BIOINFORMATICS DISCOVERY NOTE

Vol. 26 no. 23 2010, pages 2933–2935
doi:10.1093/bioinformatics/btq571

Sequence analysis

Advance Access publication October 28, 2010

Statistical distribution of amino acid sequences: a proof of Darwinian evolution

Krystian Eitner[1,2,*], Uwe Koch[3], Tomasz Gawęda[2] and Jędrzej Marciniak[1]

[1]Adam Mickiewicz University, ul. Grunwaldzka 6, 60-780 Poznań, [2]BioInfoBank Institute, Św. Marcin 80/82 lok. 355, 61-809 Poznań, Poland and [3]Lead Dicovery Center, Emil-Figge-Strasse 76a, 44227 Dortmund, Germany

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The article presents results of the listing of the quantity of amino acids, dipeptides and tripeptides for all proteins available in the UNIPROT–TREMBL database and the listing for selected species and enzymes. UNIPROT–TREMBL contains protein sequences associated with computationally generated annotations and large-scale functional characterization. Due to the distinct metabolic pathways of amino acid syntheses and their physicochemical properties, the quantities of subpeptides in proteins vary. We have proved that the distribution of amino acids, dipeptides and tripeptides is statistical which confirms that the evolutionary biodiversity development model is subject to the theory of independent events. It seems interesting that certain short peptide combinations occur relatively rarely or even not at all. First, it confirms the Darwinian theory of evolution and second, it opens up opportunities for designing pharmaceuticals among rarely represented short peptide combinations. Furthermore, an innovative approach to the mass analysis of bioinformatic data is presented.

Contact: eitner@amu.edu.pl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 5, 2010; revised on September 4, 2010; accepted on October 1, 2010

1 INTRODUCTION

According to the evolution theory of life development on Earth (Darwin, 1859; Kaplan *et al.*, 2000), the process of species' assimilation to the environment has generated genetic changes leading to optimum acclimatization. As a result, the 20 standard amino acids encoded by nucleic acids, which form proteins found in all species, occur in precisely defined quantities in each protein. The distribution of polypeptides is a good measure of the randomness of changes in the environment (Lenski, 1998, Bennett *et al.*, 1990) for respective species and it also represents the statistical distribution of selected polypeptides. Homology-based methods for designing protein tertiary structures are a useful consequence of this.

2 METHODS

A script was developed which counts the number of polypeptides with a defined length. Due to software limitations, only the listing of all the available di- (400) and tripeptides (8000) could be visualized. The script sorts polypeptides according to the number of polypeptide occurrences for the defined amino acid sequence or totally at random. The counting is performed by single amino acids or by a defined polypeptide length.

Table 1. Example for searching for three amino acid long sequences (sorting by the number of occurrences)

>Sample Fasta File; ATAATTTAGGATTTAC	
Normal search	Offset search
TTT 2	ATA 1
ATT 2	TTT 1
TTA 2	AAT 1
GAT 1	TAG 1
TAG 1	ATT 1
GGA 1	GGA 1
AGG 1	TTA 1
ATA 1	
TAA 1	
AAT 1	
TAC 1	

2.1 fastaUniqAASeq.pl (Fasta Unique Sequences Amino Acids Search Script) description

FASTA is a linear format for the description of proteins and DNA/RNA. Files in the FASTA format contain information about the types of amino acids in a protein. The format is very useful for searching for similar proteins in databases (BLAST, CLUSTALW projects). The largest FASTA databases are AS follows: UNIPROT, Sprot (Wu *et al.*, 2006) and Protein Data Bank (PDB). Owing to the software developed (fastaUniqAASeq.pl), we can search for recurring sequences of any length in FASTA files in two modes:

- Normal search (precise search, searching sequences one by one amino acid).
- Offset search (search with jumps by a defined sequence length).

The precise search involves the listing of all amino acid sequences with a defined length (shift by one amino acid). The offset search, in turn, involves shifting by a defined length (offset) with fragments at a distance of the offset being included. This is best explained by an example in Table 1.

Furthermore, the script can filter results (w, withoutIgnored option) to avoid situations when unknown amino acids (X) occur in the listing. One of the options is to generate sequences which do not occur in a group of sequences. This is useful for statistical analysis. The programme can sort results of searching for unique amino acid sequences in various ways:

- key-asc—alphabetic sorting, ascending by the amino acid name.
- key-random—no sorting.

[*] To whom correspondence should be addressed.

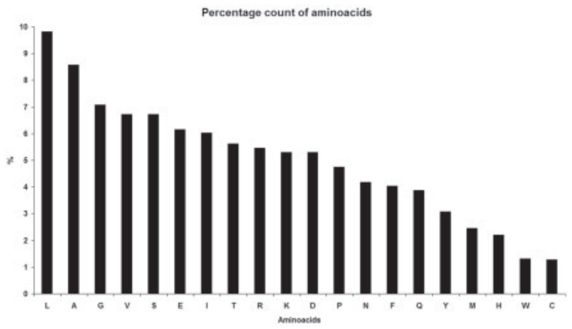

Fig. 1. Amino acid content (%) in the UniProt TREMBL database.

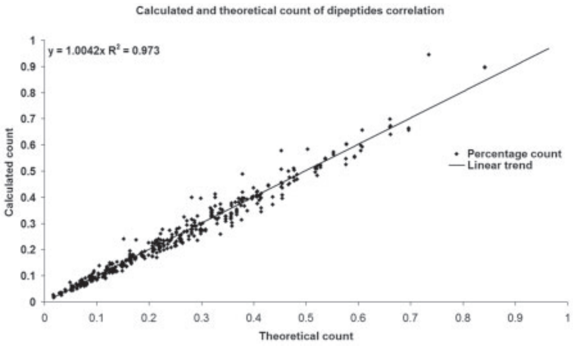

Fig. 2. The correlation between theoretical and calculated numbers (%) of dipeptides in the TREMBL database.

- key-usr—sorting by a user-defined sequence (changed in the programme source code).
- val-count—sorting by the frequency of an amino acid, descending.

The programme visualizes results using three methods. The most important is to count the physical number of sequences. In addition, it provides normalized quantity and percentage of a sequence. The programme has been used on the UNIPROT–TREMBL database. The sequences for species and selected enzymes were obtained using fastaGrep.pl. FastaGrep.pl is a programme used to retrieve respective protein or enzyme types or any combinations of characters in the description of a sequence (following the '>' character) from large FASTA files. When analyzing the whole UNIPROT–TREMBL file, fastaSplit.pl (a programme which splits FASTA files while retaining their structure, according to the number of entries in the resulting FASTA file or size of the resulting file) was also used; the programme was developed to enable the analysis of large protein sets using computers without sufficient memory by fastaUniqAASeq.pl. When processing files resulting from the splitting of large FASTA files, fastaStatResultsMerge.pl is used to merge the results.

The available UNIPROT–TREMBL database (Release 40.0 of 24 March 2009) was used for the global count analysis and a global histogram of amino acid triplets was generated (Figs 1–3).

3 CALCULATION

The calculation was performed for the whole UNIPROT TREMBL sequence database by counting amino acids, dipeptides and tripeptides. Histograms for percentage values were generated and the results of theoretical calculation were correlated with the counting results. Subsequently, species (*Arabidopsis thaliana*, *Danio rerio*, *Escherichia coli*, *Homo sapiens*, *Mus musculus*, *Oryza sativa*, *Saccharomyces cerevisiae*) were searched for. The selection of species resulted from the quantity of available sequences and the number of enzymes within the species. Hydrolases, polymerases and transferases were listed for each of the test species. Enzymes for which more than 100 protein sequences were available were used for comparative analysis. The Table 2 shows the number of sequences and enzymes for the species. Hydrolases from *A.thaliana* and *D.rerio* were rejected due to small number of sequences.

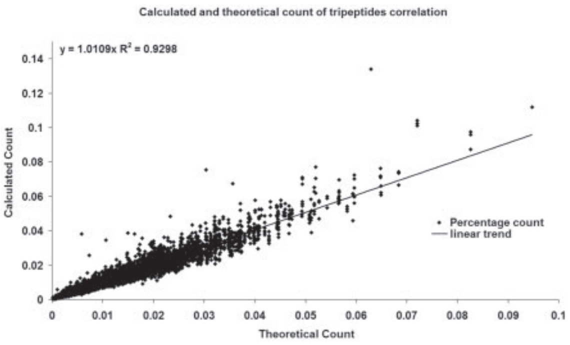

Fig. 3. The correlation between theoretical and calculated numbers (%) of tripeptides in the TREMBL database.

Table 2. Number of analyzed protein sequences within the FASTA files

Species	Proteins	Hydrolase	Polymerase	Transferase
Arabidopsis thaliana	42245	73	111	525
Danio rerio	26761	77	116	540
Escherichia coli	234128	3006	1432	10188
Homo sapiens	71093	280	267	1342
Mus musculus	48082	105	113	437
Oryza sativa	141121	307	147	1000
Saccharomyces cerevisiae	28824	103	208	436

4 RESULTS

According to the theory of independent events (Cramer, 1946), the change of an observed quantity with respect to the reference quantity is proportional to the observed quantity. The theory is reflected for example in thermal conductivity, statistical thermodynamics, chemical kinetics and the Lambert–Beer law. In mathematical terms, this means logarithmic decay of the observed quantity, dx. In our computational experiment, the distribution of dipeptides and tripeptides represented in the TREMBL database is logarithmic with the regression coefficient R^2 higher than 0.9. The percentage correlation between the theoretical number of di- (Fig. 2) and tripeptides (Fig. 3) and values calculated using the programme shows a linear relation with the regression coefficient R^2 higher than 0.9. The theoretical number of di- and tripeptides was calculated from the formula:

$$\%nAA = \frac{\prod_{i=1}^{n} \%AA}{(n-1)^{*}100\%}$$

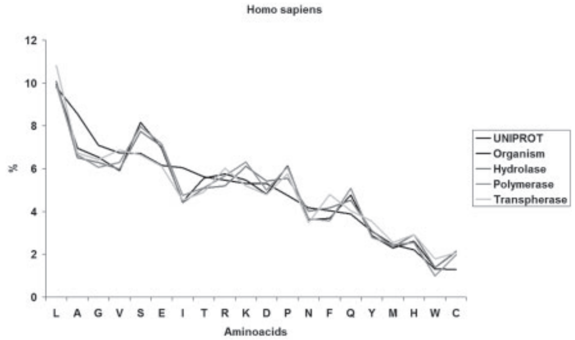

Fig. 4. Amino acid distribution for *Homo sapiens* and its selected enzymes.

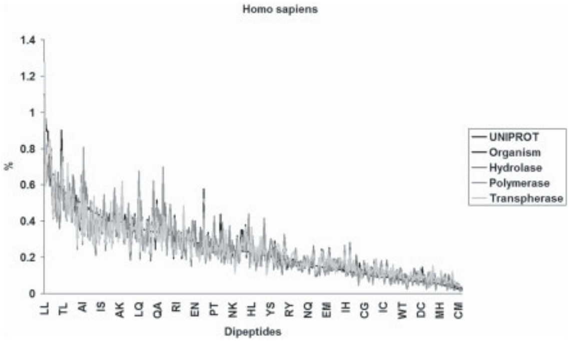

Fig. 5. Dipeptides distribution for *H.sapiens* and its selected enzymes.

where: %AA—percentage amino acid content in the whole TREMBL database (Fig. 1), %nAA—theoretical content of a dipeptide ($n=2$) or a tripeptide ($n=3$).

The major reason for the discrepancy of correlation between the theoretically calculated number of di- or tripeptides is that a lot of sequences with unknown amino acids (‘X’ or Alx ‘B’ and Glx ‘Z’), not included in the analysis, occur in the UNIPROT–TREMBL database. The programme deletes all di- or tripeptides which contain X, B or Z amino acids.

Amino acid and di- and tripeptide counting was performed for selected species similar to that for the whole TREMBL database. Figures 4–6 shows amino acid, di- and tripeptide distribution for *H.sapiens* and its selected enzymes. The distribution for the other species from the Table is available in the Supplemental Material.

When the TREMBL database is divided into species and enzyme classes, amino acid distribution close to theoretical number is obtained.

It is noted that amino acid content in a species decreases with the length of its biosynthesis pathway. Due to this, certain sequences occur more frequently, while other very rarely or not at all. A number of natural toxins, metabolites and protein inhibitors are composed of polypeptides or their natural modifications. It can be then argued that processed and highly hydrolyzed food was the first source of inhibitors and natural medicines which protected species against fungi, bacteria and viruses. Many authors claim that food processing (heat treatment) contributed to the rapid brain development (Gibbons, 2007; Milton, 1999).

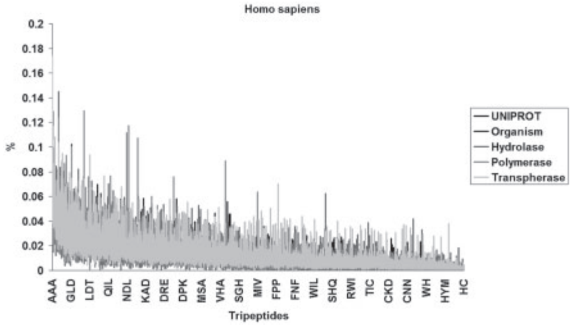

Fig. 6. Tripeptides distribution for *H.sapiens* and its selected enzymes.

5 SUMMARY

Our calculation results prove that the number of di- and tripeptide sequences is consistent with their theoretical number calculated based on the amino acid content in the TREMBL database. This results from the fact that amino acids have similar or even identical biosynthesis pathways in different species. Therefore, the number of amino acids and their combinations (polypeptides) is generally similar. Short sequences do not occur or do so only rarely and this opens up new opportunities for searching for natural protein inhibitors based on polypeptide structures and their modifications. Many toxins, immunosuppressants and inhibitors have a well-defined polypeptide fragment in their structure. Our calculations provide a new strategy for inhibitor design based on short natural amino acid sequences (Braddon *et al.*, 2010; Eitner *et al.*, 2009; Cai *et al.*, 2002).

ACKNOWLEDGEMENTS

K.E. would like to thank the Foundation for Polish Science for the support through the FOCUS programme.

Conflict of Interest: none declared.

REFERENCES

Bennett,A.F. *et al.* (1990). Rapid evolution in response to high temperature selection. *Nature*, **346**, 79–81.

Braddon,K.L. *et al.* (2010). Exploring the potential of template-based modelling. *Bioinformatics*, **26**, 1849–1856.

Cai,Y. *et al.* (2002) Information-theoretic analysis of protein sequences shows that amino acids self-cluster. *J. Theor*. *Biol.*, **218**, 409–418.

Cramer,H. (1946) *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ.

Darwin,C. (1859) In Murray,J. (ed.) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. Reprinted in Ernst Mayr, ed., On the Origin Of Species: A Facsimile of the First Edition, 1964, Harvard University Press, Cambridge, MA, pp. 495.

Eitner,K. and Koch,U. (2009) From fragment screening to potent binders: strategies for fragment-to-lead evolution. *Mini. Rev. Med. Chem.*, **9**, 956–961.

Gibbons,A. (2007) Paleoanthropology: food for thought. *Science*, **316**, 1558–1560.

Kaplan,H. *et al.* (2000) A theory of human life history evolution: diet, intelligence, and longevity. *Evol. Anthropol.*, **9**, 156–185.

Lenski,R.E. (1998) Bacterial evolution and the cost of antibiotic resistance. *Int. Microbiol.*, **1**, 265–270.

Milton,K. (1999) A hypothesis to explain the role of meat-eating in human evolution. *Evol. Anthropol.*, **8**, 11–21.

Wu,C. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.,* **34**, D187–D191.