Available online at www.sciencedirect.com
ScienceDirect
www.sciencedirect.com/science/journal/16720229

GENOMICS
PROTEOMICS &
BIOINFORMATICS

Article

Follow the Leader: Preference for Specific Amino Acids Directly Following the Initial Methionine in Proteins of Different Organisms

Ronen Shemesh*, Amit Novik, and Yossi Cohen

Compugen Ltd., Tel Aviv 69512, Israel.

Genomics Proteomics Bioinformatics 2010 Sep; 8(3): 180-189 DOI: 10.1016/S1672-0229(10)60020-4

Abstract

It is well established that the vast majority of proteins of all taxonomical groups and species are initiated by an AUG codon, translated into the amino acid methionine (Met). Many attempts were made to evaluate the importance of the sequences surrounding the initiation codon, mostly focusing on the RNA sequence. However, the role and importance of the amino acids following the initiating Met residue were rarely investigated, mostly in bacteria and fungi. Herein, we computationally examined the protein sequences of all major taxonomical groups represented in the Swiss-Prot database, and evaluated the preference of each group to specific amino acids at the positions directly following the initial Met. The results indicate that there is a species-specific preference for the second amino acid of the majority of protein sequences. Interestingly, the preference for a certain amino acid at the second position changes throughout evolution from lysine in prokaryotes, through serine in lower eukaryotes, to alanine in higher plants and animals.

Key words: amino acid, initial methionine, translation ribosome, computational biology, bioinformatics

Introduction

In recent years it has become clear that the initiation of translation is affected by many factors (*1-3*). These factors might be embedded in the mRNA sequence surrounding the initiation codon [usually AUG, coding for methionine (Met)], especially upstream (*4, 5*), or following the initiating codon (*6, 7*) and even further downstream (*8, 9*). Many of these studies have shown that the sequence directly following the initiation codon might be an important factor in both initiation of translation (*10, 11*) and efficiency of translation in prokaryotes (*12-16*) and eukaryotes (*17-19*). Although directly affecting the protein sequence and function, as well as its translation rate, the encoded amino acids directly following the most abundant Met residue were rarely studied. Yet, it was established that the amino acid sequence as well as the related coding RNA sequence and codon usage are important for control of translation and stability of proteins (*20*). More specifically, the amino acid following the initial Met is important in determining the rates of methionine aminopeptidase (MAP) function in proteins where the initial Met is cleaved (*21, 22*). Some researchers even attribute the functional features of the amino acid following the Met to the size of the side-chain (*23*). It was also

---

*Corresponding author.
E-mail: ronens@compugen.co.il

found that there is a direct relation between the metabolic stability of a protein and the identity of its N-terminal residue, mediated by different degradation mechanisms (*24*).

Analysis of the sequence surrounding the initial AUG codon in different organisms revealed a surprising diversity. In *Escherichia coli*, the most common second amino acid following the Met (AUG) is lysine (Lys), mostly with its AAA codon (*25*). This specific codon is also the most expression promoting. In contradiction, the presence of XGG codons in several positions in the open reading frame sequence, including the second position, considerably reduces the level of expression (*26*). In the yeast *Saccharomyces cerevisiae*, the most abundant codon following the AUG was found to be UC U/C (*8, 27*), giving rise to serine (Ser) as the second amino acid. However, a recent publication (*28*) indicates that when examining highly expressed proteins in a set of bacteria (*9*), archeae and unicellular eukaryotes (*3*), the most abundant amino acid following the initial Met was alanine (Ala). This finding is further supported by a study in which replacement of the three downstream codons following the initial AUG of tobacco (bases +4 to +12) with the sequence GCU UCC UCC (coding for Ala, Ser, Ser) increased the expression level of a reporter gene up to 40 fold (*19*).

The importance and potential functionality of the amino acids directly following the initial Met, as well as the apparent difference between species, raises the question of the evolutionary conservation of the second amino acid and the mechanisms by which the control of translation and protein expression co-evolved with other elements sharing the same system or mechanism. It is therefore interesting to look at the abundance of the second amino acid throughout evolution and try to make sense of the probable mechanisms by which this feature controls the expression of proteins. In this work, we attempted to examine the abundance of the second amino acid in a large dataset of known/predicted proteins from prokaryotes and lower eukaryotes, such as bacteria and yeast, throughout different evolutionary steps, up to higher eukaryotes, such as mammals and higher plants. By doing this we wish to explore the importance of the second amino acid in determining protein characteristics like expression, stability and function, as well as establish a tool for functional and evolutionary predictions.

Results

Distribution of second amino acid changes between different taxa

We have conducted a survey on all available Swiss-Prot protein sequences (see Materials and Methods). The results presented in Table S1 show vast differences between phylogenic taxa. **Figure 1** shows the differences in amino acid distribution at the second position through three major groups. In bacteria, represented by *E. coli* (Figure 1A), the most abundant amino acid in the second position is Lys (16.7%), followed by Ser (13.6%). In unicellular eukaryotes, represented by *S. cerevisiae* (Figure 1B), the most common second amino acid is Ser (23.7%), with Ala in second place (7.9%). However, in mammals, represented by *Homo sapiens* (Figure 1C), the order is reversed as the most common second amino acid is Ala (22.9%) followed by Ser (11.4%); the effect is even more apparent in plants, represented by *Arabidopsis thaliana* (Figure 1D), in which the most common second amino acid is Ala (29.6%) followed by Ser (10.8%). The results presented clearly indicate that there are considerable differences in the distribution and selection of the amino acid at the second position of protein sequences. These differences might be related to evolutionary and developmental stages, or might indicate different translation-related mechanisms.

Distribution of the second amino acid gradually changes throughout evolution

Further analysis of the data presented in Table S1 indicates that the pattern is conserved in related species. For example, some bacteria species (*Bacillus subtilis*, *Haemophilus influenzae* or *Streptococcus pneumoniae*) keep the same preference for the amino acid at the second position. However, a closer examination reveals that there are other bacteria species (*Salmonella sp.*, *Shigella flexneri* and *Pseudomonas aeruginosa*) that show a different pattern in which the most preferred amino acid in the second position is Ser