Bioinformatic approach to identify penultimate amino acids efficient for N-terminal methionine excision

Sam Khalouei, Xiaoquan Yao, Jan Mennigen, Malisa Carullo,
Pinchao Ma, Ziyu Song, Huiling Xiong and Xuhua Xia
Department of Biology, University of Ottawa, 30 Marie Curie, Ottawa, Ontario, Canada, K1N 6N5

Abstract

More than half of proteins in prokaryotes and eukaryotes undergo N-terminal methionine excision (NME). While it is known that the penultimate amino acid affects the efficiency of NME in several bacterial and eukaryotic species, it is experimentally difficult and tedious to verify which amino acid at the penultimate site (the site after initiator Met) is the most efficient for NME in different species. Here we present a new bioinformatic approach to identify penultimate amino acids that are efficient for NME. Amino acids most efficient for NME are alanine, serine and proline in human, and alanine, glycine, valine, proline and serine in the yeast Saccharomyces cerevisiae. *This finding also helps resolve the two hypotheses that have been proposed to explain the presence of +4G site in the Kozak consensus for translation initiation.*

Keywords: N-terminal methionine excision, penultimate amino acid, posttranslational modification

1. Introduction

N-terminal modifications of nascent peptides occur in in more than half of proteins in both prokaryotes and eukaryotes [1-4] as well as in mitochondria and plastids [2]. N-terminal methionine excision (NME), which occurs soon after the amino terminus of the growing polypeptide chain emerges from the ribosome, is not only an important amino-terminal modification in itself, but also required for further N-terminal modifications. For example, it is required for myristoylation where glycine at the amino terminus, after the removal of the initiator methionine, is needed to attach to a myristoyl (C14H28O2) fatty acid side chain [5].

NME is carried out by methionine aminopeptidase (MAP). Eubacteria contain only one type of MAP whereas eukaryotes contain two (MAP1 and MAP2). The efficiency of NME depends heavily on the penultimate (the second) amino acid. In the yeast, *Saccharomyces cerevisiae*, NME occurs most efficiently when the penultimate amino acid is small [6]. Such studies have contributed significantly to the understanding of not only NME itself, but also eukaryotic translation initiation.

The optimal context for translation initiation in mammalian species is GCC**R**CCaug**G** (where R = purine and “aug” is the initiation codon), with the -3R and +4G being particularly important [7, 8]. The presence of +4G has been interpreted as necessary for efficient translation initiation [8, 9], and this interpretation is featured in virtually all textbooks of molecular biology. The finding that a small amino acid is needed to facilitate NME leads to an alternative hypothesis invoking amino acid constraint [10]. Because alanine and glycine happen to be the smallest amino acids, we expect the initiator Met to be followed often by alanine or glycine. The resulting overuse of Ala and Gly codons (GCN and GGN) following the initiation codon AUG leads to the prevalence of +4G in protein-coding genes. An extensive bioinformatic study [10] lends strong support for this alternative hypothesis.

In spite of the scientific importance of NME, characterizing the efficiency of NME conferred by different amino acids at the penultimate site in different species is experimentally difficult and tedious. For this reason, only a few studies have been carried out in Escherichia coli [11], Saccharomyces cerevisiae [6] or both [12]. In this paper we propose a bioinformatic approach to characterize the efficiency of NME conferred by different amino acids at the penultimate site and apply the method to the study of human and yeast proteins.

2. Methods and Materials

For a genome with N protein-coding genes (representing nascent proteins before any N-terminal processing), there are N amino acids at the penultimate site, with its frequency distribution specified by N_i where i = 1, 2, …, 20 corresponding to the 20 amino acids. Suppose we have M proteins known to undergo NME, with its frequency distribution specified by M_i. Define $p_i = N_i/N$ and $q_i = M_i/M$. If all amino acids at the penultimate site lead to equal efficiency in NME, then we expect $q_i = p_i$. If amino acid i is very conducive to NME, then $q_i > p_i$ and vice versa.

We use $E_{NME\cdot i}$ defined below as a quantitative measure of the NME efficiency for amino acid i

$$E_{NME.i} = \log_2 \frac{q_i}{p_i} = \log_2 \frac{M_i}{N_i} + \log_2 \frac{N}{M} \qquad (1)$$

1-4244-1120-3/07/$25.00 ©2007 IEEE

The interpretation of $E_{NME \cdot i}$ is straightforward: $E_{NME \cdot i} = 0$ when $q_i = p_i$, $E_{NME \cdot i} > 0$ when $q_i > p_i$; $E_{NME \cdot i} < 0$ when $q_i < p_i$.

To obtain N_i for human, we retrieved the rna.gbk.gz file at ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/RNA/, dated Sept. 3, 2006, and extracted all 34169 annotated coding sequences (CDSs). For the yeast, we retrieved the orf_coding.fasta from NCBI which contained 5888 yeast CDSs. Translating these CDSs to amino acid sequences and computing the N_i values were done with DAMBE [13, 14].

To obtain N_i for human, we retrieved the rna.gbk.gz file at ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/RNA/, dated Sept. 3, 2006, and extracted all 34169 annotated coding sequences (CDSs). For the yeast, we retrieved the orf_coding.fasta from NCBI which contained 5888 yeast CDSs. Translating these CDSs to amino acid sequences and computing the N_i values were done with DAMBE [13, 14].

To obtain M_i for human and yeast, we extracted all the N-terminal methionine-cleaved protein sequences from the UniProtKB/Swiss-Prot database [15], by using the search interface at (http://us.expasy.org/srs5). We limited the results to reviewed sequences in Swiss-Prot database by excluding the computationally annotated sequences in TrEMBL database. *Saccharomyces cerevisiae* and *Homo sapiens* were used respectively as the species name, and “INIT_MET” as the “FtKey (Feature)” in the info menu. According to the Swiss-Prot specifications, INIT_MET indicates evidence of NME. From a total of 6093 Swiss-Prot yeast protein sequences, 267 of them contained evidence for initiator methionine cleavage. These 267 sequences were manually inspected. Those proteins without direct experimental evidence and being flagged as "Potential", "Probable", or "By similarity" are not excluded. The remaining 232 proteins were experimentally verified to undergo NME. These 232 sequences were downloaded in FASTA format and the penultimate amino acid frequency was obtained using DAMBE [13, 14]. The same is done for human with 484 proteins having experimental evidence of NME.

3. Results and Discussion

For yeast proteins, $q_i > p_i$ for alanine (Ala), glycine (Gly), valine (Val), proline (Pro), serine (Ser) and threonine (Thr), i.e., these amino acids are overrepresented in the proteins that have undergone NME (Figure 1). The list of amino acids, ranked by their $E_{NME \cdot i}$ values, is shown in columns 2-4 in Table 1 which includes details of calculating $E_{NME \cdot i}$ values. The six amino acids with positive $E_{NME \cdot i}$ values (Ala, Gly, Val, Pro, Ser and Thr), with radii of gyration of 1.29 Å or less, were also found experimentally to result in complete cleavage of initiator Met in yeast [6]. Cys was also found previously to result in complete cleavage of the initiator Met [6], but no protein with Cys at the penultimate site is represented in our yeast proteins with known experimental evidence of NME. This finding, however, does not reject the hypothesis that Cys is as efficient as other six amino acids with positive $E_{NME \cdot i}$ values because of the rarity of Cys at the penultimate site in proteins encoded in the yeast genome (only 38 out of 5888 proteins encoded in yeast genome, Table 1). We need more data to evaluate whether Cys at the penultimate site can lead to efficient NME *in vivo* in yeast.

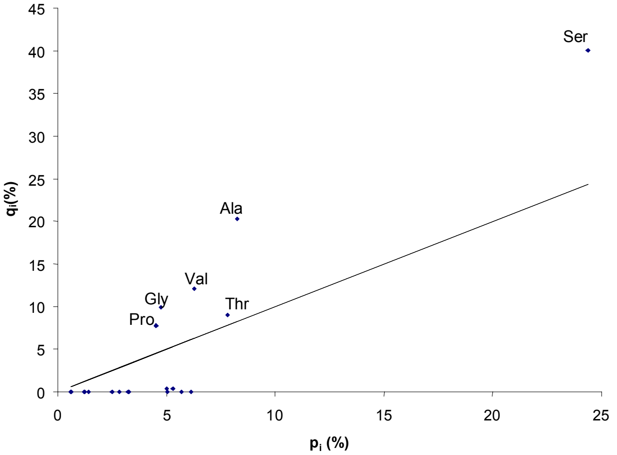

Figure 1. Different amino acids at the penultimate site result in dramatically different NME Efficiencies in yeast proteins. p_i and q_i are proportions of amino acid i at the penultimate site of unprocessed proteins and NME-processed proteins, respectively. The line indicates the position when $q_i = p_i$.

Cys is rare not only at the penultimate site but also at other sites in yeast proteins, accounting for only 0.645% of the amino acids at the penultimate and 1.263% at other sites excluding the first two sites. This may be related to the fact that the *S. cerevisiae* genome contains only 4 $tRNA^{Cys}$ genes, the fewest among all 20 amino acids. Because the number of tRNA genes in the genome is strongly correlated with the tRNA concentration in the cell [16-19], the few $tRNA^{Cys}$ genes implies relatively few $tRNA^{Cys}$ in the cytoplasm to translate Cys codons. This should result in selection against Cys usage in proteins because its usage would decrease translation efficiency. It has been theoretically suggested and empirically substantiated that tRNA availability is linearly correlated with the square root of amino acid usage [20]. Because translation initiation is often the rate-limiting step during protein translation [21, 22], it is disadvantageous to code for an amino acid that is slow to translate.

Ser is the most frequently found amino acid at the penultimate site (accounting for 24.4% of all amino acids at the penultimate site, Table 1). This is partially explained by the fact that it is also the most frequently used amino acids at other sites in yeast proteins among those amino acids with positive $E_{NME \cdot i}$ values, accounting for 8.97% of all amino acids excluding the first two amino acids (i.e., the initiator Met and the penultimate amino acid) in the proteins. In contrast, Ala, Gly, Val, Pro, and Thr account for only 5.50%, 4.98%, 5.57%, 4.38%, and 5.93%, respectively. It is likely that Ser is more abundant, and consequently translated faster, than other amino acids with positive $E_{NME\cdot i}$ values, which would account for its overuse at the penultimate site to accelerate the movement of the ribosome downstream.

Table 1. Details of computing NME efficiency ($E_{NME\cdot i}$) for amino acid i, based on yeast and human proteins. AA – amino acids in 3-letter code, N_i – number of amino acid i at the penultimate site of proteins before any N-terminal processing, M_i – number of amino acid i in the penultimate site of proteins known to undergo NME. $E_{NME\cdot i}$ is specified in Eq. **Error! Reference source not found.**.

	Yeast			*Human*		
AA	N_i	M_i	$E_{NME\cdot i}$	N_i	M_i	$E_{NME\cdot i}$
Ala	484	47	1.3013	6904	204	1.0548
Gly	278	23	1.0702	2983	59	0.4757
Val	370	28	0.9416	1576	25	0.1574
Pro	265	18	0.7857	1955	53	0.9306
Ser	1437	93	0.7159	3664	108	1.0513
Thr	459	21	0.2155	1674	32	0.4265
Asn	296	1	-3.5439			
Glu	311	1	-3.6152			
Arg	167	0		1788	1	-4.6685
Asp	297	0				
Cys	38	0		376	2	-1.4190
Gln	147	0				
His	71	0				
Ile	191	0				
Leu	360	0				
Lys	335	0				
Met	84	0		495	2	-1.8157
Phe	190	0				
Trp	35	0				
Tyr	73	0				

An alternative explanation for the overuse of Ser at the penultimate site invokes the recognition of translation initiation signal. The consensus sequence including the initiation codon aug is (A/U)A(A/C)A(A/C)AaugUC(U/C) for highly expressed yeast genes. If the +4U, +5C and/or +6Y sites are part of the recognition sequence for the ribosome-dependent scanning model of translation initiation [7, 8, 23, 24], then the use of Ser at the penultimate site will be increased as a consequence because UCY codons code for Ser. However, this explanation appears unnecessary. After all, it is tenuous to argue that, while +4U is important for translation initiation in yeast, it is +4G that is important for translation initiation in mammals. Empirical evidence suggests that the +4 site is not really important for translation initiation [10].

Given that most proteins undergo NME, one naturally would expect that most proteins should have a small amino acid with a positive $E_{NME\cdot i}$ at the penultimate site. This is true, the sum of N_i values for the six amino acids with $E_{NME\cdot i} > 0$ accounts for 55.93% of all amino acids at the penultimate site in yeast (Table 1). Given that Ala, Gly and Val are all coded by G-starting codons, the over-representation of these amino acids at the penultimate site is sufficient to explain the presence of +4G site in protein-coding genes.

For human proteins, $q_i > p_i$ for Ala, Gly, Ser, Pro, Thr and Val (Figure 2). This list of six amino acids is the same as that in the yeast (Figure 1). In contrast to the yeast proteins where Ser is used most often at the penultimate site, human proteins have Ala as the most frequently used amino acid at the penultimate site, with Ser being the second (Figure 2). $E_{NME\cdot i}$ values and the computational details are shown in the last three columns of Table 1.

The reason for Ala to be found more frequently at the penultimate site than Ser (Figure 2 and Table 2) may be related to the relative availability of $tRNA^{Ala}$ and $tRNA^{Ser}$. There are 43 $tRNA^{Ala}$ genes and only 28 $tRNA^{Ser}$ genes in the human genome (http://lowelab.ucsc.edu/GtRNAdb).

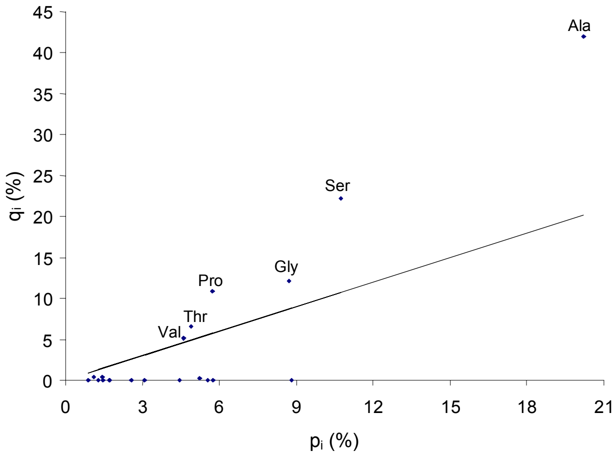

Figure 2. Different amino acids at the penultimate site result in dramatically different NME efficiencies in human proteins, with the meaning of symbols as in Figure 1.

There are several differences between the yeast and the human results. First, although eukaryotes have both MAP1 and MAP2 proteins, MAP1 appears to be the dominant isoform whose loss leads to dramatic decrease in growth whereas loss of MAP2 decreases growth only slightly [25]. In contrast, MAP2 is more important than MAP1 in higher eukaryotes [2]. MAP2 is less efficient than MAP1 in NME when the penultimate site is occupied by Gly. A yeast MAP1 mutant that over-expressed MAP2 proteins can restore most of the NME activity observed in the wild-type except for the test peptide with Gly at the penultimate site [Table I in 25]. This implies that the r_i value for Gly should be smaller in human (where MAP2 is more important) than in yeast (where MAP1 is more important). Our result confirms this prediction, with $E_{NME \cdot i}$ = 1.0702 and 0.4757, respectively, for yeast and human (Tables 1).

In conclusion, we have demonstrated the utility of a bioinformatic approach to characterize NME efficiencies of different amino acids at the penultimate site that can be adapted for any new species with empirical NME data.

4. Acknowledgement

This project is supported by NSERC's Discovery, Strategic and RTI Grants to XX. Ivo Pedruzzi of Swiss-Prot helped us in identifying protein sequences with experimental evidence for NME. The first two authors contributed equally.

5. References

[1] T. Meinnel, Y. Mechulam, and S. Blanquet, "Methionine as translation start signal: a review of the enzymes of the pathway in *Escherichia coli*," *Biochimie*, vol. 75, pp. 1061-75, 1993.

[2] C. Giglione, A. Boularot, and T. Meinnel, "Protein N-terminal methionine excision," *Cell Mol Life Sci*, vol. 61, pp. 1455-74, 2004.

[3] A. Serero, C. Giglione, A. Sardini, J. Martinez-Sanz, and T. Meinnel, "An unusual peptide deformylase features in the human mitochondrial N-terminal methionine excision pathway," *J Biol Chem*, vol. 278, pp. 52953-63, 2003.

[4] C. Giglione, O. Vallon, and T. Meinnel, "Control of protein life-span by N-terminal methionine excision," *Embo J*, vol. 22, pp. 13-23, 2003.

[5] T. A. Farazi, G. Waksman, and J. I. Gordon, "The Biology and Enzymology of Protein N-Myristoylation," *J. Biol. Chem.*, vol. 276, pp. 39501-39504, 2001.

[6] R. P. Moerschell, Y. Hosokawa, S. Tsunasawa, and F. Sherman, "The specificities of yeast methionine aminopeptidase and acetylation of amino-terminal methionine in vivo. Processing of altered iso-1-cytochromes c created by oligonucleotide transformation," *J Biol Chem*, vol. 265, pp. 19638-43, 1990.

[7] M. Kozak, "Initiation of translation in prokaryotes and eukaryotes," *Gene*, vol. 234, pp. 187-208, 1999.

[8] M. Kozak, "Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6," *Embo J*, vol. 16, pp. 2482-92, 1997.

[9] M. Kozak, "Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes," *Cell*, vol. 44, pp. 283-92, 1986.

[10] X. Xia, "The +4G site in Kozak consensus is not related to the efficiency of translation initiation," *PLoS ONE*, vol. 2, pp. e188, 2007.

[11] F. Frottin, A. Martinez, P. Peynot, S. Mitra, R. C. Holz, C. Giglione, and T. Meinnel, "The Proteomics of N-terminal Methionine Cleavage," *Mol Cell Proteomics*, vol. 5, pp. 2336-49, 2006.

[12] C. Flinta, B. Persson, H. Jornvall, and G. von Heijne, "Sequence determinants of cytosolic N-terminal protein processing," *Eur J Biochem*, vol. 154, pp. 193-6, 1986.

[13] X. Xia, *Data analysis in molecular biology and evolution.* Boston: Kluwer Academic Publishers, 2001.

[14] X. Xia and Z. Xie, "DAMBE: Software package for data analysis in molecular biology and evolution," *Journal of Heredity*, vol. 92, pp. 371-373, 2001.

[15] C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek, "The Universal Protein Resource (UniProt): an expanding universe of protein information," *Nucleic Acids Res*, vol. 34, pp. D187-91, 2006.

[16] T. Ikemura, "Correlation between codon usage and tRNA content in microorganisms," in *Transfer RNA in protein synthesis.*, D. L. Hatfield, B. Lee, and J. Pirtle, Eds. Boca Raton, Fla.: CRC Press, 1992, pp. 87-111.

[17] S. Kanaya, Y. Yamada, Y. Kudo, and T. Ikemura, "Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis," *Gene*, vol. 238, pp. 143-155, 1999.

[18] R. Percudani, A. Pavesi, and S. Ottonello, "Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae," *J Mol Biol*, vol. 268, pp. 322-30, 1997.

[19] L. Duret, "tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes," *Trends Genet*, vol. 16, pp. 287-9., 2000.

[20] X. Xia, "How optimized is the translational machinery in Escherichia coli, Salmonella typhimurium and Saccharomyces cerevisiae?" *Genetics*, vol. 149, pp. 37-44., 1998.

[21] M. Bulmer, "The selection-mutation-drift theory of synonymous codon usage.," *Genetics*, vol. 129, pp. 897-907, 1991.

[22] H. Liljenstrom and G. von Heijne, "Translation rate modification by preferential codon usage: intragenic position effects," *J Theor Biol*, vol. 124, pp. 43-55., 1987.

[23] M. Kozak, "The scanning model for translation: an update," *J Cell Biol*, vol. 108, pp. 229-41, 1989.

[24] M. Kozak, "A consideration of alternative models for the initiation of translation in eukaryotes," *Crit Rev Biochem Mol Biol*, vol. 27, pp. 385-402, 1992.

[25] X. Li and Y. H. 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*, vol. 92, pp. 12357-61, 1995.