An Entropy-Based Alignment-Free Mathematical Framework for Protein Sequence Similarity Analysis

D.Vijayalakshmi[1*], *M.Muralidharan*[2], *K.Rameshwar*[3]
[1]Department of Mathematics, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya Kanchipuram, Tamil Nadu, India
guruviji97@gmail.com
[2]Department of Mathematics, Kings Engineering college, Irungattukottai, Chennai, Tamil Nadu , India
muralidharan@kingsedu.ac.in
[3]Department of Mathematics, Bharath Instutue of Higher Education and Research, Chennai, Tamil Nadu, India
Krameshmath@gmail.com

Abstract. In this paper, a mathematical, alignment-free framework is proposed for quantifying similarity and dissimilarity among protein sequences using their physicochemical characteristics. The primary structure – the sequences of amino acids is modelled numerically by mapping each amino acid to a vector in a real-valued feature space determined by physicochemical properties, namely the hydropathy index (hI), first dissociation constant (pka1), and second dissociation constant (pka2). The considered properties play a key part in protein folding, stability, and function, and hence forms a meaningful basis for comparative analysis. The protein sequences are represented as discrete distributions using this numerical representation in a multidimensional parameter space. A Relative Distance Entropy measure is employed to compare the proteins independent of sequence alignment and length, thereby overcoming the limitations inherited in conventional homology-based methods. This enablessimilarity-measurement even in cases of low sequence identity while preserving functional characteristics. The proposed approach provides a computationally efficient and mathematically sound alternative for large-scale protein similarity analysis and functional classification. This method produces similarity values from 47%-100% providing a comparable trend with BLAST sequence identity percentage values for tested protein pairs. The proposed method emphasizes mathematical modelling and computational efficiency, making it suitable for large scale data.

1 Introduction

Protein Structure Comparison (PSC) is the key problem in structural biology and drug discovery; it allows researchers, for instance, to infer protein evolution (to understand better the relationship between protein structure and function) and to transfer knowledge about known proteins to a novel protein.

Graphical representation provides not only visual qualitative inspection of gene data but also mathematical characterizations through objects such as matrices. An important problem in bioinformatics is the analysis and comparison of DNA and protein sequences efficiently to depict their evolutionary relationship. The principal objective of evolutionary studies is to follow the state of the system through long periods of time. However, it is impractical to repeat these evolutionary events in the laboratory. Therefore, the approaches for comparing biological

* Corresponding author:guruviji97@gmail.com

sequences are mainly based on computational and statistical methods.

The pioneering work by Hamori and Ruskin [1] in 1983 established that the graphical techniques for DNA sequences are a powerful tool for visualizing and analysing sequence data. It has been subsequently used by other researchers [2-14] for DNA and protein sequences [15-,24] incorporating the algorithmic and computational statistics. However, the center of attention of these problems is the consideration of the sequence as a string. Hence, all four nucleotide bases in DNA sequences and twenty amino acids in protein sequences are equally treated. In fact, the physicochemical properties of amino acids are found to have strong effects on amino acid substitution rates [25]. Hence these properties directly determine the estimation of distance between two amino acid sequences.

In most of these existing methods, the main drawbacks are that the higher the dimension of the protein sequence graphs, the heavier the computation complexity of the methods or the lower the recognition degree of the protein sequence graphs. For example, in the methods proposed in [27, 29], the main drawback is that the lines will cross each other, which will decrease the visibility of the graphics. Protein sequence comparison approach called PVC is explained in [31]. This is done by encoding sequence data and physicochemical properties of the amino acids. In [32], the sequence comparison study is done by fuzzy integral. The parameters of Markov chain are estimated by considering the frequencies of occurrence of all possible amino acid pairs. In [26], new embedding based sequence alignment approach is explained in detail. This method refines residue level embedding similarity using K-means clustering and double dynamic programming(DDP).

The pioneering graphical and numerical representations of sequences [1-5] have laid the foundation for such comparison methods. In this work, the focus is on using the physicochemical features of amino acids to transform protein sequences into a form suitable for entropy-based analysis.

In this paper, a novel framework lies in the integration of three physicochemical properties into a unified entropy similarity measure. The properties considered are the hydropathy index, first dissociation constant and second dissociation constant. These properties are chosen because they play the key roles in determining protein structure and function. These properties act as biologically meaningful descriptors, as the Hydropathy index represents the hydrophobic or hydrophilic properties of side chain of amino acids, Dissociation constants represents the ionization behaviors of amino acids [33]. Protein embedding based alignment is developed in [30],this uses dynamic programming algorithm. The matching score of amino acid is purely based on embedding from protein language model. The numerical values of these properties are obtained from standard biochemistry property tables in biochemistry literature. Existing alignment-free methods rely on amino acid composition, single property descriptor whereas this method models protein using normalized physicochemical values and compares using a symmetric relative entropy measure. This offers a mathematically rigorous and biologically interpretable alternative to existing protein similarity measures.This work primarily presents a mathematical framework for protein sequence comparison purely based on information-theoretic principles. The contribution consistently lies in development of a computational and entropy-based model rather than biochemical experimentation.

i.e This study explores a property-based method neglecting the limitations of high dimension and computational complexity. This is done by applying information theory concepts to numerical representations of amino acid sequences.

2 METHOD

The proposed method is the formulation of mathematical model that transforms biological sequences into numerical representation for quantitative analysis.

Procedure for Analysis of Protein Sequences.

In this section, the mathematical formulation and computational steps for analysing protein sequences are presented.

Let P1 and P2 be protein sequence of length m and n. The numerical value of physicochemical properties hydropathy index, pka1, pka2 of amino acids are used to represent each amino acid in the sequence. Each amino acid is mapped to a real valued three dimensional vector as below.

$v_i = (hI_i, pka1_i, pka2_i)$

Thus each amino acid is transformed into a three dimensional vector.

The protein sequence P1 and P2 are represented as

$P1 = (v_1, v_2, \ldots\ldots v_n)$ and $P2 = (w_1, w_2, \ldots\ldots w_m)$

$v_i = w_i = (x_i, y_i, z_i)$ where x = hI value of amino acid

y = pka1 value of amino acid

z = pka2 value of amino acid

To apply entropy, the 3-D vector is converted into a single positive scalar weight as given below.

$$s_i = \sqrt{x_i^2 + y_i^2 + z_i^2} \quad (1)$$

ie. $$s_i = \sqrt{hI_i^2 + pka1_i^2 + pka2_i^2} \quad (2)$$

Then the protein sequences P1 and P2 are represented as $P1 = (s_1, s_2 \ldots\ldots s_n)$ and $P2 = (t_1, t_2, \ldots\ldots t_m)$. The numerical representation is then normalised to probability distribution to ensure length-independent comparison and to satisfy the mathematical conditions of relative entropy. By transforming the vectors using Euclidean magnitude, the combined physicochemical properties are preserved. This captures the overall physicochemical characteristic of amino acids on entropy computation. Then the normalised values are represented as discrete probability distribution using the formula

$$pi = \frac{si}{\sum s_j} \quad qi = \frac{ti}{\sum t_j} \quad (3)$$

Since entropy measures require non-negative values, the protein sequences are represented as $P = (p_1, p_2, \ldots\ldots p_n)$ and $Q = (q_1, q_2, \ldots\ldots q_n)$. (4)

where $$\sum p_i = \sum q_i = 1 \quad (5)$$

To remove the effect of sequence length and sequence alignment limitations the normalised values are used on comparing proteins

Based on normalised representation of protein sequence the symmetric relative entropy measure derived from Kullback-Leibler divergence is calculated.

$$D(P,Q) = \frac{1}{2}\sum_{i=1}^{n}\left[p_i \log\left(\frac{p_i}{q_i}\right) + q_i \log\left(\frac{q_i}{p_i}\right)\right]. \quad (6)$$

As the classical KL divergence is asymmetric, this symmetric KL divergence is preferred as it provides a reliable metric irrespective of the order of comparison satisfying the key aspect for similarity analysis.

This distance considers all selected physicochemical properties and is independent of sequence alignment and length of protein sequence.

To have an easy interpretation of similarity measured, the entropy distance value is transformed into similarity percentage using the formula below

$$Similarity\ (\%) = \frac{1}{1+D(P,Q)} * 100 \quad (7)$$

This makes the comparison with BLASTP much meaningful and easier, similarity analysis is performed to validate the entropy measure. The result reported by BLASTp is used as BLAST similarity score for comparison with entropy method.

The overall procedure for computing protein similarity mathematically is summarised in the following algorithm.
Algorithm- Entropy-Based Protein Similarity Measure
Input: PDB ID of protein P1, P2
Output: Similarity percentage

1. Protein sequence corresponding to the given PDB IDs are retrieved.
2. Assign physicochemical properties (hydropathy index, pka1,pka2) to each amino acid.
3. Amino acids are represented as three dimensional physicochemical vector
4. Convert each 3D vectorinto scalar value using Euclidean distance.
5. The normalised scalar sequence is then converted to probability distribution.
6. Symmetric relative entropy distance between the distribution is computed.
7. The relative entropy measure is then converted into a similarity percentage.

The proposed method is implemented in python to measure the applicability of the mathematical model and to generate the similarity score. Protein data are retrieved from Protein Data Bank using Biopython. Details of numerical values of physicochemical properties are obtained from standard bio literature. Using Numpy scalar transformation, probability, normalisation and symmetric relative entropy computations were calculated. This implementation proves to be computationally efficient and suitable for large scale protein similarity study.

3 Result and Discussion

Based on the symmetric relative entropy distance obtained, the similarity is measured using the following criteria
1 If the distance is zero or very small then there exist high similarity
2. If the distance is larger, then there exist greater dissimilarity.
To have a concept proof demonstration the proposed method is performed on limited datasets. Future work can be extended to large scale protein datasets.

Table 1. Symmetric relative entropy % value and blast value

Protein 1 (PDB ID)	Protein 2 (PDB ID)	Entropy-based Similarity (%)	BLASTP Percent Identity (%)
1NBL	1JXX	99.67	100
1NBL	1CBN	99.67	96.7
1NBL	1JXW	99.67	98.48
1NBL	1JXY	99.67	98.48
1NBL	2VGH	47.54	53
1JXX	1CBN	100	96.2
1JXX	1JXW	100	98.46
1JXX	1JXY	100	98.46
1JXX	2VGH	48.11	50
1CBN	1JXW	100	95.45
1CBN	1JXY	100	85.45
1CBN	2VGH	48.6	50
1JXW	1JXY	100	96.97

Protein 1 (PDB ID)	Protein 2 (PDB ID)	Entropy-based Similarity (%)	BLASTP Percent Identity (%)
1JXW	2VGH	48.11	58
1JXY	2VGH	48	58

From the results, several protein pairs show similarity values close to 100%. This is because the entropy measure is calculated based on physicochemical property distributions rather than sequence alignment. Proteins belonging to closely related families may exhibit similar physicochemical distributions, resulting in similarity values close to 100%.

4 Conclusion

In this work, a mathematical and information- theoretic framework for protein sequence comparison is presented. The proposed entropy-based similarity method for protein sequences are presented in this paper. Three physicochemical properties are integrated into a unified entropy measure. The method proves to be a biologically meaningful and a simpler mathematical alternate to traditional alignment based technique. The methodshows that the similarity trend is consistent with alignment-based approaches. This method shows a similarity trend comparable with BLASTP results. The simplicity, scalability, interpretability of the method makes it suitable for large protein datasets and functional classification task.

References

1. E. Hamori, J. Ruskin, H Curves, A novel method of representation of nucleotide series especially suited for long DNA sequences, J. Biol. Chem. 258 ,1318–1327,(1983).
2. M. A. Gates, Simpler DNA sequence representations, Nature 316, 219–219,(1985).
3. H. J. Jeffrey, Chaos game representation of gene structure, Nucleic Acids Res. 18 ,2163–2170, (1990).
4. P. M. Leong, S. Morgenthaler, Random walk and gap plots of DNA sequences, Comput. Appl. Biosci. 11,503–507,(1995).
5. M. Randić, M. Vračko, A. Nandy, S. C. Basak, On 3-D graphical representation o DNA primary sequences and their numerical characterization, J. Chem. Inf. Comput. Sci. 40, 1235–1244. (2000).
6. M. Randić, Condensed representation of DNA primary sequences, J. Chem. Inf. Comput. Sci. 40,50–56,(2000).
7. C. Yu, M. Deng, S. S. T. Yau, DNA sequence comparison by a novel probabilistic method, Inform. Sci. 181,1484–1492,(2011).
8. Y. H. Yao, Q. Dai, X. Y. Nan, P. A. He, Z. M. Nie, S. P. Zhou, Y. Z. Zhang, Analysis of similarity/dissimilarity of DNA sequences based on a class of 2D graphical representation, J. Comput. Chem. 29,1632–1639, (2008).
9. C. Yu, Q. Liang, C. Yin, R. L. He, S. S. T. Yau, A novel construction of genome space with biological geometry, DNA Res. 17,155–168,(2010).
10. S. Ding, Q. Dai, H. Liu, T. Wang, A simple feature representation vector for phylogenetic analysis of DNA sequences, J. Theor. Biol. 265,618–623,(2010).
11. M. K. Gupta, R. Niyogi, M. Misra, A new adjacent pair 2D graphical representation of DNA sequences, J. Biol. Syst. 21 (2013).
12. G. Huang, H. Zhou, Y. Li, L. Xu, Alignment–free comparison of genome sequences by a new numerical characterization, J. Theor. Biol. 281,107–112,(2011).
13. N. Jafarzadeh, A. Iranmanesh, A novel graphical and numerical representation foranalyzing DNA sequences based on codons, MATCH Commun. Math. Comput. Chem. 68, 611–620, (2012).
14. J. Song, Analysis of similarity of DNA sequences based on a novel 3-D graphical representation, in: T. Chang (Ed.), Advances in Biochemical Engineering, Inf. Engin. Res. Inst., Newark, 29–36,(2012).

15. H. J. Yu, D. S. Huang, Novel 20-D descriptors of protein sequences and it's applications in similarity analysis, Chem. Phys. Lett. 531,261–266,(2012).
16. Z. H. Qi, J. Feng, X. Q. Qi, L. Li, Application of 2D graphic representation of protein sequence based on Huffman tree method, Comput. Biol. Med. 42,556–563,(2012).
17. L. Z. X. Xie, Y. Yu, L. Liang, M. Guo, J. Song, Z. Yuan, Protein sequence analysis based on hydropathy profile of amino acids, J. Zhejiang Univ. Sci. B 13,152–158,(2012).
18. B. Liao, B. Liao, X. Lu, Z. Cao, A novel graphical representation of protein sequences and its application, J. Comput. Chem. 32 ,2539–2544,(2011).
19. C. Yu, S. Y. Cheng, R. L. He, S. S. T. Yau, Protein map: An alignment–free sequence comparison method based on various properties of amino acids, Gene 486 ,110–118,(2011).
20. M. I. Abo el Maaty, M. M. Abo–Elkhier, M. A. AbdElwahaab, 3D graphical representation of protein sequences and their statistical characterization, Physica A 389, 4668–4676,(2010).
21. Z. C. Wu, X. Xiao, K. C. Chou, 2D-MH: A web–server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids, J. Theor. Biol. 267,29–34,(2010).
22. Y. H. Yao, Q. Dai, L. Li, X. Y. Nan, P. A. He, Y. Z. Zhang, Similarity/dissimilarity studies of protein sequences based on a new 2D graphical representation, J. Comput. Chem. 31,1045–1052,(2010).
23. J. Wen, Y. Zhang, A 2D graphical representation of protein sequence and its numerical characterization, Chem. Phys. Lett. 476,281–286,(2009).
24. M. Randić, K. Mehulić, D. Vukičević, T. Pisanski, D. Vikić–Topić, D. Plavšić, Graphical representation of proteins as four–color maps and their numerical characterization, J. Mol. Graph. Model. 27,637–641,(2009).
25. X. Xia, W. H. Li, What amino acid properties affect protein evolution? J. Mol. Evol. 47, 557–564,(1998).
26. Robert Spicer, NilanjanaRaychawdhary, ShivaramDanwada, Priscilla Udomprasert, Cheryl Seals &Sutanu Bhattacharya "Evaluating the significance of embedding-based protein sequence alignment with clustering and double dynamic programming for remote homology", Scientific Reports volume 15, Article number: 39601 (2025).
27. M. Randi´c, J. Zupan, and A. T. Balaban, "Unique graphical representation of protein sequences based on nucleotide triplet codons," Chemical Physics Letters, vol. 397, no. 1–3, 247–252, (2004).
28. F. Bai and T. Wang, "A 2-D graphical representation of protein sequences based on nucleotide triplet codons," Chemical Physics Letters, vol. 413, no. 4–6, 458–462, (2005)
29. Y.h. Yao, F. Kong, Q. Dai, and P.-a. He, "A sequence-segmented method applied to the similarity analysis of long protein sequence," MATCH: Communications in Mathematical and in Computer Chemistry, vol. 70, no. 1,431–450, (2013).
30. Benjamin Giovanni Iovino & Yuzhen Ye "Protein embedding based alignment"BMC Bioinformatics, 25, 85 (2024).
31. Saeedeh Akbari RoknAbadi,Azam Sadat AbdosalehiFaezehPouyamehr&SomayyehKoohi"An accurate alignment-free protein sequence comparator based on physicochemical properties of amino acids".Scientific Reports volume 12, Article number: 11158 (2022)
32. Ajay Kumar Saw, Binod Chandra Tripathy&Soumyadeep Nandi," Alignment-free similarity analysis for protein sequences based on fuzzy integral".Scientific Reports volume 9, Article number: 2775 (2019)
33. Yongbin Zhao, Xiaohong Li, Zhaohui Qi," Novel 2D Graphic Representation of Protein Sequence and Its Application" Journal of Fiber Bioengineering and Informatics 7:1 23–33 doi:10.3993/jfbi03201403(2014).