Research Article
Open Access

Distribution of short oligopeptides in a dataset of selected polypeptides

Abstract

DNA bases act as alphabets and nucleotide triplets, each representing an amino acid, or a punctuation mark, dictate the order and frequency of occurrence for different amino acids in the newly synthesized polypeptide. The presence of the triplet code in DNA raises the possibility that there may be another code or linguistic formulation composed of 20 amino acids as different alphabets dictating the frequency and the serial order in which 20 amino acids are arranged on different polypeptide strings. With this in mind, we have created a database of di-, tri-, tetra- and pentapeptides and examined the distribution and frequency of occurrence of different types of short oligopeptides in a set of 51,685 polypeptide sequences selected from the Swiss Prot database.

Keywords: di-, tri-, tetra- and pentapeptides, oligopeptide matrices, forbidden oligopeptides, clustering algorithm

Volume 9 Issue 4 - 2020

Varun Ravishankar,[1] Natasha Kelkar,[1,2] Nachiket Pathak,[1] Rutuj Kolhe,[1] Onkar Ghuge,[1] Shantanu Madiwale,[1] Dhanashree Deore,[1] Anupam Saraph,[3] Milner Kumar,[4] Anil Gore,[5] SP Modak[6,7]

[1]Institute of Bioinformatics and Biotechnology, Savitribai Phule Pune University, India
[2]Program in Molecular and Cellular Biology, Dartmouth college, Hanover, USA
[3]Adjunct Professor at the Symbiosis Institute of Computer Studies and Research, India
[4]BlackBuckTechnologies1F-B, Raja Garden, 51 Amman Koil Street, Thelliaragaram Porur, India
[5]Cytel Statistical Software and Services Pvt Ltd, India
[6]Open Vision, 759/75, Deccan Gymkhana, India
[7]Visiting Professor, Institute of Bioinformatics and Biotechnology, Savitribai Phule Pune University, India

Correspondence: Sohan P Modak, Open Vision, 759/75, Deccan Gymkhana, Institute of Bioinformatics and Biotechnology, Savitribai Phule Pune University, Pune 411004, India, Email spmodk@gmail.com

Received: October 16, 2020 | **Published:** November 18, 2020

Introduction

After nucleic acids, proteins are the next most important macromolecules conducting life processes. Proteins give rise to specific structures and perform, albeit catalyse, specific functions in a cell. The size and composition of the proteome may vary among different organisms depending on their evolutionary position. Proteins are composed of single or multiple polypeptides wherein each one is a chain of covalently linked amino acids. The information for the composition, the number of amino acid residues and their order is encoded by an orderly sequence of nucleotide triplets in DNA, from among 64 possible combinations of 4 bases A, T, G and C, wherein 61 triplets or "codons" consign the identity of 20 amino acids, while the remaining stand for punctuation marks. The genetic information is stored in the genome in the form of DNA, while genomes of some viruses are composed of RNA.[1] In DNA, the polypeptide-specific informational segment is called as the "coding sequence". When the information is being retrieved, RNA polymerase copies the coding sequence in DNA, along with accompanying upstream and downstream regulatory noncoding segments. The product of transcription, RNA, has a limited life span after which it is degraded, while DNA is retained as the master copy. The subsequent process of information retrieval, or translation of the coding sequence into an amino acid sequence is catalysed by peptide ligases or peptidyl transferases. For translation, the ribosome serves as the platform. As the ribosome slides along the messenger RNA, it translates the coding sequence into a string of polypeptide, catalysed by peptidyl transferases that sequentially join amino acids brought to the ribosome by specific transfer RNAs. The final product is released from the mRNA-ribosome complex after moving beyond the termination codon acting as the full stop. [1]

In addition to the coding sequence, each gene contains contiguous stretches of noncoding segment acting as sites for binding gene-specific promoters required to position RNA polymerase well upstream of the coding sequence before initiating the transcription so that the actual transcript contains not only the coding sequence but also upstream noncoding regions to include the recognition sites for promoters, site/s to dock RNA polymerase and ribosome, as well as a noncoding region downstream facilitating the release of the RNA polymerase and ribosome, and processing signals to trim or modify the transcript. Thus, the primary transcript or pre-mRNA, copied from the 'Transcriptional unit', is considerably longer than the coding sequence.[2]

During the evolution from prokaryotes to eukaryotes, the number and types of proteins encoded by the genome has increased causing an increased length of the genome. However, during this process, the non-coding regions increased disproportionately greater to include, in addition to the transcriptional units, a variety of inter-genic spacers, DNA repeats, satellites, chromatin anchoring sites, and sites for recombination and other processing signals. The distance between neighbouring transcriptional units has increased so much that a major portion of the genome contains non-coding sequences.

[1]The frequency of occurrence of oligopeptide depends on the clustering model. Thus, compared to clustering at 91.7%, clustering at 99% reveals a smaller number of oligopeptides occurring once.

In order to better understand protein sequences, Erhan & Erhan et al.,[3,4] constructed matrices for the frequency of occurrence of di- and tripeptides within polypeptide sequences and suggested that certain dipeptides are favoured in proteins. Tuller et al.,[5] examined the occurrence of penta-peptides in the proteome of 368 organisms and reported that there are certain pentapeptides that are forbidden. However, Poznański et al.,[6] suggested that the so called forbidden penta-peptides may actually show up with the annotation of increasing number of protein sequences.

Short oligopeptides studied for their role in different biophysical and biochemical processes could either be complete sequences coded for by genes, or fragments formed by enzymatic degradation of proteins. For example, enkephalin is a pentapeptide formed by proteolytic cleavage of the hormone proenkephalin.[7–10] These types of oligopeptides are called endogenous oligopeptides.[11] Similar short peptides, secreted by plants, play a role in cell signalling.[10] Cationic antimicrobial peptides secreted by immune cells act as anti-bacterial agents. Synthetically produced short oligopeptides have been shown to be useful in cancer therapy.[12] Further, short pentapeptides have been used to predict secondary structure. With the increasing number of reports on oligopeptides, databases and repositories have been made which store information on food derived peptides[12,13] bioactive peptides[14] and regulatory oligopeptides.[11] The EROP-MOSCOW database provides biophysical and biochemical properties of all endogenous oligopeptides.[6]

We have been interested in the possibility that some of the oligopeptides, composed of combination of residues bearing unusual physical properties, may act as signals in defining the positional differences in the structure and the functional/active sites in polypeptides. Therefore, we have constructed a database of and examined their frequency distribution within polypeptide sequences selected from the SWISSPROT database.

Methodology

The 20 naturally occurring amino acids can form 20 x 20 or 400 types of dipeptides, 8,000 types of tripeptides, 1,60,000 types of tetrapeptides and 32,00,000 types of pentapeptide sequences. We have created sequence matrices for di-, tri-, tetra- and pentapeptides and analysed a detailed frequency distribution all possible combinations of such short oligopeptides in select polypeptide sequences.

Database preparation

A dataset of 104,267 polypeptide sequences, with evidence of existence at a protein level, was downloaded from UniProt/SwissProt. This dataset was further sorted to remove redundancy as follows: all strains, excepting the wild types, were removed. We also removed duplicates, sequences shorter than 60 amino acids, fragments, putative, probable, and uncharacterized proteins so as toobtain a dataset of well-studied proteins. It must be noted that a few polypeptides shorter than 60 amino acids,although excluded from our dataset, do act as signalling elements or neuropeptides.

Finally, the dataset was clustered at 91.7% identity using the CD-Hit server. This was done in order to remove near identical sequences. This cut-off of 91.7% was selected based on the fact that we wanted a difference of at least 5 amino acids between any pair of sequences. As the shortest sequence in our dataset is 60 amino acid long, 5 amino acids represent (5/60) x 100 = 8.3% of the sequence.

The clustering algorithm carried out sequence alignment of all sequences to estimate similarity indices. The algorithm clusters the proteins on the basis of the extent of similarity or the similarity index of their sequence. The longest sequence in the cluster is retained as the representative sequence. In our case, sequences which had 91.7% similarity or higher were grouped into a single cluster. From these only the representative sequences appear in our final dataset.

Generation of matrices and frequency distribution of oligopeptide types in the selected sequences

Codes for generating matrices were written separately in both Python and in C in order to cross check the results. These codes were run for the dataset of 51,685 sequences to generate counts of occurrences of all dipeptides, tripeptides, tetrapeptides and pentapeptides. In the 2D matrices generated, the row headings indicate the amino acid(s) from the N-terminal, while the column headings indicate the amino acid(s) towards the C-terminal. The entire oligopeptide hence, must be read from row to column. This polarity must be taken into account as two types of dipeptides AC and CA would have different frequencies of occurrence. It should be noted that we have considered all the possible oligopeptides of length 2 to 5 amino acids that can be formed by combination of the 20 amino acids. We have excluded the oligopeptides which contain modified amino acids like selenocysteine (U), asparagine (B), glutamine (Z), pyrrolysine (O). We also haven't included the combinations that contained the amino acid 'X' (X being an unidentified/ unknown amino acid).

Results and discussion

Database generation

A dataset of reviewed sequences was downloaded from UniProt/ Swiss Prot (https://www.uniprot.org/) which had evidence at protein level. This dataset contained 1,04,267 sequences, as of August, 2020. Out of this dataset, we selected 51,685 protein sequences from 3,489 organisms as per the methodology indicated in section 2.1, and the rest were removed including 29,302 from strains of bacteria and viruses keeping only interested in the wild type species. A detailed breakdown of the sequence selection is shown in Table 1. The dataset of 51,685 sequences was distributed among 3489 organisms, of which 46,260 sequences were from eukaryotes, 3997 from bacteria, 146 from archaea and 1282 from viruses. The total length of the proteins in the dataset was 2,73,72,120 amino acids.

Table 1 The table depicts how the dataset downloaded from Swiss Prot/ UniProt containing 104,267 protein sequences was curated to database containing 51,685 polypeptide sequences

Sequence Selection steps	No. of sequences
Initial set (Swiss Prot evidence at protein level)	104,267
Sequences from strains other than wild types	29,302
Duplicated sequences	2041
Sequences shorter than 60 amino acids	6409
Sequence fragments	1167
Putative, probable and uncharacterized sequences	1830
Clustering at 91.7% identity cut-off	11,860
Total sequences removed during curation	52,609
Final set of sequences after subtracting the above	51,685

Citation: Ravishankar V, Kelkar N, Pathak N, et al. Distribution of short oligopeptides in a dataset of selected polypeptides. *MOJ Proteomics Bioinform.* 2020;9(4):91–94. DOI: 10.15406/mojpb.2020.09.00288

Frequency of occurrence of oligopeptides

We examined our dataset of polypeptide sequences for the presence and frequency of occurrence of different types of dipeptides, tripeptides, tetrapeptides and pentapeptides. In the dipeptide matrix, we observed that the dipeptides with lowest counts, contained cysteine, methionine and/or tryptophan. A similar trend was observed with tri-, tetra- and pentapeptides. It is notable that as the length of oligopeptides increases from 2 to 5, the frequency of occurrence of oligopeptides decreases (Tables 2&3). For example, we find 3,57,463 types of pentapeptides were present only once. The most significant outliers were the pseudouridine synthase from *Plasmodium falciparum* (ID: Q8I3Z1) and the bat coronavirus replicase polyprotein 1ab (ID: P0C6W5) which had many such pentapeptides. We are examining longer oligopeptides for their frequency of occurrence. The overall frequency distribution of 3.2 million possible pentapeptides in the dataset of 51,685 sequences has been depicted in Figure 1. The amino acids tryptophan and cysteine prevailed in the pentapeptides regarded as universally forbidden by Tuller et al.[5] They proposed that this may be due to reactivity of cysteine residues and bulkiness of tryptophan side chain.

Table 2 Oligopeptides in 51,685 sequences with highest lowest frequency

	Oligopeptides occurring least frequently		Oligopeptides occurring most frequently		Total no. of oligopeptides
	Sequence	Occurrence	Sequence	Occurrence	
Dipeptides	WW	533	LL	271557	2,73,19,024
	WC	6227	SS	252496	
	MW	6564	SL	211301	
	CW	6925	LS	210010	
	WM	7986	AA	205927	
Tripeptides	WWC	83	SSS	42552	2,72,66,831
	WCM	92	AAA	33298	
	MWW	93	LLL	32190	
	CWW	97	EEE	32052	
	WMW	98	PPP	26686	

Table 3 Types of tetra- and pentapeptides from 51,685 sequences occurring at low frequency

	Frequency of occurrence				
	0	1	2	3	4
No. of tetrapeptide combinations	37	108	163	215	269
No. of pentapeptide combinations	4,15,348	3,57,463	3,12,554	2,67,974	2,27,302

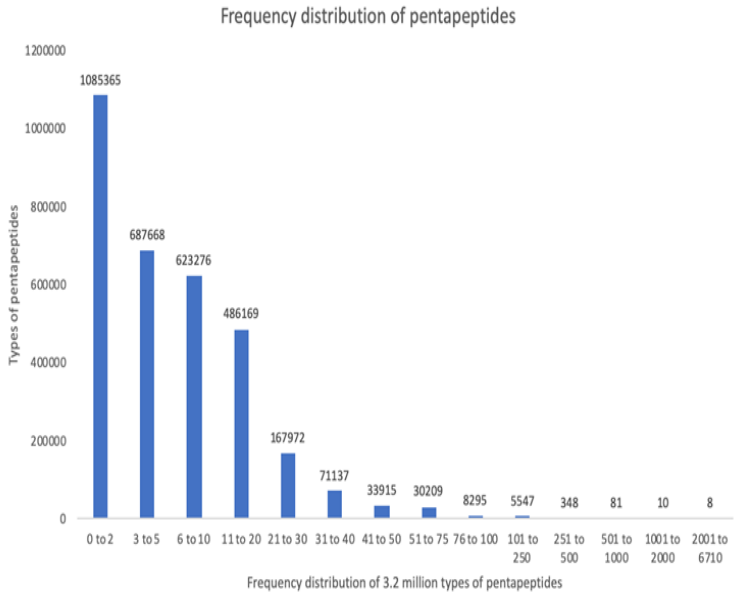

Figure 1 Frequency distribution of the 3.2 million types of pentapeptides in the dataset of 51,685 polypeptide sequences. The first bar shows that out of 32,00,000 types of pentapeptides, 10,85,365 exhibited a count of 0,1 or 2.

Citation: Ravishankar V, Kelkar N, Pathak N, et al. Distribution of short oligopeptides in a dataset of selected polypeptides. *MOJ Proteomics Bioinform.* 2020;9(4):91–94. DOI: 10.15406/mojpb.2020.09.00288

Tulleret al.,[5] examined proteomes of 386 organisms including 27 archaea, 313 bacteria and 28 eukaryotes and reported that 5000 types of pentapeptides were absent in these proteomes. They termed these as 'universally forbidden pentapeptides'. Though their dataset had a much larger number of protein sequences, tallying up to a total of 6.39 x 10^8 amino acids, as compared to 2.73 x10^7 amino acids in our dataset, it should be noted that not only was their dataset not curated, but it also contained different strains of the same wild type bacterial species thereby dealing with repetitious occurrences of the same oligopeptides. We further observed that a few pentapeptides which they had claimed to be universally forbidden were present in our dataset.For example, the pentapeptide FFMCT claimed to be universally forbidden was observed in the VP-7 glycoprotein of the African horse sickness virus - 4 (ID: P36325), while RNMFC was present in a snake venom prothrombin activator (ID: A6MFK7). This clearly means that the notion of 'forbidden sequences' depends on the size and diversity of the curated dataset of polypeptides. Hence, we term pentapeptides absent from our dataset as 'missing pentapeptides'. As the number of annotated protein sequences increases, the number of missing pentapeptides will decrease.[16–18]

Conclusions

We have developed a highly selective dataset of proteins curated by SwissProt, and generated di, tri, tetra and pentapeptide frequency of occurrence matrices and shown that there the so-called 'forbidden pentapeptides' is an erroneous nomenclature as we find some among them present in our dataset. Therefore, we call the pentapeptide sequences absent in any given dataset as 'missing' and not 'forbidden" because the oligopeptides absent in a dataset may show up in another dataset as shown above. From the matrices hence generated, we also saw that a number of polypeptide sequences showed the presence of a few oligopeptides only once. We are now examining the frequency distribution of longer oligopeptides in polypeptide sequences.

Acknowledgments

We thank Prof. Smita Zinjarde, Director Institute of Bioinformatics and Biotechnology and Dr. Vaijayanti Tamhane, for their enthusiastic support during this work.

Conflicts of interest

Author declares that there is no conflicts of interest.

References

1. Boyle J. Lehninger principles of biochemistry: Nelson, D., and Cox, M. *Biochemistry and Molecular Biology Education*. 200;33(1):74–75.
2. Scherrer, Klaus. Primary transcripts: From the discovery of RNA processing to current concepts of gene expression– Review. *Exp Cell Research*. 2018;373(2):1–33.
3. Erhan S. A unique method to represent proteins. *International Journal of Bio–Medical Computing*. 1980;11(1):77–82.
4. Erhan S, Marzolf T, Cohen L. Amino–acid neighborhood relationships in proteins. Breakdown of amino–acid sequences into overlapping doublets, triplets and quadruplets. *International Journal of Bio–Medical Computing*. 1980;11(1):67–75.
5. Tuller T, Chor B, Nelson N. Forbidden penta-peptides. *Protein Science*. 2007;16(10):2251–2259.
6. Poznański J, Topiński J, Muszewska A, et al. Global pentapeptide statistics are far away from expected distributions. *Scientific reports*. 2008;8(1):1–18.
7. Daliri EBM, Oh DH, Lee BH. Bioactive peptides. *Foods*. 2017;6(5):32.
8. Zamyatnin AA, Voronina OL. Food protein fragments are regulatory oligopeptides. *Biochemistry (Moscow)*. 2012;77(5):502–510.
9. Johanning K, Juliano MA, Juliano L, et al. Specificity of prohormone convertase 2 on proenkephalin and proenkephalin–related substrates. *Journal of Biological Chemistry*. 1998;273(35):22672–22680.
10. Murphy E, Smith S, De Smet I. Small signalling peptides in Arabidopsis development: how cells communicate over a short distance. *The Plant Cell*. 2012;24(8):3198–3217.
11. Zamyatnin AA, Borchikov AS, Vladimirov MG. The EROP–Moscow oligopeptide database. *Nucleic acids research*. 2006;34(1):D261–D266.
12. Abusara OH, Freeman S, Aojula HS. Pentapeptides for the treatment of small cell lung cancer: Optimisation by Nind–alkyl modification of the tryptophan side chain. *European journal of medicinal chemistry*. 2017;137:221–232.
13. Minkiewicz P, Iwaniak A, Darewicz M. BIOPEP–UWM database of bioactive peptides: Current opportunities. *International journal of molecular sciences*. 2019;20(23):5978.
14. Wang J, Yin T, Xiao X, et al. StraPep: a structure database of bioactive peptides. *Database. 2018*.
15. Li Q, Zhang C, Chen H, et al. BioPepDB: An integrated data platform for food–derived bioactive peptides. *International Journal of Food Sciences and Nutrition*. 2018;69(8):963–968.
16. Figureau A, Soto MA, Toha J. A pentapeptide–based method for protein secondary structure prediction. *Protein Engineering*. 2003;16(2):103–107.
17. Huang Y, Niu B, Gao Y, et al. CD–HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*. 2010;26(5):680–682.
18. www.uniprot.org