duty. This protein customization is performed by post-translational modifications (PTMs) such as acetylation, phosphorylation, glycosylation and others, which are steps in biosynthesis that alter structure at the atomic level. Shown in Figure 1.1, post-translational steps are responsible for the creating a mature, finalized protein product that is fit for some specific task or duty.

As with any building material, the structure grants function. This is also true in the case of protein and any modification to structure is likely to bring functional adaptations. The study of PTMs is important for understanding the cellular responses to environmental stresses, as utilized by different organisms since they serve to alter protein structures. PTMs may be considered as, “first-responders”, to stress management in protein. During an exposure to stress, a specialized modification of protein structure may work to make structural changes to enhance its ability to withstand the stress. In effect, this action enables the protein to

continue its function in spite of the stress condition (described in Figure 1.2). In order to understand how these stress response systems work in the presence of different kinds of stresses, we must study their mechanisms which enable proteins to tolerate stress factors. To make this study, it is important to start by understanding general mechanisms where PTMs are involved. We will later combine this knowledge, with that gathered from their signals, to infer more of their modes of operation.

1.2.1 Protein Regulation And Modification

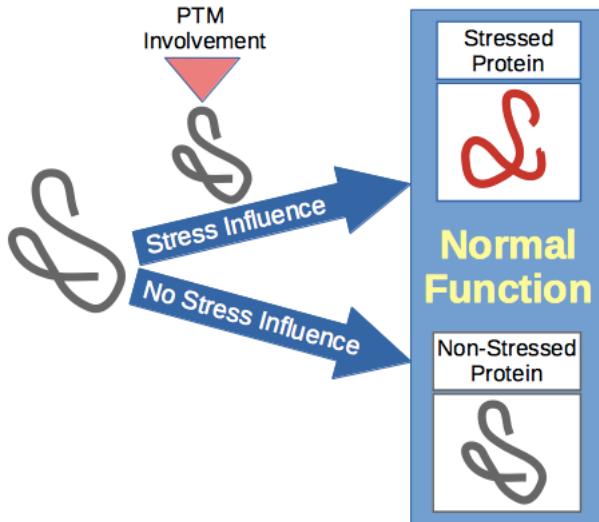


Figure 1.2: A stressed protein undergoes a conformational adaptation by one or more PTMs to allow it to continue its function unabatedly while still in stress. Once the stress has elapsed, PTMs are again able to restore the protein to its original conformation.

In the human proteome, more than 400 human PTMs have been observed^[148] and it is very likely that all proteins are regulated by PTMs at some point in their lives. The entire set of proteins expressed by the human genome at a particular moment contains many more types of proteins than the number of genes in the genome. This difference in the number of proteins is contradictory to the

one-gene-one-protein hypothesis, and is noted by the functional diversity in the proteome. The various steps to obtain functionally active proteins are shown in Figure 1.3 where we note that a PTM is applied to one of the amino acids in the chain. Upon this application, a diversification is made possible and a protein may likely have new functions since its structure has been altered. Depending on where protein chain is modified, any number of unique functions may be obtained after protein translation. These modifications may have consequences other than to create functional reassessments: PTMs may be involved in cell signaling pathways, increasing system stability, protein localization, and also regulation.

A prominent example of regulation by PTMs occurs to the human p53 protein which functions as a tumor suppressor and is active in other processes such as apoptosis, genome integrity maintenance, metabolism and several others. When subjected to cellular stresses that risk inflicting damage on DNA, p53 is functionally activated by the PTM, when phosphorylation at the C- and N-terminal series amino acids occur^[138;139]. The phosphorylation of p53 occurs at thirteen serine and five threonine amino acids which are distributed in the protein's (functional) domain regions^[107].

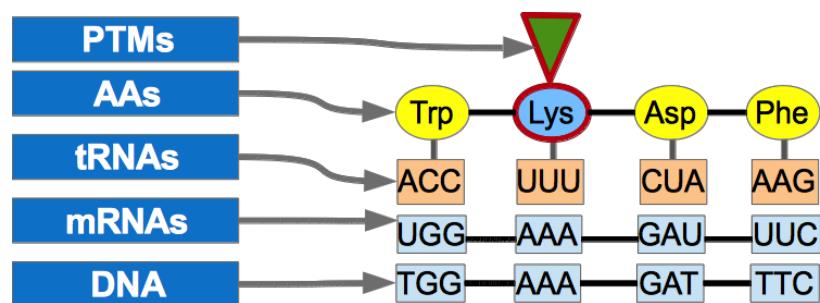


Figure 1.3: During protein synthesis, codons of mRNA's attract tRNAs which place a unique amino acid in a string. PTMs target specific AAs for modification.

As we have mentioned above, one of the main interests in PTMs is that they are active in protein stress responses. PTMs such as phosphorylation specific amino

acid residues in the protein chain called modification sites (MSs). For example, the majority of MSs of particular proteins are rapidly phosphorylated following cellular exposures creating stresses (i.e., including DNA damage, abnormal oncogenic events, telomere erosion and hypoxia^[124]), although a few MSs such as, serine (S) or threonine (T) phosphorylation, are phosphorylated constitutively in unstressed cells and are then dephosphorylated following stress^[26;180]. It is interesting to note that these phosphorylation sites exhibit redundancy since a particular site can be phosphorylated by multiple kinases (an enzyme catalyzing phosphate transfer) and a single kinase can phosphorylate multiple sites^[161]. Additionally, the p53 protein also contains nine acetylation sites, which activate transcriptional activation mechanisms and enhances the protein's stability^[161]. The acetylation levels are enhanced in the C-terminal of the protein, when the protein is stressed. Since this protein is constantly responding to a wide variety of stresses, it has been suggested that diverse stresses initiate alternative pathways for stress response^[161].

Glycosylation is a significant type of PTM that alters protein function by changing its folded structure. Different types of glycosylation have been observed (i.e., O and C-glycosylation)^[95;134;291], but much about its mechanisms are still unclear^[105]. This PTM involves the addition of glycosyl groups (carbohydrate sugars) to the protein at several MSs: asparagines (N), hydroxylysines (Hyl), serines (S), or threonines (T). Phosphorylation implies the addition of a phosphate group to serine (S), tyrosines (Y), threonines (T) or histidines (H). PTMs are also involved in the localization of the protein to different compartments of the cells or body of the organism to perform its function. For example, S-Nitrosylation, a reversible PTM that occurs at a cysteine (C) modification site, localizes proteins by regulating nitric oxide reactions^[98].

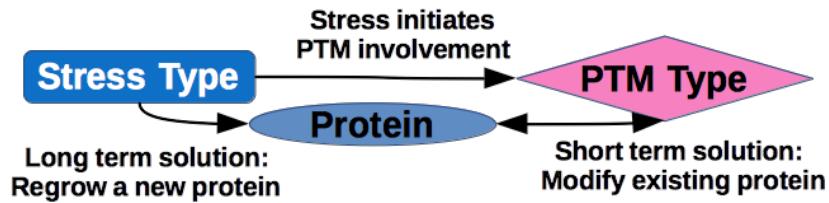


Figure 1.4: The short term solution implies that protein is temporally re-purposed to cope with stress whereas, the long term solution implies that the stress is constant and protein evolution may be involved.

1.2.2 Sudden Stresses

When a protein is stressed, PTMs are involved to create a short-term solution^[332]. This solution implies that the response to stress is rapid and economical - an existing protein is re-purposed by modifiers to cope with a sudden stress. Noted in^[3], short-term solutions in plant proteins subjected to drought are protein alterations for osmotic adjustment, defense signaling and programmed cell death. A long-term solution may be described by replacing a failing or dysfunctional protein by a freshly synthesized one for a specific task. We note that such a long-term solution implies much time and resources to implement and would not be appropriate since common stresses are ephemeral. Instead, such a solution may, instead, impress evolutionary forces on the stressed protein to change its configuration, if the stresses were of a longer duration. We note this basic arrangement in Figure 1.4. Protein folding (the creation of a mature protein) is regulated by changing points of ionic attraction in the protein to ensure that necessary folds are either made or broken. This action serves to diversify structure, and hence to modify the function of the protein while adding stability^[106].

PTMs are generally reversible modifications of amino acids which can be made when necessary to the protein. Stresses such as heat shock, salinity, microgravity and others, impress hardships on mature proteins which may prevent the protein

from normal function^[229]. Discussed in Phillips *et al.*, PTMs offer a convenient and rapid solution to changing environmental conditions since they are able to quickly convert the cellular protein into complexes that are able to survive and function under diverse conditions^[229]. When one considers resources necessary to recreate a new protein from DNA, the reuse of existing protein material is likely a better mode of adaptation, although the complete protein re-synthesis may play a part in response to a chronic stress^[13]. PTMs offer this quick and often reversible protein transformation^[3]. Once the stress has elapsed, PTMs are often able to convert the protein back into its original state. For example, in mice, rabbits, and pigs, Phillips *et al.*^[228] noted that during an exposure to metabolic stress, the heart (where the stresses are greater and more sudden) was more reliant on the PTM phosphorylation to withstand metabolic stress than the liver (where the stresses are minimal). In a related study, the relative right and left values of the heart were observed to have alternative protein expression levels for healthy heart function.

1.2.3 PTM Localization And Study

Advances in mass spectrometry technology have allowed for the increased ability to routinely identify new PTMs and observe their involvements with protein^[62]. However, the functional characterization of these modifications is confused by studies which note divergences across species where modification sites are naturally shifted (as in the case of phosphorylation, for example) or their regulation appears to affect organismal proteins in non-uniform ways^[18;123;165].

Sirt1 and Sir2L1 are two closely related proteins that are found in human and mouse (and other organisms) which perform similar types of regulation tasks. By studying their protein data available from UniProt^[9;194], we have observed that a PTM bias exists between Sirt1 and Sir2L1. For example, the PMT usage frequencies of N-acetylalanine, phosphoserine, S-nitrosocysteine and phosphothreonine in both, human

(according to UniProt: Q96EB6 has 19 PTMs total which are associated to this Sirt1 protein) and mouse (UniProt: Q923E4, 13 PTMs sum total), the most frequently used PTM is phosphoserine. The second most used PTM is phosphothreonine (human) and S-nitrosocysteine (mouse) which implies a deviation of general usage between both samples.

Serving as possible regulators of PTM interaction with proteins, PTM modification sites have been shown to be largely located in 60 percent of functional regions in proteins^[191]. Furthermore, in the same study it was found that in trans-membrane proteins, the *N-linked (GlcNAc...)* glycosylation sites were located in the extracellular regions, as well as the O-linked and C-linked glycosylation sites. Phosphorylation sites were mainly located in cytoplasmic regions, which induce signal transduction and ion transport. This suggests that the placement of modification sites along proteins may help to control PTM initiations for cellular function in certain protein regions^[58;172].

We note that an overwhelming majority of functional roles of *Homo sapien* genes are largely unknown in the 35K to 50K of the discovered genes^[52], it is likely that about half can be assigned functional roles based on homology to proteins having a known function^[135]. This lack of knowledge may be attributed to a lack of sufficient data due to underdeveloped computational tools for the annotation of protein function^[238].

1.2.4 Protein Domains

At the heart of protein function are protein domains: the conserved parts of protein functional structures which can evolve and exist independently of the rest of the protein chain. The functions of domain structures were thought to be context dependent and directed by PTM activity (i.e., phosphorylation)^[191]. However more recently, protein regulation has been made possible by vastly interconnected and

functionally associated PTM operations^[200], and other factors such as, alternative splicing and intrinsically disordered protein fragments^[83;210]. This evidence points to a contextual aspect of regulation where the locational information of PTMs and protein domains is absolutely necessary for further study of regulation mechanisms. In^[19], regions which were rich in PTM activity were mapped across organisms and found to be conserved which suggested a common and conserved regulatory mechanism.

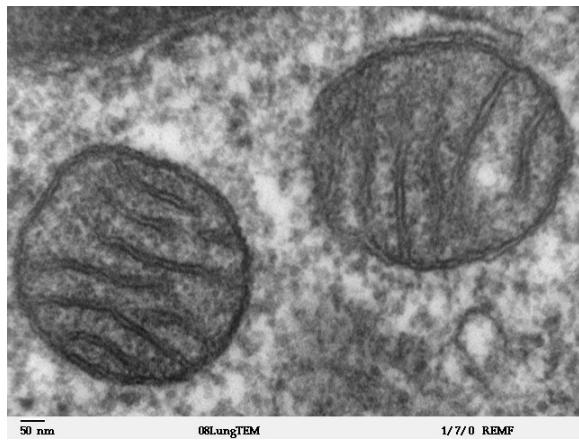


Figure 1.5: Two mitochondrial organelles located within a mammal lung tissue. Here we note their matrix and membranes as shown by electron microscopy.

Image: <https://en.wikipedia.org/wiki/Mitochondrion> (21 March 2016).

1.3 Mitochondrial And Non-Mitochondrial Protein

We differentiate mitochondria (Mt), shown in Figure 1.5, and non-Mt proteins for much of the work presented in this thesis. Mt and non-Mt protein may be differentiated by the fact that only the Mt proteins build energy and perform many diverse tasks in cellular regulation. We note that the two sets of protein already have different types of lives. For instance, the production of energy (adenosine

triphosphate, ATP) in Mt creates an environment where the types and frequencies of encountered stresses are very different from those of non-Mt proteins.

In our beginning contributions, we discuss trends occurring in each set (i.e., Mt and non-Mt) in terms of DNA and RNA data. Towards in the latter contributions, we concentrate on their interactions in protein and stress. Each type of protein may be exposed to entirely different types of environmental stresses which must be handled by internal stress-response systems and specialized PTMs. Discussed in Section 1.3 in more detail, ailments are not uniform across both these sets of proteins and protein failure may arise from one of the other type of protein. As we will see in our work in Chapter 10 (and others) the very nature of the protein, in addition to its associated types of PTM, may make the difference between the life or death of a protein, tissue or organism when exposed to specific types of stress.

Disorders such as Leigh's syndrome (i.e., a neurometabolic disorder that affects the central nervous system and is characterized by seizures and muscular debilitation), aging-related problems, Parkinson's disease^[117] and heart disease^[142] may have been initiated of disorders of Mt protein as a result of stress exposures^[50;57]. Impaired Mt function is likely to increase oxidative stress and might render cells more vulnerable to pathogenesis of the Parkinson's disease, Alzheimer's disease^[121] and other related processes, including excitotoxicity. These ailments, and others, are thought to be caused by stresses which excite and drive oxidative carbonylation (i.e., a kind of adverse protein PTM) which signifies that coping with oxidative stresses (and other types of stresses) may create the stage upon which, the PTM actors play to perform protein health maintenance.

1.3.1 Mitochondria: Energy Production And Stress Response

Chemical energy Adenosine triphosphate (ATP) is created by mitochondrial processes and serve as the chemical energy currency to support all cellular processes. Mt supply the majority of ATP in eukaryotic cells by respiration and oxidative phosphorylation. Mt are also key players in many other types of cellular processes as well, such as intercellular Ca^{2+} homeostasis, biosynthesis of pyridine nucleotides and amino acids and β -oxidation of fatty acids. The importance of Mt is so great to the maintenance of general eukaryotic cellular function that there are only a few isolated cases of “amitochondrial” organisms, in which the Mt organelles are absent. Discussed in Regoës *et al.*^[241], organisms such as *Giardia intestinalis*, having no Mt organelles, make up for their absent, yet essential, functions by containing double membrane-bounded structures involved in iron-sulfur cluster biosynthesis.

During energy production by cellular respiration, oxidative stresses take the form of reactive oxygen species (ROS) and general oxidation. Since these stresses likely originate from the creation of energy deep within the Mt organelle, its proteins cannot avoid the influences and hardships that the stresses may bring. However, in spite of this stress, there are fewer protein failures and interruptions due to stress-related disorders than we would expect. It may even appear that Mt proteins are more robust to resisting oxidative stresses. Over evolutionary time, the Mt proteins appear to have resolved this dilemma by evolving their amino acid compositions to allow for fewer regions where oxidative stresses are able to be destructive. This phenomenon is further discussed in Section 10. It is significant to note that if Mt proteins were to fail due to these stresses, then there would be far-reaching complications, possibly resulting in cellular death and so we begin to appreciate the importance of the abilities of proteins to handle the influences of stress.

1.3.1.1 An Energy Crisis

For a protein interaction to occur, there must be an energy imbalance that will “fund” the processes that make life possible. A default on the production of ATP has been shown to create a failure of protein function and cause disease. For instance, in Cha *et al.*^[51], the retardation of ATP production has been found to contribute to the on-set of Alzheimer’s disease (AD). Mitochondrial ATP synthase is a multiprotein complex that synthesizes ATP from ADP and other elements and was found to be decreased in the brains of AD patients and in transgenic mouse models, as well as in A β -treated cells. Mitochondrial dysfunction is an early and causal event in neurodegeneration. Epilepsy and various unprovoked seizures may commonly result from Mt dysfunction and its failure to deliver energy. Shown in Figure 1.6 neurons require much ATP energy from Mt and are likely to malfunction in its absence. A disorder of energy production would be crippling since these cells also work closely with other cells to transfer signals.

The lack of energy production from faulty Mt proteins has also contributed to disorders. Mitochondrial encephalomyopathies are common disorders that are the result of mutations of genes that affect Mt encoding proteins that act in many leading Mt functions^[175]. These mutations have been observed to reduce energy production in proteins where Mt are abundant (i.e., muscle and brain tissue), which initiated diseases such as HSD10, affecting muscle control, hearing and epilepsy^[55]. Furthermore, in these tissues (i.e., heart, brain and skeletal muscle) the disorders of proteins are typically characterized by weakness and types of developmental retardation. At the clinical level, the manifestation of these disorders have symptoms such as loss of ataxia, cardiomyopathy, deafness, decreased cognitive function, dementia, exercise intolerance, fluctuating encephalopathy, migraine pains, optic atrophy, proximal myopathy, seizures, spasticity (continuous muscle contraction), stroke-like episodes and others^[284;315].

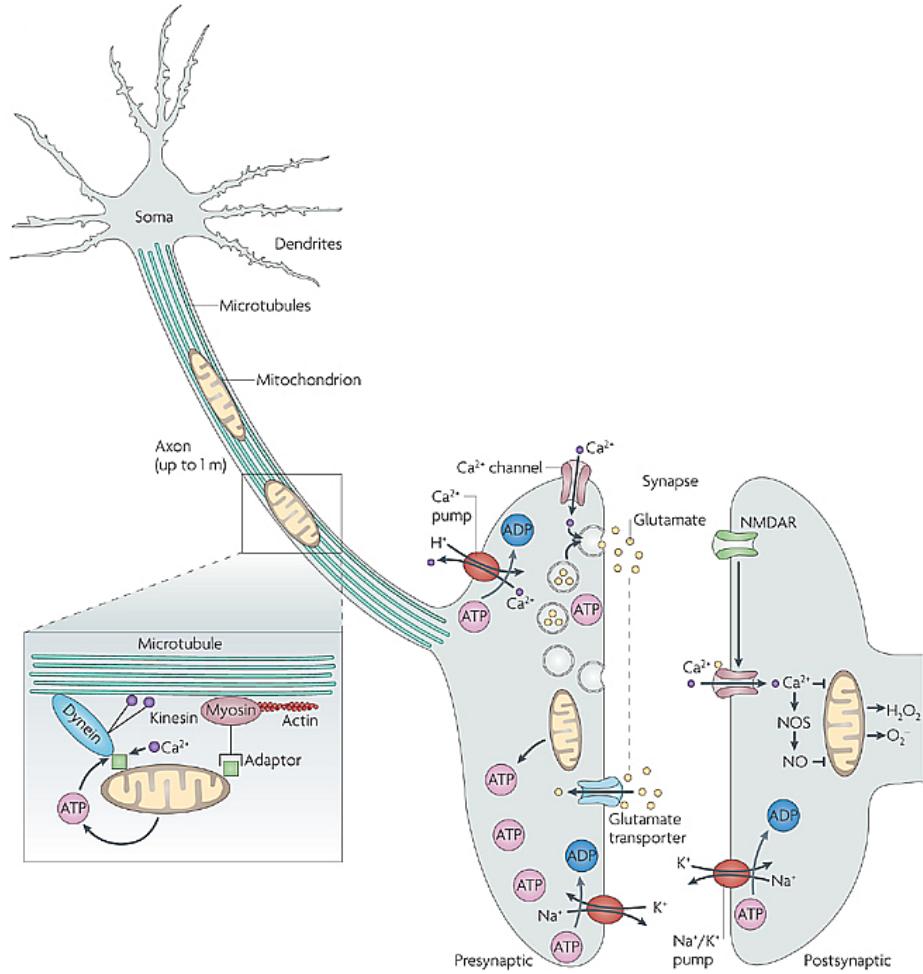


Figure 1.6: Neurons require more ATP to drive their signals and associated processes and mechanisms. We note that all of these mechanisms require energy to function and an absence of ATP is likely to cause malfunctions and the onset of ailments. Graphic taken from Knott *et al.* [151]

Mt also help to regulate many cellular processes which consume energy. Critical processes may cease during the depletion of ATP such as defects in biosynthesis and metabolism of neurotransmitters, disruption of Ca^{2+} homeostasis, generation of ROS (oxidative stress), and other complications^[94]. Oxidative stress serves to disrupt the cable-like morphology of functional Mt and cause disorders^[151]. We remark that the failure of Mt function may result in cell death as these above complications are by

themselves already very dangerous to cell health.

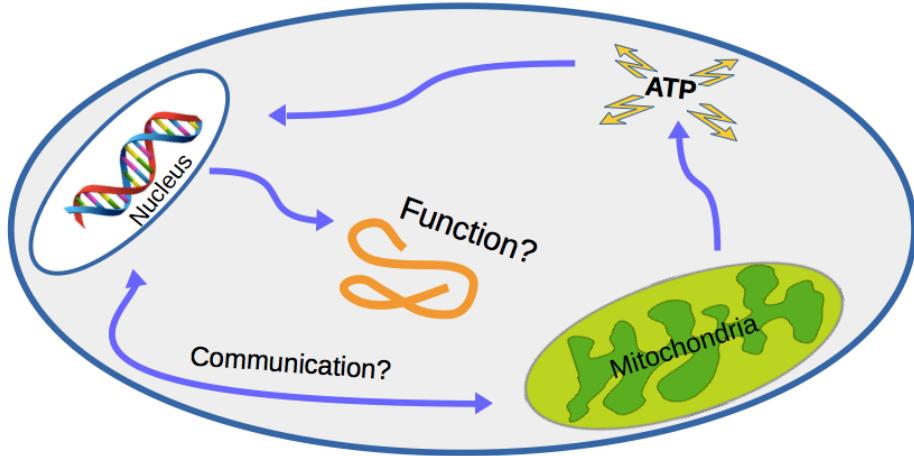


Figure 1.7: In addition to all the communication between the Mt and the nucleus, the Mt also provides the energy to support these mechanisms. Here we note that even the basic mechanism of synthesizing protein requires much communication with the Mt to gain the energy for the process.

In the above discussion, we described some of the disastrous consequences resulting from the lack of energy to drive biological processes. Signals, are emitted from mechanisms to coordinate the actions of proteins for biological processes. Since signals require energy to send, it is possible that when they are not sent due to energy deficiencies, for example, then disorders result. As we have already noted, environmental stress may work to prevent proper function of specialized proteins involved with energy production. In terms of stress, if a protein was prevented from sending a signal which caused harm, then the stress may have played a pivotal role in causing a resulting ailment.

As described in Figure 1.7, communication between Mt and the nucleus is evidenced by the fact that some of the Mt proteins are synthesized by nuclear processes (i.e., the transfer of most mitochondrial genes into the nucleus has occurred). Since protein synthesis cannot occur in absence of the Mt's energy supply, evidence for advanced communication may be found in the “ordering” and “delivery” of energy for protein synthesis and to drive all cellular processes.

Furthermore, since Mt are also directly responsible for many important cellular regulatory processes, we note that there is must be an incredible amount of Mt communication to detect and analyze in order to understand how these regulatory mechanisms function. In Yin *et al.*^[320], it was suggested that when Mt communications for cellular regulation are understood, then therapies could be efficiently administered to correct the problems left in the wake of poor or absent Mt signal communication. This therapy may also serve to correct other types of ailments which are indirectly related to poor Mt communication.

1.3.2 Mechanisms From Signal Detection

Most of the work of this thesis has been focused on detecting and analyzing some of the signals from biological processes so that their cellular mechanisms may be eventually understood. Although Mt energy production signals are important to this work, we are concentrating on the signals which originate from stress response systems and their functions involving PTMs, MSs, protein types and environmental stresses. Discussed and demonstrated in the chapters of this thesis, we offer the details of philosophies, techniques and automated technologies for the detection and study of signals from seemingly any type of biological processes where mechanisms are concerned.

In Chapter 4, we discuss the kinds of mathematical and statistical tools necessary to isolate signals and then we begin our investigation by isolating some of the signals from the DNA and RNA levels, which are discussed in Chapters 5 to 8. Here we determine some of the non-random trends in code and structure which could only be the result of mechanisms at play. We then turn our attention to the protein level where we apply our knowledge to detecting the signals of PTM mechanisms. This work is discussed in Chapters 10 to 13.

The work on the detection of PTM signals is extremely relevant since there is

still much to learn about the mechanisms used by PTMs. As we will see in Chapters 9 through 13, there are many hundreds of unique PTMs found in nature (i.e., acetylation, glycosylation, phosphorylation, methylation, and many others^[148]), that all appear to play unique roles in cellular regulation^[265], protein conformation change^[222], response to environmental stresses and other functions^[157].

In the new era, thought itself will
be transmitted by radio

Guglielmo Marconi

Chapter 2

The Ubiquitous Nature Of Signals In Biology And Living Systems

In addition to perpetual change, the real constant across all living systems and aspects of biology, is that *signals* prevail. Communicated from one party to another, a signal may take the form of a hand gesture, a sound, a smell, an excreted chemical compound, a bark, a meow, an electrical impulse, an email, a symbol, a word, or simply anything at all. What ever the signal that has been sent, its existence has been created by a member of the environment to covey some type of concept, sentiment, information, instructions or meaning to another member. We will now explain what is meant by a “signal” in the setting of living systems and biology.

2.1 Traffic Signals

At the city level, many types of signals are obvious and extremely demanding of one’s attention. For instance, on a busy road, we may often hear the sounding of *horns*

from cars and buses. A blaring car horn in traffic is a car's signal to demand attention. Many cases of loud horns may be attributed to the forewarning of danger: perhaps a careful driver has noticed that a careless turn has been made by another car. In this case, the horn may likely be used to signal caution for the avoidance of an accident. The signal may be unconditional and destined to all by-standers and cars in traffic or it may be simple designed for a single particular vehicle.

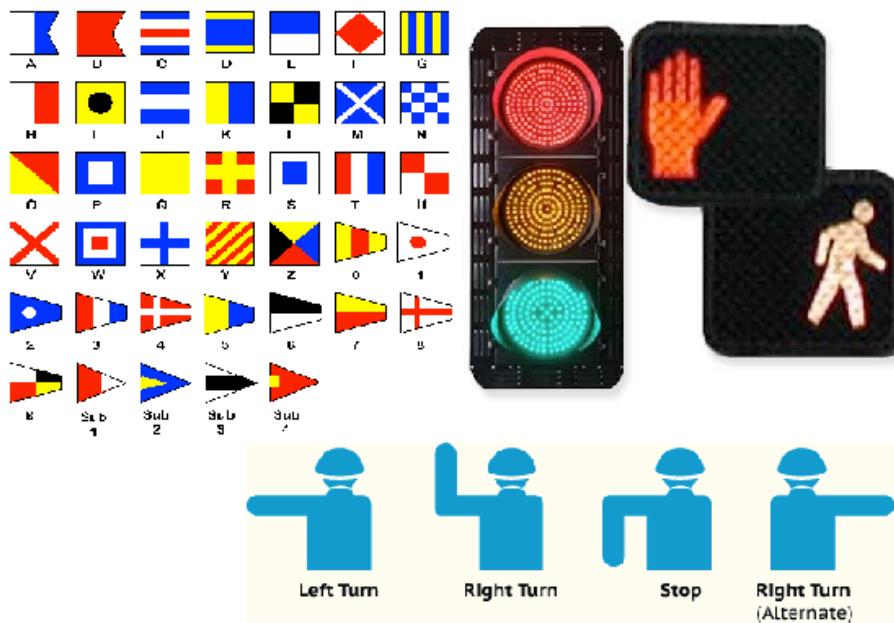


Figure 2.1: One may see signals all over and if one is not familiar with their mechanisms, then their meanings may not be obvious.

The signal may be witnessed, but not understood. When a horn is emitted to a general assembly, there is likely to be some confusion about its intentions. For instance, was the signal made for the prevention of an accident to a pedestrian? Was the signal emitted to prevent a car accident? Was the signal an alert of a collision to the driver's car? It is entirely likely that there was no danger at all and the signal was actually intended as a greeting to some by-stander or car.

Since car horns have been used to send both messages of urgency and greeting, it is often hard to determine why a horn was sounded at all. In this case, an interested

party must look for visual cues at the location of the sound's origin to investigate its possible reasons. If there is (or was) no obvious danger, then one may be inclined to believe that the message was a simple greeting to someone. However, if the sound of the horn was heard, followed by a crash of metal, then it would be obvious that the horn was one of urgency. In either case, an investigation was necessary once the *ambiguous* signal had been perceived.

Other types of signals in the city never have ambiguous meanings, as described in Figure 2.1. For example, the sirens of emergency vehicles, such as ambulances and fire engines, are unlike car honks because they are generally *unambiguous*. These signals are continuous and do not stop sounding until the emergency vehicle has arrived at its destination which ensures that all by-standers will be well aware of them. Any emission of this signal, even if not from a recognized emergency vehicle, will carry the same message of urgency – all members of the community must allow this vehicle a safe, hurried and uninterrupted passage.

The final type of signal in a city scenario is the sound of a bell from an ice cream truck. Although fire engines used bells in the late 1800's and early 1900's, they currently use loud sirens to announce their presence in traffic. Unlike the siren, the bell does not convey a message of urgency but it is none-the-less *unambiguous* in its own right (like the emergency vehicle siren) and requires little further investigation about its implied meaning, unlike the car horn.

We discuss these signals as they are common to our world, yet they are never truly analyzed for what they are - signals to remind the members of their environments of an existence of meanings. One learns at an early age that these signals originate from a mechanism composed of a vehicle and a driver. Additionally, this knowledge is supplemented by the degree of urgency of the mechanisms that they represent. Unlike the siren and the bell, the car horn is ambiguous and, even though one may know much about cars, drivers and their relationships, this signal always requires attention

to completely understand this mechanism. The siren and the bell are unambiguous and, with each sounding, they always signify the same messages and require little attention to understand their meanings.

2.2 Signals Of The Heart

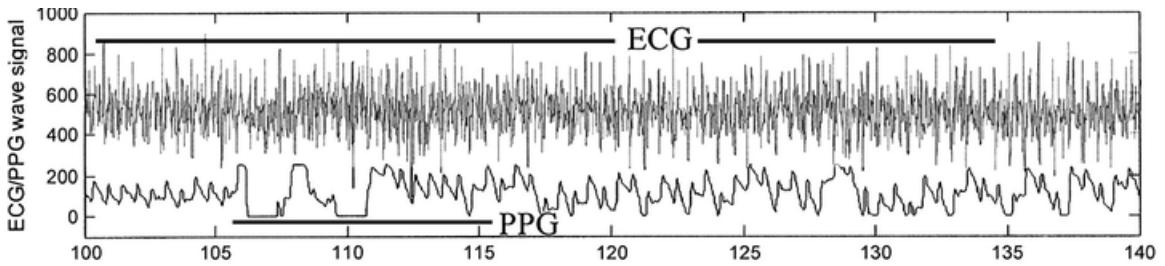


Figure 2.2: The waveforms from PPG and ECG containing noise. The ECG waveform is mostly noisy over a 40-second interval and the PPG waveform is partially noisy. The poor-quality waveform regions have been detected by SVMs and are outlined by thick horizontal bars. This image was taken from Yu *et al.*^[321].

In medicine, the use of signals has proven to be extremely helpful when caring for patients. For instance, one of the most commonly studied signals in medicine is the heart beat because it provides immediate insight into general health. The photoplethysmogram (PPG) provides a volumetric measurement of blood moving under the skin and is obtained by a pulse oximeter that evaluates changes in light absorption due to circulating blood. In this technology, the perfusion of blood to the dermis and subcutaneous tissue of the skin creates the signals that indicate the general health condition of the heart.

Traditionally, PPG signals were determined in small, isolated wave contours on a beat-by-beat basis to ascertain the state of the heart. Since heart beat rhythms may be different from patient-to-patient (especially in cases of atrial fibrillation^[111]), reading isolated regions of waveforms may provide misleading information. Furthermore, analyses taken from automatic beat-reading technologies,

concentrating on these specific regions of waveforms, may be easily confused by noise and artifacts. In these cases, the signals of the heart cannot be correctly studied and understood.

In Antink *et al.*^[7] the interference of noise is discussed. The authors propose an algorithm to remove types of noise for the robust detection of heart beats in multimodal data. In Yu *et al.*^[321], support vector machines (SVMs) are discussed as a method for improving the quality of the electrocardiogram (ECG) and PPG waveforms that have been distorted by noise. The noisy wavelengths, as detected by the SVM approach of Yu *et al.*, are shown in Figure 2.2.

Although some of these methods may concentrate on specific regions of waveforms, it was noted in Elgendi *et al.*^[86], that more information may be extracted about the heart when its PPG recordings are evaluated in entirety (i.e., taken over durations where major trends may be discovered). The authors concluded overall patterns from these recordings may be determined which are sufficient to detect types of heat stress from their effects on the heart's waveforms. Furthermore, knowledge learned from the study of these signals may eventually be used to explore some of the impacts on health from the stresses of global warming.

There are many diverse technologies, algorithms and methods for reading trends from heart beat signals. In a review by Silva *et al.*^[273], these technologies are explored in the effort to accelerate their development and to facilitate the comparison of robust methods for locating heart beats in long-term, multi-channel recordings.

In the above discussion, signals of the heart were used to explore its mechanism of beating to describe general heart health (and, therefore, the condition of the patient's health). Disorders also have mechanisms emitting signals. In the case of the heart, these signals may be noted in the ECG and PPG waveforms where the waves are malformed or, perhaps even missing. Studies, such as those mentioned above (having goals of improving the reception capabilities of natural signals) are hence, important

since they may likely uncover some of the disorders that impact the mechanisms maintaining life.

Below, we describe other types of signals which stem from diverse and functional mechanisms in nature. These signals, like those of the heart, may provide information into the nature of the organism's health or condition of living.

2.3 Natural Signals Of Organismal Biology

2.3.1 Mobile Organisms

At the organismal level, humans and animals often use visual and audible signals to express all feelings of happiness, sadness, peace and aggression. A visual signal between humans may be a hand gesture in the form of a salute to indicate respect. The hand could also be used to make a wave to convey types of greetings. Animals use visual cues from each other to determine types of aggression. For instance, dogs in packs show respect for their superiors and masters by bowing their head and avoiding direct eye contact to display their non-intention of aggression.

Signals are also used to identify dangerous types of organisms which could cause harm if bothered. For example, the black widow spiders of the family *Theridiidae* often have markings on their bodies, such as red spots which indicate their poisonous nature. In addition to sending threatening signals to others, these spiders, working in groups, also send signals to each other and were described by Krafft and Pasquet^[159] who studied hunting activities. During the hunting phase, multiple spiders are spread around the webs. When a thrashing of a struggling prey is detected in the web, the spiders venture to the epicenter to entangle, wrap and bite the prey. As they travel towards the prey, they move in a synchronized and rhythmical venture - each spider moves in function of the others. This interaction may likely prevent their combined movements of creating violent trembles in the web that could shake the prey free. It

was by studying their actions of the signals of this coordinated hunting tactic could be recognized.

Snakes also exhibit distinguishing signals to threaten attackers. The python of the family *Pythonidae* has colored markings, resembling wavy segments and the *Eunectes*, a genus of boas, have similar types of blotchy markings to intimidate attackers. There are countless examples of these types of visual signals made up of bright, colorful patterns to describe types of jeopardy. Any camper or animal who happens to encounter such a pattern in nature will quickly associate this type of signal to danger. It is likely that this signaling system has facilitated the survival of these snakes (as well as, spiders, and other types of perilous organisms) by exhibiting messages which are associated to pain. These snakes occasionally prey on porcupines which have long prickly spines which could be dangerous to snake as they are able to puncture snake guts when a porcupine is swallowed. In Duarte^[82] it was noticed that these porcupines are often ambushed by snakes in Africa, America, and Asia. During the hunt, the porcupine is mistaken for a rodent as the natural signals (i.e., quills and the extension of spines and other warnings) are unnoticed by the snakes. After they are consumed, surviving snakes may learn to never again eat porcupines due to the distress of the indigestible spines. This mechanism of porcupine survival was made evident by the visual cues of the snake - porcupine interactions.

Wild bears have many signals of aggression to keep other animals (and humans) away, yet some of children's favorite possessions are teddy bears. Interestingly, teddy bears are popular because this particular (stuffed) animal lacks any of the symbolism for aggression and natural fear. For instance, real bears have strong claws, big teeth and powerful growls, however, a teddy bear has no claws, teeth or growl. Instead the teddy bear is soft and inviting of bed-time cuddling. In absence of the signals of jeopardy (and its likely onset of suffering), a fuzzy teddy bear emits

signals of vulnerability to make it extremely inviting to small children who are especially talented at noticing these signals^[294]. The adoption of teddy bears was thought to originate from the mechanism of natural aversion to danger.

2.3.2 Immobile Organisms

Although they are unable to move and speak, plants convey very specific signals to their environments to prevent the interference of others that likely facilitates their rate of survival. To ward off the animals who may feed off their stems, leaves and fruit, raspberry plants defend themselves by displaying several visual signals – typically bright-red thorns and prickly stems. The color red has often signified toxicity across biology, such as in the case of poisonous berries from types of bushes. On raspberries, although the berries are not toxic, the plants defend themselves by their thorns to invite thoughts of suffering and deter those who would pick the berries.

Interestingly, in McMenemy *et al.*^[199] it was discovered that types of viruses, targeting raspberry plants, are able to manipulate the plant's glutamate levels to deter aphid attacks on the plant. The reduced glutamate, coupled with virus infection and abnormal aphid attacks, provided the signals necessary to determine a mechanism of a virus-orchestrated defense of plant and habitat.

The *Amanita muscaria* mushroom is a psychoactive *basidiomycete* fungus native to temperate and boreal regions of the Northern Hemisphere which also uses signals to deter its attackers. Commonly known as the *fly agaric* or *fly amanita*, this mushroom exhibits a bright red top with white speckles to signal its potent danger to humans and to animals who may graze on it. Since the mushroom is highly poisonous when raw, it is conceivable that if its bright color fails to deter, then the poison will have an immediate effect on those who interfere with its survival tactics.

Plant signals for defense may be observed in *Acacia raddiana* and *A. tortilis* trees which are found in semi-arid savannas and also in the deserts of northern Africa

and the Middle East. Despite their thorns, when a tree in a population is grazed upon, there is a production of tannin which serves two purposes: (1) a protection from the nearby animals and (2) a warning to other trees of the population that there are grazers are in the neighborhood^[251]. The airborne scent of the tannin is detected by other *acacia* trees, which start their own productions of tannin to deter grazing. This chemical compound serves to deter grazers by interfering with digestion. For instance, condensed tannins, polymers composed of 2 to 50 (or more) flavonoid molecules, inhibit herbivore digestion by binding to consumed plant proteins and making them more difficult for animals to digest, and by interfering with protein absorption and digestive enzymes^[295].

Protection of plants by signals is discussed in Lev-Yadun *et al.*^[176]. Since many of the insect grazers receive food and nourishment from plants, the visual perception (i.e., the colors, textures and patterns) of both herbivores and predators co-evolved with plants. During this evolution, these organisms developed crypsis (i.e., camouflage or, the ability to avoid observation or detection by others) which was aligned to the dominant colors, textures and patterns of plants. However, plants have evolved on their own to become generally too colorful to enable the herbivorous insects and other invertebrates to employ a universal camouflage system to hide from predators. Similar to the concept of the peppered moth, which was able to hide from predators and flourish only when its natural color was similar to the grey hue of post-industrial buildings where it lives, the parasitic insects of plants were unable to hide with the alteration of plant color. As they were unable to hide from predators, the insect populations declined and the plant populations were able to flourish. We note here that it was simply a strategy of switching signals which changed the survival rates of the plants and the insects.

At the microscopic-level, signals also prevail. For example, one of the most famous examples of signalling in biology is found in the brain. Defined here as gap

junction-mediated connections, many neurons in the mammalian central nervous system communicate through electrical synapses all around the brain. Discussed in Connors and Long^[64] is the existence of countless undiscovered electrical synapses throughout the central nervous system which serve to move electrical signals from cell to cell much like a computer network. Each process of the brain is likened to an interaction where a gap junction protein, connexin36 (Cx36) is applied for robust electrical coupling to transfer electrical signals. It is often not the signal itself (electrical or chemical) which is the focus of these brain pathways. The speed of an impulse passing through the different types of synapses and receptor mechanisms (including ionotropic receptors, metabotropic receptors, in addition to the gap junctions) describes the absolute necessity for signals in brain science and, perhaps in molecular biology, alike^[78].

There are many other examples of signals of mechanisms in nature which were first noticed by investigators who were interested in discovering and exploring their mechanisms. It is suggested here that the isolation of signals must come before any confirmation may be made of a mechanism. As we have seen above, once these signals are found by observation or in the experimental data, then the search and study of their mechanism may be performed.

2.4 Signals, Semantics And Comprehension

Most, if not all signals are at one time uncertain and their meaning must be comprehended before their messages may be interpreted. Often, the context of a signal provides much about its intended meaning. Here, we claim that the signal arises from some mechanism (known or unknown). It is by an initial examination of the signals (in its own environment) that allows one to better understand its meaning. For instance, a car honk placed in a quiet library would create much

confusion about its meaning. Furthermore, even in its rightful setting, the car horn's intended meaning may not be fully understood until its environment has been fully explored to understand its context. By itself, the car horn has no meaning when there are so many meanings that the noise could imply.

On the other hand, the sirens of the emergency vehicles create a context of emergency, as recognized by others in the environment. From their experience with other sirens of emergency vehicles, a traveler would likely have recognized that these signals carry a sense of urgency and caution. One may hear a car horn many times each day but it is only when context is known which caused the signal (i.e., to give an alert or a greeting) that one may uncover more about the actual meaning for the signal.

In the same way, signals stemming from dangerous plants also imply a degree of urgency which must be understood from a contextual standpoint. A human or grazing animal may recognize a pattern of thorns on the raspberry plant or the bright patterns on snakes (and perhaps, even their hissing sounds). After they are understood, these signals cue the recollection of discomfort or fear from previous experiences and deters interference, which may be an advantage for both dangerous organisms (such as a snake) and also for the traveler or grazer who encounters it.

Management by objective works - if you know the objectives. Ninety percent of the time you don't.

Peter Drucker

Chapter 3

Objectives

3.1 Motivation

In this thesis, we create a framework from tools and applications of analysis to discover signals from biological mechanisms. We maintain the idea that the isolation of signals is the first step in understanding various cellular mechanisms. We offer a framework, built from computational tools and analysis, to investigate signals to describe the existence of mechanisms and their general natures. In each chapter below, we develop our framework to study different types of signals originated in DNA, RNA and protein. We then apply this framework to specifically investigate the signals of PTMs, associated MSs interactions and the stresses that motivate them during protein stress responses.

3.2 General Organization

We group the contributions of this thesis into chapters to imply the types of signals they discuss: DNA, RNA and protein. In the sequences of DNA, RNA and protein, there is a grammar and syntax necessary to hold their information and so in Chapter 4, we describe the mathematical and statistical manipulations for the detection of signals from any genetic code. In this chapter, we demonstrate how mathematics may be used to differentiate sequences by comparing information content. In Chapter 5, we investigate the signals in DNA of restriction enzyme locations which resemble palindromic words (i.e., motifs) in the code. In Chapter 6, we show how the signals from permutations of DNA motifs may be used to determine the origins of contigs from an assembly. In Chapter 7, we show how the natural signals of the central dogma of biology (i.e., DNA gives RNA gives protein), has signals which can be used to encrypt human-language text.

Next, we move our attention to the RNA code. In Chapter 8, we show how the signals left over from DNA processes may eventually interfere with those of RNA. In Chapter 9 we study signals from protein post-translational modifications (PTMs). We show these signals may likely be affected by those from RNA (the phase before protein synthesis) and also influenced by typical protein code syntax signals.

Lastly, we focus on protein signals. In Chapter 10 we show how the signals due to PTMs may be used to indicate types of protein folding. In Chapter 11, we discuss and illustrate that the PTM signals become increasingly complicated in higher developed organisms. In Chapter 12, we demonstrate how these and other signals may be isolated from the literature and used to determine who the actors are in a particular stress response. Finally, in Chapter 13, we discuss that signals may be found from the distances from PTM sites to the protein domains which are thought to make the conformational changes necessary for the maturation of the protein for the completion of specific tasks.

3.3 Thesis Contributions

3.3.1 Contribution 1

One of the leading problems in bioinformatics is to compare sequences to determine their similarities. However, determining these likenesses is often confused by issues such as synteny, or the physical co-localization of genetic loci on the same chromosome within an organism. Dynamic programming techniques, such as the Smith-Waterman and Needleman-Wunsch algorithms, are often confused when working with sequences where key regions are not found in the same orders. Another approach to detecting similarity is to apply alignment-free methods (i.e., non-dynamic programming and statistically-based) to calculate the amount of common information between the sequences that is not concerned with the exact location of the information in the sequence. In Chapter 4, we review many of the popular mathematical and statistical methods to measure informational content content between DNA, RNA or protein sequences.

3.3.2 Contribution 2

When a virus infects a cell, its DNA is injected and mixed in with that of the cell's so that the cell will begin to manufacture the virus to continue the infection. However, one of the earliest defenses that cells use to protect themselves from this kind of infection is to lacerate all foreign DNA by deploying restriction enzymes which cut at specific regions of DNA. The cuts to kill the invading parties happen at programmed motifs (i.e., a specific word) in the code which, if found, is cut. These words are palindromic, meaning that they appear to be the same backwards as forwards on each strand of the DNA. Since these words could equally appear in the host's DNA, they are automatically methylated to prevent a restriction enzyme from being able to cut it. In Chapter, 5, we investigate the locations of these words using statistical

tools to compare their populations between coding (i.e., regions of DNA that code for protein) and non-coding regions of DNA. We determine that there are fewer of these words in coding regions to suggest that these stretches of sequence are intolerant of non-coding information.

3.3.3 Contribution 3

To obtain the exact code of DNA for an organism, the DNA is prepared from a sequencing machine in tiny segments called reads. Contigs are reads which have been joined together like a jigsaw puzzle pieces. Before it returns one large, continuous DNA sequence, all the tiny reads must be assembled into larger contigs and then into a larger sequence by the automation of a computer. This task is computationally intensive and may take hours to days to complete. In Chapter, 6, we describe that this assembly task can be preprocessed to speed-up the process. Our method clusters the reads according to their informational content and generates smaller pools from which the processor has to compare each read to all others. When the task of comparing all the reads can be reduced to smaller tasks, then the work proceeds faster with less computation. This work introduces the notion of spectrum sets, (i.e., sets of signaling motifs) used to cluster the sequence reads during this pre-processing step.

3.3.4 Contribution 4

In Chapter, 7 we introduce an encryption system which uses the central dogma of biology. Since the method of converting DNA to protein must follow rules for this natural system to work each time throughout biology, we adapt these rules to convert human-language messages into DNA which is then “translated” into protein. The protein sequence is the encrypted text which can be safely sent through public communication channels. The keys to the encryption are taken from the massive amounts of restriction enzyme data which is publicly available to

resolve the problem of passing keys from sender to receiver. This contribution supports the notion that signals from biology may be used in computer science for secure communications and security.

3.3.5 Contribution 5

Naturally occurring along the genomes of many viruses and other pathogens, short palindromic restriction sites (<14bps) are often exploited by bacterial restriction enzymes as autoimmune defenses to end pathogen threats. These motifs may also appear in the host's genome where they are methylated so as not to attract restriction enzymes to the host's genetic material. Since these motifs in the host's genome may pose a significant danger, it is likely that their numbers have been reduced due to possible failures of methylation during evolutionary time. To reduce the chances that methylation failure could happen, there are generally fewer of these palindromic words in DNA. During translation, tRNAs, as directed by the information in DNA, are responsible for the delivery of unique amino acids into the protein sequence. In this contribution, we noted that these missing words correlated to missing tRNA proteins as well. In Chapter 8, we describe this phenomenon after having studied the DNA motif signals of eight organisms and concluded that DNA mechanisms affect the other mechanisms down the chain of events for building protein.

3.3.6 Contribution 6

In Chapter 9, we begin our exploration of the signals stemming from the interaction between PTMs and the amino acid modification sites (MS) in protein. By studying the exact types of PTMs and the MS of nine different organisms, we note that the usage of PTMs in protein was biased since each organism had a unique collection of PTMs involved with its proteome. In this contribution, we discussed that PTMs

are not applied uniformly across organismal proteins and differing PTM preferences and usages may often exist between proteins of the same organism. We study the frequency of factors (PTM predominance and their associated active sites, tRNAs and amino acids) which likely influence a PTM bias and conclude that many signals of one particular organism are dissimilar from those of another, in terms of protein modification.

3.3.7 Contribution 7

In protein, there are specific types of motifs made up of amino acids which attract oxidative carbonylation (i.e., a biological rust). The literature has suggested that motifs composed of R, K, T, P, E and S residues in protein sequence are more likely to degrade when a protein is exposed to a stress such as microgravity (i.e., weightlessness) which is known to cultivate oxidation. Since Mt also generate natural oxidation, this study was designed to determine whether the composition of these words was fewer for organelles than that of the non-Mt proteome. In Chapter 10 we confirm that Mt protein has adapted itself over evolutionary time to have fewer motifs where oxidation would interfere with general protein function.

3.3.8 Contribution 8

Previously, in our work, we noticed that organisms have different arsenals of PTMs to apply to their proteomes for protein stress or other types of protein modifications. In Chapter 11 we noted that the increasing complexity of the 11 organisms of our study appeared to correlate with an increase in available PTMs and MSs of the proteome. Since PTMs enable proteins to withstand types of stresses and different stressed proteins may require different types of PTMs for adaptation, we concluded that complexity may be a function of the types of stresses that an organism may tolerate in its environment. For instance, the organisms having the most PTMs were

the human and the plants. We speculated that plants are immobile and cannot leave their environment. They, therefore, must be able to withstand any stresses of their environment for survival. Humans, we noted, appear to be able to live under a wide variety of stress conditions (i.e., diverse habitats) as well which is likely supported by their collection of PTMs.

3.3.9 Contribution 9

In Chapter 12 we introduce a text mining application (called “Lister”) that is able to search through the entire collection of PubMed literature (hundreds of thousands of articles) to determine the associations between common stresses, proteins and PTM types (i.e., the actors of the study). The tool is highly customized for this task and uses a supervised approach. Understanding how the literature links these three actors involved with any stress response in the proteome is pivotal to determining more about the mechanism in which these actors play a role. In addition, since their roles are determined from curated literature, they are also established with some certainty. To help form some understanding how their relationships, our tool outputs network graphs to visual show how stresses, PTMs and proteins have been connected in the literature.

3.3.10 Contribution 10

PTMs in a protein leads to conformation changes. It is believed that domains are being influenced by PTMs, and in this scenario, it is likely that the exact position of the amino acid MS, relative to the domain, plays a major part in how these structural changes are influenced by PTMs. In Chapter 13 we present a system called “PTM Tracker”, to study the general distances of MS amino acids which are relative to (i.e., before, inside and after) known functional domains of proteins. The data for this project was obtained from public databases. In this contribution, we study the

distances of these relevant MSs to their corresponding domains across the 11 diverse organisms of our study. Using our computational tool for automating this study for determining the signals of mechanisms involved in PTM - domain interactions, we conclude that many domains of each organism are generally situated at specific distances from the MS with which they are likely to interact. Each organism has slightly different measurements to imply that the mechanism across organisms is biased.

If I were again beginning my studies, I would follow the advice of Plato and start with mathematics.

Galileo Galilei

Chapter 4

Alignment-Free Genetic Sequence Comparisons: A Review of Recent Approaches By Word Analysis

4.1 Abstract

Modern sequencing and genome assembly technologies have provided a wealth of data which will soon require an analysis by comparison for discovery. Sequence alignment, a fundamental task in bioinformatics research, may be employed but with some caveats. Seminal techniques and methods from dynamic programming are proving ineffective for this work due to their inherent computational expense when processing large amounts of sequence data. These methods are prone to giving misleading information, because of genetic recombination, genetic shuffling and other inherent biological events. New approaches from information theory,

frequency analysis and data compression are available and provide powerful alternatives to dynamic programming. These new methods are often preferred since their algorithms are simpler and are not affected by synteny-related problems.

In this contribution of the thesis we provide a detailed discussion of computational tools, which stem from alignment-free methods based on statistical analysis from word frequencies. We provide several clear examples to demonstrate applications and the interpretations over several different areas of alignment-free analysis such as base-base correlations, feature frequency profiles, compositional vectors, an improved string composition and the D_2 statistic metric. Additionally, we provide detailed discussion and an example of analysis by Lempel-Ziv techniques from data compression.

4.2 Introduction

Gene structure, function and phylogenetic relations are discovered by the basic comparison of known to unknown genetic material across organisms. Sequence comparison is pivotal to the success of basic phylogenetic and metagenomics research. For instance, large portions of common genetic material between organisms provide much evidence to suggest that they are somehow related. Furthermore, similar sequence data fuels conjecture that the associated functions are also similar.

Comparative research came from computer science which provided tools and algorithms to find specific substrings in larger sequences^[109] for discovery. For instance, the Knuth-Morris-Pratt algorithm^[152], the Boyer-Moore^[40] algorithm was used initially in the 1970's^[126] to locate regions of common DNA by exact matching of larger sequences. Later, a modified version of the Boyer-Moore^[8] was applied in the 1980's. Since these algorithms assumed that the input strings contained exact matches, tiny mismatches found in DNA interrupted performance. This led to

algorithms for approximate pattern matching^[292] and others^[59;207].

Due to the growth of inexpensive computing and improvements in sequence assembly technologies, there is now more sequence data available to bioinformatics research than ever before. Comparative genomics has been an obstacle to discovery^[156;311] and still manages to be a major factor in more current applications. Some of these applications include sequence assembly^[179], evolutionary history comparison involving complications from synteny^[184], horizontal gene transfer discovery^[25;75], analysis by gene-shuffling^[67] and many other applications where proper sequence comparison must be used^[214].

Dynamic programming^[84] has often been applied to comparing sequences in the above-mentioned applications. Since global and local alignment algorithms^[208;279] work base-by-base, they stand to be confused by the inherent mismatches, gaps, alternating blocks of sequence material and inversions, that are easily found in genetic material. These methods may erroneously conclude that the functionally related sequences are largely unrelated since they do not demonstrate any statistically significant alignment. Sequence length is also important to address when running an alignment from dynamic programming. For example, local and global, implemented in softwares such as ClustalW^[166], have complexities of $O(mn)$ and so it is clear that their resource requirements quickly escalate for larger sequences of lengths, m and n . It is often infeasible to perform comparisons of complete genomes by this approach due to the large amount of time this would involve. For this reason, technologies requiring databases for speed such as BLAST^[5], BLASTZ^[56] and BLAT^[147] have gained popularity. Other methods to help overcome some of the limitations of dynamic programming have come from diverse fields such as: cloud computing^[257], distributed computing,^[60] and parallel computing for multiple sequence comparison,^[145].

Frequency-based algorithms, which are driven by the statistics of word usage or

similar, are becoming popular in research for discovery. This is because these approaches are not typically confused by the complexities caused by mismatches, gaps and sequence inversions that are often found between sequences for comparison^[114]. For example, these methods function by evaluating the informational content between sequences and so alternating blocks of DNA between two sequences will not be problematic. This form of alignment does not depend on where the features are found in the sequence, only that the sequence contains the features. Methods using frequency analysis also do not suffer from high algorithmic complexities as they are generally linear. They are, therefore, able to process larger sequences with fewer resources than dynamic programming algorithms and do not rely on having database support, as would, BLAST, BLASTZ and BLAT. There is clearly a call for an alternative approach for sequence comparison done by methods which are not of dynamic programming and so, alignment-free methods are becoming very attractive to bioinformatics research where the data is substantial and naturally dynamic.

For this contribution, we discuss some of the prominent methods stemming from vector or frequency-based analysis such as: base-base correlations, feature frequency profiles, compositional vectors, improved string composition and the D_2 statistic metric. These methods have been chosen for discussion because of their simplistic nature and ease of application to research. We provide clear examples for the implementation of these methods and discuss their interpretation. We also provide discussion and an example of a method inspired by the Lempel-Ziv compression techniques. This contribution aims to show how these alignment-free methods are integral to the quantification and discovery of sequence function and structure.

4.3 Background

Methods for differentiating sequence data by using statistical concepts (factor frequencies and approaches from data compression) have attracted much interest. In their often cited 2003 publication, Vinga *et al.*^[296] reviewed some related methods, metrics and algorithmic implementations. S. Mantaci *et al.*^[278] continued by illustrating other methods recently introduced for the alignment-free comparison which were also based on a statistical approach. The authors organize the comparison algorithms in the following basic groups:

- Count factor frequencies
- Data compression
- Edit distances, or on block edit distance - a special case involving moving entire blocks of a sequence.

Recent developments and the release of new technologies from the scientific community have caused the above references to become out-dated. Here, we discuss some of the more recent statistical methods which involve frequency data for comparison. The approaches that we cover were chosen based on their simplicity of application and can be divided into the following categories: *factor frequencies* (^[190]), *composition vectors* (^[54]), *improved compositional vectors* (^[192]), *data compression* (^[48;215;293]) and *common substrings* (^[76;292]).

4.4 Factor Frequencies

Producing seminal ideas in 1948, C.E. Shannon's *Information Theory* is the branch of mathematics which is concerned with quantifying information and signal processing^[267]. Since DNA contains observable structures and patterns^[149;304;327], tools from information theory (e.g., mutual entropy *et al.*) are appropriate for

frequency analysis. Many of these methods break each sequence for comparison into numeric parts such as frequencies from the occurrence of types of words or k -mers (substrings of length k) occurring in the sequences. If two sequences are similar, then the derived k -mer frequencies would have similar distributions to reflect this likeness. If the sequences are different, then so are the frequency distributions.

To perform a k -mer study, the size of the motif is an important factor to consider. When collecting word frequencies from motifs, the size of the motif does make a difference to the results. According to^[313] where the length of motif or window size is extensively discussed, there is a general rule of play when collecting word frequencies. When the sequences are obviously very different (they are not related, for example), then size of k -mers or window-size should be short. However, when the sequences are very similar (known to be related) then the k -mers or window sizes can be longer. The reader is invited to consult the above reference for the details behind their general rule.

4.4.1 BBC By Analysis Of Mutual Information

Mutual information is a tool from information theory, which measures the amount of common information (or interaction) between two entities. Liu *et al.*^[190], described the development of *Base-Base Correlations* (BBC), an algorithmic approach for determining sequence similarity by mutual information to infer phylogenetic relationships from complete genomes. In their work, an interval is established containing r -bases, making up strings of DNA to be used for multiple sequence comparison. In this interval, a vector is created from all possible joint probabilities of DNA pairs, since the total possible pairs = $4 * 4 = 4^2 = 16$. In their paper, they showed that the interval containing these joint probabilities in the sequence can often be expanded to get a better measurement of the difference between sequences.

For $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) \equiv (\text{A}, \text{C}, \text{G}, \text{T})$, the probability of finding base α_i is denoted p_i for $1 \leq i \leq 4$. For $T_{ij}(r)$, the average relevance of the two-base combination (the feature) with different gaps from 1 to r (a range of r), the authors define a BBC by the following:

$$T_{ij}(r) = \sum_{d=1}^r p_{ij}(d) \cdot \log_2 \left(\frac{p_{ij}(d)}{p_i p_j} \right) \quad (4.1)$$

for $i, j \in \{1, 2, 3, 4\}$ where $p_{ij}(d)$ signifies the joint probabilities (e.g., the $4^2 = 16$ possible length-2 DNA words which we refer to as *features*) of bases i and j at a distance of d . A BBC feature constitutes a 16-dimensional feature vector, V_{S_1} for a sequence S_1 having a length of n_1 .

The statistical independence of two bases for a sequence of length- l is defined by $p_{ij}(l) = p_i p_j$ and its deviation is defined, $D_{ij} = p_{ij}(d) - p_i p_j$. Let $S_1 = \text{ACGTGCTATG}$ and $S_2 = \text{ACGGCGCTA}$. We find the joint probabilities to populate the vector,

$$(\text{AA}, \text{AC}, \text{AG}, \text{AT}, \text{CA}, \text{CC}, \text{CG}, \text{CT}, \text{GA}, \text{GC}, \text{GG}, \text{GT}, \text{TA}, \text{TC}, \text{TG}, \text{TT}),$$

with the following equation for frequency, $f(W_k)$, from^[69]:

$$f(W_k) = \frac{c(W_k)}{n - k + 1}, \quad (4.2)$$

where $c(W_k)$ signifies the number of occurrences of a length- k word in a sequence of length- n_1 . The finalized vectors are the following.

$$\begin{aligned} V_{S_1} = & (0.0, 0.2, 0.0, 0.1, 0.0, 0.0, 0.1, 0.1, \\ & 0.0, 0.1, 0.0, 0.0, 0.1, 0.0, 0.3, 0.0) \end{aligned}$$

$$\begin{aligned} V_{S_2} = & (0.0, 0.3, 0.0, 0.0, 0.0, 0.0, 0.0, 0.5, 0.5, \\ & 0.0, 0.5, 0.0, 0.0, 0.5, 0.0, 0.0, 0.0) \end{aligned}$$

For two sequences S_1 and S_2 having the same length n_1 , the authors define the distance $H_{S_1S_2}$ in the following equation.

$$H_{S_1S_2} = \sqrt{\sum_{i=1}^{16} (V_{S_1i} - V_{S_2i})^2} \quad (4.3)$$

By this calculation, we find that $H_{S_1S_2} = 0.8890$ for the example above. Higher values for this metric indicate a greater spread in the frequency distribution and increasing dissimilarity, however, lower values indicate levels of increasing similarity (e.g., 0 if and only if the distributions being compared are equivalent). The authors note that $H_{S_1S_2}$ satisfies the definition of a sequence distance because (i) $H_{S_1S_2} > 0$ for different sequence lengths: $n_1 \neq n_2$; (ii) $H_{S_1S_2} = 0$; (iii) $H_{S_1S_2} = H_{S_2S_1}$ (symmetric); (iv) $H_{S_1S_2} \leq H_{S_1} + H_{S_2}$ (triangle inequality).

Liu *et al.* used phylogenetic trees, employing branch weights gained from their BBC mutual information calculations. From the sequence data of 48 different *Hepatitis E* viruses, they constructed a phylogenetic tree which was consistent with previous studies by diverse approaches^[190].

4.4.2 Feature Frequency Profiles (FFPs)

In^[10], a feature frequency approach (*UVWORD*) was presented which compares the DNA words from two sequences. Known as *oligonucleotide profiling*, the sliding-window method compared the encountered word frequencies of one sequence (the *target*) to another (the *source*). The sequence similarity was determined by how many words were common to both sequences. Word-based statistical models were also presented in^[69] which investigated the occurrence, type and frequency of overlapping and embedded DNA words for sequence comparison.

Sims *et al.*^[275] were interested in comparing whole genomes, even in situations where there are no common genes with high homology. To do this, they developed

a variation of text compression, where the distance between word frequency profiles of two texts would be taken as a measure of dissimilarity. They substituted relative k -mer frequencies (Feature Frequency Profiles, or FFPs) for word frequencies.

A sliding window of size k is run through the sequence from position 1 to $n - k + 1$ and counts the number of all $t = 4^k$ possible k -mers (the total number of features, for example) where four is the number of DNA bases. Although the k -mers extend themselves throughout the entire genome, the window is only allowed to span over the regions which are completely free of sequencing gaps. The vector $C = \langle c_1, \dots, c_t \rangle$ holds the t number of raw frequency counts for all possible words of length- k and is conventionally found by the following equation:

$$\mathbf{F} = \mathbf{C} / \sum_i c_i. \quad (4.4)$$

It is important to note that the length of the genome must be considered carefully at this vector-forming stage. If the genomes are of approximately equal length, and a less than 4-fold difference exists between sequences (four is the number of bases), then the method is conveniently employed. However, if the sequences for comparison have extremely different lengths, then it is necessary to implement the block-FFP method which is similar to the method described by^[313]. This pre-processing step works to ensure that diverse genome lengths do not yield misleading results.

This step breaks up each sequence into smaller, manageable fragments of length- n_1 (called FFP-blocks). In the case where the length of the shorter sequence is evenly divisible by the length of the longer sequence, the intervals (e.g., blocks) are made so that they have the same length as the shorter sequence. If a sequence (length n_2) is not evenly divisible by the shorter sequence (length n_1) then the total number of possible blocks for comparative analysis that can be made is n_2 modulus n_1 .

4.4.2.1 A Comparison By Frequencies And The Jensen-Shannon Divergence Test

Comparing genomes is actually comparing the sets of frequencies which have been taken over an interval of sequence data. To make this comparison, we will follow Sims *et al.*^[275] approach to employ the Jensen-Shannon Divergence (JSD) test. The JSD test is a close relation to the Kullback-Leibler Divergence (KLD) test, an information theoretic, non-symmetric divergence measure of two probability distributions, that is extensively discussed in^[186].

Once the vectors have been properly created, we are ready to apply the calculations which determine their distance apart. For two arbitrary vectors, V_{S_1} and V_{S_2} , prepared from sequences S_1 and S_2 for t , the number of features collected, the JSD is given below:

$$JS(V_{S_1}, V_{S_2}) = \frac{1}{2}KL(V_{S_1}, V_M) + \frac{1}{2}KL(V_{S_2}, V_M), \quad (4.5)$$

where,

$$V_{M_i} = \frac{V_{S_{1i}} + V_{S_{2i}}}{2} \quad (4.6)$$

for $i = \{1, \dots, t\}$ and KL is the KLD, below.

$$KL(V_{S_1}, V_M) = \sum_{i=1}^t V_{S_{1i}} \log_2 \frac{V_{S_{1i}}}{V_{M_i}}, \quad (4.7)$$

where t is the number of features.

We now return to our earlier example of the two sequences $S_1 = \text{ACGTGCTATG}$ and $S_2 = \text{ACGCGCTA}$, which we compared by this JSD analysis. In this example, we populate vectors for these sequences using all length-2 words (2-mers) in the sequences. The possible 2-mers, are ordered in the following order:

AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT.

The FFP vectors V_{S_1} and V_{S_2} are created and populated by all available 2-mers from sequences S_1 and S_2 .

$$V_{S_1} = \langle 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 2, 0 \rangle * \frac{1}{9}$$

$$V_{S_2} = \langle 0, 1, 0, 0, 0, 0, 2, 1, 0, 2, 0, 0, 1, 0, 0, 0 \rangle * \frac{1}{7}$$

At each position i of both vectors, we apply $V_{M_i} = \frac{V_{S_1i} + V_{S_2i}}{2}$ to get the an average vector V_M . The calculated values for all three vectors, are shown in Table 4.1.

Table 4.1: Positions 1 through 16 of the table of vectors for V_{S_1} from $S_1 = \text{ACGTGCTATG}$ and V_{S_2} from $S_2 = \text{ACGCGCTA}$, aligned with position. The elements of combined vector \mathbf{M} by index is also shown. Frequencies of each 2-mer are made by normalizing the occurrences of each 2-mer in S_1 and S_2 , respectively, by the total number of 2-mer occurrences in each sequence.

2-mers	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
Position i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
V_{S_1}	0	$\frac{1}{9}$	0	$\frac{1}{9}$	0	0	$\frac{1}{9}$	$\frac{1}{9}$	0	$\frac{1}{9}$	0	$\frac{1}{9}$	$\frac{1}{9}$	0	$\frac{2}{9}$	0
V_{S_2}	0	$\frac{1}{7}$	0	0	0	0	$\frac{2}{7}$	$\frac{1}{7}$	0	$\frac{2}{7}$	0	0	$\frac{1}{7}$	0	0	0
V_M	0	$\frac{8}{63}$	0	$\frac{1}{18}$	0	0	$\frac{25}{126}$	$\frac{8}{63}$	0	$\frac{25}{126}$	0	$\frac{1}{18}$	$\frac{8}{63}$	0	$\frac{1}{9}$	0

To help the reader to keep track of the vectors and their frequencies at each position, we offer Table 4.1. We apply vectors V_{S_1} and V_M (and then vectors V_{S_2} and V_M) to Equation 4.7 which we illustrate below.

$$\begin{aligned}
KL(V_{S_1}, V_M) &= \sum_{i=1}^t V_{S_1 i} \log_2 \frac{V_{S_1 i}}{V_{M_i}} \\
&= \frac{1}{9} * \log_2 \left(\frac{\frac{1}{9}}{\frac{8}{63}} \right) + \frac{1}{9} * \log_2 \left(\frac{\frac{1}{9}}{\frac{1}{18}} \right) \\
&\quad + \frac{1}{9} * \log_2 \left(\frac{\frac{1}{9}}{\frac{25}{126}} \right) + \frac{1}{9} * \log_2 \left(\frac{\frac{1}{9}}{\frac{8}{63}} \right) \\
&\quad + \frac{1}{9} * \log_2 \left(\frac{\frac{1}{9}}{\frac{25}{126}} \right) + \frac{1}{9} * \log_2 \left(\frac{\frac{1}{9}}{\frac{1}{18}} \right) \\
&\quad + \frac{1}{9} * \log_2 \left(\frac{\frac{1}{9}}{\frac{8}{63}} \right) + \frac{2}{9} * \log_2 \left(\frac{\frac{2}{9}}{\frac{1}{9}} \right) \\
&= 0.1943
\end{aligned}$$

Following this theme for sequence S_2 (vector V_{S_2}), we find that $KL(V_{S_2}, V_M) = 0.3734$.

$$\begin{aligned}
JS(V_{S_1}, V_{S_2}) &= \frac{1}{2}KL(V_{S_1}, V_M) + \frac{1}{2}KL(V_{S_2}, V_M) \\
&= \frac{1}{2} * (0.1943) + \frac{1}{2} * (0.3734) \\
&= 0.2839
\end{aligned}$$

Provided the base 2 logarithm is employed, the JSD is bounded below by 0 and 1^[186]. Higher values indicate increasing dissimilarity but lower values indicate increasing similarity (e.g., 0 if and only if the distributions are identical). Since $JS(V_{S_1}, V_{S_2}) = 0.2839$ is close to zero, we conclude that the sequences S_1 and S_2 are similar by this test.

Sims *et al.*^[275] reconstructed phylogenies from concatenated mammalian “intronic genomes” by this method and found that their method closely reflected the accepted evolutionary history, and agreed to results from a codon-sequence-based alignment

technique^[235].

4.4.3 Suffix Trees By k -mer Frequencies

The abundance of sequence noise (insertions, mismatches and similar, for example) often necessitates frequency-based analysis. Similar to the work of^[275] above, another method of applying frequency data have been extensively explored by Soares *et al.*^[280] to measure Euclidean distance between sequence data. When collecting frequency data, typically a window is opened at the beginning of the sequence and the frequencies are found for all encountered words. The authors depart from this method by presenting a new approach that determines a single optimal word length (k -mers) from which to generate a frequency distribution for application to suffix trees.

To collect these optimal k -mer frequencies, Soares *et al.* began by determining all words in the DNA alphabet (e.g., {A,C,G,T}) of length- k . An optimal resolution range of k -mers for the given set of genomes was described in^[275] and later applied to the work of Soares *et al.*, for instance, $k_{H \text{ max}} = \log_4(n_1)$ for a sequence of length- n_1 . In order to find a value applicable to all sequences under analysis, we choose n_1 as the length of the greater sequence and K as the smaller integer greater than $\log_4(m)$. For m sequences of different lengths, the peak value of word length (K) that is applicable to all sequences of the study is described by the following two equations:

$$n_1 = \max\{\text{length}(S_i), 1 \leq i \leq m\},$$

$$K = \lceil \log_4(n_1) \rceil$$

In the logarithmic equation, L is given the smallest integer not less than the calculated value.

The exhaustive lists of DNA L -words for n sequences were created by combinatorial means. For words of length- L , the size of the list can be described

mathematically: $t = 4^L$. The frequencies of these words are similarly found as in^[275]. Once amassed, they are added to an $n \times t$ matrix to create a global profile of all L -word frequencies of all input sequences. Next is the development of the genetic distance for the suffix trees. This pairwise standard Euclidean distance between pairs of sequences is calculated by the following:

$$SED(S_1, S_2) = \sqrt{\sum_{w \in t} (f_{S_1 w} - f_{S_2 w})^2} \quad (4.8)$$

for w , representing the k-mer and t representing the exhaustive list of words, respectively, and $f_{S_i w}$ represents the relative frequency of w in the sequence S_i . These values may be applied to suffix trees for convenient sequence analysis across large sets of sequences.

4.4.4 Composition Vectors Based On k -mer Frequencies

There are two main string composition vectors that we will discuss; the compositional vector (CV) and the complete composition vector (CCV). Discussed in^[113;237;314], a k -mer frequency CV for a genomic sequence is a distribution of frequencies of length- k motifs which are used for comparison across sequences. The CV method contains motif frequencies of the same length whereas the CCVs contain motif frequencies of unequal length.

The basic steps of creating CVs are the following; (1) find the frequencies of the motifs in a sequence, (2) create a vector by organizing the frequencies in some order, (3) compute the distance between every two composition vectors to form a distance matrix, and, optionally (4) construct the phylogenetic tree based on the differences. This last step is not essential but may be helpful when evaluating the degree of closeness between a set of sequences.

4.4.4.1 Creation Of Composition Vectors, (CVs, CCVs)

The creation of a CCV is very similar to that of a CV except that the input frequencies are made from strings of differing lengths. Despite its extra computational expense, the CCV method was found to provide finer evolutionary information than the CV method^[192].

Following the discussion from^[54;192], we define S_1 to be a sequence consisting of n_1 nucleotides and let $f(\alpha_1 \dots \alpha_k)$ to be the observed frequency of the length- k motif in S_1 . We define α_i for $1 \leq i \leq k$ to be a nucleotide such that $\alpha_i \in \{A, C, G, T\}$ for $1 \leq k < n_1$. Next, for some constant K , the largest string length we consider, we define $V_{S_1} = (f_1, f_2, \dots, f_{4^K})$ as the combined vector. Finally, we define $V_S = (S_1, S_2, \dots, S_K)$ as the combined vector. The vectors, V_{S_1} and V_S , reflect both random mutation and selection. Lu *et al.* noted that there is an underestimation of selective evolution for both these vectors when the data is normalized according to Equation 4.9, which is also discussed in^[42] and^[112]. For the observed frequency of $\alpha_1 \alpha_2 \dots \alpha_k$, the normalizing equation is described by the following:

$$a(\alpha_1 \dots \alpha_k) = \frac{f(\alpha_1 \dots \alpha_k) - f_e(\alpha_1 \dots \alpha_k)}{f_e(\alpha_1 \dots \alpha_k)} \quad (4.9)$$

where f_e represents the expected frequency and is defined by:

$$\begin{aligned} f_e(\alpha_1 \dots \alpha_k) &= \frac{f(\alpha_1 \dots \alpha_{k-1}) f(\alpha_2 \dots \alpha_k)}{f(\alpha_2 \dots \alpha_{k-1})} \\ &\times \frac{(n_1 - k + 1)(n_1 - k + 3)}{(n_1 - k + 2)^2}. \end{aligned}$$

for $k \geq 3$ and $f_e(\alpha_1 \dots \alpha_k)$. Lu *et al.*^[192] note that this normalization method underestimates the actual reality of the data. Next we describe their modified method which overcomes this problem.

4.4.4.2 Creation Of Improved CCV's

To overcome this setback, Lu *et al.*^[192] propose an improvement to CVs and CCVs. The improved complete composition vector (ICCV) is made assuming that the sequence bases all occur with equal probability, according to the expected frequency of a k -mer string. The variance of frequency of a given sequence S_1 of length- n_1 is also based upon this assumption. We first define the expected motif frequencies and their variance in the vectors. For any given k -mer, a position in the sequence is given as:

$$x_i = \begin{cases} 1, & \text{if the } k\text{-mer begins at position } i \\ 0, & \text{otherwise} \end{cases}$$

for integer i such that $1 \leq i \leq (n_1 - k + 1)$. This upper bound is the maximum observed frequency for a string $\alpha_1 \cdots \alpha_k$ in S_1 . Therefore it can be shown that;

$$f(\alpha_1 \cdots \alpha_k) = \sum_{i=1}^{n_1-k+1} x_i. \quad (4.10)$$

The expectation and variance of $f(\alpha_1 \cdots \alpha_k)$ are described in the following equations.

$$E[f(\alpha_1 \cdots \alpha_k)] = \sum_{i=1}^{n_1-k+1} E[x_i] = \frac{n_1 - k + 1}{4^k} \quad (4.11)$$

and the variance;

$$\begin{aligned} \text{Var}[f(\alpha_1 \cdots \alpha_k)] &= \frac{n_1 - k + 1}{4^k} \left(1 - \frac{1}{4^k}\right) \\ &\quad - \frac{2}{4^{2k}} (k-1)(n_1 - \frac{3}{2}k + 1) \\ &\quad + \frac{2}{4^k} \sum_{i=1}^{k-1} (n_1 - k + 1 - t) \frac{J_r}{4^t}, \end{aligned}$$

where J_r , is defined by the following.

$$J_r = \begin{cases} 1, & \text{if } (\alpha_1 \cdots \alpha_{k-r}) = (\alpha_{r+1} \cdots \alpha_k) \\ 0, & \text{otherwise} \end{cases}$$

for integer r such that $1 \leq r \leq k - 1$. See^[102] for a full derivation. One of the problems with the original CV and CCV concerns the denominator which requires a square root operation without which, Lu *et al.* warn of a problem of over-estimation. To mitigate the over-estimation problem, the authors apply the data's expectation and variance to the normalizing equation given below and complete the construction of the ICCV. The normalization of each observed frequency of a k -mer string, k_{norm} is given by the following equation:

$$k_{norm} = \frac{f(a_1 \cdots a_k) - E[(a_1 \cdots a_k)]}{\sqrt{\text{Var}[f(a_1 \cdots a_k)]}} \quad (4.12)$$

for $k \geq 1$.

4.4.4.3 Distance Measurement

We next discuss how the distance between vectors is measured. For two sequences, S_1 and S_2 , let their vectors be defined, $V_{S_1} = (\alpha_1, \alpha_2, \dots, \alpha_k)$ of length- k and $V_{S_2} = (\beta_1, \beta_2, \dots, \beta_k)$, also of length- k . We define the normalized distance between the vectors by the following:

$$D(V_{S_1}, V_{S_2}) = \frac{1 - C(V_{S_1}, V_{S_2})}{2}, \quad (4.13)$$

where $C(V_{S_1}, V_{S_2})$ is the cosine distance of the angle between V_{S_1} and V_{S_2} and is described by the following.

$$C(V_{S_1}, V_{S_2}) = \frac{\sum_{i=1}^k \alpha_i \cdot \beta_i}{\sqrt{\sum_{i=1}^k \alpha_i^2 \cdot \sum_{i=1}^k \beta_i^2}} \quad (4.14)$$

Lu *et al.* show that the ICCV method fixes the observed overestimation problems with the previous method, and generates more accurate and robust results. They also show that its results are consistent with methods based on alignment by dynamic programming in phylogeny.

4.4.5 A Revised String Composition Method

Chan *et al.*^[54] revisit the composition vector method and apply an analysis of entropy from information theory and operations research. Their method begins by finding the frequencies of each base of a k -string sequence. For example, from ACTGCTATGC, the base frequencies are the following: $f(A) = \frac{1}{5}$, $f(C) = \frac{3}{10}$, $f(G) = \frac{1}{5}$, $f(T) = \frac{3}{10}$.

The second step is to estimate the expected frequency $q(u)$ for each k -string. For this step, the authors suggested determining the relationship between $q(\cdot)$ and $f(\cdot)$ by maximizing the following system of equations from Hua *et al.*^[128]. Here, the entropy in $q(\cdot)$ is maximized given the frequency $f(v)$ for all $(k - 1)$ -strings v .

$$\begin{cases} q(vA) + q(vC) + q(vG) + q(vT) = f(v), \\ q(Av) + q(Cv) + q(Gv) + q(Tv) = f(v) \end{cases} \quad (4.15)$$

Chan *et al.* depart from the work of Hua *et al.* by making no assumptions between $q(\cdot)$ and $f(\cdot)$. Instead, they maximized the following equations which estimate the expected frequencies $q(u)$.

$$q(LwR) = \frac{f(Lw)f(wR)}{f(w)} \quad (4.16)$$

for $k \geq 3$, which was introduced by, Qi *et al.*^[237] and,

$$q(LwR) = \frac{f(L)f(wR) + f(Lw)(R)}{2} \quad (4.17)$$

for $k \geq 2$, from Yu *et al.*^[322]. For any k -string u , L and R represent the left and right nucleotides of the word and w represents the middle $(k - 2)$ -string located between them. In the second equation, all these elements are assumed to occur independently. From these equations, the authors created a new system of equations (below) to solve where the right-hand side concerns sequence frequencies and the left-hand side concerns the estimations:

$$\left\{ \begin{array}{l} q(vA) + q(vC) + q(vG) + q(vT) \\ = \frac{f(Lw)}{f(w)} [f(wA) + f(wC) + f(wG) + f(wT)] \\ q(Av) + q(Cv) + q(Gv) + q(Tv) \\ = \frac{f(xR)}{f(x)} [f(Ax) + f(Cx) + f(Gx) + f(Tx)]. \end{array} \right. \quad (4.18)$$

When this system is maximized, Chan *et al.* note that the system generates a set of all possible estimation formulas $q(\cdot)$ from which, one can be selected to maximize the entropy. In general, from any existing estimation formula $q(\cdot)$ given in terms of $f(\cdot)$, the authors note that the set of constraints such as the following can be derived:

$$\left\{ \begin{array}{l} q(vA) + q(vC) + q(vG) + q(vT) = l(v), \\ q(Av) + q(Cv) + q(Gv) + q(Tv) = r(v) \end{array} \right. \quad (4.19)$$

where the left and right side frequency values, $l(v)$ and $r(v)$ are derived from frequency information ($f(v)$) for each length- $(m - 1)$ motif v . To obtain the unique $q(u)$ for all u , the following optimization problem is solved:

$$\begin{aligned}
& \text{maximize: } - \sum_{i=1}^{4^k} q_i \log q_i \\
& \text{subject to: } \begin{cases} q_i \text{ satisfies the system of equations} \\ q_i \geq 0 \text{ for all } i \end{cases}
\end{aligned}$$

where $-q_i \log q_i$ is Shannon's entropy calculation. The authors apply this information to phylogenetic tree analysis in a similar fashion as we saw in Lu *et al.* [192].

Maximum Entropy Principle (MEP) After solving the problem above, a system of noise estimation formulas is provided. Note: a motif appears as the following: $(\alpha_1 \cdots \alpha_m \alpha_n \cdots \alpha_k)$ and can be split into to sub words.

$$q^{MEP}(\alpha_1 \cdots \alpha_m \alpha_n \cdots \alpha_k) = \frac{l(\alpha_1 \cdots \alpha_m) r(\alpha_n \cdots \alpha_k)}{\sigma}, \quad (4.20)$$

where, q^{MEP} is the maximized entropy principle score for the sequence data and,

$$\sigma = \sum_{L \in \{A, C, G, T\}} l(\alpha_1 \cdots \alpha_m) = \sum_{R \in \{A, C, G, T\}} r(\alpha_n \cdots \alpha_k). \quad (4.21)$$

We note that $q^{MEP} = 0$ if $\sigma = 0$ and that $l(\cdot)$ and $r(\cdot)$ are parametric functions. Different $l(\cdot)$ and $r(\cdot)$ will give different estimation formulas and will have varying levels of success. The authors applied this test to create phylogenetic trees from simulated data sets. Their results showed differentiation of "closely related" sequences.

4.4.6 D_2 Statistic

The statistic D_2 , is the number of approximate word matches of length k between sequences $S_1 = (\alpha_1, \dots, \alpha_k)$ and $S_2 = (\beta_1, \dots, \beta_k)$, with α_i and β_j belonging to an alphabet \mathcal{A} (in this case, the DNA bases), which is distributed according to a letter

distribution parameterized by $\eta^{[91]}$. This statistic is applied to two populations of differing means, but identical dispersion matrices^[307], to determine distance. Recently, the statistic has evolved to provide more exact approximations by asymptotic regimes for uniform and non-uniform distributions^[90;187]. Mathematically, the D_2 statistic is defined by the following. From^[242], given sequences $S_1 = (\alpha_1, \dots, \alpha_{n_1})$ of length- n_1 and $S_2 = (\beta_1, \dots, \beta_{n_2})$ of length- n_2 and $W = \{w_1, \dots, w_k\} \in \mathcal{A}^k$, then D_2 is defined by the following:

$$D_2 = \sum_{W \in \mathcal{A}^k} C_{s_1}(W)C_{s_2}(W) \quad (4.22)$$

where $C_{s_i}(W)$ is the number of occurrences of W in sequence S_i .

The $D2Z$ statistic^[143] was developed to compare gene regulatory sequences and offered an improvement in performance to D_2 , but could still fail due to noise complications^[242;302]. To combat this problem of noise, Reinert *et al.*^[242] propose a new statistic D_2^S , which is a self-standardized D_2 .

$$D_2^S = \sum_{W \in \mathcal{A}^k} \frac{\tilde{C}_{s_1}(W) \tilde{C}_{s_2}(W)}{\sqrt{\tilde{C}_{s_1}(W)^2 + \tilde{C}_{s_2}(W)^2}} \quad (4.23)$$

For $p_W = \prod_{i=1}^t p_{w_i}$, the probability of occurrence of w_i for $1 \leq i \leq k$ and $\tilde{n}_i = n_i - k + 1$ for i sequences, the centralized count variables, $\tilde{C}_{s_1}(W)$ and $\tilde{C}_{s_2}(W)$, are therefore denoted by the following.

$$\tilde{C}_{s_1}(W) = C_{s_1}(W) - \tilde{n}_1 p_W \text{ and } \tilde{C}_{s_2}(W) = C_{s_2}(W) - \tilde{n}_2 p_W$$

Reinert *et al.* also proposed a second statistic, D_2^* , which we shall presently define. To introduce this statistic, we replace $p(a)$, the unobserved feature probabilities, by $\tilde{p}(a)$ (the observed) for the relative count of letter a in the concatenation of the two sequences that are based on the assumption that the two sequences are independent.

We note that these sequences are both independent and contain identically distributed (i.i.d.) bases. The estimated probability of occurrence of $W = \{w_1, \dots, w_k\}$ is obtained by $\tilde{p}_W = \prod_{i=1}^k \tilde{p}_{w_i}$. We now define D_2^* by the following.

$$D_2^* = \sum_{W \in \mathcal{A}^k} \frac{\tilde{C}_{s_1}(W) \tilde{C}_{s_2}(W)}{\sqrt{n_1 n_2} \tilde{p}_W} \quad (4.24)$$

The authors found that the D_2^* statistic out-performed both the D_2 and D_2^S statistics in terms of accurate detection of relatedness between two sequences. The statistical power of both D_2^* and D_2^S increases with sequence length and tends to 1 as the sequence length tends to infinity under a common motif model. When applied to organizing sequence reads of Next Generation sequence assembly tasks, and to phylogeny tasks, the D_2^S statistic provided a powerful alignment-free comparison tool^[282]. However, when studying phenomena in the pattern transfer model such as horizontal gene transfer, the power of these statistics declines and converges to a limit that is generally less than 1 as the sequence length tends to infinity. The primary reason for this limitation is that the means of the word counts in these statistics eventually become increasingly similar to each other. This resemblance works to desensitize the detection of patterns between the sequences.

To improve the detection of relationships across sequences using alignment-free methods in the pattern transfer model, Liu *et al.*^[189] developed new statistics (T^* , T^S and T_{sum}^* , described below) which they claim have a better statistical power. The authors present them with simulations to demonstrate their power and to show that they are more appropriate for applications where long sequence-lengths are a concern.

Based on approximating the mean by a sample mean, the approach of the new statistic is to partition a long sequence of length- n_1 into consecutive, non-overlapping (discrete) subintervals of length- r , $d_{sub} = \lfloor \frac{n_1}{r} \rfloor$. Then the D_2^* and D_2^S values are calculated over each i^{th} subinterval for word counts w and are denoted $D(i)_2^*$ and $D(i)_2^S$, respectively. For, two sequences of length n_1 where, $S_1 = \{\alpha_1, \dots, \alpha_k\}$ and

$S_2 = \{\beta_1, \dots, \beta_k\}$, these statistics are defined by the following equations.

$$T^* = \sum_{i=1}^{d_{sub}} D_2^*(i) \quad (4.25)$$

and

$$T^S = \sum_{i=1}^{d_{sub}} D_2^S(i) \quad (4.26)$$

The final statistic from^[189] is drawn over two sequences S_1 and S_2 of lengths n_1 and n_2 , respectively, to conclude the degree of relatedness.

$$T_{sum}^* = \sum_{i=1}^{n_1-k+1} S_{1i}^* + \sum_{i=1}^{n_2-k+1} S_{2i}^* \quad (4.27)$$

for,

$$S_{1i}^* = \max_{\{1 \leq j \leq n_1-k+1\}} M^*[i, j, k] \quad (4.28)$$

and

$$S_{2i}^* = \max_{\{1 \leq i \leq n_1-k+1\}} M^*[i, j, k] \quad (4.29)$$

where,

$$M^*[i, j, k] = D_2^*(S_1[i, i+k-1], S_2[j, j+k-1]) \quad (4.30)$$

While D_2^* and D_2^S are generally more powerful statistics than T_{sum}^* and T_{sum}^S for the common motif model, this is not the case for studies concerning the pattern transfer model. For this reason, the statistics presented by Liu *et al.* are desirable in pattern transfer model applications when the sequence data is very long.

4.5 Data Compression And Dictionaries

Alignment-free methods, involving data compression and dictionaries, are based on the idea that the more similar two sequences are to each other, then the better one sequence can be created from the parts of another. Inspired by Lempel-Ziv(LZ)

compression technologies^[331], we offer an example of sequence comparison, from Otu *et al.*^[215].

For the sequences S_1 , S_2 and S_Q , we define $H_E(S_1)$, $H_E(S_2)$ and $H_E(S_Q)$ to be the exhaustive sets of all words found using an approach from LZ-compression. We then analyze the sets of *sequence histories* to determine how much of one sequence can be built out of the sequence histories of another. We define $c_H(\cdot)$ to be the number of components in a history of a sequence S and $c_{min}(\{c_H(S)\})$ over all histories of S .

For S_1 and S_2 , we have $c_{min}(S_1 S_Q) \leq c_{min}(S_1) + c_{min}(S_Q)$, by the sub-additivity of the LZ-complexity. To compute the closest similarity of S_1 and S_Q , $d(S_1, S_Q)$, and S_2 to S_Q , $d(S_2, S_Q)$, we take the smallest value of $\max\{c_{min}(S_1 S_Q) - c_{min}(S_1), c_{min}(S_Q S_1) - c_{min}(S_Q)\}$ and $\max\{c_{min}(S_2 S_Q) - c_{min}(S_2), c_{min}(S_Q S_2) - c_{min}(S_Q)\}$, respectively.

Compare the sequence similarity of S_1 to S_Q and S_2 to S_Q . We first find the sequence histories to compare distances. We introduce an example to demonstrate how this is performed.

$S_1 =$	ATGGC
$S_2 =$	ACGGT
$S_Q =$	ATGGC

- $S_1 = \text{ATGGC}$
 - $H_E(S_1) = \text{A, T, G, GC}$
 - $c_{min}(S_1) = 4$

- $S_2 = \text{ACGGT}$
 - $H_E(S_2) = \text{A, C, G, GT}$
 - $c_{min}(S_2) = 4$

- $S_Q = \text{ATGGC}$

- $H_E(S_Q) = \text{A, T, G, GC}$
- $c_{min}(S_Q) = 4$
- $S_1 S_Q = \text{ATGGCATGGC}$
 - $\text{A, T, G, GC, ATGGC}$
 - $c_{min}(X S_Q) = c_{min}(S_Q S_1) = 5$
- $S_2 S_Q = \text{ACGGTATGGC}$
 - $\text{A, C, G, GT, AT, GGC}$
 - $c_{min}(S_2 S_Q) = c_{min}(S_Q S_2) = 6$
- $d(S_1, S_Q) = \max \{c_{min}(S_1 S_Q) - c_{min}(S_1), c_{min}(Q S_1) - c_{min}(S_Q)\} = 1$
- $d(S_2, S_Q) = \max \{c_{min}(S_2 S_Q) - c_{min}(S_2), c_{min}(S_Q S_2) - c_{min}(S_Q)\} = 2$

By the author's method, we conclude that S_1 and S_Q are more similar since $1 = d(S_1, S_Q) < d(S_2, S_Q) = 2$. The authors used this method to populate phylogenetic trees from simulated sequences to show clusterings of "related" sequences.

4.5.1 Text Compression Algorithms

Data compression is nearly out of the scope of this contribution, however, they are worth mentioning because they also provide an alignment-free approach to comparing sequence data. These general purpose compression algorithms may be based on the Ziv and Lempel^[331] methods (as seen above). Recent advances have been developed in^[66;203] and^[158]. Cao *et al.*^[48] proposed a memory-based algorithm called *expert model* (XM) to compress DNA by applying statistical information, gained from previous encounters of a particular symbol.

4.5.2 Average Common Substring (ACS)

Ulitsky *et al.*^[293] built on information theoretic tools such as, Kullback-Leibler relative entropy, to find a distance between entire genomes, even if their lengths vary. The ACS measure that they proposed is based on computing the average lengths of maximum common substrings. They used these average lengths between the sequences to construct phylogenetic trees from an efficient algorithm.

Let S_1 and S_2 be sequences, of lengths n_1 and n_2 where, $S_1 = (\alpha_1, \dots, \alpha_{n_1})$ and $S_2 = (\beta_1, \dots, \beta_{n_2})$. For any position i , let $r(i)$ be the length of longest substring in S_1 that *exactly matches* a substring in S_2 starting at some position j . These lengths $r(i)$ are averaged to get a measure, $L(S_1, S_2) = \sum_{i=1}^n r(i)/n_1$. Since $L(S_1, S_2)$ represents a common sequence found in both sequences, then the longer it is, the more similar the sequences are to each other. This value is only a *similarity* measure and must still be converted to a distance value. The inverse is taken to get the distance and then a “correction term” is subtracted to ensure that the distance $d(S_1, S_1) = 0$ (will always be zero). This allows for, $d(S_1, S_2) = \frac{\log n_2}{L(S_1, S_2)} - \frac{\log n_1}{L(S_1, S_1)}$ where $L(S_1, S_1) = \frac{n_1}{2}$ to provide the correctional term, $2 \cdot \frac{\log(n_1)}{n_1}$ which converges to 0 as $n_1 \rightarrow \text{infty}$. Since the measure, $d(S_1, S_2)$ is not symmetric, the authors compute the final ACS measurement between the two strings, $d_s(S_1, S_2)$ by the following.

$$d_s(S_1, S_2) = d_s(S_2, S_1) = \frac{d(S_1, S_2) + d(S_2, S_1)}{2} \quad (4.31)$$

We now show how to apply this method to determine the distance between two sequences. Let $S_1 = \text{ACGTGCTATG}$ and $S_2 = \text{ACGCGCTA}$, of lengths $n_1 = 10$ and $n_2 = 8$, the method finds all common substrings as shown in Table 4.2:

Table 4.2: The similar and different chunks, taken in order from each sequence.

Sequence	Same	Different	Same	Different
S_1	ACG	T	GCTA	TG
S_2	ACG	C	GCTA	

$$\begin{aligned} L(S_1, S_2) &= \frac{(1+2+3)+(1)+(1+2+3+4)+(1)+(1)}{10} \\ &= \frac{19}{10} = 1.9 \end{aligned}$$

and

$$\begin{aligned} L(S_1, S_1) &= \frac{1+2+3+4+5+6+7+8+9+10}{10} \\ &= \frac{55}{10} = 5.5 \end{aligned}$$

Then the distance between the two sequences is;

$$d(S_1, S_2) = \frac{\log 8}{1.9} - \frac{\log 10}{5.5} = 0.293$$

Similarly, we can calculate $D(B, A)$ as follows:

$$\begin{aligned} L(S_2, S_1) &= \frac{(1+2+3)+(1)+(1+2+3+4)+(1)+(1)}{8} \\ &= \frac{19}{8} = 2.375, \\ L(S_2, S_2) &= \frac{1+2+3+4+5+6+7+8}{8} \\ &= \frac{36}{8} = 4.5 \end{aligned}$$

and,

$$d(S_2, S_1) = \frac{\log 10}{2.375} - \frac{\log 8}{4.5} = 0.220.$$

For our example above, where $d_s(S_1, S_2)$ is not symmetric, the symmetric distance is: $d_s(S_1, S_2) = d_s(S_2, S_1) = \frac{d(S_1, S_2) + d(S_2, S_1)}{2} = \frac{0.293 + 0.220}{2} = 0.257$. This value, can be used as a weight for a sequence in a phylogenetic tree to show relations between sequences of a set.

4.6 Applications Of Alignment-Free Methods

4.6.1 Biological Data And Sequence Assembly

In genetic sequence assembly work, alignment technologies are very important for determining the adjacency of reads (or contigs which are partially combined reads) to reconstruct the original sequence. During a typical *de novo* assembly task, a sequencing machine may split the genome into many millions (trillions) of reads that must be reassembled like from a jigsaw puzzle. This reconstruction task is computationally intensive since each piece must be compared with every other piece in the pool to determine adjacency. This task is frustrated when there are foreign reads of other sequences to be assembled in the same data pool. The extra sequence data serves to massively broaden the search space when determining the adjacency of a read since there are many more comparison operations to perform. To reduce the workload of the assembly project, it is therefore desirable to place all related reads into a unique groups (*bins*) and apply the main assembly algorithms to each organism separately.

A novel approach, requiring no database support, was introduced by^[28;30] to order the organisms in the pool into separate bins. The authors' method creates CVs from restriction sites^[29] to determine inter-sequence relatedness, and place the sequences from the mixed pool into separate groups. This type of proposed alignment is for a global analysis as it is able to process and compare sequences in a pool of arbitrarily size. They applied their work to the sequence assembly reads and

contigs of *Bifidobacterium longum*, *Mycobacterium bovis*, *Clostridium tetani*, *Staphylococcus aureus*, *Burkholderia pseudomallei* and *Campylobacter jejuni*. Based on the similarity of proportional values contained in the CVs, the authors were able to differentiate the sequence material by organism.

The method uses *spectrum sets* which are lists of motifs made up of permutations of restriction enzymes which are specific and unique sites in DNA where enzymes are able to cleave. To create a spectrum set from the bacterial restriction site, **GAATTC**, we observe that the motif contains, two A's, two T's, one C and one G. A spectrum set contains all motifs which have exactly the same number of each base. For example, for the bacterial restriction site, **GAATTC**, there are (156 motifs in the spectrum set which have the same base composition. A vector of length-156 is constructed from the proportions of each of these motifs which are contained in the sequence data. For example, to populate the vector V_{S_1} of the motif proportions of w_i for $i = \{1, \dots, 156\}$ for sequence S_1 of length- n_1 , the following equation is employed,

$$V_{S_1} = \frac{c(w_i) * |w_i|}{n_1}, \quad (4.32)$$

where $c(\cdot)$ represents the number of occurrences of the motif in the sequence. This equation serves to normalize the proportions so that the values can be compared across diverse data sets. The authors noted that similar sequence data gave rise to similar vectors which they used to organize the sequence data.

4.6.2 Chromosomal Data And Phylogeny

In addition, in^[30], it was shown that the method could also be applied to create phylogenetic trees which were extremely similar to trees created by NCBI's taxonomy tree making software. In this work, they used chromosomal sequences of arbitrarily chosen organisms (*Caenorhabditis elegans*, *Canis lupus familiaris*,

Drosophila melanogaster, *Mus musculus*, *Mycoplasma hyorhinis*, *Oryctolagus cuniculus* and *Rattus norvegicus*) and built a tree which replicated that of NCBI's taxonomy analysis software (available at <http://www.ncbi.nlm.nih.gov/guide/taxonomy/>).

4.6.3 Horizontal Gene Transfer

Horizontal Gene Transfer (HGT) is the phenomenon where genetic material is shared between unrelated organisms. Evolutionary^[286] and Phylogenetic studies^[204] have observed common material between unrelated bacterial organisms which suggests a parallel evolutionary history. The discovery of similar regions of DNA between two enormous genomes is not a trivial task and so alignment-free methods have proven to be helpful in this field. In^[76], the authors present *Alignment-Free Local Homology* (*alfy*), a method to determine HGT by an alignment-free approach. Since determining evolutionary distances from word frequency data is a non-trivial task, the authors report that their method is conveniently able to make this determination.

We cite and discuss the method and example presented in^[76] where the query sequence, denoted as S_Q of Table 4.3, is compared to the subject sequence, S_1 . For each position in the query S_Q , the *alfy* algorithm determines the shortest substring that starts in query which is absent from the subject sequence.

In Tables 4.4 and 4.5, this comparison task is shown by a string of numbers (match scores) which show the length of the substring starting in (S_Q) that are absent in (S_1). If the consecutive intervals created by these matching scores are wide (e.g., long strings of uninterrupted consecutive integers), then the sequences are closely related (similar), however, if the intervals are generally short, then the sequences are not closely related (dissimilar).

We cite an example from another study concerning HGT by the same authors^[76]. This method is applied to locating regions of common genetic material in *E. coli* and

Table 4.3: The sequences to compare by the *alfy* method. To compare sequences, we find the shortest sequence in the query (S_Q) which is absent from a subject.

Query	$(S_Q) = \text{CGCGATTACT\$}$
Subject	$(S_1) = \text{CGCCCGGACT\$}$
Subject	$(S_2) = \text{TGAGATTCAAG\$}$

Table 4.4: S_1 is compared to S_Q to determine the shortest substring in S_Q which is absent from S_1 . The matching numbers indicate the shortest unique substring starting at this position that is absent from the subject.

Subject	(S_1)	$\text{CGCCCGGACT\$}$
Query	(S_Q)	$\text{CGCCCTGACT\$}$
Matching Score		6543325432

Table 4.5: Sequences S_1 and S_2 are compared to S_Q . The matching numbers indicate the shortest unique substring starting at this position that is absent from the subject. The HGT is described by a string of S_1 and S_2 characters to indicate where the subsequences likely originated.

Subject	(S_1)	$\text{CGCCCGGACT\$}$
Subject	(S_2)	$\text{TGAGATTCAAG\$}$
Query	(S_Q)	$\text{CGCCCTGACT\$}$
Matching Score		4325432432
Implied HGT	(S_1) and (S_2)	$bbbccccbbb$

Table 4.6: We wish to determine the sequence relations based on common sequence material. The query sequence is S_Q and subjects are S_1 through S_3 .

	1	2	3	4	5
S_Q	T	A	G	C	\$
S_1	G	A	\$		
S_2	G	C	C	\$	
S_3	T	A	\$		

recombinant HIV-1 strains. This method is similar because it locates local regions in subject sequences which are closely related to the query sequence. In Table 4.6,

sequences S_1 to S_3 are the subjects and S_Q is the query. We find which parts of S_Q most closely resemble the subject sequences. The sections of sequence material are written in an interval notation: $S_Q[1, 2] = \text{TA}$ matches $S_3[1, 2]$, $S_Q[3, 4] = \text{GC}$ matches $S_2[1, 2]$. By this system, we claim that $S_Q[1, 2]$ is most closely related to S_3 and $S_Q[3, 4]$ is most closely related to S_2 .

During the sequence comparison task of query-to-subject, in^[75], the authors denote the length of the *shortest* query sequence prefix by $h_{i,p}$. The query suffix, $Q[p, |Q|]$ denotes the sequence starting at position p , which is absent in subject sequences. The length of the *longest* subsequence starting at $Q[p]$ taken over all subject sequences is denoted, $H_p = \max_{\{1 \leq i \leq n\}} h_{i,p}$, where h_p is bounded by the query length: $H_p \leq |Q| - p + 1$. For example, in Table 4.6, $H_{1,1} = \text{T} = 1$; $H_{2,1} = \text{T} = 1$; $H_{3,1} = \text{TAG} = 3$; and $H_1 = \max_{\{1, 1, 3\}} = 3$. Conversely, the longest subject subsequences which start at $Q[p]$ are found in a subject sequence, are denoted by $S_p = \{S_i \in S | h_{i,p} = H_p\}$. Based on these properties, the authors note that the longest sequence from Table 4.6 is S_3 (the most similar subject to sequence S_Q).

4.7 Advantages And Disadvantages Of Methods

The method that an algorithm uses to gain its statistical data for an analysis is an important part of the whole operation. A fault at this stage would travel throughout the comparison task and upset the conclusion. In this section, we describe the generation of the motif frequency distributions and we discuss how this initial statistical work may not always be appropriate for a particular data set.

The methods of Section 4.4 (Factor Frequencies) are powerful methods to employ in sequence comparison tasks since they do not concern the location of the motifs they analyze. Their algorithms are efficient since they are generally of a linear complexity.

They contrast to the general high complexity of the algorithms that are based on dynamic programming. The results of factor frequency methods are adaptable and can be conveniently applied to an analysis by mutual information (Section 4.4.2.1), k -mers (Section 4.4.3) or by compositional vectors (Sections 4.4.4 and 4.4.5).

As the factor frequency methods are generated by word occurrences in a sequence, it is important to choose words which are not likely to commonly appear in a sequence. As a general rule in DNA, the shorter the word, then the more likely it will appear randomly in a sequence. In Sections 4.4.1 and 4.4.2.1, vectors were created out of pairs of DNA bases. While this may be a simple way to illustrate the concept, frequencies made up of these short pairs have less meaning than frequencies made up by longer words because any particular DNA pair has a probability of $\frac{1}{4^2} = \frac{1}{16}$ to occur randomly. We note that for sequences which are largely dissimilar, then shorter words (hence shorter compositional vectors) should be used to create the feature frequency distributions. However, longer words, (hence, longer compositional vectors, assuming an exhaustive list of motifs) may be used when the sequences are known to be similar, such as when they are related,^[313]. The methods of Section 4.4 are well suited for this application using both long and short motifs. They also function well when the location of the motif in the sequence is not important, as in the case of synteny.

Unlike the approaches of Section 4.4 where the frequency distributions were generated by user-specified motifs, the methods of Section 4.5 ‘choose’ their own sizes of words for their sequence comparison task. In the methods proposed by^[215] (LZ compression based) and^[76] (horizontal gene transfer), the word size is not a parameter set by the researcher. These kinds of algorithms are useful to comparison tasks where it is not clear about the “correct” kinds of motifs to employ. In the case of factor frequency methods, when designing a list of motifs from which to generate a frequency distribution, an exhaustive list is likely used. As we have previously mentioned, the longer the motif, then the larger the exhaustive list. Finding the

frequencies of these extra motifs may add additional computational time to the task. Therefore, compression based methods may be more suitable to comparisons where longer motifs are desirable, such as when the sequences are similar. This might be because the words of varying size and composition will be more similar across related sequences.

4.8 Conclusion

Table 4.7: Summary of the discussed methods in this contribution. The column “Alignment” contains the best suggested use of the method.

Sec.	Method	Author	Alignment	Citation
4.4.1	Base Base Correlations	Lui <i>et al.</i>	Global	[190]
4.4.2	Oligonucleotide profiling	Arnau <i>et al.</i>	Local	[10]
4.4.2	Feature Frequency	Sims <i>et al.</i>	Local	[275]
4.4.3	k -mers Frequencies	Soars <i>et al.</i>	Global	[280]
4.4.4	Composition Vectors	Lu <i>et al.</i>	Global	[192]
4.4.5	Composition Vectors	Chan <i>et al.</i>	Global	[54]
4.4.6	D_2^* Statistic	Reinert <i>et al.</i>	Global	[242]
4.4.6	Improved D_2 Statistics	Lui <i>et al.</i>	Global	[189]
4.5	Sequence distance	Otu <i>et al.</i>	Global	[215]
4.5.1	DNA Compression	Cao <i>et al.</i>	Global	[48]
4.5.2	Avg Common Substring	Ulitsky <i>et al.</i>	Global	[293]
4.6.3	ALign. Free local homologY	Domazet-Lošo <i>et al.</i>	Local	[75;76]
4.6.1	Sequence Assembly	Bonham-Carter <i>et al.</i>	Global	[28]
4.6.2	Phylogeny	Bonham-Carter <i>et al.</i>	Global	[30]

Comparison of sequence data represents a large problem in computational biology research. Discovery is often frustrated by obstacles such as synteny or other forms of genetic recombination, preventing methods of dynamic programming from working effectively. We provide a summary of the methods that we have discussed in Table 4.7, listed by sections, references and authors. When confronted with a large number of comparison tasks, which are unsuitable for traditional forms of

alignment from dynamic programming, these alignment-free methods may be the only feasible approach for completing the tasks to permit discovery. This is because the alignment-free methods do not function based on the location of genes or regions in each sequence. When the location of these regions is not important for the analysis, alignment-free methods like the ones included in this work may accomplish the goal of comparing genetic sequences.

Since there is more sequence data available today than ever before, there are many more projects that depend on sequence comparison. In order for discovery to be made, this work will have to be done by other technologies such as those based on dynamic programming which have obvious limitations. Alignment-free methods generally require less computational resources and use algorithms that are typically of linear complexity. These incorporated elements are appropriate for advancing comparative bioinformatics research.

It is our hope that this review provides useful information for researchers who are studying alignment-free methods and are using them in the analysis of genomic sequences and metagenomes. Since the mathematical aspects of the above tools are themselves an obstacle, it is also our hope that this review helps to introduce the reader to some of the more complicated calculations that are associated with these alignment-free tools for discovery. Furthermore, we envisage that this contribution of the thesis will serve as a useful reference in identifying open problems and driving future research in sequence comparison.

4.9 Article Details

This contribution was published in *Briefings in Bioinformatics*, 2013.

- Bonham-Carter, Oliver, Joe Steele, and Dhundy Bastola. “Alignment-free genetic sequence comparisons: a review of recent approaches by word

analysis.” *Briefings in bioinformatics* (2013): bbt052.

Words have no power to impress
the mind without the exquisite
horror of their reality.

Chapter 5

Edgar Allan Poe

An Analysis Of Palindromic Content In The Coding And Non-coding DNA Regions Of Bacteria

5.1 Abstract

DNA palindromes, the reversed and complemented genetic words, are read the same in the 3' to 5' as the 5' to 3' direction, and can form a unique restriction sites (RSs) where enzymes are able to cut DNA. Several studies have confirmed that short palindromes, behaving as active RSs, are few when compared to statistically expected values in bacterial genomes. These studies suggest that palindromes bring potential instability to intolerant coding regions of the genomes which appears to alter their

concentrations. While this palindrome-avoidance phenomenon has been observed in bacteria, the exact location in the genome where palindromes are most rare has not been investigated. In this contribution of the thesis, we provide evidence to suggest where the palindromic content is the least by comparing the content in *coding* and *non-coding* regions of bacterial DNA. We study the exhaustive lists of palindromes (lengths 4, 6, 8, and 10) to conclude that at least half of the motifs of each set (and sometimes, nearly all of the motifs of a set) show similar trends of reduced presence in the coding regions, when compared to the non-coding regions of bacteria.

5.2 Introduction

A DNA palindrome (here called a *palindrome*) is a word which is equivalent to itself when in its reversed and base-complemented form. Palindromes have been shown to be key actors in bacterial auto-immune defense systems as they often form the restriction sites for type II restriction endonucleases; highly specific restriction enzymes which cleave the DNA at these sites^[49;218;226].

In palindromic avoidance studies across several bacterial groups, Koonin *et. al.*^[101] found type II restriction-modification enzymes tended to be under-represented when compared to expected levels. Since it is conceivable that natural restriction sites can fail to be methylated (and are unprotected from enzymes) on occasion, the authors explain that avoidance is likely an evolved damage-control system.

5.3 Methods

The data for this study was drawn from common bacterial chromosomal DNA which was downloaded from Genbank^[23]. We developed a software tool written in Python, employing Biopython version 1.58 that, for each sequence, calculates the GC-content, isolates the coding and non-coding regions of sequence material from the inputted

genomes, and determines an exhaustive list of palindromes which are then parsed in each preprocessed region of the input sequences. Finally, the results are organized for Mann-Whitney, non-parametric statistical tests (discussed later) to determine the final motif distributions of the genomes.

To obtain the coding and non-coding datasets, the protein-coding segments of each genome were found based on the CDS features given in the organism's Genbank record. All the segments which were associated with CDS regions were joined together to create a unified and continuous string for each organism. We secured all the non-coding material for each organism in a similar way; starting with a complete genome, we removed all the CDS regions. The remaining code was the non-coding material for the organism.

Our genera data was divided into two groups based on GC-content of the genomes. The GC rich group was made up of sequences with more than 60% GC content. The genera in this group are; *Bifidobacterium*, *Burkholderia*, *Caulobacter*, *Desulfovibrio*, *Geobacter*, *Xanthomonas*. The other group, GC-poor, contained the following; *Agrobacterium*, *Bifidobacterium*, *Brucella*, *Chloroflexus*, *Corynebacterium*, *Erwinia*, *Geobacter*, *Pantoea*.

The palindromes for our study were prepared by first creating an exhaustive list of all possible DNA words of lengths- $\{4, 6, 8, 10\}$. The complement of a base is one which is found on the opposite strand in the helix (i.e. $A \Leftrightarrow T, C \Leftrightarrow G$). Each word w in the list was tested for palindromy by determining whether $w == reversed[complemented(w)]$ was true. By the nature of this function, only even palindromes, where $\text{length}(w) \bmod 2 \equiv 0$, are considered in this study. There are $4 * 4 * 1 * 1 = 16$ possible palindromic words of length-4. Expressed mathematically, the number of possible palindromes of a length L_p is; $n^{\frac{L_p}{2}}$, where n is the size of the alphabet.

Across each organism's coding and non-coding material, we determined the

proportion of sequence code made up by each palindrome. We use proportions, not frequencies, in our study of palindromic content because proportions are naturally normalized and facilitate comparison of content between regions. For these readings, there is no overlap between palindromes in the sequences and we do not consider nested palindromes. The proportion is given the following equation; $S_L = \frac{\text{count}(m_i)*|m_i|}{|S_L|}$, where m_i is a motif, S_L is the sequence space, $\text{count}(m_i)$ is number occurrences of m_i in S_L , $|m_i|$ and $|S_L|$ are the lengths of the motif and the sequence respectively. This equation determines how much of the coding or non-coding sequence is actually composed from the current motif by finding the number of occurrences. This value is divided the length of each region. The higher the value of the proportion, the greater the content of the motif in the region.

Null Hypothesis 1. *The proportions of palindromic motifs of length 4 are the same between the coding and non-coding regions of all evaluated sequence material.*

Null Hypothesis 2. *The proportions of palindromic motifs of length 6 are the same between the coding and non-coding regions of all evaluated sequence material.*

Null Hypothesis 3. *The proportions of palindromic motifs of length 8 are the same between the coding and non-coding regions of all evaluated sequence material.*

Null Hypothesis 4. *The proportions of palindromic motifs of length 10 are the same between the coding and non-coding regions of all evaluated sequence material.*

The Mann-Whitney, non-parametric test, was selected to determine which of the two regions had more content for each palindromic. This test was appropriate for our data since there is no requirement of a normal distribution of the data. We test the palindromic motifs of lengths {4, 6, 8, 10} by the Hypotheses: [1](#), [2](#), [3](#) and [4](#).

The significant value from the outcome of these tests indicate that the alternative hypothesis was satisfied for the particular palindromic motif set under evaluation (i.e. a higher proportion of the palindrome in the non-coding region than the coding

region by evidence in all evaluated genomes). The approach that we chose may be conveniently reproduced. The other methods capable of performing a similar study, such as those involving dynamic programming (i.e. methods from global alignment and similar) are more computationally expensive. For this reason, we opted to use an efficient statistical approach which we discuss below.

5.4 Results And Discussion

We used the Mann-Whitney tests to determine which palindromes of the lengths- $\{4, 6, 8, 10\}$ had significantly greater concentrations in the non-coding regions than in the coding regions. Our working alternative hypothesis (that there is more short palindromic content in the non-coding regions than in the coding regions) was concluded by the observation that long palindromes are generally found in the non-coding regions of mitochondrial DNA^[193].

Table 5.1: The percentage of the exhaustive lists of all possible palindromes (lengths 4, 6, 8 and 10) which are found in higher proportions in the non-coding regions than the coding regions, according to their significant p-values (Mann-Whitney tests). The row “ $p < 0.05$ only” excludes the set from $p < 0.01$ and indicates that these palindromes were not as significant as the $\alpha = 0.01$ group. Each column of this table correlate to our listed Hypothesis 1 through 4.

		Motif Length							
GC		4	%	6	%	8	%	10	%
Rich	$p < 0.01$	14	87.5	54	84.4	183	71.5	431	42.1
	$p < 0.05$ only	-	-	4	6.3	18	7.03	123	12
Poor	$p < 0.01$	13	81.3	43	67.2	118	46.1	501	48.9
	$p < 0.05$ only	-	-	10	15.6	43	16.8	166	16.2
Size Of Exhaustive List		16		64		256		1024	
Hypothesis		1		2		3		4	

Table 5.1 shows the results of the Mann-Whitney tests for the GC-rich and poor data sets with lengths of 4, 6, 8, and 10 bases. There are two α values given, for both, GC-rich and poor sequence data. In each column (length) and row

(significance), the number of palindromes out of the total (i.e the exhaustive set) satisfying our alternative hypothesis, is given for our Hypotheses 1, 2, 3 and 4. The *Size of Exhaustive List* represents the number of palindromes, taken from the exhaustive list, which passed the Mann-Whitney test, used to determine that the palindrome had larger proportions in the non-coding data than the coding data. The, “ $p < 0.05$ only” row, indicates the number of significant counts but not significant at the $\alpha = 0.01$ level. A percentage is also given to describe how much of the total number of palindromes for this length were able to satisfy our alternative hypothesis. By the nature of the Mann-Whitney test, we only learn whether the proportion of a particular palindrome is greater in the non-coding data than the coding data and so it could be that the proportions were low in both areas, but less so in the non-coding regions.

5.4.1 Lengths-{4,6,8,10} Palindromes

From the GC-rich sequences we note that 14 of 16, length-4 palindromes (87.5%) are abundant in the non-coding regions at the $\alpha = 0.01$ significance level (also significant at the 0.05 level). In the GC-poor set, we have 13 out of the total 16 (81.3%) were found in greater proportions in the non-coding regions. All were significant at the $\alpha = 0.01$ level. A large percentage of the exhaustive list of palindromes of length-6 was found to have higher proportions in the non-CDS regions. From the GC-rich sequences, 54 of 64 (84.4%) at $\alpha = 0.01$ and 43 of the total 64 (67.2%) for the same alpha in the GC-poor data. For the GC-Rich dataset, four palindromes were significant only at $\alpha = 0.05$ level and 10 in the GC-poor. The majority of the possible palindromes of length-8 are still found in abundance in the non-coding regions of the GC-rich dataset; 183 of 256 (71.5%) at $\alpha = 0.01$. For the GC-rich dataset, 118 palindromes of the total 256 (46.1%) were significant also for the same alpha level in the GC-poor set. Palindromes of length-10 are almost too long to be called “short”

palindromes and since the normal RSSs is on average length-6, we expect now to see some changes in the general trends of palindromic abundance in the non-coding regions. For example, for the GC-rich set, 431 or the total 1024 there is 42.1% of the total set of all palindromes at the $\alpha = 0.01$. For the GC-poor set, we have 501 of the total 1024 (48.9%) at the $\alpha = 0.01$.

5.5 Conclusions

Short palindromic sequences play important roles as restriction sites for cleaving enzymes. Various studies have provided evidence that these palindromes occur in reduced numbers along the bacterial genome but they do not provide evidence about where palindromic avoidance is actually happening. In this study, we hypothesized that avoidance of short palindromes (for lengths {4, 6, 8, 10 }) is concentrated in the coding regions which is thought to be less tolerant of the palindromic instability^[100]. Our argument was further motivated by observations in the literature that longer palindromes have been found performing their structural duties in the non-coding regions^[193].

The results described in this contribution can be used to determine strategies for finding and studying biological mechanisms which depend on palindromic involvement such as, auto-immune function, restriction enzyme activity and methylation systems. More importantly, a sequence property such as the one observed here that was obtained from the analysis of complete genomes, would be very important in whole genome sequence assembly and annotation problems. Similar to pieces of sky in jigsaw puzzles, reads belonging to certain regions in a genome are difficult to position correctly. In particular, when current assembly algorithms use common overlapping substrings of letters as a basis to assemble sequence reads, presuming the two reads likely originated from the same

chromosomal regions in the genome. Consequently, most of the assembly algorithms are greedy or graph based. Incorporation of sequence specific features observed from biological samples is expected to overcome the limitations that arise during sequence assembly.

In the future, we will study the GC content of the palindromes to find their distribution properties. In greater detail, we plan to analyze the role of sequence specific feature in the development of sequence assemblers.

5.6 Article Details

This contribution was published in the 8th International Symposium on Bioinformatics Research and Applications (ISBRA), 2012.

- Oliver Bonham-Carter, Lotfollah Najjar, Ishwor Thapa and Dhundy Bastola, “Distributions of palindromic proportional content in bacteria”, 8th International Symposium on Bioinformatics Research and Applications (ISBRA 2012).

The past, like the future, is
indefinite and exists only as a
spectrum of possibilities.

Stephen Hawking

Chapter 6

A Base Composition Analysis Of Natural Patterns For The Pre-Processing Of Metagenome Sequences

6.1 Abstract

On the pretext that sequence reads and contigs often exhibit the same kinds of base usage that is also observed in the sequences from which they are derived, we offer a base composition analysis tool. Our tool uses these natural patterns to determine relatedness across sequence data. We introduce spectrum sets (sets of motifs) which are permutations of bacterial restriction sites and the base composition analysis framework to measure their proportional content in sequence data. In this

contribution of the thesis, we suggest that this framework will increase the efficiency during the pre-processing stages of metagenome sequencing and assembly projects.

Our method is able to differentiate organisms and their reads or contigs. The framework shows how to successfully determine the relatedness between these reads or contigs by comparison of base composition. In particular, we show that two types of organismal-sequence data are fundamentally different by analyzing their spectrum set motif proportions (coverage). By the application of one of the four possible spectrum sets, encompassing all known restriction sites, we provide the evidence to claim that each set has a different ability to differentiate sequence data. Furthermore, we show that the spectrum set selection having relevance to one organism, but not to the others of the data set, will greatly improve performance of sequence differentiation even if the fragment size of the read, contig or sequence is not lengthy.

We show the proof of concept of our method by its application to ten trials of two or three freshly selected sequence fragments (reads and contigs) for each experiment across the six organisms of our set. Here we describe a novel and computationally effective pre-processing step for metagenome sequencing and assembly tasks. Furthermore, our base composition method has applications in phylogeny where it can be used to infer evolutionary distances between organisms based on the notion that related organisms often have much conserved code.

6.2 Introduction And Related Work

During a DNA sequencing task, the nucleotides of the reads or contigs must be placed in the correct order to reconstruct the original sequence. This sequencing task is particularly challenging when working with a metagenomic task, which requires one to gather and order similar sequences from a number of different organisms. This metagenomic technique has been extensively discussed in [162;316] and a framework to

infer phylogenetic relationships (patterns) among assemblages of microorganisms has been developed^[146]. This approach is expected to help improve assembly projects by reducing search spaces when grouping related sequence fragments. Massively parallel *next-generation* sequencing technologies (a major technological rebirth of the former Sanger methods of the 1980's^[264]) provide ultrahigh throughput results at a low cost but the reads are often too short to be able to determine their adjacency. In^[183], the authors describe a novel method for *de novo* assembly of large genomes from short read sequences which they used to assemble two giant genomes: the Asian and African human genome sequences.

Some of the limitations encountered in the assembly process include read coverage and size. The absence of placement information such as read coverage forms a bottleneck in the reassembly process^[80;127]. When the read sequences are very short, then special procedures must be taken to maximize their informational content to achieve placement evidence. For this work, it may be necessary to form contigs by *de novo* assembly methods as in^[305]. Despite these limitations, technologies such as *Velvet* and *Oases* have been used for many genome assembly projects^[96;136] and^[324]. Assembling reads using approaches from probability theory, or from the memory-based, are gaining popularity. This was determined by Zhang *et. al.*^[328] who compared the performance of eight distinct tools (i.e., SSAKE, VCAKE, QSRA, SHARCGS, Edena, Velvet, SOAPdenovo, and Taipan) against eight groups of simulated datasets.

In metagenomic studies, where there are different kinds of reads or contigs mixed together into the same pool, the task of separating them back into n -distinct groups, becomes an NP-hard problem. Although a researcher may choose to determine their order using some computational tools, as described in Figure 6.1, this still is an NP-hard problem to separate the sequence data.

Furthermore, this difficulty of separating the reads may prevent the assembly tools

from ever being used optimally. In [225], the authors discuss the problem of filtering the reads or contigs into smaller groups for better management. Time and productivity can be saved by these pre-processing steps where related sequence material is placed into a bin (here called, *binning*) to reduce search spaces for reconstructing entire sequences or genomes. It is therefore important to perform efficient binning steps to save costs in the sequencing task to reduce the work-load in an assembly project.

Chromosomal material across different genera were organized into species-specific groupings by virtue of the motif composition which was contained in the DNA [164]. In our study, we present a similar framework of organizing samples of DNA by their motif content. Our method differs from the authors' work, however, because it could be applied to smaller sequence fragments than chromosomes and it also employs motifs of similar base-composition to *associate* (e.g., bin) sequences of different organisms into related groups. Our set of motifs are biologically relevant since they were derived from known bacterial restriction sites. We permuted the base composition of the bases found in a particular restriction site to generate a list of all possible motifs of the same composition. Here, all the motifs belonging to a set of the same base composition is said to form a *spectrum set*. We show that an organism's recognition sequence belongs to only one of the four possible non-palindromic spectrum sets. Furthermore, each set must be strategically selected for successful sequence binning.

Our hypothesis is that a restriction site base composition algorithm can be used to separate and bin the sequence material from several different organisms. Our method compares the spectrum set motif proportions between sequences and uses this knowledge to separate them. For instance, if the motifs have similar proportions across two sequences, then there is evidence to suggest that the sequences are related to each other in some way. Here, this relation is called an association. In summary, our work stands apart from the traditional assembly pre-processing methods found in wet-labs since our method relies on statistics alone to find likenesses across sequence

material to discover associations and bin the sequence data.

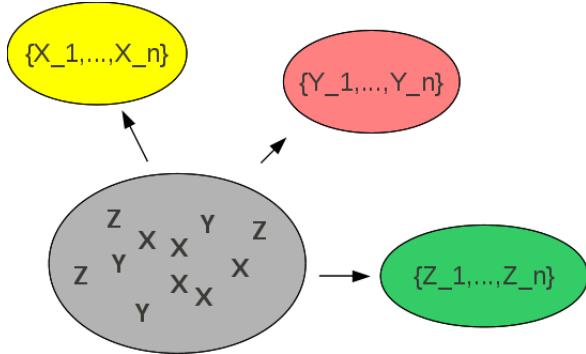


Figure 6.1: Sequence fragments are separated into groups (called, “bins”) of relatedness by a quick pre-processing step. This graphic taken from our previous work in^[28].

6.3 Methods

6.3.1 Genome Sequences

The genomic DNA sequences were studied from six different phylogenetic groups (Actinobacteria, Firmicutes and Proteobacteria) shown in Table 6.1. The genomes and chromosomes for this study were downloaded from Genbank (a public and international online database) and were manipulated with tools which we describe below. All sample genomes were at least 1Mb in length. There is evidence to suggest that GC-groups have a tendency to mutate to AT groups^[119;120]. Furthermore, it is thought that similar GC composition implies similar genomic structure^[185]. In light of this knowledge, our analysis was drawn from bacteria comprising many raised and lowered levels of GC-content.

In each experiment, ten trials of five freshly drawn reads were studied. We used MetaSim^[245] to create artificial contigs (or reads) similar to those of an actual assembly project. Each set of contigs (or reads) was extracted from the *Contig Originator* organism of Table 6.1. We applied all four spectrum sets to determine

Table 6.1: This table displays the genera used in our study. The *Read Originator* column displays the sequences which we processed via MetaSim for its reads. To determine their general associative behaviors, we studied ten trials of five freshly drawn reads. We chose two organisms from each of the three divisions from which to draw our contigs.

Organism	Contig Originator	Division
<i>Bifidobacterium longum</i>	NC_004307	Actinobacteria
<i>Mycobacterium bovis</i>	NC_002945	Actinobacteria
<i>Clostridium tetani</i>	NC_004557	Firmicutes
<i>Staphylococcus aureus</i>	NC_007622	Firmicutes
<i>Burkholderia pseudomallei</i>	NC_012695	Proteobacteria
<i>Campylobacter jejuni</i>	NC_008787	Proteobacteria

the proportional distributions used for the leaf weights in our heatmap trees. We placed randomly selected contigs in each test. There were also several other related organisms added to each pool to test and further determine the association behaviors.

We found very similar trends in each division. We illustrate them by discussing the arbitrarily chosen the organisms *Staphylococcus* and *Clostridium* of the Firmicutes division. The results from the other divisions featured in Table 6.1, Proteobacteria and Actinobacteria, were very similar to the findings of the Firmicutes.

6.3.2 Read And Contig Sequences

The synthetic data was made up of shorter reads of less than 1kb and were generated utilizing the 454 framework that was offered by the MetaSim software tool. MetaSim selects its reads by a statistical approach according to user input. The software simulates the approaches of both Sanger sequencing and Roche's 454 (sequencing-by-synthesis). The maximum allowed length of contigs by MetaSim is 1Kbp and so the longer reads or contigs for this study (1Kbp - 30Kbp) had to be generated by our own tool, which also follows the 454 (sequencing-by-synthesis) method. We created longer

reads or contigs of lengths 2kbp, 5kbp, 10kbp and 30kbp for an exhaustive study using this tool. Although it may appear that some of these reads are unnaturally long, we note that the typical lengths of reads appear to be growing as the sequencing technology improves and evolves.

In our experiments, we ran binning tests containing many reads or contigs but due to redundancy in the outcome of the analysis, our tests required only about five to ten reads or contigs to display the relevant trends. This small set of sequence data was acceptable to our work because we often observed that nearly all of the reads of a larger set had very similar distributions of motifs content from the spectrum sets.

6.3.3 Motifs

REBASE^[248] is an online database of information concerning bacterial restriction enzymes and their recognition sites. Each of the organisms (*Campylobacter*, *Burkholderia*, *Bifidobacterium*, *Mycobacterium*, *Clostridium*, *Staphylococcus*) were queried at REBASE for their organism-specific, palindromic recognition site sequences of length-6. This length was desirable for our work because (1) it is a common size in bacteria, mitochondria and plasmids and, (2) it is statistically interesting. For example, let A be the size of the DNA alphabet {A,C,G,T} (four elements) and let L be the motif length. There are $A^{\frac{L}{2}} = 4^{\frac{6}{2}} = 4^3 = 64$ possible palindromic sites available from the set of all possible length-6 words, $A^L = 4^6 = 4096$. When compared to the seemingly spontaneous occurrence rates of the shorter motifs, these longer words are less likely to be random occurrences along the genome.

6.3.3.1 Base Compositions And Spectrum Sets

There are usually several uniquely spelled, palindromic recognition sequences of length-6 for each bacterial organism according to REBASE. For example,

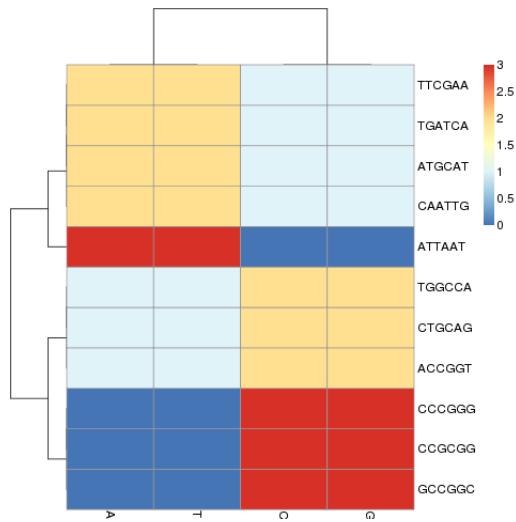


Figure 6.2: The spectrum set taken from the four restriction sites of the *Clostridium* genera. There are ten unique recognition sites covering all four spectrum sets (shown in Figure 6.4). This graphic taken from our previous work in^[28].

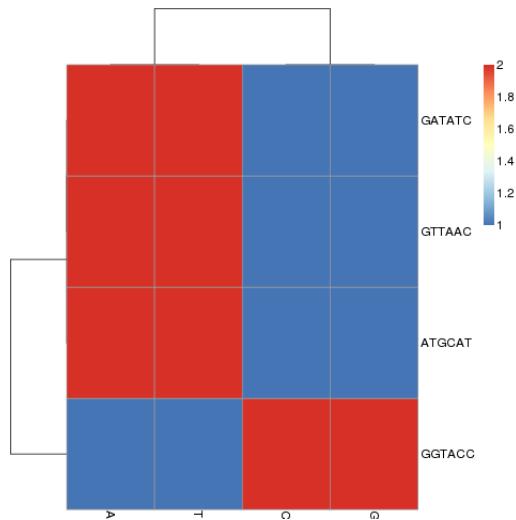


Figure 6.3: The spectrum set taken from the four restriction sites of the *Staphylococcus* genera. The motif ATGCAT is common to *Clostridium*. This graphic taken from our previous work in^[28].

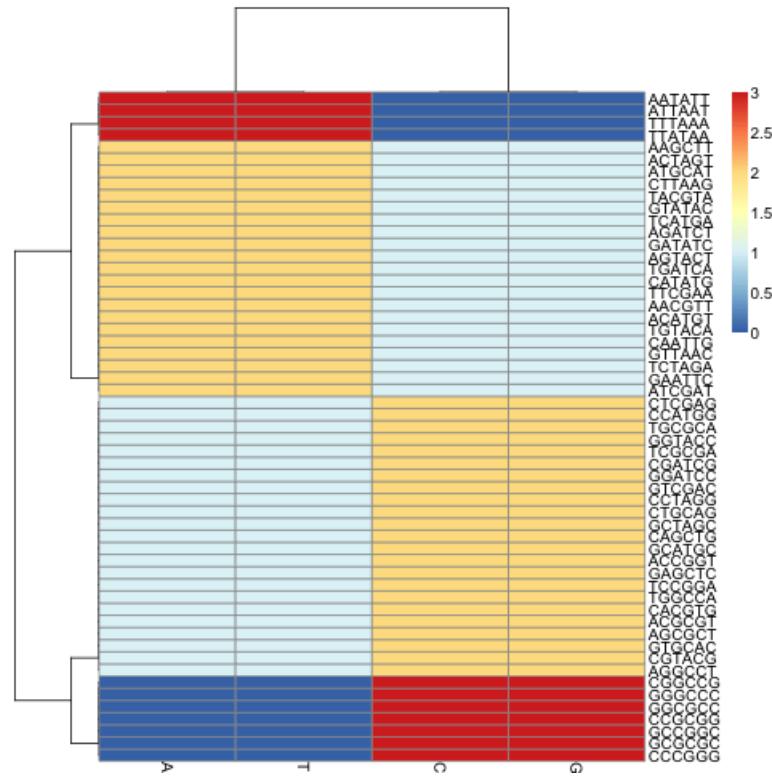


Figure 6.4: From its base composition, each bacterial restriction site fits into only one of the four spectrum sets, featured by unique color patterns. The motifs of each set are made up by the permutations of one of the following words; *AAATTT*, *AATTAG*, *CCGGAT* or *CCCGGG*. This result taken from our previous work in [28].

Clostridium has eleven recognition sequences (Figure 6.2), and *Staphylococcus* (Figure 6.3) has only four. It is typically rare to find common recognition sequences between two organisms however, in this case, *ATGCAT* is common to both. Consulting REBASE, we found all the known restriction sites and placed each into one of four unique sets according to their DNA compositions. In Figure 6.4, we show this grouping of all restriction sites. We call these sets, *spectrum sets* where each element of a set contains the same count of each base. We name each set by the following motifs: *AAATT*, *AATTG*, *CCGGAT* and *CCCGGG*. For example, the motifs, *ATTTAA*, *AATTTA*, *TAAATT* and *TTAAAT*, are all elements belonging to the *AAATT*-spectrum set. A DNA word, w is palindromic if $w == \text{reversed}[\text{Complemented}(w)]$. We do not consider palindromes in our spectrum sets, although many of the restriction sites of restriction modification systems are naturally palindromic, since they are thought to be avoided in the genome^[100]. This avoidance property may confuse our results since we are investigating their occurrences in a sequence. Table 6.2 lists the sizes of each set.

Table 6.2: The numbers of available motifs belonging to each spectrum. The motifs in the spectrum set are non-palindromic and are permutations of the set seeds. The set created from the permutation of *AATTG* is called, the *AATTG*-spectrum, for example.

Set Seed	Available Motifs
AAATT	12
CCCGGG	12
AATTG	156
CCGGAT	156

6.3.4 Proportions

We use proportions, not frequencies, in our study of motif content because we are only using a subset of the set of all possible motifs of length-6. We ignored overlapping

palindromes (no nested motifs) in the sequences for simplicity. The motif occurrence data in the sequences was normalized to make the comparisons meaningful. We determine the proportionality for each motif in a set across a genome by the following Equation 6.1:

The proportion of,

$$m_i \text{ in } S_L = \frac{\text{count}(m_i) * |m_i|}{|S_L|} \quad (6.1)$$

where m_i is a motif, S_L is a sequence fragment (a read, contig or genome), $\text{count}(m_i)$ represents the number of occurrences of m_i found in S_L , and $|m_i|$ and $|S_L|$ are the lengths of the motif and the sequence, respectively. For each motif in a spectrum set, the proportion of sequence that is made up of the motif is calculated by this equation. For each spectrum set, a vector is created from all proportions to be applied to a clustering analysis by `hclust`: a command in the R Statistical software^[288]. The result of the analysis is a heatmap^[154] to determine the associations.

We used the motif proportions, to make vectors from each sample sequence. Comparing the vectors across the organisms determined likenesses and relatedness. If the vectors of the spectrum set motifs were similar between sequences, then this may have been an indication of much common DNA between both sequences. This may also suggest a degree of relatedness between the organisms. Since a contig comes from a sequence, then the contig and the sequence will both share all their DNA and so our analysis will locate these similarity patterns and bin them together. Our analysis code was written in Python. In Figure 6.5, we provide a summary of how our method is applied.

6.4 Results And Discussion

According to their proportions of motif content, the clustering in heatmaps describes a tree of relatedness between the organisms. Similar proportions between the sequences

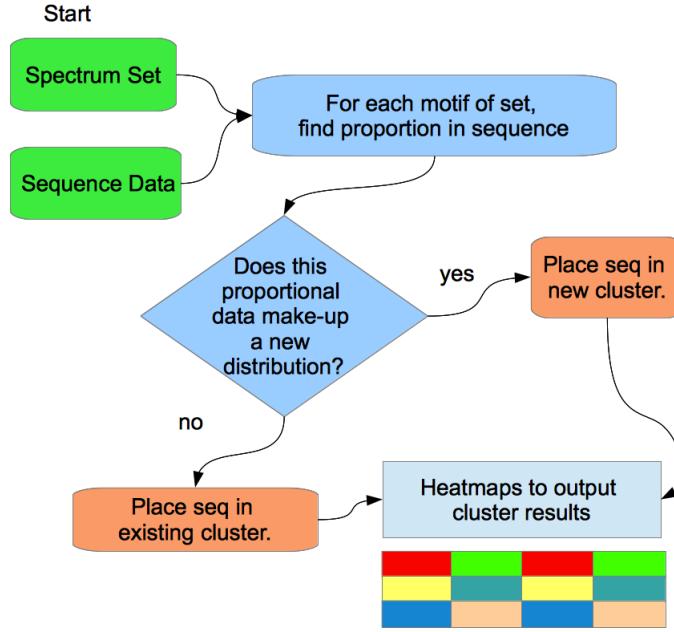


Figure 6.5: The flowchart that we applied to the clustering operation using heatmaps.

are indicated by their close proximity in a *subtree* of the main tree of relatedness. A *parent* sequence is one which is closely related to the sequence from which the reads or contigs were derived. Since these fragments may contain large regions of common code with parent sequence(s), they will associate with them and will be found in its subtree in our heatmaps. By association, we imply that there is ample evidence to suggest that the reads or contigs are more similar to their parent(s) than any other genome in the tree of relatedness. We, furthermore, suggest that these fragments make up a sequence that is related to the parent(s). This property can be utilized to create bins from which to begin assembling each sequence in the reassembly task.

6.4.1 Sequence Data

In the following, we discuss the task of binning long reads or contigs. Here, we choose to use DNA strands of a length 5000bps. These strands shall be contigs for the purpose of describing the tool that manipulates them. Our method is a tool to determine the proportions of motifs occurring in sequence data. The tool requires

enough information from each strand to make correct decisions about relatedness and if there is an insufficient amount of sequence material for comparisons to others, then our base composition tool will make poor determinations. Sequence fragments of 700bps were often enough to show the trends we discuss in this contribution, but we found some errors. We found that longer sequence data provided clearer and more accurate results due to having enough base information upon which our method relies.

This size of sequence material may seem large if the sequences were reads and not contigs. However, we note that sequencing and assembly technologies appear to continually create longer reads than previous technologies. Very large sizes may soon be a reality since read pre-processing methods and various read alignment technologies are already being used to create larger contigs^[85;183;219;259;274;325].

6.4.2 *Clostridium* And *Staphylococcus*

Clostridium and *Staphylococcus* typify the kinds of phenomena we observed after of ten trials of each experiment, using the arbitrarily selected pairs of organisms from Table 6.1. Here will describe the typical kinds of observed phenomena using spectrum sets on these two organisms. We will begin by showing that the two genera groups, *Clostridium* and *Staphylococcus*, are unrelated by the analysis of their motif proportions. We note from Figures 6.3 and 6.4 that only *Clostridium*, having the recognition sites *ATTAAT* and *CCCGGG*, can be discriminated by the *AAATT* and *CCCGGG*-Spectrum sets (*Staphylococcus* does not have restriction sites of this composition). By our analysis of motif proportions of this spectrum set, we see that both organisms have very different proportions of these spectrum sites.

We note from Figures 6.6 and 6.7 that there were two clearly contrasted subtrees in the heatmap to separate the two organisms. There was similar contrast between the sequences of our other heatmaps of the other organisms. In the present two organisms, we noted that the heatmaps are nearly opposite from each other: the

Clostridium family members tend to have warmer colors (elevated proportions) and the *Staphylococcus* members have colder colors (low proportions) in the *AAATTT*-spectrum set. This trend is the inverse for the *CCCGGG*-spectrum set.

The *AATTAG*-spectrum set was also successful in showing two different family subtrees but there was much less apparent contrast between the organisms than there was when using the *AAATTT*-spectrum set. We attribute this high contrast to the phenomenon that a spectrum set may perhaps be more biologically relevant to one of the organisms than the other, according to their recognition sequence usage. The *CCGGAT*-spectrum set was not typically very successful in showing contrasts for binning in our trials for these organisms. This same experiment was performed ten times with different (i.e., newly selected) contigs and we observed similar results in the heatmaps as those discussed. We suggest that since the *Staphylococcus* group appears to have a higher proportion of *CCCGGG* content than *Clostridium*, this contrast helps to associate the reads by relations.

It is clear that the proper use of the correct spectrum set can neatly differentiate one organism group from another for binning. Above, we saw that there are differences in the amounts of the spectrum sets which are found in the organisms. This made a high contrast which helped to determine one organism from another. We will now discuss how this method can discriminate between only read or contig sequence data.

6.4.3 Proportional Differences In Contigs By Spectrum Sets

We shall now discuss an application of separating reads originating from three different organisms that have been mixed together into the same pool. Incidentally, a part of this process comprises the separation of contigs belonging to two different organisms. For our test, we arbitrarily selected another organism (featured in our organism group in Table 6.1) *Burkholderia pseudomallei* to be added to the contigs from *Clostridium tetani* and *Staphylococcus aureus*. The contigs are of length 5000bps which we chose

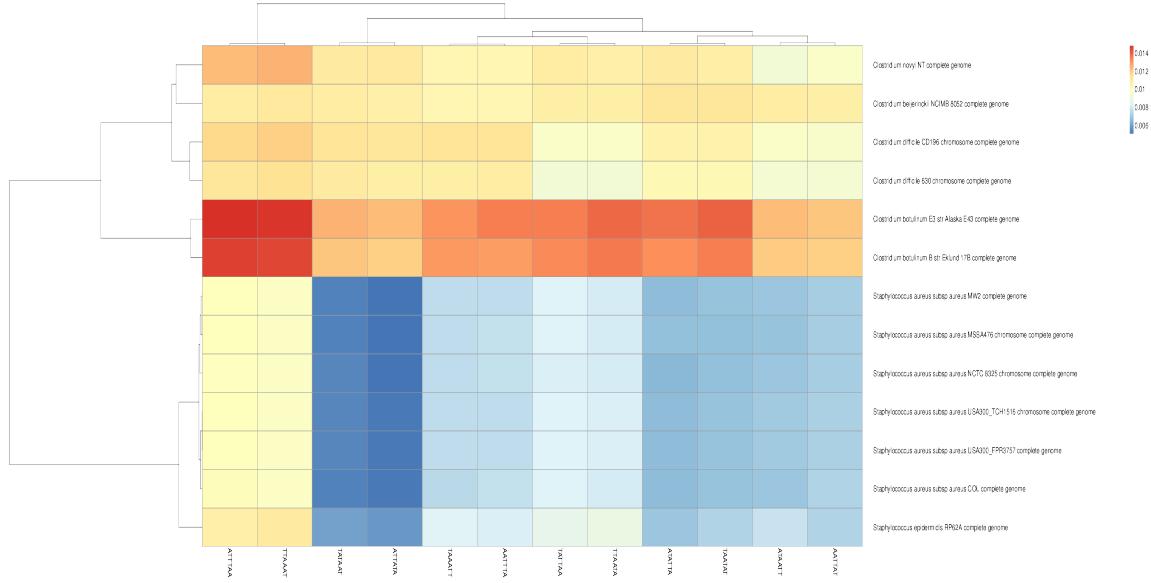


Figure 6.6: Separation by the *AAATT*-Spectrum set. There is a clear distinction between each bin; *Clostridium* and *Staphylococcus* of the Firmicute division. The data is segregated except for the two middle sequences forming a separate group. We had similar results from the *AATTCG*-Spectrum set. This result from our previous work in [28].

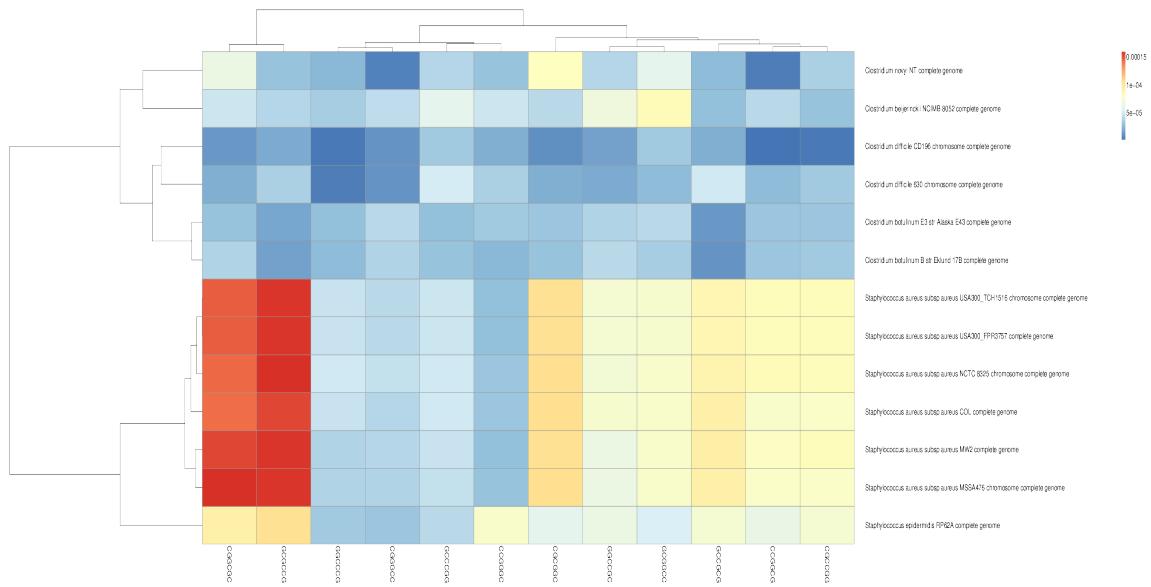


Figure 6.7: Separation by the motifs of the *CCCGGG*-Spectrum set. Note a clear distinction between each bin. In addition, we note that there is no longer a color pattern showing that *Clostridium botulinum* are closely related, as we saw in Figure 6.6. This result taken from our previous work in [28].

to illustrate the test and to showcase its performance.

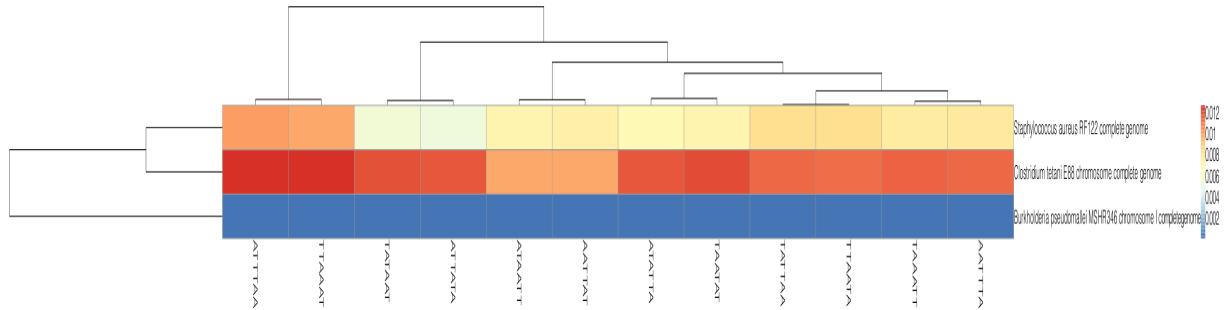


Figure 6.8: The *AAATTT*-Spectrum set test. The sequence data is applied to our base composition analysis to determine its relatedness.

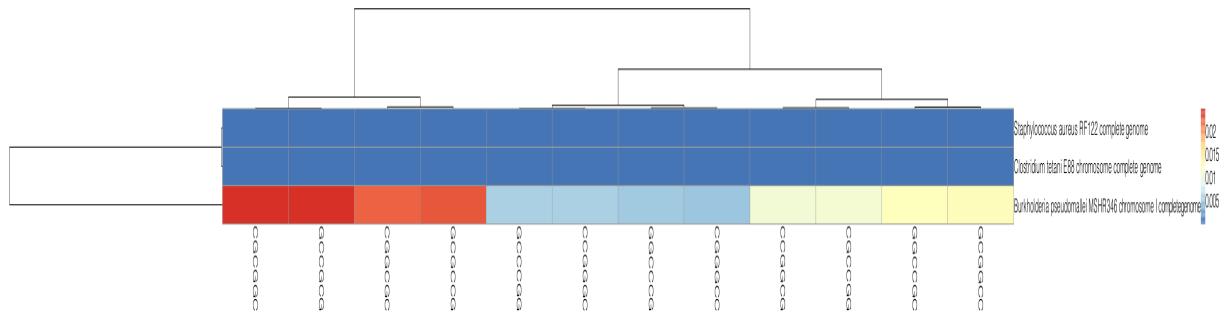


Figure 6.9: The *CCCGGG*-Spectrum set test. The sequence data is analyzed by base composition to determine relatedness.

6.4.3.1 Tests To Determine Pliable Spectrum Sets

When working with the contigs of two organisms, a spectrum set could be selected based on the restriction sites which are inherent to the involved organisms. However, a sequencing task may combine contigs of three or more organisms together. The contigs of each organism will have to be separated from those of the other organisms to make the sequence assembly more feasible. Due to the large number of contigs in the whole project, it may not be convenient to run a base composition analysis over all sequence data and so, to determine the spectrum set for the binning task, it is suggested to use the spectrum set test as shown in Figure 6.8. This test is

a base composition analysis taken only across the organisms who are known to be close relatives of the contigs (parents) in the pool. In Figure 6.8, we note that *Burkholderia* has the lowest proportions of the *AAATT*-spectrum set. Conversely, in Figure 6.9, *Staphylococcus* and *Clostridium* have the lowest proportions of the *CCCGGG*-spectrum set. When either of these spectrum sets are applied to the pool of all contigs, we note that the *Burkholderia*, *Staphylococcus* and *Clostridium* contigs reflect the same trends observed at the genome-level. For instance, Figures 6.10 and 6.11 reflect the underlining trends of Figures 6.8 and 6.9, respectively, in terms spectrum set motif coverage.

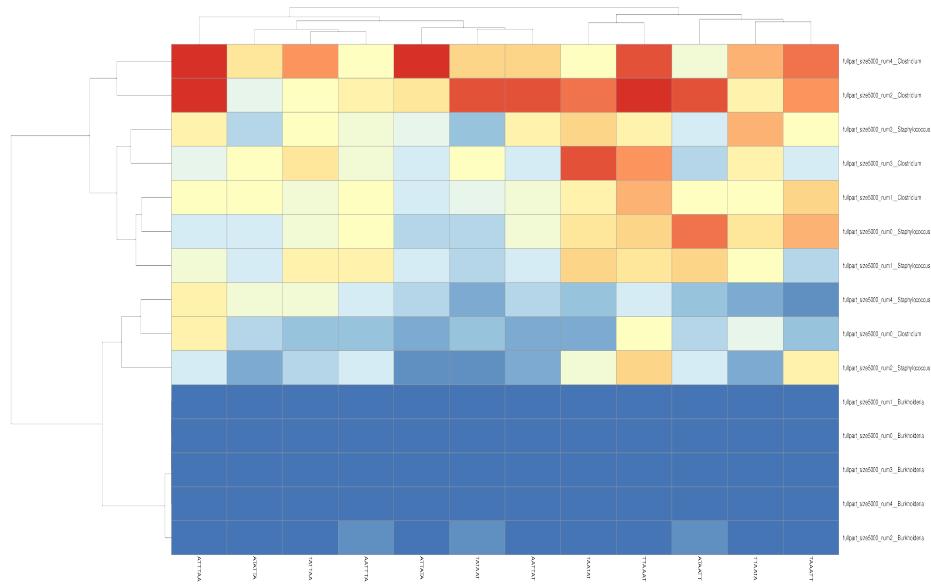


Figure 6.10: The *AAATT*-spectrum set analysis taken across all sequence data in a pool. The *Burkholderia pseudomallei* sequence data, having elevated proportions of the motifs of this spectrum set, create a contrast from those of *Clostridium tetani* and *Staphylococcus aureus* and have mixed proportions.

6.4.3.2 Removal Of The Contrasting Contig Group

In Figure 6.8 (spectrum set *AAATT*), we noted that *Burkholderia* had low proportions of this set, and also in Figure 6.9 (spectrum set *CCCGGG*, the opposite was true. In Figures 6.10 and 6.11, we see that the *Burkholderia* contigs also show

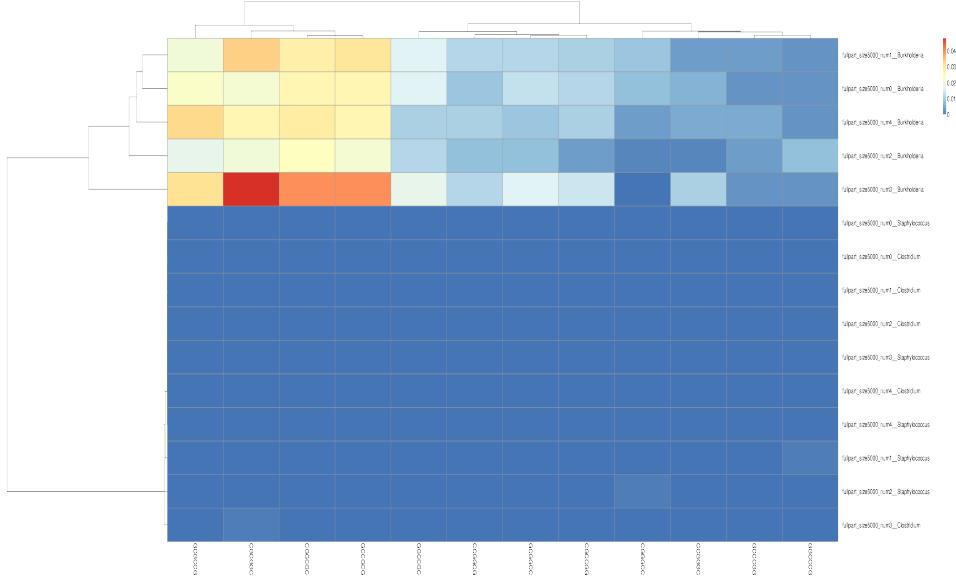


Figure 6.11: The *CCCGGG*-spectrum set analysis taken across all the contigs in the pool. We note that the *Burkholderia pseudomallei* sequence data, having low proportions of the motifs of this spectrum set, create a contrast from those of *Clostridium tetani* and *Staphylococcus aureus*. These organisms are observed to have mixed proportions by this heatmap.

this same pattern. Therefore, by this strong contrast, we could remove all contigs which show these strong contrasts and in doing so we would likely be binning the *Burkholderia* contigs. We note that the spectrum set *AATTAG* was unable to show contrasts between two of three organisms (Figure 6.12) but *Burkholderia* was still a contrasting group. Interestingly, without this organism, the *AATTAG* spectrum set clearly differentiated *Staphylococcus* and *Clostridium* contigs as shown in Figure 6.13. This suggests that the addition of *Burkholderia* (having such low proportions of the spectrum set motifs) to the set may change the parameters of the heatmap software.

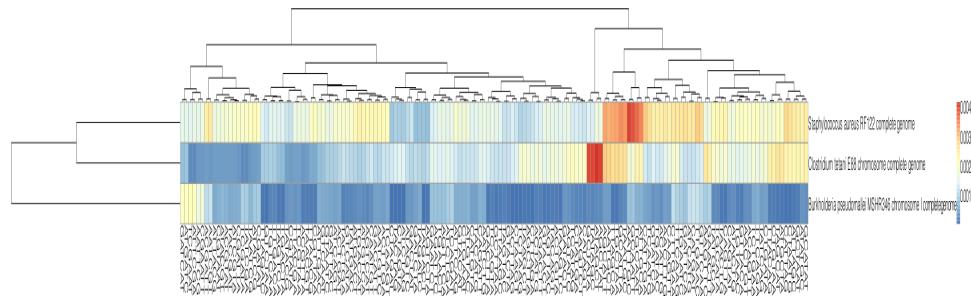


Figure 6.12: The *AATTCG*-Spectrum set test: The genomes or chromosomes are analyzed by base composition to determine the expected clustering behavior of their contigs.

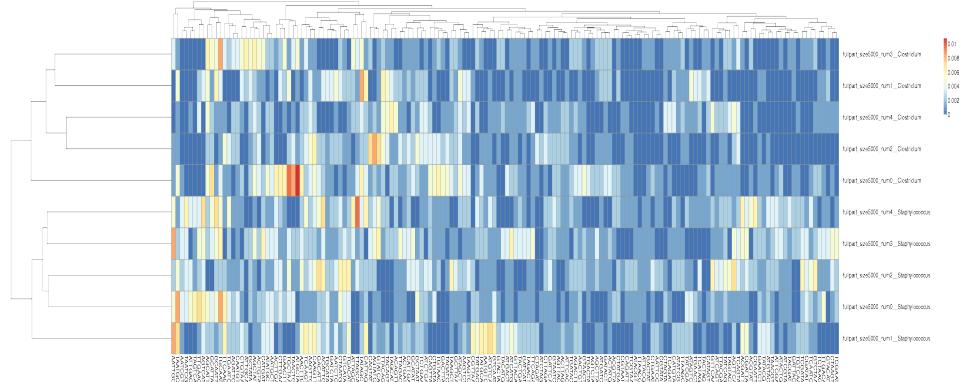


Figure 6.13: Separation of contigs of *Clostridium tetani* and *Staphylococcus aureus* by the *AATTCG*-spectrum set. We found that this spectrum set worked well to separate the contigs. The *AAATT*-spectrum set did not perform as well as we had expected from our work in Figure 6.6. We suggest that the contigs of these particular organisms followed trends shown in Figure 6.12.

6.5 Phylogeny From Full Chromosomes

To demonstrate its ability to differentiate sequence data into biologically relevant groups, we show that our method is able to form phylogenetic trees which conform to NCBI's taxonomy tool (available at <http://www.ncbi.nlm.nih.gov/taxonomy>). In our example, we arbitrarily selected a chromosome from each of seven diverse organisms listed in Table 6.3. We then applied our framework to extract the distributions of each spectrum set and compared the results to the taxonomy tree in Figure 6.14 from NCBI which is based on the classification of their taxonomy database and other resources.

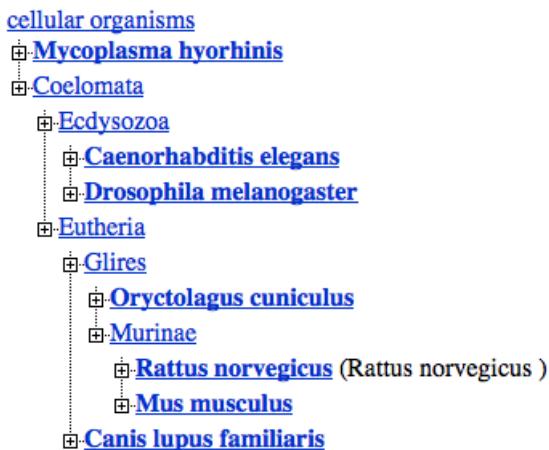


Figure 6.14: NCBI's Taxonomy Tree used for validation and comparison. This phylogenetic tree was used to compare the results of the spectrum set analysis of the organisms listed in Table 6.3. We ranked the results on a scale of highest to lowest resemblance in Table 6.4.

We remind the reader that the subtrees in this example contain organisms that may be related by basic evolutionary phenomena. If we had contigs in the pool from each of these organisms, then these fragments would associate to form more specific family subtrees. Instead, this data is chromosomal sequence material which group by relatedness.

We inspected the resulting trees of this example with the following criteria: the

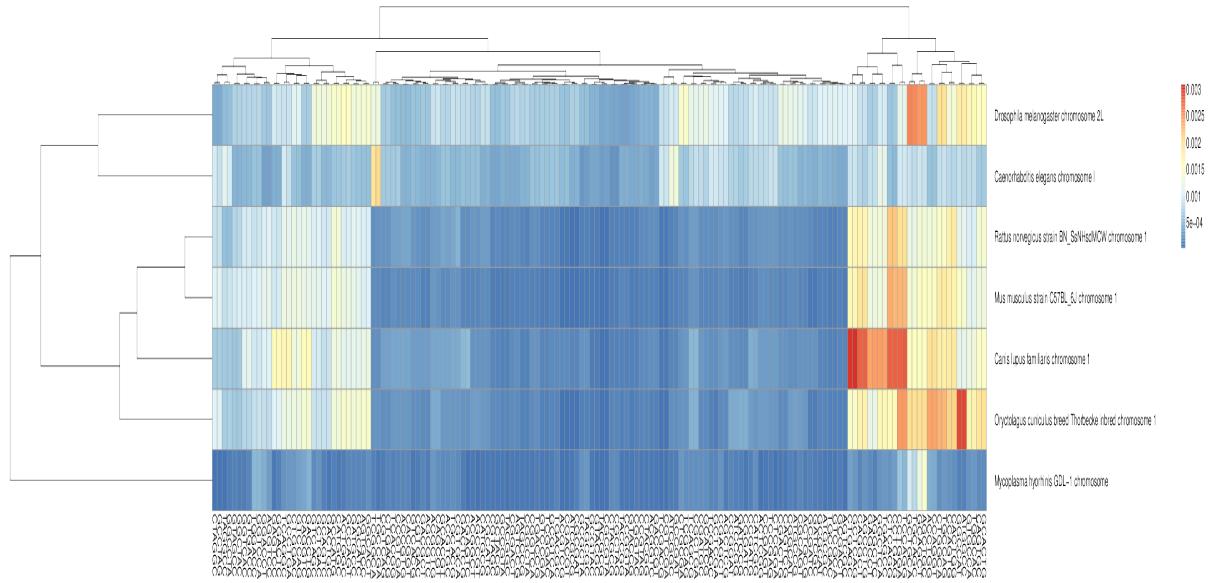


Figure 6.15: The *CCGGAT*-spectrum set. This tree perfectly resembles the taxonomy tree of Figure 6.14 and shows the great evolutionary distances between the organisms. The rat and mouse are found to be closely related. We note three distinct subtrees: one containing the bacterium, one for the mammals and one containing the worm and fruit fly. The location of these subtrees conforms to taxonomy tree.

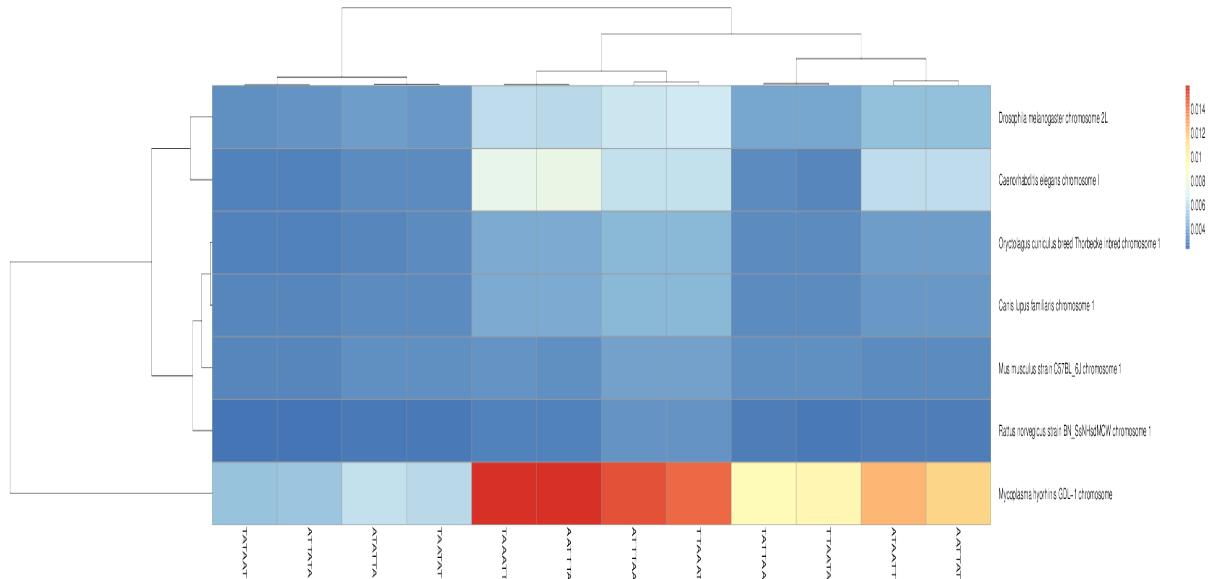


Figure 6.16: The *AAATT*-spectrum set. This tree also resembles the taxonomy tree but there is a slight distance between mouse and rat which is not found in Figure 6.15. We note three distinct subtrees conforming to the taxonomy tree.

Table 6.3: The organisms used in the base composition experiment. We note that rabbit, dog, mouse and rat are seemingly more closely related than the bacteria, fruit fly and the worm. This observation is used as a *first-glance* assessment of the heatmaps below.

Locus	Organism	Common
NC_003279	<i>Caenorhabditis elegans</i> , chrm1	Worm
NC_006583	<i>Canis lupus familiaris</i> , chrm 1	Dog
NT_033779	<i>Drosophila melanogaster</i> , chrm 2L	Fruit fly
NT_039169	<i>Mus musculus</i> , chrm 1 genomic contig	Mouse
NC_016829	<i>Mycoplasma hyorhinis</i> , GDL-1 chrm 1	Bacteria
NW_003159226	<i>Oryctolagus cuniculus</i> , breed Thorbecke inbred chrm1	Rabbit
NW_047544	<i>Rattus norvegicus</i> , chrm 1	Rat

bacterium should be the most evolutionarily distinct organism. The mammals (i.e., the dog, rabbit, rat and mouse) should be the most evolutionarily similar group of the set. The worm and the fruit fly should be found in a subtree which is evolutionarily between the bacterium and the mammals. Indeed, the worm and the fruit fly are quite diverse organisms, however, for this example they are clearly more similar to each other (than to the bacterium) and do not belong to the set of mammals. Therefore, our inspection involved checking for three basic subtrees: one for the mammals, one for the worm and fruit fly, and a subtree containing only the bacterium. In other words, the subtrees had to be arranged similarly to those of NCBI's taxonomy tree shown in Figure 6.14.

In Figures 6.15 through 6.18, we note the phylogenetic trees from each spectrum set. By inspection, the closest trees to the one in Figure 6.14 are from the *CCGGAT* and *AAATT* spectrum sets, Figures 6.15 and 6.16, respectively. Both of these trees show that the bacterium is most evolutionarily distant from rest of the organisms and that the fruit fly and the worm form a subtree which is distinct from that of the mammals. The locations of the subtrees in both figures are in the same configuration as illustrated in NCBI's taxonomy tree however, the tree of the

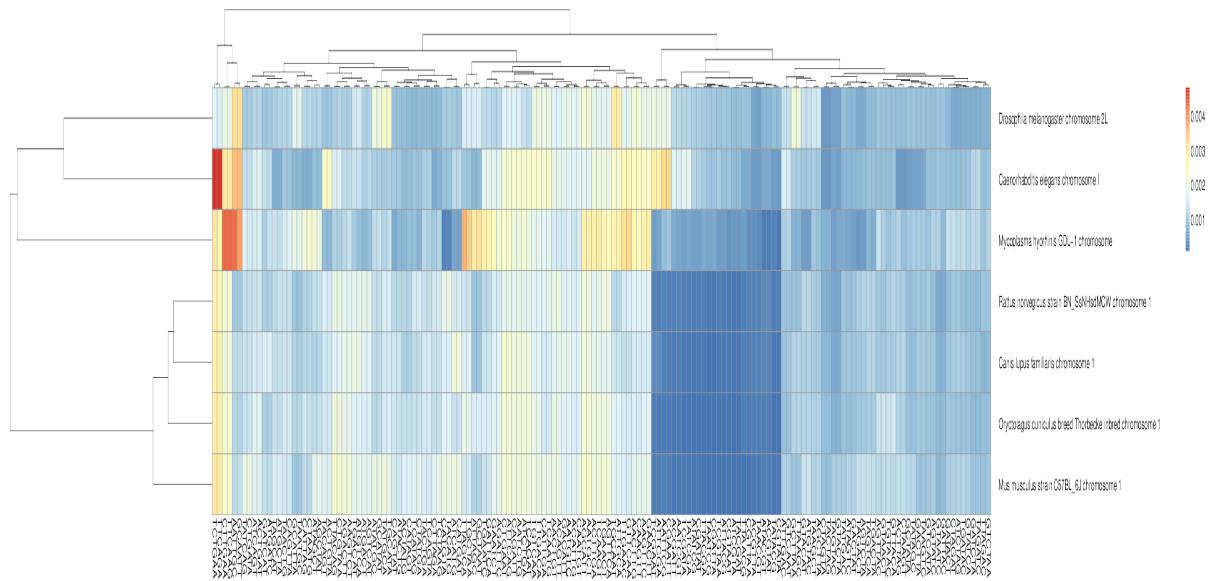


Figure 6.17: The *AATTG*-spectrum set. We note that mouse and rat are not closely related. The bacterium is also evolutionarily located between the mammals and the subtree containing the worm and fruit fly.

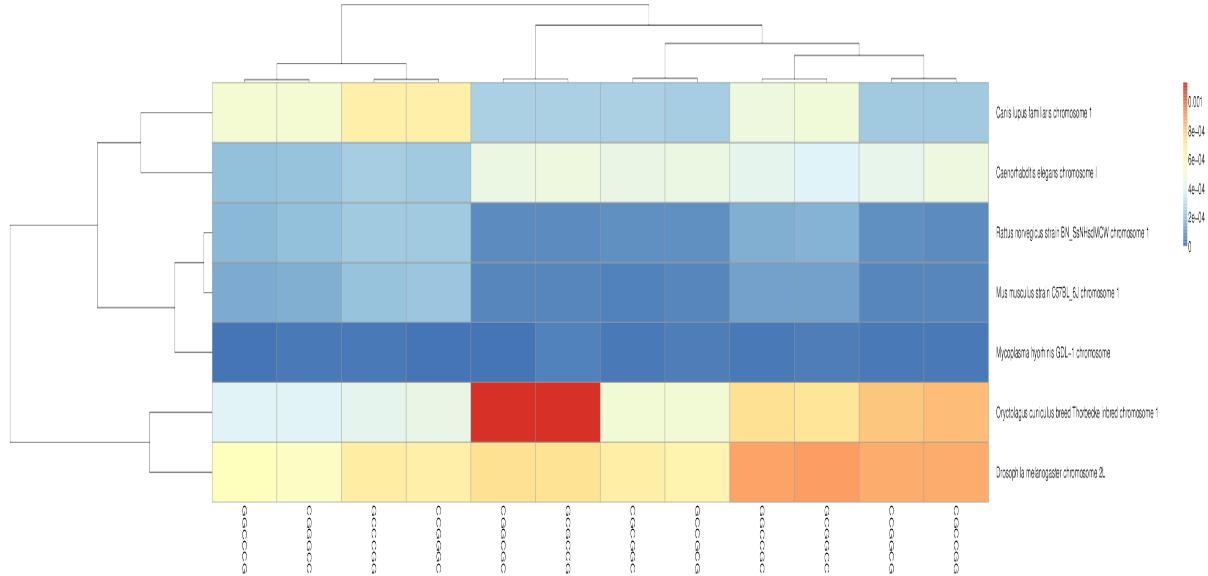


Figure 6.18: The *CCCGGG*-spectrum set. This tree is inaccurate because it indicates that the rabbit and the fruit fly are closely related.

AAATT-spectrum set is not as accurate as that of the *CCGGAT*-set due to the displayed shorter evolutionary distances (for instance, longer branch lengths indicate more distance). In addition, the distance between rat and mouse is expectedly closer by the *CCGGAT*-spectrum set than by the *AAATT*-set.

The tree from the *AATT**CG*-spectrum set in Figure 6.17 shows that the bacterium is evolutionarily found between the mammal’s subtree and that of the worm and fruit fly. This is inaccurate by the taxonomy tree Figure 6.14. In addition, the tree from the *CCCGGG*-spectrum set (Figure 6.18) is also inaccurate since it shows that the fruit fly is closely related to the rabbit. These results confirm our earlier findings that the choice of the correct spectrum set is of paramount importance for a successful analysis.

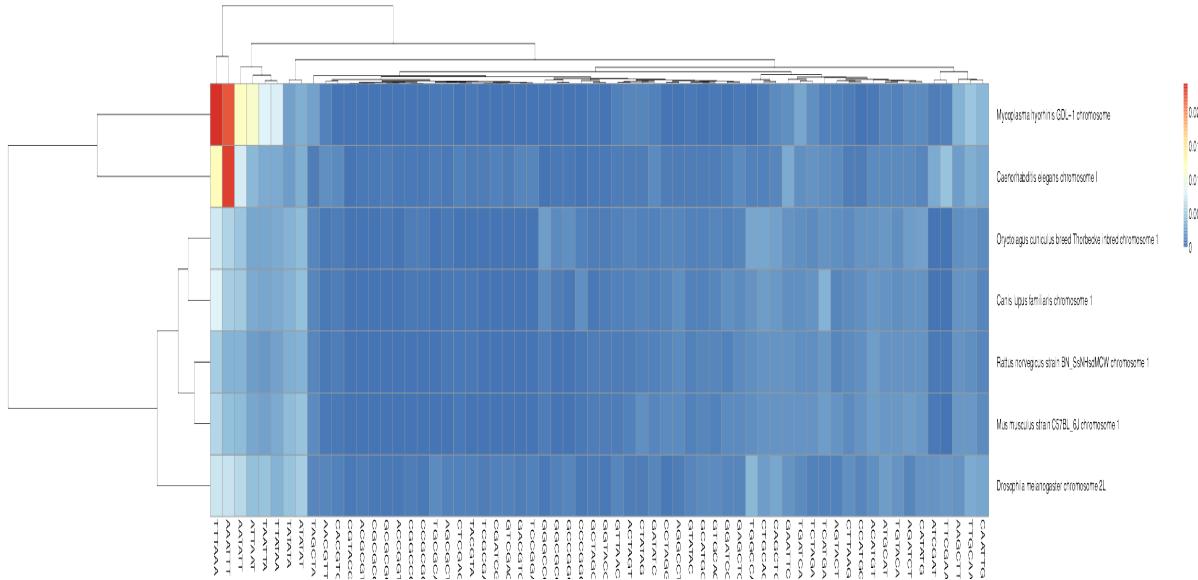


Figure 6.19: Length-6 palindromic spectrum set. Here we note that this tree does not conform well to the validation tree in Figure 6.14. Rat and mouse are shown to be closely related but inaccurately, the tree shows that the bacterium and the worm are also closely related.

Since the spectrum set motifs were originally inspired from palindromic restriction sites, we also studied the proportions of an exhaustive list of length-6 palindromic motifs (64 in total) across the sequence data. Interestingly, although palindromic

motifs have been known to successfully differentiate chromosomes, as shown in^[164]. However in Figure 6.19, we note that palindromes do not successfully recreate the taxonomy tree from Figure 6.14.

Table 6.4: Ranking of Spectrum Sets over Chromosomal Data: We note the best to worse resemblance of spectrum set trees to actual taxonomy data. For this data set, the *CCGGAT*-spectrum set created a tree which most closely resembled the one based on the classification in NCBI taxonomy database in Figure 6.14.

Ranking	Set Seed	Figure
1	CCGGAT	6.15
2	AAATT	6.16
3	AATTG	6.17
4	Palindromes, Length-6	6.19
5	CCCGGG	6.18

To summarize these results, we offer Table 6.4 which contains the highest to lowest resemblance to the tree in Figure 6.14. We note from their ranking that the spectrum sets do not behave uniformly and that further study is required to understand how they should be applied to a particular set of organismal data for classification.

6.6 Conclusion

As in playing with jigsaw puzzles, if there are the pieces of several different puzzles in the same box, then the completion of any one of the puzzles is a sizable undertaking. In the same way, during a sequence assembly task where the contigs of different organisms are mixed together in the pool, much time can be spared by first sorting the contigs into their own bins from which to work. Our method places many of the unknown contigs into their corresponding bins to drastically reduce the search space.

Most of this contribution discussed working with contigs which are typically longer than reads. Our base composition analysis tool works by quantifying the amount the spectrum set motifs which are contained in the sequence data. When there is not

enough sequence data, then our method may produce erroneous results and so we suggest using contigs of at least 1000bps because they should contain enough sequence information for a good analysis. We should mention here that we have had very good results when using contigs of 700bps in length and so 1000bps is not always absolutely necessary.

By sorting the contigs with related sequence data which is based on motif proportions, our method aims to accomplish the binning task. We used heatmaps to show the contig clusters by organism-types. Furthermore, we illustrated that there are only four spectrum sets, which can be created from length-6 recognition sites to apply to differentiate by contrasting the sequence data. For instance, we used the *AAATTT* and *CCCGGG*-spectrum sets to show that one set had high proportional values in one organism, but not the other. This created the contrast that would help to bin the contigs of these two organisms. We then showed how the contigs of three organisms can be binned in a two-step process. We first removed the most contrasting set of contigs in the pool and then reapply our method to the remaining contigs. An analysis by base composition can also be used to determine evolutionary orders of organismal sequence data. For instance, we showed that our method could create phylogenetics trees which were very similar to those produced by NCBI's taxonomy tool.

One of the leading benefits to our method is that there is no setup required as there would be for other sequence recognition softwares such as BLAST^[5] or BLAT^[147]. While these methods provide powerful sequence analysis, they require expansive hardware requirements for use (i.e., memory, storage and fast computational power). Our method is a statistical approach, programmed in Python to run on basic hardware and does not require a database for operation.

Our goals for the future are to test this base composition framework using synthetic and biological data to further analyze its performance and levels of

sensitivity. This work will be conducted using MetaSim, to generate contigs of a 10 X coverage for two or more genomes which we shall apply to our binning method. This study will help to give us a more realistic interpretation of its power for discriminating contigs and how best to use it as a pre-processing step to sequence assembly.

6.7 Article Details

This contribution was published in BMC Bioinformatics, 2013.

- Oliver Bonham-Carter, Hesham Ali and Dhundy Bastola, “A base composition analysis of natural patterns for the preprocessing of metagenome sequences”, BMC Bioinformatics 14. Suppl 11 (2013): S5.

The only real security that a man can have in this world is a reserve of knowledge, experience and ability.

Henry Ford

Chapter 7

sEncrypt - An Encryption Algorithm Inspired From Biological Processes

7.1 Abstract

As a fourth contribution of this thesis we present a new conceptual methodology for realizing encryption involving trap-door functions built from biological processes. Many standard encryption methods such as RSA security, for example, utilize functions that are easy to compute in one direction but the reverse is a computationally hard problem without a key. In biology, a trap-door like functions can be created from natural phenomena such as the process of creating protein sequences. A fragment of DNA can be transformed to protein easily however given a protein sequence, it is very hard to convert the protein information back to DNA. In

essence, protein creation is a lossy function and if we keep certain side-information secret, then a trap-door like function can be constructed from this mechanism that is ideal for encryption.

We propose sEncrypt (sequence Encrypt), a model inspired by the central dogma of biology to encode, encrypt, decrypt and decode plain text using publicly-available sequence data from bioinformatics research. We evaluate the entropy of the cipher text to show randomness of characters and show by autocorrelation tests that the encrypted text of our method contains no repetition which could form potential weaknesses. These tests and results show that the sEncrypt framework constitutes a good encryption framework for use in information exchange.

7.2 Introduction And Related Work

In many modern information security technologies such as encryption, key exchange, password protection and similar kinds of security, the protocols which provide the actual security are likely built out of functions similar to trap-door functions^[72]. When computing in one direction across one-way functions, the task is trivial, but computing in the reverse direction generally cannot be performed in feasible time^[178]. However, trap-door one-way functions are easy to invert when using a key.

To secure information, there are many different kinds of algorithms available which rely on trap-door functions or other functions for which the inverse is very difficult to find without a key. For instance, the RSA algorithm^[247] is based on the presumed difficulty of factoring large integers (the factoring problem).

A different kind of encryption, the Advanced Encryption Standard (AES), originally called Rijndael, is a cryptographic algorithm using symmetric block ciphers for protecting electronic data^[68]. This forms a part of the Federal Information Processing Standard. Serpent is a similar encryption system using

symmetric key block ciphers,^[6]. Twofish also uses symmetric key block cipher but with a block size of 128 bits and key sizes up to 256 bits^[260].

The AES, serpent and Twofish algorithms employ the substitution-permutation network method to confuse and diffuse output bits based on the input bits of the plain text. This network forms a series of operations which are hard to invert due to the near impossibility of constructing the input information from the output bits of the substitution-permutation network without the key. Additionally, these described methods also satisfy Shannon's *confusion and diffusion* properties^[267] which imply that the inter-character associations have been removed when constructing the cipher text.

Other low power consuming algorithms have been developed. Among notables are RC series of algorithms, RC4 being the most popular. Quasigroup based encryption algorithms have been explored in^[14;15;103]. At the same time quantum cryptography^[21;216] has shown promise for perfectly secure communication, however, remain far away from practical use because of limitations in hardware design. "Encryption less" secure storage of data by dividing it into partitions has been explored in^[217].

DNA watermarking, a system to identify fraudulent sequences or the unauthorized use of genetically modified organisms, has recently gained attention. In^[116], DNA-Crypt, a method to secretly mark sequences, is proposed using concepts from both encryption and steganography. By their method, the authors propose that information be conveniently binary-encrypted using algorithms such as AES, RSA or Blowfish. This encrypted information may eventually be placed into the DNA of a living organism for long term storage. Since DNA is likely to undergo mutations, the information must first be protected by correction codes such as the Hamming-code or WDH-code^[287] to ensure that the information is unaltered. Finally, the encrypted information is converted to DNA form and is placed into

organismal DNA where it cannot be found, except by those that placed it there.

Encryption is another area where biological processes have contributed. The literature contains many methods which involve wet-lab techniques. In^[174], a wet-lab method of encryption is presented employing primers (i.e., small and unique strands of DNA used for locating specific regions of DNA in a solution and here used as keys) to find a specific region of DNA corresponding to a binary-encoded plain text. Also employing primers, Gehani *et al.*^[99] proposed DNA-driven cryptography methods based on one-time pads (encryption by modular addition) of nearly unlimited size using DNA as a data structure. Encryption techniques based on number conversion, DNA digital coding and PCR (polymerase chain reaction) amplification are being explored by^[234]. Although wet-lab methods have been well-received by the community, they may be much slower than computer-algorithmic approaches such as the ones proposed in^[116] and^[253].

In this contribution, we propose sEncrypt (*sequence Encryption*), a new conceptual technique for encryption of plain text that uses the process of DNA to protein translation as its foundation. The proposed technique leverages two observations: first, is the existence of publicly available databases containing DNA sequences of millions of organisms and second, the many-to-one mapping between DNA codons and corresponding amino acid.

The mounting availability of publicly available DNA sequence data allows us to use terra-bytes of DNA data to form a part of our encryption key. In other words, if one were to randomly choose one of these sequences (belonging to a particular organism), to encrypt the message (plaintext) to be sent, then the identity of this organism could serve as a part of the secret key that would need to be conveyed to the recipient. Upon receiving the encrypted data and the secret key (organism ID), the recipient could go to one of these public repositories and download the corresponding DNA sequence to decrypt the data. The existence of terra-bytes of DNA data belonging to millions of

different organisms makes it difficult to determine the DNA of which organism was used for encryption.

The second observation, many-to-one mapping, enables us to create a trap-door like function in which some amount of side information is needed to invert the function. This many-to-one mapping arises from the fact that for 64 combinations of DNA bases only 20 possible amino-acids exist. Amino-acids form the building blocks of proteins. Further, for different organisms, this mapping differs in frequency of use. In other words, the frequency with which a given DNA codon maps to a given amino acid is unique for different algorithms. This information forms the second part of the secret key.

We show that the proposed encryption algorithm performs well to increase the entropy of the input sequence (indicating a highly random looking output sequence). This is one of the properties a good cipher text should have. Further, we perform auto-correlation tests between the outputs of different input sequences to determine if the proposed technique introduces any similar looking structure.

7.2.1 Background On DNA To Protein Translation

DNA (Deoxyribonucleic acid) is comprised of a sequence of “bases” (molecules) called adenine (A), guanine (G), cytosine (C) and thymine (T), sometimes also referred to as nucleotides. Therefore, a strand of DNA can be represented as a string of characters A, G, T and C. Further, DNA exists in the form of double stranded helical structure and each strand runs anti-parallel to the other based on certain pairing rules of molecules.

DNA is first converted into mRNA by a process called transcription and then mRNA is translated into amino acids that form proteins. There are 20 amino acids. However, the translation of DNA to an amino acid involves forming groups of three bases called codons that correspond to one amino-acid. Since there are four possible

bases (A, G, T and C) and a codon consists of three bases, a total of 64 possible codons can be constructed. However, only 20 amino acids exist. As a result, multiple codes translate to a single amino acid while some codes act as stop and start signals for the translation process.

7.2.2 Lossy Biological Functions

In essence the DNA to protein mapping is a many-to-one function. Therefore, obtaining the original sequence of DNA from the protein is not trivial because of its lossy nature. As we will see later, a function similar to a trap-door function may be constructed from this system to obtain the original sequence of DNA using a key.

In the set of 20 protein amino acids, most have between two and six different codons from DNA that encode them. For instance, leucine, a protein amino acid, can be encoded by six different DNA codons. By this redundancy, it becomes increasingly hard to determine the exact DNA sequence when starting at the protein sequence and working backward. Basic problems cannot always be answered, such as, whether the same codon is used each time to encode a particular protein amino acid, or whether there is no such logic. As the protein sequence gets longer the complexity eventually diverges since a sequence of n leucines has 6^n possible DNA formations. This assumes a uniform probability of mapping a codon to its corresponding protein amino acid. In [206], codon-use frequency tables have been created which shows that the translation process varies by organism.

7.3 Methods

Although each step is discussed in detail below, we provide Figure 7.1 to summarize each step of the encryption and decryption steps. In addition, we provide Table 7.1 to help the reader keep tract of the information as it transitions between the steps.

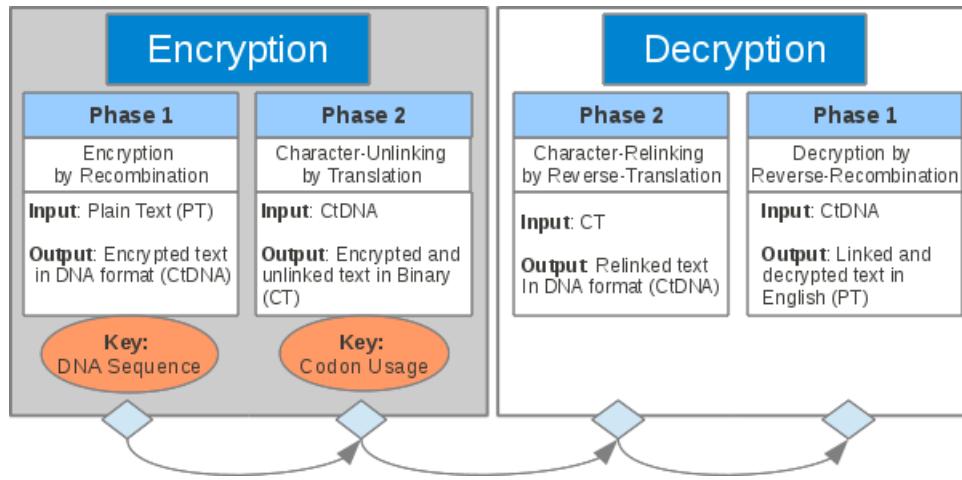


Figure 7.1: The flow chart of the encryption and decryption phases.

Table 7.1: A Summary of Steps. The plain text moves through phases one and two before becoming the cipher text.

Phase	Stage	Description	Sequence
One	PT	Plain text written in a language (here, English)	A, B, C
	PtDNA	Binary PT encoded into DNA	CAA CCA AGC AAT
	keyDNA	The sequence of DNA which is used by the Latin square to encrypt PtDNA	AGC TTT TCA TTC
	CtDNA	Cipher text in DNA form having completed the Latin square of phase one	TGC GGT TTT TTG
Two	CtProtein	Amino acids, A translated version of CtDNA	['C', 'G', 'F', 'L']
	CtFinal	Text version of the encoded amino acids and triplets	[('0001011', '1'), ...,]
	CT	CtFinal in binary format containing the protein encoding followed by the corresponding triplet codes.	[00010110010..., 011011110101...]

7.3.1 Phase One

7.3.1.1 Latin Squares

Latin squares are extensively discussed in [196]. A Latin square is an n by n matrix filled with n unique symbols such that no symbol occurs more than once in any row or column. The Latin square has elements on the top row and left-side column which, at their intersection points, yield a specific element. Although any permutation of the Latin square using the four bases of DNA {A,C,G,T} can be used as long as the same square is used for both encryption and decryption operations, we designed our Latin square for this study using a rotational ordering as shown in Table 7.2. The

Table 7.2: The quasigroup table used.

	A	C	G	T
A	a	c	g	t
C	t	a	c	g
G	g	t	a	c
T	c	g	t	a

notation of the Latin square is the following: each cell of a Latin square is written using a coordinate system of three members: $\{row, column, symbol\}$. The notation begins at the top, left-most cell and finishes at the bottom, right-most cell, covering the rows similarly in-between. Our own Latin square of Table 7.2 is therefore written: $\{(1,1,a), (1,2,c), (1,3,g), (1,4,t), \dots, (4,1,c), (4,2,g), (4,3,t), (4,4,a)\}$.

7.3.1.2 Sequence Encoding

In keeping with the biological concepts of this study, we mapped the English language plain text into a DNA form as done in Pedersen *et al.*[221]. For this effort, we imply that the data is manipulated as DNA, using modeled biological functions and processes. To create this unambiguous sequence of DNA from the plain text,

the plain text was first encoded in its binary form. Since the English-language PT characters and their punctuation can be encoded by a length-8 binary word, the concatenation of these words creates a single long binary sequence. Moving in pairs down this binary sequence, we assign DNA bases to the encountered pairs as shown in Figure 7.2. In this way, we obtained the plain text in a DNA form (PtDNA) for the entire English language sequence.

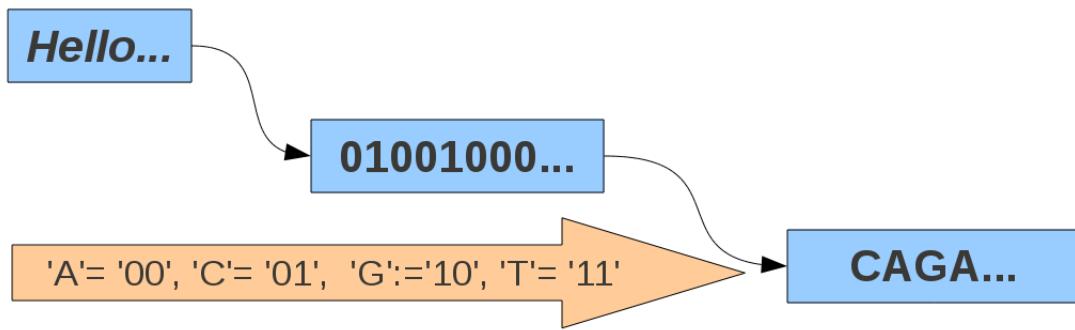


Figure 7.2: The process to convert PT into PtDNA. The plain text is encoded to binary words of length eight. This binary sequence is read a pair at a time and each is encoded by the correlating DNA bases.

At this point, we have only converted the PT into a PtDNA by an encoding procedure. It should be stressed here that *encoding* is the action of mapping the alphabet of the message to another via an arbitrary, but logical, one-to-one function. The logic behind this mapping is not hidden as it is in encryption.

7.3.1.3 Choosing The Encryption Key

The KeyDNA is a fragment of biologically relevant (or even synthetically created) DNA whose length is equal that of the PtDNA. This fragment can be obtained from Genbank^[23] or similar databases such as Ensemble^[89] or SwissProt^[133]. The same strand of keyDNA must be used by both the encryption and the decryption stages. We suggest using the biological DNA of some organism having an established genome in a public database. If a key from non-biological or synthetic DNA is desired, then

this sequence must be transferred to the receiving party, which creates concerns of how the exchange will be made. It is therefore easier to construct the key from a genome from a database which both the sender and the receiver can access.

Once a DNA sequence is chosen, we can extract the encryption key from it using one of the two techniques. The first method is the *inclusive base-to-base* where the key begins at some randomly chosen base in the genome and includes all bases for a distance of the necessary length. Another key selection method is the *periodicity of the n^{th} base* system, where an arbitrarily base is selected and each n^{th} base after this location is appended to create the key of the desired length. Since selecting the n^{th} base in the natural DNA to create a KeyDNA sequence likely removes base structure, the resulting key may not resemble actual DNA. If it is desirable that the key retain some biological structure, then we recommend the inclusive base-to-base method instead to create the KeyDNA as the chances of including some basic structure are better than employing the periodicity of the n^{th} base method.

For this study, our KeyDNA was arbitrarily created from the genome of *Escherichia coli* (LOCUS: NC_008253) which is publicly available from Genbank^[23]. Although we could have chosen any point for the start, we chose the first base (position 1) of the genome, for simplicity, using the *inclusive base-to-base* method to create a key sequence of length of 12 – the length of our PtDNA.

7.3.1.4 Encryption

After the KeyDNA has been created, we aligned both strings (the PtDNA and the keyDNA) to achieve a pairing of their bases at each position (i.e., PtDNA_{*i*} AND KeyDNA_{*i*}, for $0 \leq i \leq$ the length of the PtDNA). Taken together, each position gives a base from the KeyDNA and one from the PtDNA, as described in Figure 7.3. When using the Latin square, there are rows and columns for which the keyDNA and PtDNA sequence data must be applied. In our study, we arbitrarily chose to use the

rows for the key data and the columns for the PtDNA data. We note: the Latin square on the receiving end must apply the same sequence data to the same row and column for the decryption to work.

To encrypt the data using the Latin square, we located the keyDNA base-character in its left-most column. We then found the intersection of the column containing the PtDNA base-character (found in the top row). This intersection between the KeyDNA (row) and the PtDNA (column) is the cipher text base-character as illustrated in Figure 7.4.

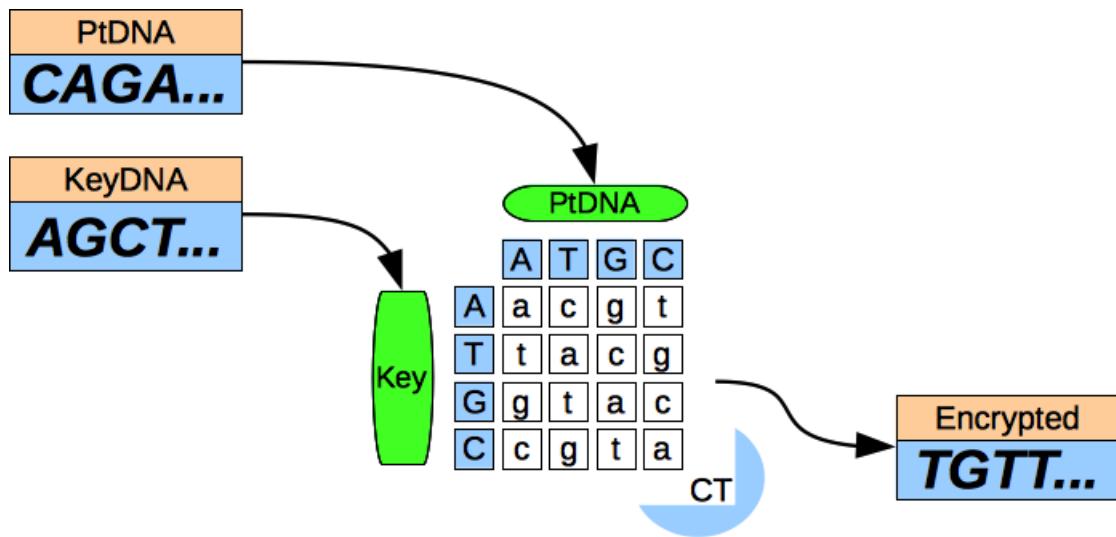


Figure 7.3: The application of the Latin square to the PtDNA and the KeyDNA. This step serves to encrypt the information by recombination in function of two input-sequences.

7.3.2 Phase Two: Translation Of DNA To Protein Code

According to the central dogma of biology, DNA is encoded into RNA to be translated into protein code. In the literature, it is well known that there is much redundancy in the triplets which encode a particular protein amino acid. In addition, it is also understood that organisms have varying habits of how they encode proteins from triplets. In this phase, sEncrypt converts CtDNA to protein amino acids (CtProtein)

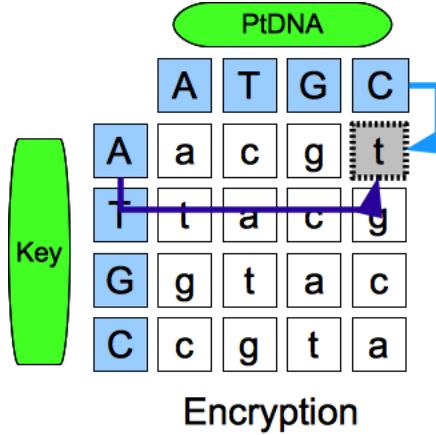


Figure 7.4: Encryption using the Latin square. Here the KeyDNA character is ‘A’ from the left column and the plain text character is ‘C’ of the top row. At the intersection of these two characters is the cipher text character ‘t’.

to further conceal the original PT using natural biological processes.

CtProtein is then converted to binary using efficient Huffman codes to make CtFinal (explained below). We use Huffman codes for efficient transmission of encrypted messages.

7.3.2.1 Huffman Codes From Triplet Frequencies

Each listed organism appears to have a unique preference of triplet usage to build its proteins during translation. From these preferences, frequency distributions have been derived^[206] which we applied to constructing Huffman codes^[130]. These prefix codes are further optimized as they are able to be read without the need of a delimiter separating them.

To translate our CtDNA, we arbitrarily chose an organism, *Bacillus phage PBS2* from which to use these frequencies to create codes. A casual glance will show that the triplet, UUA for leucine (L) is the only triplet used to create this amino acid by this organism. However, in a related organism, *Bacillus phage SP82*, UUC has the highest frequency of use for this protein amino acid. In using these frequencies to make codes, the task of determining the original DNA from our protein sequence is

Amino Acids	Proline (P)	Alanine (A)
A C . P S . Y	10101 0001011 00011 0101 01110	CCA CCC CCG CCT 0 11 110 10
Original DNA:	CCG GCC	
Protein:	PA	
Protein Code:	00011, 10101	
Triplet Code:	110, 110	
Final:	{00011, 110}, {10101, 110}	

Figure 7.5: The transformation of DNA code to Huffman binary codes. Here we coded these triplets and their amino acids according to the codon usage frequencies of *Bacillus phage PBS2*. We note that the final product of phase two contains two levels of code: the amino acid from the translation and its exact RNA triplet. Since there are redundant triplets encoding the same amino acid, coding the triplets all together make lengthy codes. To maintain shorter codes, we made sets of triplet codes, corresponding to each unique protein amino acid.

further complicated without the knowledge of the exact organism for its frequencies.

Each amino acid was given a Huffman code to record the exact sequence of protein amino acids. For each protein, the sum of its associated triplet frequencies was used to create its Huffman code. For example, according to *Bacillus phage PBS2*, proline is created by the following triplets and their associated frequencies (i.e., probabilities): $\{P('CCA') = 0.353, P('CCT') = 0.118, P('CCC') = P('CCG') = 0.0\}$. We ranked proline's frequency by the following: $Rank(proline) = \sum_{j=0}^m freq(c_j)$ for each of the m triplets, associated to an amino acid, c . Therefore, $Rank(P) = 0.353 + 0.118 + 0 + 0 = 0.471$. Each protein amino acid was treated similarly.

7.3.2.2 Encoding Triplet Codes By Protein Amino Acid Codes

During translation, we split up CtDNA into 3-mers. Each triplet group of the sequence was converted to RNA and then translated to its protein amino acid by a

biological codon table. To make a new binary sequence, each protein amino acid was Huffman encoded according to codes prepared by the codon use frequencies of *Bacillus phage PBS2*. Simply having the knowledge of a protein amino acid is insufficient information to return to the original sequence of RNA or DNA. The exact triplet data must be used and so we kept a record of these triplets. All triplets, corresponding to each protein amino acid, were encoded as a set according to frequency data. Each triplet code must be written with the knowledge of its own protein amino acid to avoid confusion with the same arbitrary code which is associated with a difference amino acid.

The CtFinal contains two binary sequences, one for the protein amino acids and another for their triplets. To decode the triplets, the same codon-usage table must be used to reconstruct the protein and triplet codes. Since each protein is encoded using a prefix code, the string can be read in absence of code delimiters. This is also the case for the triplet codes but they can only be prefix-free codes once their corresponding protein code (their code set) is known. Therefore, to decode this string, the triplets are decoded in function of the protein codes which are read first. We include a summary of phase 2 in Figure 7.5.

Since CtProtein will likely be sent over a computer network, it would be convenient to have the data in a file-format. If we were to save the file as a simple text file, then the size of the file would soon become large but would be reduced in size when in a binary format. To prepare the binary, the binary sequences of CtFinal were split into length-8 words which were written to a binary file (CT).

7.3.3 Decryption

When the cipher text CtFinal has been decoded, the work involving Huffman codes of phase two is undone. Here, we return to the CtDNA which is the encrypted sequence from phase one. To decrypt this sequence and obtain plain text in DNA (PtDNA), we

apply the Latin square in the reverse direction using the KeyDNA. The KeyDNA and the CT sequences are aligned to locate the base pairings by position in the sequences. For each position, the KeyDNA base is found in the left-most row. The CT base is then found along this row and the PtDNA character is the entry at the top of this column. Figure 7.6 describes how this method is performed using a KeyDNA and CT base. This concludes the encryption and decryption steps of phases one and two of the sEncrypt framework.

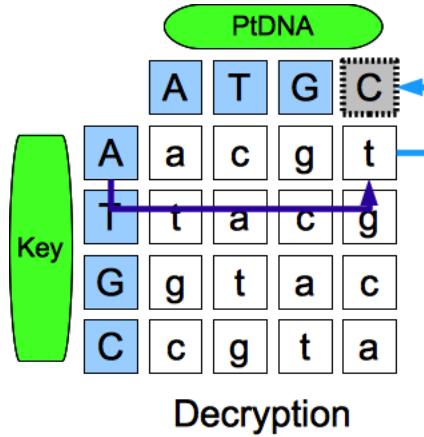


Figure 7.6: Decryption using the Latin square. Here the KeyDNA character is ‘A’ from the left column and the cipher text character is ‘t’. At the top of the column is the plain text character is ‘C’.

7.4 Results And Discussion

When encrypting data, the resulting cipher text must be made to look as random as possible to defeat any statistical tests which work to break the CT code. Below we discuss how we measured the randomness of the CT, compared to the English language PT.

7.4.1 Entropy

Shannon's entropy^[267] is a measure the predictability of information by the frequency of content occurrence. We measured the unpredictability of characters in the PT and CT text sequences using normalized entropy, bounded by zero and one for low or high randomness of sequence data, respectively.

For a set of probabilities P such that $p_i \in P$ for $0 \leq i \leq m$ and $\sum_{i=1}^m p_i = 1$, Shannon's entropy is defined as, $h(P) = -\sum_{i=1}^m p_i * \log_2 p_i$. The upper bound of entropy (i.e., complete unpredictability of the m characters) is reached when the m frequencies are identical (i.e., $p_1 = p_2 = \dots = p_m$). Mathematically, this upper bound can be written, $h_{max} = \log_2(m)$. We define normalized-entropy, $h_{norm} = h(P)/h_{max}$, which we used to compare the frequencies of characters occurrence each PT and the corresponding CT. This measurement was applied in the same style by Minosse *et al.*^[202].

To test the randomness of sEncrypt's CT data, we chose PT data which was made up of about 500 to 3000 characters of the following kinds of arbitrary text: a data table, a fragment of biological gene code, a sample of legal text (an end-user agreement), a piece of poetry (Alfred Tennyson), a news article, a piece of prose (Conan Doyle's, *The Red Headed League*), a paragraph of random words and a technical abstract (one of the papers in the references).

Figure 7.7 illustrates the entropy scores for each PT and its corresponding CT. We note that the normalized entropy for the PT of each text was between 0.65 and 1 (the upper bound). Although genetic code for making protein is highly structured^[16;239;244;317], the structure of our sample was not apparent. As we expected, the CT of each of our samples of text obtained maximum entropy scores after being processed by sEncrypt. We recall entropy scores, approaching the upper bound, imply that the individual frequencies of elemental occurrence are similar. Thanks to the Huffman encoding process, these similar elemental frequencies are

important for thwarting statistical attacks which exploit their differences.

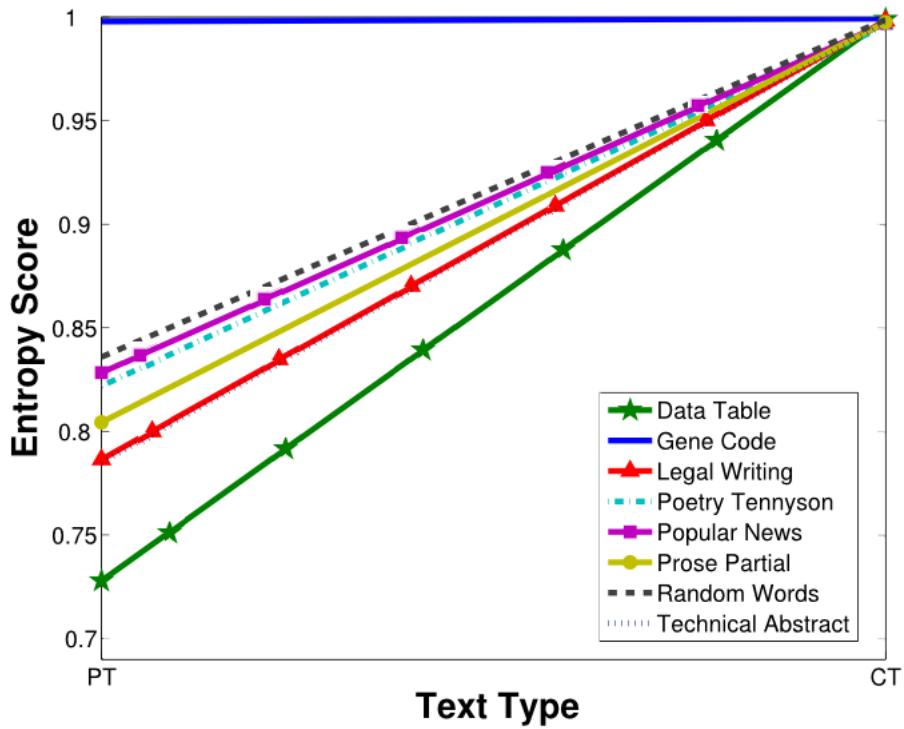


Figure 7.7: Raising entropy in CT from PT forms. We note an incline in entropy from sEncrypt. The Gene code was the only information type that already had high entropy when PT form. There was an entropy increase of all other forms of information we tested.

7.4.2 Autocorrelation Evaluation

It is undesirable to have large fragments of repetition in a cipher text file. To detect repetitive sequence data in the cipher text we performed an autocorrelation test. Here, we utilized the *dottup* tool, available from Emboss^[243] which displays a word match dot-plot of two sequences on each axis. The diagonal line indicates that the sequence characters were the same at each position. Any deviation would be expressed as a mark away from the diagonal line. In Figure 7.8 we show the output of the poetry text which has no repetition except for a few random marks. We obtained similar results from the other tests.

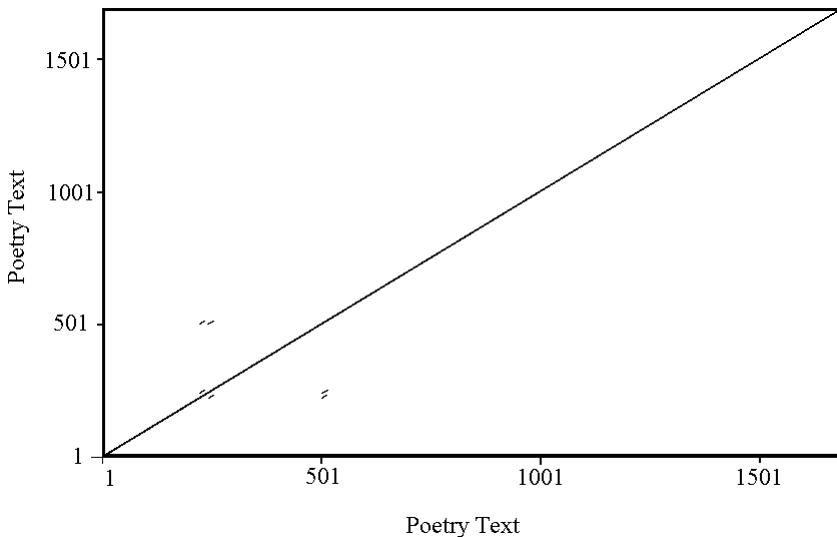


Figure 7.8: Autocorrelation of the poetry CT sequence was tested by *dottup* from Emboss. We note some tiny regions where the cipher text has repetition but these are too small to be significant.

7.4.3 General File Reductions

When we converted a sequence of English plain text to one of DNA via binary, the resulting sequence length was dramatically increased. This was because each character of the original text had to be first converted to a length-8 binary word from which pairs corresponding to DNA bases were created. There were three DNA bases for each character of English language text. In phase two, the Huffman encoding was efficiently constructed according to an arbitrarily chosen organism and additional bulk was added to the sequence information of the CtFinal file size. When the sequence data from this file was split into length-8 words to be saved to a file in binary, we noted a significant drop in size as noted in Figure 7.9. Interestingly, we note that only the genetic data had the largest CT file size as a binary file which was likely an observation connected to its high entropy value from Figure 7.7. While sequences of high entropy are generally challenging to compress,[137;163;231], our method is able to significantly reduce the size of the file's binary version.

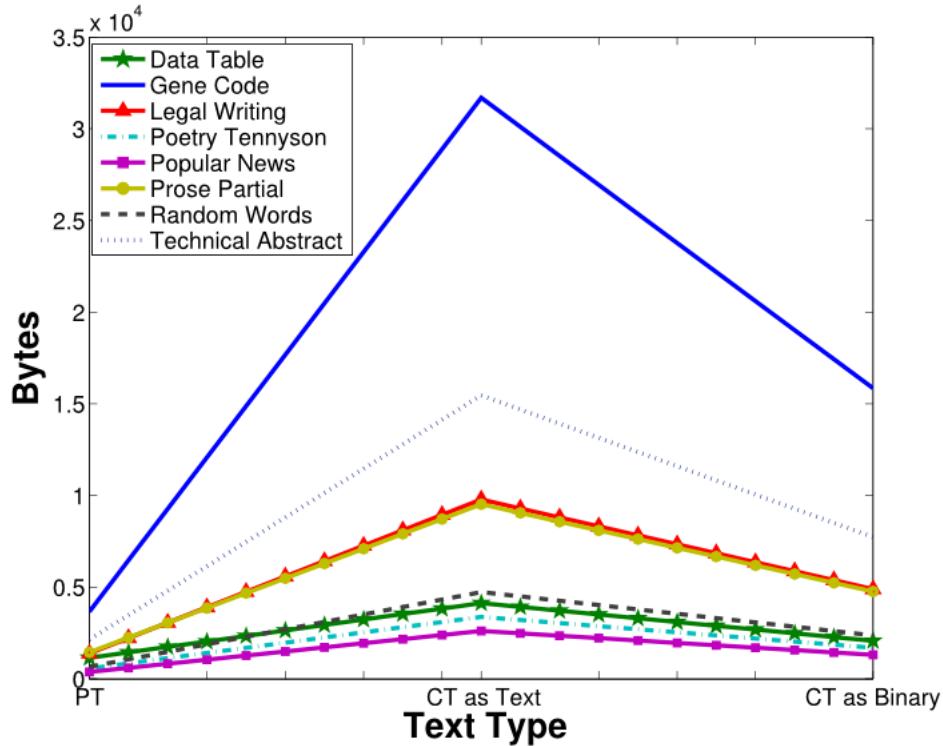


Figure 7.9: There is a significant reduction in the file size containing the CT when saved in binary.

7.5 Conclusion And Future Work

We've proposed an encryption algorithm that utilizes a structure similar to biological process of conversion of DNA to proteins. The key is formed from two pieces of information, the identity of the organism whose DNA was chosen in phase 1 and the identity of the organism whose DNA codon to amino acid mapping table was used in phase 2. As a result, we make use of the already present and accessible tera-bytes of public DNA information. The security arises from the fact that given millions of organisms in the databases, without the correct identities of the right organisms, the reverse mapping is very difficult. Further, we note that the actual encryption sequences arise from nature and not from a computer algorithm. Therefore, without the correct DNA sequence and amino acid mapping, cryptanalysis of the cipher text is very nearly impossible.

In the future, we intend to further analyze sEncrypt’s algorithmic complexity, strength of encryption and determine the amount of computational power and time necessary to break the codes. In addition, we will study the subtle changes in informational content between the stages of biologic data (PtDNA, CtDNA, and CtProtein) of different organisms. We also plan to compare our algorithm to some of the standard encryption algorithms such as those mentioned in the introduction.

7.6 Article Details

This contribution was published in the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2013.

- Oliver Bonham-Carter, Abhishek Parakh and Dhundy Bastola, “sEncrypt: An Encryption Algorithm Inspired from Biological Processes”, 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2013, IEEE, 2013.

The universe is wider than our views of it.

Henry David Thoreau

Chapter 8

Evidence Of A Pathway Of Reduction In Bacteria: Reduced Quantities Of Restriction Sites Impact tRNA Activity In A Trial Set

8.1 Abstract

Occurring naturally along the genomes of many viruses and other pathogens, short palindromic restriction sites (<14bps) are often exploited by bacterial restriction enzymes as autoimmune defenses to end pathogen threats. These motifs may also appear in the host's genome where they are methylated so as not to attract

restriction enzymes to the host's genetic material. Since these motifs in the host's genome may pose a significant danger, it is likely that their numbers have been reduced due to possible failures of methylation during evolutionary time.

These palindromes are composed of bases likely containing information relating to codons used for protein translation. If palindromes are reduced in the genome, then its sequence composition making up the codons may also be found in reduced quantities. Furthermore, during translation codons are associated with tRNAs for protein fabrication which may also occur in reduced numbers.

We suggest that a *pathway of reduction* that can be followed from the onset of these missing palindromes to the reduction (or absence) of specific tRNAs correlated to the codons from the palindromes. To create evidence for this pathway, we studied the bacterial genomes of *Bacillus subtilis*, *Escherichia coli*, *Haemophilus influenzae*, *Methanococcus jannaschii*, *Mycoplasma genitalium*, *Synechocystis sp.* and *Marchantia polymorpha*. Across these organisms, we applied statistical data from reduced palindromic populations (biological and non-relevant words) to regression models and performed an analysis of genomic tRNA presence from their compositions. In this contribution of the thesis we illustrate a pathway of reduction that extends from palindromes to tRNAs which may follow from evolutionary pressures concerning restriction site handling.

8.2 Introduction

8.2.1 Palindromes And Restriction Enzymes

A short DNA palindrome of four to eight bases (here called a *palindrome*) is a word which is equivalent to itself when in its reversed and base-complemented form. On double stranded DNA, palindromes create identical motifs when read in the 5' to 3' direction at this region. Palindromes have been observed to maintain various

roles in the regulation and cellular processes such as gene expression, replication and DNA recombination^[122]. Palindromes tend to make up the recognition sites for most restriction endonucleases in prokaryotic genomes^[218]. In this theater, they have been shown to be key actors in bacterial auto-immune defense systems. Since the genome of an invading pathogen (e.g., a virus) may likely contain many of these palindromic sites, upon its invasion of the cell, the host may deploy restriction enzymes to cleave the foreign DNA at these unique sites to end the threat.

8.2.2 Methylation And Damage Control

The methylation process, also responsible for rendering gene transcription inoperative^[182;258], is an important step for controlling the activity (e.g., the level of danger) of palindromic content in the host's genome. During methylation, the host cellular machinery adds a methyl group to encountered palindromic sites which coincidentally occur in the host's own DNA to make these motifs inert to local restriction enzymatic activity.

In addition to the natural palindromic content in the genome, it has been observed that these motifs readily occur in prophages^[170;249]. Since it is conceivable that natural restriction sites may fail to be methylated on occasion, and may instead form dangerous cleavage sites for restriction enzymes in bacteria, it is suggested in^[45] and^[71] that palindromic *avoidance* is likely to have evolved as a damage-control system. For example, in studies across several bacterial groups, Koonin *et al.*^[101] found that type II restriction-modification binding sites tended to be under-represented when compared to their statistically expected levels in bacterial genomes. Here, the authors concluded that these palindromes were generally selected-against in their sample population since the un-methylated motifs could potentially trigger self-destruction.

8.2.3 The Pathway Of Reduction

Here, we suggest that the pathway of reduction of this study is a representation of the central dogma of biology. We maintain that the pathway can be observed from these early palindromic avoidance trends, reflected in codon quantities and traced down to the pool of tRNAs. In this contribution, we show this pathway by investigating the effects of low palindromic content over codons and tRNA. We first show that many of our organisms avoid palindromes in a similar fashion and then we show that the codon and tRNA content (taken from the palindromic content) from all organisms is similar and has been reduced in tandem.

Palindrome distributions may deviate from expected levels due to other kinds of evolutionary pressures such as, alterations of gene structure for competitive advantage^[153], or were possibly affected by uneven distributions of GC content and base compositions, as noted in *Drosophila melanogaster* by Liu *et al.*^[188]. Palindromes of four to eight bases contain nested DNA triplets (e.g., codons), corresponding to specific amino acids during translation. If the palindromes are avoided, then the codons from their compositional triplets may also appear in below-expected quantities.

Codon bias is the preferential use of specific codons for translation^[233]. Biases related to palindromic content have been studied in^[92]. For instance, they have been shown to reduce the amount of close-by palindromic content in the genome, according to experimental models^[93]. Here, the author produced evidence to suggest that codon succession in *Escherichia coli* was correlated to reduced palindromic content in genes.

In *Escherichia coli* and *Saccharomyces cerevisiae* it was noted that biases were observed to exist according to gene expression levels: genes that are strongly expressed have more codon bias than genes having lower levels of expression^[22]. In addition, a strong positive correlation existed between codon usage and genomic tRNA content. For instance, in these organisms as well as *Bombyx mori* (a

multicellular organism), it was observed that certain prominent genes adapt a codon-choice pattern to better fit the tRNA pool to maintain efficient translation processes^[132]. Other studies suggest that the tRNA pool itself may be regulated to optimize gene expressions under specific growth conditions^[24;77]. In^[299], codon-pairs were observed to be biased towards codons that form a perfect Watson-Crick pairing with tRNAs. In this study the authors reported that codons favored stable interactions to weak interactions with tRNAs, that decreased the chances of mistranslations. In *Drosophila* studies, codon bias was explained by tRNA availability and was likely influenced by developmental changes in the organism^[205].

This study concerns tRNA selection and use which is relevant to cellular regulation and general cell health. In^[220] it is discussed that Trm9-catalyzed tRNA modifications promote fidelity during the translation of specific transcripts. Modification of the wobble nucleotides in tRNA by enzyme-catalyzing may impact the positioning of ribosome anticodons which creates changes in codon-dependent translation of specific transcripts. The final result of this scenario is that protein errors may be caused by the activation of unfolded protein and heat shock responses resulting in threats to cellular health. Although this study does not concern wobbling-bases where the third base is different between tRNAs, we discuss the interchanging of tRNAs where the third base is still unique between the tRNAs corresponding to the same amino acid. Here we study some of the reasons for the natural modifications of tRNAs as a result of an external impact.

8.3 Methods

8.3.1 Data Collection

To investigate the possible connection between the avoidance of short palindromes and restriction-modification systems in bacteria, Gelfand and Koonin^[101] organized

their data by palindromes of lengths-4, 5 and 6. They show that the most avoided length-4 and length-6 palindromes are likely to be recognition sites for two novel restriction-modification systems. This finding was made by their comparison of the palindromic content in *Haemophilus influenzae*, *Mycoplasma genitalium*, *Synechocystis sp.*, *Methanococcus jannaschii*, *Escherichia coli* and *Bacillus subtilis*. Since the literature notes that mitochondria and chloroplasts have significantly low counts of palindromic material and do not encode restriction-modification systems^[144], the authors used *Marchantia polymorpha* mitochondria and chloroplast genomes as a control set. The literature also notes that it is rare to find foreign DNA mixed into these genomes^[213]. Table 8.1 summarizes the list of Gelfand and Koonin’s organisms which provided the data we used in our study.

Table 8.1: The organisms, their abbreviations and the type of data used in our study. This selection of organisms is the from Gelfand and Koonin’s published results^[101]. We note that “Mito” and “Chloro” indicate “mitochondria” and “chloroplasts,” respectively.

Organism Name	Locus	Abbrev.	Material
<i>Bacillus subtilis</i>	NC_000964	BS	genome
<i>Escherichia coli</i>	NC_010498	EC	genome
<i>Haemophilus influenzae</i>	NC_000907	HI	genome
<i>Methanococcus jannaschii</i>	NC_000909	MJ	genome
<i>Mycoplasma genitalium</i>	NC_000908	MG	genome
<i>Synechocystis sp.</i>	NC_000911	Ssp	genome
<i>Marchantia polymorpha</i>	NC_001660	Mit-MP	Mito
<i>Marchantia polymorpha</i>	Z98094.1	Chlor-MP	Chloro

8.3.2 Regression Models

Stepwise regression models are powerful tools of analysis which create models where the independent and dependent sets of data share a statistically-significant likeness or growth trend. In our study, we created stepwise regression models from the avoidance data of Gelfend and Koonin’s work^[101] to ascertain which independent and dependent

sets would fit into a model and relate to a significant relationship. The sets (i.e., the avoidance data from each organism) that enter the models together may provide evidence that they avoid palindromes with some common trend. Furthermore, since these avoided palindromes are made up of DNA code which could correspond to certain translatable codons, we analyzed the distribution of amino acids found in the palindromic DNA code, to be compared to that of genomic tRNA content.

8.3.3 An Analysis Of tRNAs

To determine the amount of avoided (and possibly absent) codon content in the palindromic sequences and genomes, we analyzed the tRNA content of each organism in Table 8.1. We compared this tRNA content with the codons from the avoided palindromic code. Chloroplasts and mitochondria were excluded from this study as they do not encode restriction-modification systems^[101]. We parsed the Genbank records of each organism for its tRNA code. To get the exact tRNA codons, we employed the online Genomic tRNA Database^[53] to BLAST these fragments over bacterial code where the tRNA was known.

8.3.4 Biological Importance

Since any encountered DNA triplet may be translated into a codon, we maintain that certain codons may also be under-expressed in the genome. If there are codons missing, then there may not be much evolutionary purpose for the host to maintain the corresponding tRNA content. We shall call this phenomenon the pathway of reduction which is described by a reduction of genetic material beginning at the avoided palindromes, continuing to the codons, and finally ending at the missing tRNA content.

The relationships between the organisms of this study are shown in Figure 8.1 which we obtained from^[309]. In Gelfand and Koonin's paper, the authors organized



Figure 8.1: The taxonomy tree of the organisms of this study. This information was obtained from NCBI taxonomy.

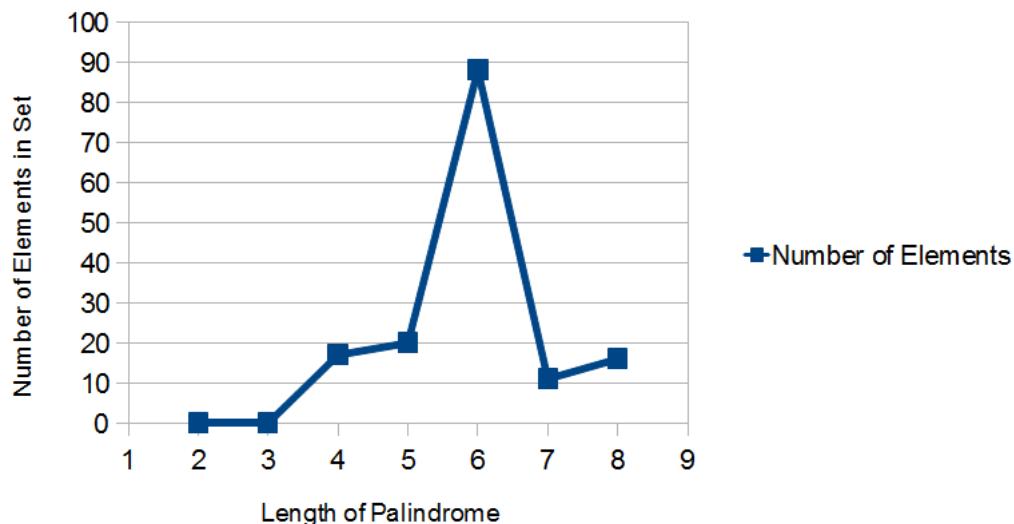


Figure 8.2: The size of each set of palindromes as arranged by word length. We note that there also are some odd-length relevant palindromes. This data taken from REBASE.

their avoidance data into three tables for length-4, 5 and 6 of biologically relevant palindromes (i.e., restriction sites). By data collected from the online restriction site database, REBASE^[248], Figure 8.2 indicates that length-6 palindromes are the most common kinds of restriction sites in bacteria. An exhaustive list of palindromes will have $n^{\frac{L_p}{2}}$ elements, where n is the size of the alphabet {A, C, G, T} and L_p is the

length of the words of the list. For example, there are $4 * 4 * 1 * 1 = 4^{\frac{4}{2}} = 4^2 = 16$ possible palindromic words of length-4. Not all of the possible palindromes make up biologically relevant restriction sites although many happen to have length-6. In our study, we selected data from biologically relevant motifs.

8.3.5 Model Building By Stepwise Regression

Linear regression models are used to describe, control and/or predict relations between variables. In our analysis, we only built models to determine which predictors are significant to the response variable. The variables in model may show that a relationship exists, but we note that this relation does not necessarily imply causation, only that there is a statistical correspondence described by the data sets of both variables.

To determine the statistically significant trends which exist across the avoidance data sets from the genomes, we built stepwise regression models to automate the process. We treated our independent variables as potential *predictors* of one dependent variable. From any set of $p - 1$ predictors, there are 2^{p-1} alternative models that can be constructed. This calculation is based on the concept that each predictor can either be included or excluded from the model. Since our pool contained eight sets of palindromic avoidance data, a total of $2^{p-1} = 2^{8-1} = 2^7 = 128$ possible models would have to be tested. Therefore, in light of the many different models which would have to be created and tested in the data (i.e. for each dependent variable, choose all other variables separately as independents to check for significance), the automated stepwise regression analysis by SPSS software suite (IBM Corp. Released 2010. IBM SPSS Statistics for Windows, Version 19.0. Armonk, NY: IBM Corp.) was desirable.

During stepwise regression model building, each dependent variable (i.e., the data set from palindromic avoidance by organism) was placed into a model with

another variable from the pool. If the corresponding *p*-value of the test-statistic was significant, then there was a presumed relation between the variables of the model. If another variable was introduced into the model having test statistic indicating a “better” relationship, then the previous variable would be excluded. After the removal of the former variable, if the new test statistic is still significant (to indicate a better model), then the new model would contain only the recently added variable. After each predictor variable has been tested by this modeling process, only the most significant predictors remain with the dependent variable. Incidentally, the search for the *best* model may sometimes be misleading since the *good* models from the pool are disqualified when a *better* model is found. Below in Table 8.2 we give the SPSS syntax code that we used in our study.

Table 8.2: Our SPSS code for stepwise regression. We did nine experiments where each organism was a *Dependent* variable to be regressed over all the others of the pool (the *Predictors* variables).

<pre> REGRESSION /DESCRIPTIVES MEAN STDDEV CORR SIG N /SELECT=id EQ 2 /MISSING LISTWISE /STATISTICS COEFF OUTS R ANOVA /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT {DependentVariable} (i.e. MitMP) /METHOD=STEPWISE {Predictors}(i.e Ssp HI EC MJ BS MG ChlorMP). </pre>

8.4 Results

8.4.1 Predictor Significance

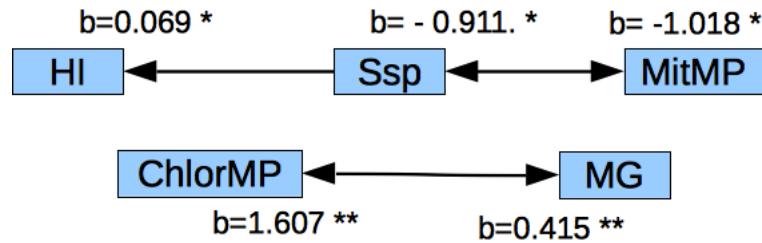


Figure 8.3: Palindrome Avoidance Data, Length-4. Our notation, * and ** specifies a significance at the $\alpha = 0.01$ and $\alpha = 0.05$ levels, respectively. The listed value is the unstandardized b coefficient. The variable names are given in Table 8.1.

The graphs created from the regression model outputs are shown in Figures 8.3 and 8.4 (upper and lower). An arrow points to a dependent variable from a predictor variable. This connection indicates that a relationship exists (i.e., a model is formed) between the variables. In particular, a dependent variable is a function of its predictor variable(s) and this extent is noted by the un-standardized coefficients (i.e., b values which function as regression weights) next to each arrow head. All exhibited variables had significant p-values. The α -value significance of each relationship is indicated by * or ** for $\alpha = 0.01$ or $\alpha = 0.05$, respectively.

To illustrate an example of how to read the graphs, we turn to Figure 8.3. The variable names are given in Table 8.1. Here we note the predictor variable, Ssp (*Synechocystis sp.*), has an arrow to its dependent variable, HI (*Haemophilus influenzae*). Next to the head of the arrow between these variables, we note, $b = 0.069 *$ to signify that this model by regression analysis has an unstandardized coefficient (b) of 0.069 and was significant at $\alpha = 0.01$ due to the single asterisk (*).

In Figures 8.3 and 8.4 (upper and lower), all model-building tests are described. In Figures 8.3 and 8.4 (lower), we note that these graphs are much smaller than the one in Figure 8.4 (upper) containing data of length-5 palindromic avoidance which

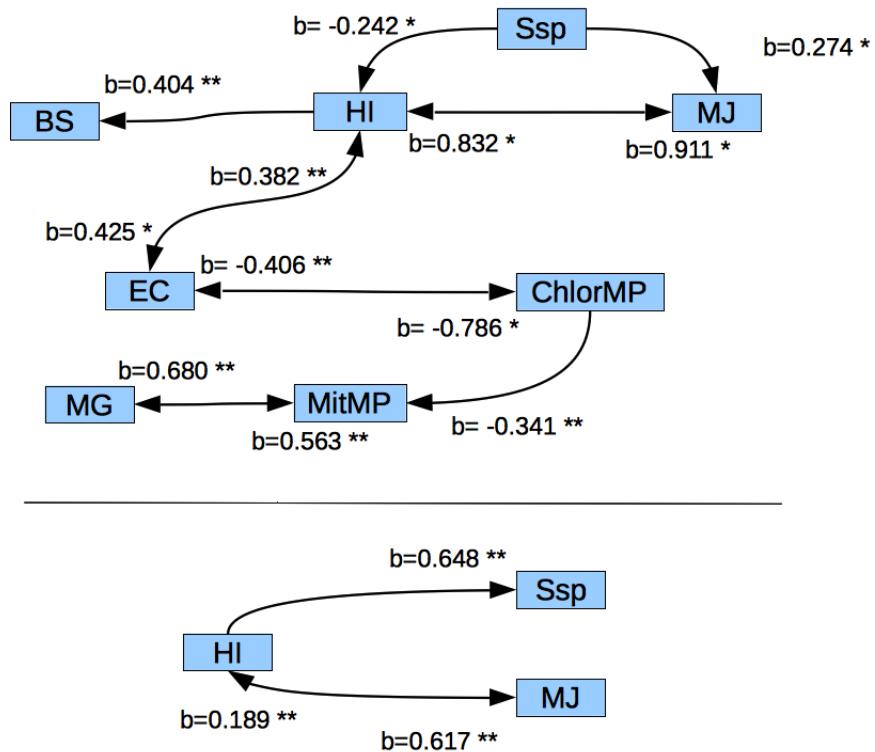


Figure 8.4: Palindrome Avoidance Data, Length-5 (upper) and Length-6 (lower). Our notation, * and ** specifies a significance at the $\alpha = 0.01$ and $\alpha = 0.05$ levels, respectively. The listed value is the unstandardized b coefficient. The variable names are given in Table 8.1.

contains predictors pointing to several different dependents simultaneously (i.e., the predictor HI points to three dependents at the same time, and the predictors EC and ChlorMP point to two dependent variables to indicate relationships). In Figure 8.3 there is one predictor, Ssp, is pointing to two dependents. In Figure 8.4 (upper), HI points to three dependents. A multi-directional arrow between two data sets may be suggestive of a much closer relationship than when there is only a unidirectional arrow. These close relationships indicate variables having very similar patterns of palindromic avoidance.

Of these significant predictors for palindromes of length-4, there were three cases of significance at the $\alpha = 0.01$ and two at the $\alpha = 0.05$ levels. Interestingly, for the length-4 palindrome data, there were three organisms having no significant predictors that were not included in any models. For the avoidance data concerning motifs of length-5 in Figure 8.4 (upper), we note that there were six predictors having a significance at $\alpha = 0.01$ and six at $\alpha = 0.05$. For palindromes of length-6, all predictors were significant at the $\alpha = 0.05$ level. The smaller the α value for the model, the *better* the evidence of the relationships between the data sets. If we ignored the b values contained in these models which often describe a positive relationship between the data sets, the fact that there are three occurrences of $\alpha = 0.01$ in Figure 8.3 and six in Figure 8.4 (upper), indicates a strong general connection in the data. All models in Figure 8.4 (lower) were not as significant and may be because length-6 palindromes often make-up the unique organismal restriction sites, seen in Figure 8.2.

8.4.2 An Analysis By Transfer RNA Composition

We formulated Table 8.3 showing which amino acids where obtained from the DNA of the avoided palindromes. We note that as the palindromes get longer, there were generally more triplets found (these numbers are given in the table) and variety (length of the table) of amino acids to be found. In the table, the gray cells indicate

Table 8.3: A complete listing all codons for amino acids (AAs) that were extracted from the DNA of the avoided palindromes (APs). The columns contain the counts of codons correlating to each extracted amino acid. The gray cells indicate that a triplet from the AP code was also missing a corresponding tRNA (listed in Table 8.4) according to our analysis using BLAST. These cells are evidence for the pathways of reduction of our study.

In Len 4 APs	Count	In Len 5 APs	Count	In Len 6 APs	Count
A	3	A	2	A	15
D	1	E	2	C	2
G	1	D	3	E	3
I	1	G	6	D	5
STOP	1	F	1	G	9
L	1	I	2	F	1
P	1	L	2	I	5
R	3	N	1	H	3
		Q	1	K	1
		P	6	STOP	2
		S	2	M	2
		R	3	L	7
		T	2	N	2
		W	1	Q	2
		V	2	P	10
				S	10
				R	17
				T	5
				W	2
				V	7
				Y	4

Table 8.4: The tRNAs which were absent across all organisms of our study according to a BLAST analysis. The above tRNAs, by extension from reduced codon content according to avoided palindromic DNA, are the end-points of the pathway of reduction.

Amino Acid	Codons	Length 4 missing codons	Length 5 missing codons	Length 6 missing codons
Cysteine (C)	2			tgt
Phenylalanine (F)	2		ttt	ttt
Leucine (L)	6	tta, ttg	tta, ttg	tta, ttg
Asparagine (N)	2		aat	aat
Serine (S)	6		tcc, tcg, tct	tcc, tcg, tct
Tryptophan (W)	1		tgg	tgg
Tyrosine (Y)	2			tat, tac

which of the extracted triplets were correlated to absent tRNA content, according to our analysis using BLAST. Additionally, Table 8.4 provides the absent tRNAs which were correlated to the amino acids. These absent tRNAs are the end-points to the pathway of reduction since they extend from avoided material.

In Table 8.3, we note the evidence of the pathway. Extracted from the length-4 palindromic content, leucine (L) has a triplet encoded by the avoided palindromic content, a missing codon (gray cell) and two absent tRNAs (noted in Table 8.4). In the larger palindromes (lengths-5 and 6), there are several cases of missing codons (gray cells) to indicate pathways of reduction. For instance, in the table for length-5 palindromes, we note that phenylalanine (F), leucine (L), asparagine (N), serine (S), and tryptophan (W) were included in the avoided palindromic DNA and had missing correlated tRNAs (noted in Table 8.4).

In the length-6 set, we note many generally higher numbers of triples extracted from the avoided palindromes. This is so because the palindromes are much longer and may also be due to there being more length-6 restriction sites than any other size (see Figure 8.2). For instance, cysteine (C), phenylalanine (F), leucine (L), asparagine (N), serine (S), Tyrosine (Y) and Tryptophan (W) showed evidence of

pathways since they have missing tRNA material which could be traced back to the avoided palindromes. Interestingly, by the codon table in any biology text book, Tryptophan (W) and Tyrosine (Y) have only one available triplet and tRNA. The triplets were in the avoided palindromic DNA and the tRNA was absent according to our BLAST analysis. Here, it would appear that both were examples of absent tRNAs extended from avoided palindromes of length 5 (only Y) and length 6 (W and Y). In another case, half of the tRNAs of serine (S) were absent in all genomes and ten of its triplets were found in the length-6 avoided palindromes.

8.4.3 Available tRNAs

In our analysis, we were able to determine the available (present) tRNAs in the genomes which code for the amino acids. For instance, across all genomes, Cycsteine (C) : {tgc}, Phenylalanine (F) : {ttc}, Leucine (L) : {cta, ctc, ctg, ctt}, Asparagine (N) : {aac}, Serine (S) : {tca, agc, agt}, Tryptophane (W) :{ \emptyset }, and Tyrosine (Y) : { \emptyset }.

8.5 Discussion

The arrows in Figures 8.3 and 8.4 (upper and lower), connect variables which have a common general avoidance of palindromic content. We note that Figure 8.4 (upper, palindromes of length 5) shows the most connections between variables. This may be because this size of palindrome (length-5) is not as avoided as are lengths-4 and 6 which tend to make up restriction sites. This lack of high avoidance, may provide more numerical data to be applied to making the graph. In Figures 8.3 (length-4) and 8.4 (lower, length-6), there may have been so much motif avoidance that there was not much data to graph.

If palindromic avoidance is a result of restriction modification pressures alone,

then odd-length motifs may not endure the same evolutionary stresses as would the motifs of even-length. For instance, an even-length palindrome is the same motif on both strands at the same location of two complementary DNA sequences. This is not the case for the odd-length palindromes which are actually not words that are equivalent to themselves when in reversed and complemented forms. On complementary strands, the odd-length palindrome on one strand is a different word by one base. Here, we claim that there are restriction modification pressures (and others that are unknown) which may limit the population sizes of the palindromes and bring avoidance qualities to the genome.

Palindromes contain codons embedded in their sequences which correlate to unique tRNAs. We noted that there exist connections between avoided palindromic content and these corresponding tRNAs which make up the pathways of reduction. A good example of this pathway concerns Leucine that can be fabricated by six different codons according to a codon table. In all three sets in Tables 8.3 and 8.4, we noted that the palindromes contained codons for constructing leucine. Two tRNAs (TTA and TTG) were consistently absent from the entire set of organisms to create a pathway of reduction. Consulting Subsection 8.4.3 (*Available tRNAs*) we note that neither of these codons was present in the study's genomes. This indicates a carry-down of avoidance from the palindromes to tRNAs.

Tryptophan has only one correlating codon TGG which was found once and twice in the length-5 (CCTGG) and length-6 (CCATGG and TGGCCA) data, respectively. This tRNA was completely absent (a pathway) in both of these sets, but not in the length-4 set. Serine also indicated a pathway existence since three of its codons were found in the palindromes and only three tRNAs were present in the genomes.

8.6 Conclusions

After we intersected all the codon lists taken from avoided palindromes over our data sets, we discovered that there were several codons and tRNAs which were omitted across all genomes. For instance, two of the six codons for leucine (L) were absent in all genomes, however all palindrome sets (lengths 4, 5 and 6) contained DNA code which was associated with this amino acid. Similarly, three of the six serine (S) codons were encoded by the palindromes (including `tcc`) which has no correlating tRNA. This finding marks another pathway of reduction for this amino acid. Most notable, the pathway can be observed in the tryptophan and tyrosine for which their single codons are encoded in the palindromes and their tRNAs are absent in our genome set.

From the above discussion, we can see that a pathway exists which begins at avoided palindromic code and continues past the lost codons to their absent but correlating tRNAs. We note that the pathway suggests that the avoided palindromes of Koonin and Gelfand's study is actually a much deeper study than previously thought. As with any scientific research project, there are usually more questions than conclusions at the end. In this study, we asked whether the missing palindromes were only a part of the avoidance question. We concluded that there was a string of missing genetic material down to the tRNA level which we called the pathway of reduction.

On a final note, organisms are often unable to manufacture all the amino acids that they require to live. With this analysis and perhaps in lieu of other forms of study, we may be able to predict their nutritional requirements to be able to restrict or even prevent growth of organisms causing serious health concerns (bacteria during the course of an infection, for example).

8.7 Article Details

This contribution was published in Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics, ACM, 2013.

- Oliver Bonham-Carter, Lotfollah Najjar and Dhundy Bastola, “Evidence of a Pathway of Reduction in Bacteria: Reduced Quantities of Restriction Sites Impact tRNA Activity in a Trial Set.”, Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics, ACM, 2013.

We shall meet in the place where
there is no darkness.

George Orwell, 1984

Chapter 9

Evidence Of Post-translational Modification Bias Extracted From The tRNA And Corresponding Amino Acid Interplay Across A Set Of Diverse Organisms

9.1 Abstract

A post-translational modification (PTM) describes a form of biosynthesis for the task of initializing proteins for specific functions. PTMs are complexes which are involved in developing or customizing proteins to increase their functional diversity. In times of protein stress, PTMs may be involved in altering protein structures to

allow for better chances of survival. Once the stress-condition has elapsed, PTMs are able to transform the protein's structure back to its original form for the continued survival of the protein. PTMs are not applied uniformly across organismal proteins and differing PTM preferences and usages may often exist between proteins of the same organism. Here, we study the frequency of factors (PTM predominance and their associated active sites, tRNAs and amino acids) which likely influence a PTM bias. We extract and study these factor frequencies across both mitochondrial (Mt) and non-Mt proteins of nine diverse organisms (closely following two, *Arabidopsis thaliana* and *Caenorhabditis elegans*, due to space limitations) to illustrate their remarkable differences which may strongly influence natural PTM selection. By the work in this contribution of the thesis, we offer evidence to argue that this PTM bias may be the result of these factors which combine in a poorly understood system to affect and control PTM interactions. Our analysis is made up of an application of frequency information concerning PTMs, active sites, tRNA and amino acids and is used to create network models for the clear visualization of its mechanisms for this PTM natural selection.

9.2 Introduction

9.2.1 PTM Bias

It is extremely likely that all proteins in nature undergo some level of post-translational modification (PTM) for a structural, and therefore functional, alteration. Such an alteration may occur where a specific amino acid or active site is triggered by a complex to induce changes in enzymatic activity, localization, or to be marked for degradation, as noted in the case of a failing protein^[32]. Proteins may also be altered without the aid of enzymes such as in deamidation, glycation and isomerisation.

The modification may be necessary for the survival of the protein during times of stress. To add functional diversity and adaption to their alternative environments^[271], proteins may respond to stresses by a transformation of structure and hence, function. For instance, a protein stress may result from an event or treatment which leads to failure when the protein is forced to sustain duties under unnatural circumstances. Exposure to severe heat or lack of moisture, for example, may cripple cellular proteins when they are unused to these conditions. PTMs imply a transformation by biosynthesis to initialize proteins for specific functions and allow for the regulation of protein activity. PTMs may be employed by cellular mechanisms to quickly alter stressed proteins (or proteins which are unable to handle their tasks) to enable them to maintain their duties and sustain life under diverse environmental conditions. This immediate change is thought to allow for rapid adaptation since a new protein will not have to be regrown (from DNA) to cope with the new environment. In addition, when the stress is removed, the protein may undergo another modification to restore it to its original form^[20]. During severe types of stress, interestingly, it may be likely that PTMs enjoy an interplay with other post-transcriptional regulatory mechanisms. This may help to explain why there are so many different types of PTMs in nature – about 87308 different PTMs have been experimentally identified in^[148].

Since countless and diverse proteins depend on PTMs such as *acetylation*, *glycosylation* and *phosphorylation*, they are likely to be ubiquitous across all domains of life and are also likely to share a universal common ancestor. In fact, *lysine acetylation* is likely used for much gene expression regulation and may be evolutionarily conserved from bacteria to mammals^[326]. Furthermore, a method to quickly adapt protein complexes is thought to be a major aid to large-scale survival and evolution.

The protein sequences of our study were downloaded from the Uniprot

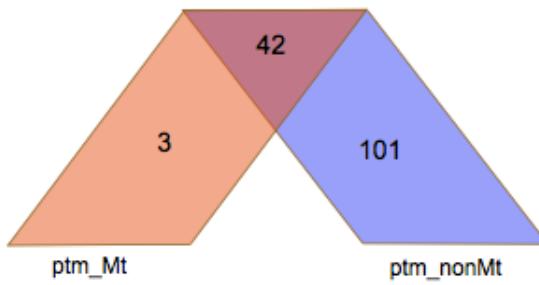


Figure 9.1: A comparison between the number of PTMs between our Mt and non-Mt sequence data.

Knowledge Base^[9]. At the time of our study, our downloaded set was the most current available and was dated by Uniprot: 11 June, 2014. The proteins were divided into mitochondrial (Mt) and non-Mt sets which we parsed to gain an idea of their populations of unique PTMs. In Figure 11.1, we applied this PTM data to the Venn diagram tool at bioinformatics.psb.ugent.be/webtools/Venn/. We note, there are only three PTMs which are unique to Mt protein (*Tele-8alpha-FAD histidine*, *N6-lipoyllysine*, *Methylhistidine*). Since the organization of the Mt genome is highly conserved in insects, as in most other bilateral animals^[37;168], this small listing of unique PTMs in Mt may be explained by its conserved nature. The data of our study is shown in Table 11.1 where we indicate the numbers of proteins analyzed and the two most frequently occurring PTMs we encountered in the combined Mt and non-Mt proteins. We noted that the ranking of both PTMs changed. Here we show that PTMs do not always have the same degree of usage across organisms. In our arbitrarily chosen set of organisms, we build on this bias to show some of the other factors (PTM predominance and their associated active sites, tRNAs and amino acids) which work to affect the selection of PTMs for interaction with organismal proteins.

Table 9.1: Diverse organisms of the study. The two PTMs indicated are the two most frequently encountered (from both Mt and non-Mt sequences simultaneously).

Organism	N. Seqs.	Top PTMs
Mustard Plant <i>Arabidopsis thaliana</i>	707	<i>Glycosylation</i> <i>Phosphoserine</i>
Nematode worm <i>Caenorhabditis elegans</i>	199	<i>Glycosylation</i> <i>Lipidation</i>
Domestic Dog <i>Canis familiaris</i>	60	<i>Glycosylation</i> <i>Phosphoserine</i>
Zebrafish <i>Danio rerio</i>	202	<i>Glycosylation</i> <i>Phosphoserine</i>
Human <i>Homo sapiens</i>	1027	<i>Phosphoserine</i> <i>Glycosylation</i>
House Mouse <i>Mus musculus</i>	973	<i>Phosphoserine</i> <i>Glycosylation</i>
European rabbit <i>Oryctolagus cuniculus</i>	46	<i>Glycosylation</i> <i>Phosphoserine</i>
Norway Rat <i>Rattus norvegicus</i>	571	<i>Glycosylation</i> <i>Phosphoserine</i>
Bakers Yeast <i>Saccharomyces cerevisiae</i>	1056	<i>Phosphoserine</i> <i>Glycosylation</i>

9.2.2 tRNA Bias

In a general sense, each amino acid has a corresponding tRNA which enables their incorporation into a protein complex. Going back further, each tRNA is signaled at the DNA level by a unique codon and we note from the codon table that there is often a multitude of codons (each interacting with a unique tRNA) to place the same particular amino acid. For instance, lysine (K) has two DNA codons **AAA** and **AAG**, each of which interact with a unique tRNA, correlated with the same amino acid (K). Since there is a selection of codons available for many of the amino acids, it is not surprising that over evolutionary time, an organism-specific tRNA preference (or bias) has been introduced to prefer one codon over another for a specific amino acid.

In the 1980's, codon bias was explored in yeast in [22]. In early works of Ikemura *et al.*, it was observed that tRNA usage varied between organisms [131;132;141]. For

instance, experiments over *E. coli* confirmed an adaptation to the codon pool in nature and inspired the plausible explanation of the presence of codon usage bias in highly expressed genes.

To address this phenomenon, it was suggested more recently by^[270] that codon usage patterns may likely be influenced by natural selection for particular codons that are translated more accurately and/or efficiently in bacteria. To explore this phenomenon more deeply, the authors introduced a population genetics-based model for quantifying how natural selection may play a role and concluded that species exposed to selection for rapid growth have more rRNA operons, more tRNA genes and a highly selected codon usage bias. Another explanation for this tRNA bias may stem from the contributions of mutations, drifts and other general factors of natural selection. Although variation in mutational bias is very likely a strong influence of codon usage, translational selection may act as a weak additional factor to influence synonymous codon usage^[74]. In^[31], it was suggested that a bias may have resulted from avoided DNA motifs, resembling restriction enzymes, which could otherwise contribute negatively to the genome.

9.2.3 Amino Acid Bias

A prominent study of nucleotides and amino acid bias was performed by^[276] which surveyed the genes in 21 completely sequenced eubacterial and archaeal genomes, as well as, the *Saccharomyces cerevisiae* genome and two *Plasmodium falciparum* chromosomes. In their study, the authors observed a nucleotide bias which encoded biased proteins on a genome-wide scale and noted a positive correlation between the degree of amino acid bias and the magnitude of protein sequence divergence. Furthermore, due to the imposed selective constraints of inhabiting harsh environments, in^[277], an amino acid bias (among other factors) was found to aid in the survival of thermophiles in high temperature environments.

An amino acid bias may impact sequence diversity which is necessary for protein folding, function and evolution. Interestingly, in^[308], it was noted that natural protein sequences are statistically indistinguishable from strings of amino acids which have been randomly placed, with the exception of larger additions of leucine (most abundant) and lesser additions of tryptophan and cysteine (least abundant). Although randomly placed, amino acid abundances are fairly well conserved across organisms to indicate a mechanism to explain this seemingly random nature. In^[160] a simple model is presented to help explain these relative abundances of amino acids across a diverse set of proteomes. The model follows a premise that there may be a trade-off between the minimization of protein synthesis cost and the degree of achieved protein sequence diversity in natural proteomes. In their study, the authors suggest that this cost, derived from amino acid decay during protein modifications, maintains a particular distribution which is economical (implying biological stability) for an often changing biological setting. Therefore, to maximize sequence entropy, for the production of proteins that ameliorate survival rates, an amino acid bias may be important for evolution.

9.3 Methods

9.3.1 Diverse Organisms

The organisms of our study were diverse and represented a wide spectrum of biology^[309] as shown in the taxonomy tree of Figure 9.2. We focus only on the results and graphs for the data originating from *Arabidopsis thaliana* and *Caenorhabditis elegans*.

The protein data was downloaded from the Uniprot Knowledge Base^[9], dated 11 June, 2014. Having both mitochondrial (Mt) and non-Mt origins (according to Uniprot), were divided the protein data into these two sets. Unlike the Mt genome

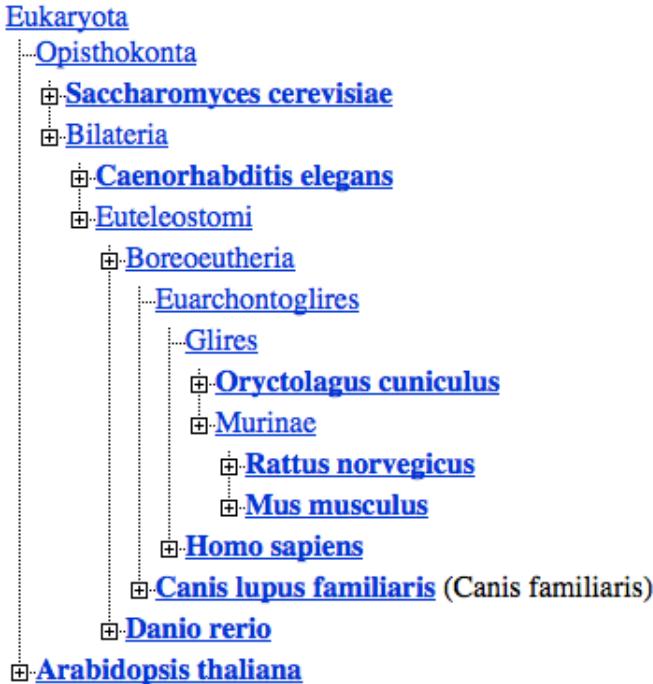


Figure 9.2: The taxonomy tree of our study’s data.

which may be highly conserved, the non-Mt protein may be diverse in nature and could be more revealing of bias in nature. In addition to this sequence data, the downloaded data also contained details of the PTMs that have been observed to interact with particular proteins in an organism. The PTMs of our study are those that interact with a single amino acid (a length-1 motif) called an *active site* and so all PTM information we extracted concerned only these kinds of PTMs, although others exist. The exact amino acid of the active site was found by searching the protein sequence for the location of the active site.

The data that we used to derive our statistical information may have had incomplete references due to the difficult nature of extracting PTM information from physical protein samples in a wet lab setting. In light of this limitation, we believe that this study is accurate and is able to illustrate the factors behind PTM bias. In Table 11.1, we list the organisms and the numbers of protein sequences that we parsed (by our in-house program written in Python) to get frequency

information. Also in this table, we provide the top two most frequently occurring PTMs (first and second rankings) that were found across each organismal protein set. These PTMs are often the same across all organisms (*glycosylation*, *phosphoserine* and *lipidation*), yet their ranks differ between sets to suggest a subtle PTM bias.

9.3.2 Computing Frequencies

A frequency analysis is well suited for comparing large datasets^[28;30]. To calculate the PTM frequencies, we parsed the Uniprot protein records concerning protein modification. Nearly all protein records contained information about PTM usage and its active site. To find the PTM information, we isolated the *amino acid modifications* and the modified residue information in each protein record. According to Uniprot, *glycosylation* was listed as *carbohyd* and so each encountered *carbohyd* was counted accordingly. A count of PTM type was created for each organism and the PTM frequencies were calculated from this information in each of the Mt and non-Mt protein datasets for each organism. We used this information to populate Table 11.1.

The frequency calculations for PTMs, active sites and the amino acids will now be discussed. Across each organism j , the frequency of a particular $PTM_{(i,j)}$ and its associated active site, $actSite_{(i,j)}$, were calculated by Equations 11.1 and 11.2, respectively. We note the use of the *count()* function which determines the number of occurrences of the element in the current dataset. Across all PTMs of organism j , the frequency of a particular $PTM_{i,j}$ may be found by the following.

$$freq(PTM_{(i,j)}) = \frac{count(PTM_{(i,j)})}{\sum_{i=1}^{N_{(PTMs)}} count(PTM_{(i,j)})} \quad (9.1)$$

Across all active sites found associated with the PTMs of organism j , the frequency

of a particular site, $actSite_{i,j}$, may be found by the following.

$$freq(actSite_{(i,j)}) = \frac{count(actSite_{(i,j)})}{\sum_{i=1}^{N_{(actSites)}} count(actSite_{(i,j)})} \quad (9.2)$$

The amino acid frequency magnitudes in the protein sequences of organism j (for which PTM details were available by Uniprot), were calculated using the following equation. We note that this frequency is normalized by the concatenated lengths of protein sequences of the organism.

$$freq(SeqASite_{(i,j)}) = \frac{count(aminoAcid_{(i,j)})}{|\sum_{i=1}^{N_{Proteins}} Seq_{(i,j)}|} \quad (9.3)$$

9.3.3 Network Models

In this work, we applied normalized frequency information to network analysis as it allowed us to conveniently compare magnitudes (and perhaps degrees of correlation) between the variables of our study (i.e., PTMs, active sites and their associated tRNAs across our organismal protein data). Our network analysis yielded network models which clearly describe the relationships and connections between these three elements in both the Mt and non-Mt protein data. For each organism, we show the network models which illustrate the prominence of PTM, tRNA frequencies, as well as, the magnitudes of this PTM-tRNA pairing across the entire set of proteins.

Reading the Network Models: In Figures 9.7 to 9.10, the left-side nodes are sized represent the PTM frequency magnitudes (Equation 11.1). The edges of the nodes (both thin to thick), represent the frequency magnitudes of the PTM active sites (Equation 11.2). The right side nodes represent the tRNAs and their typeface sizes represent the tRNA frequencies which were taken directly from the Codon Usage Database (<http://www.kazusa.or.jp/codon/>^[206]) which offers codon usage information by organism. Since codon usage is closely connected to tRNA

usage, this data is well suited for our purposes. Larger frequency magnitudes of tRNA are indicated by the larger typefaces.

Heatmaps: Later on, we will introduce the heatmaps of Figures 9.5 and 9.6 which describe the frequencies of the amino acids making up the active sites of the protein sequence data. Equation 11.3 was applied to prepare these heatmaps.

9.4 Results and Discussion

In Table 11.1, we provided a listing of organisms which are a part of our study. We listed the number of protein sequences that we analyzed. We also listed the top two most frequently occurring PTMs for each organism (combined Mt and non-Mt data). This information was gained by a simple tally of PTMs concerned with the proteins of each organism and the highest frequencies indicate PTMs which are likely playing big roles in organismal protein survival.

Interestingly, our preliminary study noted an apparent preference for individual PTMs across the organisms. For instance, we noted that although *glycosylation* and *phosphoserine* were popular PTMs for many organisms, they do not appear to always achieve the same first and second rankings in the organisms. In other words, PTM rankings were often different between organisms. In addition, we noticed that *Caenorhabditis elegans* was the only organism of our set which had a high frequency for *lipidation*.

Across each organism, proteins having PTM information from Uniprot were analyzed for their PTM frequencies (we attributed general higher frequencies to an increased prominence). In Figures 9.3 and 9.4, we summarize PTM frequencies for Mt and non-Mt data. The organisms are displayed along the bottom and along the left, we note the PTMs found in the organismal proteins. The displayed frequencies met our criteria of having a value of at least 0.1. Although our analysis found many

other PTMs involved with each organism having low frequencies, a threshold was used to help focus on the highly occurring data. The shaded cells visually indicate the frequency magnitudes that are defined in the legends. The degree of shade has been normalized between graphs and so a direct comparison is possible.

Cells of equal shades show similar PTM frequency values between organisms. For instance, shown in Figure 9.3, *phosphoserine* appeared to have a relevance to most organisms. This suggested that this PTM may have had a very early beginning in evolutionary history since it is so well conserved in the Mt genome (that is already highly conserved). In Figure 9.4, we noted that phosphoserine was also a prominent PTM for nearly all organismal non-Mt proteins. In addition to *phosphoserine*, *glycosylation* was also found to play a prominent role in both the Mt and non-Mt protein datasets. This finding corroborates the results of Table 11.1 where nearly all organisms were closely associated with these two PTMs.

In the heatmaps in Figures 9.5 (Mt) and 9.6 (non-Mt), we present the amino acid frequencies across all the organisms. The frequencies were calculated using Equation 11.3 and describe the prominence of the amino acids in the protein sequences. Since active sites are a subset of the set of amino acids, we may determine that their reduced frequencies could be linked to reduced tRNA frequencies. We note that cells of equivalent shades show similar values between organisms. Across all organisms, we only focused on amino acids having frequency values which were greater than 0.1. We used these values to determine the impact of tRNA frequencies on the active sites. In general, we noted from the heatmaps that the amino acid frequencies for lysine (K) and serine (S) indicated a predominance across all organisms. This predominance was also noted in our network models which are discussed below. In Mt proteins, Figure 9.5 illustrates that many fewer amino acids had elevated frequencies than those of the non-Mt set of Figure 9.6.

The larger number of elevated amino acids in non-Mt appeared to support the

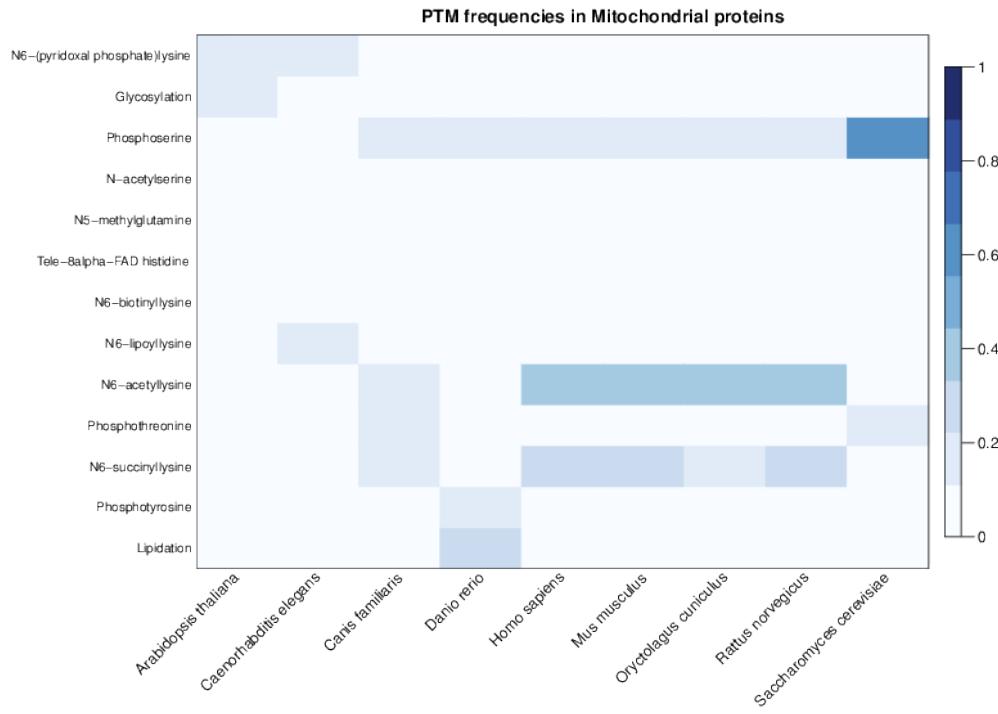


Figure 9.3: A heatmap of PTM frequencies in Mt protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.

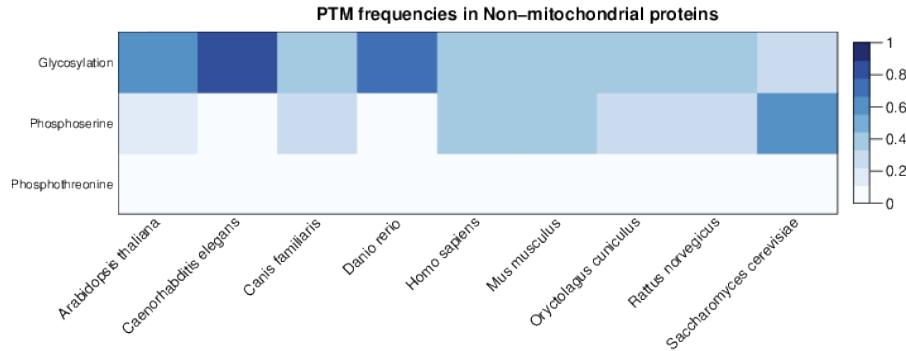


Figure 9.4: A heatmap of PTM frequencies in non Mt protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.

observation of PTMs which interact with several unique active sites. For example, *glycosylation* appeared to interact with several unique active sites in non-Mt which was not often observed in Mt (this PTM usually interacted with only one active site

in Mt). In Figure 9.4, we noted that *glycosylation* was a more frequently occurring PTM in the non-Mt set but this was not the case in the Mt set of Figure 9.3. We suggest that this discrepancy may be due to reduced amino acid quantities which we observed in the Mt protein. These results are supported in our discussion of network models below.

The organisms which are closely related, shown in Table 11.1, display some similarity according to our heatmaps. For instance in Figure 9.5 and 9.6, we note that *Homo sapiens*, *Mus musculus*, *Oryctolagus cuniculus* and *Rattus norvegicus* have similar shades of each amino acid. This likely describes an evolutionary trait.

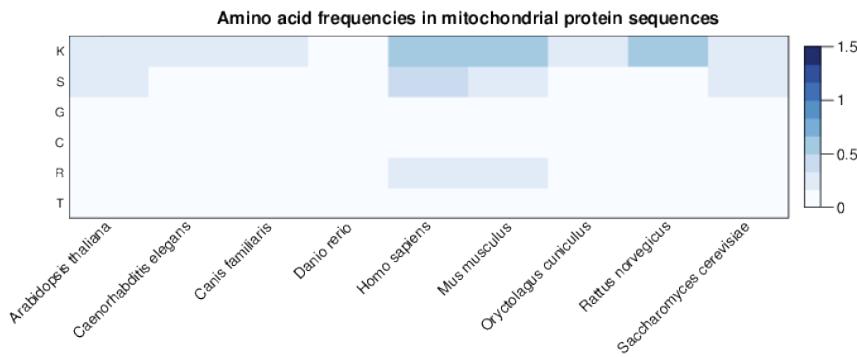


Figure 9.5: Mt amino acid frequency heatmaps prepared using Equation 11.3. Only the active site frequencies greater than 0.1 are displayed.

Network models help analyze complex relationships between these entities. We now introduce our network models of Figures 9.7 to 9.10 describing the frequencies of PTMs, their active sites and the associated tRNAs across all proteins by organism. Due to space limitations, we are only including models of *Arabidopsis thaliana* and *Caenorhabditis elegans* for Mt and non-Mt protein.

For each protein set (Mt and non-Mt), we have two networks for each organism. Each network contains two types of nodes: PTMs (left) and tRNA (right). Each PTM of our study physically interacts with protein at a specific active site. The edge between both nodes links a PTM to the tRNA which is responsible for the active site.

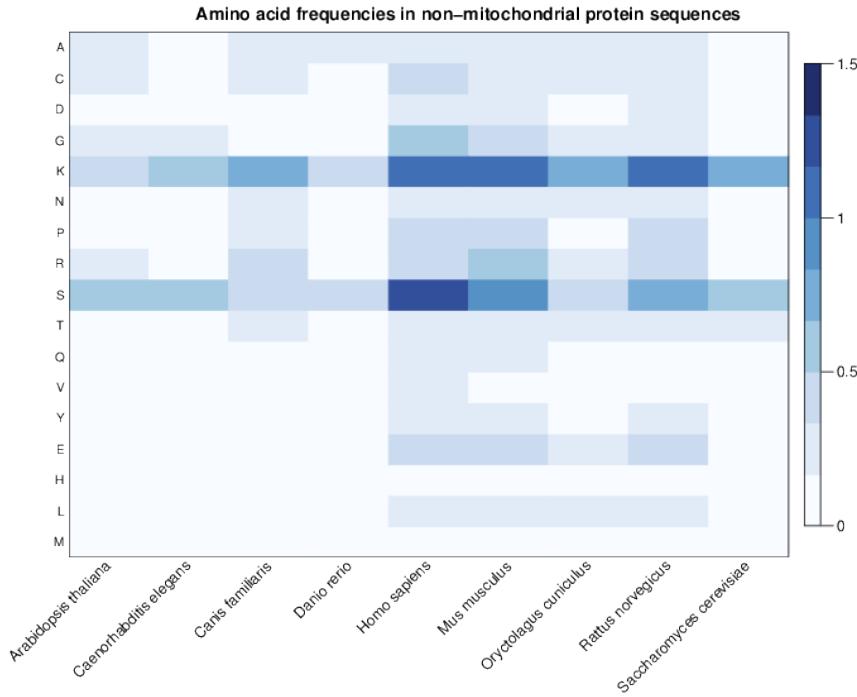


Figure 9.6: Non-Mt amino acid site frequency heatmaps prepared using Equation 11.3. Only the active site frequencies greater than 0.1 are displayed.

In this relationship, we note which active sites play larger roles in PTM mechanics, by organism. The thickness of the edge represents how often a PTM interacts with a particular active site.

Typically, in the Mt models, a particular PTM interacts only with a unique active site and this may often be noted in both the Mt and non-Mt datasets. For instance, in *Arabidopsis thaliana* in Figures 9.7 and 9.8, we noted that PTMs such as *N6-biotinyllysine*, and *N6-Carboxylysine* (as well as others) interact uniquely with lysine (K). However, in the non-Mt network model, we noted that there were other PTMs (not found in the Mt data) which interact simultaneously with several different active sites. In Figure 9.8, we noted that *lipidation* commonly interacted with Q (glutamine), N (asparagine), D (aspartic acid) and C (cysteine) in the same organism. Observations where a particular PTM interacts with multiple active sites in non- Mt sets (generally not present in the Mt sets) may be found in the other

figures as well. These PTMs of the non-Mt sets, which are able to interact with a multitude of different active sites, suggest that there may be a protein-level bias of PTM. This bias may affect the active site interaction pathways or mechanisms which do not exist in the Mt data. Here we imply that different active sites may be biochemically similar across different proteins of the same organism which enable this phenomenon^[272].

It is well-known that Mt export much of their protein production to their host cells. In fact, Mt are only able to produce a very limited number of their own proteins which may help to explain why there were so few unique Mt PTMs compared to those of the non-Mt. This discrepancy may also be attributed to the highly conserved nature of Mt genetics requiring only a few PTMs for its regulation and perhaps stress response. On the other hand, the host responsible for creating most of the proteins in the cell, may require a wider variety of PTMs to support the diverse functionality of its proteins. Furthermore, since Mt live inside the cell where there are perhaps fewer stresses, they may not need as many PTMs to modify proteins for stress responses. The cell, on the other hand, under the constant threat of stress, may rely more on its PTMs for its response tactics.

In addition to having many more PTMs in the non-Mt sets, there were more active sites available than in Mt, as mentioned earlier. We have also discussed that some PTMs are able to interact with a multitude of different active sites. For example, *glycosylation* is a critical function of the biosynthetic secretory pathway in the endoplasmic reticulum and Golgi apparatus. Nearly half of all cellular proteins typically expressed in the cell may undergo such a modification by this PTM during their lifetime. We note that *glycosylation* is active in both the MT and non-Mt sets of the following organisms: *Arabidopsis thaliana*, *Canis familiaris*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus* and *Saccharomyces cerevisiae*, (*Arabidopsis thaliana*, Figures 9.7, 9.8 and *Caenorhabditis elegans*, Figures 9.9, 9.10).

In these Mt sets, this PTM typically interacts with only one active site (except for *Homo sapiens*, *Mus musculus* and *Rattus norvegicus*). However, in the non-Mt data, this PTM generally interacts with several different active sites of the same organism. Since nearly half of all cellular proteins are thought to undergo such a modification during their lifetime, interactions with different active sites may be necessary to accommodate all the diverse protein functionalities. The observation of the use of different active sites describes a clear bias.

Also in *Arabidopsis thaliana* of Figures 9.7 and 9.8, a PTM may appear to be more active in Mt protein than in non-Mt protein. For example, *glycosylation*, *N6-(pyridoxal phosphate)lysine* and *N6-lipoylllysine* appear to be prominent PTMs in Mt but not in non-Mt. This may be explained by the density of non-Mt networks that hosts many other PTMs which may be able to assume the same PTM duties in Mt. In both sets, *phosphoserine* appears to be prominent which may indicate that this PTM is useful to both by performing some unique task. For instance, serine (and threonine) *phosphorylation* is known to directly result in the formation of multimolecular signaling complexes^[319]. *Phosphoserine* may be involved with the formation of these specific signaling complexes which are later controlled by kinases and binding modules for regulation and thus, is equally important in both.

9.4.1 Notable PTMs

The most frequently occurring PTM in our network models was *phosphoserine* among both the Mt and the non-Mt proteins. This particular PTM represents the phosphorylation of serine base in a protein's amino acid sequence and is one of the most common modifications to proteins that can alter functionality. Among other sites such as threonine, tyrosine and histidine residues, serine is the most common type of *phosphorylation*. *Serine phosphorylation* like other *phosphorylations* can cause structural changes in proteins to activate or deactivate them.

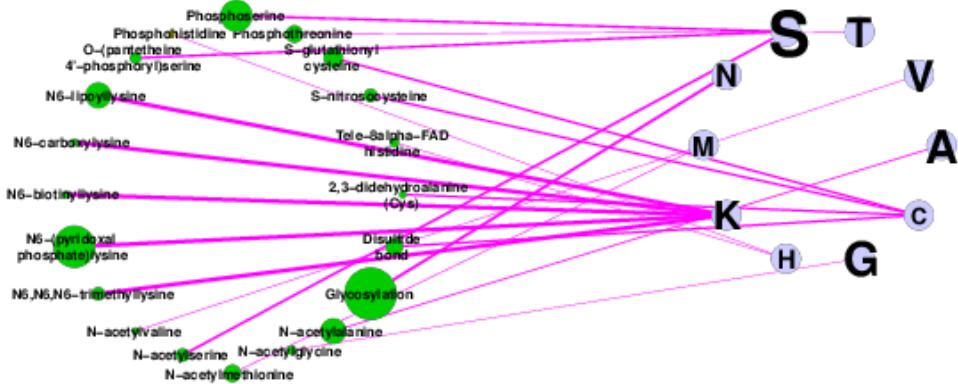


Figure 9.7: The *Arabidopsis thaliana* Mt Network: PTM frequencies (left) and associated active sites by tRNA frequencies (right). These models tend to be *less dense* than those of non-Mt proteins.

Glycosylation results when a carbohydrate molecule is added to hydroxyl or other molecules in a protein. The majority of the proteins synthesized in the rough endoplasmic reticulum undergo *glycosylation*, hence most of the *glycosylation* is found in non-Mt proteins. However, in plants, Mt have an additional function of photo respiration. This discrepancy may explain why we noted *glycosylation* in Mt proteins in *Arabidopsis thaliana*.

Acetyllysine is another important PTM that adds an acetyl group to a lysine residue in proteins. The *acetylation* of lysine (K) residue is considered as a regulating mechanism for various epigenetic factors^[41;118]. We observed higher amount of *N6-acetyllysine* in Mt proteins and was conserved across *Homo sapiens*, *Mus musculus*, *Oryctolagus cuniculus* and *Rattus norvegicus*.

The PTM, *succinylation*, refers to addition of succinyl group to a lysine residue in proteins. It is one of the newest PTMs discovered in last few years^[329]. Much of its role is unknown, however it is believed to cause a significant structural change in the protein as this PTM changes a positively charged lysine to a negative charge^[318]. Like *acetylation* of lysine, this PTM is also seen in higher abundance in Mt proteins and is conserved across higher order organisms.

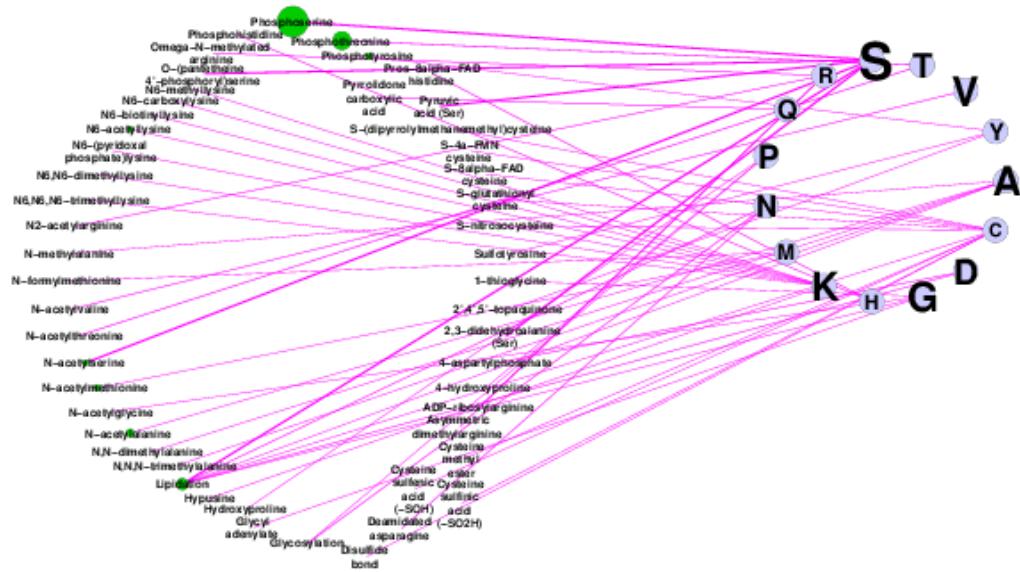


Figure 9.8: The *Arabidopsis thaliana* non-Mt Network: PTM frequencies (left) and associated active sites by tRNA frequencies (right). These models tend to be *more dense* than those of non-Mt proteins.

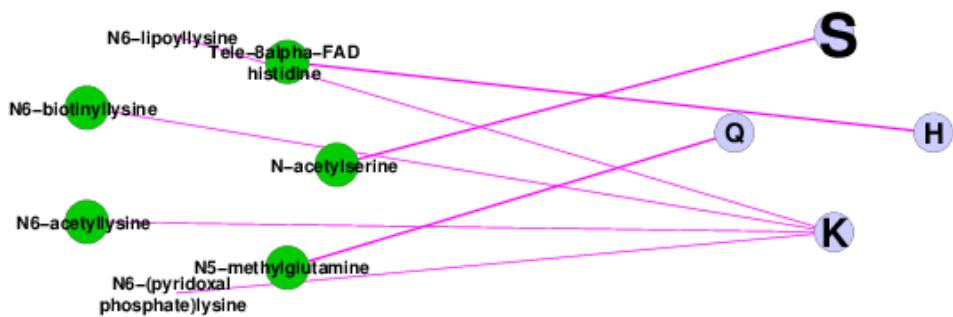


Figure 9.9: The *Caenorhabditis elegans* Mt Network: PTM frequencies (left) and associated active sites by tRNA frequencies (right).

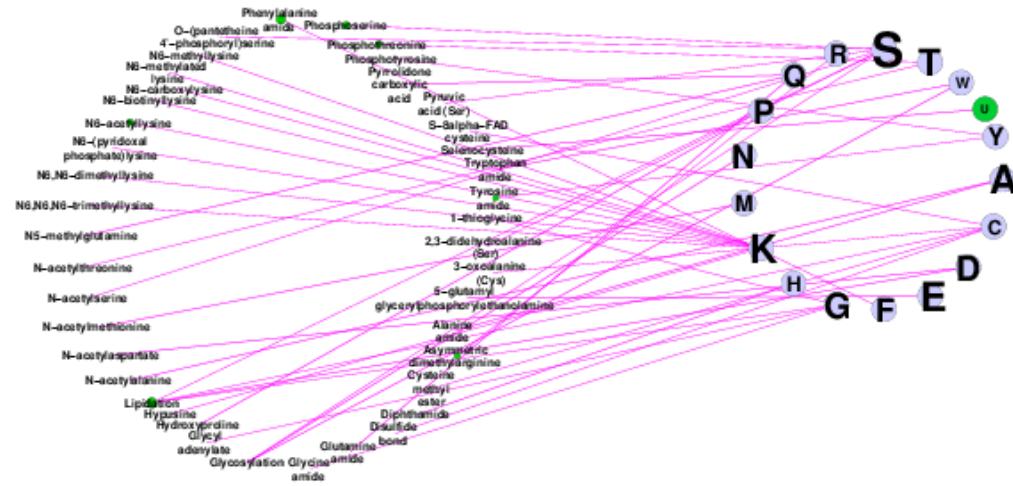


Figure 9.10: The *Caenorhabditis elegans* non-Mt Network: PTM frequencies (left) and associated active sites by tRNA frequencies (right).

9.5 Conclusions

A bias is a preferential treatment of some element. In this contribution, we used basic frequency information to produce evidence of a bias in PTM, active site and amino acids. We began the study by showing that the top two most common PTMs in nine diverse organisms (i.e., *glycosylation* and *phosphoserine*) had different preferential treatment by organism. Here we observed that although the PTMs were the same, their ranks of occurrence were often different between organisms. We then enriched this work by presenting the heatmaps of Figures 9.3 and 9.4 and the network models of Figures 9.7 to 9.10, to display how the PTM bias was present across the organisms and was closely linked to the biases of tRNA and active site occurrences.

Our data in the heatmaps and the network models had been separated into Mt and non-Mt datasets to show that proteins of both sets (of the same organism) were discriminate in their PTM usage. While *Homo sapiens*, *Mus musculus* *Oryctolagus cuniculus* and *Rattus norvegicus*, were associated with the PTMs *N6-Acetyllysine* and *N6-succinyllysine* of Figure 9.3 (Mt), the other organisms of this heatmap were clearly associated with other PTMs. Conversely, the organisms of the non-Mt dataset

of Figure 9.4 appeared to be associated with only *glycosylation* and *phosphothreonine*.

In our discussion, we noted that Mt proteins tended to have many fewer PTMs than the non-Mt proteins. Here we observed that the non-Mt networks were more dense with more PTM interactions than Mt models. This, we suggested, may have been attributed to the fact that Mt have highly conserved genetics which may make better use of fewer PTMs which have been there since early life. Mt also import many of their proteins from the nuclear mechanisms of their hosts. This observation sponsors evidence that there is less need of PTMs in Mt proteins because they do not require them to modify many proteins. Since Mt may not experience many different kinds of protein stress by being protected inside a cell, there may be no need for PTMs (again) to regulate and respond to stresses. Because non-Mt proteins have many more PTMs than Mt proteins, there is reason to suggest that the PTMs are available to respond to stresses which may constantly threaten the cell. In addition, since most of the non-Mt and Mt proteins are made by nuclear processes, these added PTMs, found associated to non-Mt proteins, may be available to help with this production. In this study, the evidence for this PTM bias is clearly presented by these findings and others.

PTMs interact with proteins at specific physical locations (active sites) and were also found to be affected by the bias (noted by the tRNA and amino acid frequencies). In addition, the heatmaps of Figures 9.5 (Mt) and 9.6 (non-Mt), illustrated that there is much discrepancy between the active site occurrence across the organisms. For instance, in Figure 9.5 (Mt), lysine (K) and serine (S) were common active sites for many of the organisms. In all the Mt and non-Mt network models we noted that lysine and serine often interacted with only one PTM in Mt models. However, in the non-Mt models, these two were often linked to PTMs which also interacted with other active sites as well.

The active sites are also simple amino acids and we investigated their amino acid

frequencies. In our data, we noted that their frequencies were not uniform and that the tRNAs which are associated to each amino acid were also not of uniform distribution. In our network models, the edge widths between the PTM (left) and tRNA nodes (right) described the amino acid frequency magnitudes. Here we observed from these differing edge widths that there is another bias which may impact the PTM bias.

In conclusion, unlike the network models of the non-Mt proteins, those of the Mt data were less dense and were always filled with fewer PTMs which appeared to interact with few active sites. These Mt networks appeared to us to be more efficient since their PTMs may have had more established roles than those of the non-Mt networks. In contrast, the roles of PTMs in non-Mt networks were seemingly ad-hoc and contained many unique PTMs which appeared to interact at times with many different active sites but often infrequently. This suggests that these non-Mt networks might be filled with PTMs which may have part-time roles in protein regulation. Many of these PTMs appeared to have low frequencies of occurrence which supports the notion that they were themselves modified complexes for some fleeting purpose.

In our extended work, we intend to provide all the graphs and figures of the organisms of this study which we were unable to include due to space limitations. In this work, we intend to investigate the PTMs which appear to be dormant. In another future work, we will investigate some of the factors that may cause them to begin regulation (in non-Mt), to compare these kinds of roles with equivalent ones in Mt. In addition, we intend to further study these PTM biases across extreme life forms (animals, plants, fish and insects) to determine whether some of the bias may also be explained by life style factors as well.

9.6 Article Details

This contribution was published in the Proceedings of the 8th International Workshop on Biomolecular Network Analysis (IWBNA), ACM-BCB 2014.

- Oliver Bonham-Carter, Ishwor Thapa and Dhundy Bastola, “Evidence of post-translational modification bias extracted from the tRNA and corresponding amino acid interplay across a set of diverse organisms”, In the Proceedings of the 8th International Workshop on Biomolecular Network Analysis (IWBNA) at ACM-BCB 2014.

Any fool can know. The point is to understand.

Albert Einstein

Chapter 10

A Content And Structural Assessment Of Oxidative Motifs Across A Diverse Set Of Life Forms

10.1 Abstract

Exposure to weightlessness (microgravity) or other protein stresses are detrimental to animal and human protein tissue health. Protein damage has been associated with stress and is linked to aging and the on-set of diseases such as Alzheimer's, Parkinsons, sepsis, and others. Protein stresses may cause alterations to physical protein structure, altering its functional identity. Alterations from stresses such as microgravity may be responsible for forms of muscle atrophy (as noted in returning astronauts), however, protein stresses come from other sources as well.

Oxidative carbonylation is a protein stress which is a driving force behind protein

decay and is attracted to protein segments enriched in R,K,P,T,E and S residues. Since mitochondria apply oxidative processes to produce ATP, their proteins may be placed in the same danger as those that are exposed to stresses. However, they do not appear to be impacted in the same way.

Across fourteen diverse organisms, we evaluate the coverage of motifs which are high in the amino acids thought to be affected by protein stresses such as oxidation. For this contribution of the thesis, we study RKPT and PEST motifs which are both responsible for attracting forms of oxidation across mitochondrial and non-mitochondrial proteins. We show that mitochondrial proteins have fewer of these oxidative sites compared to non-mitochondrial proteins. Additionally, we analyze the oxidative regions to determine that their motifs preferentially tend to make-up the connection points between the four kinds of structures of folded proteins (helices, turns, sheets, and coils).

10.2 Introduction

10.2.1 The Effects Of Weightlessness On Mitochondrial Function

The effects of exposure to microgravity or weightlessness for extended periods of time have proven to have negative impacts on mitochondrial protein function. For instance, in Philpott *et al.*^[230] it was found that morphological changes were observed in the left ventricle of rat hearts after space flight for 12.5 days. After this short time in weightlessness aboard the Cosmos 1887 bio-satellite, many of the rats in the experiment acquired damaged and irregular-shaped mitochondria and generalized myofibrillaredema which contributed to heart failures and death. Mitochondria, which are unable to orient themselves in the cell, have been studied^[57] where these dynamics were linked to several major neurodegenerative

diseases-including Alzheimers, Huntingtons, Parkinsons and other diseases. The animals also exhibited myofibrils (rod-like units of muscles) which were abnormal after this short time of exposure. In addition, the rats in the study by Philpott *et al.* (1990) exhibited loss of filament protofibrils (e.g., actin and myosin). The literature notes that protofibrils may be responsible for cell death in the organism, as noted in Caughey and Lansbury^[50] and may have been implicated as the toxic species responsible for cell dysfunction and neuronal loss such as in Alzheimer's disease and other protein aggregation diseases, explored in^[110].

Oxidative stresses on Earth may be very similar to those noted during space flight due to naturally created free radicals and reactive oxygen species, as noted by Nikawa *et al.*^[209]. In their study it was discovered that altered gravity conditions may be responsible for the onset of skeletal muscle atrophy in rat models, where rats were subjected to two forms of simulated weightlessness and also to actual space-flight conditions. Their study concluded that the distribution of muscular mitochondria had become diminished as a consequence of the damage to muscle fibers in all three conditions. They suggested that the muscular atrophy could be traced down to the interactions of free radicals and reactive oxygen species as a result of space-flight stresses.

Since stresses and their accompanying free radical and reactive oxygen species damage also exist on Earth, the study of their interaction sites in protein may provide insight into how similar damage may be incurred in space and on Earth. In this contribution, we show that there are generally less oxidative motifs in mitochondrial protein (from our data set of enzymatic and non-enzymatic proteins) than in non-mitochondrial proteins (for the same two sub groups of protein). We draw our evidence from the reduced occurrence of carbonylation motif *hot-spots* which were defined by^[195]. We contrast the scarcity of mitochondrial protein oxidative sites by showing that nuclear protein code contains many motifs which

were probably not lethal to the cell since they continued to exist, or were embedded in folded protein at locations where they were allowed to prevail.

We analyze the functional make-up of the existing regions of oxidation in mitochondrial and non-mitochondrial protein to determine that the oxidative sites tend to be located in the connection points between two structural events in folded proteins (helices, turns, sheets, and coils). We determine that these functional regions hold some protein structural importance which may explain why they still exist in mitochondrial protein which produces high levels of dangerous oxidative activity.

10.2.2 Carbonylation And PEST Protein Regulation Mechanics

RKPT Sequences and General Carbonylation: Carbonyl derivatives are the result of direct metal-catalysed oxidation interaction with the carbonylatable amino-acid side chains of arginine (R), lysine (K), threonine (T) and proline (P) residues and were explored in^[195]. Carbonyl derivatives of cysteine, histidine, and lysine may also be formed by the addition of reactive aldehydes which are derived from the metal-catalysed oxidation of polyunsaturated fatty acids. In^[70], it was noted that the residues of lysine carbonyl derivatives may be formed by secondary reactions with reactive carbonyl compounds on carbohydrates and advanced glycation/lipoxidation end products.

Proteolysis is the process of naturally removing proteins that are non-functional due to stresses of aging and related kinds of natural damage. Here, a region of protein sequence signals a natural removal by cellular processes. The literature suggests that the age of the protein may not always be the needed trigger for protein carbonylation^[171]. The same authors also studied insulin resistance (e.g., a symptom of protein degeneration) in mouse models in which they discovered that

mice, having an over-expression of the human catalase gene to mitochondria, are protected from an age-induced decrease in muscle mitochondrial function and muscle insulin resistance. Furthermore, the study suggested that age-associated reductions in mitochondrial function are due to organelle-generated reactive oxygen species production, contributing to the pathogenesis of age-associated muscle insulin resistance (protein degeneration). We note that insulin-resistance, and perhaps the above-mentioned diseases associated with aging, may be avoided by therapies that reduce mitochondrial oxidative damage.

Previously mentioned, aging or non-functional proteins are marked for destruction to avoid risks of failing protein in tissues. Oxidative carbonylation may not always be beneficial when it is due to environmental stresses that could create ailments such as: Alzheimers, cancer, cataractogenesis, diabetes, sepsis and others. Carbonylation may be central to these misfortunes since they all exhibit marked protein structures for degradation.

PEST Sequences: PEST sequences are hydrophilic, at least 12 amino acids in length and are rich in proline (P), glutamic acid (E), serine (S), and threonine (T). As in the case of carbonylation motifs, these regions also contain proline and threonine which may be attractors of protein degradation due to an associated short intracellular half-life. In Rechsteiner and Rogers^[240], it was noted that PEST sequences are involved in proteolytic signaling for rapid protein degradation by cellular regulation and its associated control systems. The PEST sequences typically signal the protein which contains the motif(s) for quick proteolytic degradation by the 26S ubiquitin proteasome system. It was also noted that this mechanism is active after the ubiquitination at the lysine residues within the PEST sequence. Rechsteiner and Rogers^[240] maintained that the PEST sequence generally acts as a signal peptide since its phosphorylation is likely necessary for protein degradation noted in Salmerón *et al.*^[255]. These sequences have also been noted to be a

stabilizing factor for L-type calcium channel proteins explored in Rogers *et al.*^[250].

PEST sequences are involved in the regulation of proteins in plants,^[167]. Dehydration responsive element binding is an important transcription factor that regulates environmental (abiotic) stress tolerance in plants. It was noted in Sakuma *et al.*^[254] that a central region of the DREB2A transcription factor in *Arabidopsis*, acting as a negative regulatory domain, and when deleted, activates its protein under stress conditions and also allows for the up-regulation of genes associated with salt or heat-stress responsive genes. Furthermore, the authors have suggested that this mechanism involves a PEST sequence acting as a negative regulatory domain that contains phosphorylation target sites for protein kinases such as PKC and CK2.

10.2.3 Mitochondria

Mitochondria play a part in cellular signaling, cellular differentiation and are able to initiate cellular death. Because they are important to the cellular house-keeping and the general health of the cell, any alterations to prevent normal function in mitochondria may be lethal to the cell. The host is also at risk in the event of the dysfunction of the organelle – functional mitochondrial respiration and energy homeostasis are critical for normal heart function and skeletal muscle maintenance,^[140]. General muscular atrophy is also a result of impaired mitochondria,^[198]. Interruptions to normal mitochondrial function are often associated to other ailments and disorders such as myopathies and cardiomyopathies, diabetes mellitus, neurodegenerative illnesses such as Alzheimers disease, diabetes,^[171] and aging,^[61;73;300;301].

Mitochondria are also responsible for the energy production of eukaryotic cells. In their absence, the cell would depend entirely on the anaerobic glycolysis as a source of ATP (Adenosine triphosphate). When glucose is converted to pyruvate by glycolysis, only a marginal quantity of total free energy is released from the glucose

which makes this an inefficient process for energy production. In the metabolism of sugars by mitochondria, the pyruvate is imported into the organelle where it is oxidized by molecular oxygen to carbon dioxide and water. The release of free energy from this operation makes an efficient process: 30 molecules of ATP are produced for each molecule of oxidized glucose, whereas, only two molecules are released by glycolysis in absence of the energy-making organelles. Mitochondria are mobile, able to change shape in the cytoplasm, and are able to drift around the cell while apparently associated with the microtubules. In some cells, they have been observed to anchor themselves to cellular locations where large amounts of ATP are necessary, such as in-between the myofibrils in a cardiac muscle cells or at the base of the flagellum of sperm cells.

In muscle cells, much ATP energy is required for function which is provided by the mitochondrial matrix enzymes of inner membrane along the respiratory chain. Since the mitochondria produce these sizable amounts of energy by oxidation processes, it is likely that proteins (or regions along the proteins) which attract oxidative carbonylation would not provide an evolutionary advantage and may be removed due to evolutionary pressures. Furthermore, since mitochondria are able to merge with other like-organelles, a failure in energy production may be introduced to the unified pair to provide negative impacts to both of the original organelles. On the other hand, such a point of oxidative damage attraction may not be detrimental to other kinds of proteins (e.g., nuclear) which do not function with such a significant profile in the cell. In this case, these carbonylation sites may be allowed to exist, especially if they occur in the middle of a folded protein where they cannot communicate with outside agents of biochemistry.

10.3 Methods

10.3.1 Protein Sequence Data From Organisms

We are interested in determining the general trends of motifs which attract carbonylation across a wide set of organismal protein sequence data. In Tables 10.1 and 10.2 we give a complete listing of all the organisms that provide the sequence data for both our mitochondrial and non-mitochondrial (enzymatic and non-enzymatic) protein comparative sequence analysis, as well as the number of protein sequences in all sets. This organismal set was chosen for two main reasons: (1) the sequence data for the organisms and as well as their mitochondrial genomes were freely available for download from the public international database, Uniprot (Protein Knowledgebase (UniProtKB), www.uniprot.org/); (2) these organisms represent a diverse group of life forms which may be sent on missions in space by NASA. For instance, during a mission of long duration, it may be desirable to send plants into space to provide much-needed nutrition for the crews. The animal, reptile and insect organisms provide more evidence from a wider variety of protein data which can be used for further comparison. By studying these organisms, we set the stage to understand how oxidation from microgravity or zero-gravity may affect them.

The source of protein sequence data for this study came from the 3rd May, 2013 release of curated protein definitions from the UniProt-SwissProt knowledge base^[88] (<http://www.uniprot.org/>). Within these protein definitions, the annotation keyword *Mitochondrion* is applied to proteins considered local to this organelle. The existence of an enzyme number was used to determine whether a particular protein is an enzyme. The BioPerl package (<http://www.bioperl.org/>) supports processing Swiss-Prot files and can extract protein sequences, annotation keywords and general information for each curated protein. We used the protein definitions

Table 10.1: The data used for the study. Organismal mitochondrial and non-mitochondrial protein which was divided into enzymatic and non-enzymatic data sets. The numerical values indicate the number of protein sequences that were selected from each of the four groups.

Organism	Mt Enzym.	Mt Non-enzym.	Non-Mt Enzym.	Non-Mt Non-enzym.
African clawed frog <i>Xenopus laevis</i>	61	108	615	2587
Amoeba <i>Acanthamoeba castellanii</i>	12	20	3	14
Mustard Plant <i>Arabidopsis thaliana</i>	247	460	3997	7520
Aspergillus <i>Aspergillus fumigatus</i>	31	56	453	341
Bakers Yeast <i>Saccharomyces cerevisiae</i>	296	760	1506	5238
Domestic Dog <i>Canis familiaris</i>	30	30	163	580
Fruit Fly <i>Sophophora melanogaster</i>	81	123	721	2273

Table 10.2: The data used for the study. Organismal mitochondrial and non-mitochondrial protein which was divided into enzymatic and non-enzymatic data sets. The numerical values indicate the number of protein sequences that were selected from each of the four groups.

Organism	Mt Enzym.	Mt Non-enzym.	Non-Mt Enzym.	Non-Mt Non-enzym.
House Mouse <i>Mus musculus</i>	415	558	3300	12352
Human <i>Homo sapiens</i>	431	596	3496	15744
Maize <i>Zea mays</i>	17	21	231	449
Norway Rat <i>Rattus norvegicus</i>	274	297	1771	5516
European rabbit <i>Oryctolagus cuniculus</i>	22	24	251	592
Nematode worm <i>Caenorhabditis elegans</i>	87	112	774	2458
Zebrafish <i>Danio rerio</i>	70	132	560	2136

from the Swiss-Prot release to create and populate our own SQLite database containing the protein information. Based on their genus and species, the protein sequences for the 14 target organisms of the study were extracted from the database. These protein sequences were partitioned into four protein classes based on the four possible combinations of the mitochondrial and enzyme properties.

For each target organism, four combined sequences were created for the four protein classes. Each combined sequence consisted of all the proteins of the same protein class for the particular organism. Note that a delimiter character was inserted between individual protein sequences to prevent new motifs from appearing in the joins (e.g., between protein sequences) in their concatenated sequence. This delimiter served to increase the size of each protein by one which caused the coverage percent of motifs to be slightly underestimated for all cases. The protein definitions are almost all larger than 100 amino acids and thus the added space adds no more than 1% to the size of each protein. In addition, although the current contribution reports results over protein data where a sequence similarity may exist across some of the sequences, we also provide all results in the supplementary data using protein where the sequence similarity has been reduced by less than by 40 percent. These supplementary results are available from the publisher's website.

10.3.2 RKPT Motifs - Attractors Of Oxidative Carbonylation

We now discuss the carbonylation content which we study across the concatenated protein content. Carbonylation and other forms of oxidative damage to cellular and mitochondrial proteins leave observable traces in the protein code,^[212;289]. In the study by Maisonneuve *et al.*^[195], a system of *rules* was created to find carbonylation sites (e.g., profiles of oxidation attractors) in protein sequence data. The authors built these rules on the concept that carbonyl derivatives may be formed by direct

metal-catalyzed oxidation attacks on the carbonylatable amino-acid side chains of arginine (R), lysine (K), threonine (T) and proline (P) residues (amino acids) in a protein sequence. These rules help to describe how to profile the RKPT-enriched carbonylation sites to be able to detect sites which are susceptible to oxidative attraction.

Maisonneuve *et al.* suggested that their profile system could be used to generate motifs which are known to attract types of oxidation. Since mitochondria produce free radicals and reactive oxygen species (known agents of oxidation) as a result of respiration, a study of the quantity of sites that could initiate oxidation would help to explain why these proteins do not appear to oxidize more readily. By finding that this content is generally lower in mitochondrial proteins than in non-mitochondrial proteins, our study provides a deeper understanding into how nature may resist natural dangers. Furthermore, since free radicals and reactive oxygen species are also thought to be initiators of protein damage during other exposures to stresses such as microgravity^[209], a study of content may explain where these protein failures are likely to occur.

We used these rules to create profiles of the protein sequence sites of oxidative damage. Although^[195] observed that aspartic acid (D), glutamic acid (E), tyrosine (Y), histidine (H) and cysteine (C) may be located near the hot-spot carbonylation motifs, we were only studying a distribution of 256 motifs themselves from an exhaustive list. We required the full set because the authors found that their rules do not always predict the oxidative nature of motifs by an analysis using mass spectrometry. By the same method, we created an alternative set of motifs attracting oxidation (PEST sequences) which are rich in proline (P), glutamic acid (E), serine (S), and threonine (T).

10.3.3 A Sequence Analysis By l -Word Proportions

We used proportions (for example, the coverage of a particular motif in the concatenated sequence data), not frequencies, to determine impact of these motifs. A statistical tool was developed in Bonham-Carter *et al.*^[28;30] that is able to determine the coverage across sequence material by an analysis of content of an arbitrarily selected motif set across a wide set of sequence data. Using this tool, the measurements of potential carbonylation sites were automatically normalized to allow comparison of the motif content between the sequences. To determine coverage of a particular motif in a sequence, we applied the following equation: $m_i \text{ in } S_L = (\text{count}(m_i) * |m_i|) / |S_L|$, where m_i is a motif, S_L is a sequence, $\text{count}(m_i)$ represents the number of occurrences of m_i found in S_L , and $|m_i|$ and $|S_L|$ are the lengths of the motif and the sequence, respectively. A proportion was computed for each of the n motifs of a set to create an n -length vector which would be processed by heatmaps. Each vector represented the coverage of the 256 RKPT motifs with respect to a particular protein class (mitochondrial, non-mitochondrial: enzymatic and non-enzymatic). By the same method, four vectors of size 256 are also created for the PEST motifs with respect to each organism.

The Kruskal-Wallis one-way analysis of variance by ranks and its post-hoc pairwise comparison tests are sufficient to order the motif coverage over our sequence data (e.g., smallest to largest coverage). Traditional statistical tests such as the ANOVA test of means could also be applied to our data, however, we could not be sure that each test assumption (normally distributed, for example) could always be met with our biological data and so non-parametric tests were appropriate.

10.4 Cluster Analysis

The *l*-word analysis developed in Bonham-Carter *et al.*^[28;30] was originally used to determine similarity between genetic sequence material. We used the author's method and tool to create vectors of motif proportion in our protein content. To create these vectors, the proportions of motifs from each RKPT and PEST set across each of the four protein class sequences is calculated and normalized. These vectors were then clustered and illustrated by heatmaps also used by the above authors.

The heatmaps helped to visualize the proportions of oxidative motifs across our protein samples. In each graphic, there are four different classes of proteins shown for the organism (the enzymatic and non-enzymatic content for each of mitochondrial and non-mitochondrial sets). Running along the bottom of the figures are the proportions (coverages) of the RKPT or PEST motif set. The size of the proportion for each motif across each of the protein classes are illustrated by the brightness of cell colors directly above each motif. The lighter colored cells indicate larger proportions for a particular motif where-as darker cells indicate that the proportion is either zero (dark blue) or nearly-zero.

Reduced Mitochondrial Oxidative Content: For both of our motif sets (RKPT and PEST) we noted that there was an overwhelming reduction of oxidative motif content in the mitochondrial (enzymatic and non-enzymatic) content. For instance, in the RKPT trials, all but three organisms indicated this reduction in mitochondrial protein content. Across the heatmaps of nearly all organisms of both the RKPT and PEST experiments we noted that in the mitochondrial proteins, there were generally more darker cells (indicating reduced motif proportions and absences). By contrast, the non-mitochondrial proteins tended to have many lighter colors to imply that these proteins contained many more motifs and generally higher concentrations. We note that this evidence suggests that mitochondrial proteins have fewer sites where oxidative carbonylation is likely to be initiated. Interestingly,

in both the mitochondrial and non-mitochondrial sets, the enzymatic material tended to have fewer or lower motif concentrations than the non-enzymatic material.

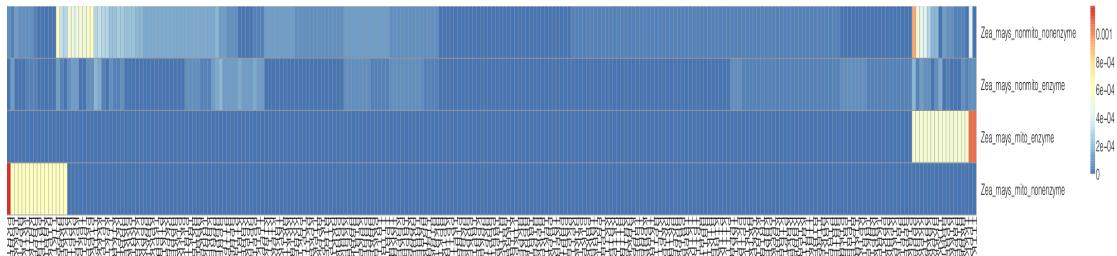


Figure 10.1: The four protein classes from *Zea mays* (maize) sequence data describe a typical heatmap from our data. Each bar is made up of colored cells which represent the amount of motif coverage in a particular protein class. The darker the shade of cell (blue) then the closer to zero is the coverage of motif which is thought to attract oxidative activity. We noted that the darker regions were usually more pronounced in the mitochondrial proteins than the non-mitochondrial proteins and were often more pronounced in the enzymatic data sets. The other heatmaps from this work are included in the supplemental data.

Figures 10.1 and 10.2 typify our findings of the RKPT and PEST sets: there were large portions of mitochondrial protein sequence code (enzymatic and non-enzymatic) which were devoid of the oxidative motifs , noted by the large expanses of darker colors, zero or near-zero proportions (no coverage), in the heatmaps. The other heatmaps (available in the supplemental data) were similar and showed that a large part of the RKPT and PEST motifs sets also had zero or near-zero proportions in the mitochondrial sequence data. There was a striking similarity between both the RKPT and PEST sets in that there were only four organisms for which it was not completely obvious that the mitochondrial protein data had the least oxidative motifs : *Rattus norvegicus* (Norway rat), *Homo sapiens* (human), *Mus musculus* (house mouse) and *Saccharomyces cerevisiae* (bakers yeast).

Evolutionary Importance: During the course of this study, we have noticed that there is often a link between the amount of absent RKPT and PEST material and the perceived longevity of the organism. For instance, in the above two heatmaps for the organisms maize and fruitfly which have shorter lifespans, the trends of motif absence

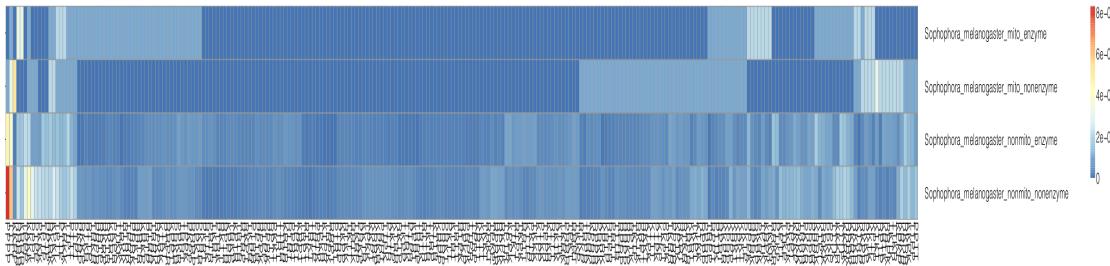


Figure 10.2: The four protein classes from *Sophophora melanogaster* (fruitfly) sequence data in a heatmap. The heatmaps from the other organisms discussed in this work are included in the supplemental data.

are established. We suggest that since human lifespans are longer, the evolution of motif trends is not as pronounced as shown in Figure 10.3. The heatmaps in the supplementary data may allow other examples of this evolutionary interest to be found, although it does not appear to be a consistent finding throughout the entire set.

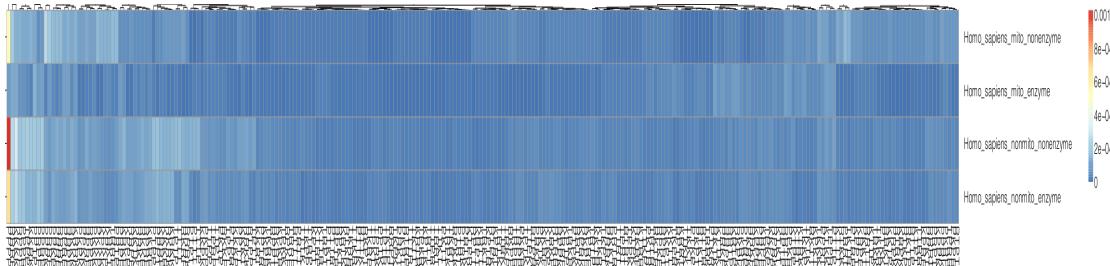


Figure 10.3: The four protein classes from *Homo sapiens* (human) sequence data in a heatmap.

10.4.1 A Comparison Across Organism Sequence Data

When all the organismal vectors (56 total) were compared to each other in the same heatmap where they were clustered based on their motif content, we discovered that there were two basic sub-trees created in the RKPT and the PEST heatmaps. We also noted that across both oxidative motif sets, there were corresponding subtrees where the numbers of each type of sequence data was approximately the same. For instance,

Table 10.3: RKPT: First and Second subtrees of all the data when clustered together.

	RKPT _{1st}		PEST _{1st}	
	Mito	non-Mito	Mito	non-Mito
Enzy	1	9	0	6
Non-Enzy	7	13	5	12
	RKPT _{2nd}		PEST _{2nd}	
Enzy	14	5	14	8
Non-Enzy	6	1	9	2

in the first subtree of the RKPT motif set, summarized in Table 10.3 (RKPT_{1st}), and that of the PEST motif set (PEST_{1st}) the counts of the mitochondrial (both enzymatic and non-enzymatic) sequence data was eight (RKPT) versus five (PEST). The counts of the non-mitochondrial sequences for the same two subtrees were: 22 (RKPT) and 18 (PEST). Conversely, in the second two subtrees summarized in Table 10.3 (RKPT_{2nd}) and (PEST_{2nd}), there were 20 (RKPT) and 23 (PEST) mitochondrial (enzymatic and non-enzymatic), however there were six (RKPT) and 10 (PEST) non-mitochondrial (enzymatic and non-enzymatic) sequences. These findings suggest that the mitochondrial and non-mitochondrial sequence data (both enzymatic and non-enzymatic), had very comparable motif composition in terms of the RKPT and PEST sets, taken across in the protein sequence data.

10.4.2 Statistical Analysis

All 256 RKPT-enriched motifs are considered equally likely to promote carbonylation, we examine the coverage percentages of all such motifs to a particular combined protein class sequence in a single vector of 256 percentages. We are interested whether the percentages vary based on protein class. Populations with significantly higher percentages would represent a protein class that is more susceptible to carbonylation. A similar argument can be made for PEST motifs. The resulting data sets of coverage percentages by the RKPT and PEST motifs did

not follow a standard distribution – many of the 256 motifs had a zero coverage percentage. The resulting distributions of values were skewed by the abundance of zero values which created non-normal distributions. These distributions necessitated non-parametric tests of analysis. For each organism, we examined the data sets using the non-parametric Kruskal-Wallis rank test which was used primarily for comparing the distributions of multiple populations. Gao's nonparametric multiple comparison procedure was used for this purpose.

The result of the non-parametric Kruskal-Wallis rank test comparing the RKPT motif coverage rate of the four protein classes in all fourteen target organisms was that the null hypothesis was rejected for each organism using a 5% error allowance ($\alpha = 0.05$). This indicates that the distributions of coverage values are not considered to be equivalent in any of the organisms. The result of Gao's non-parametric multiple comparison post-hoc tests showed that in 13 of 14 organisms, the RKPT-coverage ratios in the mitochondrial proteins (both types) were found to differ significantly from those in the non-mitochondrial proteins. Between the two types of mitochondrial proteins (enzymatic and non-enzymatic), there was never a statistically significant difference in their RKPT-coverage ratios. In 13 of 14 cases, we noted that the mitochondrial proteins had RKPT-coverage ratios less than that of their non-mitochondrial counterparts (shown by comparison of the left and right sides of Table 10.4).

Rankings of Oxidative Content: The premise of this contribution is that mitochondrial protein sequence data holds the least oxidative motifs since oxidation is commonly performed in this organelle. In Table 10.4, we show the rankings from lowest to most oxidative motifs for both RKPT and PEST, according to our analysis. When we were ranking, if there was a tie between protein sequences, then we assigned the averages of all ranks of the tied sequences. In the case of *amoeba*, our Kruskal-Wallis results were inconclusive since all values were just above zero

Table 10.4: Ranking scores from lowest to highest PEST and RKPT (oxidative carbonylation) content of the four protein classes: ME *mitochondrial enzymatic*, MN *mitochondrial non-enzymatic*, NE *non-mitochondrial enzymatic*, NN *non-mitochondrial non-enzymatic*.

Organism	ME		MN		NE		NN	
	PEST	RKPT	PEST	RKPT	PEST	RKPT	PEST	RKPT
African clawed	1.5	1.5	1.5	1.5	3	4	4	3
Amoeba	3	2	1.5	4	1.5	1	4	3
Aspergillus	2	1.5	1	1.5	3	3	4	4
Bakers yeast	1	1	2	2	3	3	4	4
Domestic dog	1.5	1.5	1.5	1.5	3	4	4	3
European rabbit	1.5	1.5	1.5	1.5	3	4	4	3
Fruit fly	1.5	1.5	1.5	1.5	3	3	4	4
House mouse	1	1	3	2	2	3	4	4
Human	1	1	3	3	2	2	4	4
Maize	1.5	1.5	1.5	1.5	4	3	3	4
Mustard plant	1	3	2	1	3	2	4	4
Nematode worm	1	1	2	2	3	3	4	4
Norway rat	1	1	3	3	2	2	4	4
Zebrafish	2	1.5	1	1.5	3	3	4	4
Averages	1.46	1.46	1.86	1.96	2.75	2.86	3.93	3.71

and too low to be accurately determined. For their ranking, we examined the RKPT and PEST heatmaps for the organism and made our rankings from the observed number of motifs contained in the protein samples. By our premise, we note that the average rankings of the mitochondrial enzymatic sequence data (ME) for RKPT and PEST were both 1.46. However, for the mitochondrial non-enzymatic (MN) data, we noted that the averages increased slightly for each: 1.96 (RKPT) and 1.86 (PEST). The mitochondrial values are much smaller than those of the non-mitochondrial sequences – the non-mitochondrial enzymatic (NE) sequence data had average rankings of: 2.86 (RKPT) and 2.75 (PEST). The non-mitochondrial non-enzymatic (NN) sequence appeared to contain the most oxidative motifs from the entire set: 3.71 (RKPT) and 3.93 (PEST).

Individual Residues: We noted from the average content of the RKPT and PEST sequence motifs increased in the mitochondrial to non-mitochondrial (both, enzymatic

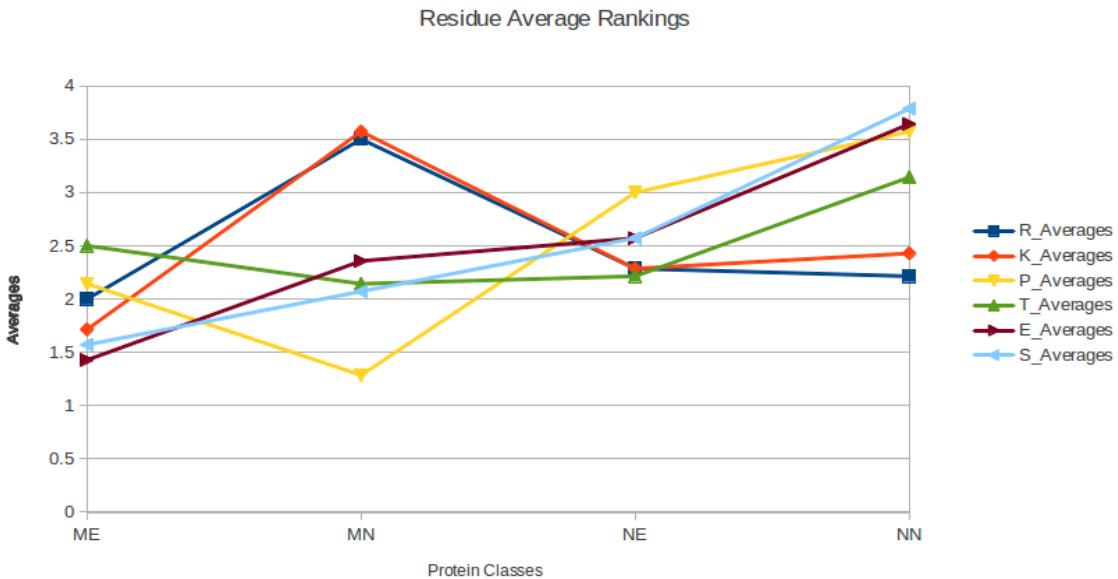


Figure 10.4: Rankings of R, K, T, P, E and S residues across the protein classes of all organisms. We note that P and T, common to both the PEST and RKTP motif sets, have general upward trends but several residues do not.

and non-enzymatic) sequence data. Since there are two residues (P and T) in common with both the RKPT and PEST motif sets, we asked whether this trend was due to an overlap of residues. We note in Figure 10.4.2 that the upward-averages from the motif sets were probably not likely due to both P and T uniquely since the trend we were looking for was largely absent. Although some of the residue averages across the organismal set appear to follow the same upward trend noted in the motif averages, the other residues do not. This finding suggests that the conjunction of some residues when found with others (forming motifs) may have been responsible for the trend of increasing averages in the protein sequences.

10.5 Structural Analysis

In its primary structure, a protein appears as a sequence of amino acids which describe its functional identity. Since our study concludes that there are fewer regions which

may incite oxidative activity in mitochondrial proteins, we now examine the kinds of structural features occurring at these RKPT and PEST oxidation-attracting regions. As implied already, one of the features of proline (P) in the RKPT and PEST data sets is that it is able to provide a flexible joint-like feature in the protein secondary structure. This amino acid is also likely to attract oxidative activity. Since the folded state of a protein is critical to its functional role, motifs which may allow for the alteration of the protein structure may be dangerous yet, necessary commodities. In this case, the proline component of the RKPT and PEST motif sets could be a necessary risk for the protein since it may have helped to determine some functional importance.

Natural selection has appeared to favor the reduction in sites which are susceptible to oxidative activity in mitochondria. This may be due to the fact that mitochondria perform many oxidative reactions related to respiration which expose its proteins to large amounts of stress. It is therefore logical that regions which are not necessary to the protein's structure may not be conserved especially when they may cause danger. We suspect that these sites, although dangerous, may have been retained for some reason such as their structural contributions to the protein formation.

In the next part of the study, we show that the few oxidative sites found in mitochondrial protein may also originate in regions where the oxidative content is unable to react when in contact with stress. In this work, we study the structural compositions of the regions of the sites and take into account their proportions (e.g., coverage) to show that they are often found to make-up the bends or joining regions that separate different states (e.g., the coils, sheets, helices and turns) of a folded protein. To show their bending or joining regions, we compare the number of oxidative motifs to all the other sequence content, across mitochondrial and non-mitochondrial protein. We utilize non-parametric Kruskal-Wallis rank tests to compare the proportions of structural features between these sets to draw our

conclusions. Below, we list the individual tests (1 through 9) for our study where we weigh oxidation on one hand to non-oxidation on the other across mitochondrial and non-mitochondrial protein data. In Section 10.5.2, we explain the outcomes and significance of this work.

1. RKPT oxidative regions within mitochondria compared to non-oxidative regions within mitochondria
2. PEST oxidative regions within mitochondria compared to non-oxidative regions within mitochondria
3. RKPT oxidative regions within non-mitochondria versus non-oxidative regions within non-mitochondria
4. PEST oxidative regions within non-mitochondria versus non-oxidative regions within non-mitochondria
5. RKPT oxidative regions within mitochondria compared to oxidative regions within non-mitochondria
6. PEST oxidative regions within mitochondria compared to oxidative regions within non-mitochondria
7. RKPT oxidative regions versus PEST oxidative regions in mitochondria
8. RKPT oxidative regions versus PEST oxidative regions in non-mitochondria
9. All words from mitochondria compared to all words from non-mitochondria protein content

We first describe the method utilized to obtain the structural features located within the oxidative and non-oxidative protein of mitochondrial and non-mitochondrial sequence data. The Garnier tool^[97], available in Emboss'

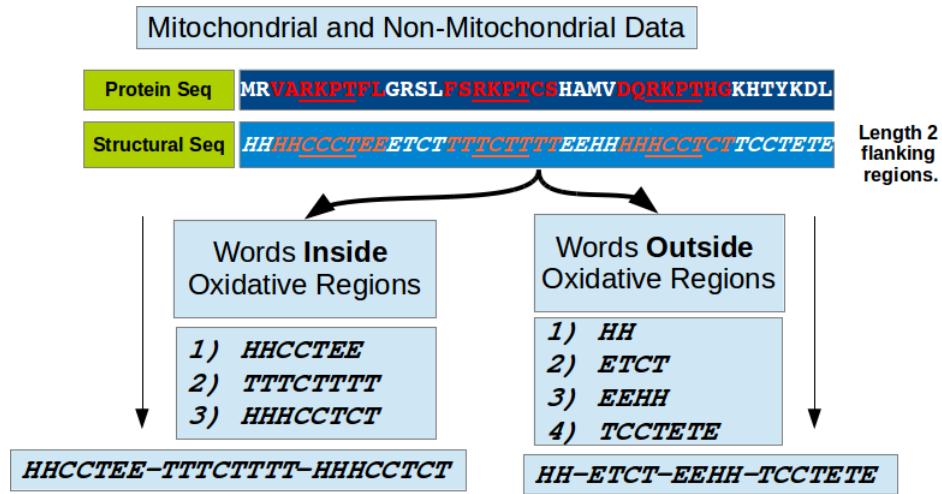


Figure 10.5: The flowchart for the method (Part 1). Here we describe our method using an sample protein sequence which is processed by EMBOSS' protein prediction tool to ascertain the protein's functional structure (C = Coils, E = Sheets, H = Helices, T = Turns). The individual locations where oxidation sites were detected are extracted from the structural sequence to uncover the structure-feature words at these locations. These words are then used to build a new sequence containing all oxidation site structural information.

bioinformatics toolkit^[243], is a prominent software for the convenient prediction of protein secondary structural sequences^[232;330]. Each organismal protein of our study was applied to this tool to determine a likely protein structure. We refer to the *structural sequence* as a that which was prepared by Garnier to predict the structures of the protein. This sequence is comprised of the alphabet {C, E, H and T} for {Coils, Sheets, Helices and Turns}, respectively. The term, *word* is reserved for regions which have been isolated from this structural sequence and are composed of protein structural information at oxidative sites. We use the term, *motif*, to imply a region that is made up of amino acid information and is derived from a protein sequence. We will use these two terms in the description of our method.

In each protein, we determined the locations of the oxidation of the RKPT and PEST motifs. Since the relationship between an amino acid sequence (protein) and a structural sequence is one-to-one, we determine their structural nature from the

structural sequences. At each oxidation site of length-4 (according to the location of the motif in original protein sequence), we recorded the length-4 word with a length-2 flanking region at both ends from the structural sequence. By adding this flanking region, we amassed more information concerning its structure. Each structural word was recorded and then removed from the structural sequence which we maintained for further testing. To avoid introducing new words into the sequence from its joining regions, we inserted a delimiter at the locations in the structural sequences. The words that were removed were recorded in lists and were later used to create a new sequence where each listed word was placed adjacently with the next listed word, separated by delimiters. In Figure 10.5, we see the words of length-4 (and its length-2 flanking regions) being extracted from a structural sequence. These extracted words are shown to be used to create a new sequence which only contains the structural details of the oxidative regions. The structural sequences (devoid of oxidative motifs) are also maintained for later use.

10.5.1 Grouping Structural Elements

We define a *grouping* to be a specific word cluster where all members share a common prominent structure (e.g., a coil, helix, sheet or turn). In order to quantify the groupings which were derived from the sequence containing only structural content from the oxidative regions, we employed a *sliding-window* which moved from left to right down the sequence to extract all words of length-2 through 8. For each obtained word of a set for a particular length, we extracted its proportion (e.g., the sequence coverage) using the same equation as the one mentioned above in Section 10.3.3. This proportion was used to drive our non-parametric statistical tests which are discussed later. We extracted the word groupings (the dominant structural groups) by determining the most abundant character in the words using a majority ruling. In the case where there was a tie between two abundant characters in a word we

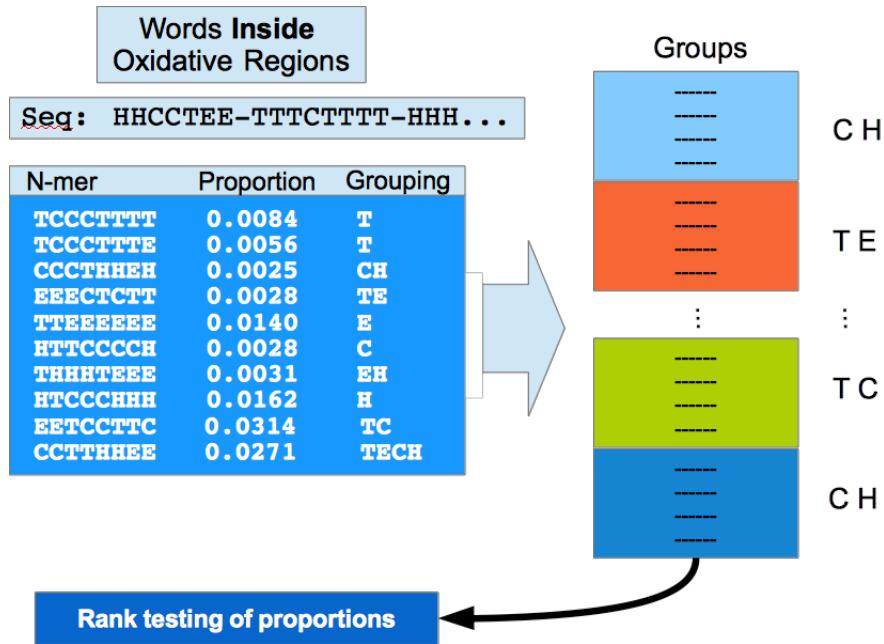


Figure 10.6: The flowchart for the method (Part 2). We create a distribution of functional groupings using the majority rule where the dominant structures (C = Coils, H = Helices, E = Sheets, T = Turns) determine the word's form.

attributed both dominant characters to the words.

For example, the word, TTTCCC was characterized as contributing to both groups T and C (providing a *turn* and *coil* structure), however the word TTTT was characterized as only contributing to the group T (providing the word's *turn* structure). In Figure 10.6, we illustrate the method used to assign the word groupings.

10.5.2 Structural Results

We now describe the results of the outlined tests to compare the groupings of length-8 structural feature words. In each test, we determined the most prominent structural features which were in one kind of a set (oxidative or non-oxidative) and not in the other. In Figures 10.7 and 10.8, we compared the groupings between oxidative regions within mitochondrial to non-oxidative regions within the mitochondrial proteins. For the RKPT set, we noted that the most prominent grouping was HT, suggesting that

helices and turns were the most abundant kinds of structural features making up RKPT oxidative regions. These features, by comparison, were not prominent in the non-oxidative regions. In Figure 10.8, CH (coils and helices) were the most prominent groupings of the PEST oxidative regions. From these two tests, we esteem that helices which become curves (or curves which become helices) are very important structures making up the oxidative regions. We suggest that any oxidative regions which are located in mitochondrial protein may have survived so long by virtue of the fact that they were likely hidden deep within helices and curves and were therefore unable to interact with agents of carbonylation when under conditions of stress. All graphs concerning this work are available in the supplementary data.

In the tests of oxidative regions within non-mitochondria compared to non-oxidative regions within non-mitochondrial protein, of Figures 10.9 and 10.10, we note the appearance of a length-4 grouping, CEHT. We note that this word implies that all four structures were found close to each other that made up some of the oxidative regions in non-mitochondrial proteins. In Figures 10.9 and 10.10, curves and helices (especially for PEST motifs) were again important features in non-mitochondria which may appear to attract the oxidative motifs . Since complex structures of length-4 were not found mitochondrial oxidative motifs (from Figures 10.7 and 10.8) we may conclude that mitochondrial oxidative content was not found in these features. This adds support to a notion that oxidative regions exist in the simple joins between two kinds of basic structures.

In the tests of oxidative regions within mitochondria compared to the oxidative regions within non-mitochondria (Figures 10.11 and 10.12) we determine that RKPT motifs appeared to form joins of helices and turns (or vice versa) in mitochondria since HT in Figure 10.11 was exceptionally prominent. This supports the preference of the motif set for helices and turns of Figure 10.7. In the PEST motif set of non-mitochondria, joins between coils and helices were important structural features in

Figure 10.12, as they were in Figure 10.8.

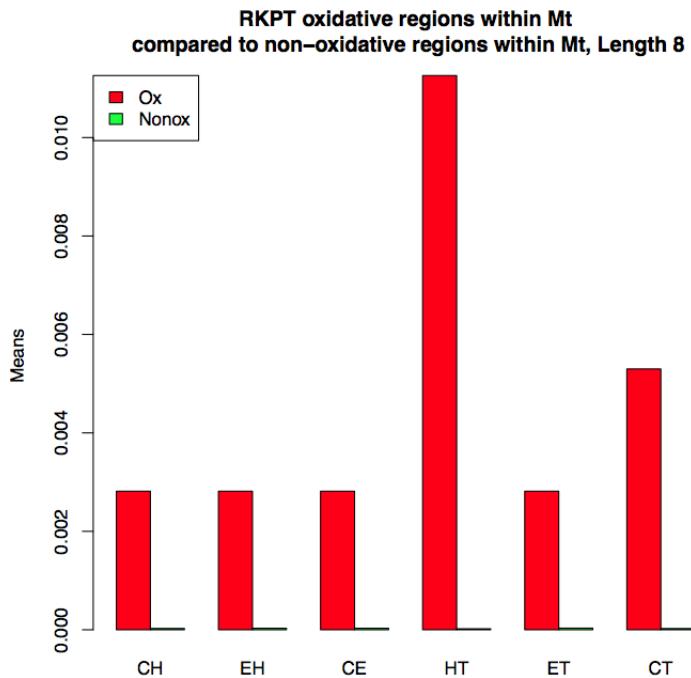


Figure 10.7: Test Number One: The structural words made up of coils (C), sheets (E), helices (H) and turns (T) are along the x-axis. The y-axis represents the mean coverage of the structural words of all proteins in the set. The *Ox* and *Nonox* refer to oxidative and non-oxidative content, respectively.

When we compared RKPT structural features to those of the PEST set in mitochondrial protein (Figure 10.13) we noted, that both sets form the joins of sheets and turns ET, although RKPT motifs appeared to inhabit more of these joins than the PEST motif set. In non-mitochondrial protein, there were fewer differences between both sets (Figure 10.14) and we found that many of the structural features were present, although also in differing amplitudes. This implies that the oxidative content of non-mitochondria is able to survive in many more kinds of structural joining regions than in mitochondria.

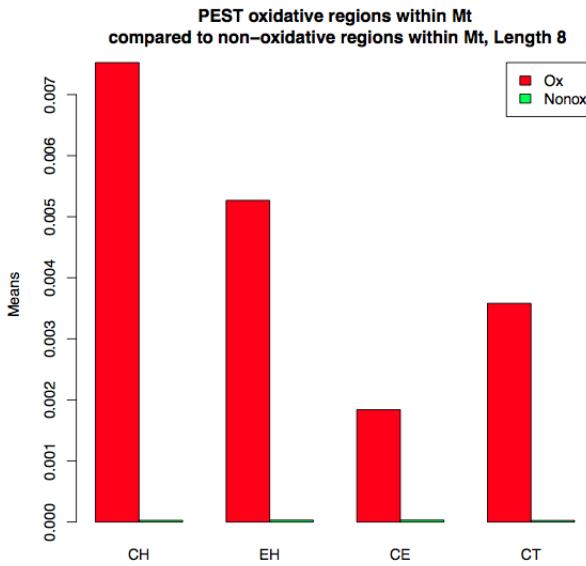


Figure 10.8: Test Number Two: The structural words made up of coils (C), sheets (E), helices (H) and turns (T) are along the x-axis. The y-axis represents the mean coverage of the structural words of all proteins in the set. The *Ox* and *Nonox* refer to oxidative and non-oxidative content, respectively.

10.6 Discussion

Similar to the flying buttresses found in medieval cathedral construction offering powerful structural support to walls, similar biological constructs in protein construction may also offer similar kinds of support. Furthermore, due to the extreme sensitivity of the protein's functional identity to its folded form, these biological *flying-buttresses* may play more of a role than previously thought. For instance, in [79], it was found that human protein, hHep1 oligomerizes (bonding to protein) in a concentration-dependent fashion and that its zinc ion is thought to have an important protein-structural stabilizing effect. The RKPT and PEST oxidative content in both mitochondria and non-mitochondria was generally found in the joins of structural regions and so these regions may also provide some form of structural support, as well.

In mitochondria we claim that there are fewer sites which attract oxidative activity

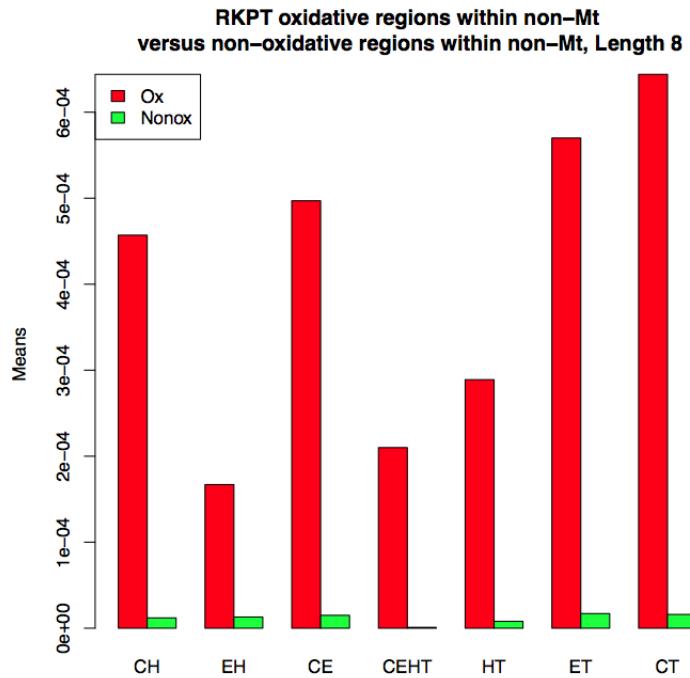


Figure 10.9: Test Number Three: The structural words made up of coils (C), sheets (E), helices (H) and turns (T) are along the x-axis. The y-axis represents the mean coverage of the structural words of all proteins in the set. The *Ox* and *Nonox* refer to oxidative and non-oxidative content, respectively.

due to these proteins being so close to a source of internally created oxidation during respiration. This pressure to reduce their numbers may be frustrated by a greater need for these regions as they may allow for specific folding configurations and may therefore be indispensable for the formation of a functional identity. In mitochondria, there appeared to be less diversity in the numbers and kinds of these joining regions, however in non-mitochondria, we generally saw more unique kinds of joins in which we found oxidative content. In mitochondria, the RKPT motifs were generally found in the joining regions of helices and turns (Figure 10.7, HT), but the PEST motifs were generally found in the joins of coils and helices (Figure 10.8, CH).

Although the PEST motif set appeared in some of the same structural regions as the RKPT set, we noted that the two groups did not generally inhabit the same kinds of joining regions. In the case of mitochondrial protein, only sheets and turns (Figure

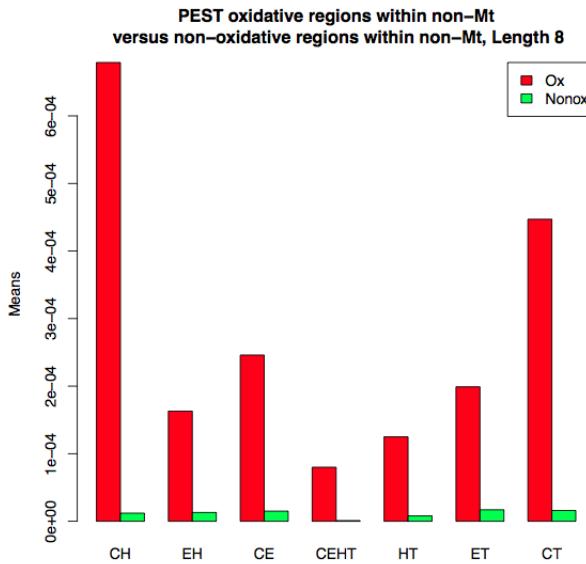


Figure 10.10: Test Number Four: The structural words made up of coils (C), sheets (E), helices (H) and turns (T) are along the x-axis. The y-axis represents the mean coverage of the structural words all proteins in the set. The *Ox* and *Nonox* refer to oxidative and non-oxidative content, respectively.

10.13, ET) were locations where both motif sets could be found however, in Figure 10.14, we note that there were many more places where both oxidative sets could be found together. We also note that there was no significant difference between the word proportions for structures in the full sequences of mitochondrial and the non-mitochondrial proteins. The differences only became visible when we focused on the differences between oxidative and non-oxidative content.

10.7 Conclusions

In this study, we discussed general protein damage from oxidation carbonylation which has also been associated with muscular degeneration as observed from stress exposure to microgravity or zero-gravity environmental conditions. Since free radicals and reactive oxygen species are thought to be closely linked to muscular damage as a result of space-flight stresses according to Nikawa *et al.*^[209]. Since they are

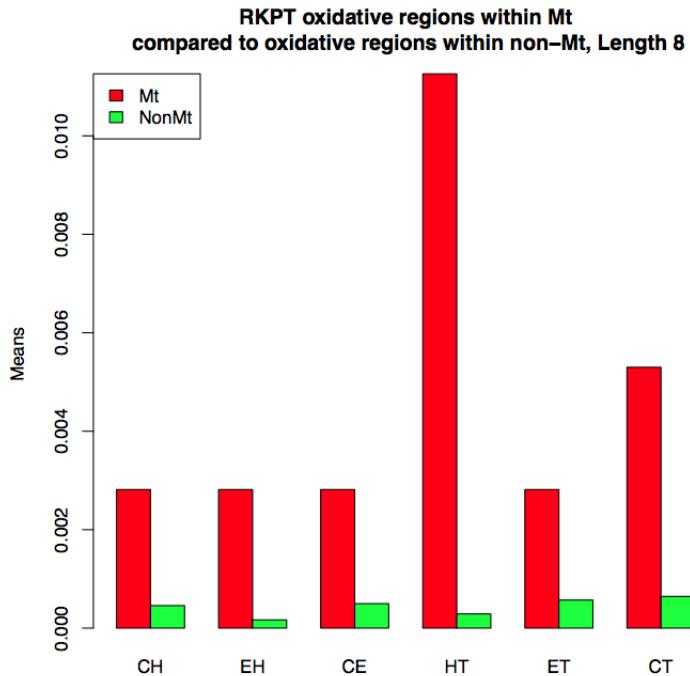


Figure 10.11: Test Number Five: The structural words made up of coils (C), sheets (E), helices (H) and turns (T) are along the x-axis. The y-axis represents the mean coverage of the structural words of all proteins in the set. The *Mt* and *NonMt* refer to mitochondrial and non-mitochondrial content, respectively.

also existent on Earth and initiated by other kinds of stresses, the study of protein interaction sites (both the RKPT and PEST sets) has provided knowledge about a possible mechanism for damage. In this contribution, we showed that there were generally fewer sites for oxidative carbonylation in mitochondrial protein than in non-mitochondrial proteins (further divided into two discrete groups: enzymatic and non-enzymatic content). For this effort, we studied the differences of sequence composition in both datasets to locate and count profiled reaction site motifs (from the RKPT and PEST sets) which we applied to a statistical analysis.

We noted from the average rankings of RKPT and PEST motif content across each of the four concatenated sequence sets that there were generally differing amounts of the motif proportions. For example, in Table 10.4 it was clear that mitochondrial

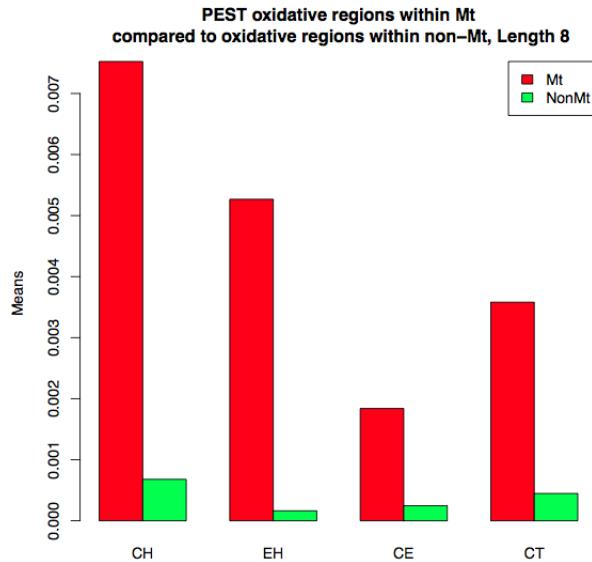


Figure 10.12: Test Number Six: The structural words made up of coils (C), sheets (E), helices (H) and turns (T) are along the x-axis. The y-axis represents the mean coverage of the structural words of all proteins in the set. The *Mt* and *NonMt* refer to mitochondrial and non-mitochondrial content, respectively.

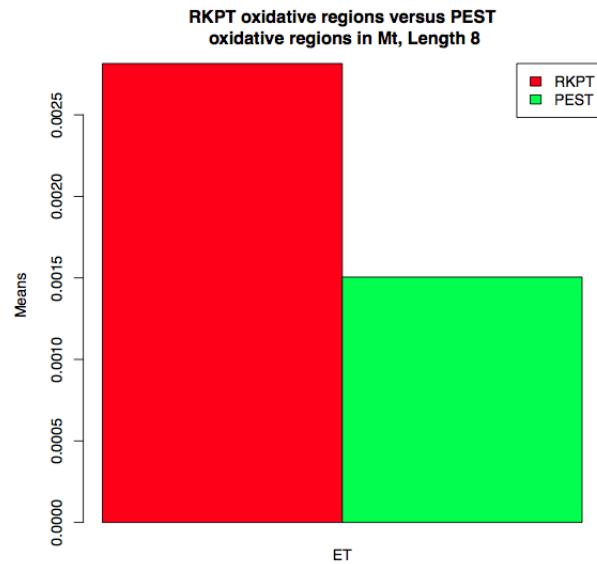


Figure 10.13: Test Number Seven: The structural words made up of coils (C), sheets (E), helices (H) and turns (T) are along the x-axis. The y-axis represents the mean coverage of the structural words of all proteins in the set. The *RKPT* and *PEST* refer to these motif contents, respectively.

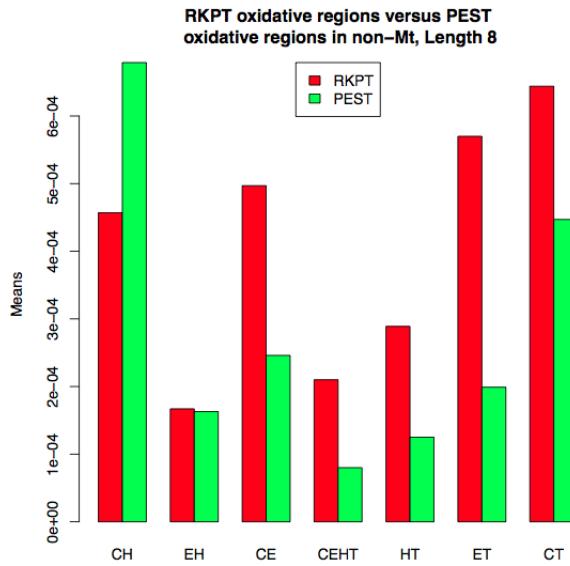


Figure 10.14: Test Number Eight: The structural words made up of coils (C), sheets (E), helices (H) and turns (T) are along the x-axis. The y-axis represents the mean coverage of the structural words of all proteins in the set. The *RKPT* and *PEST* legends refer to these motif contents, respectively.

enzymatic protein sequence data contained the least amount of motifs which attract oxidative activity (the RKPT set). This finding was also the same for the PEST motif set. At the other end of the scale, the non-mitochondrial, non-enzymatic protein sequence data contained the most oxidative motifs (in both the RKPT and PEST sets).

This phenomenon of having the least number of oxidative motifs in mitochondria, may likely be explained by the fact that these organelles regularly perform procedures involving oxidation to harvest energy for their proteins. Furthermore, in muscular proteins, which require massive amounts of energy, there is generally more pressure on the mitochondria to perform their energy-amassing functions by oxidation. If certain mitochondrial proteins were attractors of *unauthorized* oxidative activity, then these proteins may place the organelle in the danger of being unable to fulfill their energy duties. Therefore, we suggest that the addition of elements of the RKPT

and PEST motif sets, which are responsible for types of protein carbonylation and degradation, would not enrich or positively contribute to the survival rates of the associated organelles. In this case, we maintain that they would likely be avoided by the protein sequence structure and may even be removed by evolutionary processes.

We also note that there was no significant difference between the word proportions for structures mitochondrial and non-mitochondrial proteins. The differences only became visible when comparing the oxidative and non-oxidative content. We noted that the joining points where structural elements met were generally found in the oxidative regions of both protein sets. We found that there was a lower percentage of these joining-regions in the mitochondrial set compared to the non-mitochondrial set. For instance, in mitochondria, the largest percentage of structural joins in the RKPT oxidative set tended to be made up of helices joining with turns (see Figure 10.7). However, the PEST set was found in fewer types of joining regions with the largest percentage occurring in the joins of coils to helices. There are four types of structural elements and every possible join between them occurs in the RKPT regions. This is not the case for the PEST content which was contained in four of the six possible joins (see Figure 10.13). In particular, this set is missing the HT variety of join which is the prominent in the RKPT regions.

Oxidation sites in non-mitochondrial protein may not be exposed to a constant source of oxidative carbonylation. For this reason, their numbers may not have the same evolutionary pressures to be reduced as imaginable in mitochondria. On a final note, these sites in non-mitochondrial proteins may also reside deep within the cell where they are protected from the sources of oxidation. In future works, we will study a wider variety of proteins from diverse organisms and tissues-types to determine levels of carbonylation content. We esteem that similar tissue-types may have comparable levels of oxidative motifs .

10.8 Article Details

This contribution was published in Computers in biology and medicine, 2014

- Bonham-Carter, Oliver, Jay Pedersen, and Dhundy Bastola. “A content and structural assessment of oxidative motifs across a diverse set of life forms.” Computers in biology and medicine 53 (2014): 179-189.

Life is really simple, but we insist
on making it complicated.

Confucius

Chapter 11

A Study of Bias And Increasing Organismal Complexity From Their Post-Translational Modifications And Modification Site Interplays

11.1 Abstract

Post-translational modifications (PTMs) are important steps in the biosynthesis of proteins. Aside from their integral contributions to protein development that perform specialized proteolytic cleavage of regulatory subunits, the covalent addition of functional groups of proteins, or the degradation of entire proteins,

PTMs are also very involved in enabling proteins to withstand and recover from temporary environmental stresses (i.e., heat shock, microgravity and many others).

The literature supports evidence of thousands of recently discovered PTMs, many of which may likely contribute similarly (perhaps, even, interchangeably) to protein stress response. Although there are many PTM actors upon the biological stage, our study determines that these PTMs are generally cast into organism-specific, preferential roles. For this contribution of the thesis we study the PTM compositions across the mitochondrial (Mt) and non-Mt proteomes of eleven diverse organisms to illustrate that each organism appears to have a unique list of PTMs, and an equally unique list of PTM-associated residue modification sites (MSs) where PTMs interact with protein.

Despite the present limitation of available of PTM data across different species, we apply existing and current protein data to illustrate particular organismal biases. We explore the relative frequencies of observed PTMs, the MSs, and general amino acid compositions of Mt and non-Mt proteomes. We apply this data to create networks and heatmaps to illustrate the evidence of bias. We show that the number of PTMs and MSs appears to grow along with organismal complexity which may imply that environmental stress could play a role in this bias.

11.2 Introduction

11.2.1 PTMs

It is extremely likely that all proteins naturally undergo some level of structural and, therefore, functional alteration by post-translational modification (PTMs). Although many thousands of PTMs have been discovered (7,308 experimentally identified PTMs and 234,938 putative modifications on 530,264 proteins according to^[148]), there are some (i.e., acetylation, glycosylation, phosphorylation, proteolysis,

lipidation, methylation, nitrosylation, ubiquitination, and others) which commonly interact with proteins at specific modification sites (MSs). The amino acids are at precise locations of the protein chain and their interactions with PTMs inspire changes in protein conformations. PTMs have been shown to play prominent roles in protein alteration for destruction^[32;34], general regulation^[224;312] and stress response^[177;227;281]. In this way, PTMs are able to greatly expand the functional diversity of the proteome and disprove the *one-gene-one-protein* hypothesis.

To add functional diversity and adaption to their alternative environments^[271], proteins may respond to stresses by a transformation of structure and hence, function. For instance, a protein stress may result from an event or treatment which leads to protein failure when the protein is forced to sustain its duties under unnatural circumstances such as environmental stress. Across seemingly all proteins, PTMs offer an extremely rapid solution for withstanding naturally occurring environmental stresses such as microgravity^[17], drought^[108], thermic shock^[223;283] and others. Stress responses resulting from the proteins themselves are also conducted by phosphorylation as in the case of initiation and regulation of tumor suppression by the p53 complex^[177] and SUMOylation for its response to oxidative stress^[227]. Furthermore, by intervening with PTM function, therapies may be created to treat types of cancer^[266] or be used to maintain cellular homeostasis^[87;281]. A modification is immediate as it does not necessitate the re-synthesis of a new protein to cope with environment stresses. Once the stress is removed, PTMs are often able to restore the protein to its previous conformation^[20].

11.2.2 Biases

PTMs also show evidence of preferential treatment. Discussed in Khoury *et al.*^[148], PTM activities from acetylation, glycosylation and phosphorylation were frequently observed in their data, however, there were many PTMs which were rarely exerted

(i.e., FAD, bromination and many others). A PTM bias between related proteins may conveniently be observed using public data such as UniProt^[9]. For example, *Sir1*, also known as the *NAD-dependent protein deacetylase sirtuin1*, is a regulatory protein found in human and mouse. The *Sir1* Human (UniProt: Q96EB6) and *Sir1* Mouse (UniProt: Q923E4) had a listing of 19 and 13 observed PTM interactions, respectively. Although there were 16 phosphorylation sites in *Sir1* Human and only 10 in *Sir1* Mouse, acetylation was observed only in mouse protein. Here, we note that these two similar proteins offer high granularity evidence for the existence of PTM bias between human and mouse.

11.2.2.1 Research Statement

In this contribution, we extend and advance our earlier study,^[35], which described some of the initial patterns of PTM bias inherent in some of the organisms of the present study. We studied the proteomes of eleven diverse organisms shown in Table 11.1 to show that each organism has unique PTM biases and an associated MS bias. We present evidence that the number of observed PTMs and MSs by organism appears to increase with organismal complexity. To clearly describe these biases, we employ heatmaps and networks which are built from relative frequency data that we harvested from parsing data available from UniProt. Since mitochondria (Mt) have unique genomes and therefore unique proteomes, we extend our study of protein PTM biases to these organelles to describe their PTM and MS biases by organism. Since Mt are highly conserved across biology, we show that their PTM employment is not a conserved entity. Finally, we show that the trends of increasing PTM and MS bias are both observed in Mt and non-Mt with similar degrees of clarity.

Table 11.1: Diverse organisms of the study. We noted that glycosylation and phosphorylation were commonly the most frequently occurring PTMs in our data.

Organism	Number of Proteins	Top PTMs
Mustard Plant <i>Arabidopsis thaliana</i>	3155	Glycosylation Phosphorylation
Fungi <i>Aspergillus nidulans</i>	254	Glycosylation Methylation
Nematode worm <i>Caenorhabditis elegans</i>	572	Glycosylation Lipidation
Domestic Dog <i>Canis familiaris</i>	616	Glycosylation Phosphorylation
Zebrafish <i>Danio rerio</i>	622	Glycosylation Phosphorylation
Human <i>Homo sapiens</i>	11884	Phosphorylation Glycosylation
House Mouse <i>Mus musculus</i>	10388	Phosphorylation Glycosylation
European rabbit <i>Oryctolagus cuniculus</i>	641	Glycosylation Phosphorylation
Norway Rat <i>Rattus norvegicus</i>	5413	Glycosylation Phosphorylation
Bakers Yeast <i>Saccharomyces cerevisiae</i>	3013	Phosphorylation Glycosylation
African clawed frog <i>Xenopus laevis</i>	671	Glycosylation Phosphorylation

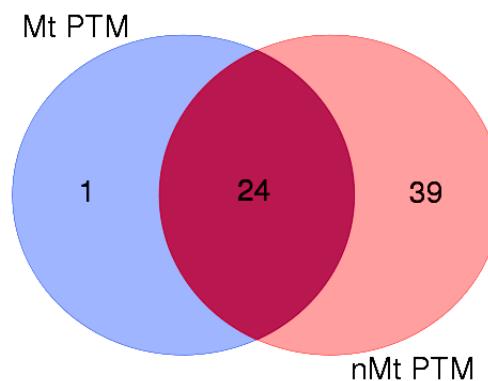


Figure 11.1: A comparison between the number of PTMs in our Mt and non-Mt sequence data. Here we exclude all PTMs that are labeled by UniProt as *InterChain* due to a lack of information available for our study.

11.3 Methods

For the organisms of Table 11.1, protein data was downloaded in June 2015 from UniProt, a public protein knowledge base that provides curated data. At the time of our study, our downloaded set was the most currently available. The curated protein records were divided into Mt and non-Mt sets, depending on their origins for each organism. For every protein of each set, the PTM data was assembled – the type and number of PTM as well as their associated MSs which were often unique to each particular PTM. In Figure 11.1 (created by <http://bioinformatics.psb.ugent.be/webtools/Venn/>), we illustrate the counts of PTMs which were obtained across the organismal Mt and non-Mt sets, taken all together. Since the organization of the Mt genome is highly conserved in insects, as in most other bilateral animals^[37;168], we maintain that the patterns that we were able to find in Mt may likely be extended to other types of organisms as well, although the nuclear proteins may not be similar.

We noted that there were often cases where a specific PTM type was given by UniProt that actually fell into a more general category. For example, N-acetylalanine and N-acetylaspartate are actually two specific types of acetylation. There were many other cases where specific PTMs (often specifically named due to their associated MSs) could be reduced to more general denominations. In order to simplify PTM quantifications during our analysis, we followed the PTM conversion documentation available by UniProt to record the general PTM denominations. These frequently occurring PTMs describe evidence of PTM bias even from a high granularity. Specifically, the order of the first and second most frequently observed PTMs (generally, glycosylation and phosphorylation) were not unanimously conserved across the organisms, as noted in Table 11.1.

11.3.1 Organismal Protein Samples

In Table 11.1, we indicate the actual number of proteins analyzed as well as the two most commonly occurring PTMs from the Mt and non-Mt protein sets of each organism. In the second and third columns of Table 11.2, we display the number of processed Mt and non-Mt UniProt protein records, respectively. In the fourth column, we present the size of the exhaustive list of organism-specific (curated) proteins from UniProt from the time of our study. In this column, there are records containing PTM information, as well as many where PTMs are not discussed. By comparing the numbers of records where PTM information is known to the numbers where it is lacking, it is obvious that much work is yet to be done to complete our knowledge of PTMs.

In the fifth column we illustrate an estimation for the number of scientific articles available from the National Center for Biotechnology Information (NCBI), where PTM information may be extracted to populate protein records with PTM information (likely by UniProt and others). To estimate the number of NCBI articles, we applied the text mining analysis implemented in^[27;47] that served to locate all article abstracts containing relevant keywords: protein names, PTM types and other words for syntax.

The organisms were diverse and represented a wide spectrum of biology^[309]. We divided the protein data between Mt and non-Mt sets. Unlike the Mt genome which may be highly conserved across biology, the non-Mt protein is generally more diverse and may be more revealing of natural bias from organism to organism. Proceeding protein by protein for each organism, we determined the types of PTMs, the count of each and their associated MS type. We note that although there are many different kinds of PTMs in nature, we restricted our study only to those PTMs that have been observed to interact with single amino acids (i.e., a length-1 motif) along the protein sequence (a single MS). Figure 11.2 describes the procedure for capturing the data

Table 11.2: The table to show the number of protein records by organism available for our work. The second and third columns display the number of Mt and non-Mt UniProt protein records, respectively. The forth column describes the exhaustive number of protein records where PTM are discussed in some of the articles. The fifth column provides an estimation of the number of scientific articles from the literature that may have been sources of PTM information for protein records. This data was furnished by text mining the NCBI body of literature.

Organism	Unique PTM records (Mt)	Unique PTM records (non- Mt)	Total protein records (Mt and non-Mt)	Total NCBI articles
<i>A. thaliana</i>	116	3809	13943	260
<i>A. nidulans</i>	4	283	914	21
<i>C. elegans</i>	16	711	3537	339
<i>C. familiaris</i>	24	613	812	2
<i>D. rerio</i>	22	611	2945	8
<i>H. sapiens</i>	589	11419	20207	191
<i>M. musculus</i>	564	10032	16718	21
<i>O. cuniculus</i>	30	661	889	0
<i>R. norvegicus</i>	374	5115	7923	7
<i>S. cerevisiae</i>	212	3005	7900	461
<i>X. laevis</i>	22	692	3394	67

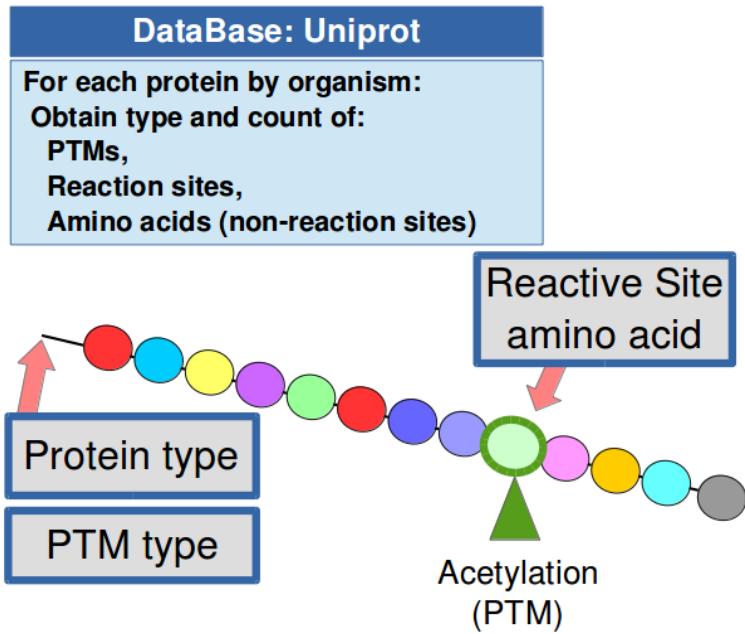


Figure 11.2: All Mt and non-Mt proteins were examined in each organism of our study. We recorded the protein type (Mt or non-Mt), the PTMs of the protein and their associated MSs. This information was used to assemble relative frequency data.

which we then used to calculate frequencies (explained in Section 11.3.2). We note that the data used to calculate these frequencies may have had incomplete references due to the general difficulties of extracting PTM information from physical protein samples in a wet lab. In light of such a limitation, however, we believe that the design of this study is still worthy of providing detailed patterns of PTM bias across the organismal data. Furthermore, as more data becomes available, our method may again be applied to discover new patterns.

11.3.2 Computing Frequencies

Due to the common hardships of applying limited computing resources to processing voluminous quantities of data, a statistical analysis is often appropriate^[28;30]. Furthermore, frequency analysis is especially well suited for comparing large datasets and discovery as it embraces convenient techniques of network analysis to

ascertain natural patterns^[33;296]. Here we discuss the collection of frequency information which is later used to build networks to discover PTM and MS biases.

We used relative frequencies to determine all PTM occurrence magnitudes for elements that have been observed to interact within the Mt and non-Mt proteomes of an organism. We note that frequency distributions are collected in isolation for each proteome of each organism. This implies that the frequency distribution of any proteome may be compared to any other distribution. All records of proteins were downloaded from UniProt which were parsed using an in-house program. Across all the organisms of Table 11.1, we made a tally of the number of PTMs and MSs that had been observed throughout the proteins of each proteome for each organism. Additionally, we also collected the occurrence magnitudes for each non-MS amino acid for the later comparison of MS distributions to ordinary amino acids in each proteome.

We note that PTMs have specific names which generally imply information about their MSs. In our work, we generalized the PTM names into basic rubrics (i.e., N-acetylalanine, N-acetylaspartate, N-acetylatedlysine and N-acetylcysteine are all kinds of acetylation) since we were also collecting the associated information about MSs. Once all the proteins of an organism were parsed for their PTM, MS and amino acid tallies, we applied this data to three equations to derive relative frequency information. Using Equation 11.1, we calculated the PTM frequencies. This equation determined the occurrence magnitude of each unique type of PTM by dividing the number of its counts into the combined number of all observed PTMs by proteome. For example, glycosylation generally appeared many times in a proteome and our calculation combined all its observations by proteome to create one relative frequency value.

The information concerning PTM and MS interactions was recorded. Similar to how we calculated PTM frequencies, we employed Equation 11.2 to calculate MS

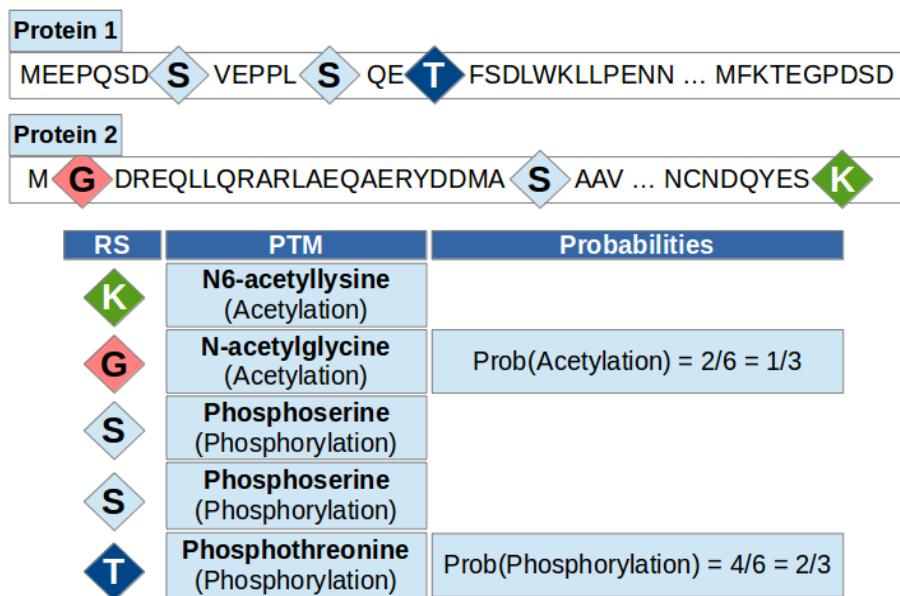


Figure 11.3: An example of how relative frequency information was extracted from protein data. For each organism, all Mt and non-Mt protein records were queried to ascertain their observed PTMs that have been curated by UniProt. The type and count of each PTM, including its associated MS was recorded to calculate frequencies by Equations 11.1 and 11.2. Not shown, the occurrence magnitudes of all amino acids (non-MSs) were also obtained and applied to Equation 11.3 to determine the general amino acid compositions of each proteome.

frequencies. The relative frequency of a particular MS type was found by dividing its tally into the combined number of all observed MSs by proteome. Visualized in Figure 11.3, a count of each PTM type was created for each organism and the PTM frequencies were calculated from this information in each of the Mt and non-Mt protein datasets for each organism. We used this information to populate Table 11.1. We noted an apparent preference for individual PTMs across the organisms. For instance, although glycosylation and phosphoserine were popular PTMs for many organisms, they do not appear to always achieve the same first and second rankings in the organisms. In addition, we noticed that *Caenorhabditis elegans* was the only organism of our set which had a high frequency for lipidation and *Aspergillus nidulans* was the only organism to exhibit methylation.

We now discuss the equations. Across each organism j , for a specific element i (i.e., PTM, MS, or Amino Acid), the relative frequency of a particular $PTM_{(i,j)}$ and its associated modification site, $MS_{(i,j)}$, were calculated by Equations 11.1 and 11.2, respectively. We note the use of the `count()` function which determines the number of occurrences of the element in the current dataset. Across all PTMs of organism j , the relative frequency of a particular $PTM_{i,j}$ may be found by the following;

$$\boxed{freq(PTM_{(i,j)}) = \frac{count(PTM_{(i,j)})}{\sum_{i=1}^{N_{(PTMs)}} count(PTM_{(i,j)})}} \quad (11.1)$$

Across all reactive sites found associated with the PTMs of organism j , the frequency of a particular amino acid modification site, $MS_{i,j}$, may be found by the following equation.

$$\boxed{freq(MS_{(i,j)}) = \frac{count(MS_{(i,j)})}{\sum_{i=1}^{N_{(MS)}} count(MS_{(i,j)})}} \quad (11.2)$$

The counts of each amino acid of each proteome were also tallied to determine relative frequencies for each organism, j . Akin to simply placing all the protein

sequences of a proteome end-to-end to create one sequence, Seq , we determined the amino acid composition and frequencies using Equation 11.3.

$$freq(AA_{(i,j)}) = \frac{count(AA_{(i,j)})}{|\sum_{i=1}^{N_{Proteins}} Seq_{(i,j)}|} \quad (11.3)$$

11.3.3 Building Heatmaps And Networks

Heatmaps: A heatmap is a color-coded matrix of numerical values which have been clustered across the top and the side. We used heatmaps to determine the amino acid compositions across proteomes for comparison to the PTMs biases in their proteomes. Heatmaps are useful in comparing different large sets of data together in terms of their frequency or other numerical information. Our general heatmaps of Figures 11.4 and 11.5 were created from the relative frequency data from Equations 11.1 and 11.2, respectively, and applied to the method described by^[155]. Equation 11.3 was utilized to calculate the frequency of occurrence of each amino acid, regardless of also being an MS. These results are shown in Figure 11.6.

Since PTMs (i.e., phosphorylation and others, for example) may interact with several different MSs simultaneously^[323], we determined that the details of their relationships would be obvious when described in networks where individual interactions between PTMs and MSs may be explored in detail.

Networks: Our networks were built from relative frequency data by applying Equations 11.1 and 11.2 using^[261]. In the networks of each proteome, we determine the frequency magnitude by the size of the node: larger nodes describe more common occurrences. The left and right sides of the networks represent the PTM and MS populations (respectively) which were found in a proteome. The edges between the PTMs and MSs were calculated by the product of the PTM and MS frequencies. Since an interaction is not mutually exclusive, this calculation describes the interaction magnitude between the pair. Here, we note that the heavier edge weights describe

more common interactions. The networks are read from the left side PTMs, which interact with the MSs on the right side. We summarize the main results from the networks in Table 11.5 and Figure 11.18.

11.4 Results And Discussion

11.4.1 Heatmaps

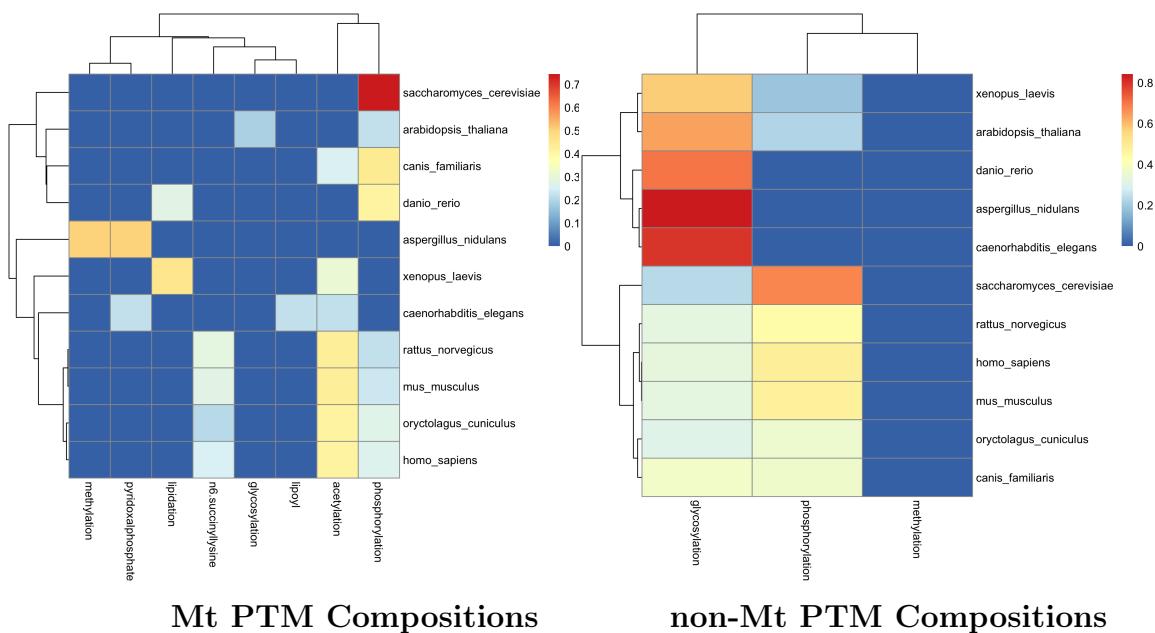


Figure 11.4: Mt and non-Mt PTM compositions prepared using Equation 11.3. High magnitudes of frequency are described by warmer colors. We note that phosphorylation and acetylation were common PTMs across the organisms. We note that all frequency values greater than 0.18 (threshold) are included here.

We note that a single PTM observed in an isolated protein in the proteome may provide misleading information about its relevance to the proteome. We therefore, apply a threshold to the frequency value to be able to distinguish the higher frequencies from the lower ones. In the Mt set, the range of PTM frequencies was 0.0003 to 0.21 and in non-Mt, the range was 0.0004 to 0.5 and we therefore defined the threshold to be 0.18, or the average of the midpoints of both ranges.

In Figure 11.4, we display the Mt and non-Mt heatmaps. Here, the counts of PTMs were eight and three for Mt and non-Mt, respectively. It is interesting to note that in non-Mt, the three PTMs, glycosylation, phosphorylation and methylation are some of the more adaptive PTMs that are able to modify many different types of proteins^[211] and have been observed to commonly interact together. We note from these heatmaps that related organisms generally appeared to have similar types and frequencies of PTMs. For example, the mammals of our data, *Rattus norvegicus* (rat), *Mus musculus* (mouse), *Oryctolagus cuniculus* (rabbit) and *Homo sapiens* (human) are closely clustered according to their PTM frequencies. In non-Mt, all mammals, including *Canis familiaris* (dog), were clustered together with the inclusion of *Saccharomyces cerevisiae* (yeast). Although Mt are highly conserved across organisms, we find that there is enough difference between PTM populations in the data to suggest that sequence similarity may not play much of a role. Extending this idea to the non-Mt protein data, we suggest that the clustering of mammal data in Figure 11.4 could be due to environmental conditions.

Mt is highly conserved across organisms and so we may expect to see less diversity in PTMs in this set, however we found that of the eight PTMs (shown in Figure 11.4), only glycosylation and phosphorylation were also common to the non-Mt set. The other PTMs may be involved in Mt-specific activities such as lipoyl for metabolism^[256] and acetylation that has been known to target large macromolecular complexes involved in diverse cellular processes for regulation^[63].

In Figure 11.5 we note the associated amino acids which play roles as MSs in Mt and non-Mt. No frequency threshold was necessary since nearly all amino acids had strong frequencies (note the warmer colors). We observed that there was much more variety in selection for MSs in the non-Mt set than the Mt set. This signifies that there is more promiscuity in terms of PTMs interacting with diverse MSs in non-Mt and suggests that amino acids may have few restrictions in terms of their roles with

PTMs. The PTMs of Mt appeared to interact with a specific type of MS. We will return to this observation in the networks where we will see that PTMs generally interact with multiple types of MS in the non-Mt set.

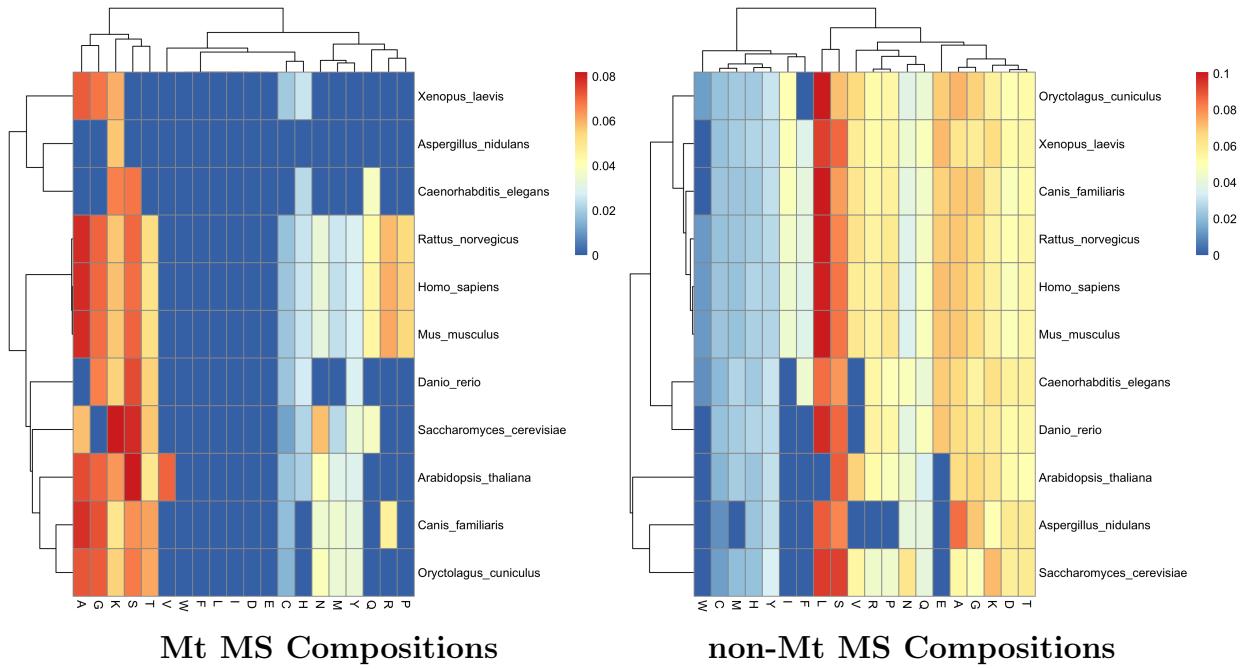


Figure 11.5: Mt and non-Mt MS compositions prepared using Equation 11.2. High magnitudes of frequency are described by warmer colors. Unlike the non-Mt heatmap where nearly all amino acids played a roles as MSs, there were many AAs in the Mt proteomes that were never involved with the PTMs.

In Figure 11.6, we note the compositions of all amino acids across the organismal proteomes. We note that the prominent MSs of Figure 11.5 are not necessarily the prominent amino acids of the same organisms. For instance, by the heatmap in the *Homo sapiens* Mt proteome, the glycine (G), alanine (A), serine (S) and lysine (K) were prominent amino acids as MSs, however, they are not so as amino acids. The highest frequency magnitude in this organism was leucine (L) which was not found to be an MS. There are other similar observations to make from these heatmaps.

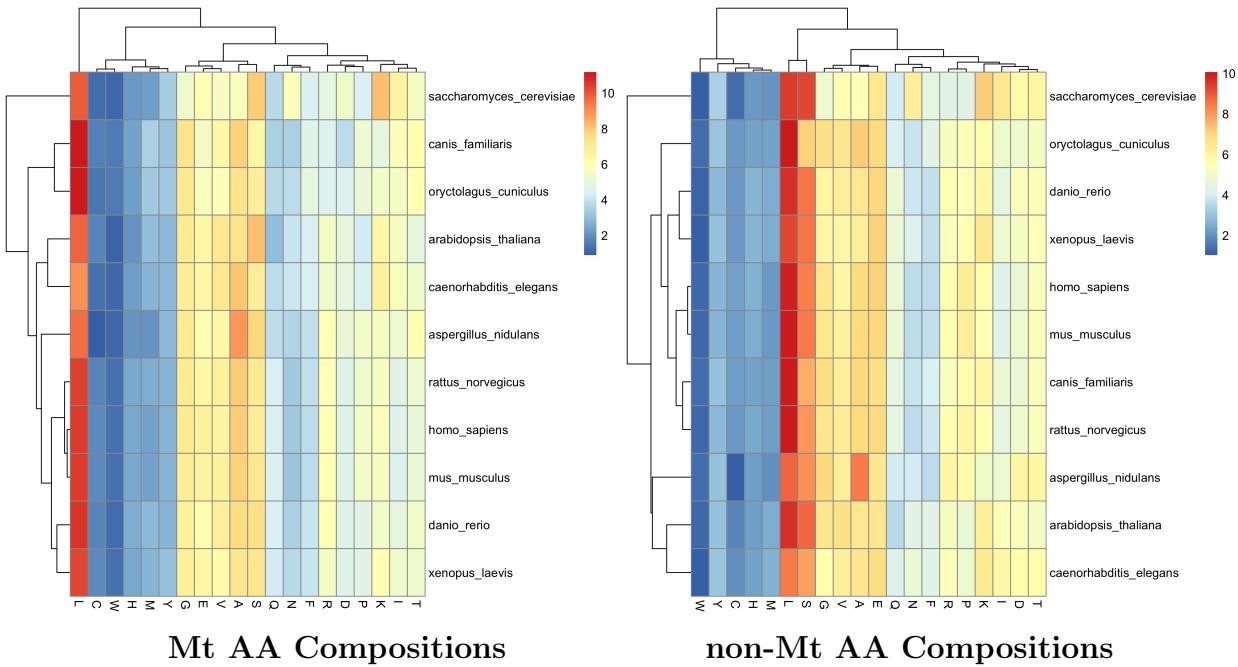


Figure 11.6: Mt and non-Mt amino acid compositions prepared using Equation 11.3. High magnitudes of frequency are described by warmer colors. Although all organisms display a common theme of color bands indicating that their amino acid composition is similar, we note that related organisms have especially similar patterns of color indicating that the amino acid distributions are similar.

11.4.2 Domains And PTMs

At the heart of protein function are domains: the conserved parts of protein functional structures which can evolve and exist independently of the rest of the protein chain. The functions of domain structures are thought to be context dependent and directed by PTM activity (i.e., phosphorylation)^[191]. The difference between the prominences of the MS and amino acid frequencies supports evidence that the location of an MS (perhaps, found near protein domains) may be more important than its basic biophysical properties. For example, in human, phosphorylation of p53 occurs at thirteen serine and five threonine amino acids which are distributed in the protein's (functional) domain regions^[107].

Protein conformations by PTMs create changes in behavior. For example, the DNA binding domain of p53 is heavily influenced by changes in conformation from

ubiquitination^[107]. In^[43] and^[173], phosphorylation has been observed to disrupt FoxOs interaction with 14-3-3 proteins (likely at ww-domains^[263]) to allow nuclear translocation of FoxO^[2] and initiate programmed cell death (apoptosis).

PTMs that influence domains have been studied in the context of heart failure and arrhythmia as a result of functional defects in cardiac type 2 ryanodine receptors on the internal sarcoplasmic reticulum (SR). Specifically, the disease of this contractile protein (muscular) machinery has been attributed to regulation failure of the Calcium (Ca^{2+}) release channels in the SR. Shao *et al.* discovered that carbonylation (a PTM) may be responsible for the (Ca^{2+}) dysfunction which was observed to disable two main lysine amino acid sites (at positions 2190 and 2887), flanking the RyR2 (ryanodine receptor: a Ca^{2+} release channel) subdomain site^[269]. By disabling these lysine sites, the N-terminal and central protein domains of RyR2 (near two subdomains at position 2000 - 2500 and 2234 - 2750) were observed to be destabilized and unable to properly regulate Ca^{2+} for normal muscle function.

In a related study,^[268], the activity of SERCA2a (a protein that undergoes a series of timed conformational changes to hydrolyze ATP and transport Ca^{2+} ^[290]) was studied in heart tissue. Here, the authors found that Ca^{2+} transport (regulated by the SERCA2a) may be reduced or disabled when amino acid sites are neutralized. For instance, four sites were studied which reside in the protein's domains; A-domain: {R164}, N-domain: {K476, K481}, and P-domain: {R636}. The study found that Ca^{2+} transport was reduced or prevented by the paired modification (carbonylation / charge neutralization by conversion to glycines) of {R164, K481}, {K476, R636}, and {R164, R636} (carbonylation / charge neutralized by conversion to tyrosines) to suggest that these amino acids behave as functional switches. In addition, the carbonylation / charge neutralization of {R164, K476, K481, R636} by conversion to tryptophan (which also increased the hydrophobic bulk of the sites), reduced the ability of SERCA2a to transport Ca^{2+} . The above two findings support the notion

that amino acids, making-up MSs residing near or inside protein domains, may be modified by PTMs to regulate the functions of domains. In our next section, we use networks to help us visualize some of these kinds of interactions.

11.4.3 Networks

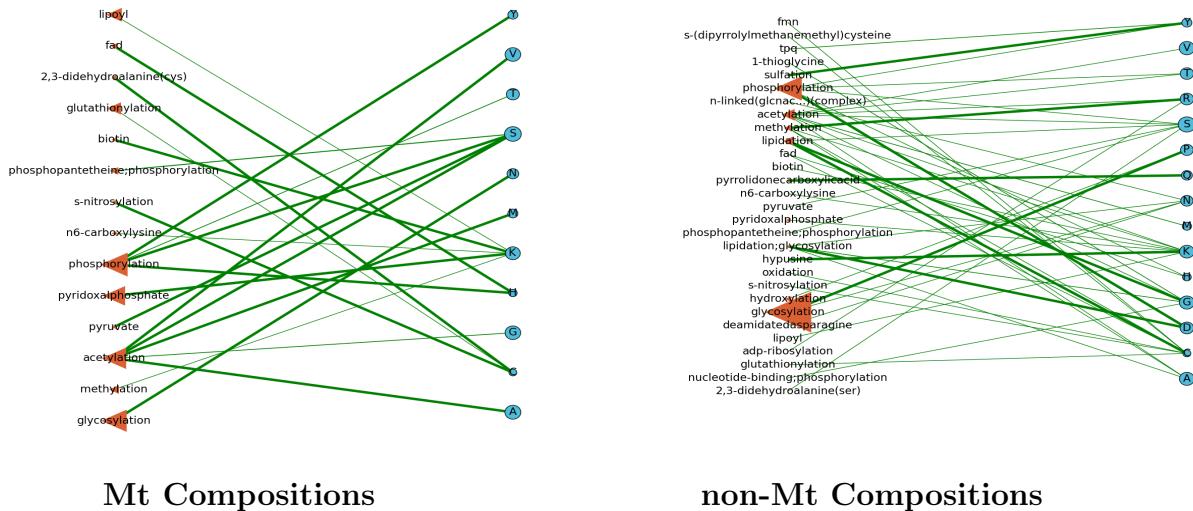


Figure 11.7: *Arabidopsis thaliana*: A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.

Network models help analyze the complex relationships between the interacting entities of our study. The models of Figures 11.7 to 11.17 (inclusive) describe the frequencies of PTM and MS occurrences, and the frequencies of their interactions. In Figure 11.18, we summarize the results by showing the number of PTMs and MSs per organism in effort to describe their increase by organismal complexity. We show two networks for each organismal proteome, each of which have two types of nodes: PTMs (left) which interact with the MSs on the right. In these networks, we may determine the number of different interactions that a particular element may have by

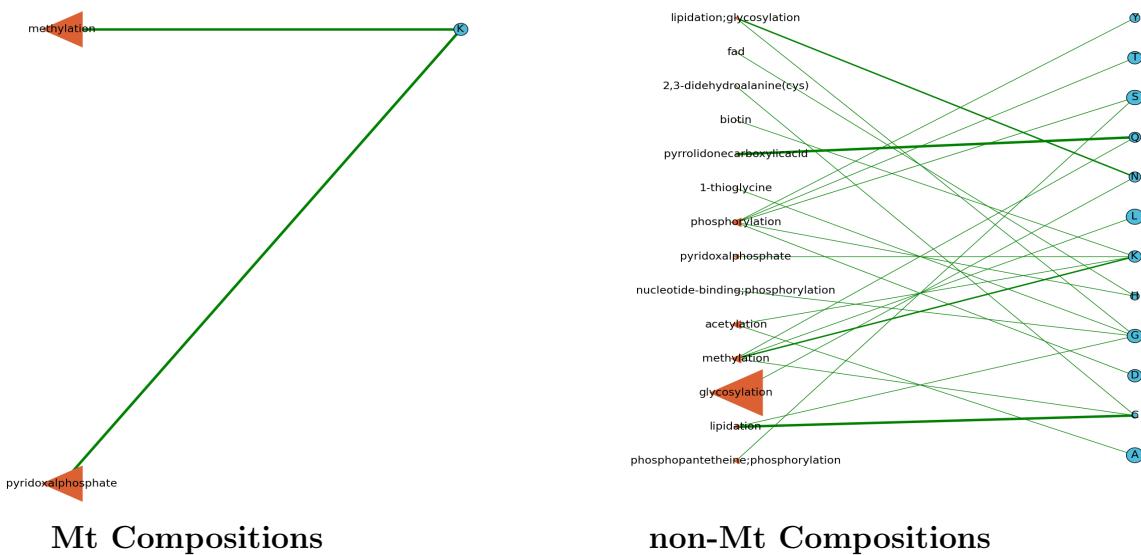


Figure 11.8: *Aspergillus nidulans*: A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.

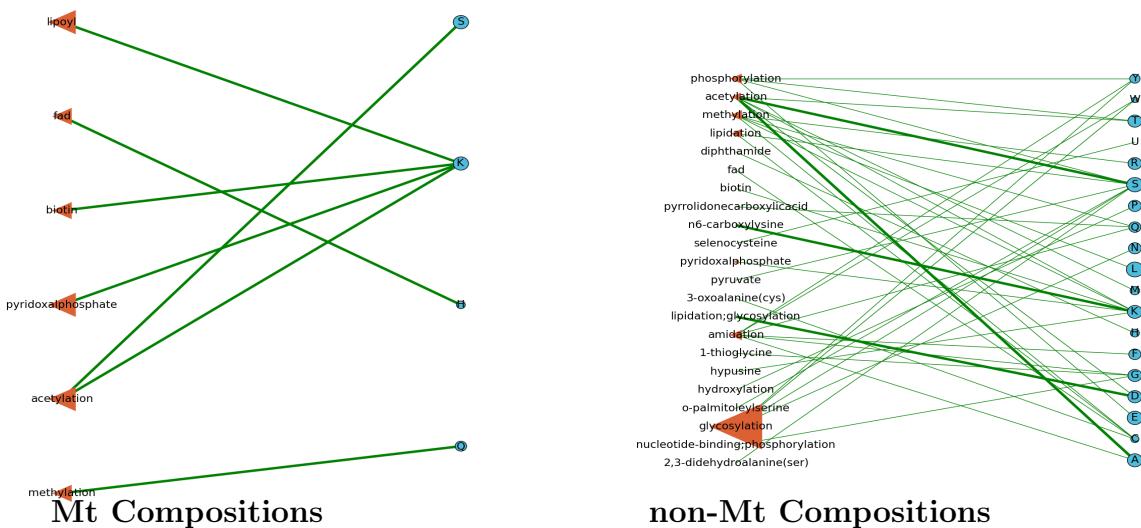


Figure 11.9: *Caenorhabditis elegans*: A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.

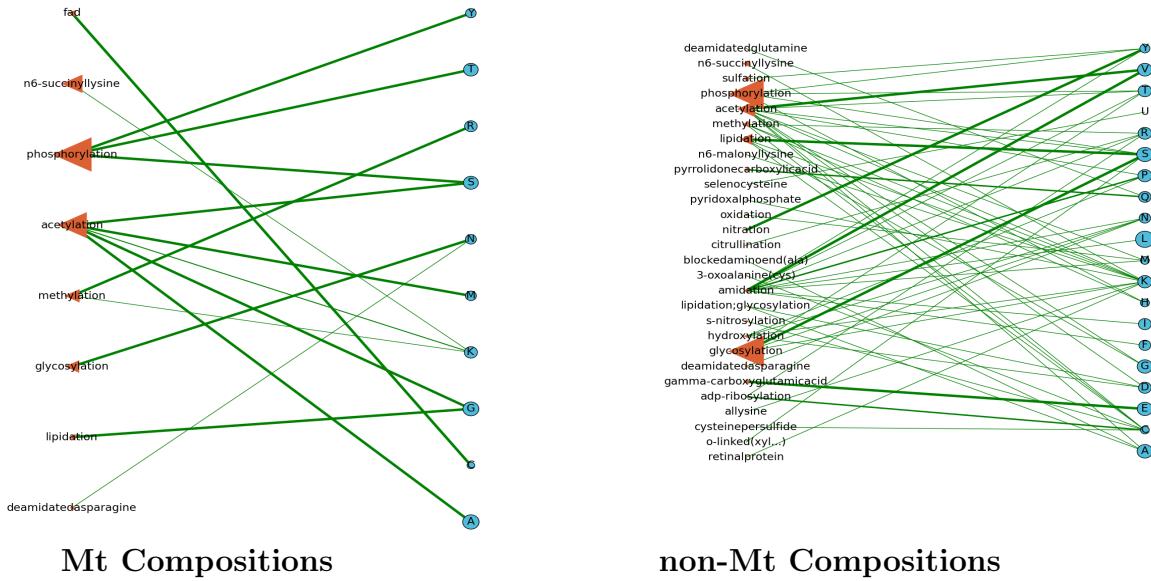


Figure 11.10: *Canis familiaris*: A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.

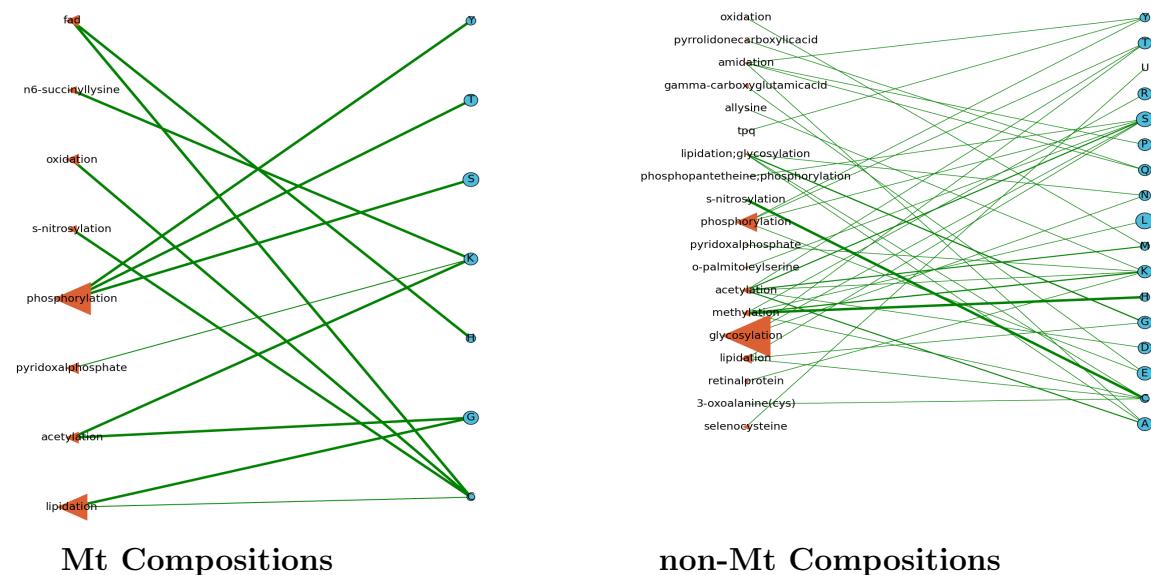


Figure 11.11: *Danio rerio*: A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.

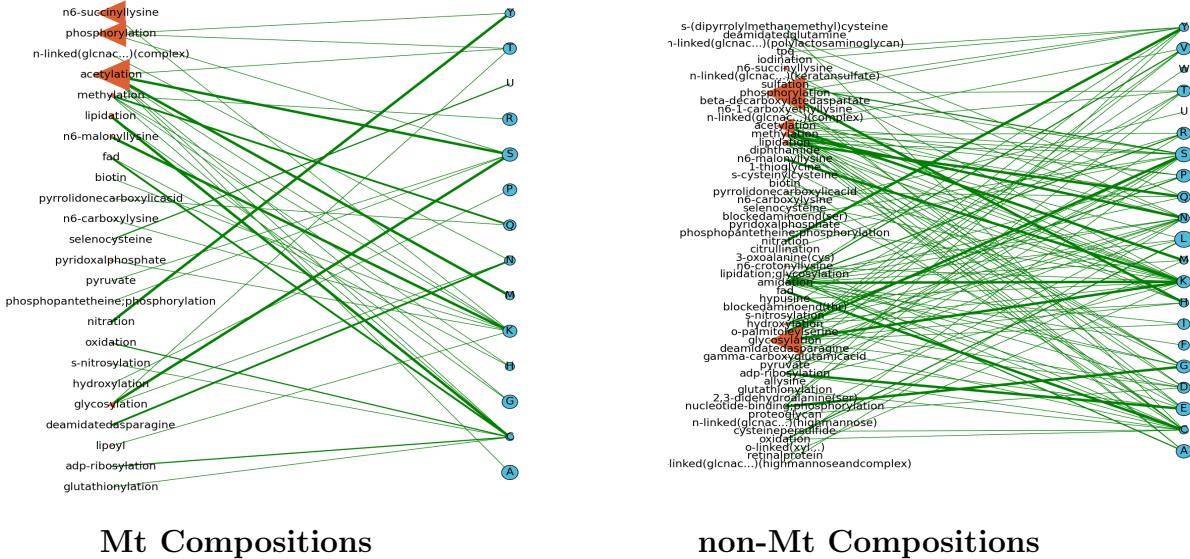


Figure 11.12: *Homo sapiens*: A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.

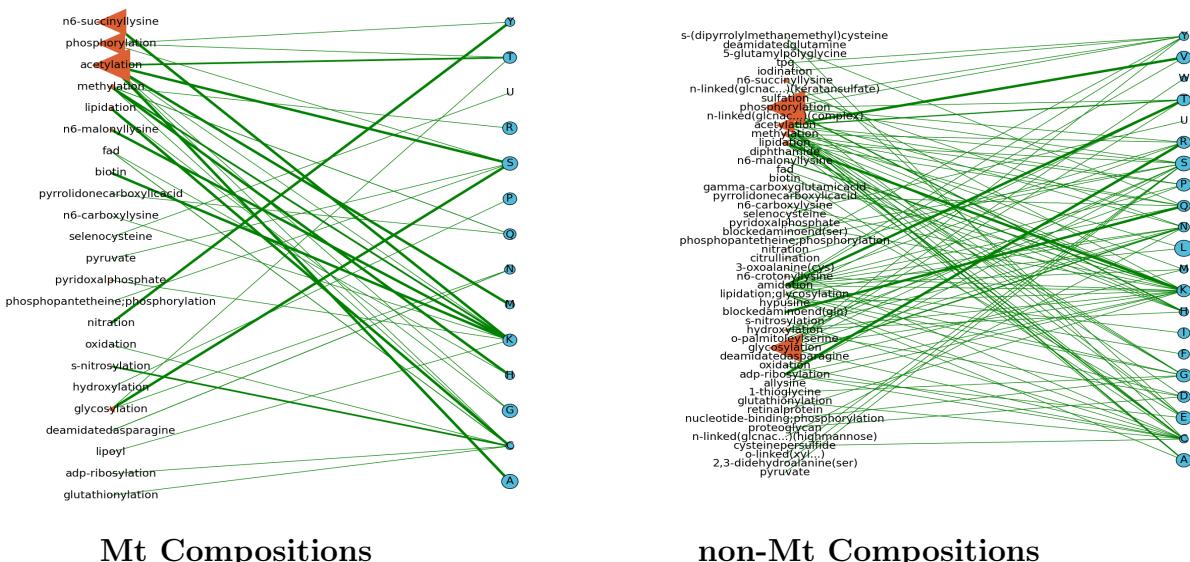


Figure 11.13: *Mus musculus*: A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.

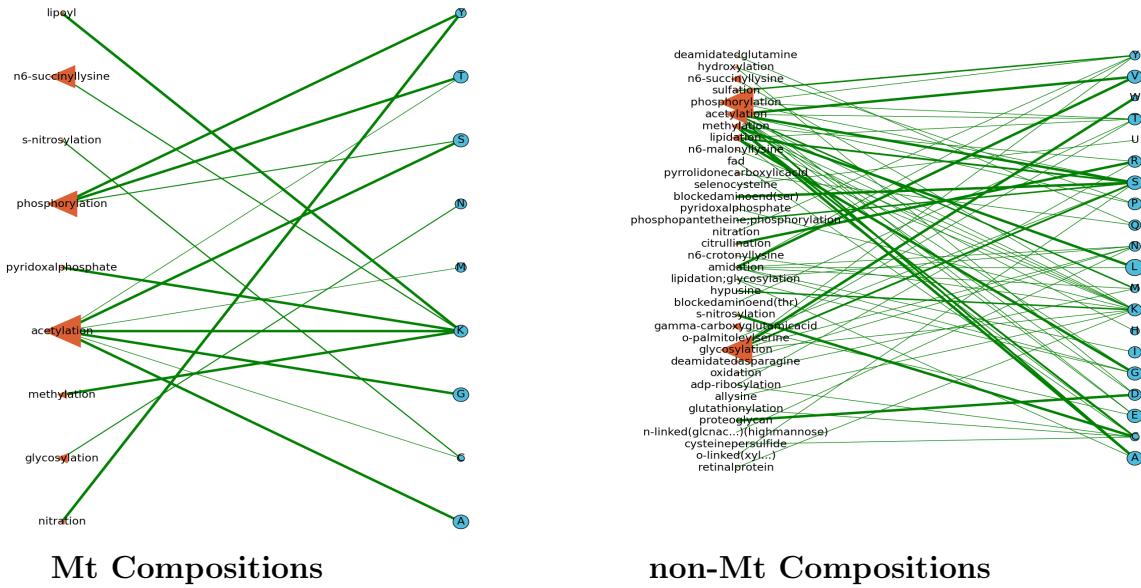


Figure 11.14: *Oryctolagus cuniculus*: A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.

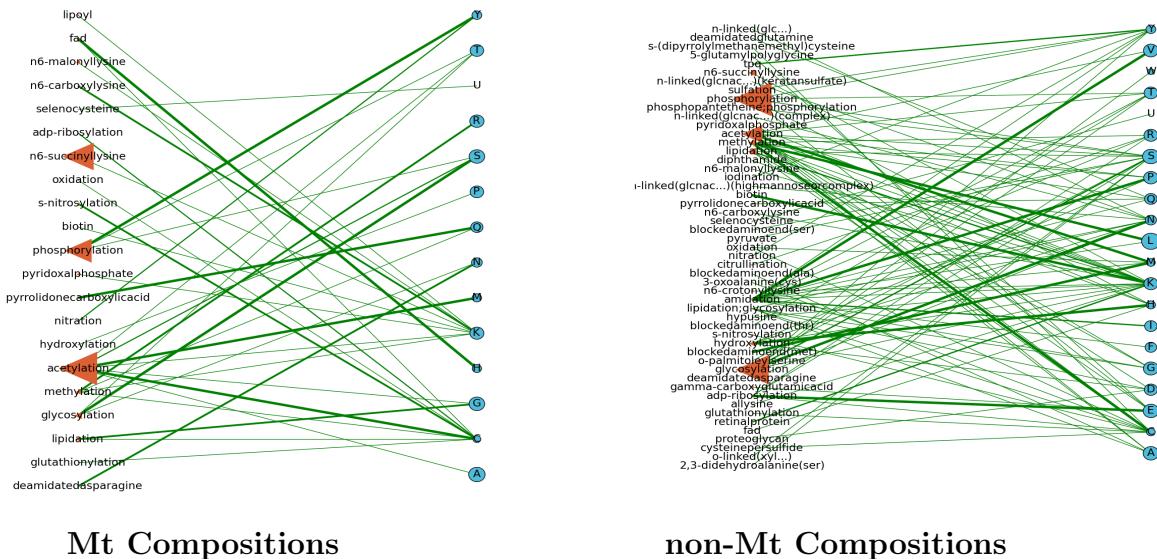


Figure 11.15: *Rattus norvegicus*: A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.

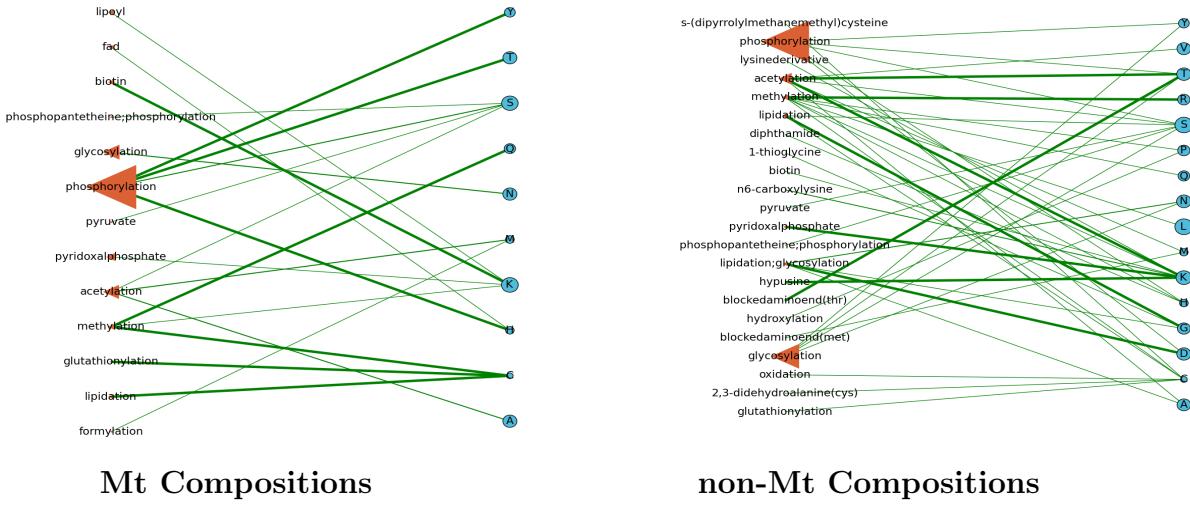


Figure 11.16: *Saccharomyces cerevisiae*: A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.

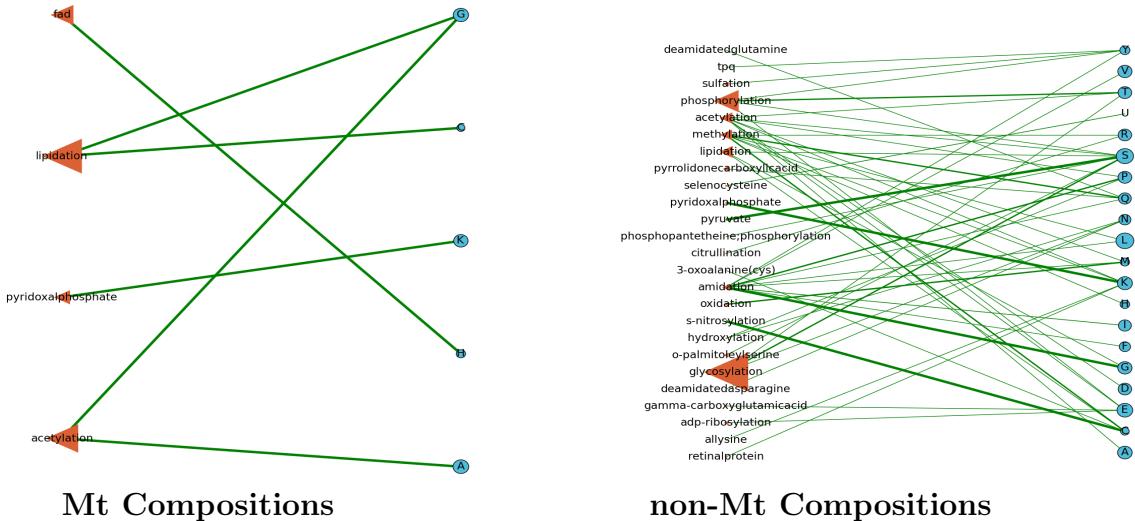


Figure 11.17: *Xenopus laevis*: A Network of PTM frequencies in Mt (left) and non-Mt (right) protein. We display the prominent PTMs across all organisms of our study having a frequency of at least 0.1.

studying the degree (or number of connections) stemming from its node. In addition, the weight of the edge describes the relative frequency of the interaction between two

elements.

11.4.4 Pearson's And Kendall-Tau Correlation

Due to the insufficient number of data points, we used both parametric (Pearson's) and non-parametric (Kendall-tau) correlation tests to calculate a statistic between the set of organism rankings and the sets of PTM and MS counts, for each protein type (i.e., Mt and non-Mt). We tested Hypothesis 5, 6, 7 and 8 and our summarized data can be found in Table 11.3. In the table, we show that the set of complexity rankings (1 through 11) is consistently correlated to PTMs of Mt protein and MSs of non-Mt protein.

Null Hypothesis 5. *No relationship exists between the organism rank and the PTMs in Mt.*

Null Hypothesis 6. *No relationship exists between the organism rank and the MSs in Mt.*

Null Hypothesis 7. *No relationship exists between the organism rank and the PTMs in non-Mt.*

Null Hypothesis 8. *No relationship exists between the organism rank and the MSs in non-Mt.*

11.4.4.1 Mt And Non-Mt Networks

Although there were often cases of common PTMs observed between Mt proteomes (i.e., phosphorylation and acetylation), the Mt networks themselves presented a decided difference between kinds and types of PTM usages. In terms of PTM and MS connections from organism to organism, we found low to high levels of interactions between these nodes. For instance, in Figure 11.8 (*Aspergillus nidulans*)

Table 11.3: The Pearson and Kendall-tau correlation values of implied organism complexity, and PTM and MS magnitudes. In this table, we assigned complexity values based on inspection of the general number of connections between PTMs and MSs in the networks. We computed a parametric (Pearson) and a non-parametric (Kendall-tau) correlation coefficient between the set of organism ranks and the set of PTM counts for each protein type (i.e., Mt and non-Mt). These two different correlation tests were employed to provide a wider view of correlation between the limited number of data points of each set. Three out of the four Pearson and Kendall-tau rank correlation tests were found to be significant. However, in both tests, the Mt and non-Mt PTMs were consistently correlated with the organism rankings. The Mt and non-Mt MSs were correlated, but not consistently across both tests.

	Mt PTMs	MSs	Non-Mt PTMs	MSs
Hypothesis	5	6	7	8
Pearson Correlation	0.970	0.935	0.850	0.40
Significance ($p < \alpha$)	0.01	0.05	0.01	0.05
Pearson correlation significant?	yes	yes	yes	no
Kendall tau Rank Correlation	0.661	0.213	1.000	0.841
2-Sides p-Value	0.006	0.441	0.000	0.001
Significance ($p < \alpha$)	0.01	0.05	0.01	0.05
Kendall-tau correlation significant?	yes	no	yes	yes

(fungi), the Mt proteome had only two Mt PTMs (methylation and pyridoxalphosphate) which interacted at the same lysine site. This was a sharp contrast to the Mt proteome of *Homo sapiens* (Figure 11.12), where over 20 Mt PTMs were observed interacting with a host of different types of MSs and seven (of this set) were observed to interact only with lysine. In Figure 11.9, *Caenorhabditis elegans* (worm), has six Mt PTMs interacting with a variety of MSs, of which four interacted with lysine. In Figure 11.7, *Arabidopsis thaliana* (mustard plant), had 14 Mt PTMs, of which five interacted with lysine. The networks describe many other similar features in terms of PTMs and MSs to show that the employment of PTMs in the above fungi, worm and human proteomes are quite different.

For each organism, the rule of always having fewer Mt PTMs when compared to non-Mt data had few exceptions. Such an exception may be noted between the comparison of *Mus musculus* (mouse) and *Rattus norvegicus* (rat) of Figures 11.13

and 11.15, respectively, which showed that mouse had more Mt PTMs but less non-Mt PTMs than Rat.

From their enlarged node sizes throughout nearly all Mt networks, acetylation and phosphorylation were prominent PTMs (often including glycosylation) that tended to interact with a few specific MSs. In the non-Mt networks, glycosylation and phosphorylation, were prominent PTMs which tended to interact with a diverse set of MSs. We refer the reader once again to the Figures 11.4 and 11.5 (heatmaps of PTM and MS compositions, respectively) to note the prominence of the above PTMs. We direct the reader to Table 11.5 and Figure 11.18 to show that higher organisms tended to utilize more PTMs and associated MSs than the others.

Table 11.4: A ranking of proteomes in terms of number of unique PTMs observed (Mt and non-Mt). The gray fields indicate that the ranking is not the same for both the Mt and non-Mt sets. The majority of proteomes have the same ranking in both sets.

Mt Rank	Org	Non-Mt Rank	Org
2	<i>A. nidulans</i>	14	<i>A. nidulans</i>
4	<i>X. laevis</i>	22	<i>D. rerio</i>
6	<i>C. elegans</i>	27	<i>C. elegans</i>
8	<i>C. familiaris</i>	34	<i>S. cerevisiae</i>
9	<i>D. rerio</i>	37	<i>X. laevis</i>
11	<i>O. cuniculus</i>	42	<i>C. familiaris</i>
15	<i>A. thaliana</i>	46	<i>A. thaliana</i>
16	<i>S. cerevisiae</i>	49	<i>O. cuniculus</i>
31	<i>R. norvegicus</i>	114	<i>R. norvegicus</i>
33	<i>M. musculus</i>	131	<i>M. musculus</i>
34	<i>H. sapiens</i>	157	<i>H. sapiens</i>

In Table 11.4, the proteomes have been ranked according the number of observed PTM types. We note that the majority of the proteomes have a similar ranking (see the non-gray cells). We also note that the organisms appear to become more complex as more PTMs are observed in the proteomes. For example, in Figure 11.9, *Caenorhabditis elegans* has fewer PTMs than the mammals such as *Canis familiaris* and *Homo sapiens* of Figures 11.10 and 11.12, respectively.

We remark here that inequality may be the result of the types of environmental stresses which each organism may encounter in its environment. Since PTMs enable proteins to adapt to stress, the number of PTMs in an organism's proteome could be a measurement of the kinds of stresses to which the organism may respond. It may be that there are many habitats where humans and mammals may thrive but are quite lethal to *Aspergillus nidulans*, *Arabidopsis thaliana*, *Caenorhabditis elegans* and *Danio rerio*, such as arid environments. The correlation, between the number of individual PTMs and organismal complexity may help to explain why the environments of lower organisms appear to be specialized in terms of warmth, moisture, humidity and others, in addition to the availability of food sources. Furthermore, in Figure 11.7, we note that *Arabidopsis thaliana* has more PTMs than *Caenorhabditis elegans* (Figure 11.16). This could be explained by the concept that plants and fungi cannot easily remove themselves from their hostile environments and must, therefore, be able to survive their stresses by applying their arsenal of PTMs. In Table 11.5, we note the counts of PTMs and MSs which exist according our data and UniProt. This data is also displayed in Table 11.5 and as a scatter plot in Figure 11.18.

Table 11.5: The number of PTMs and MSs, associated with each organism. Organisms without PTM or MS data are absent from this list. These results are displayed as a scatter plot in Figure 11.18.

Num	Organism	PTMs in Mt	MSs in Mt	PTMs in Non-Mt	MSs in Non-Mt
1	<i>A. nidulans</i>	2	0	14	10
2	<i>X. laevis</i>	4	5	26	20
3	<i>C. elegans</i>	7	4	23	19
4	<i>C. familiaris</i>	8	10	28	20
5	<i>D. rerio</i>	9	7	20	17
6	<i>O. cuniculus</i>	10	9	38	20
7	<i>S. cerevisiae</i>	14	10	23	16
8	<i>A. thaliana</i>	15	11	30	15
9	<i>R. norvegicus</i>	24	14	53	20
10	<i>M. musculus</i>	25	14	52	20
11	<i>H. sapiens</i>	26	14	56	20

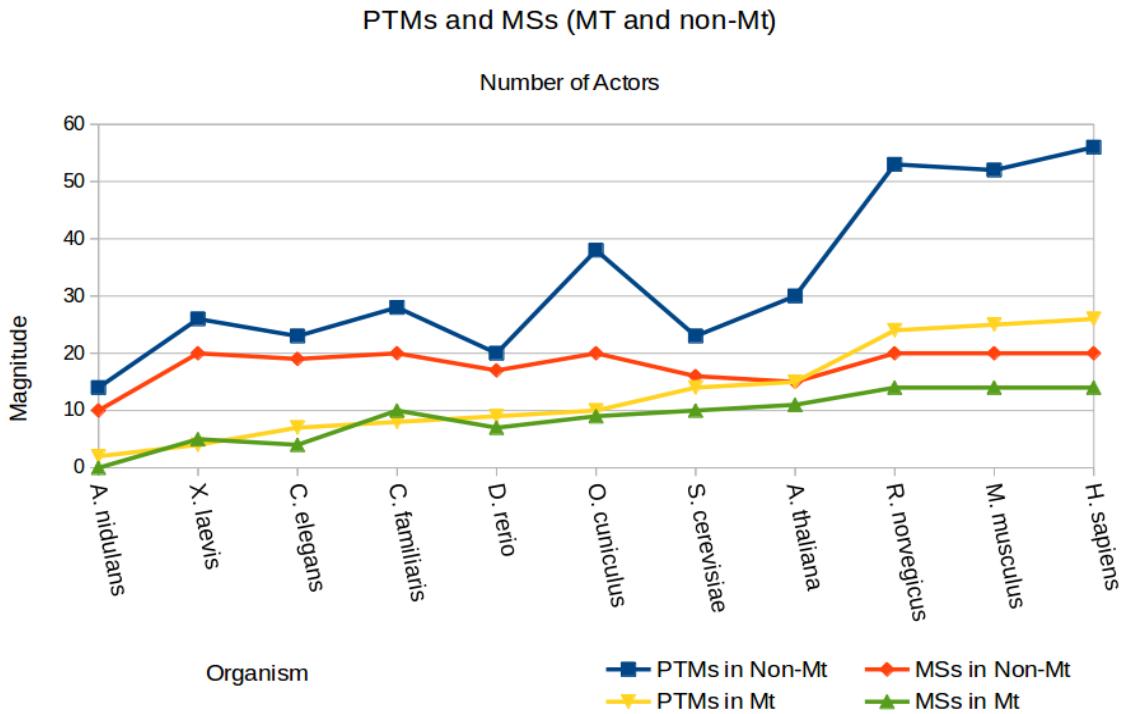


Figure 11.18: The number of PTMs and MSs, associated with each organism. This data is also shown in Table 11.5.

11.4.5 Bias Analysis Using Gene Ontologies

Environmental stresses may force proteins to alter their conformation by PTMs during stress responses. Because each protein has a specific function, they are likely to react in diverse ways to stresses. Some proteins, such as those which are found in Mt, may also have evolved adaptations to easily cope with types of oxidative stress^[34]. To determine the types of PTMs which are involved with particular stresses, we studied the *Homo sapiens* Mt and non-Mt proteins for their functions, as a function of their PTM interactions. We chose acetylation and phosphorylation since they were commonly encountered PTMs and would likely to be found in many different proteins. We made a list of all Mt and non-Mt proteins which were found to interact with acetylation (at least once) and another list of Mt and non-Mt proteins interacting only with phosphorylation (at least once). Using the method by^[303] we extracted a

list of functions from each protein on the acetylation and phosphorylation lists and applied them to Venn diagrams, found in the supplementary data: Figures 1 and 2, and in Tables 1 to 16.

We determined that PTMs interacted with proteins which had some function in stress response. According to the diagrams, in Mt proteins, both acetylation and phosphorylation appeared to interact with proteins handling oxidative stress. Acetylation alone was specific to oxidative stresses responses whereas, phosphorylation alone was specific to general cellular responses to oxidative stresses. This is logical to have these PTMs in Mt proteins since these organelles produce cellular energy by oxidative processes.

In the non-Mt proteins, we noted that acetylation was involved uniquely with proteins which regulated cellular responses to stress, signaled apoptosis as a response to oxidative stress and was involved with the cellular response to heat. On the other hand, phosphorylation was typically involved with proteins which function to regulate cellular processes and control kinase signaling pathways, in addition to some general cellular responses to stresses. In the supplementary data, we provide this general functions information for our protein data.

11.4.6 Poisson Approximation By The Chen-Stein Method

Here we discuss the increasing complexity of the networks of each organism of Figure 11.7 to Figure 11.17. The complexity of a particular network increases according to the number of connections that exist between the PTMs and the MSs. Following a statistical approach similar to^[11;236], we provide *p*-values to support the notion that higher organisms tended to have more complex networks (i.e., having more interactions between their PTMs and MS nodes).

We now describe the test. Because more complex networks are characterized by having more PTMs, MSs and the connecting edges between them (PTM-MS pairs),

we used the unique counts of these elements to compare the complexities of networks across the organisms by an, “all-against-all” test. The data we collected is the following: counts of MSs (Mt and non-Mt), counts of PTMs (Mt and non-Mt) and counts of PTM-MS pairs (Mt and non-Mt). All this raw data is available in the supplementary data.

We discuss the test for the connecting edges between the PTM-MS pairs. For each of the sets of data, and that of *Acanthamoeba castellanii* for comparison, one count total for each of the 12 organisms was given. Let x_1, x_2, \dots, x_{12} denote these 12 counts. We assume each x_i is the sum of independent Bernoulli (binary) random variables. The numbers of such random variables equals the number of the (PTM, MS) combinations, which we will denote N .

We calculate each x_i by the following. Where, $i = 1, 2, \dots, 12$, and $j = 1, 2, \dots, N$, we used the following equation.

$$B_{i,j} = \begin{cases} 1 & \text{if PTM acts upon MS} \\ 0 & \text{otherwise} \end{cases}$$

$$x_i = \sum_{j=1}^N B_{i,j},$$

(11.4)

It is well known that if N is large, we may approximate the distribution of x_i by either the Poisson distribution or a normal distribution. We also assume that independence of x_1, x_2, \dots, x_N . For each pair of organisms (I and J) and provided that not both x_i and x_j equal zero, we compute an absolute Z-value $|Z_{I,J}|$ where,

$$Z_{I,J} = \frac{x_i - x_j}{\sqrt{x_i + x_j}}$$

(11.5)

Here we used the consistent approximation that the variance of x_i is (or is very close to) x_i . In fact, x_i is an estimated upper bound on the variance of x_i since this

variance is less than the mean of this variable. We therefore note that the $|Z_{I,J}|$ value is a *conservative* test statistic.

We have also used a Bonferroni inequality adjustment for simultaneous comparison of all 66 pairs of organisms, which makes these statistical tests even more conservative. The Poisson approximation element of the test can be used even in some cases of dependence. The conservative nature of the tests should allow for more than just slight dependence. Organisms I and J were considered different in terms of their complexity if the two-sided p -value was less than $\frac{\alpha}{66}$, where α is the level of significance. For our purposes, $\alpha = 0.05$ was sufficient and the p -value for a single test for a particular pair I,J is given by the following.

$$p\text{-value} = 2 \int_{|Z_{I,J}|}^{\infty} \frac{1}{\sqrt{2\pi}} \exp^{-\frac{x^2}{2}} dx \quad (11.6)$$

We noted that a majority of the p -values were less than $\frac{\alpha}{66} = 0.00076$ and we therefore concluded that most pairs of organisms differed in complexity. In Table 11.6, for the (PTM, MS) connections of the Mt organismal proteins, nearly all the p -values were significant to describe major differences in network complexities. Here, these tables are read starting from an organism in the left column which is compared to those in its row. A significant value supports the notion that there are more edges in the network of the former organism than the latter. Only three tests were not significant (i.e., *Aspergillus nidulans* by *Oryctolagus cuniculus*, *Caenorhabditis elegans* by *Xenopus laevis* and *Homo sapiens* by *Mus musculus*). In these three tests, it was found that the latter of the pair of organisms presented a more complex network than the former according to the (PTM, MS) edges. In Table 11.7, we note the results for the non-Mt data. Only two tests (*Aspergillus nidulans* by *Saccharomyces cerevisiae* and *Danio rerio* by *Oryctolagus cuniculus*) were not significant to support a departure of complexities between networks. The full results for the general PTM and MS complexity comparisons using this same statistical test are offered in the

supplementary data.

Table 11.6: The p -values of from our Poisson approximation by the Chen-Stein Method over (PTM, MS) pairs in Mt networks. Significant p -values (i.e., values less than $\frac{\alpha}{66}$) are denoted by stars (*) to suggest that these pairs of organisms differ in complexity according to their networks.

Table 11.7: The p -values of from our Poisson approximation by the Chen-Stein Method over (PTM, MS) pairs in non-Mt networks. Significant p -values (i.e., values less than $\frac{\alpha}{66}$) are denoted by stars (*) to suggest that these pairs of organisms differ in complexity according to their networks.

11.4.7 Protein Isoforms In Organisms

Since much evolutionary time separates the complex organisms from the lower ones, conserved yet divergent isoform proteins are likely to exist. These isoforms may have originated from paralogous and alternatively spliced mRNA to create alternative gene products and functions from single coding sequences. It is known that alternative splicing is likely to encourage transcriptome diversity^[38]. For instance, in^[44], it was discussed that alternative splicing may have led to the larger divergence noted between the higher and lower organisms. In our own study, we also noted a wider divergence in PTM and MS usages between the higher and lower organisms.

During this evolutionary time, there is more opportunity for the generation of isoforms which may utilize PTMs in diverse ways. For instance, in Figure 11.19, we note that the mammals, notably *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, have more isoform proteins than the other organisms according to UniProt. To gather these results, we searched for all proteins corresponding to each organism and then we counted the number of isoforms. In addition to their abilities to respond to stresses, increasing PTM populations by organismal complexity may also be explained by their involvements with specific isoform-specific functionalities such as RAS protein isoforms^[1] and 14-3-3 protein isoforms^[252]. Also in Figure 11.19, we note that *Arabidopsis thaliana* had a large number of isoforms (compared to the others) which may help to explain its complicated networks shown in Figure 11.7.

11.4.8 Notable PTMs

The most frequently occurring PTM in our network models was phosphoserine, among both the Mt and the non-Mt proteins. This particular PTM represents the phosphorylation of serine base in a protein's amino acid sequence and is one of the most common modifications to proteins that can alter functionality. Among other sites such as threonine, tyrosine and histidine residues, serine is the most common

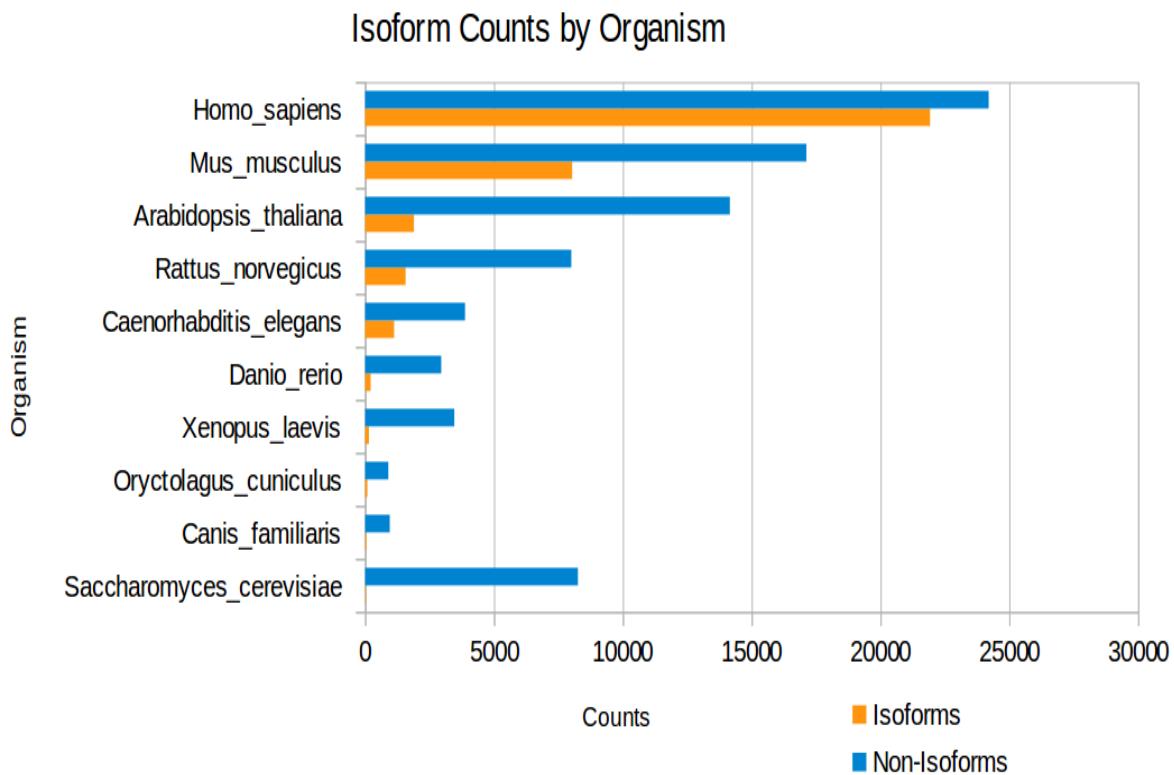


Figure 11.19: The number of isoforms of the organisms in our study. These counts were prepared by querying all organismal proteins in UniProt and then determining how many isoform proteins were present. The increasing number of isoforms may help to explain the increasing number of PTMs in higher organisms. Note that *Aspergillus nidulans* has been omitted due to the lack of isoform information.

type of phosphorylation. Serine phosphorylation like other phosphorylations can cause structural changes in proteins to activate or deactivate them.

Glycosylation is the result of a carbohydrate molecule that is added to a hydroxyl group, or another functional group of another molecule, (a glycosyl acceptor) in protein. The majority of the proteins synthesized in the rough endoplasmic reticulum undergo glycosylation. Although we found traces of glycosylation in the Mt proteomes of our data, we generally found more glycosylation in the non-Mt proteomes. In plants, however, Mt have the function of photo respiration requiring glycosylation. Although there may be other more pertinent reasons, the rich glycosylation observed in the Mt proteome of *Arabidopsis*

thaliana (of Figure 11.7) may be due to performing aerobic respiration. Interestingly, no glycosylation was observed in the Mt proteome of *Caenorhabditis elegans* of Figure 11.9.

Acetyllysine is another important PTM that adds an acetyl group to a lysine residue in proteins. The acetylation of lysine (K) residue is considered as a regulating mechanism for various epigenetic factors^[41;118]. We observed higher amount of N6-acetyllysine in Mt proteins and was conserved across *Homo sapiens*, *Mus musculus*, *Oryctolagus cuniculus* and *Rattus norvegicus*.

11.5 Conclusions

A bias is a preferential treatment of some element. Despite the present limitation of available PTM data across different species, our goal was to describe biases of PTM usage inherent to organisms using the most current available data. In this contribution, we used basic frequency information to produce evidence of the bias in the usage of PTMs, MSs and amino acids. We applied this frequency data (PTMs and MSs) to create heatmaps and networks which gave clear details about the differences between organismal proteomes. From the heatmaps and the networks, we noted that PTMs and MSs have very different compositions between proteomes. We observed that the non-Mt networks were generally more dense and more populated by PTMs and MSs than the Mt network of the same organism.

We noted that an organism's PTM and MS bias is not likely explained by its amino acid composition (from Figure 11.6) since the compositions were too similar between all organisms. Instead, this bias must come from another source which we suggested was related to environmental stress response. Since PTMs enable stress response in protein, our study supports the notion that the environmental stresses of its habitat may likely play a role in an organism's PTM and MS bias. This finding is strengthened

by the discussion of the high number of PTMs and MSs that were observed in the networks of *Arabidopsis thaliana* and *Aspergillus nidulans* (Figures 11.7 and 11.8, respectively). We note that the survival of these, and other plant organisms, may be based on their ability to tolerate their environmental stresses. Furthermore, we noted that the organismal complexity increased in tandem with the number of observed PTMs in both the Mt and non-Mt proteomes. This, we speculated, may be due in part to the ability of the more complicated organisms to inhabit regions that host a wider variety of environmental stresses than those habitats of the less complicated organisms, with the exception of plants.

Our study showed that PTMs such as acetylation and phosphorylation were common to the Mt proteomes and glycosylation and phosphorylation were prominent across the non-Mt proteomes. Although many of the PTMs were common throughout our organismal data shown in Table 11.1, we noted that the individual organisms tended to interact with MSs in different ways. For instance, in all networks (Mt and non-Mt) the PTMs themselves did not interact consistently with the same MSs, across the organisms. In non-Mt proteomes, we observed that PTMs were more promiscuous in their interactions with MSs and often a particular PTM would interact with several different MSs simultaneously. This was generally not the case in the Mt networks. Here these PTMs tended to interact with the same MSs across the organisms. This finding may be partially explained by the conserved nature of Mt but says nothing about the environmental stresses which the organelles may have to tolerate from the organism's habitat. Importantly, the differences in the frequencies of PTM and MS usage across the data may suggest a unique organismal mechanism which supports our contention that environmental stress is likely the motivator of bias.

In future work, we intend to investigate the influence of stress on PTMs. In particular, in protein stress response systems, we intend to study the relationships

between stresses and the PTMs of proteins which are related to specific functional groups across this and other organismal data.

11.6 Article Details

This contribution was published in *Briefings in Bioinformatics*, 2016.

- Oliver Bonham-Carter, Ishwor Thapa, Steven From and Dhundy Bastola, “A study of bias and increasing organismal complexity from their post-translational modifications and reaction site interplays”, *Briefings in Bioinformatics*, 2016, bbv111.

It is a capital mistake to theorize
before one has data. Sherlock
Holmes, A Study in Scarlet

(*Arthur Conan Doyle*).

Chapter 12

A Text Mining Application For Linking Functionally Stressed-Proteins To Their Post-Translational Modifications

12.1 Abstract

In the proteome, stresses may work against optimal protein function and PTMs play roles in protein stress responses. Many peer-reviewed articles are available to bioinformatics research in the literature, however, the details of stress, protein and their PTM interactions have been scattered throughout the literature and these concepts are mentioned amongst the other details of respective studies. In each publication, for instance, there are many small pieces of knowledge which could be

combined to build a better understanding. Since it is impossible to harvest all of its available knowledge using manual means, text mining methods are an attractive approach to assemble ideas from articles where these concepts may not have been a main focus.

In this contribution of the thesis we present a text mining method to harvest and assemble a knowledge base relating to the relationships of stresses, proteins and PTMs from the literature. Although we also studied the stresses, proteins and PTMs which were associated with apoptosis, diabetes and Parkinson's diseases in the literature, to introduce our method, we address these concepts as they are related to Alzheimer's disease. We use the results from our text mining tool to process article abstracts to build networks which suggest how functional proteins may be linked to environmental stresses and their PTMs. We discuss how networks of biologically relevant keywords may eventually be used to describe directions in research which could be further explored to forecast new trends of studies. We also show how our method may help to predict stress, protein and PTM associations which may be included in these future studies.

12.2 Introduction

Post-translational modifications (PTMs) are steps in biosynthesis that alter a protein's physical confirmation. Since the protein structure and function are intimately connected, a change in its conformation will lead to a new function. In the human proteome, where more than 400 human PTMs have been observed^[148], it is very likely that all proteins are regulated by PTMs at some point in their lives. Apart from preparing proteins for particular duties in specific locals, one of the most fascinating aspects of PTMs is that they generally enable stress responses throughout seemingly all proteins. For example, PTMs enable tolerance to heat

shock in eukaryotic cells^[157], resistance to types of aging^[81] and are active in overcoming the stress caused by reactive oxygen species^[306]. When a protein is exposed to stress, common PTMs such as phosphorylation, for example, may rapidly phosphorylate specific amino acids of the protein to initiate physical changes. Once the stress has elapsed, these sites are often dephosphorylated to return the protein back to its original conformation.

Apoptosis, as well as ailments such as Alzheimer's, Parkinson's diseases, and diabetes, are characterized by the alteration of proteins where PTMs also play roles^[104;297]. In the diseases, it has been suggested that proteins have succumb to environmental stresses and failed due to faults in PTM-driven stress responses. For example, mitochondrial (Mt) disorders stemming from protein misfolding (implicating PTM failures) are also linked to the onset of Alzheimer's disease^[285]. The studies of PTMs have gained much traction and the beginnings of a firm knowledge base has already begun. In spite of this effort, there is much to learn about the relationships between stresses, proteins, and PTMs, and about how they are connected to the ailments (such as those described above). Although there are prominent studies coming out which focus on these actors upon the stages of other ailments, it is hard to find many such studies which bring all these elements together. Fortunately in the literature, there are countless articles where stresses, proteins and PTMs have been mentioned in studies where these concepts are relevant but not primary focuses. As it is impractical to manually read all papers where some mention of these elements may be found, text mining is used to harvest this knowledge.

Previous text mining tools have concentrated on extracting information for convenient use (text summarization, document retrieval), assessing document similarity (document clustering, key-phrase identification), and extracting structured information (entity extraction, information extraction)^[310].

Unfortunately, many of these tools are unable to capture the relevance from bioinformatics studies where the main goals of text mining are to discover inherent relationships between the located concepts. Furthermore, these relationships may remain elusive due to the high volumes of domain-specific literature to process, the complexities of managing many hundreds of keywords to be retrieved in the corpus and the extreme difficulty of finding the connections between these keywords to discover relationships. There are, however, many good text mining software tools available in bioinformatics, such as gene searching^[12], retrieval of protein relationships^[246], for example. However, the lack of interoperability between many of these unique tools (i.e., incompatible output formats) frustrates the ability to use multiple tools during single studies.

In this contribution, we describe a novel text mining method, called *Lister* which was further applied from a previous study^[47]. We show how this method meets our stringent demands of being customizable, able to work with large sets of data and is suitable for finding connections between tiny pieces of relevant information (i.e., stresses, proteins, and PTMs, as in our field of study). Our method may process a corpus of seemingly any size. We show how open source database software is applied to organize and find connections and patterns between the harvested concepts and details. Applying our work to proteins which are associated to Alzheimer's disease, we use network models to visualize the discovered connections and patterns. In addition to determining these relationships in the data, we explain how our method may be used to predict new studies which could contribute profoundly to a scientific field using linked knowledge from previous works in the literature. Finally, we discuss how this contributed knowledge may originate from the related and unrelated work, in terms of a literature review of a novel study.

12.3 Methods

Text Mining: Our text mining method (called, *Lister*) is built from open source software (Python (<https://www.python.org/>) and Sqlite3 (<https://www.sqlite.org/>)). Lister provides convenient customization and its output has been especially formatted to create input files for populating the Sqlite3 database. In time, we plan to release the Lister’s source code to the community.

The corpus data for our work was downloaded from NCBI (URL: <ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/>) and is their most recent compilation (dated: 27 June 2015). Uncompressed, there are about 100 GB of articles to process (about 1,137,842 articles). Our method is different from traditional approaches to text mining since we determine associations between keywords by employing a supervised, *bag-of-words* approach to isolate all articles containing particular user-selected keywords. Lister was designed to scan only the abstracts of these articles as they are typically relevant summaries that have been carefully written to reflect their article’s contents and include all non-ambiguous and relevant keywords. All records are in an `nxml` format for convenience.

Lister scans all abstracts to find the occurrence of selected keywords across several different and, perhaps, unrelated articles. A relationship has been predicted to exist between these keywords when they are all found in the same article(s). This implies that these keywords were relevant to some study where they all played a role. Several keywords which are found together in a bioinformatics article is very likely to signify that these were actors for roles of an experiment. For example, in bioinformatics, learning that a particular protein and a stress have been found in the same article is very likely to suggest that the protein has been studied in some context of the stress. Furthermore, if it is found that a type of PTM is also mentioned in the text, then we have reason to suggest that the PTM may be a part of the stress response for the protein. In general, we have evidence to suggest that the stress, protein and

PTM are likely to share some form of relationship. Although further exploration is necessary to determine the exact details of the discovered relationship, this is not a limitation because all relationships, no matter their strengths, may be important parts of a rigorous review. Furthermore, many discoveries have been suggested by simple *guilt by association* scenarios.

The associations between keywords are made by connecting their sources. Since each article from NCBI has a unique PMID number (identification reference for PubMed citations), all encountered keywords are recorded with the PMID number of the article where they were found. Our method connects these keywords to each other by finding the intersections of the lists of PMID numbers from each keyword.

Database Support: Databases allow for finding connections by queries in their data. For this task, we used Sqlite3, which was chosen for its simplicity, power, open source nature and ability to keep an entire database in a single file. Sqlite3 is also provides for a convenient way to setup and populate a base from the outputs of Lister by use of a basic script. For an occurrence of a keyword in the text, a database table held the PMID number, article references, the occurrence number, and its associated blurb of text. The database had six main tables: **Protein Function** (the functional group from which our proteins were extracted), **Mt Proteins**, **non-Mt Proteins**, **PTM-General** (general PTM names such as *acetylation* or *phosphorylation*) and **PTM-Specific** (specific cases of general PTMs). The table for **Stress** concerns the types of stresses that we had observed from a manual literature review. Since any type of relationship may be found with the use of an appropriate query, there are countless ways to exploit this information for a variety of studies.

Keywords: Our interest was to find the relationships between the non-organism-specific proteins associated with the ailment groups (i.e., Alzheimer's, in addition to apoptosis, diabetes and Parkinson's), their stresses and their corresponding PTMs according to the literature. Since proteins associated with

apoptosis and the ailments tended to include Mt proteins, our study included these complexes, in addition to non-Mt proteins. Lister was designed to use user-specified keywords and so each table of our database contained the results from each set of keywords. The PTMs that we were interested in studying were those that interact with proteins at a modification site which is made-up of only one residue. These keywords are: *{acetylation, glycosylation, methylation, phosphorylation, and others}*. The types of stresses that we were interested in were: *{carbonylation, cold/heat shock, oxidation, microgravity and others}*. The protein keywords came from our previous work^[32;34]. The number of unique keywords we employed were the following: {stresses: 45, general PTMs: 33, specific PTMs: 29, Mt proteins: 589, and non-Mt proteins: 10,041}. We amassed a total of 1976 (Mt) and 36,854 (non-Mt) references of our proteins.

Networks: To find the associations between our keywords, we built networks from the output of our text mining system using^[261]. There are three types of nodes featured in each of our networks: green pentagons (*PTMs*), red circles (*protein types*), and blue squares (*stresses*). An edge connecting nodes signifies that at least one study exists where the keywords have been mentioned in an abstract. The networks that we are primarily interested in studying are those featuring cliques – where all keywords are included in a common abstract or have the same article PMID. Each network we created concerns a specific ailment group of proteins. For instance, the proteins of Figure 12.1 have been ordered according to their relationship to abstracts concerning *Alzheimer's* and we call this network an *ailment protein network* (APN) to describe the linked keywords by ailment according to the literature. APN networks provide strong evidence for a relationship between nodes because they imply that studies exist to link its elements. In addition, discussed in Section 12.4, we may study these networks to infer trends in bioinformatics research and likely predict the publication of future studies where these keywords are included all together.

12.4 Results And Discussion

In our previous work^[32;34;35], we used public databases to determine PTM interactions and their frequencies of occurrence in proteomes. Although the data of this study is from a corpus of disconnected literature, and not a database of observed PTM interactions with proteins, we note that there are several fundamental similarities between the observations of this current work and those of our previous studies. For instance, we note from the results of our text mining method that the commonly encountered PTMs, *acetylation*, *glycosylation* and *phosphorylation* and often, *methylation*, are often present in each the ANPs (both in the Mt and non-Mt networks). This observation agrees with the fact that these PTMs are also very commonly encountered across organisms in nature. Since we have made the same discovery of PTM compositions in protein from two separate sources, we have a reason to suggest that other findings may be discovered here which are also biologically relevant using our text mining system.

We note that an edge between nodes signifies that a study exists where the nodes representing both keywords are included in the abstract of the study. In the Mt APN of Alzheimer's disease (Figure 12.1, left), there were fewer proteins found to be linked to PTMs than the number of proteins in the non-Mt network. This is consistent with our work in^[34]. Two reasons may explain this largely unequal number of proteins. The first reason is that Mt produce a limited number of their own proteins internally, while the rest are synthesized by nuclear mechanisms of the cell. Although this phenomenon may be well known to any student of biology, it is also an observation in the output of this text mining task. The second reason for fewer Mt proteins in Alzheimer's disease may be explained by the science itself. Perhaps fewer proteins have been observed thus far to explain this lack of documentation. Proteomics (or the study of proteins with a focus on structures and functions) is a young and developing field which is only just able to answer its research questions thanks to advancing

technologies. The lack of proteins in the Mt network may also indicate that there is much more work to be done in this discipline in order to discover proteins that are connected to Alzheimer's disease.

We noted the similarity between the APNs (both Mt and non-Mt) of Alzheimer's and Parkinson's diseases. For instance, in the Mt network of Figure 12.1, left, we note the appearance of the protein *OGT* (the forth node down in protein types) which has been recently studied by^[169] where the abnormal glycosylation by *OGT* of an essential mammalian enzyme (O-linked β -N-acetylglucosamine transferase), has been linked to insulin resistance, diabetes, cancer and neurodegenerative diseases including Alzheimer's disease (in human and other models). In the Mt APN of Parkinson's disease (not shown due to space limitations), *OGT* also appeared. In both networks (Alzheimer's and Parkinson's) this protein is connected to the same PTM and stress: glycosylation and oxidative stress, respectively. There are similar findings in diabetes and Parkinson's disease networks. This describes an overlap between both ailments. Furthermore, this common protein may imply that there could be a wealth of information from one type of study which could be used to complement another

By following the work of one type of ailment, we may likely be able to infer observations for other ailments where these keywords are included. Sadly, accustomed as we are to reading only the journals which discuss our keywords in particular settings, we often maintain our ignorance of the real purpose of our work which is to understand the mechanism. Thanks to the conserved nature of biology, many of the discoveries concerning one kind of function may also be applied to other functions after a simple text mining operation which is able to connect the ideas from the corpus.

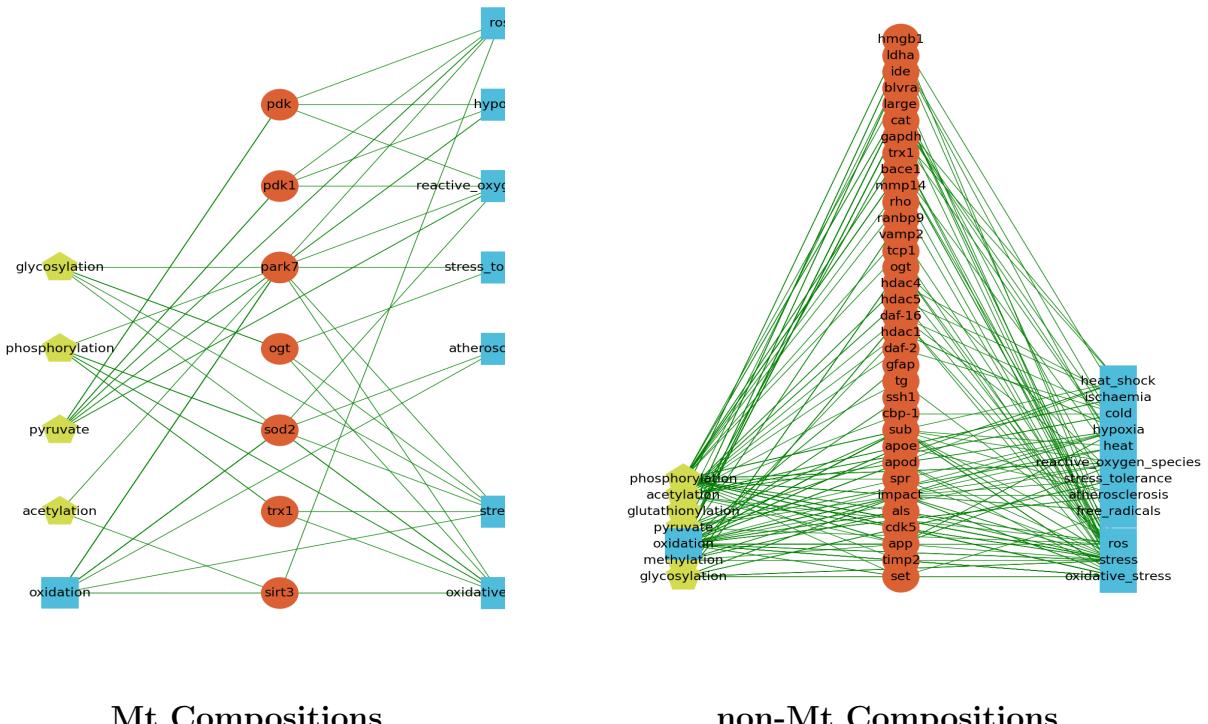


Figure 12.1: *ALN*: The stressed Mt proteins associated by **Alzheimer’s** disease. The three types of nodes featured the networks are: green pentagons (*PTMs*), red circles (*protein types*), and blue squares (*stresses*).

12.5 Conclusion

In this study, we showed how relationships between keywords can be made from isolated articles of NCBI’s extensive corpus using simple means. We introduced our tool (Lister) for this purpose and described how it is able to process large amounts of information, while being fully customizable and able to simplify relationship-finding between keywords. Our tool mines its information from abstracts that we noted is plausible since their terms are carefully chosen to provide the non-ambiguous details of the articles they represent. Using the power of database queries, our method isolated keywords in the studies and found relationships which were biologically relevant and met the expectations of our previous work in stresses, proteins and PTMs, as actors on the functional stages of Alzheimer’s, Parkinson’s, diabetes and apoptosis. In our

APNs, we explained that an edge between nodes signifies that a study exists where the connected nodes are both mentioned and we described how missing edges may likely predict the beginning of new studies to complete similar cliques where the connections between two of the three nodes (keywords) are already known for other ailments. We suggest that understanding these links will enrich the study of these ailments.

In the future, we intend to extend our Lister tool to add statistical power into its framework. This enhanced analysis would allow the use of our tool to discern between strong and weak types of relationships between keywords. We also expect to publish all results from this work in a more expansive article which will also include other applications of Lister. Finally, we plan to make our tool open source and to release it to the bioinformatics community in hopes that it can be used to build more connections among the literature and help predict new and parallel studies.

12.6 Article Details

This contribution was published in the IEEE International Conference on Bioinformatics and Biomedicine (Semantics and Ontology Track), 2015

- Oliver Bonham-Carter and Dhundy Bastola, “A text mining application for linking functionally stressed-proteins to their post-translational modifications”, 2015 IEEE International Conference on Bioinformatics and Biomedicine (Semantics and Ontology Track), 2015.

What makes the desert beautiful is
that somewhere it hides a well.

Antoine de Saint-Exupery

Chapter 13

PTM Tracker: A System For Determining Trends Of PTM Modification Sites Relative To Protein Domains

13.1 Abstract

Post-translational modifications (PTMs) increase protein functional diversity by modifying an amino acid at specific locations called modification sites (MSs) in protein. It is believed that domains are being influenced by PTMs at interacting MSs to determine the unique functional changes in protein and, in this scenario, it is likely that the exact position of the MS, relative to the domain, plays a major part in structural changes by PTM influence. In this study, we present a system

called “PTM Tracker”, built from two main parts to study the general distances of MS amino acids which are relative to protein functional domains. In the first part, we apply our system to illustrate that unique organisms appear to have distinguishing locations where PTMs may be found in the proteome. These crowded locations of MS sites (called, “neighbourhoods”) are relative to protein domains. We describe how these MS neighbourhoods may be a conserved extension of the already-conserved domain. Since specific protein domain types may be found in diverse proteins, the second part of our system studies MS neighbourhood clusters, relative to user-selected domains. From the study of many different proteins containing the same domain type, we conclude trends to suggest that MS neighbourhoods have specific locations in protein, relative to the domains (where-ever the domain occurs naturally), with which they are likely to interact. We conclude that the study of these distances may help to understand interaction mechanisms and describe types of protein folding requirements.

13.2 Introduction

Post-translational modifications (PTMs) are steps in biosynthesis to regulate protein and enable them to perform specialized tasks within the cell. PTMs broaden the functional diversity of the proteome, by interacting with protein domains to initiate specific functions such as regulation^[268] in both, mitochondrial (Mt) and non-Mt protein. PTMs are also involved in stress response as they allow for the adaptation of protein to function in short-term, stressed situations and once the stress has elapsed, PTMs generally restore the protein to its original conformation state.

Protein, such as the *Tau* variety, stabilize microtubules and are highly regulated by PTMs. This type of protein is also closely linked to the onset of Alzheimer’s disease^[46] and its improper regulation by PTMs may strongly contribute to the

onset of disease^[197]. Alzheimer's is only one of many ailments (i.e. Parkinson's, Huntington's and other age-related disorders) and it is necessary to study and understand the roles played by PTMs during the regulation of protein function. Since cellular stress appears to demand much PTM activity, the onset of these diseases may be partially explained by studying the role of PTMs as initiated by stress.

The functions of proteins are controlled by a diverse assortment of PTMs from specific binding sites that control domains. Studies of the influence of PTMs on protein domains may be facilitated by many database-driven tools such as UniProtKB^[39], SysPTM 2.0^[181], PhosphoSiteSlus^[125], dbPTM^[129] and others. However, considering that domains are distinct, highly conserved, functional and/or structural units in a protein, which could be found in a variety of diverse biological contexts, an analysis may be complicated when individual proteins are studied in isolation.

We present a system called ("PTM Tracker") which allows for the study of PTM relations with general and specific protein domains, regardless of the individual proteins where they are encountered. Since PTMs occur at specific amino acid modification sites (MS) in protein, our system focuses on this phenomenon and allows us to study the actual distances between the MSs and the domains that they likely influence. For instance, an MS may have to be located at a specific location in the sequence to ensure that it is able to influence in the domain for function. Here we suggest that this distance between MS and domain is a conserved part of the domain. Furthermore, we suggest that common domains, even if originating from diverse organismal protein, may also conserve this spacing of MSs for function. Our method exploits this likely conserved spacing to describe unique clusters of MSs (called *neighbourhoods*) which are common when studying like-domains, regardless of their origins. We note that these MSs are curated and have been observed in the

literature to interact with at least one of these domains^[148].

Our system has two main applications: (1) to study the distances between MSs and all domains located in an organism’s proteome, and (2) to study the distances between MSs and user-selected domains which are encountered in a wide variety proteins (Mt and non-Mt). In Section 13.3.1 we detail this organism-specific study to show that organisms exhibit general trends of MS spacings which are likely an extended part of their PTM bias as described in^[34;35]. We study the MSs of sequences relative to the domains in locations situated *before* (upstream), *inside* and *after* (downstream) of encountered domains in protein. The general trends become apparent when we consider increasingly larger sets of protein which serve to eclipse noise. In Section 13.3.2, we discuss the MS neighbourhoods which were uncovered by studying specific domains, across all proteins of our data. Here we discuss the global observations of particular domains that appear to have distinct preferences for MSs at predictable locations.

PTM Tracker allows for an in-depth study of these MS distances, with a focus on domains. The aim of this work is to provide an analysis platform for the study of mechanisms where PTMs influence domain function. Furthermore, we imply that the understanding of these PTM-domain relations may likely help to explain some of the reasons for a potential protein failure or disorder.

13.3 Methods

Following our previous work of PTM involvement with proteomes^[34;35], we limited the current study to the same protein data of the 11 organisms listed in Table 13.1. The protein data was downloaded from the UniProt Knowledge Base^[65] protein database in March 2016. The details of each organism’s protein were contained in a record from which we extracted the sequence, and the observed types of PTMs with the

Table 13.1: Diverse organisms of the study. The number of MS encountered in Mt and non-Mt are shown in the organism's row.

Organism	Common Name	Mt	nonMt
<i>A. thaliana</i>	Mustard Plant	138	11086
<i>C. elegans</i>	Nematode worm	9	2403
<i>C. familiaris</i>	Domestic Dog	130	2599
<i>D. rerio</i>	Zebrafish	33	1916
<i>H. sapiens</i>	Human	3285	50871
<i>M. musculus</i>	House Mouse	3413	42470
<i>O. cuniculus</i>	European rabbit	221	2788
<i>R. norvegicus</i>	Norway Rat	2684	22021
<i>S. cerevisiae</i>	Bakers Yeast	412	10520
<i>X. laevis</i>	African clawed frog	13	2086

protein (i.e., *acetylation*) at specific amino acid MSs. We also collected a listing of observed protein domains as well as their exact locations in the sequence. We note that any mention of an MS suggested that it likely played some role to influence at least one of the domains in the protein.

Mitochondria (Mt) is responsible for the cellular energy production and is highly conserved across organisms. In spite of its conserved nature, organismal Mt, in addition to non-Mt protein, show striking differences in PTM usage (bias) across organisms^[36]. Since domains may simultaneously exist in Mt or non-Mt proteins, we created two classes to organize each type of protein for a comparative study. We note that UniProt conveniently labels Mt protein when the origins are known, however the lack of this label does not necessarily imply that the protein is not Mt.

13.3.0.1 PTMs

In each protein record, are the details of all its observed PTMs. For instance, we counted 47 (Mt) and 249 (non-Mt) unique individual PTM MS types from our 11 organisms. Because some of these PTMs may be categorized into more general classes of common PTMs (i.e., *N6-acetyllysine* is actually a general type of *acetylation*), we

have considered most of these specific instances of PTMs by their more general types (i.e., in this case, all types of acetylation are considered as, simply, *acetylation*).

Here, we focus on *acetylation*, although any PTM could be studied with this method. We chose this particular PTM since there is currently lots of available data to our study (via UniProt) thanks to recent bio-medical work in Bioinformatics. Interestingly, since *acetylation* plays a role a great variety of unrelated organisms, by evolutionary time, this PTM may be one of the oldest and also more critical for supporting life^[298]. PTMs such as *phosphorylation*, *glycosylation* and *methylation* are also becoming popular research interests and data (i.e., concerning mechanisms and domain influence) is currently increasing. We hope to integrate this knowledge into our future studies of MS neighbourhoods.

13.3.0.2 Modification Sites

In our study, we consider only those PTMs which modify a protein at a single point (i.e., an MS). In each record from UniProt, we collected a list of all MSs which have been observed to be modified. *Acetylation* typically interacts with lysine but may also modify protein at other types of amino acid MSs, as well. Although our method is capable of studying any particular kind of amino acid MS, our study focused on lysine, in addition to all other amino acids (i.e., MSs) which are involved with *acetylation*.

13.3.0.3 Domains

Our research of protein domains fits into two categories: (1) the organism-centric study of PTM distances (i.e., in this work, *acetylation*) to any, non-particular, encountered domain in the proteome (discussed in Section 13.3.1) and, (2) the domain-centric study of PTM distances (i.e., *acetylation*) relative to a selected domain, across any proteome where it is naturally encountered (discussed in Section 13.3.2).

Across the proteins of our 11 organisms, we counted 264 and 2402, unique Mt and non-Mt domains, respectively. We chose a set of sample domains to study which were also observed in at least four different organisms. We note here, that the domains from Mt and non-Mt may not necessarily originate from the same organisms. Any domain that we studied was isolated from the proteins of at least four different organisms. This was done to gain a general pattern of how the domain was employed in all its proteomes. In addition to the conserved domains, we maintain by our results that particular MS usages are also conserved. For this reason, our study paid close attention to all MSs of the proteins where the domains were encountered since they may play roles in domain function.

13.3.0.4 Plots

In the organism-centric plots, all domains from a particular organism and involved with *acetylation*, were retained. In Section 13.3.1.1, we describe the neighbourhoods to be regions where the MSs of a particular PTM are generally found. These regions are relative to domains and, as noted in Figure 13.1, may be found *before*, *inside* and *after* the protein domains. For neighbourhoods found *before* or *after* domains, the distances of MSs were measured leading-away from the domain's beginning or ending positions, respectively. For MSs encountered *inside* domains, its distance to the domain-end position was collected. In our plots, created by the “NetworkX” Python library^[262], the *x*-axis represents its distance from a domain position in a protein sequence. The *y*-axis represents the total number (i.e., the magnitude) of all occurrences of MSs at particular locations. All measurements of MSs relative to domains are represented in separate plots for which proportional values (explained later) have been calculated to enable their cross-comparisons with other protein sets.

13.3.0.5 Heatmaps

It was important to determine the natural amino acid compositions of regions where MS neighbourhoods were encountered. We collected each segment of protein (i.e., the *before*, *inside* and *after* regions of the sequence as displayed in Figure 13.1) and the frequency for each of the 20 amino acids (not including the stop codons) was calculated. These values were then applied to Pheatmap R-statistic library^[155] to create heatmaps to visualize amino acid compositions across the organisms.

13.3.1 Organism-Centric Study Of PTM Distances

We assumed that PTMs occur at specific amino acid MSs to influence domains and our work concerns only the patterns of distance between MSs and these domains. This part of the method allows for the study of all domains of an organism. Building from the work of^[35], we note that this approach also describes any biases which may exist between organisms in connection to PTM usage. Other types of PTM usage biases have been uncovered and are discussed in^[34]. We maintain that the trends found in the spacing of the MS neighbourhoods are another form of usage bias. To determine this bias, we applied our system to three types of locations in protein sequences which are described below.

13.3.1.1 Three Criteria Of Proximity

We established three neighbourhoods where MSs may be found (i.e., *before*, *inside* or *after*) any domain in the same protein sequence. In Figure 13.1, the starting and ending locations of protein domains in sequences are defined as, *domStart* and *domEnd*. The MS positions are defined as, *MSLoc*, and the length of the entire protein sequence is defined as, *seqLen*. Also shown in Figure 13.1 are the chunks of sequence code from which we conducted an amino acid composition analysis (discussed in Section 13.4.0.3).

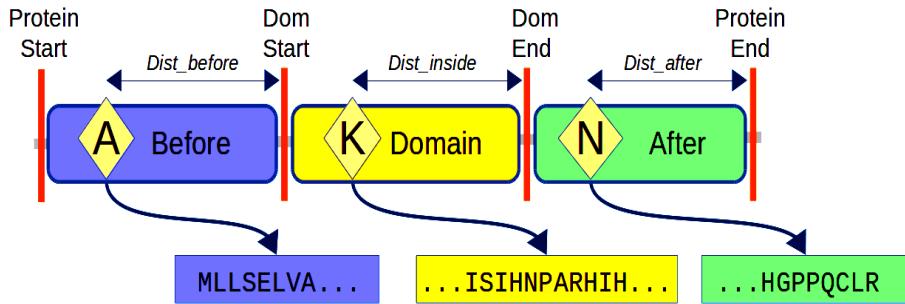


Figure 13.1: The description of where the distances of PTMs are found: *before*, *inside* and *after* a protein domain. The amino acid residues for each region were collected and analyzed for composition. The MSs are shown inside the diamonds and the arrows-above describe the collected distances.

13.3.1.2 MSs Found Before A Domain

If an MS was found to occur before, or upstream of a domain, we measured its distance from the start of the domain. Therefore, this measurement represents the number of amino acids lying between the MS and the domain's starting position (down-stream). We may expect to find noise in cases of MS-domain measurements where the two (i.e., domains and PTMs) do not naturally relate, however, we may draw a consensus from the resounding patterns of MSs if many PTMs influence function. Equation 13.1 creates a proportional measurement value allowing for a comparison between samples. Although this equation also calculates distances from amino acid residues on both sides of the *domStart*, we only apply this equation for MSs occurring before a domain. These distance values are proportional and allow for comparisons across proteins since they have been calculated by dividing their value by the length of the region.

13.3.1.3 MSs Found Inside A Domain

If the PTM was located inside a domain, then we measured its distance between the beginning of the domain to the domain's end (hence, inside the domain). Therefore, this measurement represents the number of amino acids lying between the MS and the

domain-end, *inside* the domain. The equation is shown in Equation 13.2. We chose the end of the domain from which to measure this distance so as to move downstream from the position of the MS. We note that the domain sizes are proportional to each other since we used the lengths of the domains to calculate this proportional value.

13.3.1.4 MSs Found After A Domain

Defined in Equation 13.3 is the distance of particular MSs to the end, or downstream, of encountered domains in the sequence. Therefore, this measurement represents the number of amino acids lying between the domain-end and the MS located downstream and *after* the domain. This calculation implies the number of amino acids downstream and after the domain. These distance values are also proportional and allow for comparisons across proteins since they have been calculated using the region length.

$$Dist_{\text{before}} = \frac{MSLoc}{domStart} \quad (13.1)$$

$$Dist_{\text{inside}} = \frac{(domEnd - MSLoc)}{(domEnd - domStart)} \quad (13.2)$$

$$Dist_{\text{after}} = \frac{MSLoc}{(seqLen - domEnd)} \quad (13.3)$$

13.3.2 Domain-Centric Study Of PTM Distances

Next, we apply our method to investigate the MS neighbourhoods, involved with specific (“user selected”) domains. We note that a specific domain is likely to be very conserved and available in other protein samples of un-related organisms. This part of our system aims to give a global overview of the distance patterns from the neighbourhoods and their domains.

13.3.2.1 MSs Found Relative To Domains

To achieve our results concerning amino acid residue positions, we traversed all proteins (from the organisms of our data) containing a domain of interest. We collected the beginning and ending positions of the domain in the protein sequences, along with the locations of all MSs of a PTM which were contained in the UniProt protein records. We applied Equation 13.4 below, to determine the distances between the elements (i.e., *MSLoc*, *domStart* or *domEnd*) of a protein sequence. This approach provides a global view of the proteins containing particular domains. By tracking the MS - domain distances of many different proteins, we uncover the MS neighbourhoods. We may encounter noise from some of the MSs which do not necessarily interact with a chosen domain, however this noise is minor when compared to the results of the actual domain-interacting MSs. To demonstrate this approach, we choose the domain, *atp-grasp2*, which may be found in both Mt and non-Mt organismal proteins.

$$Dist_{element} = \frac{Element_{position}}{seqLen} \quad (13.4)$$

13.4 Results And Discussion

It is possible that a protein record from UniProt may list multiple domains and modification sites involved with acetylation, as shown in Figure 13.2. We record the distance between the MS to each of its domains in separate steps. In addition, there may be cases where a single domain is influenced by two different MSs, as shown in Figure 13.3. Here, each MS is processed separately for the same domain. We are aware that noise from “false positives” may be introduced where recorded distances have no biological meaning. However, there are enough cases of conserved distances (of biological meaning) that general patterns are likely to become visible.

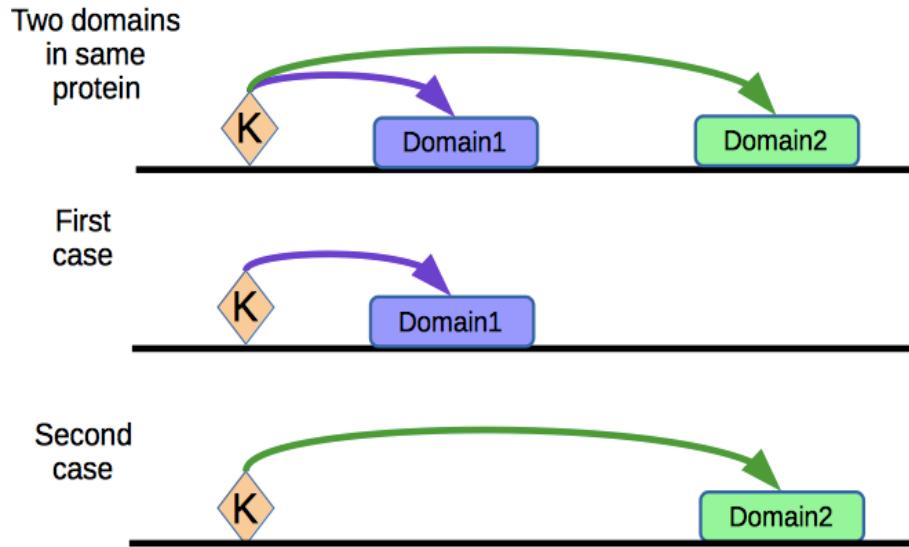


Figure 13.2: We note that a protein may exist having multiple domains as influenced by the same MS. In this case, each domain is processed separately by PTM-Tracker with the same MS.

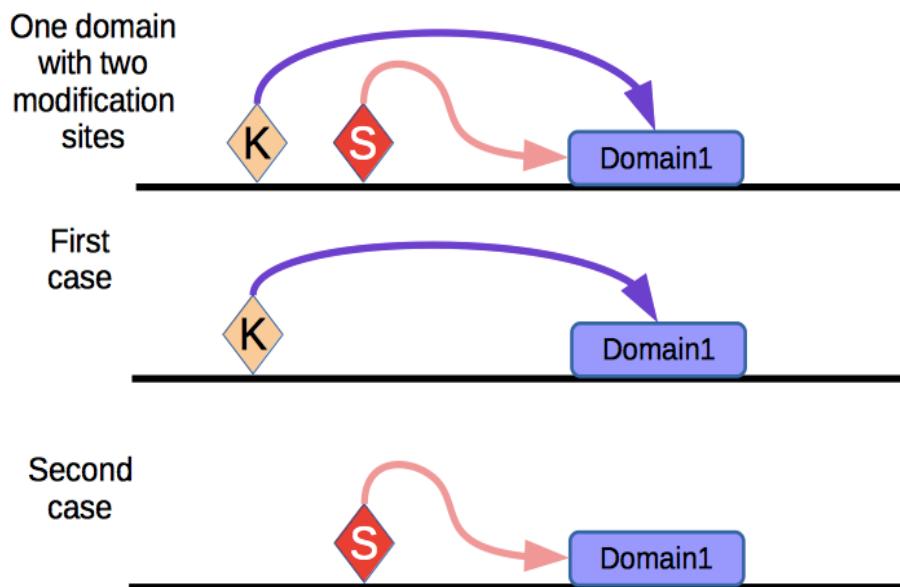


Figure 13.3: We note that a protein may exist having multiple MSs that influence the same domain. In this case, the domain is processed for each MS.

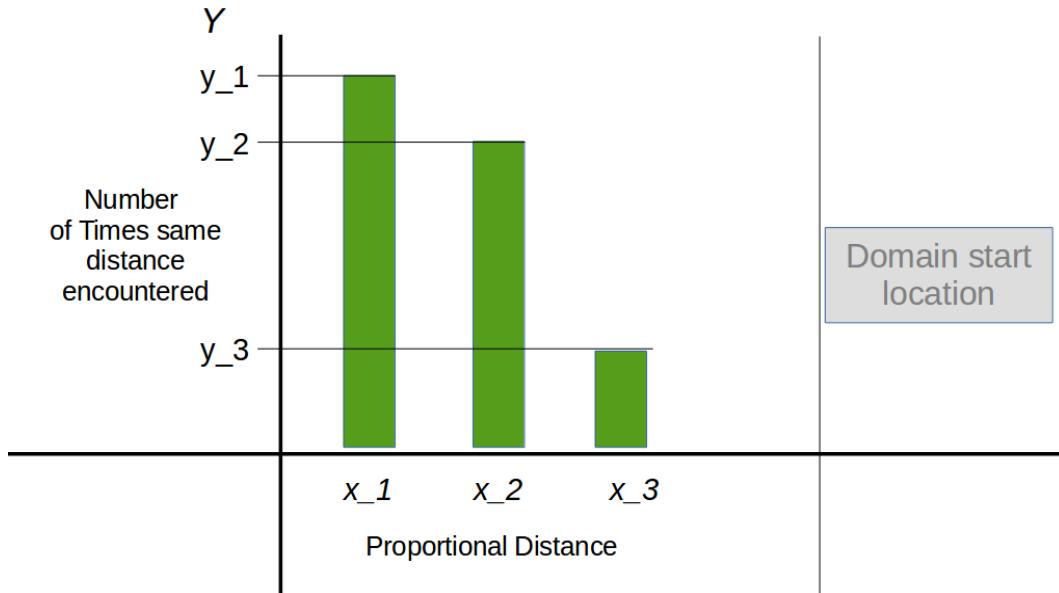


Figure 13.4: How to read plots containing organismal MS information occurring *before* domains. The domain start is imagined to be at the far right and the green bars describe the MS neighbourhoods which are located upstream of the domain. The magnitude is understood to be the number of other MSs for *acetylation* (or any single PTM of interest) found at the same locations. The plots for showing MS neighbourhoods *inside* and *after* the domains of a particular organism are similar. The plot in this example describes that MSs are found in three general locations (i.e., neighbourhoods) *before* all domains of a particular organism.

13.4.0.1 Graphical Interpretation - Organism-Centric Study Of PTM Distances

In Figure 13.4, we describe how to read the plots containing information of the MS neighbourhoods encountered *before* all domains of a particular organism. The proportional value of the location of an MS is calculated to allow for a cross comparison between proteins. The domain start locations are imagined to occur on the far-right. In the figure, we note that there are three MS neighbourhoods where populations of MSs are situated just before the domain starting positions. The green bar on the far left implies that the largest population of MSs are found at the start of the protein and are relatively far away from the beginning of domains. Reading the plots describing the MSs occurring *inside* and *after* domains is

performed in a similar way.

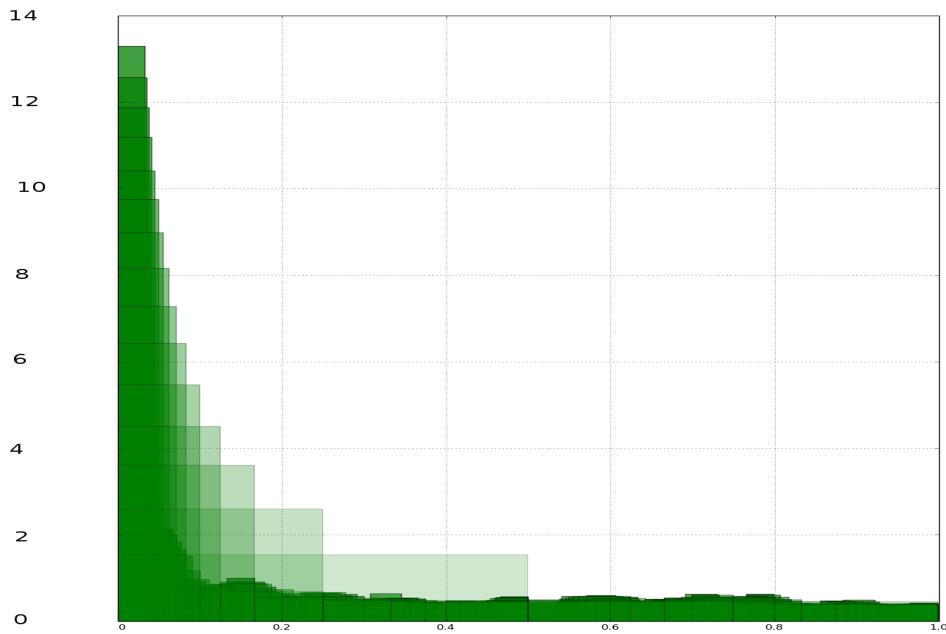


Figure 13.5: The non-Mt protein plot of all locations of *acetylation* MSs found *before* the domains in *H. sapiens* (human). The *x*-axis represents the location of the MS neighbourhoods (green). The *y*-axis describes the number of times that this same location was observed for the element across the samples. This is a typical plot for the mammal protein data.

In our organism-centric study of PTM distances, the domains are those which are encountered in the proteome of a particular organism. All distances between *acetylation* MSs (i.e., any PTM may be selected for study) and the encountered domains were collected to determine general trends in distances. We note that related organisms tended to have extremely similar trends. For instance, the plots of related organisms (i.e., across the mammal proteomes) described similar numbers of neighbourhoods and also similar types of locations of MS neighbourhoods relative to domains. This finding suggested a conservation of MS and domain placements and may be found in other related organisms. Additionally, this feature could be used to

differentiate un-related organisms.

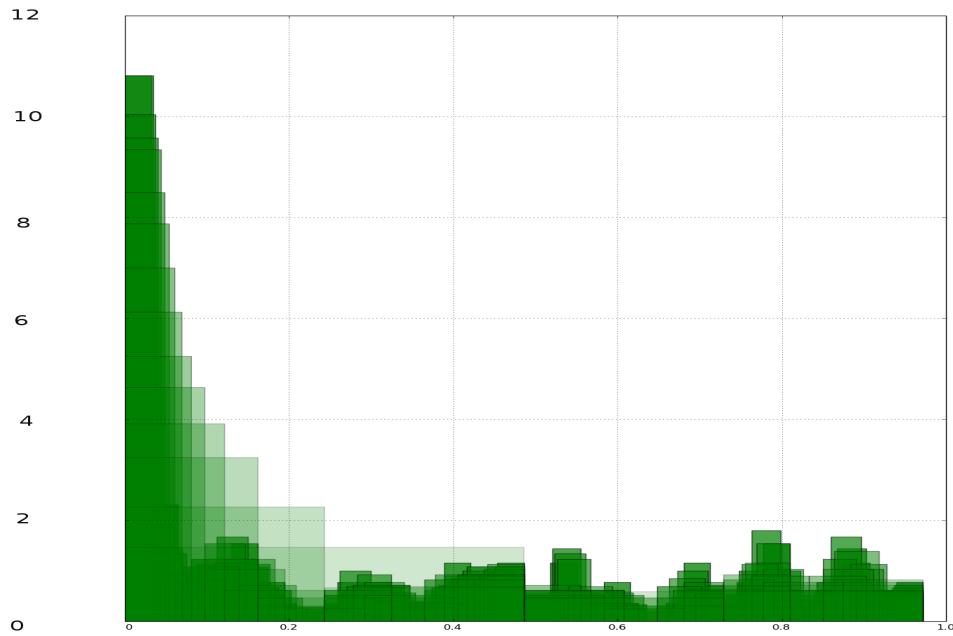


Figure 13.6: The non-Mt protein plot of all locations of *acetylation* MSs found *before* the domains in *C. familiaris* (dog). This plot resembles that of Figure 13.5 which also has the same axes.

Describing the MSs coming *before* domains, we note in Figures 13.5 (*H. sapiens*, non-Mt) and 13.6 (*C. familiaris*, non-Mt) that there are generally many MSs at the beginning of proteins and, consequently, located far upstream from the domain beginning locations. We note in Figure 13.6 that there are bundles of MS populations (neighbourhoods) that are not found in Figures 13.5 and *O. cuniculus* Figure 13.8 (*before*, non-Mt). This may support that notion that *C. familiaris* domains require that MSs be in different positions than those of *H. sapiens* and *M. musculus*, shown in Figure 13.7, in which we note budding and less pronounced neighbourhoods. These placements of MS neighbourhoods in an organism's proteome may suggest influences from environmental stresses and on protein folding^[150].

In Figure 13.9 (*H. sapiens*, Mt) and Figure 13.10 (*M. musculus*, Mt), we observed

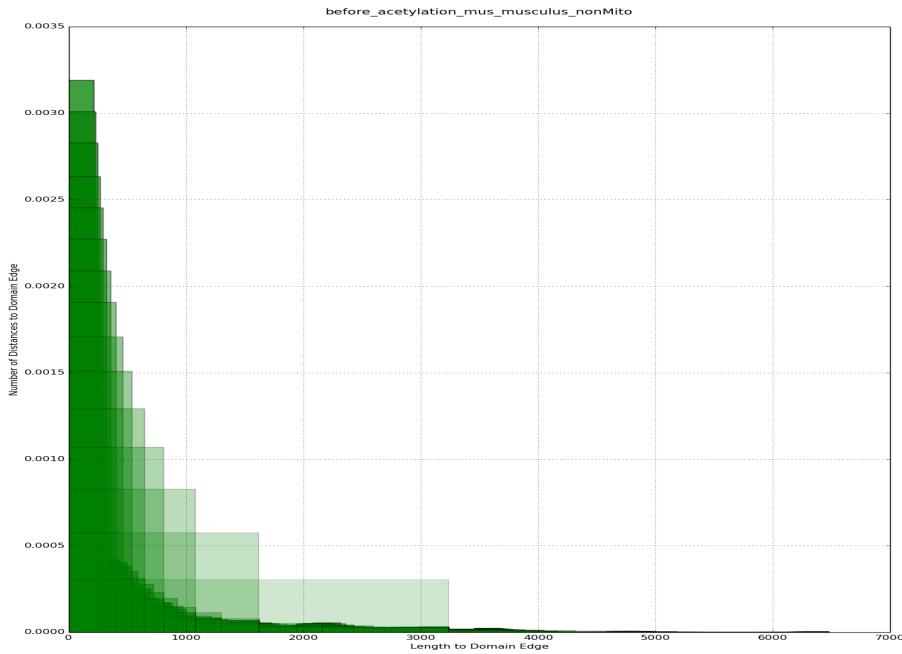


Figure 13.7: The non-Mt protein plot of all locations of *acetylation* MSs found *before* the domains in *M. musculus* (mouse).

the Mt plots of MSs encountered *before* organismal domains. We note that these plots resembled the plots of non-Mt protein, yet may be considered to have more noticeable formations of MS neighbourhoods. In Figure *O. cuniculus* of Figures 13.11 (*before*, Mt), we note these developing neighbourhoods.

We display the plots of MSs encountered *inside* domains in Figures 13.12 (*H. sapiens*, non-Mt), 13.13 (*C. familiaris*, non-Mt) and 13.14 (*M. musculus*, Mt) and we note that these plots appear to be opposite to those of the *before* data. We note that one of the general patterns from all the above non-Mt plots (i.e., *before* and *inside*) is that MSs tend to adhere to both sides of the domain starting positions.

In Figures 13.9 (*H. sapiens*, Mt), 13.10 (*M. musculus*, Mt) and, 13.11 (*O. cuniculus*, Mt) we provide some of the plots of Mt proteins where the MSs came *before* domains. Clearly, the plots from *H. sapiens* and *M. musculus* have much in

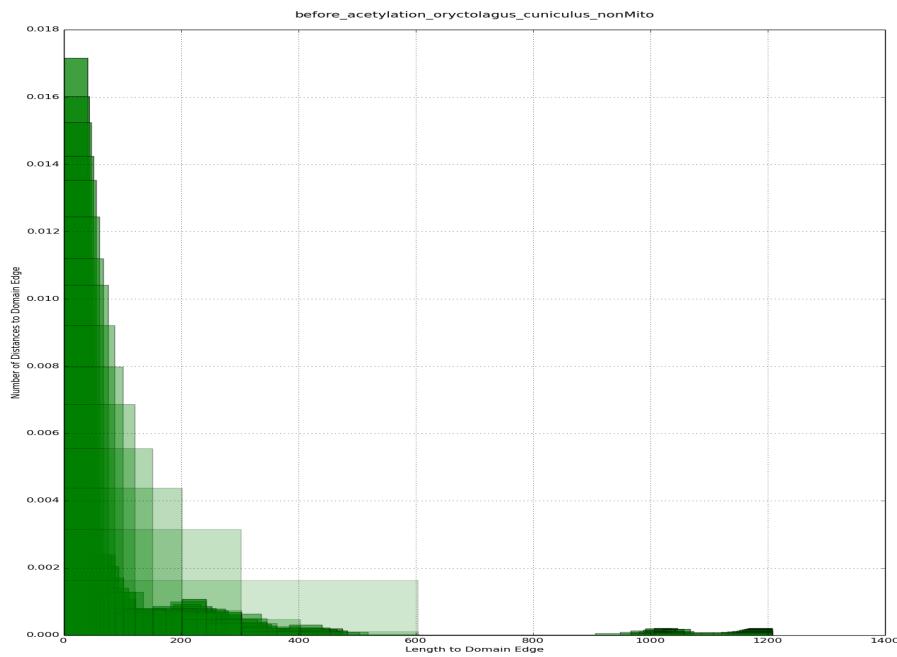


Figure 13.8: The non-Mt protein plot of all locations of *acetylation* MSs found *before* the domains in *O. cuniculus* (rabbit).

common with each other and neither of them closely resembles the plot from *O. cuniculus*. Since these three organisms are closely related, it is unexpected to discover such an unusual distribution of neighbourhoods. We maintain that the differences in neighbourhood patterns may arise by (1) different types of functional domains when compared to those of the other organisms, or (2) there are inherent biases in the data which may be related to PTM usage.

Across non-Mt protein, the plots of MSs *inside* domains, we note in Figures, 13.12 (*H. sapiens*), 13.13 (*C. familiaris*), 13.14 (*M. musculus*) and 13.15 (*O. cuniculus*) that the plot for *M. musculus* is unlike the other two plots (containing clumps of neighbourhoods), and might suggest biases of PTM usage. The plots of Figures 13.16 (*M. musculus*, Mt) and 13.17 (*O. cuniculus*, Mt) describe a striking difference between the MS neighbourhoods in Mt protein. Since Mt is highly conserved, such a contrast

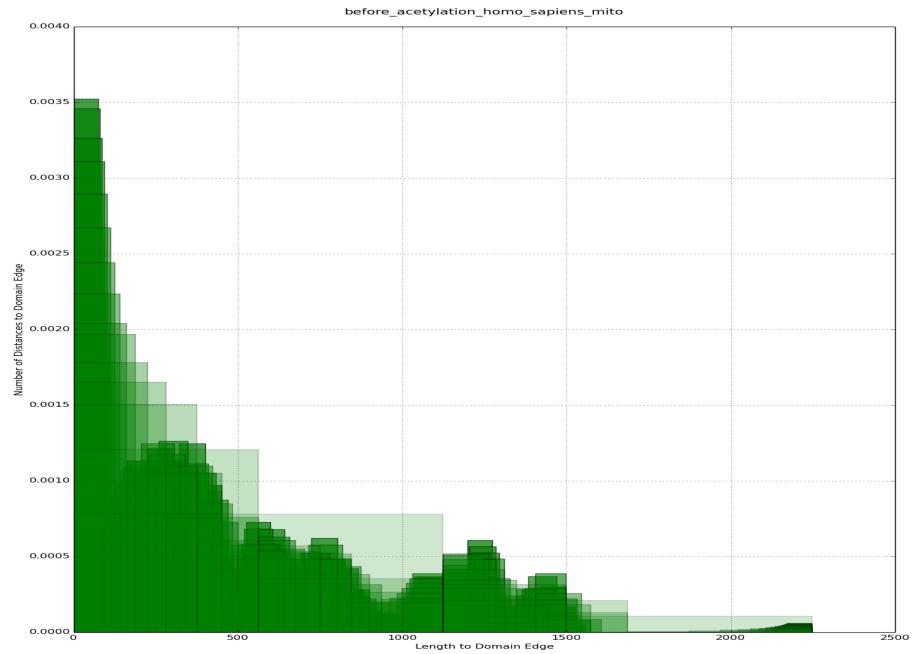


Figure 13.9: The Mt protein plot of all locations of *acetylation* MSs found *inside* the domains in *H. sapiens* (human).

in neighbourhood formations suggests a type of PTM bias.

We noticed that the plots of *acetylation* MSs occurring *before* the domains contained similar types of unifying themes, and were unlike those of the *inside* data. For example, in Figure 13.12 (*H. sapiens*) and Figure 13.13 (*C. familiaris*, dog), we note very turbulent arrangements of neighbourhoods. The pattern was completely different from those of where MSs occurred *before* domains. The similarity of the amino acid compositions shown in the heatmaps of Section 13.4.0.3 suggested that the dissimilarity of the MS neighbourhood placements was exceptional and may be related to the same kinds of PTM usage biases discussed in [34;35].

The plots of the MSs occurring *before*, *inside* and *after* the domains all had distinctive themes, yet all contained types of neighbourhoods where different PTMs (*acetylation* and *phosphorylation*) could be found together. This may suggest

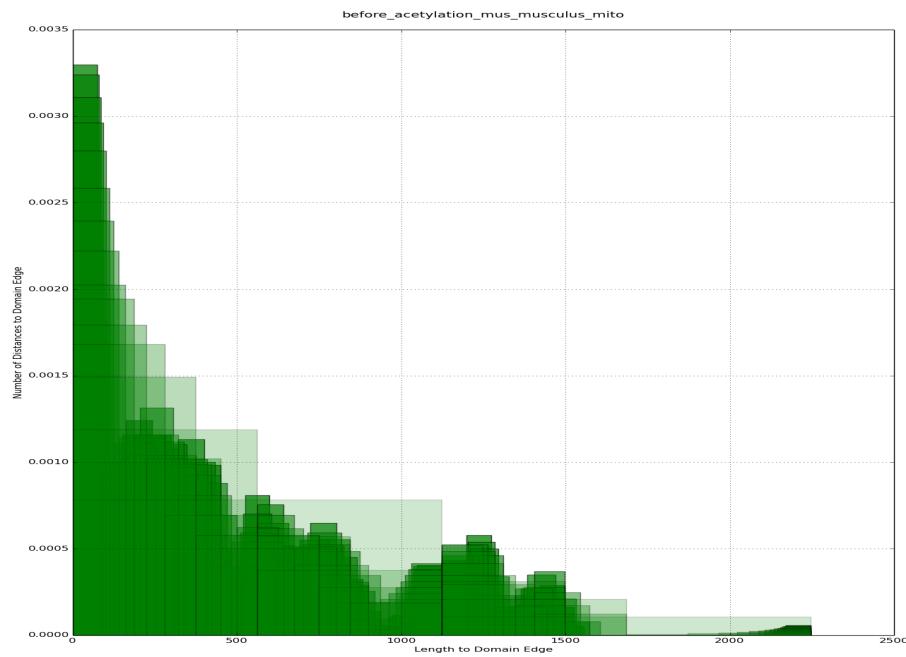


Figure 13.10: The Mt protein plot of all locations of *acetylation* MSs found *inside* the domains in *M. musculus* (mouse).

cross-talking,^[201], however, we suggest that these plotted neighbourhoods, found at similar locations in relation to domains, may suggest that these placements have some importance for protein folding. Compositions of amino acids of the *before*, *inside* and *after* data were very similar to each other and are typical of the heatmap of Figure 13.21 (the data from MSs coming *before* the domains).

13.4.0.2 Graphical Interpretation - Domain-Centric Study Of PTM Distances

In Figure 13.18, we describe how to read the plots containing domain location information and MS neighbourhoods. The proportional value of the location is calculated. Similar to a percentage of the distance that a traveler must venture to reach a point along a path, a proportional distance indicates the beginning and

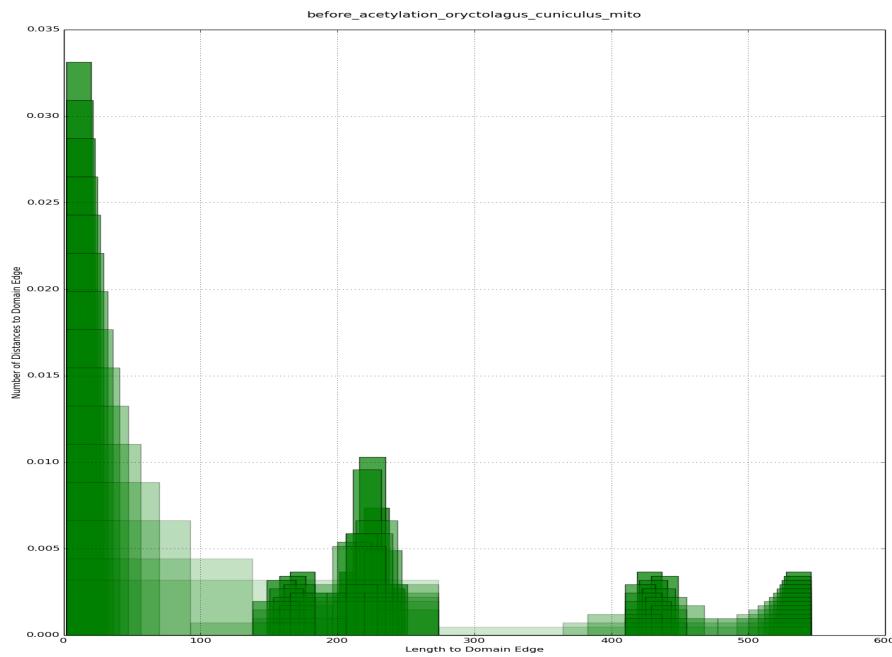


Figure 13.11: The non-Mt protein plot of all locations of *acetylation* MSs found *before* the domains in *O. cuniculus* (rabbit).

ending positions of a domain in relation to the domain beginning and ending positions of other proteins where the same and specific domain is also found. The red bar indicates the proportional distance along the protein where a domain begins and the blue represents the proportional distance to its end. The magnitudes of the bars indicate the number of other domains found in other proteins having the same locations where they begin and end. We note that this information could stem from several different organisms at a time since a domain may be found in a wide variety of organismal proteomes. The green bar represents the MS neighbourhoods and its magnitude indicates the number of other MS at this same region.

The visual representation allows for conveniently determining common distances (and magnitudes) of MS neighbourhoods in relation to domain positions. In these plots, we view all proteins from beginning to end which contain a specific domain

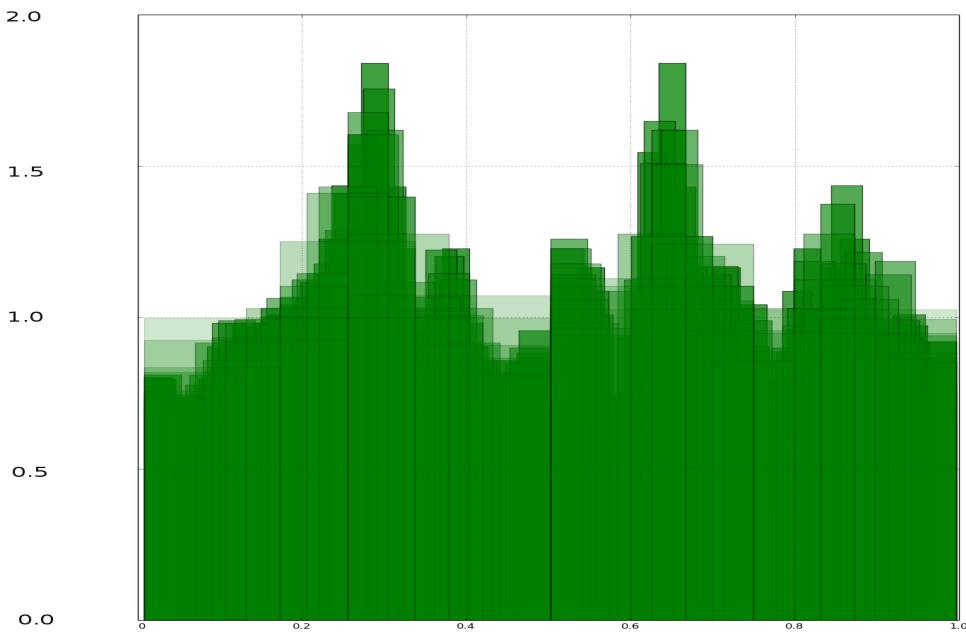


Figure 13.12: The non-Mt protein plot of all locations of *acetylation* MSs found *inside* the domains in *H. sapiens* (human).

(i.e., a user-selected domain). In the plots of Figures 13.19 (Mt) and 13.20 (non-Mt), the red and blue coloured clumps indicate the beginning and ending positions of domains, respectively. The green clumps indicate the position curated (by UniProt) MS populations which may likely influence the domains. In these plots, we note that the domain *atp-grasp2* occurs downstream of the MS neighbourhoods. This finding may be expected since domains are already highly conserved and if so, these plots suggest that these MS neighbourhoods may also be a part of the domain's mechanism in both organisms. Each of these plots in Figures 13.19 and 13.20 incorporates the proteins of at least four organisms and so this combined information from this domain creates a consensus of its trends with respect to *acetylation* from our data.

Having particular MSs found at specific regions may shed some light onto the mechanism of the domain's regulation. For instance, the domains consist of highly

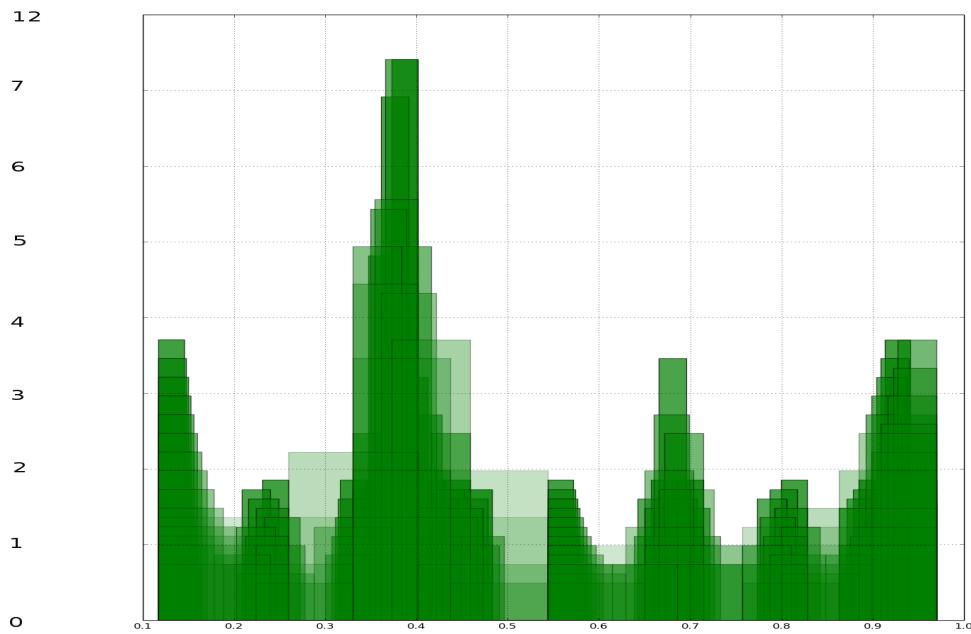


Figure 13.13: The non-Mt protein plot of all locations of *acetylation* MSs found *inside* the domains in *C. familiaris* (dog).

conserved sequence code and so it stands to reason that the MS interaction mechanisms are also conserved. In this case, their relative positions and, perhaps, the distances of MS amino acids from the domains, may also play a part in the function of a conserved functional domain. For instance, returning to Figures 13.19 (Mt) and 13.20 (non-Mt), the domains and MS neighbourhoods describe similar trends in terms of spacings and locations.

13.4.0.3 Analyses By Heatmaps

During the processing of protein samples, we recorded the MS amino acids residue and domain locations. If an MS was found to occur before a domain, then the region of amino acids (shown in Figure 13.1) making up the MS's neighbourhood was collected. We used heatmaps to describe the compositions of the sequence material which was

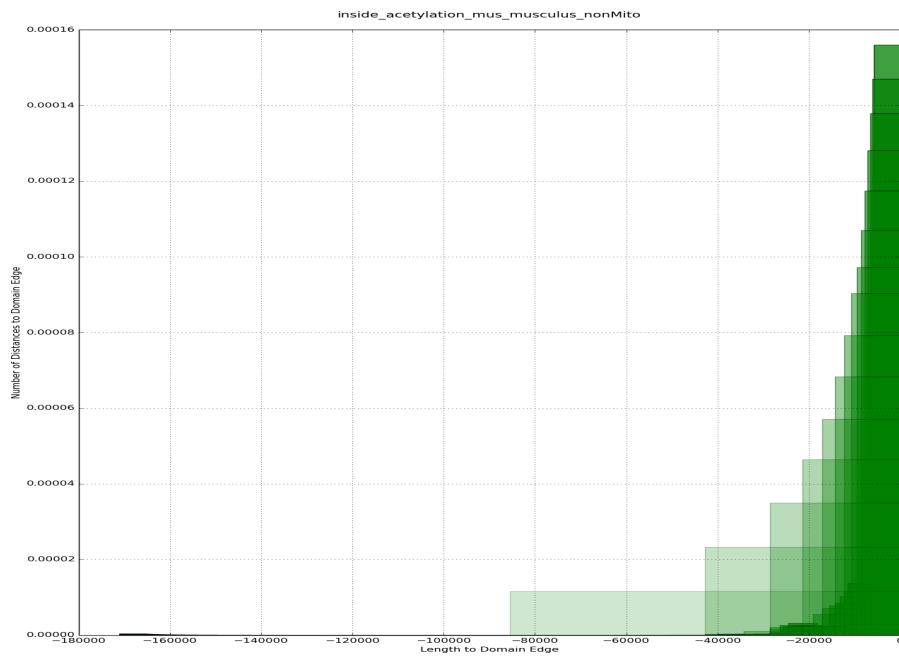


Figure 13.14: The non-Mt protein plot of all locations of *acetylation* MSs found *inside* the domains in *M. musculus* (mouse).

processed in the *before*, *inside* and *after* collected regions.

We read the heatmaps according to red to blue hues where reds and blues indicate higher and lower frequencies, respectively. Looking down the column, if the hues are all the similar, then this is an indication that the frequencies are similar. In addition, similarity indicates that the amino acid compositions are also similar across the organism protein regions.

In our organisms, we noted consistent patterns of amino acids composition. Consulting the heatmaps of all organisms of Figures 13.21 (*before*), 13.22 *inside* and 13.23 *after*, we illustrate this consistency of amino acid frequency. The existence of MS neighbourhoods containing many PTM MSs which are separated by sparse areas, cannot be adequately explained by the composition alone. It is very likely, therefore, that existence of these neighbourhoods are sponsored by the same kinds

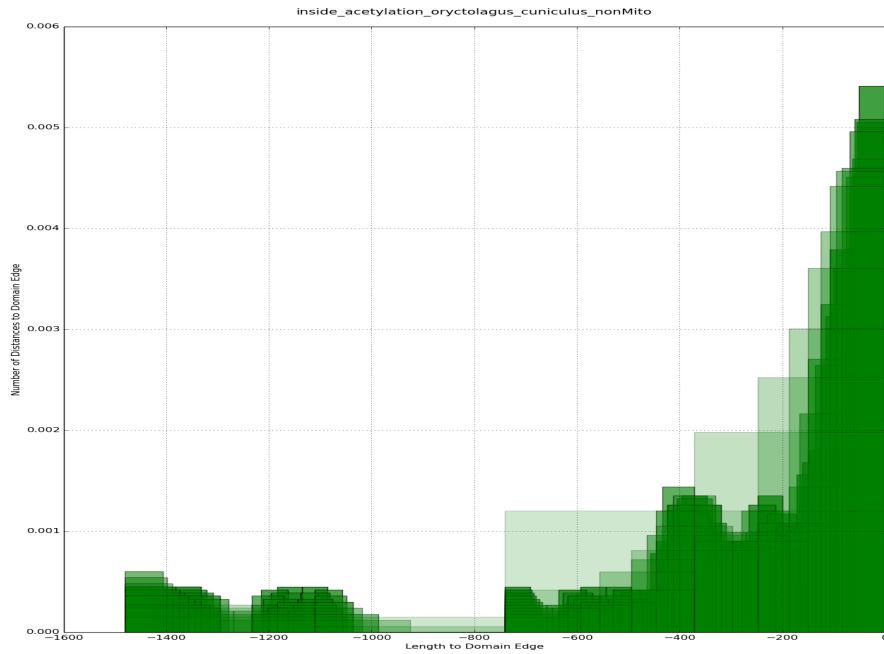


Figure 13.15: The non-Mt protein plot of all locations of *acetylation* MSs found *inside* the domains in *O. cuniculus* (rabbit).

of biases, as seen in our previous work with PTMs^[34;35].

13.5 Conclusions

We present a system called “PTM Tracker” to describe and visualize trends stemming from the natural distances between MS neighbourhoods and protein domains. Accustomed as we are to investigating phenomena on a case-by-case basis, the proposed system examines MSs and domain distances from a global view. We maintain that in addition to the conserved nature of protein domains, the distances between their MSs (perhaps serving as “switches’ for regulation) are also conserved. We provide evidence of this conservation from the study of distances between MSs and all domains encountered in an organism’s proteome, and also from the study of

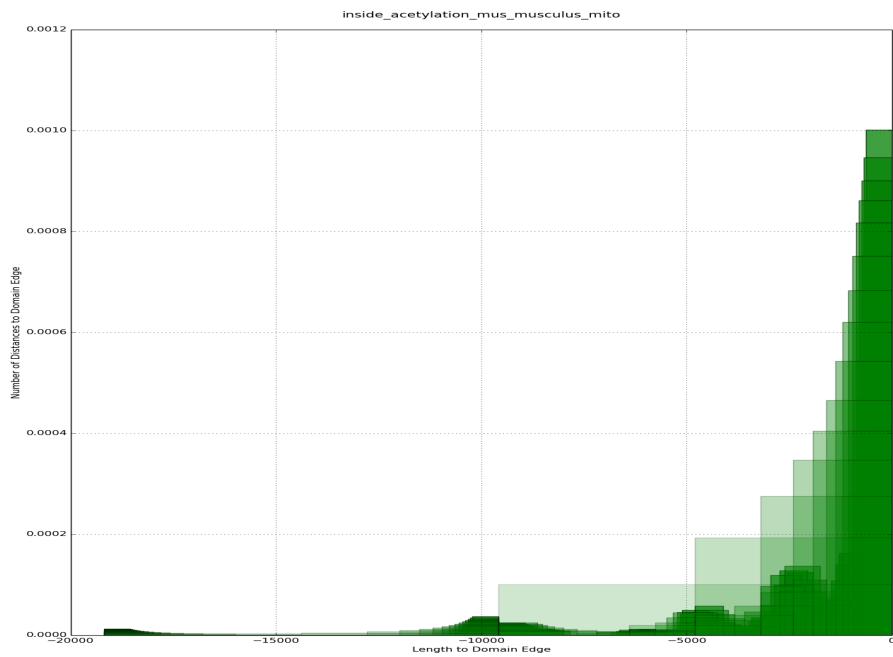


Figure 13.16: The Mt protein plot of all locations of *acetylation* MSs found *inside* the domains in *M. musculus* (mouse).

MSs from specifically chosen domains which are encountered in both Mt and non-Mt proteomes, of a wide variety of organisms. In the plots, we noted location similarities of MS neighbourhoods amongst related organisms. This evidence suggested that the conserved nature of domains may also be extended to the MS which are likely to influence them. Across these organisms, there were still unexplained dissimilarities to suggest that PTM bias was an integral factor in their proteomes.

PTM-Tracker focuses on MSs found *before*, *inside* and *after* domains where it finds patterns to help classify types of domains. For instance, we were able to determine particular domains for which the MSs were always upstream or downstream. Using this classification system, types of domain regulation may be explored and eventually explained as this information may suggest protein folding constraints.

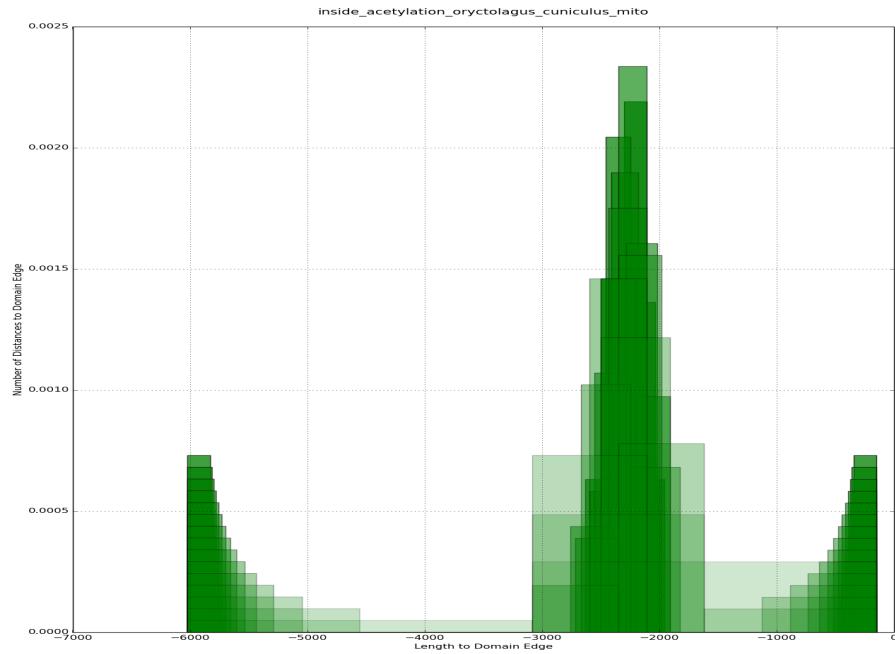


Figure 13.17: The Mt protein plot of all locations of *acetylation* MSs found *before* the domains of *O. cuniculus* (rabbit).

In future work, we intend to include more plots of these trends and themes for each of the regions relative to the domains that we have studied. In addition, we will study the common domains of varied types of protein samples to learn more about the conserved nature of the MS distances. We would also like to apply our system to study domains that function with similar types of mechanisms to determine whether MS neighbourhoods are also similar. After the publication of our extension of this work, we intend to release the software, written in Python, to the community for use and development.

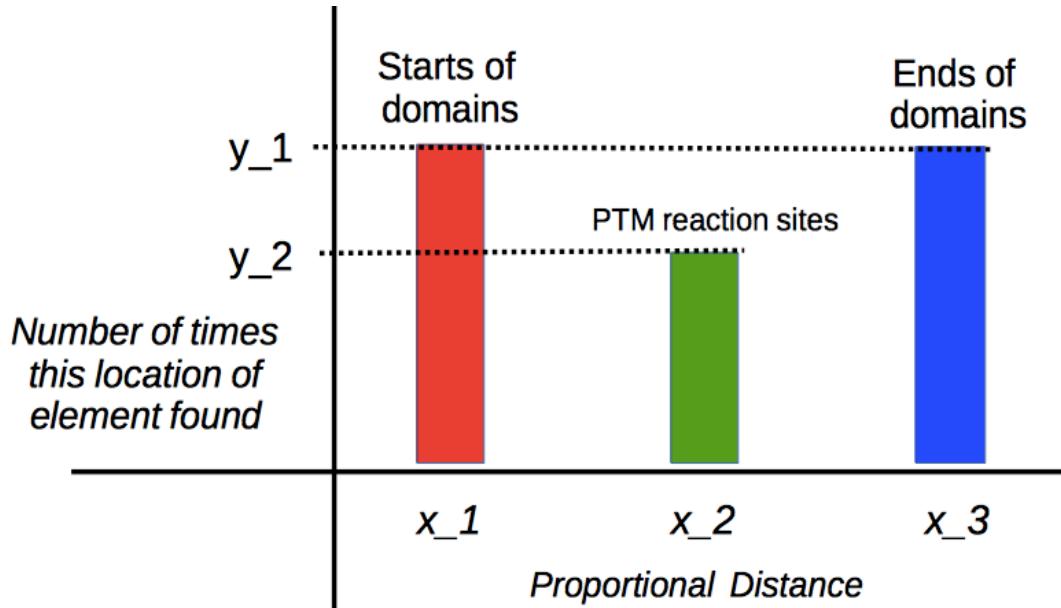


Figure 13.18: How to read plots containing domain and MS information. Red and blue bars represent the proportional distances ventured across the protein to reach the beginning and end of a domain. Their magnitudes indicate the number of other domains having the same proportional values. The green bar describes the proportional locations of the MS neighbourhoods for *acetylation* (or any single PTM of interest). The magnitude describes the number of encountered MSs at these same locations. The plot in this example describes that the MS neighbourhood is found *inside* domains.

13.6 Article Details

This contribution was accepted at the 2016 IEEE International conference on Electro/Information Technology, <http://www.eit-conference.org/eit2016/>.

- “PTM Tracker: A system for determining trends of PTM modification sites relative to protein domains”, 2016

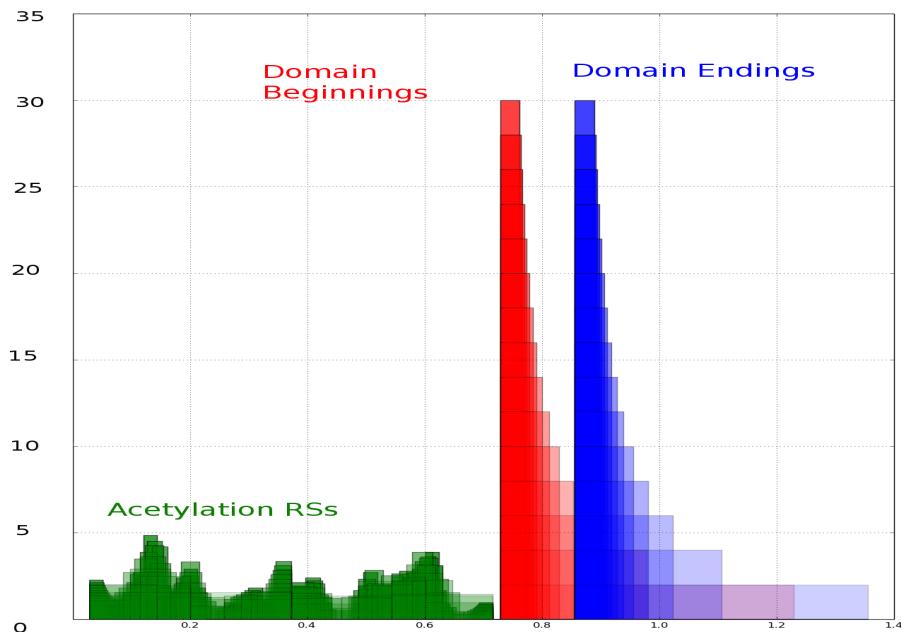


Figure 13.19: The proportional distribution of *acetylation* MSs, encountered in all proteins in UniProt for the 11 organisms of our study, before the *atp-grasp2* domain. We note how this plot is simple to that of Figure 13.20. The *x*-axis represents the location of the start (blue) and end (red) of the domain, in addition to the MS neighbourhoods (green). The *y*-axis describes the number of times that this same location was observed for the across the samples.

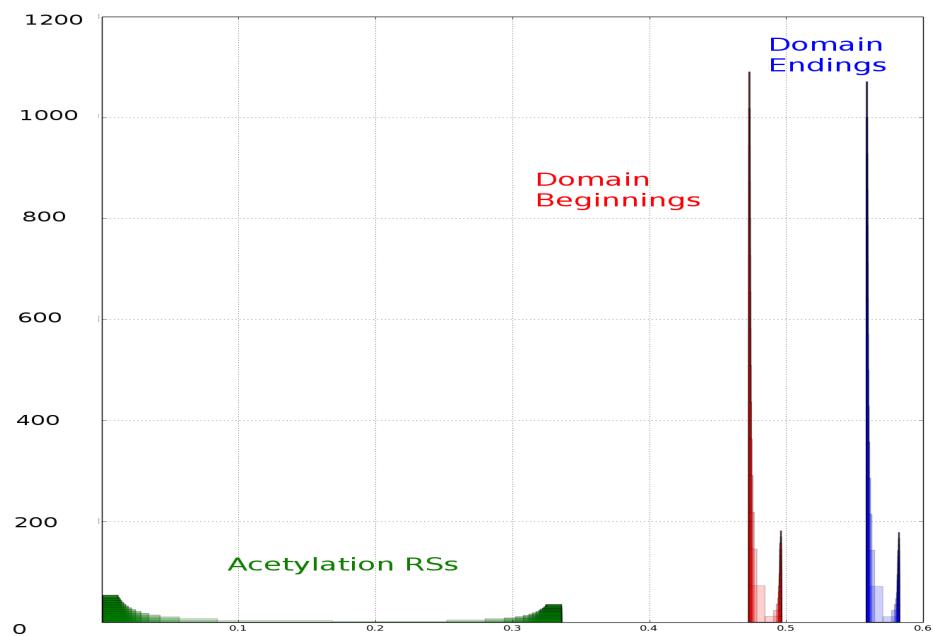


Figure 13.20: The non-Mt plot of *acetylation* MS encountered before the *atp-grasp2* domain. We note how this plot is similar to that of Figure 13.20.

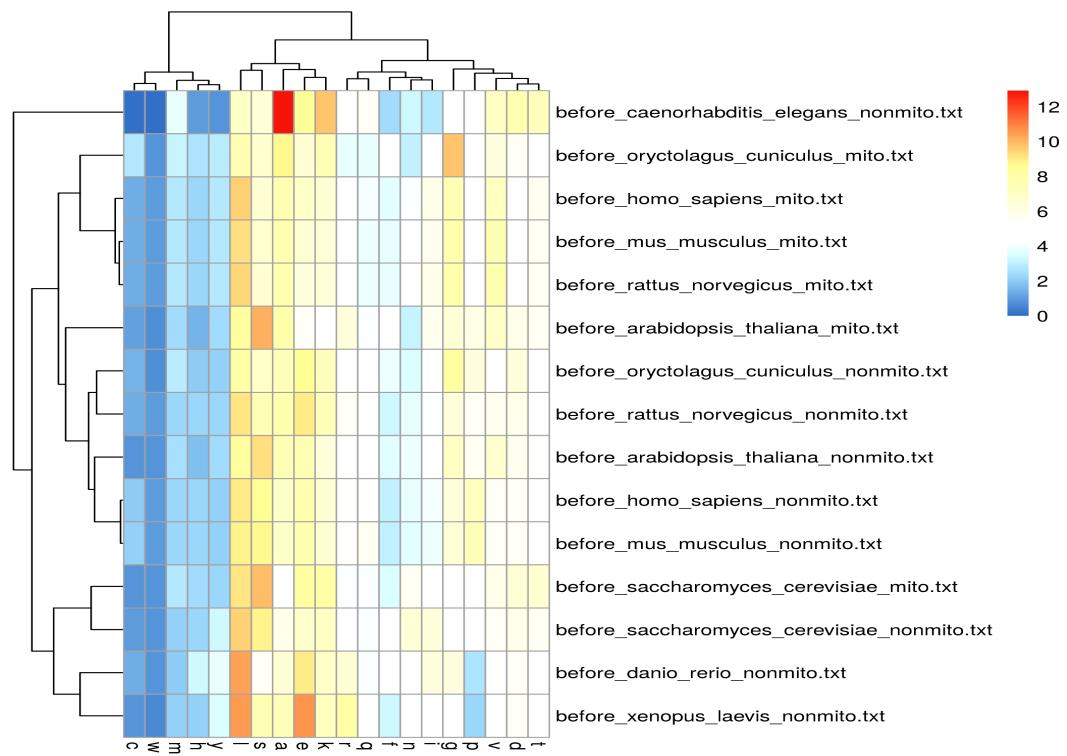


Figure 13.21: A heatmap of amino acid composition of Mt and Non-Mt protein in MS neighbourhoods *before* the domains. We note that the amino acid compositions are very similar across the samples.

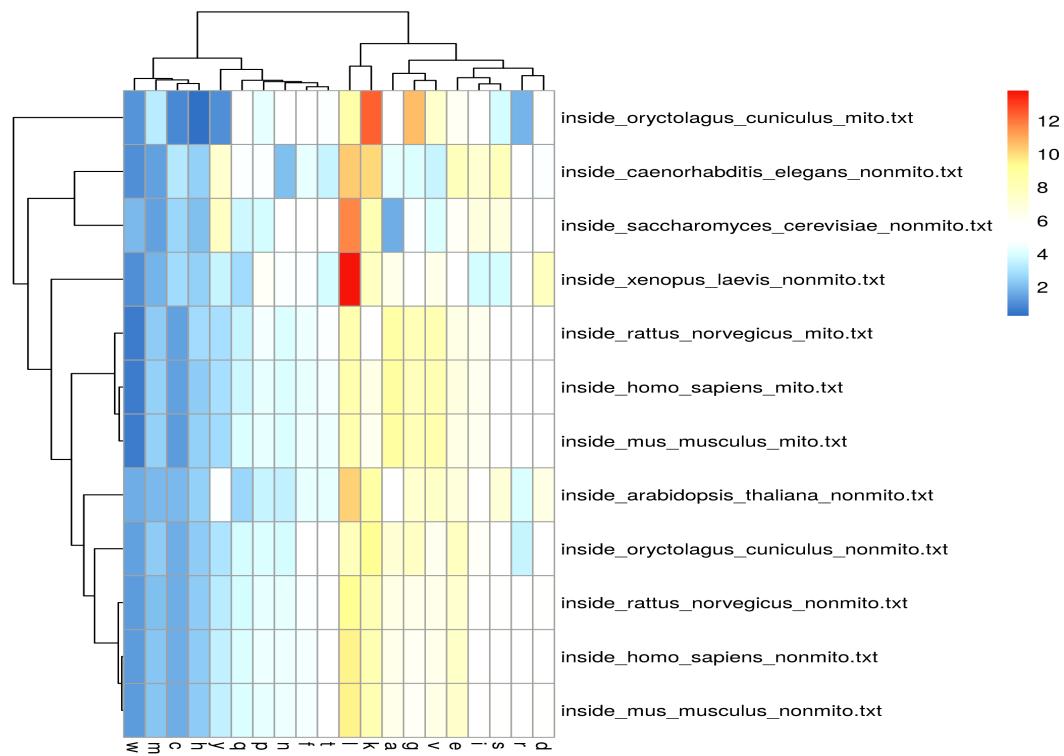


Figure 13.22: A heatmap of amino acid composition of Mt and Non-Mt protein in MS neighbourhoods *inside* the domains. We note that the amino acid compositions are very similar across the samples.

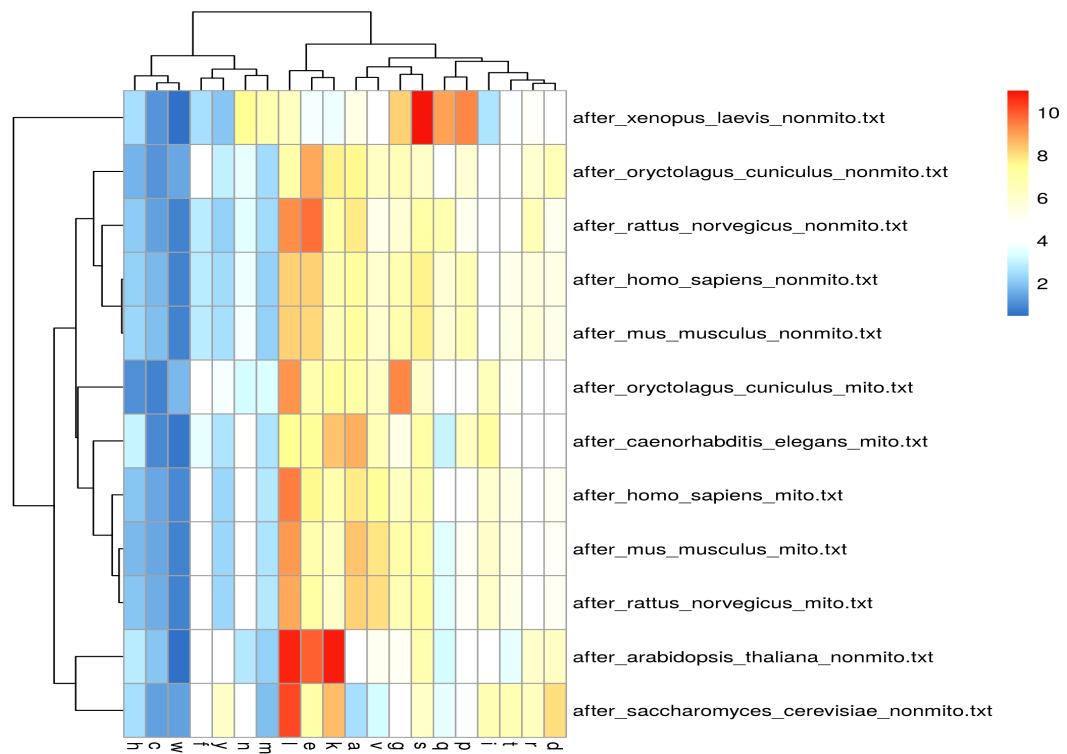


Figure 13.23: A heatmap of amino acid composition of Mt and Non-Mt protein in MS neighbourhoods *after* the domains. We note that the amino acid compositions are very similar across the samples.

Never, never, never give up.

Winston Churchill

Chapter 14

Conclusions

Signals are all around us: they are found where ever there is life, in its movements and mechanisms. At the global level, signals of ourselves are left behind us where ever we go: we mark our journeys as we walk across beaches, just as we leave digital traces of ourselves at each Internet site that we visit. In biology, signals are also left behind as a by-product of all mechanisms. In the case of a beating heart, for example, the blood-pumping action creates sounds and distinctive electrical impulses, which may be surveyed and measured to determine and track the health of the heart. At the cellular level, signals also travel between cells during inter-cellular communication. For instance, a neuron is a type of an electrically excitable cell that processes and transmits information through electrical and chemical signals. Pathways, serving important biological functions, are created from the neural networks that are formed from interacting cells having the ability to send and receive valuable signals.

The existence of a mechanism may be recognized by the basic truth that all mechanisms release signals. In this thesis, we studied signals occurring from

processes at the DNA, RNA and protein levels to detect mechanisms. At the DNA level, we investigated some of the restriction enzymatic defense systems which rely on palindromic motif for function. Here, we uncovered a wealth of evidence to suggest that DNA motif biases exist between coding and non-coding regions of DNA. In the tRNA level, we determined that the biases of DNA may introduce biases in transfer RNA mechanisms, as well. At the protein level, biases were also found to result out of necessity and may improve survival for types of tissues. For instance, in Mt where oxidation prevails, proteins were found to be structurally different from non-Mt proteins: Mt proteins tended to have fewer “RKPT” regions (i.e., sites enriched in arginine; R, lysine; K, proline; P, and threonine; T residues), and also fewer “PEST” regions (i.e., sites enriched in proline; P, glutamic acid; E, serine; S, and threonine; T residues). The mechanism behind this bias was reasoned to be one of protection for Mt proteins since these RKPT and PEST regions tended to attract alteration from oxidation, having fewer of these regions near sources of oxidation may be a naturally selected method to prevent types of protein degeneration.

It was not the scope of this work to fully explain the functions of mechanisms from their signals. Our focus, instead, was to create the basis and technology to detect the existence of mechanisms according to some of their signals. In many cases, our interests in signal detection was reserved to specific fields such as protein stress response systems where we crafted special techniques to isolate some of the tell-tail patterns of mechanisms from sources of “big-data” and other public knowledge. Our analysis provided evidence to confirm the existences of mechanisms and also to infer some of their functional abilities. In all our contributions of this thesis, we maintained plausible rationals behind these mechanisms which were based on pairing the signals, gathered using computational instrumentation, to likely biological concepts and philosophies. We ventured further to explain some of the reasons for

their existence which were aligned with natural evolution and formed from their environmental influences and general biological systems.

14.1 Gains From The Contributions

- In Chapter 4 we described some of the mathematics and statistics which may be conveniently used by bioinformaticians to determine the informational content of sequences (DNA, RNA and protein) to help differentiate them from each other. In this chapter, we argue that techniques from dynamic programming are becoming obsolete due to the larger sizes of sequences that are necessary to compare. Although heuristics may be applied, there will be a time when even this approach will become obsolete. It is therefore necessary that techniques from mathematics and statistics be applied to handle this voluminous data.

Many of the mathematical concepts that we discussed were for obtaining signal frequencies in large data sets. Once these frequencies where collected, then other concepts were recommended for comparative analyses between the sets. This contribution was aimed at putting the power of these tools in the hands of community members who may not be aware of the benefits from this type of mathematical and statistical analysis.

- In Chapter 5 we isolated some of the signals that concerned palindromic DNA motifs. A palindrome, we noted, is a DNA word that appears the same forward and backward in bound DNA. The palindrome also marks the specific region where a restriction enzyme may cut the DNA. These motifs occur naturally in viral genomes. During a common virus attack, the virus' strategy is to mix its DNA with that of the host so that the host will continue to create new viral bodies at the cell's expense. To prevent the attack, the host cell will deploy restriction enzymes to lacerate the invading DNA at locations of the motifs and

stop the threat. Since these same motifs may also appear in the host's genetic code, self-motifs are methylated to prevent them from also being cleaved.

In our work, we used non-parametric tests to discover that there were more of these palindromic motifs in the non-coding regions of the host's DNA when compared to the coding-regions (i.e., where the genes are generally found). We concluded that methylation to protect these regions in host-DNA may fail and a cut in coding regions would be fatal. In addition, we speculated that coding regions have a stringent syntax for the code which would not be favorable to any modifications, whether random or from viral infection.

- In Chapter 6 we developed a pre-processing step for sequence assembly tasks to hurry the assembly process. During a sequence assembly task, reads, or chunks of DNA, must be put back together like a jigsaw puzzle to make larger subsequences called contigs. There are many computing resources required for this task and so we developed a pre-processing step to reduce the necessary time to assemble the entire DNA sequence.

This pre-processing step was developed from our observation that informational content is not consistent throughout an organism's DNA. Instead, there are regions where the frequencies of particular words are higher and so our method exploited this concept to localize many of the reads before they were assembled using the local motif frequencies. Having knowledge of where these subsequences may fit before using "shot-gun" methods to test for adjacency, was found to reduce the amount of time to assemble the reads and contigs of an entire sequence.

- DNA is converted to protein via transcription and translation. Both of these mechanisms are simply the conversion of information from one form to another. In computer security, encryption is also a process of information conversion.

In Chapter 7, we used transcription and translation to convert plain-text into encrypted data by a system which functions similarly to the mechanism of protein synthesis in biology. By following the central dogma of biology, we showed how it was possible to use an ancient mechanism to encrypt and decrypt modern data.

- In Chapter 8 we returned to DNA research where we show that the trends of DNA palindromic word placement influence the structure of transfer RNAs. For instance, tRNAs are responsible for moving a specific amino acid into the protein sequence. Each tRNA also has a particular DNA triplet (instructions in the code) that it follows to perform this task. When there are few triplet codes for a particular amino acid, then there are a reduced number of tRNAs available in an organism's genome. We found this trend across eight organisms where we concluded that biases in DNA equated to biases in tRNA populations of an organism.
- In Chapter 9 we returned to tRNA signals to study their patterns of availability in the genome of nine organisms. We noted that the biases of the tRNA were carried into the protein level where they influenced PTM interactions with the protein amino acids. For instance, each PTM must physically interact with a specific amino acid modification site (MS). Since each amino acid must be placed in the protein sequence by a tRNA that is associated with a particular amino acid, fewer tRNAs for an amino acid may reduce the numbers of this amino acid in the protein. If these reduced amino acids happen to also be the MS for a specific PTM (i.e., lysine is the MS of acetylation), then there will also be fewer tasks performed by the interacting PTM due to a lack of available modification sites for PTMs in the protein. In this study, we studied the frequency factors (PTM predominance and their associated active sites, tRNAs and amino acids)

which likely influence a PTM bias. Our study was performed across both Mt and non-Mt proteins to offer evidence to argue that this PTM bias may be the result of these factors which combine in a poorly understood system to affect and control PTM interactions.

- In Chapter 10 we studied the stress response systems employed by Mt protein. As Mt produce energy by cellular respiration, they also produce internal oxidation which become dangerous sources of protein stress. Interestingly, these proteins do not appear to suffer as a result of this stress. In our work, we investigated the structural signals of protein to describe that there were many fewer regions (enriched in specific amino acids) in Mt where oxidation could potentially damage the protein. Similar to having fewer dangerous places around where accidents may happen, Mt protein appears to have evolved conformational changes to prevent chronic oxidation from causing stress-related damage. In comparison, there was no evidence to suggest that non-Mt proteins had also evolved this protection.
- In Chapter 11 we determined the biases of PTM usage, modification sites and amino acid composition across the proteomes of diverse organisms. In this contribution, we hypothesized that the proteins of Mt (thought to resemble primitive *alphaproteobacteria*) contain fewer PTMs than non-Mt proteins (of advanced life forms). Additionally, the trend of increasing number of nuclear encoded protein PTMs appeared to be correlated with the general complexity of the organism. For instance, we noted that higher organisms tended to have many more interactions between their PTMs and MS in their proteomes than the number of these interactions in the lower orders of organisms. We concluded that higher organisms may apply more types of PTMs to inhabit more varied terrains and, higher degrees of hostility of environments. Since PTMs are active

in many types of protein stress responses, we reasoned that having more types of available PTMs, would theoretically offer better survival chances in habitats having more diverse types of stresses.

- In Chapter 12 we presented a text mining tool to mine public, peer-reviewed literature for relationships between the actors of stress response mechanisms (i.e., proteins, stress types and PTM types). The contribution of this project was the development of a highly customizable text mining tool (called *Lister*) which is able to process the abstracts of the entire PubMed corpus to determine which actors are linked (i.e., are found having some form of relationship according to the studies from the literature). Our tool reads abstracts and performs *direct* (i.e., all actors are found in same article) and *indirect linking* (i.e., connections are made but not all actors are found in the same articles).

We studied relationships between keywords concerning stress-types, protein-types and known PTMs. This work allowed us to determine which proteins had been studied (and were therefore, associated in light of particular stress types) to particular stresses and PTMs. This information, in turn, gave us an idea about which players may likely be involved with a stress-response system. We assume that any paper where some or all of our stress response actors are mentioned together is significant and describes that they are associated.

- A Protein domain is a conserved part of a given protein sequence and (tertiary) structure that can evolve, function, and exist independently of the rest of the protein chain. Each domain forms a compact three-dimensional structure and often can be independently stable and folded Protein domains are important to PTM research as they are thought to be altered by PTMs during protein stress responses. In Chapter 13 we studied the location of PTM

MSs in relation to the protein domains with which they are likely to interact. Since domains are highly conserved, we suggest that their extended mechanisms involving PTMs are also likely to be conserved. We discovered that, for each type of PTM, there were basic distances between MSs and domains which were highly conserved. This suggests that the domains, in addition to the spacings from their modification sites, are likely equally conserved in the proteome. Understanding the conserved nature of the local and extended mechanisms of PTM protein domains alterations may help us to understand more about their roles across biology in terms of general stress responses.

In this work we presented a new tool (called *PTM-Tracker*) to analyze the distances between the domain and the MSs. The analysis of these distances presents major trends which we described in our work. We noted that these recognizing these trends may be used to isolate similar types of mechanisms in other biological samples in the case of conserved functionalities.

14.2 Concluding Thoughts

We studied signals to learn and explain the mechanisms that created them. It is by understanding these signals, and ultimately understanding their meanings, that we may begin to understand the mechanisms of their interaction. In particular, our focus was set on protein stress response systems and all efforts to study signals at the DNA, tRNA and protein levels were developed for the study of these stress response systems where PTMs, MSs, as well as, types of stresses and proteins were major actors in the response. All the contributions of this thesis are included to show that our techniques are applicable and have sufficient sensitivity to collect measurements of signals which are appropriate to study protein stress response systems.

In the former part of our work, we developed the methods and tools to discover signals from biological systems to detect the existence of biological mechanisms. In this thesis, we used these signals to gain ideas about the purpose and function of (known and unknown) mechanisms. This insight was made in context with the biological events at play and also with a concentration on the philosophies of evolutionary biology. In the latter part of our work, we apply our knowledge of detecting and exploring the natural signals pertaining to proteins, PTM influences and stress factors, to understand more about their particular mechanisms. To gain insight into their mechanisms, we approached these signals with three main *lenses*:

1. We performed a quantitative study (Chapter 11), involving frequencies and statistical data to ascertain the patterns from the data. In the data, we discovered that complex organisms such as *H. Sapiens* have many more PTM and modification interactions than more simple organisms such as pond dwelling *C. elegans*.
2. We performed an information-based study (Chapter 13) of signals from curated literature to determine how keywords concerning stresses, proteins and PTMs are associated according to the literature. Here we maintain that any connection from one keyword to another may likely represent a relationship which has been supported by some peer-reviewed study. We note that more research is continually emerging in the literature which will help our technique to grow and provide more comprehensive relationship information from the expanding literature. To determine how keywords concerning stresses, proteins and PTMs are related, we developed a text-mining method (called, *Lister*) to process countless articles to gather this relationship information. We note that the provided knowledge of actors and their associations will provide insight into where to begin new research into particular types of protein stress response systems.

3. We performed a study of protein domains in relation to MSs, and hence PTMs, with which they are likely associated (Chapter 10 and 13). The focus of this work was to uncover the extended conserved elements of the domain-MS alteration systems which may play significant roles in stress-response systems. We discovered that the domain and its general distances to an MS appeared to be a generally conserved phenomenon across the proteins of organisms and also across similar domains (from diverse, non-organism-centric, proteins). Generally, conserved mechanisms may be considered to carry a universal importance to explain why the mechanism has been conserved. Finding common patterns, in connection to these conserved systems (and domains) suggests that conservation may be extended to further reaches of the known systems. Since the resolution of stress has been linked to domain processes, we suggest that any common patterns between domains and MS may suggest knowledge may be gained by further study of these stress response systems.

Following the above three approaches, knowledge concerning biological mechanisms from their signals will emerge. Although the full details of the mechanism may take years to completely uncover, the analysis of its signals will help to guide the context and limit the search space for a more explicit study of the mechanism. Furthermore, it is often because of the detection of its signal that there was any initial notion the mechanism's existence. In these cases, we may extend the work (or leave it for others) to further explore the functions of these unknown mechanisms, in efforts to place this new knowledge into the hands of the bioinformatics and life-sciences research community.



Figure 14.1: Lenses for different depths of field and levels of granularity.

14.3 Future Works

Shown in Figure 14.1, a particular camera lens is selected for its depth of field and levels of granularity which conform to the specifications of the task. In scientific work, instruments also exist which contain diverse types of *lenses* to achieve higher (or lower) depths of field for the discovery of patterns embedded in the data to study phenomena. Summarized in Section 14.2, we discussed three such lenses that we used to discover signals from PTM mechanisms to study types of stress responses. In the future work, each of these lenses will be polished for extra clarity in spotting more diverse types of signals in our efforts to broaden the basis of our knowledge of stress response systems.

In the first lens, more statistical tests will be applied to larger sets of data to gain a firmer understanding of the prevailing trends. Larger sets of data will be applied to find trends from PTM and MS interactions and will give us a better, and more global, view for when we begin to compare and contrast mechanisms and PTM interactions in new organisms.

In the second lens, text mining has been utilized to extract information from the literature concerning the relationships between proteins, PTMs and stresses. This

work will be extended to include smarter systems for determining the relationships between the actors in PTM mechanisms. For instance, the algorithm could record all particular relationships, in addition to the types of articles where they originate. Only those relationships which are mentioned in prominent articles from a certain number of impact-factor journals would be recorded. In addition, we can extend our statistical analysis to better rank the importance of the members of the relationships which stem from in-direct linking (i.e., where the actors are not all found in the same articles but have been found to be associated across several unrelated ones.).

In the third lens, there can be a more rigorous examination into the PTM - domain interaction. If the MS and a particular domain type must be separated by a certain distance (i.e., a distinct number of amino acids), then this distancing suggests that protein folding may play a role in this conserved system. This work will be expanded to include concepts from protein folding to help explain why such a placement of an MS from a domain is of such importance.

References

- [1] AHEARN, I.M., HAIGIS, K., BAR-SAGI, D. & PHILIPS, M.R. (2012). Regulating the regulator: post-translational modification of RAS. *Nature reviews Molecular cell biology*, **13**, 39–51. [237](#)
- [2]AITKEN, A. (2011). Post-translational modification of 14-3-3 isoforms and regulation of cellular function. In *Seminars in cell & developmental biology*, vol. 22, 673–680, Elsevier. [221](#)
- [3] ALAM, I., SHARMIN, S.A., KIM, K.H., YANG, J.K., CHOI, M.S. & LEE, B.H. (2010). Proteome analysis of soybean roots subjected to short-term drought stress. *Plant and soil*, **333**, 491–505. [7](#), [8](#)
- [4] ALLEN, D.L., BANDSTRA, E.R., HARRISON, B.C., THORNG, S., STODIECK, L.S., KOSTENUIK, P.J., MORONY, S., LACEY, D.L., HAMMOND, T.G., LEINWAND, L.L. *et al.* (2009). Effects of spaceflight on murine skeletal muscle gene expression. *Journal of Applied Physiology*, **106**, 582–595. [2](#)
- [5] ALTSCHUL, S.F., MADDEN, T.L., SCHÄFFER, A.A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, **25**, 3389–3402. [39](#), [106](#)
- [6] ANDERSON, R., BIHAM, E. & KNUDSEN, L. (2000). Serpent and smartcards. In *Smart Card Research and Applications*, 246–253, Springer. [110](#)
- [7] ANTINK, C.H., BRÜSER, C. & LEONHARDT, S. (2015). Detection of heart beats in multimodal data: a robust beat-to-beat interval estimation approach. *Physiological measurement*, **36**, 1679. [22](#)
- [8] APOSTOLICO, A. & GIANCARLO, R. (1986). The Boyer Moore Galil String Searching Strategies Revisited. *Siam J. Comput.*, **15**, 98–105. [38](#)
- [9] APWEILER, R., BAIROCH, A., WU, C.H. *et al.* (2004). UniProt: the universal protein knowledgebase. *Nucleic acids research*, **32**, D115–D119. [8](#), [150](#), [153](#), [207](#)
- [10] ARNAU, V., GALLACH, M. & MARÍN, I. (2008). Fast comparison of DNA sequences by oligonucleotide profiling. *BMC Research Notes*, **1**, 5. [44](#), [70](#)

- [11] ARRATIA, R., GOLDSTEIN, L. & GORDON, L. (1990). Poisson approximation and the Chen-Stein method. *Statistical Science*, 403–424. [233](#)
- [12] BARAN, J., GERNER, M., HAEUSSLER, M. *et al.* (2011). pubmed2ensembl: a resource for mining the biological literature on genes. *PloS one*, **6**, e24716. [245](#)
- [13] BASU, K., GRAHAM, L.A., CAMPBELL, R.L. & DAVIES, P.L. (2015). Flies expand the repertoire of protein structures that bind ice. *Proceedings of the National Academy of Sciences*, **112**, 737–742. [8](#)
- [14] BATTEY, M. & PARAKH, A. (2012). An Efficient Quasigroup Block Cipher. *Wireless Personal Communications*, 1–14. [110](#)
- [15] BATTEY, M. & PARAKH, A. (2012). Efficient Quasigroup Block Cipher for Sensor Networks. In *Computer Communications and Networks (ICCCN), 2012 21st International Conference on*, 1–5. [110](#)
- [16] BEESE, L., DERBYSHIRE, V., STEITZ, T. *et al.* (1993). Structure of DNA polymerase I Klenow fragment bound to duplex DNA. *Science (New York, NY)*, **260**, 352. [123](#)
- [17] BELAVÝ, D.L., MIOKOVIC, T., ARMBRECHT, G. *et al.* (2009). Differential atrophy of the lower-limb musculature during prolonged bed-rest. *European journal of applied physiology*, **107**, 489–499. [206](#)
- [18] BELTRAO, P., TRINIDAD, J.C., FIEDLER, D., ROGUEV, A., LIM, W.A., SHOKAT, K.M., BURLINGAME, A.L. & KROGAN, N.J. (2009). Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. *PLoS biology*, **7**, e1000134. [8](#)
- [19] BELTRAO, P., ALBANÈSE, V., KENNER, L.R., SWANEY, D.L., BURLINGAME, A., VILLÉN, J., LIM, W.A., FRASER, J.S., FRYDMAN, J. & KROGAN, N.J. (2012). Systematic functional prioritization of protein posttranslational modifications. *Cell*, **150**, 413–425. [10](#)
- [20] BELTRAO, P., BORK, P., KROGAN, N.J. & NOORT, V. (2013). Evolution and functional cross-talk of protein post-translational modifications. *Molecular systems biology*, **9**. [149, 206](#)
- [21] BENNETT, C.H. & BRASSARD, G. (2014). Quantum cryptography: Public key distribution and coin tossing. *Theoretical computer science*, **560**, 7–11. [110](#)
- [22] BENNETZEN, J.L. & HALL, B. (1982). Codon selection in yeast. *Journal of Biological Chemistry*, **257**, 3026–3031. [131, 151](#)
- [23] BENSON, D.A., CAVANAUGH, M., CLARK, K., KARSCH-MIZRACHI, I., LIPMAN, D.J., OSTELL, J. & SAYERS, E.W. (2012). GenBank. *Nucleic acids research*, gks1195. [74, 116, 117](#)

- [24] BERG, O.G. & KURLAND, C. (1997). Growth rate-optimised tRNA abundance and codon usage. *Journal of molecular biology*, **270**, 544–550. [132](#)
- [25] BERKMAN, P.J., SKARSHEWSKI, A., MANOLI, S., LORENC, M.T., STILLER, J., SMITS, L., LAI, K., CAMPBELL, E., KUBALÁKOVÁ, M., ŠIMKOVÁ, H. *et al.* (2012). Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theoretical and applied genetics*, **124**, 423–432. [39](#)
- [26] BODE, A.M. & DONG, Z. (2004). Post-translational modification of p53 in tumorigenesis. *Nature Reviews Cancer*, **4**, 793–805. [6](#)
- [27] BONHAM-CARTER, O. & BASTOLA, D.R. (2015). A text mining application for linking functionally stressed-proteins to their post-translational modifications. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, 611–614, IEEE. [210](#)
- [28] BONHAM-CARTER, O., ALI, H. & BASTOLA, D. (2012). A meta-genome sequencing and assembly preprocessing algorithm inspired by restriction site base composition. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference on*, 696–703, IEEE. [xxii](#), [xxiii](#), **64**, **70**, **85**, **88**, **89**, **95**, **155**, **181**, **182**, **212**
- [29] BONHAM-CARTER, O., NAJJAR, L., THAPA, I. *et al.* (2012). Distributions of Palindromic Proportional Content in Bacteria. Short Paper. [64](#)
- [30] BONHAM-CARTER, O., ALI, H. & BASTOLA, D. (2013). A base composition analysis of natural patterns for the preprocessing of metagenome sequences. *BMC bioinformatics*, **14**, S5. [64](#), [65](#), **70**, **155**, **181**, **182**, **212**
- [31] BONHAM-CARTER, O., NAJJAR, L. & BASTOLA, D. (2013). Evidence of a Pathway of Reduction in Bacteria: Reduced Quantities of Restriction Sites Impact tRNA Activity in a Trial Set. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, 926, ACM. [152](#)
- [32] BONHAM-CARTER, O., PEDERSEN, J., NAJJAR, L. & BASTOLA, D. (2013). Modeling the Effects of Microgravity on Oxidation in Mitochondria: A Protein Damage Assessment across a Diverse Set of Life Forms. In *IEEE Data Mining Workshop (ICDMW)*, 250–257, IEEE. [148](#), [206](#), [248](#), [249](#)
- [33] BONHAM-CARTER, O., STEELE, J. & BASTOLA, D. (2013). Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Briefings in bioinformatics*, bbt052. [213](#)
- [34] BONHAM-CARTER, O., PEDERSEN, J. & BASTOLA, D. (2014). A content and structural assessment of oxidative motifs across a diverse set of life forms. *Computers in biology and medicine*, **53**, 179–189. [206](#), [232](#), [248](#), [249](#), [256](#), [260](#), [270](#), [276](#)

- [35] BONHAM-CARTER, O., THAPA, I. & BASTOLA, D. (2014). Evidence of post translational modification bias extracted from the tRNA and corresponding amino acid interplay across a set of diverse organisms. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, 774–781, ACM. [207](#), [249](#), [256](#), [260](#), [270](#), [276](#)
- [36] BONHAM-CARTER, O., THAPA, I., FROM, S. & BASTOLA, D. (2016). A study of bias and increasing organismal complexity from their post-translational modifications and reaction site interplays. *Briefings in bioinformatics*, bby111. [257](#)
- [37] BOORE, J.L. (1999). Animal mitochondrial genomes. *Nucleic Acids Research*, **27**, 1767–1780. [150](#), [209](#)
- [38] BOUE, S., LETUNIC, I. & BORK, P. (2003). Alternative splicing and evolution. *Bioessays*, **25**, 1031–1034. [237](#)
- [39] BOUTET, E., LIEBERHERR, D., TOGNOLLI, M., SCHNEIDER, M., BANSAL, P., BRIDGE, A.J., POUX, S., BOUGUELERET, L. & XENARIOS, I. (2016). UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. *Plant Bioinformatics: Methods and Protocols*, 23–54. [255](#)
- [40] BOYER, R.S. & MOORE, S.J. (1977). A Fast String Searching Algorithm. *Commun. ACM*, 762–772. [38](#)
- [41] BOYES, J., BYFIELD, P., NAKATANI, Y. & OGRYZKO, V. (1998). Regulation of activity of the transcription factor GATA-1 by acetylation. *Nature*, **396**, 594–598. [164](#), [239](#)
- [42] BRENDL, V., BECKMANN, J.S. & TRIFONOV, E.N. (1986). Linguistics of nucleotide sequences: morphology and comparison of vocabularies. *Journal of Biomolecular Structure and Dynamics*, **4**, 11–21. [51](#)
- [43] BRENT, M.M., ANAND, R. & MARMORSTEIN, R. (2008). Structural basis for DNA recognition by FoxO1 and its regulation by posttranslational modification. *Structure*, **16**, 1407–1416. [221](#)
- [44] BRETT, D., POSPISIL, H., VALCÁRCEL, J. *et al.* (2002). Alternative splicing and genome complexity. *Nature genetics*, **30**, 29–30. [237](#)
- [45] BREWER, B.J., PAYEN, C., RAGHURAMAN, M. & DUNHAM, M.J. (2011). Origin-dependent inverted-repeat amplification: a replication-based model for generating palindromic amplicons. *PLoS Genet*, **7**, e1002016. [130](#)
- [46] BUKAR MAINA, M., AL-HILALY, Y.K. & SERPELL, L.C. (2016). Nuclear Tau and Its Potential Role in Alzheimer’s Disease. *Biomolecules*, **6**, 9. [254](#)

- [47] CAMARA, M., BONHAM-CARTER, O. & JUMADINOVA, J. (2015). A Multi-Agent System with Reinforcement Learning Agents for Biomedical Text Mining. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, ACM. **210**, 245
- [48] CAO, M.D., DIX, T.I., ALLISON, L. & MEARS, C. (2007). A simple statistical algorithm for biological sequence compression. In *Data Compression Conference, 2007. DCC'07*, 43–52, IEEE. **41**, 61, 70
- [49] CASTRONOVO, M., RADOVIC, S., GRUNWALD, C., CASALIS, L., MORGANTE, M. & SCOLES, G. (2008). Control of steric hindrance on restriction enzyme reactions with surface-bound DNA nanostructures. *Nano letters*, **8**, 4140–4145. **74**
- [50] CAUGHEY, B. & LANSBURY JR, P.T. (2003). Protofibrils, Pores, Fibrils, and Neurodegeneration: Separating the Responsible Protein Aggregates from The Innocent Bystanders*. *Annual review of neuroscience*, **26**, 267–298. **11**, 172
- [51] CHA, M.Y., CHO, H.J., KIM, C., JUNG, Y.O., KANG, M.J., MURRAY, M.E., HONG, H.S., CHOI, Y.J., CHOI, H., KIM, D.K. *et al.* (2015). Mitochondrial ATP synthase activity is impaired by suppressed O-GlcNAcylation in Alzheimer’s disease. *Human Molecular Genetics*, ddv358. **13**
- [52] CHAISSON, M.J., HUDDLESTON, J., DENNIS, M.Y., SUDMANT, P.H., MALIG, M., HORMOZDIARI, F., ANTONACCI, F., SURTI, U., SANDSTROM, R., BOITANO, M. *et al.* (2014). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*. **9**
- [53] CHAN, P. & LOWE, T. (2009). GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic acids research*, **37**, D93–D97. **134**
- [54] CHAN, R.H., CHAN, T.H., YEUNG, H.M. *et al.* (2011). Composition Vector Method Based on Maximum Entropy Principle for Sequence Comparison. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. **41**, 51, 54, 70
- [55] CHATFIELD, K.C., COUGHLIN, C.R., FRIEDERICH, M.W., GALLAGHER, R.C., HESSELBERTH, J.R., LOVELL, M.A., OFMAN, R., SWANSON, M.A., THOMAS, J.A., WANDERS, R.J. *et al.* (2015). Mitochondrial energy failure in HSD10 disease is due to defective mtDNA transcript processing. *Mitochondrion*, **21**, 1–10. **13**
- [56] CHEN, C. & RAJAPAKSE, J.C. (2007). Grid-Enabled BLASTZ: Application to Comparative Genomics. *The Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, **48**, 301–309. **39**

- [57] CHEN, H. & CHAN, D.C. (2009). Mitochondrial dynamics–fusion, fission, movement, and mitophagy—in neurodegenerative diseases. *Human molecular genetics*, **18**, R169–R176. [11](#), [171](#)
- [58] CHEN, S.A., LEE, T.Y. & OU, Y.Y. (2010). Incorporating significant amino acid pairs to identify O-linked glycosylation sites on transmembrane proteins and non-transmembrane proteins. *BMC bioinformatics*, **11**, 536. [9](#)
- [59] CHENG, L.L., CHEUNG, D.W. & YIU, S.M. (2003). Approximate string matching in DNA sequences. In *Database Systems for Advanced Applications, 2003.(DASFAA 2003). Proceedings. Eighth International Conference on*, 303–310, IEEE. [39](#)
- [60] CHENNA, R., SUGAWARA, H., KOIKE, T. *et al.* (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic acids research*, **31**, 3497–3500. [39](#)
- [61] CHOMYN, A. & ATTARDI, G. (2003). MtDNA mutations in aging and apoptosis. *Biochemical and biophysical research communications*, **304**, 519–529. [175](#)
- [62] CHOUDHARY, C. & MANN, M. (2010). Decoding signalling networks by mass spectrometry-based proteomics. *Nature reviews Molecular cell biology*, **11**, 427–439. [8](#)
- [63] CHOUDHARY, C., KUMAR, C., GNAD, F. *et al.* (2009). Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science*, **325**, 834–840. [218](#)
- [64] CONNORS, B.W. & LONG, M.A. (2004). Electrical synapses in the mammalian brain. *Annu. Rev. Neurosci.*, **27**, 393–418. [27](#)
- [65] CONSORTIUM, U. *et al.* (2011). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic acids research*, gkr981. [256](#)
- [66] COX, A.J., BAUER, M.J., JAKOBI, T. *et al.* (2012). Large-scale compression of genomic sequence databases with the Burrows–Wheeler transform. *Bioinformatics*, **28**, 1415–1419. [61](#)
- [67] CRAMERI, A., RAILLARD, S.A., BERMUDEZ, E. *et al.* (1998). DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature*, **391**, 288–291. [39](#)
- [68] DAEMEN, J. & RIJMEN, V. (2002). *The Design of Rijndael*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. [109](#)
- [69] DAI, Q., LI, L., LIU, X. *et al.* (2011). Integrating Overlapping Structures and Background Information of Words Significantly Improves Biological Sequence Comparison. *PloS one*, **6**, e26779. [43](#), [44](#)

- [70] DALLE-DONNE, I., ALDINI, G., CARINI, M., COLOMBO, R., ROSSI, R. & MILZANI, A. (2006). Protein carbonylation, cellular dysfunction, and disease progression. *Journal of cellular and molecular medicine*, **10**, 389–406. [173](#)
- [71] DARMON, E., EYKELENBOOM, J.K., LINCKER, F., JONES, L.H., WHITE, M., OKELY, E., BLACKWOOD, J.K. & LEACH, D.R. (2010). E. coli SbcCD and RecA control chromosomal rearrangement induced by an interrupted palindrome. *Molecular cell*, **39**, 59–70. [130](#)
- [72] DIFFIE, W. & HELLMAN, M. (1976). New directions in cryptography. *Information Theory, IEEE Transactions on*, **22**, 644–654. [109](#)
- [73] DiMAURO, S. & SCHON, E.A. (2003). Mitochondrial respiratory-chain diseases. *New England Journal of Medicine*, **348**, 2656–2668. [175](#)
- [74] DOHERTY, A. & McINERNEY, J.O. (2013). Translational selection frequently overcomes genetic drift in shaping synonymous codon usage patterns in vertebrates. *Molecular biology and evolution*, mst128. [152](#)
- [75] DOMAZET-LOŠO, M. & HAUBOLD, B. (2011). Alignment-free detection of horizontal gene transfer between closely related bacterial genomes. *Mobile Genetic Elements*, **1**, 230–235. [39](#), [68](#), [70](#)
- [76] DOMAZET-LOŠO, M. & HAUBOLD, B. (2011). Alignment-free detection of local similarity among viral and bacterial genomes. *Bioinformatics*, **27**, 1466–1472. [41](#), [66](#), [69](#), [70](#)
- [77] DONG, H., NILSSON, L. & KURLAND, C.G. (1996). Co-variation of tRNA Abundance and Codon Usage in Escherichia coli at Different Growth Rates. *Journal of molecular biology*, **260**, 649–663. [132](#)
- [78] DONG, Y., GRAZIANE, N., GRAZIANE, N. & DONG, Y. (2016). Fast and Slow Synaptic Currents. *Electrophysiological Analysis of Synaptic Transmission*, 111–120. [27](#)
- [79] DORES-SILVA, P., MINARI, K., RAMOS, C., BARBOSA, L. & BORGES, J. (2013). Structural and stability studies of the human mtHsp70-escort protein 1: An essential mortalin co-chaperone. *International journal of biological macromolecules*, **56**, 140–148. [196](#)
- [80] DOUGLAS W BRYANT JR., W.K.W. & MOCKLER, T.C. (2009). QSRA – a quality-value guided de novo short read assembler. *BMC Bioinformatics*, **10**. [83](#)
- [81] DU TOIT, A. (2014). Post-translational modification: Sweetening protein quality control. *Nature Reviews Molecular Cell Biology*, **15**, 295–295. [244](#)
- [82] DUARTE, M.R. (2003). Prickly food: snakes preying upon porcupines. *Phyllomedusa: Journal of Herpetology*, **2**, 109–112. [24](#)

- [83] DUNKER, A.K., BONDOS, S.E., HUANG, F. & OLDFIELD, C.J. (2014). Intrinsically disordered proteins and multicellular organisms. In *Seminars in cell & developmental biology*, Elsevier. 10
- [84] EDDY, S.R. (2004). What is dynamic programming? *Nature biotechnology*, **22**, 909–910. 39
- [85] EDGAR, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, **32**, 1792–1797. 93
- [86] ELGENDI, M., NORTON, I., BREARLEY, M., FLETCHER, R.R., ABBOTT, D., LOVELL, N.H. & SCHUURMANS, D. (2015). Towards Investigating Global Warming Impact on Human Health Using Derivatives of Photoplethysmogram Signals. *International journal of environmental research and public health*, **12**, 12776–12791. 22
- [87] FENG, Y., YAO, Z. & KLIONSKY, D.J. (2015). How to control self-digestion: transcriptional, post-transcriptional, and post-translational regulation of autophagy. *Trends in cell biology*, **25**, 354–363. 206
- [88] FERNÁNDEZ-SUÁREZ, X.M. & GALPERIN, M.Y. (2013). The 2013 Nucleic Acids Research Database Issue and the online molecular biology database collection. *Nucleic acids research*, **41**, D1–D7. 177
- [89] FLICEK, P., AMODE, M., BARRELL, D., BEAL, K., BRENT, S., CARVALHO-SILVA, D., CLAPHAM, P., COATES, G., FAIRLEY, S., FITZGERALD, S. et al. (2012). Ensembl 2012. *Nucleic acids research*, **40**, D84–D90. 116
- [90] FORÊT, S., KANTOROVITZ, M. & BURDEN, C.J. (2006). Asymptotic Behavior of k-Word Matches Between Two Uniformly Distributed Sequences. *BMC Bioinformatics*, **7**, S21. 57
- [91] FORÊT, S., WILSON, S.R. & BURDEN, C.J. (2009). Characterizing the D2 statistic: word matches in biological sequences. *Statistical Applications in Genetics and Molecular Biology*, **8**, 1–21. 57
- [92] FUGLSANG, A. (2003). Distribution of potential type II restriction sites (palindromes) in prokaryotes. *Biochemical and biophysical research communications*, **310**, 280–285. 131
- [93] FUGLSANG, A. (2004). The relationship between palindrome avoidance and intragenic codon usage variations: a Monte Carlo study. *Biochemical and biophysical research communications*, **316**, 755–762. 131
- [94] FUNCK, V., RIBEIRO, L., PEREIRA, L., DE OLIVEIRA, C., GRIGOLETTO, J., DELLA-PACE, I., FIGHERA, M., ROYES, L., FURIAN, A., LARRICK, J. et al. (2015). Contrasting effects of Na⁺, K⁺-ATPase activation on seizure activity in acute versus chronic models. *Neuroscience*, **298**, 171–179. 14

- [95] FURMANEK, A. & HOFSTEENGE, J. (1999). Protein C-mannosylation: facts and questions. *Acta biochimica polonica*, **47**, 781–789. [6](#)
- [96] GARG, R., PATEL, R., TYAGI, A. & JAIN, M. (2011). De Novo Assembly of Chickpea Transcriptome Using Short Reads for Gene Discovery and Marker Identification. *DNA Research*, **18**, 53–63. [83](#)
- [97] GARNIER, J., OSGUTHORPE, D. & ROBSON, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of molecular biology*, **120**, 97–120. [190](#)
- [98] GASTON, B.M., CARVER, J., DOCTOR, A. & PALMER, L.A. (2003). S-nitrosylation signaling in cell biology. *Molecular interventions*, **3**, 253. [6](#)
- [99] GEHANI, A., LABEAN, T. & REIF, J. (2004). DNA-based cryptography. *Aspects of Molecular Computing*, 34–50. [111](#)
- [100] GELFAND, M. & KOONIN, E. (1997). Evidence of selection upon genomic GC-content in bacteria. *Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes*, **25**, 2430–2439. [79, 90](#)
- [101] GELFAND, M.S. & KOONIN, E.V. (1997). Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. *Nucleic acids research*, **25**, 2430–2439. [xvii, 74, 130, 132, 133, 134](#)
- [102] GENTLEMAN, J.F. & MULLIN, R.C. (1989). The distribution of the frequency of occurrence of nucleotide subsequences, based on their overlap capability. *Biometrics*, 35–52. [53](#)
- [103] GLICOROSKI, D. (2005). Candidate one-way functions and one-way permutations based on quasigroup string transformations. *arXiv preprint cs/0510018*. [110](#)
- [104] GONG, C.X., LIU, F., GRUNDKE-IQBAL, I. & IQBAL, K. (2005). Post-translational modifications of tau protein in Alzheimer's disease. *Journal of neural transmission*, **112**, 813–838. [244](#)
- [105] GOTO, Y., NIWA, Y., SUZUKI, T., DOHMAE, N., UMEZAWA, K. & SIMIZU, S. (2014). C-mannosylation of human hyaluronidase 1: Possible roles for secretion and enzymatic activity. *International journal of oncology*, **45**, 344–350. [6](#)
- [106] GREGERSEN, N. & BROSS, P. (2010). Protein misfolding and cellular stress: an overview. In *Protein Misfolding and Cellular Stress in Disease and Aging*, 3–23, Springer. [7](#)
- [107] GU, B. & ZHU, W.G. (2012). Surf the post-translational modification network of p53 regulation. *International journal of biological sciences*, **8**, 672. [5, 220, 221](#)

- [108] GUERRA, D., CROSATTI, C., KHOSHRO, H. *et al.* (2015). Post-transcriptional and post-translational regulations of drought and heat response in plants: a spider's web of mechanisms. *Name: Frontiers in Plant Science*, **6**, 57. [206](#)
- [109] GUSFIELD, D. (1997). *Algorithms on strings, trees and sequences: computer science and computational biology*. Cambridge University Press. [38](#)
- [110] HAASS, C. & STEINER, H. (2001). Protofibrils, the unifying toxic molecule of neurodegenerative disorders? *Nature neuroscience*, **4**, 859–860. [172](#)
- [111] HAISSAGUERRE, M., JAÏS, P., SHAH, D.C., TAKAHASHI, A., HOCINI, M., QUINIOUT, G., GARRIGUE, S., LE MOUROUX, A., LE MÉTAYER, P. & CLÉMENTY, J. (1998). Spontaneous initiation of atrial fibrillation by ectopic beats originating in the pulmonary veins. *New England Journal of Medicine*, **339**, 659–666. [21](#)
- [112] HAO, B. & QI, J. (2004). Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. *Journal of bioinformatics and computational biology*, **2**, 1–19. [51](#)
- [113] HAO, B., QI, J. & WANG, B. (2003). Prokaryotic Phylogeny Based On Complete Genomes Without Sequence Alignment. *Modern Physics Letters B*, **2**, 1–4. [50](#)
- [114] HARA, Y. & IMANISHI, T. (2011). Abundance of ultramicro inversions within local alignments between human and chimpanzee genomes. *BMC evolutionary biology*, **11**, 308. [40](#)
- [115] HARGROVE, J.L., HULSEY, M.G. & BEALE, E.G. (1991). The kinetics of mammalian gene expression. *Bioessays*, **13**, 667–674. [2](#)
- [116] HEIDER, D. & BARNEKOW, A. (2007). DNA-based watermarks using the DNA-Crypt algorithm. *BMC bioinformatics*, **8**, 176. [110, 111](#)
- [117] HENCHCLIFFE, C. & BEAL, M.F. (2008). Mitochondrial biology and oxidative stress in Parkinson disease pathogenesis. *Nature clinical practice Neurology*, **4**, 600–609. [11](#)
- [118] HERNANDEZ-HERNANDEZ, A., RAY, P., LITOS, G., CIRO, M., OTTOLENGHI, S., BEUG, H. & BOYES, J. (2006). Acetylation and MAPK phosphorylation cooperate to regulate the degradation of active GATA-1. *The EMBO journal*, **25**, 3264–3274. [164, 239](#)
- [119] HERSHBERG, R. & PETROV, D. (2010). Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet*, **6**. [85](#)
- [120] HILDEBRAND, F., MEYER, A. & EYRE-WALKER, A. (2010). Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet*, **6**. [85](#)

- [121] HIRAI, K., ALIEV, G., NUNOMURA, A., FUJIOKA, H., RUSSELL, R.L., ATWOOD, C.S., JOHNSON, A.B., KRESS, Y., VINTERS, H.V., TABATON, M. *et al.* (2001). Mitochondrial abnormalities in Alzheimer's disease. *The Journal of Neuroscience*, **21**, 3017–3023. 11
- [122] HIRATSU, K., MOCHIZUKI, S. & KINASHI, H. (2000). Cloning and analysis of the replication origin and the telomeres of the large linear plasmid pSLA2-L in *Streptomyces rochei*. *Molecular and General Genetics MGG*, **263**, 1015–1021. 130
- [123] HOLT, L.J., TUCH, B.B., VILLÉN, J., JOHNSON, A.D., GYGI, S.P. & MORGAN, D.O. (2009). Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science*, **325**, 1682–1686. 8
- [124] HORN, H. & VOUSDEN, K. (2007). Coping with stress: multiple ways to activate p53. *Oncogene*, **26**, 1306–1316. 6
- [125] HORNBECK, P.V., KORNHAUSER, J.M., TKACHEV, S., ZHANG, B., SKRZYPEK, E., MURRAY, B., LATHAM, V. & SULLIVAN, M. (2011). PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic acids research*, gkr1122. 255
- [126] HORSPPOOL, R.N. (1980). Practical fast searching in strings. *Software: Practice and Experience*, **10**, 501–506. 38
- [127] HOSSAIN, M., AZIMI, N. & SKIENA, S. (2009). Crystallizing short-read assemblies around seeds. *BMC Bioinformatics*, **10**. 83
- [128] HUA, R. & WANGA, B. (2001). Statistically significant strings are related to regulatory elements in the promoter regions of *Saccharomyces cerevisiae*. *Physica A*, **290**, 464–474. 54
- [129] HUANG, K.Y., SU, M.G., KAO, H.J., HSIEH, Y.C., JHONG, J.H., CHENG, K.H., HUANG, H.D. & LEE, T.Y. (2016). dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins. *Nucleic acids research*, **44**, D435–D446. 255
- [130] HUFFMAN, D.A. *et al.* (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, **40**, 1098–1101. 119
- [131] IKEMURA, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *Journal of molecular biology*, **151**, 389–409. 151
- [132] IKEMURA, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular biology and evolution*, **2**, 13–34. 132, 151

- [133] JAIN, E., BAIROCH, A., DUVAUD, S., PHAN, I., REDASCHI, N., SUZEK, B.E., MARTIN, M.J., MCGARVEY, P. & GASTEIGER, E. (2009). Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC bioinformatics*, **10**, 1. [116](#)
- [134] JARRELL, K.F., DING, Y., MEYER, B.H., ALBERS, S.V., KAMINSKI, L. & EICHLER, J. (2014). N-linked glycosylation in Archaea: a structural, functional, and genetic analysis. *Microbiology and Molecular Biology Reviews*, **78**, 304–341. [6](#)
- [135] JENSEN, L.J., GUPTA, R., BLOM, N., DEVOS, D., TAMAMES, J., KESMIR, C., NIELSEN, H., STÆRFELDT, H.H., RAPACKI, K., WORKMAN, C. *et al.* (2002). Prediction of human protein function from post-translational modifications and localization features. *Journal of molecular biology*, **319**, 1257–1265. [9](#)
- [136] JI, Y., SHI, Y., DING, G. & LI, Y. (2011). A new strategy for better genome assembly from very short reads. *BMC Bioinformatics*, **12**. [83](#)
- [137] JI, Z., ZHOU, J., ZHU, Z. & CHEN, S. (2012). Self-configuration single particle optimizer for DNA sequence compression. *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, 1–8. [125](#)
- [138] JIMENEZ, G.S., KHAN, S.H., STOMMEL, J.M. & WAHL, G.M. (1999). p53 regulation by post-translational modification and nuclear retention in response to diverse stresses. *Oncogene*, **18**, 7656–7665. [5](#)
- [139] JIN, L., LI, C., XU, Y., WANG, L., LIU, J., WANG, D., HONG, C., JIANG, Z., MA, Y., CHEN, Q. *et al.* (2013). Epigallocatechin gallate promotes p53 accumulation and activity via the inhibition of MDM2-mediated p53 ubiquitination in human lung cancer cells. *Oncology reports*, **29**, 1983–1990. [5](#)
- [140] JOZA, N., OUDIT, G.Y., BROWN, D., BÉNIT, P., KASSIRI, Z., VAHSEN, N., BENOIT, L., PATEL, M.M., NOWIKOVSKY, K., VASSAULT, A. *et al.* (2005). Muscle-specific loss of apoptosis-inducing factor leads to mitochondrial dysfunction, skeletal muscle atrophy, and dilated cardiomyopathy. *Molecular and cellular biology*, **25**, 10261–10272. [175](#)
- [141] KANAYA, S., YAMADA, Y., KUDO, Y. & IKEMURA, T. (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, **238**, 143–155. [151](#)
- [142] KANE, L.A. & VAN EYK, J.E. (2009). Post-translational modifications of ATP synthase in the heart: biology and function. *Journal of bioenergetics and biomembranes*, **41**, 145–150. [11](#)

- [143] KANTOROVITZ, M.R., ROBINSON, G.E. & SINHA, S. (2007). A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics*, **23**, i249–i255. [57](#)
- [144] KARLIN, S., BURGE, C. & CAMPBELL, A. (1992). Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic acids research*, **20**, 1363–1370. [133](#)
- [145] KATOH, K. & TOH, H. (2010). Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics*, **26**, 1899–1900. [39](#)
- [146] KEMBEL, S.W., EISEN, J.A., POLLARD, K.S. & GREEN, J.L. (2011). The Phylogenetic Diversity of Metagenomes. *PLoS ONE*, **6**, e23214+. [83](#)
- [147] KENT, W.J. (2002). BLAT: the BLAST-like alignment tool. *Genome research*, **12**, 656–664. [39, 106](#)
- [148] KHOURY, G.A., BALIBAN, R.C. & FLOUDAS, C.A. (2011). Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Scientific reports*, **1**. [4, 17, 149, 205, 206, 243, 256](#)
- [149] KIM, I., WATADA, J., PEDRYCZ, W. *et al.* (2012). Pattern Clustering With Statistical Methods Using a DNA-Based Algorithm. *IEEE Transactions on NanoBioscience*, **11**, 100–110. [41](#)
- [150] KING, A.M. & MACRAE, T.H. (2015). Insect heat shock proteins during stress and diapause. *Annual review of entomology*, **60**, 59–75. [267](#)
- [151] KNOTT, A.B., PERKINS, G., SCHWARZENBACHER, R. & BOSSY-WETZEL, E. (2008). Mitochondrial fragmentation in neurodegeneration. *Nature Reviews Neuroscience*, **9**, 505–518. [xxii, 14](#)
- [152] KNUTH, D.E., MORRIS, J.J.H. & PRATT, V.R. (1977). Fast pattern matching in strings. *SIAM journal on computing*, **6**, 323–350. [38](#)
- [153] KOBAYASHI, I. (2001). Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Research*, **29**, 3742–3756. [131](#)
- [154] KOLDE, R. (2012). pheatmap: Pretty Heatmaps. R package version 0.6.1. <http://CRAN.R-project.org/package=pheatmap>. [91](#)
- [155] KOLDE, R. (2012). pheatmap: Pretty Heatmaps. R package version 0.7.7. [216, 260](#)
- [156] KOONIN, E. (1999). The emerging paradigm and open problems in comparative genomics. *Bioinformatics*, **15**, 265–266. [39](#)

- [157] KOURTIS, N., MOUBARAK, R.S., ARANDA-ORGILLES, B. *et al.* (2015). FBXW7 modulates cellular stress response and metastatic potential through HSF1 post-translational modification. *Nature cell biology*, **17**, 322–332. [17, 244](#)
- [158] KOZANITIS, C., SAUNDERS, C., KRUGLYAK, S. *et al.* (2011). Compressing genomic sequence fragments using SlimGene. *Journal of Computational Biology*, **18**, 401–413. [61](#)
- [159] KRAFFT, B. & PASQUET, A. (1991). Synchronized and rhythmical activity during the prey capture in the social spider *Anelosimus eximius* (Araneae, Theridiidae). *Insectes Sociaux*, **38**, 83–90. [23](#)
- [160] KRICK, T., VERSTRAETE, N., ALONSO, L.G. *et al.* (2014). Amino acid metabolism conflicts with protein diversity. *Molecular biology and evolution*, **31**, 2905–2912. [153](#)
- [161] KRUSE, J.P. & GU, W. (2009). Modes of p53 regulation. *Cell*, **137**, 609–622. [6](#)
- [162] KUNIN, V., COPELAND, A., LAPIDUS, A., MAVROMATIS, K. & HUGENHOLTZ, P. (2008). A bioinformatician's guide to metagenomics. *Microbiology and Molecular Biology Reviews*, **72**, 557–578. [82](#)
- [163] KURUPPU, S., BERESFORD-SMITH, B., CONWAY, T. & ZOBEL, J. (2012). Iterative dictionary construction for compression of large DNA data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, **9**, 137–149. [125](#)
- [164] LAMPREA-BURGUNDER E, L.P.M.P. (2011). Species-specific Typing of DNA Based on Palindrome Frequency Patterns. *DNA Res.*, **18**, 117–24. [84, 105](#)
- [165] LANDRY, C.R., LEVY, E.D. & MICHNICK, S.W. (2009). Weak functional constraints on phosphoproteomes. *Trends in genetics*, **25**, 193–197. [8](#)
- [166] LARKIN, M., BLACKSHIELDS, G., BROWN, N. *et al.* (2007). ClustalW and ClustalX Version 2. *Bioinformatics*, **23**, 2947–2948. [39](#)
- [167] LATA, C. & PRASAD, M. (2011). Role of DREBs in regulation of abiotic stress responses in plants. *Journal of experimental botany*, **62**, 4731–4748. [175](#)
- [168] LAVROV, D.V. (2007). Key transitions in animal evolution: a mitochondrial DNA perspective. *Integrative and Comparative Biology*, **47**, 734–743. [150, 209](#)
- [169] LAZARUS, M.B., NAM, Y., JIANG, J., SLIZ, P. & WALKER, S. (2011). Structure of human O-GlcNAc transferase and its complex with a peptide substrate. *Nature*, **469**, 564–567. [250](#)
- [170] LEACH, D.R., OKELY, E.A. & PINDER, D.J. (1997). Repair by recombination of DNA containing a palindromic sequence. *Molecular microbiology*, **26**, 597–606. [130](#)

- [171] LEE, H.Y., CHOI, C.S., BIRKENFELD, A.L., ALVES, T.C., JORNAYVAZ, F.R., JURCZAK, M.J., ZHANG, D., WOO, D.K., SHADEL, G.S., LADIGES, W. *et al.* (2010). Targeted expression of catalase to mitochondria prevents age-associated reductions in mitochondrial function and insulin resistance. *Cell metabolism*, **12**, 668–674. [173](#), [175](#)
- [172] LEE, T.Y., CHEN, S.A., HUNG, H.Y. & OU, Y.Y. (2011). Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. *PLoS One*, **6**, e17331. [9](#)
- [173] LEHTINEN, M.K., YUAN, Z., BOAG, P.R. *et al.* (2006). A conserved MST-FOXO signaling pathway mediates oxidative-stress responses and extends life span. *Cell*, **125**, 987–1001. [221](#)
- [174] LEIER, A., RICHTER, C., BANZHAF, W. & RAUHE, H. (2000). Cryptography with DNA binary strands. *Biosystems*, **57**, 13–22. [111](#)
- [175] LEONARD, J. & SCHAPIRA, A.H. (2000). Mitochondrial respiratory chain disorders I: mitochondrial DNA defects. *The Lancet*, **355**, 299–304. [13](#)
- [176] LEV-YADUN, S., DAFNI, A., FLAISHMAN, M.A., INBAR, M., IZHAKI, I., KATZIR, G. & NE’EMAN, G. (2004). Plant coloration undermines herbivorous insect camouflage. *BioEssays*, **26**, 1126–1130. [26](#)
- [177] LEVAV-COHEN, Y., GOLDBERG, Z., TAN, K.H. *et al.* (2014). The p53-Mdm2 Loop: A Critical Juncture of Stress Response. In *Mutant p53 and MDM2 in Cancer*, 161–186, Springer. [206](#)
- [178] LEVIN, L. (2003). The tale of one-way functions. *Problems of Information Transmission*, **39**, 92–103. [109](#)
- [179] LI, H. & HOMER, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, **11**, 473–483. [39](#)
- [180] LI, H.H., CAI, X., SHOUSE, G.P., PILUSO, L.G. & LIU, X. (2007). A specific PP2A regulatory subunit, B56 γ , mediates DNA damage-induced dephosphorylation of p53 at Thr55. *The EMBO journal*, **26**, 402–411. [6](#)
- [181] LI, J., JIA, J., LI, H., YU, J., SUN, H., HE, Y., LV, D., YANG, X., GLOCKER, M.O., MA, L. *et al.* (2014). SysPTM 2.0: an updated systematic resource for post-translational modification. *Database*, **2014**, bau025. [255](#)
- [182] LI, L.C., OKINO, S.T. & DAHIYA, R. (2004). DNA methylation in prostate cancer. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, **1704**, 87–102. [130](#)
- [183] LI, R., ZHU, H., RUAN, J., QIAN, W., FANG, X., SHI, Z., LI, Y., LI, S., SHAN, G., KRISTIANSEN, K. *et al.* (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome research*, **20**, 265–272. [83](#), [93](#)

- [184] LIAO, B.Y., CHANG, Y.J., HO, J.M. *et al.* (2004). The UniMarker (UM) method for synteny mapping of large genomes. *Bioinformatics*, **20**, 3156–3165. 39
- [185] LIGHTFIELD, J., FRAM, N. & ELY, B. (2011). Across Bacterial Phyla, Distantly-Related Genomes with Similar Genomic GC Content Have Similar Patterns of Amino Acid Usage. *PLoS ONE*, **6**, 85
- [186] LIN, J. (1991). Divergence Measures Based On the Shannon Entropy. *IEEE Transactions on Information Theory*, **37**, 145–151. 46, 48
- [187] LIPPERT, R., HUANG, H. & WATERMAN, M. (2002). Distributional Regimes For The Number of k-Word Matches Between Two Random Sequences. *Proc. Natl Acad Sci*, **99**, 13980–13989. 57
- [188] LIU, G., LIU, J. & ZHANG, B. (2012). Compositional Bias is a Major Determinant of the Distribution Pattern and Abundance of Palindromes in *Drosophila melanogaster*. *Journal of molecular evolution*, **75**, 130–140. 131
- [189] LIU, X., WAN, L., LI, J. *et al.* (2011). New powerful statistics for alignment-free sequence comparison under a pattern transfer model. *Journal of theoretical biology*, **284**, 106–116. 58, 59, 70
- [190] LIU, Z., MENG, J. & SUN, X. (2008). A novel feature-based method for whole genome phylogenetic analysis without alignment: application to HEV genotyping and subtyping. *Biochemical and biophysical research communications*, **368**, 223–230. 41, 42, 44, 70
- [191] LU, C.T., HUANG, K.Y., SU, M.G. *et al.* (2012). DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. *Nucleic acids research*, gks1229. 9, 220
- [192] LU, G., ZHANG, S. & FANG, X. (2008). An Improved String Composition Method For Sequence Comparison. *BMC Bioinformatics*, **9**, S15. 41, 51, 52, 56, 70
- [193] LUKIĆ-BILELA, L., BRANDT, D., POJSKIĆ, N., WIENS, M., GAMULIN, V. & MÜLLER, W.E. (2008). Mitochondrial genome of *Suberites domuncula*: palindromes and inverted repeats are abundant in non-coding regions. *Gene*, **412**, 1–11. 77, 79
- [194] MAGRANE, M., CONSORTIUM, U. *et al.* (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database*, **2011**, bar009. 8
- [195] MAISONNEUVE, E., DUCRET, A., KHOUEIRY, P., LIGNON, S., LONGHI, S., TALLA, E. & DUKAN, S. (2009). Rules governing selective protein carbonylation. *PLoS One*, **4**, e7269. 172, 173, 179, 180

- [196] MANN, H.B. (1942). The Construction of Orthogonal Latin Squares. *The Annals of Mathematical Statistics*, **13**, 418–423. [115](#)
- [197] MARTIN, L., LATYPOVA, X. & TERRO, F. (2011). Post-translational modifications of tau protein: implications for Alzheimer's disease. *Neurochemistry international*, **58**, 458–471. [255](#)
- [198] MARZETTI, E., HWANG, J.C., LEES, H.A., WOHLGEMUTH, S.E., DUPONT-VERSTEEGDEN, E.E., CARTER, C.S., BERNABEI, R. & LEEUWENBURGH, C. (2010). Mitochondrial death effectors: relevance to sarcopenia and disuse muscle atrophy. *Biochimica et Biophysica Acta (BBA)-General Subjects*, **1800**, 235–244. [175](#)
- [199] McMENEMY, L.S., HARTLEY, S.E., MACFARLANE, S.A., KARLEY, A.J., SHEPHERD, T. & JOHNSON, S.N. (2012). Raspberry viruses manipulate the behaviour of their insect vectors. *Entomologia Experimentalis et Applicata*, **144**, 56–68. [25](#)
- [200] MINGUEZ, P., PARCA, L., DIELLA, F., MENDE, D.R., KUMAR, R., HELMER-CITTERICH, M., GAVIN, A.C., VAN NOORT, V. & BORK, P. (2012). Deciphering a global network of functionally associated post-translational modifications. *Molecular systems biology*, **8**. [10](#)
- [201] MINGUEZ, P., LETUNIC, I., PARCA, L. & BORK, P. (2013). PTMcode: a database of known and predicted functional associations between post-translational modifications in proteins. *Nucleic acids research*, **41**, D306–D311. [271](#)
- [202] MINOSSE, C., CALCATELLA, S., ABBATE, I., SELLERI, M., ZANIRATTI, M.S. & CAPOBIANCHI, M.R. (2006). Possible compartmentalization of hepatitis C viral replication in the genital tract of HIV-1-coinfected women. *J Infect Dis.*, **194**, 1529–1536. [123](#)
- [203] MOHAMMED, M.H., DUTTA, A., BOSE, T. *et al.* (2012). DELIMINATE—A fast and efficient method for loss-less compression of genomic sequences. *Bioinformatics*. [61](#)
- [204] MORENO-LETELIER, A., OLMEDO, G., EGUIARTE, L.E. *et al.* (2011). Parallel evolution and horizontal gene transfer of the pst operon in Firmicutes from oligotrophic environments. *International journal of evolutionary biology*, **2011**. [66](#)
- [205] MORIYAMA, E.N. & POWELL, J.R. (1997). Codon usage bias and tRNA abundance in *Drosophila*. *Journal of molecular evolution*, **45**, 514–523. [132](#)
- [206] NAKAMURA, Y., GOJOBORI, T. & IKEMURA, T. (2000). Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic acids research*, **28**, 292–292. [113, 119, 156](#)

- [207] NAVARRO, G. (2001). A Guided Tour To Approximate String Matching. *J. ACM Comp. Surveys (CSUR)*, **33**, 31–88. [39](#)
- [208] NEEDLEMAN, S. & WUNSCH, C. (1970). A General Method Applicable To The Search For Similarities In The Amino Acid Sequence of Two Proteins. *J Mol Biol.*, **48**, 443–453. [39](#)
- [209] NIKAWA, T., ISHIDOH, K., HIRASAKA, K., ISHIHARA, I., IKEMOTO, M., KANO, M., KOMINAMI, E., NONAKA, I., OGAWA, T., ADAMS, G.R. *et al.* (2004). Skeletal muscle gene expression in space-flown rats. *The FASEB journal*, **18**, 522–524. [2](#), [172](#), [180](#), [198](#)
- [210] NIKLAS, K., BONDOS, S., DUNKER, A. & NEWMAN, S. (2015). Rethinking Gene Regulatory Networks in Light of Alternative Splicing, Post-Translational Modifications, and Intrinsically Disordered Protein Domains. *Name: Frontiers in Cell and Developmental Biology*, **3**. [10](#)
- [211] NOTHAFT, H. & SZYMANSKI, C.M. (2010). Protein glycosylation in bacteria: sweeter than ever. *Nature Reviews Microbiology*, **8**, 765–778. [218](#)
- [212] NYSTRÖM, T. (2005). Role of oxidative carbonylation in protein quality control and senescence. *EMBO J.*, **24**, 1311–1317. [179](#)
- [213] ODA, K., KOHCHI, T. & OHYAMA, K. (1992). Mitochondrial DNA of Marchantia polymorpha as a single circular form with no incorporation of foreign DNA. *Bioscience, biotechnology, and biochemistry*, **56**, 132. [133](#)
- [214] OROBITG, M., CORES, F., GUIRADO, F. *et al.* (2012). Enhancing the scalability of consistency-based progressive multiple sequences alignment applications. In *Parallel & Distributed Processing Symposium (IPDPS), 2012 IEEE 26th International*, 71–82, IEEE. [39](#)
- [215] OTU, H.H. & SAYOOD, K. (2003). A new sequence distance measure for phylogenetic tree construction. *Bioinformatics*, **19**, 2122–2130. [41](#), [60](#), [69](#), [70](#)
- [216] PARAKH, A. (2013). A probabilistic quantum key transfer protocol. *Security and Communication Networks*. [110](#)
- [217] PARAKH, A. & KAK, S. (2009). Online data storage using implicit security. *Information Sciences*, **179**, 3323–3331. [110](#)
- [218] PARK, C.K., JOSHI, H.K., AGRAWAL, A., GHARE, M.I., LITTLE, E.J., DUNTEM, P.W., BITINAITE, J. & HORTON, N.C. (2010). Domain swapping in allosteric modulation of DNA specificity. *PLoS Biol*, **8**, e1000554. [74](#), [130](#)
- [219] PASZKIEWICZ, K. & STUDHOLME, D.J. (2010). De novo assembly of short sequence reads. *Briefings in bioinformatics*, **11**, 457–472. [93](#)

- [220] PATIL, A., CHAN, C.T., DYAVAIYAH, M., ROONEY, J.P., DEDON, P.C. & BEGLEY, T.J. (2012). Translational infidelity-induced protein stress results from a deficiency in Trm9-catalyzed tRNA modifications. *RNA biology*, **9**, 990–1001. [132](#)
- [221] PEDERSEN, J., BASTOLA, D., DICK, K., GANDHI, R. & MAHONEY, W. (2012). Blast your way through malware malware analysis assisted by bioinformatics tools. In *Proceedings of the International Conference on Security and Management (SAM)*, 1, The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp). [115](#)
- [222] PEJAVER, V., HSU, W.L., XIN, F. *et al.* (2014). The structural and functional signatures of proteins that undergo multiple events of post-translational modification. *Protein Science*, **23**, 1077–1093. [17](#)
- [223] PENFIELD, S. (2008). Temperature perception and signal transduction in plants. *New Phytologist*, **179**, 615–628. [206](#)
- [224] PENG, L., YUAN, Z., LI, Y. *et al.* (2015). Ubiquitinated Sirtuin 1 (SIRT1) Function Is Modulated during DNA Damage-induced Cell Death and Survival. *Journal of Biological Chemistry*, **290**, 8904–8912. [206](#)
- [225] PENG, Q. & SMITH, A. (2011). Multiple sequence assembly from reads alignable to a common reference genome. *IEEE/ACM Trans Comput Biol Bioinform.*, **8**. [84](#)
- [226] PERONA, J. (2002). Type II restriction endonucleases. *Methods*, **28**, 353–364. [74](#)
- [227] PEUGET, S., BONACCI, T., SOUBEYRAN, P. *et al.* (2014). Oxidative stress-induced p53 activity is enhanced by a redox-sensitive TP53INP1 SUMOylation. *Cell Death & Differentiation*, **21**, 1107–1118. [206](#)
- [228] PHILLIPS, D., APONTE, A.M., COVIAN, R., NEUFELD, E., YU, Z.X. & BALABAN, R.S. (2011). Homogenous protein programming in the mammalian left and right ventricle free walls. *Physiological genomics*, **43**, 1198–1206. [8](#)
- [229] PHILLIPS, D., COVIAN, R., APONTE, A.M., GLANCY, B., TAYLOR, J.F., CHESS, D. & BALABAN, R.S. (2012). Regulation of oxidative phosphorylation complex activity: effects of tissue-specific metabolic stress within an allometric series and acute changes in workload. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, **302**, R1034–R1048. [8](#)
- [230] PHILPOTT, D., POPOVA, I., KATO, K., STEVENSON, J., MIQUEL, J. & SAPP, W. (1990). Morphological and biochemical examination of Cosmos 1887 rat heart tissue: Part I—Ultrastructure. *The FASEB Journal*, **4**, 73–78. [171](#)

- [231] PINHO, A., PRATAS, D. & GARCIA, S. (2012). GReEn: a tool for efficient compression of genome resequencing data. *Nucleic Acids Research*, **40**, e27–e27. 125
- [232] PIROVANO, W. & HERINGA, J. (2010). Protein secondary structure prediction. In *Data Mining Techniques for the Life Sciences*, 327–348, Springer. 191
- [233] PLOTKIN, J.B. & KUDLA, G. (2010). Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics*, **12**, 32–42. 131
- [234] PRABHU, D. & ADIMOOLAM, M. (2011). Bi-serial DNA Encryption Algorithm (BDEA). *arXiv preprint arXiv:1101.2577*. 111
- [235] PRASAD, A.B., ALLARD, M.W. & GREEN, E.D. (2008). Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Molecular Biology and Evolution*, **25**, 1795–1808. 49
- [236] PRESMAN, E. (1986). Approximation in variation of the distribution of a sum of independent Bernoulli variables with a Poisson law. *Theory of Probability & Its Applications*, **30**, 417–422. 233
- [237] QI, J., WANG, B. & HAO, B.I. (2004). Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *Journal of molecular evolution*, **58**, 1–11. 50, 55
- [238] RADIVOJAC, P., CLARK, W.T., ORON, T.R., SCHNOES, A.M., WITTKOP, T., SOKOLOV, A., GRAIM, K., FUNK, C., VERSPOOR, K., BEN-HUR, A. *et al.* (2013). A large-scale evaluation of computational protein function prediction. *Nature methods*, **10**, 221–227. 9
- [239] RÄDLER, J., KOLTOVER, I., SALDITT, T. & SAFINYA, C. (1997). Structure of DNA -cationic liposome complexes: DNA intercalation in multilamellar membranes in distinct interhelical packing regimes. *Science*, **275**, 810–814. 123
- [240] RECHSTEINER, M. & ROGERS, S.W. (1996). PEST sequences and regulation by proteolysis. *Trends in biochemical sciences*, **21**, 267–271. 174
- [241] REGOES, A., ZOURMPANOU, D., LEÓN-AVILA, G., VAN DER GIEZEN, M., TOVAR, J. & HEHL, A.B. (2005). Protein import, replication, and inheritance of a vestigial mitochondrion. *Journal of Biological Chemistry*, **280**, 30557–30563. 12
- [242] REINERT, G., CHEW, D., SUN, F. *et al.* (2009). Alignment-Free Sequence Comparison (I): Statistics and Power. *J Comput Biol.*, **16**, 1615–1634. 57, 70
- [243] RICE, P., LONGDEN, I., BLEASBY, A. *et al.* (2000). EMBOSS: the European molecular biology open software suite. *Trends in genetics*, **16**, 276–277. 124, 191

- [244] RICHMOND, T. & DAVEY, C. (2003). The structure of DNA in the nucleosome core. *Nature*, **423**, 145–150. [123](#)
- [245] RICHTER, D., OTT, F., AUCH, A., SCHMID, R. & HUSON, D. (2008). MetaSim: A Sequencing Simulator for Genomics and Metagenomics. *PLoS ONE*, **3**. [85](#)
- [246] RINALDI, F., CLEMATIDE, S., MARQUES, H., ELLENDORFF, T., ROMACKER, M. & RODRIGUEZ-ESTEBAN, R. (2014). OntoGene web services for biomedical text mining. *BMC bioinformatics*, **15**, S6. [245](#)
- [247] RIVEST, R.L., SHAMIR, A. & ADLEMAN, L. (1978). A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*, **21**, 120–126. [109](#)
- [248] ROBERTS, R.J., VINCZE, T., POSFAI, J. & MACELIS, D. (2010). REBASE: a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic acids research*, **38**, D234–D236. [87](#), [135](#)
- [249] ROCHA, E.P., VIARI, A. & DANCHIN, A. (1998). Oligonucleotide bias in *Bacillus subtilis*: general trends and taxonomic comparisons. *Nucleic acids research*, **26**, 2971–2980. [130](#)
- [250] ROGERS, S., WELLS, R. & RECHSTEINER, M. (1986). Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. *Science*, **234**, 364–368. [175](#)
- [251] ROHNER, C. & WARD, D. (1997). Chemical and mechanical defense against herbivory in two sympatric species of desert Acacia. *Journal of Vegetation Science*, **8**, 717–726. [26](#)
- [252] ROSENQUIST, M., SEHNKE, P., FERL, R.J. *et al.* (2000). Evolution of the 14-3-3 protein family: does the large number of isoforms in multicellular organisms reflect functional specificity? *Journal of Molecular Evolution*, **51**, 446–458. [237](#)
- [253] ROY, B., RAKSHIT, G., SINGHA, P., MAJUMDER, A. & DATTA, D. (2011). An Improved Symmetric Key Cryptography with DNA Based Strong Cipher. In *Devices and Communications (ICDeCom), 2011 International Conference on*, 1–5, IEEE. [111](#)
- [254] SAKUMA, Y., MARUYAMA, K., OSAKABE, Y., QIN, F., SEKI, M., SHINOZAKI, K. & YAMAGUCHI-SHINOZAKI, K. (2006). Functional analysis of an *Arabidopsis* transcription factor, DREB2A, involved in drought-responsive gene expression. *The Plant Cell Online*, **18**, 1292–1309. [175](#)
- [255] SALMERÓN, A., JANZEN, J., SONEJI, Y., BUMP, N., KAMENS, J., ALLEN, H. & LEY, S.C. (2001). Direct phosphorylation of NF- κ B1 p105 by the I κ B kinase complex on serine 927 is essential for signal-induced p105 proteolysis. *Journal of Biological Chemistry*, **276**, 22215–22222. [174](#)

- [256] SANKARANARAYANAN, R., DOCK-BREGEON, A.C., ROMBY, P. *et al.* (1999). The structure of threonyl-tRNA synthetase-tRNA Thr complex enlightens its repressor activity and reveals an essential zinc ion in the active site. *Cell*, **97**, 371–381. [218](#)
- [257] SCHADT, E.E., LINDERMAN, M.D., SORENSEN, J. *et al.* (2010). Computational solutions to large-scale data management and analysis. *Nature Reviews Genetics*, **11**, 647–657. [39](#)
- [258] SCHAYEK, H., BENTOV, I., JACOB-HIRSCH, J., YEUNG, C., KHANNA, C., HELMAN, L., PLYMATE, S. & WERNER, H. (2012). Global methylation analysis identifies PITX2 as an upstream regulator of the androgen receptor and IGF-I receptor genes in prostate cancer. *Hormone Metab Res*, **44**, 511–519. [130](#)
- [259] SCHMUTZ, J., CANNON, S.B., SCHLUETER, J., MA, J., MITROS, T., NELSON, W., HYTEN, D.L., SONG, Q., THELEN, J.J., CHENG, J. *et al.* (2010). Genome sequence of the palaeopolyploid soybean. *nature*, **463**, 178–183. [93](#)
- [260] SCHNEIER, B. *et al.* (1999). *The twofish encryption algorithm: a 128-bit block cipher*. New York: J. Wiley. [110](#)
- [261] SCHULT, D.A. & SWART, P. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conferences (SciPy 2008)*, vol. 2008, 11–16. [216, 248](#)
- [262] SCHULT, D.A. & SWART, P. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conferences (SciPy 2008)*, vol. 2008, 11–16. [259](#)
- [263] SCHUMACHER, B., SKWARCZYNSKA, M., ROSE, R. & OTTMANN, C. (2010). Structure of a 14-3-3 σ -YAP phosphopeptide complex at 1.15 Å resolution. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, **66**, 978–984. [221](#)
- [264] SCHUSTER, S. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, **5**, 16–18. [83](#)
- [265] SERRANO-GOMEZ, S.J., MAZIVEYI, M. & ALAHARI, S.K. (2016). Regulation of epithelial-mesenchymal transition through epigenetic and post-translational modifications. *Molecular cancer*, **15**, 1. [17](#)
- [266] SHAH, S.P., LONIAL, S. & BOISE, L.H. (2015). When Cancer Fights Back: Multiple Myeloma, Proteasome Inhibition, and the Heat Shock Response. *Molecular Cancer Research*, molcanres–0135. [206](#)
- [267] SHANNON, C. (1948). The Mathematical Theory of Communication. *Bell Systems Tech. J.*, **27**, 379–423, 623–656. [41, 110, 123](#)

- [268] SHAO, C.H., CAPEK, H.L., PATEL, K.P., WANG, M., TANG, K., DESOUZA, C., NAGAI, R., MAYHAN, W., PERIASAMY, M. & BIDASEE, K.R. (2011). Carbonylation contributes to SERCA2a activity loss and diastolic dysfunction in a rat model of type 1 diabetes. *Diabetes*, **60**, 947–959. [221](#), [254](#)
- [269] SHAO, C.H., TIAN, C., OUYANG, S., MOORE, C.J., ALOMAR, F., NEMET, I., D'SOUZA, A., NAGAI, R., KUTTY, S., ROZANSKI, G.J. *et al.* (2012). Carbonylation induces heterogeneity in cardiac ryanodine receptor function in diabetes mellitus. *Molecular pharmacology*, **82**, 383–399. [221](#)
- [270] SHARP, P.M., BAILES, E., GROCOCK, R.J., PEDEN, J.F. & SOCKETT, R.E. (2005). Variation in the strength of selected codon usage bias among bacteria. *Nucleic acids research*, **33**, 1141–1153. [152](#)
- [271] SÓTI, C. & CSERMELY, P. (2007). Protein stress and stress proteins: implications in aging and disease. *Journal of biosciences*, **32**, 511–515. [149](#), [206](#)
- [272] SIERACKI, N.A. & KOMAROVA, Y.A. (2013). *Studying Cell Signal Transduction with Biomimetic Point Mutations*. INTECH Open Access Publisher. [162](#)
- [273] SILVA, I., MOODY, B., BEHAR, J., JOHNSON, A., OSTER, J., CLIFFORD, G.D. & MOODY, G.B. (2015). Robust detection of heart beats in multimodal data. *Physiological measurement*, **36**, 1629. [22](#)
- [274] SIMPSON, J.T., WONG, K., JACKMAN, S.D., SCHEIN, J.E., JONES, S.J. & BIROL, İ. (2009). ABYSS: a parallel assembler for short read sequence data. *Genome research*, **19**, 1117–1123. [93](#)
- [275] SIMS, G., JUN, G.W.S. & KIM, S. (2008). Alignment-Free Genome Comparison With Feature Frequency Profiles (FFP) and Optimal Resolutions. *Proc Natl Acad Sci*, **106**, 2677–2682. [44](#), [46](#), [48](#), [49](#), [50](#), [70](#)
- [276] SINGER, G.A. & HICKEY, D.A. (2000). Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Molecular Biology and Evolution*, **17**, 1581–1588. [152](#)
- [277] SINGER, G.A. & HICKEY, D.A. (2003). Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene*, **317**, 39–47. [152](#)
- [278] S.MANTACI, RESTIVO, A. & SCIORTINO, M. (2008). Distance Measures For Biological Sequences: Some Recent Approaches. *Internat. J. Approx. Reason*, **47**, 109–124. [41](#)
- [279] SMITH, T. & WATERMAN, M. (1981). Identification of Common Molecular Subsequences. *J Mol Biol.*, **147**, 195–197. [39](#)

- [280] SOARES, I., GOIOS, A. & AMORIM, A. (2012). Sequence Comparison Alignment-Free Approach Based on Suffix Tree and L-Words Frequency. *The Scientific World Journal*, **2012**. [49](#), [70](#)
- [281] SOMMERS, J.A., SUHASINI, A.N. & BROSH, R.M. (2015). Protein Degradation Pathways Regulate the Functions of Helicases in the DNA Damage Response and Maintenance of Genomic Stability. *Biomolecules*, **5**, 590–616. [206](#)
- [282] SONG, K., REN, J., ZHAI, Z. *et al.* (2012). Alignment-free sequence comparison based on next generation sequencing reads. In *Research in Computational Molecular Biology*, 272–285, Springer. [58](#)
- [283] SPÄTH, G.F., DRINI, S. & RACHIDI, N. (2015). A touch of Zen: post-translational regulation of the Leishmania stress response. *Cellular microbiology*, **17**, 632–638. [206](#)
- [284] STÖLLBERGER, C. & FINSTERER, J. (2006). Autonomic dysfunction in left ventricular hypertrabeculation/noncompaction. *International journal of cardiology*, **109**, 286–287. [13](#)
- [285] SWERDLOW, R.H., BURNS, J.M. & KHAN, S.M. (2014). The Alzheimer's disease mitochondrial cascade hypothesis: progress and perspectives. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, **1842**, 1219–1231. [244](#)
- [286] SYVANEN, M. (2012). Evolutionary Implications of Horizontal Gene Transfer. *Annual review of genetics*. [66](#)
- [287] TANENBAUM, A.S. (1996). Computer Networks, Ch. 5. [110](#)
- [288] TEAM, R.D.C. (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. [91](#)
- [289] TEMPLE, A., YEN, T.Y. & GRONERT, S. (2006). Identification of specific protein carbonylation sites in model oxidations of human serum albumin. *Journal of the American Society for Mass Spectrometry*, **17**, 1172–1180. [179](#)
- [290] TOYOSHIMA, C. & NOMURA, H. (2002). Structural changes in the calcium pump accompanying the dissociation of calcium. *Nature*, **418**, 605–611. [221](#)
- [291] UEMATSU, S., GOTO, Y., SUZUKI, T., SASAZAWA, Y., DOHMAE, N. & SIMIZU, S. (2014). N-Glycosylation of extracellular matrix protein 1 (ECM1) regulates its secretion, which is unrelated to lipid proteinosis. *FEBS open bio*, **4**, 879–885. [6](#)
- [292] UKKONEN, E. (1985). Finding Approximate Patterns In Strings. *J. Algor.*, **6**, 132–137. [39](#), [41](#)

- [293] ULITSKY, I., BURSTEIN, D., TULLER, T. *et al.* (2006). The Average Common Substring Approach to Phylogenomic Reconstruction. *Jour. Of Comp. Bio.*, **13**, 336–350. [41](#), [62](#), [70](#)
- [294] VALTONEN, A. (2015). Teddy bears. *Consumption Markets & Culture*, 1–6. [25](#)
- [295] VAN SOEST, P.J. *et al.* (1982). *Nutritional ecology of the ruminant. Ruminant metabolism, nutritional strategies, the cellulolytic fermentation and the chemistry of forages and plant fibers..* O & B Books, Inc. [26](#)
- [296] VINGA, S. & ALMEIDA, J. (2003). Alignment-free sequence comparison - a review. *Bioinformatics*, **19**, 513–523. [41](#), [213](#)
- [297] VUCIC, D., DIXIT, V.M. & WERTZ, I.E. (2011). Ubiquitylation in apoptosis: a post-translational modification at the edge of life and death. *Nature reviews Molecular cell biology*, **12**, 439–452. [244](#)
- [298] WAKEFIELD, M.J., KEOHANE, A.M., TURNER, B.M. & GRAVES, J.A.M. (1997). Histone underacetylation is an ancient component of mammalian X chromosome inactivation. *Proceedings of the National Academy of Sciences*, **94**, 9665–9668. [258](#)
- [299] WALD, N., ALROY, M., BOTZMAN, M. & MARGALIT, H. (2012). Codon usage bias in prokaryotic pyrimidine-ending codons is associated with the degeneracy of the encoded amino acids. *Nucleic Acids Research*, **40**, 7074–7083. [132](#)
- [300] WALLACE, D.C. (1999). Mitochondrial diseases in man and mouse. *Science*, **283**, 1482–1488. [175](#)
- [301] WALLACE, D.C. *et al.* (1997). Mitochondrial DNA in aging and disease. *Scientific American*, **277**, 40–59. [175](#)
- [302] WAN, L., REINERT, G., SUN, F. *et al.* (2010). Alignment-free sequence comparison (II): theoretical power of comparison statistics. *Journal of Computational Biology*, **17**, 1467–1490. [57](#)
- [303] WANG, J., ZHOU, X., ZHU, J. *et al.* (2012). GO-function: deriving biologically relevant functions from statistically significant functions. *Briefings in bioinformatics*, **13**, 216–227. [232](#)
- [304] WANG, J.D. (2011). A comparison study of virus classification by genome sequences. In *Bioinformatics and Bioengineering (BIBE), 2011 IEEE 11th International Conference on*, 270–273, IEEE. [41](#)
- [305] WANG, Y., ZENG, X., IYER, N.J., BRYANT, D.W., MOCKLER, T.C. & MAHALINGAM, R. (2012). Exploring the switchgrass transcriptome using second-generation sequencing technology. *PloS One*, **7**, e34225. [83](#)

- [306] WASZCZAK, C., AKTER, S., JACQUES, S. *et al.* (2015). Oxidative post-translational modifications of cysteine residues in plant signal transduction. *Journal of experimental botany*, **66**, 2923–2934. [244](#)
- [307] WATERMAN, M. (1995). *Introduction to Computational Biology: Maps, Sequences and Genomes*. Chapman and Hall. [57](#)
- [308] WEISS, O., JIMENEZ-MONTANO, M.A. & HERZEL, H. (2000). Information content of protein sequences. *Journal of theoretical biology*, **206**, 379–386. [153](#)
- [309] WHEELER, D., BARRETT, T., BENSON, D., BRYANT, S., CANESE, K., CHETVERNIN, V., CHURCH, D., DiCUCCIO, M., EDGAR, R., FEDERHEN, S. *et al.* (2007). Database resources of the national center for biotechnology information. *Nucleic acids research*, **35**, D5–D12. [134](#), [153](#), [210](#)
- [310] WITTEN, I.H. (2005). Text mining: Practical Handbook of Internet Computing. [244](#)
- [311] WOOLEY, J. (1999). Trends In Computational Biology: A Summary Based On A RECOMB Plenary Lecture. *J. Comput. Biol.*, **6**, 459–474. [39](#)
- [312] WRIGHT, P.E. & DYSON, H.J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology*, **16**, 18–29. [206](#)
- [313] WU, T.J., HUANG, Y.H. & LI, L.A. (2005). Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences. *Bioinformatics*, **21**, 4125–4132. [42](#), [45](#), [69](#)
- [314] WU, X., WAN, X., WU, G. *et al.* (2006). Phylogenetic Analysis Using Complete Signature Information of Whole Genomes And Clustered Neighbour-Joining Method. *Int J Bioinform Res Appl.*, **2**, 219–248. [50](#)
- [315] WU, Y.T., WU, S.B., LEE, W.Y. & WEI, Y.H. (2011). *The Cross-Talk Between Mitochondria and the Nucleus in the Response to Oxidative Stress Associated with Mitochondrial Dysfunction in Mitochondrial Encephalomyopathies*. INTECH Open Access Publisher. [13](#)
- [316] WU, Y.W. & YE, Y. (2010). A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. In *Research in Computational Molecular Biology*, 535–549, Springer. [82](#)
- [317] XIA, Q. & QIU, X. (2012). Sequence and Structure Dependent DNA-DNA Interactions. *Biophysical Journal*, **102**, 636. [123](#)
- [318] XIE, Z., DAI, J., DAI, L., TAN, M., CHENG, Z., WU, Y., BOEKE, J.D. & ZHAO, Y. (2012). Lysine succinylation and lysine malonylation in histones. *Molecular & Cellular Proteomics*, **11**, 100–107. [164](#)

- [319] YAFFE, M.B. & ELIA, A.E. (2001). Phosphoserine/threonine-binding domains. *Current opinion in cell biology*, **13**, 131–138. [163](#)
- [320] YIN, F. & CADENAS, E. (2015). Mitochondria: the cellular hub of the dynamic coordinated network. *Antioxidants & redox signaling*, **22**, 961–964. [16](#)
- [321] YU, C., LIU, Z., MCKENNA, T., REISNER, A.T. & REIFMAN, J. (2006). A method for automatic identification of reliable heart rates calculated from ECG and PPG waveforms. *Journal of the American Medical Informatics Association*, **13**, 309–320. [xxii, 21, 22](#)
- [322] YU, Z.G., ZHOU, L.Q., ANH, V.V. *et al.* (2005). Phylogeny of prokaryotes and chloroplasts revealed by a simple composition approach on all protein sequences from complete genomes without sequence alignment. *Journal of molecular evolution*, **60**, 538–545. [55](#)
- [323] ZAND, R., LI, M.X., JIN, X. *et al.* (1998). Determination of the sites of posttranslational modifications in the charge isomers of bovine myelin basic protein by capillary electrophoresis-mass spectroscopy. *Biochemistry*, **37**, 2441–2449. [216](#)
- [324] ZERBINO, D., MC EWEN, G., MARGULIES, E. & BIRNEY, E. (2009). Pebble and Rock Band: Heuristic Resolution of Repeats and Scaffolding in the Velvet Short-Read de Novo Assembler. *PLoS One.*, **4**. [83](#)
- [325] ZERBINO, D.R. & BIRNEY, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829. [93](#)
- [326] ZHANG, J., SPRUNG, R., PEI, J., TAN, X., KIM, S., ZHU, H., LIU, C.F., GRISHIN, N.V. & ZHAO, Y. (2009). Lysine acetylation is a highly abundant and evolutionarily conserved modification in Escherichia coli. *Molecular & Cellular Proteomics*, **8**, 215–225. [149](#)
- [327] ZHANG, L., MENG, J., LIU, H. & HUANG, Y. (2011). Clustering DNA methylation expressions using nonparametric beta mixture model. In *Genomic Signal Processing and Statistics (GENSIPS), 2011 IEEE International Workshop on*, 170–173, IEEE. [41](#)
- [328] ZHANG, W., CHEN, J., YANG, Y., TANG, Y., SHANG, J. & SHEN, B. (2011). A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies. *PLoS One*, **6**. [83](#)
- [329] ZHANG, Z., TAN, M., XIE, Z., DAI, L., CHEN, Y. & ZHAO, Y. (2011). Identification of lysine succinylation as a new post-translational modification. *Nature chemical biology*, **7**, 58–63. [164](#)
- [330] ZHU, Y., LI, T., LI, D., ZHANG, Y., XIONG, W., SUN, J., TANG, Z. & CHEN, G. (2012). Using predicted shape string to enhance the accuracy of γ -turn prediction. *Amino acids*, **42**, 1749–1755. [191](#)

- [331] ZIV, J. & LEMPEL, A. (1977). A Universal Algorithm For Sequential Data Compression. *IEEE Transactions on Information Theory*, **23**, 337–343. [60](#), [61](#)
- [332] ZÖRB, C., SCHMITT, S. & MÜHLING, K.H. (2010). Proteomic changes in maize roots after short-term adjustment to saline growth conditions. *Proteomics*, **10**, 4441–4449. [7](#)