UNIVERSITY OF CALIFORNIA, SAN DIEGO

Exploration of N-terminal methionine excision via comparative proteogenomics

A thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Computer Science

by

Stefano Romoli Bonissone

Committee in charge:

Professor Pavel Pevzner, Chair
Professor Vineet Bafna
Professor Alexander Hoffmann

2010

The Thesis of Stefano Romoli Bonissone is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2010

DEDICATION

To my parents, family and friends.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Pavel Pevzner who helped guide me through the course of this project, always with insightful comments and ideas. Nitin Gupta is also to thank for a great deal of help, particularly earlier on in this project, as well as providing for additional guidance throughout. Margaret Romine is to be thanked for providing us with high-quality *Shewanella* data, as are Eugene Koonin, Yuri Wolf and Andrei Osterman who all contributed ideas, critiques and comments at various stages. I would also like to thank my committee members Vineet Bafna and Alexander Hoffmann.

Chapters 3,4 and 5 are in part also currently being prepared for submission for publication of the material. Bonissone, Stefano; Gupta, Nitin; Romine, Margaret; Pevzner, Pavel. The thesis author was the primary investigator and author of this material.

ABSTRACT OF THE THESIS

Exploration of N-terminal methionine excision via comparative proteogenomics

by

Stefano Romoli Bonissone
Master of Science in Computer Science

University of California San Diego, 2010

Professor Pavel Pevzner, Chair

This thesis will explore N-terminal methionine excision (NME), a co-translational process that occurs on virtually all proteins in all organimsms from bacteria to eukaryotes. NME is an essential process for the funtioning of proteins, with the enzymes responsible for carrying out NME belonging to the minimal bacterial cell [HPG+99]. This excision is almost exclusively determined by the second amino acid of the nascient protein. The set of seven amino acids that prompt cleavage also correspond to the seven stabilizing amino acids that Arfin and Bradshaw [AB88] connected to protein degradation. Comparative proteogenomics approaches are employed to perform an analysis of NME based on sequence as well as other available data. The ties of NME to protein degradation are questioned, with the results of experiments suggesting that only two of the seven amino acids that prompt cleavage are important and required for post-translational modifications, specifically N-acetylation.

Chapter 1

N-Terminal Methionine Excision

1.1 N-Terminal Methionine Excision

N-Terminal methionine excision (NME) is the co-translational removal of the initiating methionine residue of proteins. Although methionine is the initial amino acid for essentially all proteins, in many this first methionine is removed. Approximately two-thirds of all proteins are subject to NME and NME machinery appears in all organisms from bacteria to eukaryotes [GBM04]. There has been considerable work performed studying NME, most of which focus on the identity of the second N-terminal residue [AB88, FMP^{+}06, GBM04]. The focus of the literature is on the second amino acid since the methionine aminopeptidase (MAP) enzymes, which cleave the N-terminal methionine, appear only to utilize the second residue to determine cleavage [RM93].

Eubacteria have present an N-formyl group on their initial methionine residue. This formyl group must be removed prior to the MAP enzyme acting on the protein. The peptide deformylase (PDF) enzyme removes this formyl group and exposes the N-terminal methionine for the MAP enzyme. Giglione and Meinnel [GM01], Giglione et. al. [GBM04] discuss the role and evolution of peptide deformylase enzymes in addition to the MAP enzymes.

There have been many suggestions that NME relates to protein degradation and the N-end rule [Var96]. Giglione et. al. [GVM03] provide an example of a protein, D2 in *C.reinhardtii*, of the PSII complex. The authors replace the second Thr residue with a Glu or Asp to prevent NME. This replacement causes early degradation of D2 and instability of the PSII complex which causes other component proteins to degrade as well (proteins D1, CP43, CP47). Chen et. al. [CVC02] provide another example (GST variant expressed in yeast). From these examples, Giglione et. al. [GVM03] go on to suggest that NME determines life-span in the plastid, mitochondria and cytoplasm via currently unknown machinery.

Giglione et. al. [GVM03] claim that NME controls protein life-span. This claim is based on few examples, namely the aforementioned protein D2 in *Chlamydomonas reinhardtii* and GST in yeast, which are both unstable and degraded early when NME is prevented. NME is a universally conserved mechanism, with MAP variants having preference to excise the same set of residues across species. MAP excises the first methionine residue when the second residue is one of the following seven: Gly, Ala, Pro, Ser, Thr, Val or Cys. The MAP enzyme requires small residues in the position adjacent to the methionine in order for the excision to occur. Even within the set of residues cleaved by MAP, Ala, Cys, Gly, Pro and Ser are more efficiently cleaved than Val or Thr [KYM$^+$07, FMP$^+$06]. The claim suggests that NME occurs for the benefit of a select small set of proteins, while the others are processed merely because they are present. Few, if any, other hypothesis for the role of NME have been proposed, however all statements connecting NME to protein degradation have little concrete support. The role of NME still seems to elude us. However, it is clear that NME is a necessary process for proper cell functioning since it is included in the minimal genome set of eubacteria [HPG$^+$99]. The removal of both MAP enzymes in yeast causes the death of the organism [LC95], which shows that NME is essential in yeast as well as other organisms.

1.2 N-End rule and NME

Frequently in the NME literature does the description of Varshavsky's N-End rule appear [Var96]. The N-End rule is depicted in Figure 1.1. It relates a proteins' half-life with its N-terminal amino acid. Figure 1.1 displays what Varshavsky calls *stabilizing* residues and *destabilizing* residues as white circles and

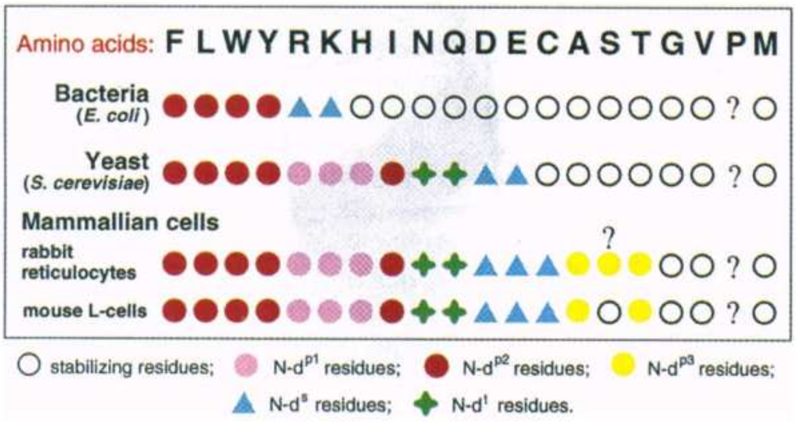

Amino acids: F L W Y R K H I N Q D E C A S T G V P M
Bacteria (E. coli)
Yeast (S. cerevisiae)
Mammalian cells
rabbit reticulocytes
mouse L-cells
?
stabilizing residues; N-d^p1 residues; N-d^p2 residues; N-d^p3 residues;
N-d^s residues; N-d^t residues.

Figure 1.1: N-End rule residues for different species. Stabilizing residues are represented as white circles. Destabilizing residues as colored shapes. The different shapes denote the level in the hierarchy of destabilizing residues. Figure taken from [Var96]

colored shapes, respectively. Stabilizing residues do not prevent early degradation of the protein while destabilizing promote degradation. The destabilizing residues are further categorized into a hierarchy depending on the molecular machinery required for identification and subsequent degradation. The degradation of secondary residues utilizes the molecular machinery of the primary residues in addition to their own specific machinery. The reuse of building blocks for degradation is the reason for describing destabilizing residues as a hierarchy. The molecular machinery for degradation differs from organism to organism, but the general hierarchy is maintained.

A connection is made by many authors between NME and the N-End rule. Earlier work by Arfin and Bradshaw [AB88] connect NME and degradation based on the identity of the residues acted upon for each process. Recent papers make similar connections and provide few examples where disruption of the normal NME process via residue substitution causes degradation of a particular protein [GVM03, CVC02].

1.3 NME and other N-terminal PTMs

NME is also invoked in the literature that studies other N-terminal modifications, such as N-acetylations, N-myristoylations, etc. One prevalent post-translational modification (PTM) in eukaryotes is N-acetylation, the addition of an acetyl group added to the N-terminus of proteins. This acetylation is carried out by N-acetyltransferases (NATs) that transfer an acetyl group from acetyl-CoA to the N-terminus. The N-acetylation event occurs in the majority of eukaryotic proteins (50% in yeast [LLS89], 80%-90% in higher eukaryotes [BR76, Bro79]), and occurs very rarely in prokaryote proteins. In yeast, three different N-acetyltransferase (NATA, NATB, NATC) are responsible for the acetylation. NATA targets Ser, Ala, Gly, Thr termini, those that are exposed after NME [Wal05]. The NATs act co-translationally, after NME has occurred if cleavage occurs at all. However, NME is not required for N-acetylation to take place [PNT+99]. N-acetylation, either the presence or absence, is required for the correct functioning of proteins, with each requirement being specific to the individual protein [Pol00].

There have been tools recently developed for the prediction of N-acetylation based on the sequence data [KBB05, CL08]. Martinez et. al. [MTV+08] also generate predictive rules for N-myristoylation and N-terminal S-palmitoylation in fungi, archea and eukaryotes. There have been few strong connections drawn between NME and N-acetylation, perhaps since acetylation can take place regardless of NME having occurred [PNT+99].

1.4 Comparative proteogenomics

Broadly, comparative genomics utilizes genomes from multiple organisms for analysis in order to draw conclusions about their similarities or differences. Comparative protegenomics takes a comparable approach utilizing protein data, often collected by mass-spectrometry, to perform similar analyses. Comparative genomics is frequently exploratory, using genetic data to generate knowledge. Cluster analysis is one employed method to deduce relationships from the data based on defined metrics. Similarly, exploratory methods are part of comparative proteogenomics as well, with the difference being in the data utilized.

Comparative Protegenomics as used in this work is used to loosely describe the approaches of analysis employed. Rather than a specific method for analysis, in this work the phrase describes an approach of utilizing multiple proteomes from different organisms for the analysis of N-terminal methionine excision. Many different types of tests are performed using data from varying organisms. By using protein together with other data from these different organisms, comparisons are made; hypothesis tested; and conclusions drawn.

1.5 Motivation for studying NME

As mentioned previously, NME has been connected to the degradation of proteins if certain amino acids are exposed due to the cleavage of the initiating methionine. However, this connection has yet to be tested thoroughly in the lab. Experimental tests are difficult to construct since mutants missing both MAP enzymes die [LC95]. Testing mutants with only one MAP enzyme present has shown to cause the remaining MAP enzyme to assume the role of the missing MAP enzyme [CVC02], thus further complicating testing. Testing one protein at a time for NME cleavage and stability has been performed [GVM03, CVC02], but it is both costly and laborious to generate sufficient data to draw conclusions on a proteome-wide scale. It is in analyzing NME using data from multiple proteomes where comparative methods show their potential.

This thesis continues in Chapter 2 by beginning to explore the importance of NME. Metrics of conservation are defined and used to show the importance of NME across species. Chapter 3 continues by focusing on proteins that are suspected to undergo NME while being highly conserved across species. That chapter attempts to identify a set of genes for which NME is required. The connection between NME and protein degradation is tackled in Chapter 4. Chapter 5 searches an avenue to potentially tie NME with other post-translational modifications, in particular N-acetylation. Conclusions and future directions for work are detailed in Chapter 6.

Chapter 2

Identifying NME and conservation

While the function of N-terminal methionine excision remains poorly understood, the rules governing NME are well studied. From the extensive studies conducted on methionine aminopeptidase (MAP) [CVC02, AB88, SST85, HSD$^+$89], there is ample information and understanding on the mechanics of NME. The MAP enzymes almost exclusively act upon the second amino acid of a nascent protein. Because of this property of MAP, the focus is brought to the beginning of the N-terminus of proteins. The set $\mathcal{X} = \{G, A, P, S, T, V, C\}$ denotes the set of amino acids in the second position which trigger MAP enzymes to cleave the N-terminal methionine. The remaining amino acids, which do not cause an excision, are denoted by $\overline{\mathcal{X}}$. Proteins are referred to as NME-proteins if their second amino acid is from $\mathcal{X}$ and $\overline{\text{NME}}$-proteins otherwise. NME-proteins can be thought of as predicted of undergoing NME and $\overline{\text{NME}}$-proteins as retaining their starting methionine.

A comparative proteogenomics approach is used to test if NME predicted proteins, ie - NME-proteins, are conserved across species and if the identifying feature is the second residue of a protein. This can be computationally tested since it implies that sequelogs (sequelogs are defined as orthologous genes with respect to sequence similarity [Var04]) of NME-proteins are themselves NME-proteins.

2.1 NME prediction rule

The two sets of NME and $\overline{\text{NME}}$ inducing residues, denoted by $\mathcal{X}$ and $\overline{\mathcal{X}}$ respectively, provide for a simple rule to label a protein sequence as undergoing NME or not. Using only the second residue of a protein as the identifying feature, a protein is labeled as undergoing NME if its residue from the second position is contained in $\mathcal{X}$. If it is instead contained in $\overline{\mathcal{X}}$, the protein is labeled as $\overline{\text{NME}}$. Recent large-scale mass-spectrometry studies (Gupta et. al. [GTJ$^+$07, GBB$^+$08]) confirmed that this simple rule is well correlated with experimental data on NME.

Since the NME mechanism is universally conserved in all species, one can conjecture that if a protein P is an NME-protein in one species, then its sequelogs in related species are also NME-proteins. Other studies have devised more detailed rules for the prediction of N-terminal methionine excision [FMP$^+$06]. The accuracy of an existing NME identification rule described in Frottin et. al. [FMP$^+$06] is investigated. The NME prediction rule defined by Frottin et. al. utilizes the amino acids in positions 2 and 3 (denoted P2 and P3 in the remainder of the text) as the features used in the classification. Under this rule, a protein does not undergo NME iff P2 $= \overline{\mathcal{X}}$ or P2 and P3 $= \{VE, VP, TE, TP\}$. The simpler rule, utilized in this work, relies solely on P2 and classifies a protein as an NME-protein iff P2 $= \mathcal{X}$. Utilizing this rule yields the confusion matrix in Table 2.1 below.

Table 2.1: Confusion matrix for rule: 'NME iff P2=$\mathcal{X}$' using mass spectrometry data from Gupta et. al. [GTJ$^+$07] for *S. oneidensis*, *S. frigidimarina* and *S. putrefaciens*

		Predicted	
		NME	$\overline{\text{NME}}$
Actual	NME	659	11
	$\overline{\text{NME}}$	96	565

The described rule results in an 8% error rate, most of which originates from the 96 false positives, as is seen in Table 2.1. A false positive means the rule predicted NME when it did not actually occur. True NME labels are determined by observing peptides via mass-spectrometry data that start at position 2 of a protein.

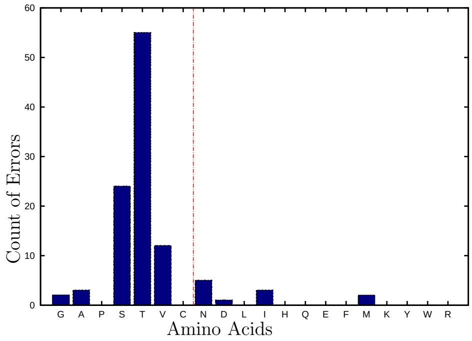


Figure 2.1: Error breakdown counts of each amino acid in position 2 for all mispredictions. False positives are shown to the left of the vertical red line, false negatives to the right.

This NME labeling is taken from Gupta et. al. [GTJ$^+$07]. Figure 2.1 shows the error breakdown of amino acids with false positives (left of the vertical red line) and false negatives (right of the vertical red line). The majority of false positives are attributed to mispredicting on residue T (55 of 96 false positive errors), followd by residue S (24 of 96 false positive errors). The true values of class labels are taken from *Shewanella* mass-spectrometry data described in Gupta et. al. [GTJ$^+$07]. The accuracy of this simple prediction rule is sufficient given the error rate of the data. The small sample size of the data, 1331 samples, along with the intrinsic labeling error rate prevents a more accurate data-driven model from being feasible. This simple rule also confers with the behavior of the MAP enzymes, which cleave when the seven $\mathcal{X}$-residues are present in position 2. The 1331 different samples are taken from three organisms *S.oneidensis*, *S.frigidimarina* and *S.putrefaciens*.

2.2 Measuring conservation

Prior to analyzing the different sequences for similarities in conservation of $\mathcal{X}$-residues, metrics for measuring conservation must be defined. A matrix is created for compiling the differences in sequelogs of two different species. This matrix is then utilized for computing different metrics of conservation for the two species in question. The matrix is referred to as a *conservation matrix*, since the metrics of conservation are computed utilizing its information. Each conservation matrix contains as many entries as sequelogs that exist between the two species being analyzed. For example, *S.oneidensis* and *S.frigidimarina* share 2729 sequelogs, as shown in Tables 2.2 and 2.3. Pairings from other species contain similar numbers of sequelogs.

2.2.1 Conservation matrix

A conservation matrix consists of the counts of an amino acid pair, occurring in a specified position in both sequences considered. For example, the conservation matrix $\mathcal{M}$ seen in Table 2.2 is read with the amino acid in the row corresponding to *S.oneidensis* and the amino acid in the column corresponding to *S.frigidimarina*. Element $\mathcal{M}_{ij}$ represents the number of counts of residue i in the first genome and residue j in the second genome, for each pair of sequelogs. For example, Table 2.2 shows $\mathcal{M}_{AG} = 4$, which is the number of counts *S.oneidensis* contains A and *S.frigidimarina* contains G for the position the conservation matrix is constructed. This section and sections to follow use the following abbreviations for *S.oneidensis*, *S.frigidimarina*, *S.putrefaciens*; SOne, SFri and SPut, respectively.

The conservation matrix is not symmetric, ie-$\mathcal{M}_{AG}$ is not necessarily equal to $\mathcal{M}_{GA}$, but this information can be aggregated in a post-processing step. A conservation matrix is created by defining the `Substitution` function. For our purposes it is defined as

```
Substitution(genome1, genome2, position, *selected)
```

which takes two genomes, and for each aligned gene in the specified position, the corresponding element in the matrix is incremented. The fourth argument, `selected` , is optional as denoted by the * symbol. This parameter allows one to specify a subset of genes from the two genomes over which to compute conservation. The specified subset is used rather than then entire set of genes. This option is used to compute conservation over the predicted NME-proteins or $\overline{\text{NME}}$-proteins. The `Substitution` function returns a filled conservation matrix $\mathcal{M}$ given a pair of genomes, `Substitution`(`SOne`, `SFri`, 2) produces Table 2.2.

Table 2.2: Conservation matrix at P2 for sequelog proteins of *S.oneidensis* and *S.frigidimarina*. The entry at the intersection of row i (corresponding to an amino acid X), and column j (corresponding to an amino acid Y), shows the number of sequelogs in which the second amino acid in *S.oneidensis* is X while the second amino acid in *S.frigidimarina* is Y.

S.frigidimarina

S.oneidensis

	G	A	P	S	T	V	C	N	D	L	I	H	Q	E	F	M	K	Y	W	R	X	-	Σ
G	29	6	1	1	0	0	0	1	1	0	0	0	0	0	0	0	2	0	0	0	0	9	50
A	4	154	4	17	11	6	0	3	2	1	2	0	0	0	2	8	0	0	0	1	0	26	241
P	0	3	43	10	5	1	0	2	0	1	0	1	3	0	0	1	3	0	0	0	0	23	96
S	1	14	5	205	33	1	0	9	1	2	4	3	0	3	2	10	4	1	0	2	0	65	365
T	0	11	1	22	102	1	0	17	2	3	2	0	1	0	0	11	7	0	0	1	0	34	215
V	1	2	0	3	2	19	0	0	1	0	4	1	1	0	0	1	0	0	0	2	0	14	51
C	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	8
N	1	0	0	10	13	0	0	82	5	1	2	1	1	2	0	9	15	0	0	0	0	25	167
D	2	0	0	4	1	1	0	4	52	0	0	0	0	11	0	4	2	0	0	0	0	8	89
L	0	0	2	2	2	1	0	4	1	97	14	1	3	0	7	10	2	0	2	1	0	22	171
I	0	1	1	3	3	2	0	1	0	6	78	0	0	0	2	7	7	0	0	1	0	12	124
H	0	0	0	2	1	0	0	3	0	0	0	14	3	0	0	0	1	0	0	2	0	1	27
Q	0	1	2	2	2	0	0	3	1	3	0	2	55	2	0	2	3	0	1	2	0	13	94
E	1	1	0	0	2	0	0	3	7	1	0	1	2	53	0	2	2	0	0	1	0	6	82
F	0	0	0	0	0	0	0	0	0	7	1	0	0	0	44	2	0	1	0	0	0	12	67
M	1	5	1	15	15	0	0	6	2	10	7	1	3	2	4	32	7	1	1	3	0	4	120
K	0	0	1	6	5	0	0	18	0	3	2	0	10	3	0	9	211	0	0	11	0	22	301
Y	0	0	1	0	0	0	0	0	0	0	0	1	0	0	2	0	0	14	0	0	0	0	18
W	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	11	0	0	0	13
R	1	0	0	3	1	0	1	3	0	1	0	1	2	0	0	0	5	0	0	63	0	13	94
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-	11	26	12	67	58	8	1	26	8	20	16	8	9	14	12	7	17	8	0	6	0	2	336
Σ	52	224	74	372	256	40	9	185	83	157	133	35	93	90	75	115	288	25	15	96	0	312	2729

Table 2.3: Conservation matrix at P3 for sequelog proteins of *S.oneidensis* and *S.frigidimarina*

S.frigidimarina

S.oneidensis

	G	A	P	S	T	V	C	N	D	L	I	H	Q	E	F	M	K	Y	W	R	X	-	Σ
G	23	3	1	5	3	0	0	0	1	1	0	1	1	0	1	2	3	0	0	1	0	8	54
A	2	47	4	3	7	3	0	3	2	2	1	0	4	1	0	4	2	0	0	0	0	9	94
P	1	3	40	4	2	2	0	1	1	2	0	2	0	0	0	0	1	0	0	1	0	18	78
S	4	6	4	69	16	1	1	16	2	2	2	0	4	5	1	4	4	0	0	4	0	30	175
T	0	10	1	12	88	4	1	13	2	3	8	2	8	2	1	5	13	1	0	2	0	21	197
V	0	3	0	0	6	65	0	0	1	5	5	0	1	2	0	3	1	0	0	0	0	17	109
C	0	0	0	0	0	0	6	0	0	1	0	0	0	0	0	0	0	0	0	1	0	3	11
N	4	1	1	15	10	0	0	90	9	2	2	5	5	1	1	2	18	1	0	6	0	21	194
D	3	3	0	3	1	0	0	5	67	0	0	0	2	7	0	1	3	0	0	0	0	15	110
L	0	2	2	2	1	7	1	2	0	155	15	1	2	0	8	8	4	0	0	3	0	22	235
I	1	0	2	0	1	18	0	4	1	11	128	0	0	0	6	8	6	2	0	0	0	24	212
H	0	2	1	1	3	2	0	3	0	2	0	29	7	0	0	3	1	2	0	1	0	8	65
Q	0	4	3	6	5	2	0	3	2	1	1	3	75	11	0	2	4	1	0	0	0	15	138
E	2	3	0	3	3	1	0	1	11	1	0	0	3	70	0	2	0	0	0	0	0	12	112
F	0	0	0	1	3	2	0	0	0	10	4	0	1	0	66	3	1	3	1	0	0	9	104
M	1	2	0	6	10	4	0	3	1	14	4	2	5	4	2	26	6	1	0	2	0	3	96
K	1	0	0	5	6	2	0	12	3	4	2	0	9	3	0	5	183	0	0	10	0	24	269
Y	0	0	0	0	1	2	1	1	0	2	0	0	2	0	3	1	3	37	0	0	0	3	56
W	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	14	0	0	2	19
R	3	3	2	6	4	1	1	4	0	4	0	1	3	1	0	1	9	1	0	74	0	2	120
X	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
-	5	17	6	26	25	14	2	19	12	31	21	8	22	7	8	4	27	7	2	11	0	6	280
Σ	50	109	67	167	195	130	13	180	115	256	193	55	154	114	97	84	289	56	17	116	0	272	2729

2.2.2 SimpleConservation metric

The data in the conservation matrix is difficult to interpret in the 20x20 matrix form. The simplest method of condensing the data in the matrix is to extract the ratio of conserved sequelogs to non-conserved sequelogs. This will represent the percentage of genes that conserve the amino acid in the position for which the matrix is created. This `SimpleConservation` metric is defined as the ratio of diagonal elements to the total number of elements in the conservation matrix $\mathcal{M}$.

$$\texttt{SimpleConservation(}\mathcal{M}) = \frac{\sum_i \mathcal{M}_{ii}}{\sum_i \sum_j \mathcal{M}_{ij}}$$

The diagonal of the matrix describes the amount of conservation of the different amino acids for the two sequences. A diagonal matrix would represent perfect conservation, `SimpleConservation` would return 1 in this case.

2.2.3 SetConservation metric

In addition to the `SimpleConservation` metric defined, it is also of benefit to measure the conservation of the residues in the $\mathcal{X}$ set. This can potentially show differences between NME-proteins and $\overline{\text{NME}}$-proteins. The `SetConservation` metric is defined as the ratio of the sum of amino acids of a specified set over the count of all amino acids. `SetConservation` requires a conservation matrix $\mathcal{M}$ and a set of amino acids $\mathcal{A}$.

$$\texttt{SetConservation(}\mathcal{M}, \mathcal{A}, \mathcal{B}) = \frac{\sum_{n \in \mathcal{A}} \sum_{m \in \mathcal{B}} \mathcal{M}_{nm}}{\sum_i \sum_j \mathcal{M}_{ij}}$$

As an example using Table 2.2 from above and amino acid set $\mathcal{A} = \mathcal{B} = \{G, A\}$, we obtain `SetConservation(Table 2.2,` $\{G, A\}$`,` $\{G, A\}$`)` $= \frac{29+6+4+154}{2083}$. This metric can also be viewed as reducing the 20x20 conservation matrix into a 2x2 matrix. In this reduced matrix, each element is the count of amino acids to transition from one set in one proteome to another amino acid set in the other proteome. Continuing with the example above, the conservation matrix from Table 2.2 is reduced to Table 2.4b.

This reduction shows more clearly the counts of genes with sequelogs that are conserved within the given set of amino acids. Table 2.4a displays the possible

Table 2.4: Reduced conservation matrix example

(a) Example format

$\mathcal{X} \rightarrow \mathcal{X}$	$\mathcal{X} \rightarrow \overline{\mathcal{X}}$
$\overline{\mathcal{X}} \rightarrow \mathcal{X}$	$\overline{\mathcal{X}} \rightarrow \overline{\mathcal{X}}$

(b) Reduced Table 2.2

725	129
119	1110

states for set $\mathcal{X}$, the diagonal elements retaining conservation within a set $\mathcal{X} \rightarrow \mathcal{X}$ and $\overline{\mathcal{X}} \rightarrow \overline{\mathcal{X}}$. The off-diagonal elements representing the move from one set to its compliment, $\mathcal{X} \rightarrow \overline{\mathcal{X}}$ and $\overline{\mathcal{X}} \rightarrow \mathcal{X}$. This metric is used to view the conservation within the NME protein set and the $\overline{\text{NME}}$ protein set in the analysis to follow.

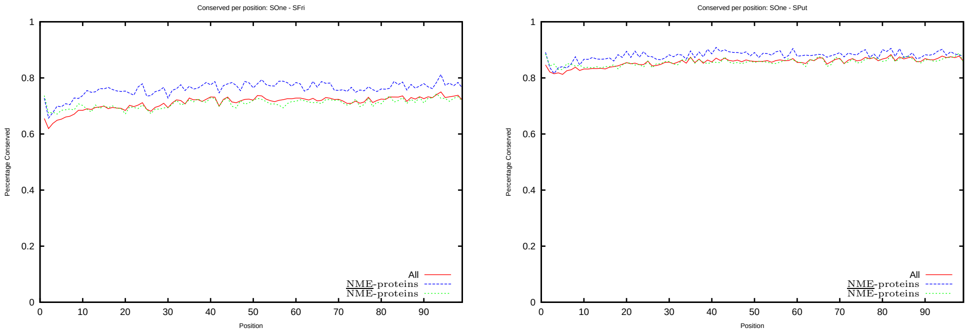
2.3 Analysis of NME-protein conservation

The two metrics of conservation defined previously, `SimpleConservation` and `SetConservation`, are now utilized to determine the importance of NME. As mentioned previously, NME-proteins serve as a proxy for NME verified proteins.

Using first the `SimpleConservation` metric, the amount of conservation over all positions from 2 to 100 is computed. Three sets of proteins are used: all proteins; $\mathcal{X} \rightarrow \mathcal{X}$-proteins (NME-proteins in both species); and $\overline{\mathcal{X}} \rightarrow \overline{\mathcal{X}}$-proteins ($\overline{\text{NME}}$-proteins in both species). These three sets are used to determine if there is a difference in conservation for NME-proteins versus $\overline{\text{NME}}$-proteins. Plotting the conservation over different positions attempts to determine if the conservation increases or decreases as the focus is moved away from the N-terminus. The 'All' curve in Figure 2.2a below is generated by calling:

```
SimpleConservation(Substitution(SOne,SFri, p))
```

for $\forall p \in \{2, 3, ..., 100\}$. The $\mathcal{X} \rightarrow \mathcal{X}$ and $\overline{\mathcal{X}} \rightarrow \overline{\mathcal{X}}$ curves are generated by a similar call, but using only $\mathcal{X} \rightarrow \mathcal{X}$ and $\overline{\mathcal{X}} \rightarrow \overline{\mathcal{X}}$ protein sets as the fourth argument to `Substitution`, respectively. Figure 2.2b is generated in a similar manner using the *S.oneidensis* and *S.putrefaciens* proteomes.



(a) SOne and SFri

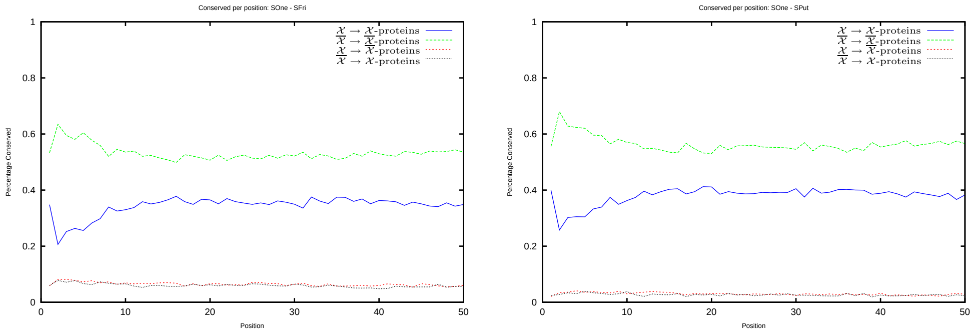
(b) SOne and SPut

Figure 2.2: SimpleConservation over positions 2 to 100 for different sets of sequelogs. (a) *S. oneidensis* and *S. frigidimarina* shown. (b) *S.oneidensis* and *S. putrefaciens* shown.

From Figure 2.2a and Figure 2.2b there is only a small difference between the ‘All’ (red) and $\overline{\text{NME}}$-protein (green) lines compared to the NME-protein line (blue). This small, appreciable difference between the lines yields little information about the conservation of NME. There is a drop in conservation among the amino acids for all three plotted sets in both figures near the N-terminus. This shows a drop in overall conservation in the N-terminus from positions 3 and on. The MAP enzymes are universally conserved and are shown to act on the residue in the second position. The drop in conservation at position 3 can be viewed as the beginning of the less universally important residues of the N-terminus. The second position residue is more highly conserved since it will affect how MAP interacts with the protein. The third position and further on the N-terminus are not involved in MAP or protein structure, and thus are likely to be less conserved, which is observed in Figures 2.2a and 2.2b.

The `SetConservation` metric will show the conservation within a set of amino acids. For this particular test the $\mathcal{X} = \{G, A, P, S, T, V, C\}$ set is used. Figures 2.3a and 2.3b show the conservations of the amino acids associated with the 2x2 reduced conservation matrix described previously in Table 2.4.

The plots in Figure 2.3 represent the set conservation for either $\mathcal{X}$-residues or $\overline{\mathcal{X}}$-residues. All sequelog sequences are considered in the creation of each con-



(a) SOne and SFri

(b) SOne and SPut

Figure 2.3: SetConservation over positions 2 to 50 for All sequelogs. (a) *S. oneidensis* and *S. frigidimarina* shown. (b) *S.oneidensis* and *S. putrefaciens* shown.

servation matrix. The $\mathcal{X} \rightarrow \mathcal{X}$ curve (blue) in the plots are generated by using:

$$\texttt{SetConservation(Substitution(SOne,SFri, p),}\ \mathcal{X},\ \mathcal{X}\texttt{)}$$

for $\forall p \in \{2, 3, ..., 50\}$. A sharp drop in the conservation of the NME-protein set is observed starting at position 3. The conservation at position 2 is the same as positions farther away from the N-terminus. A protein's N-terminus is frequently unfolded and free-floating, not affecting the structure of the protein. These conservation rates suggest that the second position is being conserved in the set $\mathcal{X}$ as much as in positions that affect the structure of the protein. Position 3 and adjacent positions drop in this conservation with respect to $\mathcal{X}$. Recalling Figures 2.2 of the `SimpleConservation`, these positions also saw lower rates of conservation of amino acids. This shows that position 2 is in fact important and position 3 likely not necessary to be conserved or play a significant role in NME.

Chapter 3

NME-Critical Proteins

3.1 NME acting on many proteins

While Chapter 2 shows that some NME-proteins are conserved in related species, we do not know which proteins require the NME process to occur. It is known that NME is a necessary process since mutants without the MAP1 and MAP2 enzymes do not survive [LC95]. However, it may be that only some proteins require the starting methionine to be removed, while the other proteins are merely 'bystanders' being cleaved since the MAP machinery already exists. This hypothesis can be tested via computational means utilizing comparative genomics methods.

3.2 Methods to test for necessary proteins

In order to properly test if such an NME necessary set exists, protein sequences from multiple species must be utilized in the analysis. Three data sets of different organisms are used, *Shewanella*, *Saccharomyces* and mammalian. The bacterial data set comprising of 19 *Shewanella* species is taken from 10 species considered in the work of Gupta et. al. [GTJ+07] and supplimented with 9 additional species. The *Saccharomyces* data set comprises of 7 species of the fungi obtained from the Saccharomyces Genome Database (http://downloads.yeastgenome.org/). The mammalian data set contains sequences from six species: human, chimpanzee, macaque, cow, opossum and rat. The mammalian data is obtained from the MSOAR project (http://msoar.cs.ucr.edu/MSOAR2.0/Rawdata/). An additional species, mouse, is provided in MSOAR but is removed from this analysis.

If a protein in one species and its sequelogs all require NME in order to properly function, the excision should be conserved across the various species. Finding NME-proteins in all sequelogs is surprising unless NME is truly important for the functioning of this particular protein. To illustrate this likely importance, if the probability of an NME-protein being conserved is $p = 0.4$, then the probability that it is conserved in all 19 *Shewanella* sequelogs by chance[1] is $(0.4)^{19} = 1.7 \cdot 10^{-8}$. This unanimity in conservation is unlikely unless there is some underlying reason for retaining $\mathcal{X}$-residues in the second position.

To determine the number of proteins for which NME is necessary, the sequence data is used to identify which genes are conserved NME-proteins. Those genes that retain an $\mathcal{X}$-residue in the second position purportedly do so because there is some reason to keep the identity of the amino acid. The process to create this list, essentially a distribution of NME-proteins among sequelogs, is the same for each of the different data sets. Using the *Shewanella* data set as an example, only the sequelogs present in all 19 species are filtered out and retained. The set of 1860 remaining proteins is iterated over to produce a table, seen as leftmost group in Figure 3.1. For each gene, all of the 19 sequelogs are labeled either 1 or 0, representing an $\mathcal{X}$ or $\overline{\mathcal{X}}$ residue respectively, for the position in question. Figure 3.1 shows these as steps 2 and 3, going from sequences to counting the number of $\mathcal{X}$-residues for position 2 in the example. The fourth and final step sums the number of 1's in the binary vector, yielding the number of $\mathcal{X}$-residues for the position and protein in question.

This procedure creates a table ranging from 0 counted $\mathcal{X}$-residues for the position, up to 19 such residues. Proteins contained in the '0' row of the table represent those that do not contain any sequelogs with an $\mathcal{X}$-residue in the position in question; ie-all $\overline{\text{NME}}$-proteins. Those contained in the '19' row represent

[1] We use $(0.4)^{19}$ since $\approx 40\%$ of proteins in *Shewanella* species have $\mathcal{X}$-residues in the 2nd position. While this estimates uses an (unrealistic) independence assumption, it still illustrates that there are few spurious NME-critical proteins.

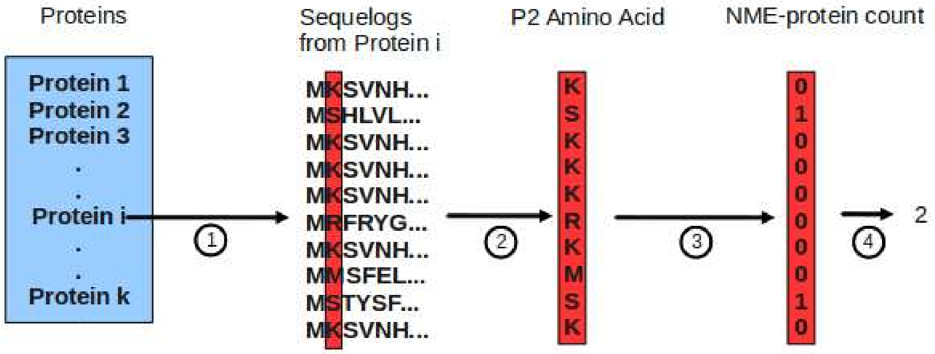


Figure 3.1: Diagram describing method to compute tables of conservation of NME-proteins. Step 1 depicts selecting a gene conserved among all available organisms and extracting all sequences. Step 2 keeps only the residues from the position in question, here position 2. Step 3 converts this into a binary vector of $\{\overline{\mathcal{X}}, \mathcal{X}\} = \{0, 1\}$. Step 4 sums the binary vector to obtain the number of NME-proteins for the particular gene.

the proteins that contain an $\mathcal{X}$-residue in all of the available sequelogs investigated; ie-all NME-proteins. The same process is used to create the tables for the *Saccharomyces* and mammalian data sets which contain fewer organisms, 7 and 6 respectively. The sequence data used in the creation of all these tables is not aligned since the true N-terminus positions are to be compared with one another.

3.3 Analysis of necessary proteins

3.3.1 NME necessary tables

The process described in Section 3.2 generates a table of $\mathcal{X}$-residue conservation for various positions along the N-terminus. The compilation for *Shewanella* is shown in Table 3.1(a). The row to note, all conserved $\mathcal{X}$-residues, in Table 3.1(a) is highlighted in bold. This row shows that over positions 2, 3, 4 and 5 of sequelogs, there is a larger number of conserved $\mathcal{X}$-residues in position 2 than in 3. As the position moves away from the N-terminus of the protein, the number of conserved $\mathcal{X}$-residues slowly increases. This is consistent with the prior finding that conservation for $\mathcal{X}$-residues drops at position 3 and increases as the positions move away from the N-terminius. This phenomena was shown previously in Figure 2.3.

Table 3.1: Table of $\mathcal{X}$-residue conservation for positions 2 through 5 for the 19 *Shewanella* species

(a) *Shewanella* species with 1860 sequelogs.

Number NME	P2 Count	P3 Count	P4 Count	P5 Count
0	558	664	617	606
1	149	189	186	187
2	76	132	101	118
3	54	117	99	82
4	34	67	62	46
5	36	47	43	42
6	27	43	55	37
7	31	18	32	36
8	26	42	36	30
9	25	21	27	23
10	23	29	36	27
11	29	35	35	43
12	34	43	33	40
13	38	43	19	44
14	41	31	46	37
15	35	33	34	40
16	61	38	48	51
17	83	45	63	49
18	125	61	82	96
19	**375**	**162**	**206**	**226**

In order to determine if a pattern of conservation of $\mathcal{X}$-residues exists with respect to the position along the N-terminus, the other data sets must also be compiled into this table form. Tables for *Saccharomyces* and the mammalian organisms are shown in Tables 3.1(b) and 3.1(c), respectively.

Tables 3.1(b) and 3.1(c) show a similar trend to that seen in Table 3.1(a) for *Shewanella*. There are a larger number of conserved $\mathcal{X}$-residues in position 2 than in positions 3, 4 or 5. From the *Shewanella* and *Saccharomyces* tables, there is a very similar distinction in the number of perfectly conserved $\mathcal{X}$-residues among the different positions. The mammalian set contains considerably more proteins than the other two, but also the ortholog mapping between organisms is more difficult and error-prone. As such it should be taken with less credibility. These

Table 3.1: Tables of $\mathcal{X}$-residue conservation among different positions of the N-terminus for *Saccharomyces* data set.

(b) 7 *Saccharomyces* species with 1502 sequelogs

Number NME	P2 Count	P3 Count	P4 Count	P5 Count
0	274	420	427	460
1	196	297	297	282
2	88	154	168	131
3	33	69	67	66
4	69	82	76	72
5	113	128	137	150
6	260	182	166	175
7	**469**	**170**	**164**	**166**

Table 3.1: Tables of $\mathcal{X}$-residue conservation among different positions of the N-terminus for mammalian data set.

(c) 6 mammalian species with 7934 sequelogs

Number NME	P2 Count	P3 Count	P4 Count	P5 Count
0	1591	1964	2105	2160
1	1018	1170	1195	1176
2	559	614	652	609
3	456	458	526	494
4	698	700	685	740
5	1389	1279	1201	1262
6	**2223**	**1749**	**1570**	**1493**

tables provide evidence to continue the search for NME-critical proteins, and also provides sets of proteins to focus on.

3.3.2 Identifying NME necessary proteins

In order to identify the proteins that are necessary to undergo N-terminal methionine excision for their function, a method of filtering out proteins must be created. Using the NME tables from Section 3.3.1, the protein IDs from each cell in the table can be extracted. The perfectly conserved $\mathcal{X}$-residues in the second position are of interest. The IDs from this set were extracted from each of the three tables. Using a mapping of orthologous proteins, specifically from the eggNOG database (http://eggnog.embl.de/), we can find which relating to NME-proteins are shared across the species. This approach unfortunately only reveals a small set of proteins (35) that are shared among *S.oneidensis* and *S.cerevisiae*. This could be due to ambiguous mappings as well as lack of mappings for some proteins.

Since this attempt to use orthologous group IDs proved unsuccessful, another approach would be needed to determine if there is a commonality among conserved NME-proteins. Another possibility is to utilize an ontology of different proteins to determine if natural groupings exist. The proteins from *S.oneidensis* and *S.cerevisiae* could be mapped onto an existing gene ontology (http://www.geneontology.org/). These three ontologies provides a hierarchical description of gene products in three domains: biological processes; cellular compartments; and molecular function. There has been work done in utilizing gene ontologies to generate measures of similarity among different proteins [LSBG03, SDRL06, SQGD08]. Schlicker et. al. [SDRL06] perform a functional comparison of yeast proteins with human proteins. Utilizing Schlicker et. al. [SDRL06]'s measures, one can compare the set of NME-proteins in *S.oneidensis* and *S.cerevisiae* to determine if these proteins perform similar functions. Should a small number of clusters emerge from such a comparison, one would be able to focus on a smaller set of proteins

Chapter 3 is in part also currently being prepared for submission for publication of the material. Bonissone, Stefano; Gupta, Nitin; Romine, Margaret; Pevzner, Pavel. The thesis author was the primary investigator and author of this material.

Chapter 4

NME for Protein Regulation

4.1 NME and regulation of proteins

Protein degradation is an often suggested reason for N-terminal methionine excision. Many authors in the literature connect NME with the N-End rule. This connection can be further examined using computational methods. This chapter explores the possibility of a connection between NME and regulation of proteins.

NME could be used for the regulation of proteins and adapting to new environments. If NME relates to degradation, with the mutation of one or a few amino acids a protein could dramatically increase its expression levels by preventing degradation. An advantageous increased presence of a particular protein that requires NME would be retained in the population of organisms.

4.2 Methods to test for regulatory properties

Regulation of proteins carrying out similar functions are done simultaneously in bacterial genomes via operons. Since proteins that are used as building blocks for one larger complex are required in similar quantities, genes comprising the operon are regulated in synchrony. Should NME play a role in the regulation of proteins, then operons would need to contain either all NME-proteins or all $\overline{\text{NME}}$-proteins. This would enable an operon to be expressed or completely repressed. In other words, all the proteins originating from an operon would need to consist of $\mathcal{X}$-residues in position 2 or all contain $\overline{\mathcal{X}}$-residues in that position. A mix of $\mathcal{X}$ and $\overline{\mathcal{X}}$ would mean that some proteins from the operon undergo NME while others do not. This would be inconsistent with the hypothesis that NME controls degradation since all genes in an operon would need to be expressed in coordination.

The hypothesis that NME controls degradation in operons is tested by using operon groupings for two organisms, *S.oneidensis* and *Escherichia coli*. The operon groupings for *S.oneidensis* are obtained from Gupta et. al. [GTJ$^+$07]. While the *E.coli* operon sets are taken from RegulonDB (http://regulondb.ccg.unam.mx) for the K12 strain of the bacteria. Operons containing more than two genes are retained for the analysis.

In order to test if NME is connected with protein regulation, there needs to be an established link between half-life rates and NME-proteins. A connection of all NME-proteins or $\overline{\text{NME}}$-proteins in operons would suggest that there is connection within a group of related proteins. A connection between individual protein half-lives and their second residue would solidify such a relationship. If NME dictates protein degradation and is connected to the N-End rule via stabilizing residues, then there should exist a correlation between NME-proteins and longer half-lives. The data set used to test this hypothesis is obtained from the supporting information of Belle et. al. [BTB$^+$06]. The authors provide protein half-lives for yeast by using ≈4200 tagged proteins (tagged on the C-terminus), collected during the exponential growth phase under standard laboratory conditions. The abundance of each tagged protein were measured during varying amounts of time after inhibition of protein synthesis, and this was used to determine half-life.

4.3 Analysis of tests

4.3.1 Operon analysis

Prior to utilizing the operon groupings for *S. oneidensis* and *E.coli*, the distances on the genome strands of perfectly conserved NME-proteins are analyzed. Since operons encompass a contiguous set of genes, one would expect that clusters of NME-proteins and $\overline{\text{NME}}$-proteins appear if NME is used as a regulation mechanism for operons. Table 4.1 shows the number of contiguous elements from NME-proteins in *S.oneidensis* for genes *perfectly conserved* in all 19 *Shewanella* organisms. A perfectly conserved protein in this context refers to one of the 375 NME-proteins in the last row of the P2 column of Table 3.1a. These 375 NME-proteins are considered *perfectly conserved* since they contain an $\mathcal{X}$-residue in each of their sequelogs for the second position. Most of the contiguous elements for perfectly conserved NME-proteins are encompassed in what appear to be smaller operons, as seen in Table 4.1.

Table 4.1: Number of Adjacent Groups from set of 375 perfectly conserved *S.oneidensis* NME-proteins

Number of contiguous elements	Counts of contiguous sets
2	118
3	41
4	16
5	6
6	3
7	1
8	0

Figure 4.1 shows the the distance between adjacent perfectly conserved NME-proteins for *S.oneidensis.* Each of the 375 perfectly conserved NME-proteins for the *Shewanella* data set are ordered according to the location of their sequelog on the strands for *S.oneidensis.* The distances between these ordered genes on the strands are plotted in Figure 4.1. The solid red curve indicates the expected number for each given distance. The experimental distribution follows the expected closely after a distance of one in the strand.

Figure 4.1 shows that many of the perfectly conserved NME-proteins are adjacent to one another in the strands or with a gap of one location. This encourages testing the hypothesis using the operon sets to determine if the operons are indeed saturated with NME-proteins or $\overline{\text{NME}}$-proteins.

Operon sets are given for *S.oneidensis* and *E.coli.* Each set contains the ID numbers of proteins belonging to the same operon. For each operon set, the number of NME-proteins is counted. Sets of size 2 are disregarded since there are many of them and are too small to be of use. Figures 4.2a and 4.2b show the

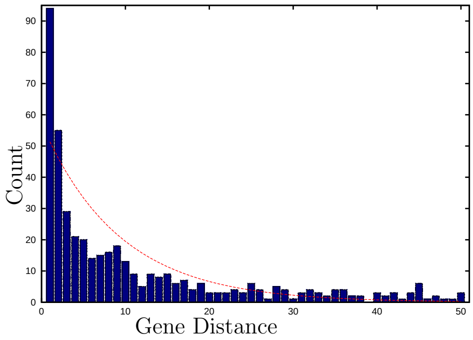
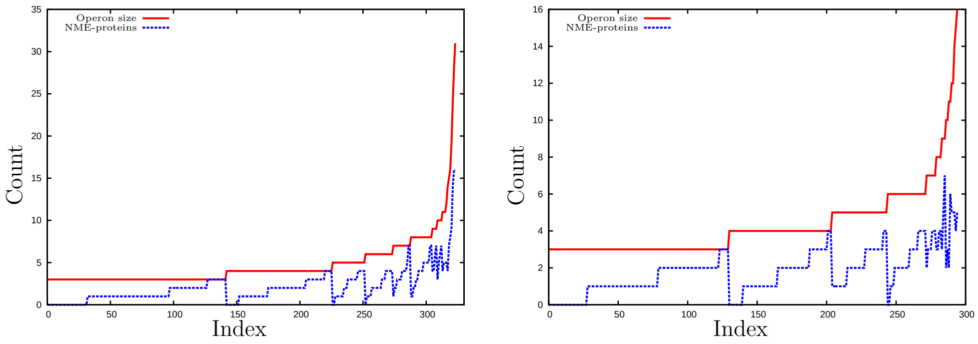


Figure 4.1: Gene distance histogram for perfectly conserved NME-proteins from *S.oneidensis* with expected distribution plotted as the solid red curve

results of the *S.oneidensis* and *E.coli* operon test, respectively. The solid red curves in each figure represent the operon size. Each figure contains all of the operons with more than 2 genes for each organism, sorted by operon size (as is seen by a monotonically increasing solid red curve). The dashed blue curve represents the number of NME-proteins in each of the operons. All of the operons, within a set of a given size, are sorted by their count of NME-proteins. This is seen by the monotonically increasing dashed blue line within each group of operon sizes.



(a) *S.oneidensis* Operons

(b) *E.coli* Operons

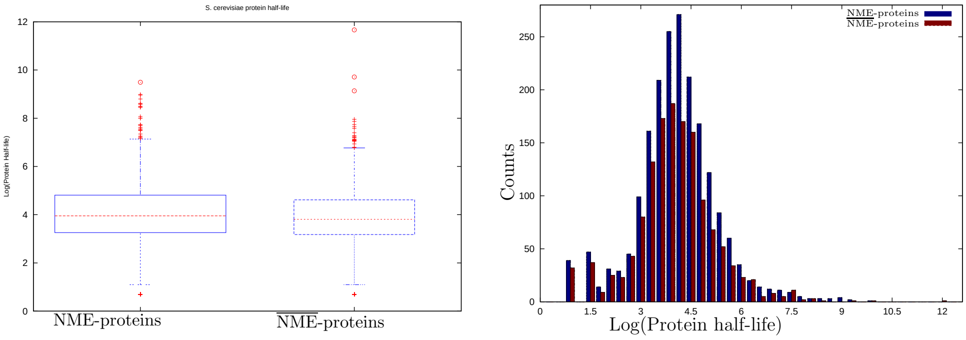
Figure 4.2: Operons sorted by sizes (solid red) and NME-protein counts (dashed blue) for (a) *S.oneidensis* (b) *E.coli*

Each operon set is sorted according to the count of NME-proteins in Figures 4.2. If the hypothesis of NME playing a role in the regulation of operons is correct, one would expect the dashed blue curve to only touch either 0 or the solid red line. This would correspond to an operon containing unanimity for $\overline{\text{NME}}$-proteins and NME-proteins, respectively. Instead, Figures 4.2a and 4.2b display the dashed blue curve for both organisms seldom touching the two extremes. For the majority of the operons, the NME-protein count is between 0 and the max for the operon size. This suggests that NME-proteins play no role in the regulation of operons.

4.3.2 Protein half-lives

Section 4.3.1 shows that there is little, if any, evidence to suggest that NME plays a role in regulating operons. However, the operon test did not attempt to show any connection between NME and protein degradation outside of an operon set. To test if NME is connected to protein half-life, the half-life data set from Belle et. al. [BTB+06] is utilized. The correlation between NME-proteins and their experimentally observed half-lives is computed. Each of the genes in the data set are labeled 0 or 1, for $\overline{\text{NME}}$-protein or NME-protein. With this labeling, the data set is now comprised of two variables, one for the quantification of the half-life and the other for an NME-protein or $\overline{\text{NME}}$-protein. The point-biserial correlation coefficient is computed since one variable a (protein half-life), $a \in \mathbb{Z}$, and the second variable b (NME-protein or $\overline{\text{NME}}$-protein), $b \in \{0, 1\}$. The resulting correlation coefficient between a anb b is 0.0466, showing that there is no correlation between the NME-proteins and half-lives. Figure 4.3a shows the lack of correlation between these two variables as a boxplot. The two boxes in the plot correspond to the NME-protein set and $\overline{\text{NME}}$-protein set. The y-axis is the natural log of the quantified half-lives. Should a correlation exist, the means of the two sets should be located in different places on the y-axis. If NME-proteins did correlate with half-life, we would expect the box on the left hand side of Figure 4.3a to have a larger mean than the box on the right.

The lack of correlation between the two sets is further exemplified in Figure 4.3b. Here the two sets are visualized as overlapping bar plots, NME-proteins as filled blue bars, $\overline{\text{NME}}$-proteins as red bars. The two distributions are indistinguish-



(a) *S.cerevisiae* half-life boxplot (b) *S.cerevisiae* protein set breakdown

Figure 4.3: Protein half-life correlation in *S.cerevisiae* for NME-proteins and $\overline{\text{NME}}$-proteins. (a) Half-life boxplot for NME-protein and $\overline{\text{NME}}$-protein sets. (b) Histogram of protein half-life for the two sets (NME-proteins and $\overline{\text{NME}}$-proteins).

able from one another, and are centered on approximately the same values. This also suggests that there is no difference between the two sets, that NME-proteins and $\overline{\text{NME}}$-proteins do not differ from one another with regard to half-lives.

This rather surprising find, that NME-proteins do not correlate with half-lives, prompted further questions. One possible explanation for the finding is that the NME-protein set utilized may not be correct. Another possible explanation is that NME-proteins do not in fact correlate to half-lives. We must also allow for the possibility that experimental half-life data is unreliable and cannot be used to test such a conjecture. In order to exclude the first possibility of an erroneous $\mathcal{X}$ set, all 2^{20} amino acids sets must be exhaustively searched. This brute-force search of all sets computed the point-biserial correlation between each candidate amino acid set and the half-life data from Belle et. al. [BTB+06]. The set with the largest correlation found had a coefficient of 0.155. This is stronger than the initial $\mathcal{X}$ set, but is still not a significant correlation.

The hypotheses tested in this chapter suggest, with consensus among one another, that NME-proteins and protein degradation do not appear to be strongly connected. Despite the connections made frequently in the literature, the evidence found does not support such a conclusion. The connection with the N-End rule appears natural because of the similarity between amino acids cleaved by methionine aminopeptidase and the set of stabilizing amino acids from the N-End rule. Tests for both operon regulation of proteins and individual protein half-lives in yeast failed to show such a connection. Other possibilities are explored, specifically a connection between NME and other post-translational modifications (PTMs), which are tested in the following chapter.

Chapter 4 is in part also currently being prepared for submission for publication of the material. Bonissone, Stefano; Gupta, Nitin; Romine, Margaret; Pevzner, Pavel. The thesis author was the primary investigator and author of this material.

Chapter 5

NME and other Post-translational modifications

5.1 N-terminal Post-translational modifications

Post-translational modifications (PTMs) occur frequently to proteins, and can act as a means of regulation. The number of PTMs that occur in humans is thought to increase the diversity of the types of proteins by approximately two fold [Wal05]. Despite their name, some PTMs, namely N-terminal acetylation in eukaryotes [Pol00], can occur co-translationally just as NME. The importance, prevalence and some of their locations on the N-terminus make some post-translational modifications another potential avenue for connection to NME.

This chapter will explore a possible connection between N-terminal methionine excision and other post-translational modifications. Tests performed include utilizing mass-spectrometry and sequence data to draw some preliminary conclusions about such a connection. Further tests on more comprehensive and larger data sets will be necessary to further stabilize and provide confidence to conclusions.

5.2 Ala/Ser-critical proteins

Earlier, while searching for NME-critical proteins, we analyzed *group* conservation of all seven $\mathcal{X}$-residues at initial positions of the proteins (checking whether the second residue in each sequelog is an $\mathcal{X}$-residue). We now test for conservation of *individual* amino acids (instead of the set $\mathcal{X}$) at the second position and compare it to the conservation at the 3rd position (where conservation is not expected) as a control.

Table 5.1 shows that Ala is conserved among all *Shewanella* species in 79 proteins at position 2 but only in 14 proteins at position 3. Ser is conserved in 75 proteins at position 2 but only in 16 proteins at position 3. Conservation levels of other amino acids are rather similar between positions 2 and 3, suggesting that Ala and Ser might be the important targets for NME in *Shewanella* (since Ala and/or Ser are exceptionally well-conserved, it is reasonable to assume that NME affects function of these proteins). Since the N-terminal amino acids that are targets for N-terminal PTMs are expected to be highly conserved, it is reasonable to conjecture that Ala and Ser are exposed (in Ala- and Ser-critical proteins) to enable further modifications.

Yeast and mammalian species also show much higher levels of conservation for Ala and/or Ser in the second position compared to other residues. The distributions, however, are not identical across these different types of organisms, perhaps a reflection of differences in the types and frequency of PTMs found in these organisms.

The phenomenon of Ala/Ser elevation can also been seen in Tables 5.5(a), 5.5(b) and 5.5(c) which display a 2-dimensional counterpart of Table 5.1 for bacterial, yeast and mammalian data sets, respectively. Each entry of the tables, indexed by (X_1, X_2), defines a set $\mathcal{X}$ of two amino acids to use for computing the conservation for that cell (the same process detailed in Figure 3.1 is used). It differs from the process depicted in Figure 3.1 in that proteins with perfectly conserved sequelogs with one amino acid are not counted. The count for $(X_1, X_2) = \{X_1, X_2\} - \{X_1\} - \{X_2\}$, where $\{X_1\}$, and $\{X_2\}$ represent the $\mathcal{X}$-residue sets that contain a single residue, X_1 and X_2, respectively (the perfectly conserved sequelogs). $\{X_1, X_2\}$ represents the $\mathcal{X}$-residue set for the pair of amino acids in question. For example the entry at (Ala, Ser) describes the number of proteins with all sequelogs being either Ala or Ser at a particular position excluding proteins for which all sequelogs have only Ala or all sequelogs have only Ser in the second position. Table 5.5(a) shows that Ala and Ser have the largest difference between positions 2 and 3 among mixed residues for *Shewanella.* This interchangeability is also present in Table 5.5(b) for *Saccharomyces*, but not present in the mammalian data set seen in Table 5.5(c)

Table 5.1: Conservation of single amino acids for *Shewanella*, *Saccromyces* and mammalian sequence data sets (in positions 2 and 3). Ala and Ser (shown in bold) are by far the most conserved residues in the 2nd position. The number of proteins with Ala in the second position in *Shewanella*, *Saccharomyces*, and mammalian datasets are 79, 24 and 541, respectively. The number of proteins with Ser in the second position in *Shewanella*, *Saccharomyces*, and mammalian datasets are 75, 130 and 268, respectively.

(a) *Shewanella*

AminoAcid	P2	P3
G	11	11
A	**79**	**14**
S	**75**	**16**
P	16	17
V	6	25
T	21	20
C	3	2
L	39	52
I	28	45
N	21	16
D	13	19
Q	20	20
K	58	55
E	19	21
M	1	1
H	6	8
F	14	27
R	23	28
Y	7	17
W	3	4

(b) *Saccharomyces*

AminoAcid	P2	P3
G	18	13
A	24	15
S	**130**	**13**
P	15	9
V	14	8
T	10	10
C	1	0
L	25	18
I	1	12
N	11	9
D	13	6
Q	3	10
K	14	19
E	3	4
M	0	0
H	2	4
F	11	14
R	7	33
Y	1	10
W	1	2

(c) Mammalian

AminoAcid	P2	P3
G	168	119
A	**541**	**186**
S	268	190
P	116	81
V	60	64
T	90	104
C	15	29
L	94	171
I	24	25
N	53	59
D	97	93
Q	28	55
K	67	97
E	177	138
M	25	30
H	18	17
F	39	49
R	45	121
Y	19	24
W	19	29

Table 5.2: Tables of $\mathcal{X}$-residue conservation among different positions of the N-terminus for *Shewanella*, *Saccharomyces* and mammalian data sets for $\mathcal{X} = \{G, P, T, C, V\}$ with Ala and Ser removed from the set of 7 stabilizing residues. For example, the second column in (a) shows that the second residue for 900 is $\mathcal{X}$ in no species, is $\mathcal{X}$ in 1 species for 247 proteins, and is $\mathcal{X}$ in all 19 species for 62 proteins.

(a) 19 *Shewanella* species (1860 proteins).

Number NME	P2 Count	P3 Count	P4 Count	P5 Count
0	900	904	830	831
1	247	228	239	253
2	129	150	145	127
3	88	98	93	69
4	57	49	53	70
5	31	41	52	40
6	44	36	49	18
7	27	28	36	34
8	20	22	34	34
9	25	13	23	20
10	24	31	27	28
11	26	28	22	32
12	29	27	24	30
13	22	23	13	25
14	25	14	20	22
15	12	13	18	21
16	16	14	25	22
17	33	23	30	28
18	43	37	38	51
19	62	81	89	105

(b) 7 *Saccharomyces* species (1502 proteins)

Number NME	P2 Count	P3 Count	P4 Count	P5 Count
0	710	742	676	696
1	298	306	367	323
2	112	123	126	121
3	65	63	64	58
4	60	60	65	62
5	80	86	74	98
6	94	73	80	83
7	83	49	50	61

(c) 6 Mammalian species (7934 proteins)

Number NME	P2 Count	P3 Count	P4 Count	P5 Count
0	4092	3555	3505	3480
1	1587	1536	1539	1521
2	531	684	653	636
3	329	411	457	480
4	362	531	573	596
5	480	652	621	675
6	553	565	586	546

5.2.1 Tests of frequency for Ala/Ser-critical proteins

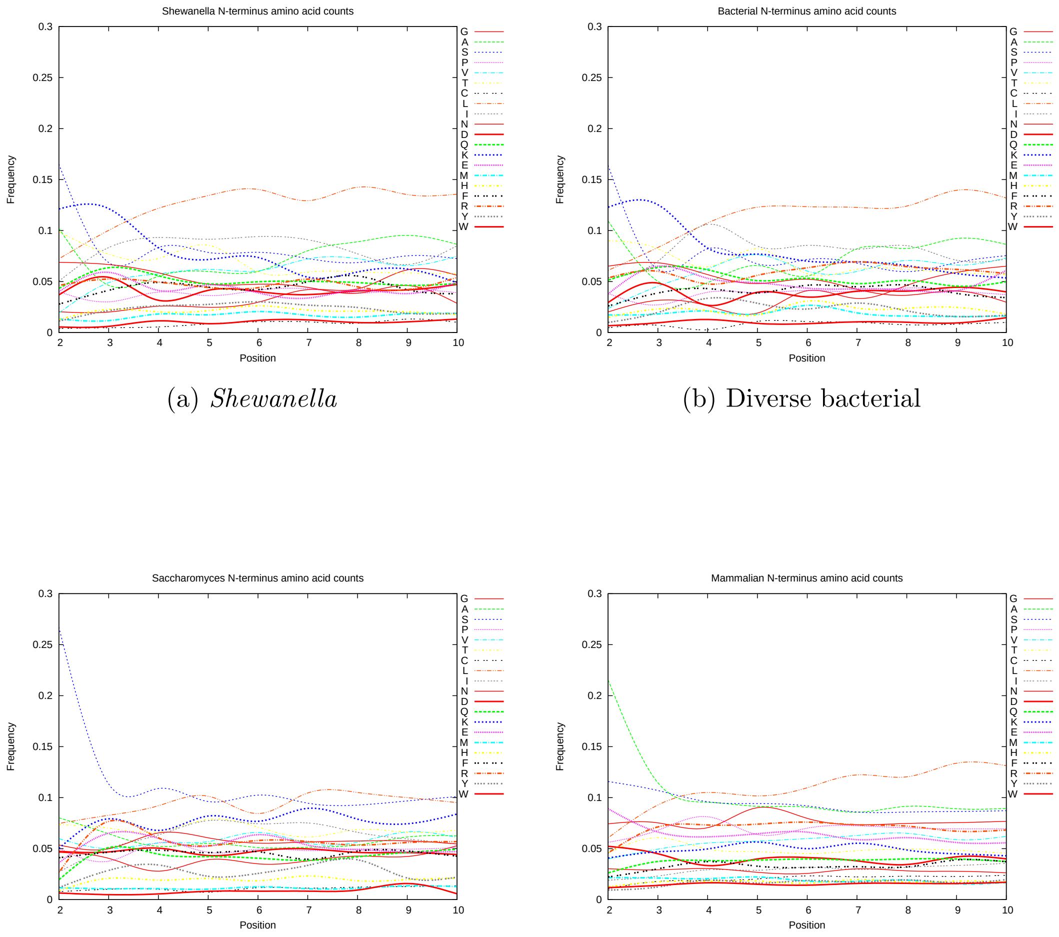
While an interesting pattern began to emerge in the previous section, further testing must be performed to determine if this pattern of Ala and Ser conservation in the second position is applicable to all species or merely the ones in the three data sets. While the three data sets do fall into three wide categories (bacterial, eukaryotic and vertebrae for *Shewanella*, *Saccharomyces* and mammalian data sets, respectively), a more diverse data set for experimentation is required. Further, the species within both the *Shewanella* and *Saccharomyces* data sets are closely related. More distantly related species should be included in these new data sets.

To this end, a diverse bacterial data set comprising of distantly related organisms is utilized. The organisms included are *Escherichia coli*, *Shewanella oneidensis*, *Vibrio parahaemolyticus*, *Vibrio vulnificus*, *Yersinia pestis* and *Vibrio cholerae*. Table 5.3 displays the counterpart of Table 5.1 for the new diverse bacterial data set. As is seen in Table 5.3, Ala is more conserved in position 2 than in position 3.

Table 5.3: Single amino acid breakdown for diverse bacterial data set.

AA	P2	P3
G	3	6
A	**36**	**5**
S	19	5
P	11	4
V	2	7
T	3	8
C	3	1
L	10	15
I	11	9
N	5	4
D	3	6
Q	5	6
K	20	21
E	5	10
M	0	1
H	1	3
F	5	3
R	7	13
Y	0	4
W	2	2

Tables 5.1 and 5.3 provide for an interesting picture of the composition of the N-terminus. To obtain a clearer notion of the amino acids that comprise the N-terminus of proteins among the four data sets (*Shewanella*, *Saccharomyces*, mammalian and diverse bacterial data sets), the frequency of each residue on the N-terminus is plotted in Figures 5.1.



(a) *Shewanella*

(b) Diverse bacterial

(c) *Saccharomyces*

(d) Mammalian

Figure 5.1: Count of amino acids for positions 2 through 10 along the N-terminus. Four data sets shown: (a) *Shewanella*, (b) Bacteria, (c) *Saccharomyces* and (d) Mammalian data sets, respectively. The *Shewanella*, Bacterial and *Saccharomyces* figures are all similar to one another with Ser being highly expressed at position 2, only to drop dramatically at positions farther from the N-terminus.

The lines in Figures 5.1(a), (b), (c) and (d) show that Ser (in three data sets) and Ala (in the fourth) have higher coverage only in position 2 among the first 10 positions along the N-terminus. This is in contrast to Lys, which shares a high frequency in Figures 5.1(a) and (b), but whose elevated frequency is in *both* positions 2 and 3. Since Lys is not one of the residues on which MAP excises methionine, it is possible that Lys is required for some other reason on the N-terminus. There is no such trend of position dependent elevations in frequency for the C-terminus (data not shown).

To further investigate the Ala/Ser elevation in frequency, a data set containing a broader range of organisms is used. Taken from the Uniprot database (www.uniprot.org) are 36 organisms with the most available protein sequences. These 36 organisms range from vertebrates to bacteria. For each one, the frequency of each amino acid in position 2 is plotted in Figure 5.2 below.

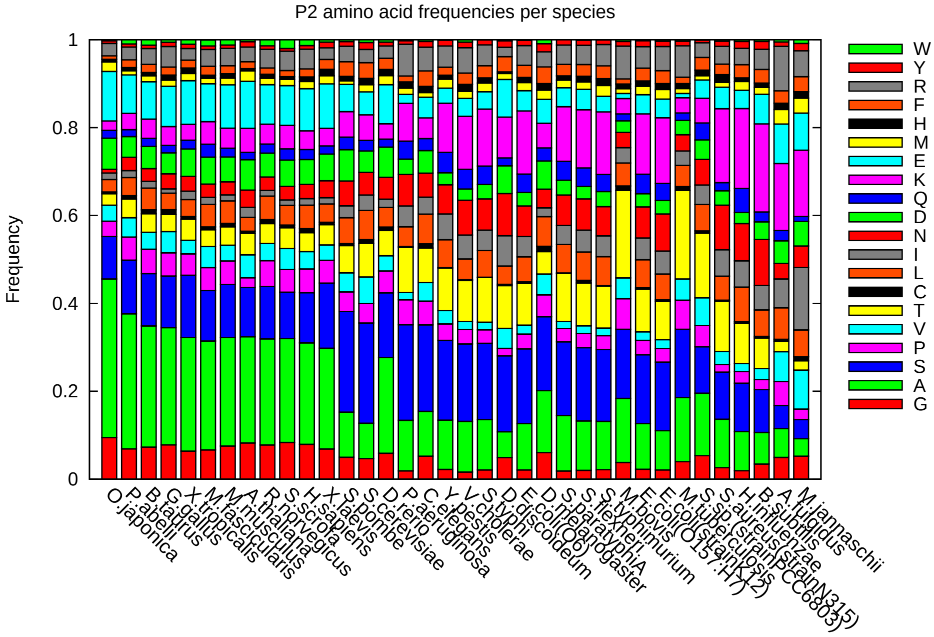


Figure 5.2: Frequency of amino acids in position 2 of all 36 considered species from Uniprot data. The species are ordered from left to right with respect to the frequency of Ala and Ser (the larger of the two frequencies).

The bars in Figure 5.2 are ordered according to the Ala or Ser frequency (the larger of the two for each organism). From this figure, it is clear that nearly all of the 36 species share this high Ala and Ser frequency in position two. All of the species except *Methanocaldococcus jannaschii* and *Archaeoglobus fulgidus* follow this tendency for high Ala/Ser frequency.

Table 5.4 displays these same 36 organisms sorted according to the ratio of frequency of amino acid in position 2 compared to position 3, as defined by Ratio_{P2P3}.

$$\text{Ratio}_{P2P3} = \begin{cases} \frac{f_{P2}}{f_{P3}} & : f_{P2} > f_{P3} \\ -\frac{f_{P3}}{f_{P2}} & : f_{P2} \leq f_{P3} \end{cases}$$

Each of the three columns in the table show the most frequent amino acid, along with the 2nd and 3rd most frequent. Each cell contains the frequency along with the amino acid and Ratio_{P2P3} in parenthesis. While Lys represents the most frequent amino acid in *Bacillus subtilis*, *Haemophilus influenzae*, and *Staphylococcus aureus*, Lys is equally frequent in the 3rd position in these species (as seen by their Ratio_{P2P3} values) thus indicating that it has little to do with NME. Ser and/or Ala, on the other hand, while showing smaller frequency than Lys in these species, show much high frequency in the 2nd position as compared to the 3rd position. Table 5.4 also reveals the elevated frequency of Thr (that can also be acetylated) in the 2nd position for some species *Mycobacterium tuberculosis*, *Mycobacterium bovis*, and *Synechocystis sp.*. For these species Ser and/or Ala also show similar contrast in frequency between the 2nd and the 3rd position.

5.2.2 N-acetylation and NME

NME has been linked to N-terminal post-translational modifications (PTMs), such as N-acetylations and N-myristoylations [Wal05, MTV^{+}08]. Some of these *post*-translational modifications, namely N-acetylation in eukaryotes [Pol00], can occur co-translationally just as NME does. One prevalent modification in eukaryotes is N-acetylation carried out by N-acetyltransferases (NATs). N-acetylation occurs in the majority of eukaryotic proteins (50% in yeast [LLS89], 80%-90% in higher eukaryotes [BR76, Bro79]), but rarely in prokaryotic proteins. In yeast, three N-acetyltransferase (NATA, NATB, NATC) are responsible for N-acetylation that act co-translationally, after NME has occurred. NATA targets N-termini with Ser, Ala, Gly, Thr, exposed after NME [Wal05]. The presence or absence of N-acetylation is important for the correct protein functioning [Pol00].

Table 5.4: Most frequent amino acids and their Ratio$_{P2P3}$. 36 species taken from Uniprot, sorted according to ratio of frequency of amino acid in position 2 comparted to position 3. Each of the first three columns shows the most frequent amino acid, along with the 2nd and 3rd most frequent. Each cell contains the frequency along with the amino acid and Ratio$_{P2P3}$ in parenthesis. Columns displaying the Ser and Ala frequency and Ratio$_{P2P3}$ are also shown.

Species	Most frequent	2nd frequent	3rd frequent	Ser	Ala
Arabidopsis thaliana	0.2419(A, 4.06)	0.1119(S, -1.32)	0.1071(E, 1.74)	0.2419(4.06)	0.1119(-1.32)
Pseudomonas aeruginosa	0.2175(S, 3.68)	0.1151(A, 1.15)	0.1024(T, 1.40)	0.1151(1.15)	0.2175(3.68)
Yersinia pestis	0.1816(S, 3.60)	0.1120(K, -1.05)	0.1120(A, 2.06)	0.1120(2.06)	0.1816(3.60)
Escherichia coli O6	0.1699(S, 3.37)	0.1435(K, 1.08)	0.1056(A, 2.75)	0.1056(2.75)	0.1699(3.37)
Shigella flexneri	0.1667(S, 3.26)	0.1324(K, -1.04)	0.1125(A, 2.75)	0.1125(2.75)	0.1667(3.26)
Salmonella typhi	0.1737(S, 3.05)	0.1295(K, -1.00)	0.1145(A, 2.43)	0.1145(2.43)	0.1737(3.05)
Vibrio cholerae	0.1763(S, 2.95)	0.1207(K, -1.22)	0.1154(A, 2.30)	0.1154(2.30)	0.1763(2.95)
Salmonella paratyphi A	0.1676(S, 2.92)	0.1265(A, 2.55)	0.1243(K, 1.07)	0.1265(2.55)	0.1676(2.92)
Salmonella typhimurium	0.1636(S, 2.89)	0.1424(K, 1.20)	0.1098(A, 2.31)	0.1098(2.31)	0.1636(2.89)
Escherichia coli O157:H7	0.1564(S, 2.75)	0.1373(K, 1.04)	0.1041(A, 2.56)	0.1041(2.56)	0.1564(2.75)
Escherichia coli (strain K12)	0.1559(S, 2.63)	0.1497(K, 1.22)	0.0892(A, 1.95)	0.0892(1.95)	0.1559(2.63)
Oryza sativa subsp. japonica	0.3608(A, 2.56)	0.1126(E, 2.23)	0.0968(S, -1.47)	0.3608(2.56)	0.0968(-1.47)
Mycobacterium tuberculosis	0.2016(T, 2.52)	0.1553(S, 2.14)	0.1461(A, 1.28)	0.1461(1.28)	0.1553(2.14)
Xenopus laevis	0.2295(A, 2.43)	0.1480(S, 1.26)	0.1013(E, 1.22)	0.2295(2.43)	0.1480(1.26)
Mycobacterium bovis	0.1995(T, 2.39)	0.1573(S, 2.41)	0.1457(A, 1.34)	0.1457(1.34)	0.1573(2.41)
Xenopus tropicalis	0.2584(A, 2.36)	0.1419(S, 1.28)	0.0994(E, 1.21)	0.2584(2.36)	0.1419(1.28)
Danio rerio	0.2180(A, 2.34)	0.1471(S, 1.21)	0.0890(E, 1.25)	0.2180(2.34)	0.1471(1.21)
Sus scrofa	0.2364(A, 2.30)	0.1054(S, -1.13)	0.0905(E, 1.69)	0.2364(2.30)	0.1054(-1.13)
Pongo abelii	0.3074(A, 2.27)	0.1222(S, -1.01)	0.0874(E, 1.16)	0.3074(2.27)	0.1222(-1.01)
Rattus norvegicus	0.2415(A, 2.26)	0.1192(S, 1.11)	0.0901(E, 1.42)	0.2415(2.26)	0.1192(1.11)
Gallus gallus	0.2666(A, 2.24)	0.1177(S, 1.16)	0.0915(E, 1.19)	0.2666(2.24)	0.1177(1.16)
Macaca fascicularis	0.2478(A, 2.18)	0.1146(S, 1.22)	0.0861(E, 1.17)	0.2478(2.18)	0.1146(1.22)
Mus musculus	0.2467(A, 2.07)	0.1209(S, 1.06)	0.0988(E, 1.49)	0.2467(2.07)	0.1209(1.06)
Bos taurus	0.2756(A, 2.06)	0.1193(S, 1.11)	0.0850(E, 1.25)	0.2756(2.06)	0.1193(1.11)
Homo sapiens	0.2309(A, 2.01)	0.1140(S, 1.08)	0.0959(E, 1.46)	0.2309(2.01)	0.1140(1.08)
Saccharomyces cerevisiae	0.2280(S, 1.79)	0.0804(A, 1.34)	0.0765(T, 1.25)	0.0804(1.34)	0.2280(1.79)
Schizosaccharomyces pombe	0.2290(S, 1.69)	0.1027(A, 1.82)	0.0713(D, 1.54)	0.1027(1.82)	0.2290(1.69)
Drosophila melanogaster	0.1682(S, 1.55)	0.1411(A, 1.63)	0.0672(L, -1.29)	0.1411(1.63)	0.1682(1.55)
Synechocystis sp. (strain PCC 6803)	0.1471(T, 1.51)	0.1422(A, 1.97)	0.1056(S, 1.03)	0.1422(1.97)	0.1056(1.03)
Dictyostelium discoideum	0.1728(S, 1.51)	0.0971(T, 1.28)	0.0961(N, -1.05)	0.0585(2.39)	0.1728(1.51)
Caenorhabditis elegans	0.1969(S, 1.46)	0.1021(A, 1.62)	0.0785(T, 1.13)	0.1021(1.62)	0.1969(1.46)
Archaeoglobus fulgidus	0.1525(K, 1.39)	0.1010(R, 1.33)	0.0899(E, -1.13)	0.0657(1.12)	0.0525(-1.23)
Bacillus subtilis	0.2005(K, 1.12)	0.1044(N, 1.52)	0.0973(S, 1.73)	0.0721(2.13)	0.0973(1.73)
Haemophilus influenzae	0.1817(K, 1.11)	0.1100(S, 2.21)	0.0920(T, 1.25)	0.0892(2.22)	0.1100(2.21)
Methanocaldococcus jannaschii	0.1510(K, 1.10)	0.1420(I, -1.07)	0.0887(V, 1.15)	0.0398(1.36)	0.0438(-1.00)
Staphylococcus aureus (strain N315)	0.1681(K, -1.29)	0.1146(T, 2.19)	0.1103(A, 3.60)	0.1103(3.60)	0.1070(2.58)

While acetylation is one of the most common modification in eukaryotes that may rival phosporylation in cell signalling [B. 02], the role of acetylation remains mysterious and the function of acetylation may be subtle and not absolute for most proteins [PNT^{+}99]. Moreover, revealing the subset of essential proteins for N-acetylation remains an open problem. Since we established that Ala and

Ser are exceptionally conserved in the second positions of Ala- and Ser-critical proteins, it is tempting to use MS/MS data for investigating the potential role of NME in acetylation. However, there are a few obstacles on this path. While some proteins require acetylation for proper functioning, other acetylated proteins do not absolutely require this modification. Polevoda and Sherman [B. 02] remark that N-terminal acetylation does not necessarily protect proteins from degradation, as has often been supposed, nor does it play any obvious role in protection of proteins from degradation by the N-End rule degradation pathway.

These complications are further compounded by the fact that only a small fraction of peptides starting in the 2nd position of proteins are detected by MS/MS experiments due to low *peptide detectability.* Moreover, the accuracy of MS/MS-based peptide identification tools further deteriorates while detecting modified (e.g., acetylated) peptides. Therefore, the acetylation status of most proteins cannot be inferred from a typical MS/MS experiment.

5.3 MS/MS data analysis

Mass spectrometry data for three *Shewanella* species and *S. cerevisiae* allows us to identify the types of PTMs and their amino acid targets at N-termini of proteins. N-acetylation is a common PTM in yeast proteins, with Ser, Ala, Gly and Thr being the acetylated residues when the initiating methionine is cleaved [Wal05]. Of the 106 N-acetylated sites found in *S. cerevisiae* by mass spectrometry, 87 are on Ser, 9 on Ala and 10 on Thr. Acetylation is viewed as a rare modification in bacteria, for example, in previous studies of *E.coli* proteins identified that only three are N-acetylated (ribosomal subunits S5, S18, and L12) [YISI, TMYI89]. Conspicuously, ribosomal subunits S5 and S18 also represent Ala-critical proteins in *Shewanella*, while L12 is a Ser-critical protein.

We emphasize that 106 N-acetylated proteins in *S. cerevisiae* are likely to represent only a fraction of N-acetylated proteins in *S. cerevisiae.* Among 2533 proteins found to be expressed in *S. cerevisiae*, 1218 have Ser (767), Ala (259), or Thr (192) at the 2nd position. However, for only 36 out of these 2533 proteins we identified (non-modified) peptides starting at the second position (9 of these 36 contain an unmodified Ala, Ser or Thr at position 2), an indication that these proteins are *not* N-acetylated. Therefore, the N-acetylation status of most yeast proteins (1218-106-9=1103) with Ser, Ala, or Thr in the 2nd position cannot be derived from the available MS/MS data. However, if the ratio 9:106 of non-modified:acetylated peptides (among all identified peptides starting with Ala, Ser, and Thr at the 2nd position of proteins) represents an accurate proxy for the ratio between unmodified and acetylated proteins (among all proteins with Ala, Ser, and Thr at the 2nd position) then a large fraction ($\approx 90\%$) of *S. cerevisiae* proteins with Ala, Ser, and Thr in the 2nd position are acetylated.

The analysis of a connection between Ala- and/or Ser-critical proteins and N-acetylation is further compounded by the fact that only 24 out of 106 N-acetylated proteins in yeast have sequelogs in all 7 yeast species (making it difficult to establish a connection due to a small sample size). Of these proteins, 16 are Ala/Ser conserved with 15 containing a Ser at position 2 and one with an Ala. A large fraction (43%) of 1502 *S. cerevisiae* proteins (with sequelogs in all 7 yeast species) have Ser, Ala, or Thr in the 2nd position (419 for Ser, 125 for Ala, and 114 for Thr). Among these 1502 proteins, 759 are found to be expressed in *S. cerevisiae* based on MS/MS analysis. The fact that only 24 such proteins *are found to be* N-acetylated represents a lower bound for the number of N-acetylated proteins rather than indicating that a small fraction of these 759 proteins are N-acetylated. Only 12 out of 759 such proteins have an *unmodified* peptide in the 2nd position identified by MS/MS, an indication that these 12 proteins are not N-acetylated. It demonstrates that the N-acetylation status for most of 759 expressed proteins ($\approx 95\%$) in *S. cerevisiae* (with sequelogs in all 7 yeast species) remains unknown.

Despite the limitations of mass spectrometry in revealing the full set of N-acetylated proteins, we nevertheless can evaluate the correlation between N-acetylated and Ser- and Ala-critical proteins. For simplicity, we limit our attention to proteins with Ser in the second position in *S. cerevisiae* (the numbers of identified Ala- and Thr-acetylated proteins are too small).

While only $\approx 31\%$ of 419 proteins with Ser at the 2nd position are Ser-critical ($130/419 \approx 0.31$), 70% of Ser-acetylated proteins are Ser-critical ($15/19 \approx 0.70$). This large contrast indicates that Ser-critical proteins may show larger propensity to being N-acetylated. Similarly, 87 out of 1438 such proteins in *S. cerevisiae* are N-acetylated resulting in $87/1438 \approx 6\%$ frequency of N-acetylated proteins among proteins with Ser at the 2nd position. Among Ser-critical proteins, the frequency of N-acetylated proteins is $15/130 \approx 11\%$.

We further limit our attention to 1502 *S. cerevisiae* proteins with sequelogs in all 7 yeast species and examine 419 proteins in this set with Ser at the 2nd position. 19 out of these 419 proteins are N-acetylated resulting in $19/419 \approx 5\%$ frequency of acetylation among proteins with Ser at the second position. Ser-critical proteins, on the other hand, have $15/130 \approx 11\%$ frequency of acetylation. Clearly, while our analysis suggest that Ser-critical proteins in *S. cerevisiae* have much higher propensity for acetylation, the MS/MS data provide an incomplete picture of N-terminal acetylations (due to limited peptide detectability) and we are able to draw only limited conclusions. However, it is clear that connecting NME to the protein half-life regulation based on the similarities in the specificities of NME and the N-End rule is not confirmed in this study, thus alternative explanations are needed.

Chapter 5 is in part also currently being prepared for submission for publication of the material. Bonissone, Stefano; Gupta, Nitin; Romine, Margaret; Pevzner, Pavel. The thesis author was the primary investigator and author of this material.

Table 5.5: Pairwise amino acid conservation for mixed residues in positions 2 and 3. Conservation of position 2 is to the left of the '/' within each cell, position 3 to the right. Given X denoting the row and Y denoting the column, each cell (X, Y) in the table contains the number of proteins conserved for $\mathcal{X} = \{X, Y\} - \{X\} - \{Y\}$

(a) 19 *Shewanella* species

	G	A	P	S	T	V	C	N	D	L	I	H	Q	E	F	M	K	Y	W	R	Σ
G	0/0	4/0	0/0	11/2	1/2	1/1	0/0	0/1	1/0	0/1	0/1	0/0	0/2	1/1	0/0	0/0	1/1	0/0	0/0	0/1	20/13
A	4/0	0/0	7/3	23/2	16/10	2/1	0/0	0/1	0/1	3/5	1/0	0/0	0/2	1/2	1/0	2/2	0/2	1/0	1/0	2/0	**64/31**
P	0/0	7/3	0/0	8/1	3/0	0/0	0/0	2/0	0/1	6/3	1/3	0/1	1/0	0/2	0/1	1/0	1/0	0/0	0/0	1/2	31/17
S	11/2	23/2	8/1	0/0	31/16	2/1	1/0	11/8	4/4	7/5	1/2	0/1	1/5	3/4	2/0	2/0	5/5	0/1	1/0	4/1	**117/58**
T	1/2	16/10	3/0	31/16	0/0	0/1	0/1	7/2	1/1	2/1	2/3	1/0	1/1	0/0	0/2	3/3	6/7	0/1	0/0	1/0	75/51
V	1/1	2/1	0/0	2/1	0/1	0/0	0/1	2/1	1/0	2/6	6/20	1/0	0/0	2/1	0/1	0/1	4/2	1/0	0/0	1/1	25/38
C	0/0	0/0	0/0	1/0	0/1	0/1	0/0	1/0	0/0	0/1	0/0	0/0	0/1	0/0	1/0	0/0	0/0	0/0	0/0	0/0	3/4
N	0/1	0/1	2/0	11/8	7/2	2/1	1/0	0/0	5/10	3/0	1/2	2/3	1/2	0/2	0/1	2/1	23/21	0/0	0/0	2/0	62/55
D	1/0	0/1	0/1	4/4	1/1	1/0	0/0	5/10	0/0	1/1	1/1	0/0	1/3	19/27	0/1	1/0	2/4	0/1	0/1	0/0	37/56
L	0/1	3/5	6/3	7/5	2/1	2/6	0/1	3/0	1/1	0/0	9/13	0/0	0/5	4/1	11/7	1/0	1/2	1/1	2/0	4/3	57/55
I	0/1	1/0	1/3	1/2	2/3	6/20	0/0	1/2	1/1	9/13	0/0	0/0	1/0	0/2	2/3	1/3	5/6	2/1	0/0	0/4	33/64
H	0/0	0/0	0/1	0/1	1/0	1/0	0/0	2/3	0/0	0/0	0/0	0/0	2/3	0/1	1/0	0/0	0/0	1/1	0/0	0/1	8/11
Q	0/2	0/2	1/0	1/5	1/1	0/0	0/1	1/2	1/3	0/5	1/0	2/3	0/0	2/10	1/0	0/1	8/6	1/1	0/0	1/1	21/43
E	1/1	1/2	0/2	3/4	0/0	2/1	0/0	0/2	19/27	4/1	0/2	0/1	2/10	0/0	2/0	0/1	4/3	0/0	0/0	1/1	39/58
F	0/0	1/0	0/1	2/0	0/2	0/1	1/0	0/1	0/1	11/7	2/3	1/0	1/0	2/0	0/0	1/0	1/4	5/4	1/0	1/0	30/24
M	0/0	2/2	1/0	2/0	3/3	0/1	0/0	2/1	1/0	1/0	1/3	0/0	0/1	0/1	1/0	0/0	3/6	0/0	0/0	0/0	17/18
K	1/1	0/2	1/0	5/5	6/7	4/2	0/0	23/21	2/4	1/2	5/6	0/0	8/6	4/3	1/4	3/6	0/0	0/2	0/1	18/18	82/90
Y	0/0	1/0	0/0	0/1	0/1	1/0	0/0	0/0	0/1	1/1	2/1	1/1	1/1	0/0	5/4	0/0	0/2	0/0	0/0	1/0	13/13
W	0/0	1/0	0/0	1/0	0/0	0/0	0/0	0/0	0/1	2/0	0/0	0/0	0/0	0/0	1/0	0/0	0/1	0/0	0/0	1/0	6/2
R	0/1	2/0	1/2	4/1	1/0	1/1	0/0	2/0	0/0	4/3	0/4	0/1	1/1	1/1	1/0	0/0	18/18	1/0	1/0	0/0	38/33

Table 5.5: Pairwise amino acid conservation for mixed residues in positions 2 and 3. Conservation of position 2 is to the left of the '/' within each cell, position 3 to the right. Given X denoting the row and Y denoting the column, each cell (X, Y) in the table contains the number of proteins conserved for $\mathcal{X} = \{X, Y\} - \{X\} - \{Y\}$

(b) 7 *Saccharomyces* species

	G	A	P	S	T	V	C	N	D	L	I	H	Q	E	F	M	K	Y	W	R	Σ
G	0/0	9/2	3/2	15/3	1/0	5/0	1/2	2/2	2/1	0/3	1/2	0/0	0/1	0/1	1/0	0/0	1/2	0/2	0/0	1/3	42/26
A	9/2	0/0	7/6	26/11	6/7	8/5	0/0	2/1	3/4	0/4	1/1	1/1	0/1	2/6	0/0	0/0	3/4	0/0	0/0	1/0	69/53
P	3/2	7/6	0/0	25/4	2/0	5/1	1/1	1/1	1/2	1/5	2/3	1/1	0/0	0/3	2/0	0/0	1/0	0/0	1/0	1/1	54/30
S	15/3	26/11	25/4	0/0	32/17	12/4	1/1	14/6	8/3	10/10	6/2	3/3	5/11	11/7	13/4	2/1	8/7	4/3	1/0	6/9	**202/106**
T	1/0	6/7	2/0	32/17	0/0	3/1	0/0	6/6	3/2	1/2	2/4	0/0	0/0	2/6	1/2	0/0	2/4	1/1	0/0	0/3	62/55
V	5/0	8/5	5/1	12/4	3/1	0/0	0/1	1/0	0/1	5/7	3/12	2/0	1/1	0/3	2/2	2/1	1/0	0/0	0/0	0/0	50/39
C	1/2	0/0	1/1	1/1	0/0	0/1	0/0	2/0	0/0	0/0	0/0	0/0	0/1	0/0	1/1	0/0	1/0	0/0	0/0	0/0	7/7
N	2/2	2/1	1/1	14/6	6/6	1/0	2/0	0/0	7/3	3/3	0/1	0/0	0/1	2/2	1/1	0/0	0/4	1/0	0/0	0/1	42/32
D	2/1	3/4	1/2	8/3	3/2	0/1	0/0	7/3	0/0	3/0	0/0	0/1	0/2	7/7	1/0	0/1	0/2	0/0	0/0	1/0	36/29
L	0/3	0/4	1/5	10/10	1/2	5/7	0/0	3/3	3/0	0/0	9/5	0/1	1/3	0/4	16/18	0/1	5/2	1/1	4/0	3/7	62/76
I	1/2	1/1	2/3	6/2	2/4	3/12	0/0	0/1	0/0	9/5	0/0	0/1	0/1	0/2	0/1	0/0	1/3	1/1	1/0	0/0	27/39
H	0/0	1/1	1/1	3/3	0/0	2/0	0/0	0/0	0/1	0/1	0/1	0/0	0/2	1/1	0/0	0/0	0/1	0/5	0/0	0/1	8/18
Q	0/1	0/1	0/0	5/11	0/0	1/1	0/1	0/1	0/2	1/3	0/1	0/2	0/0	1/1	1/3	1/1	2/4	0/2	0/0	0/3	12/38
E	0/1	2/6	0/3	11/7	2/6	0/3	0/0	2/2	7/7	0/4	0/2	1/1	1/1	0/0	0/1	0/0	0/2	0/0	0/0	0/1	26/47
F	1/0	0/0	2/0	13/4	1/2	2/2	1/1	1/1	1/0	16/18	0/1	0/0	1/3	0/1	0/0	0/1	1/2	2/4	0/0	1/1	43/41
M	0/0	0/0	0/0	2/1	0/0	2/1	0/0	0/0	0/1	0/1	0/0	0/0	1/1	0/0	0/1	0/0	0/0	0/0	0/0	0/0	5/6
K	1/2	3/4	1/0	8/7	2/4	1/0	1/0	0/4	0/2	5/2	1/3	0/1	2/4	0/2	1/2	0/0	0/0	0/0	0/0	10/23	36/60
Y	0/2	0/0	0/0	4/3	1/1	0/0	0/0	1/0	0/0	1/1	1/1	0/5	0/2	0/0	2/4	0/0	0/0	0/0	0/0	0/1	10/20
W	0/0	0/0	1/0	1/0	0/0	0/0	0/0	0/0	0/0	4/0	1/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	7/0
R	1/3	1/0	1/1	6/9	0/3	0/0	0/0	0/1	1/0	3/7	0/0	0/1	0/3	0/1	1/1	0/0	10/23	0/1	0/0	0/0	24/54

Table 5.5: Pairwise amino acid conservation for positions 2 and 3. Conservation of position 2 is to the left of the '/' within each cell, position 3 to the right. Given X denoting the row and Y denoting the column, each cell (X, Y) in the table contains the number of proteins conserved for $\mathcal{X} = \{X, Y\} - \{X\} - \{Y\}$

(c) 6 mammalian species

	G	A	P	S	T	V	C	N	D	L	I	H	Q	E	F	M	K	Y	W	R	Σ
G	0/0	88/60	23/31	41/51	19/15	11/22	5/3	11/8	22/17	28/27	13/5	4/5	12/11	38/37	8/8	7/4	14/19	2/6	3/4	14/44	363/377
A	88/60	0/0	63/46	152/87	79/71	79/55	8/10	23/13	37/18	89/44	26/10	13/6	28/18	93/35	22/14	57/18	36/13	9/5	11/3	49/24	**962/550**
P	23/31	63/46	0/0	50/72	10/22	14/16	4/5	8/3	17/7	26/51	3/7	3/13	14/12	18/16	8/8	9/4	9/17	4/2	4/3	13/26	300/361
S	41/51	152/87	50/72	0/0	41/44	17/19	10/20	36/28	18/19	41/59	11/9	7/5	19/15	38/16	21/21	15/14	21/20	8/6	7/6	24/31	577/542
T	19/15	79/71	10/22	41/44	0/0	6/14	1/5	17/12	9/9	13/20	12/10	3/4	6/6	21/12	4/5	11/21	10/11	3/3	1/3	7/16	273/303
V	11/22	79/55	14/16	17/19	6/14	0/0	2/8	3/10	10/9	24/41	4/13	0/5	7/10	13/13	8/8	12/11	7/7	2/3	2/0	9/15	230/279
C	5/3	8/10	4/5	10/20	1/5	2/8	0/0	1/8	1/4	3/14	1/3	1/2	0/3	4/7	0/3	1/0	3/1	4/3	3/2	6/4	58/105
N	11/8	23/13	8/3	36/28	17/12	3/10	1/8	0/0	17/15	8/14	6/0	1/8	3/5	10/2	2/5	6/6	7/7	0/1	1/2	1/3	161/150
D	22/17	37/18	17/7	18/19	9/9	10/9	1/4	17/15	0/0	12/17	8/3	2/3	5/10	69/57	8/7	7/7	8/11	1/2	0/1	10/13	261/229
L	28/27	89/44	26/51	41/59	13/20	24/41	3/14	8/14	12/17	0/0	7/20	3/7	18/17	28/24	20/32	15/17	18/17	2/6	8/7	14/29	377/463
I	13/5	26/10	3/7	11/9	12/10	4/13	1/3	6/0	8/3	7/20	0/0	1/0	2/7	11/8	2/4	3/5	7/8	0/0	0/2	3/7	120/121
H	4/5	13/6	3/13	7/5	3/4	0/5	1/2	1/8	2/3	3/7	1/0	0/0	8/12	10/4	1/2	1/3	3/10	2/3	0/1	7/10	70/103
Q	12/11	28/18	14/12	19/15	6/6	7/10	0/3	3/5	5/10	18/17	2/7	8/12	0/0	16/25	5/4	3/3	7/13	1/1	2/6	10/23	166/201
E	38/37	93/35	18/16	38/16	21/12	13/13	4/7	10/2	69/57	28/24	11/8	10/4	16/25	0/0	15/7	22/7	12/21	2/2	4/4	18/23	442/320
F	8/8	22/14	8/8	21/21	4/5	8/8	0/3	2/5	8/7	20/32	2/4	1/2	5/4	15/7	0/0	6/2	2/11	5/10	2/1	4/15	143/167
M	7/4	57/18	9/4	15/14	11/21	12/11	1/0	6/6	7/7	15/17	3/5	1/3	3/3	22/7	6/2	0/0	5/3	0/0	1/0	5/4	186/129
K	14/19	36/13	9/17	21/20	10/11	7/7	3/1	7/7	8/11	18/17	7/8	3/10	7/13	12/21	2/11	5/3	0/0	0/2	0/1	35/32	204/224
Y	2/6	9/5	4/2	8/6	3/3	2/3	4/3	0/1	1/2	2/6	0/0	2/3	1/1	2/2	5/10	0/0	0/2	0/0	0/1	2/5	47/61
W	3/4	11/3	4/3	7/6	1/3	2/0	3/2	1/2	0/1	8/7	0/2	0/1	2/6	4/4	2/1	1/0	0/1	0/1	0/0	6/3	55/50
R	14/44	49/24	13/26	24/31	7/16	9/15	6/4	1/3	10/13	14/29	3/7	7/10	10/23	18/23	4/15	5/4	35/32	2/5	6/3	0/0	237/327

Chapter 6

Conclusions

6.1 Discussion of Analysis

Chapter 2 shows the importance of NME through the conservation of $\mathcal{X}$-residues. The proxy for experimentally labeled NME proteins, NME-proteins, shows that the simple rule works well at predicting NME. The `SimpleConservation` and `SetConservation` metrics are created to show the importance of NME via conservation of amino acids. It is shown that the conservation of $\mathcal{X}$-residues in the second position requires deeper investigation and that the connection between NME and NME-proteins can be made.

Chapter 3 makes evident the fact that there is a set of proteins which have an affinity for NME across many species. This chapter attempted to identify the set of proteins which require NME. The NME necessary tables for *Shewanella*, *Saccharomyces* and mammalian organisms highlight a pattern. The first attempts at identifying the necessary proteins using groupings of orthologous proteins proved non-trivial. Other potential avenues are suggested.

Chapter 4 tested the connection of NME-proteins to their half-lives. The surprising lack of correlation is contrary to previous papers on NME. Analysis of NME-proteins and operons also failed to reveal a connection, showing that NME is not used as a form of regulation. These findings raise the question of the role of NME.

Chapter 5 explores one possible connection for NME, relating to other post-translational modifications. The preliminary work described shows that there is a potential connection, however no strong claim can be made at this point. The apparent universal importance of Ala and Ser for position 2 provides an interesting clue for further investigation. As put forth in Chapter 5, acetylation may be necessary for these Ala/Ser-critical proteins. Because of the propensity for Ala and Ser to be acetylated, along with these two residues comprising the most conserved residues in the $\mathcal{X}$-residue set, it is suggested that NME's function is to expose these two residues for acetylation.

While the true function of NME cannot be definitively stated in this work, its connection to protein degradation has been placed into question. The lack of correlation in *S.cerevisiae* between NME-proteins and protein half-life caused a search for other explanations. It was the hope that identification of NME-critical proteins could provide for some clues to NME's true function. Unfortunately a different approach is needed than the one described for uncovering such a set of NME-critical proteins.

6.2 Potential Directions

Identification of NME necessary proteins could provide information to determine the true function of NME. Additional methods of computationally extracting the set of proteins requiring NME is one avenue worth exploring. One approach to this could be to utilize the Gene Ontology hierarchy to draw conclusions about conserved proteins in different species. Should groupings of conserved NME-proteins appear in closely related parts of the biological process and molecular function ontologies, this could show a connection between NME and a specific type of process or function.

Another avenue to pursue is the connection between NME and other post-translational modifications. Chapter 5 shows some preliminary results, but more data is required to draw stronger conclusions.

Bibliography

[AB88] S. Arfin and R. Bradshaw. Cotranslational processing and protein turnover in eukaryotic cells. *Biochemistry*, 27:7979–7984, 1988.

[B. 02] F. Sherman B. Polevoda. The diversity of acetylated proteins. *Genome Biol.*, 3:59–107, 2002.

[BR76] J. Brown and W. Roberts. Evidence that approximately eighty per cent of the soluble proteins from ehrlich ascites cells are nalpha-acetylated. *J. Biol. Chem.*, 251:1009–1014, 1976.

[Bro79] J. Brown. A comparison of the turnover of alpha-n-acetylated and nonacetylated mouse l-cell proteins. *J. Biol. Chem.*, 254:1447–1449, 1979.

[BTB$^+$06] A. Belle, A. Tanay, L. Bitincka, R. Shamir, and E. O'Shea. Quantification of protein half-lives in the budding yeast proteome. *PNAS*, 103:13004–13009, 2006.

[CL08] Y. Cai and L. Lu. Predicting n-terminal acetylation based on feature selection method. *Biochemical and biophysical research communications*, 372:862–865, 2008.

[CVC02] S. Chen, J. Vetro, and Y. Chang. The specificity in vivo of two distinct methionine aminopeptidases in saccharomyces cerevisiae. *Archives of Biochemistry and Biophysics*, 398:87–93, 2002.

[FMP$^+$06] F. Frottin, A. Martinez, P. Peynot, S. Mitra, R. Holz, C. Giglione, and T. Meinnel. The proteomics of n-terminal methionine cleavage. *Mol Cell Proteomics*, 5:2336–2349, 2006.

[GBB$^+$08] N. Gupta, J. Benhamida, V. Bhargava, D. Goodman, E. Kain, I. Kerman, N. Nguyen, N. Ollikainen, J. Rodriguez, J. Wang, M. Lipton, M. Romine, V. Bafna, R. Smith, and P. Pevzner. Comparative proteogenomics: Combining mass spectrometry and comparative genomics to analyze multiple genomes. *Genome Res.*, 18:1133–1142, 2008.

[GBM04] C. Giglione, A. Boularot, and T. Meinnel. Protein n-terminal methionine excision. *Cellular and Molecular Life Sciences*, 61:1455–1474, 2004.

[GM01] C. Giglione and T. Meinnel. Organellar peptide deformylases: universality of the n-terminal methionine cleavage mechanism. *Trends in Plant Science*, 6:566–572, 2001.

[GTJ$^+$07] N. Gupta, S. Tanner, N. Jaitly, J. Adkins, M. Lipton, R. Edwards, M. Romine, A. Osterman, V. Bafna, R. Smith, and P. Pevzner. Whole proteome analysis of post-translational modifications: Applications of mass-spectrometry for proteogenomic annotation. *Genome Res.*, 17:1362–1377, 2007.

[GVM03] C. Giglione, O. Vallon, and T. Meinnel. Control of protein life-span by n-terminal methionine excision. *EMBO J*, 22:13–23, 2003.

[HPG$^+$99] C. Hutchison, S. Peterson, S. Gill, R. Cline, O. White, C. Fraser, H. Smith, and J. Venter. Global transposon mutagenesis and a minimal mycoplasma genome. *Science*, 286:2165–2169, 1999.

[HSD$^+$89] P. Hirel, M. Schmitter, P. Dessen, G. Fayat, and S. Blanquet. Extent of n-terminal methionine excision from escherichia coli proteins is governed by the side-chain length of the penultimate amino acid. *Proceedings of the National Academy of Sciences of the United States of America*, 86:8247–8251, 1989.

[KBB05] L. Kiemer, J. Bendtsen, and N. Blom. NetAcet: prediction of N-terminal acetylation sites. *Bioinformatics*, 21:1269–1270, 2005.

[KYM$^+$07] S. Khalouei, X. Yao, J. Mennigen, M. Carullo, P. Ma, Z. Song, H. Xiong, and X. Xia. Bioinformatic approach to identify penultimate amino acids efficient for n-terminal methionine excision. In *Bioinformatics and Biomedical Engineering, 2007. ICBBE 2007. The 1st International Conference on*, pages 386–389, 2007.

[LC95] X. Li and Y. Chang. Amino-terminal protein processing in saccharomyces cerevisiae is an essential function that requires two distinct methionine aminopeptidases. *Proc Natl Acad Sci U S A*, 92:12357–12361, 1995.

[LLS89] F. Lee, L. Lin, and J. Smith. N[alpha]-acetyltransferase deficiency alters protein synthesis in saccharomyces cerevisiae. *FEBS Letters*, 256:139–142, 1989.

[LSBG03] P. Lord, R. Stevens, A. Brass, and A. Goble. Semantic similarity measures as tools for exploring the gene ontology. *Pac Symp Biocomput*, pages 601–612, 2003.

[MTV$^+$08] A. Martinez, J. Traverso, B. Valot, M. Ferro, C. Espagne, G. Ephritikhine, M. Zivy, C. Giglione, and T. Meinnel. Extent of n-terminal modifications in cytosolic proteins from eukaryotes. *Proteomics*, 8:2809–2831, 2008.

[PNT$^+$99] B. Polevoda, J. Norbeck, H. Takakura, A. Blomberg, and F. Sherman. Identification and specificities of n-terminal acetyltransferases from saccharomyces cerevisiae. *EMBO J*, 18:6155–6168, 1999.

[Pol00] S. Polevoda B. Nalpha -terminal acetylation of eukaryotic proteins. *J Biol Chem.*, 275:4635–4639, 2000.

[RM93] S. Roderick and B. Matthews. Structure of the cobalt-dependent methionine aminopeptidase from escherichia coli: a new type of proteolytic enzyme. *Biochemistry*, 32:3907–3912, 1993.

[SDRL06] A. Schlicker, F. Domingues, J. Rahnenfuhrer, and T. Lengauer. A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, 7:302+, 2006.

[SQGD08] B. Sheehan, A. Quigley, B. Gaudin, and S. Dobson. A relation based measure of semantic similarity for gene ontology annotations. *BMC Bioinformatics*, 9:468+, 2008.

[SST85] F. Sherman, J. Stewart, and S. Tsunasawa. Methionine or not methionine at the beginning of a protein. *BioEssays*, 3:27–31, 1985.

[TMYI89] S. Tanaka, Y. Matsushita, A. Yoshikawa, and K. Isono. Cloning and molecular characterization of the gene riml which encodes an enzyme acetylating ribosomal protein l12 of escherichia coli k12. *Molecular & general genetics*, pages 289–293, 1989.

[Var96] A. Varshavsky. The N-end rule: functions, mysteries, uses. *Proceedings of the National Academy of Sciences of the United States of America*, 93:12142–12149, 1996.

[Var04] A. Varshavsky. ‘Spalog’ and ‘sequelog’: neutral terms for spatial and sequence similarity. *Current Biology*, 14:R181–R183, 2004.

[Wal05] C. Walsh. *Posttranslational Modification of Proteins. Expanding Nature’s Inventory.* Roberts and Co. Publishers, 2005.

[YISI] A. Yoshikawa, S. Isono, A. Sheback, and K. Isono. Cloning and nucleotide sequencing of the genes rimi and rimj which encode enzymes acetylating ribosomal proteins s18 and s5 of escherichia coli k12. *Molecular and General Genetics*, 209:481–488.