Alignment-Free Approaches

Subjects: Biochemistry & Molecular Biology
Contributor: Guillermin Agüero-Chapin

Alignment-free (AF) methodologies have increased in popularity in the last decades as alternative tools to alignment-based (AB) algorithms for performing comparative sequence analyses. They have been especially useful to detect remote homologs within the twilight zone of highly diverse gene/protein families and superfamilies. The most popular alignment-free methodologies, as well as their applications to classification problems, have been described in previous reviews. Despite a new set of graph theory-derived sequence/structural descriptors that have been gaining relevance in the detection of remote homology, they have been omitted as AF predictors when the topic is addressed. Here, we first go over the most popular AF approaches used for detecting homology signals within the twilight zone and then bring out the state-of-the-art tools encoding graph theory-derived sequence/structure descriptors and their success for identifying remote homologs.

Keywords: alignment-free ; popular ; remote homolgs

We compile the most popular alignment-free methods applied to the detection of homologous sequences within the twilight zone of alignment algorithms. Such homologous proteins placed at this zone or beyond are known as remote homologs.

1. Word Frequency-Based Methods

The most popular AF approaches are based on word frequency counting, known as word-based methods. They estimate how many times a letter from the DNA or protein alphabets appears along the query sequence, or alternatively, they can also count the occurrences of certain subsequence of length k, where k size must be smaller than the query sequence length. Thus, they encompass those AF methods based on nucleotide [1], amino acid [2] and pseudo compositions [3], and others related to subsequence frequencies like k-mers or k-words [4], spaced k-words [5] and k-tuples [6]. They all have been applied up to certain extent in database searching, gene annotation, comparative genomics and phylogenetics by using AF similarity measures, to cope with the previously mentioned alignment's handicaps. For example, amino acid composition (ACC) was implemented in a webserver named Composition based Protein identification (COPid) to perform protein searches and phylogenetic analysis by means of AF distances https://webs.iiitd.edu.in/raghava/COPid/ [2]; but also has been applied to detect remote homology in the G-protein coupled receptor superfamily (GPCR) [7]. The GPCR family has represented a challenging target, due to its high sequence diversity, for studying the prediction performance of several AF tools [8] including the pseudo amino acid composition (PseAAC) protein feature [9][10].

The Chou's PseAAC concept was firstly applied to predict protein cellular attributes related to the biological function regardless alignment information [11]. This AF approach incorporated the sequence order effect to the ACC to improve the quality of predictions. It was implemented in a webserver hosted at http://www.csbio.sjtu.edu.cn/bioinf/PseAA/ [12]. The performance of PseACC has been evaluated in the twilight zone by (i) identifying enzymatic signatures and delimiting their subclasses in a non-redundant subset of enzymes and non-enzymes sharing sequence similarities lower than 40% of identity [13] (ii) classifying structurally characterized proteins sharing < 30% of similarity into the four SCOPe' classes (α, β, α/β, α + β) just having sequence primary information [14] and (iii) detecting remote protein homologous using benchmark datasets [15]. In addition to the proven utility of other compositional AF features like k-mers in assembling reads from NGS technologies into contigs [16], identification of species in metagenomic samples [17][18] and improving heterologous gene expression [19]; they have been applied to overcome several handicaps found in the twilight zone such as (i) the annotation of protein families within the metagenome's diversity [20], (ii) the classification of structural protein classes in designed datasets sharing low sequence similarities just by using k-word frequencies or AF distances [21][22], or by k-mers incorporation into the general scheme of PseACC [23], and (iii) the phylogeny reconstruction for constantly-evolving viral genomes by the estimation of alignment-free distances [24][25]

Popular AF methods based on compositional features have been also applied to genome or proteome-based phylogeny reconstructions [26][27] because they circumvent some well-known problems arising when intending the alignment of large genomic sequences, finding orthologs to build species trees and dealing with low homology genes/proteins [5][28]; instead

they can estimate directly AF distances from unassembled Next Generation Sequencing (NGS) reads for phylogenetic tree building [29].

Last but not least, many of the previously-mentioned word-based methods, have been also exploited to detect, analyze and compare the less conserved blocks of the genomes made up by regulatory regions including promoters, transcription factors and enhancers [30]. In this sense, the D2z AF measure derived from k-words frequencies highlights as one of the first reports in detecting functional and/or evolutionary similarities among cis-regulatory modules (CRMs) from several tissues of human's and Drosophila's genomic sequences [31]. One year later, k-words distributions were added directly to Markov models to define new AF similarity measures to discriminate functionally related CRMs from the unrelated ones [32]. In 2010, the concept of Regulatory Region Scoring (RRS), based on the potential distribution of the transcription factors in CRMs, was introduced as an AF prediction model for the detection of related functional signals in non-alignable enhancers found in the CRMs; but could also be extended to other regulatory sequences like promoters [33]. More details about the definition and application of the most popular AF methods and measures were addressed by Vinga and Almeida in several outstanding reviews [34][35][36].

2. Information Theory-Based Methods

The runners up of most popular AF methods are those based on the information theory which measure the information contained in the organization of DNA and protein strings using different approaches. For example, the Kolmogorov complexity of a sequence is measured through the shortest description of its string. However, such abbreviated description of the string is really expressed as a "compression" measure like the ".zip files". As longer and more complex is the sequence, a larger description would be needed and therefore less compression of its string would be possible to apply [37]. Another type of complexity information measure is the Lempel–Ziv complexity that calculates the number of different substrings (occurrence rates) found along the sequence. The number of iterations needed to find such substring occurrences is related with the complexity of the sequence [38]. Once the Kolmogorov's and Lempel–Ziv's complexities are determined for the sequences, the estimation of similarity or distance metrics can be easily computed [37][39][40]. In this sense, compression-based distance measures from Lempel–Ziv's and Kolmogorov's complexities were used to detect distant protein similarities in a subset of the SCOP protein structure database [41][42], and to classify non-homologous domains into the CATH levels (class, architecture, and topology) [42], respectively.

The so-called Universal Similarity Metric introduced by Li et al. 2001 [43] lying on the Kolmogorov complexity concept showed success to cluster protein structures sharing low sequence similarity within structural families and subfamilies [44].

Another theory-based measure is the Shannon entropy defined as the uncertainty of finding a given symbol (nucleotide or amino acid) or word (L-tuples) in the analyzed sequence [45]. The Shannon entropy concept has been used to estimate Kullback–Leibler (KL) divergence measure that allowed the comparison of two sequences [46][47]. The Shannon entropy has been recently applied to relieve the perturbation caused by several biological processes such as mutations, recombinations, insertions and deletions and fast-evolving genomes on pairwise effective genome comparisons [48].

AF methods based on the information theory have been also applied to characterize/compare regulatory sequences [49][50] and to identify/compare transcription factor binding sites [51][52]. For further details about the application of information theory-based AF methods to non-coded sequence analysis, one may go through a comprehensive review published by Vinga [53]. At last, **Table 1** shows a summary of the most popular AF methods applied to datasets of low sequence similarity for remote homology detection and the clustering of similar protein structures under such conditions.

Table 1. Summary of the most popular AF features applied to detect remote homology

Word-frequency methods

AF feature	Low-similarity dataset	Web-Implementation	Ref.
Amino Acid Composition (ACC)	G-protein coupled receptor superfamily	COPid https://webs.iiitd.edu.in/raghava/COPid/	[7]

Pseudo Amino Acid (PseACC)	G-protein coupled receptor superfamily	http://www.csbio.sjtu.edu.cn/bioinf/PseAA/	[9] [10]
PseACC	Designed dataset identity from ENZYME SwissPro database in [13]	http://chou.med.harvard.edu/bioinf/EzyPred/	[13]
PseACC	Chou's designed dataset [54] from SCOP structural classes	http://www.csbio.sjtu.edu.cn/bioinf/PseAA/	[14]
k-mers	Benchmark Structural data designed based on [55][56]	No publicly available for proteins	[21]
k-mers	Benchmark Structural data designed in [57] and also used by [58]	No publicly available for proteins	[22]
Information theory-based methods			
Lempel–Ziv complexity	Subset of SCOP designed by [57]	No publicly available	[41]
Kolmogorov complexity	Subset of SCOP designed by [57]	No publicly available	[41]
Kolmogorov complexity (Universal Similarity Metric)	Benchmark Structural data <25% designed based on [55][56]	No publicly available	[42]

Kolmogorov complexity (Universal Similarity Metric)	Clustering protein structures using at low sequence similarity Benchmark Structural data [56]	http://www.cs.nott.ac.uk/~nxk/USM/protocol.html	[44]

References

1. Guo, F.-B., et al., Accurate prediction of human essential genes using only nucleotide composition and association information. Bioinformatics, 2017. 33(12): p. 1758-1764.
2. Kumar, M., V. Thakur, and G.P. Raghava,; COPid: composition based protein identification. *In Silico Biol* **2008**, *8*, 121-128, 10.2305/iucn.uk.2013-1.
3. Kuo-Chen Chou; Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of Theoretical Biology* **2011**, *273*, 236-247, 10.1016/j.jtbi.2010.12.024.
4. Gunasinghe, U., D. Alahakoon, and S. Bedingfield, Extraction of high quality k-words for alignment-free sequence comparison. Journal of theoretical biology, 2014. 358: p. 31-51.
5. Leimeister, C.-A., et al., Fast alignment-free sequence comparison using spaced-word frequencies. Bioinformatics, 2014. 30(14): p. 1991-1999.
6. Chen, W., et al., PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. Anal Biochem, 2014. 456: p. 53-60.
7. Elrod, D.W. and K.C. Chou; A study on the correlation of G-protein-coupled receptor types with amino acid composition. *Adv Exp Med Biol* **2002**, *15*, 713-715, 10.1007/bf02787095.
8. Makiko Suwa; Bioinformatics Tools for Predicting GPCR Gene Functions. *Advances in Experimental Medicine and Biology* **2013**, *796*, 205-224, 10.1007/978-94-007-7423-0_10.
9. Gu, Q., Y.S. Ding, and T.L. Zhang; Prediction of G-protein-coupled receptor classes in low homology using Chou's pseudo amino acid composition with approximate entropy and hydrophobicity patterns. *Protein Pept Lett* **2010**, *17*, 559-67, 10.1111/j.1540-8159.1985.tb05824.x.
10. Qiu, J.D., et al., Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform. Anal Biochem, 2009. 390(1): p. 68-73.
11. K C Chou; Prediction of protein cellular attributes using pseudo-amino acid composition.. *Proteins: Structure, Function, and Bioinformatics* **2001**, *43*, , .
12. Shen, H.B. and K.C. Chou; PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal Biochem* **2008**, *373*, 386-8, 10.1093/nq/193.25.545.
13. Liu, B., et al., Protein Remote Homology Detection by Combining Chou's Pseudo Amino Acid Composition and Profile-Based Protein Representation. Mol Inform, 2013. 32(9-10): p. 775-82.
14. Ding, Y.S., T.L. Zhang, and K.C. Chou; Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. *Protein Pept Lett* **2007**, *14*, 811-5, 10.1057/9781137343581.0018.
15. Liu, B., et al., Protein Remote Homology Detection by Combining Chou's Pseudo Amino Acid Composition and Profile-Based Protein Representation. Mol Inform, 2013. 32(9-10): p. 775-82.
16. Compeau, P.E.C., P.A. Pevzner, and G. Tesler, How to apply de Bruijn graphs to genome assembly. Nature Biotechnology, 2011. 29: p. 987
17. Ames, S.K., et al., Scalable metagenomic taxonomy classification using a reference genome database. Bioinformatics, 2013. 29(18): p. 2253-60.
18. Ounit, R., et al., CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics, 2015. 16: p. 236.

19. Gustafsson, C., S. Govindarajan, and J. Minshull; Codon bias and heterologous protein expression.. *Trends Biotechnol* **2004**, *22*, 346-53, 10.1177/000944556500100405.

20. Edwards, R.A., et al.; Real time metagenomics: using k-mers to annotate metagenomes. *Bioinformatics* **2012**, *28*, 3316-7, 10.1016/j.jce.2003.09.006.

21. Dai, Q. and T. Wang, Comparison study on k-word statistical measures for protein: from sequence to 'sequence space'. BMC Bioinformatics, 2008. 9: p. 394.

22. Lingner, T. and P. Meinicke; Remote homology detection based on oligomer distances . *Bioinformatics* **2006**, *22*, 2224-31, 10.1093/ptep/ptw132.

23. Qin, Y.F., et al.; Predicting protein structural class by incorporating patterns of over-represented k-mers into the general form of Chou's PseAAC. *Protein Pept Lett* **2012**, *19*, 388-97, 10.15421/40260307.

24. Domazet-Loso, M. and B. Haubold; Alignment-free detection of local similarity among viral and bacterial genomes. *Bioinformatics* **2011**, *27*, 1466-72, 10.1007/10522884_1946.

25. Hohl, M. and M.A. Ragan, Is multiple-sequence alignment required for accurate inference of phylogeny? Syst Biol, 2007. 56(2): p. 206-21

26. Cheong Xin Chan; Mark A Ragan; Next-generation phylogenomics. *Biology Direct* **2013**, *8*, 3-3, 10.1186/1745-6150-8-3.

27. Qi, J., H. Luo, and B. Hao, CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic acids research, 2004. 32(suppl 2): p. W45-W47

28. Kang, Y., et al; PVTree: A Sequential Pattern Mining Method for Alignment Independent Phylogeny Reconstruction. *Genes (Basel), 2019. 10(2)* **2019**, *10*, 283-293, xxxx.

29. Song, K., et al; Alignment-free sequence comparison based on next-generation sequencing reads. *J Comput Biol* **2013**, *20*, 64-79, 10.1007/bf00420751.

30. Song, K., et al., New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. Brief Bioinform, 2014. 15(3): p. 343-53.

31. Miriam R. Kantorovitz; Gene E. Robinson; Saurabh Sinha; A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics* **2007**, *23*, 249-55, 10.1093/bioinformatics/btm211.

32. Qi Dai; Yanchun Yang; Tianming Wang; Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison. *Bioinformatics* **2008**, *24*, 2296-2302, 10.1093/bioinformatics/btn436.

33. Koohy, H., et al., An alignment-free model for comparison of regulatory sequences. Bioinformatics, 2010. 26(19): p. 2391-7.

34. Susana Vinga; Jonas Almeida; Alignment-free sequence comparison-a review.. *Bioinformatics* **2003**, *19*, 513-523, 10.1093/bioinformatics/btg005.

35. Susana Vinga; Editorial: Alignment-free methods in computational biology. *Briefings in Bioinformatics* **2014**, *15*, 341-342, 10.1093/bib/bbu005.

36. Andrzej Zielezinski; Susana Vinga; Jonas Almeida; Wojciech M. Karlowski; Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology* **2017**, *18*, 186, 10.1186/s13059-017-1319-7.

37. Li, M. and P.M.B. Vitányi, An introduction to Kolmogorov complexity and its applications. 3rd ed. Texts in computer science. 2008, New York: Springer. xxiii, 790 p.

38. Lempel, A. and J. Ziv, On the complexity of finite sequences. IEEE Transactions on information theory, 1976. 22(1): p. 75-81

39. Otu, H.H. and K. Sayood, A new sequence distance measure for phylogenetic tree construction. Bioinformatics, 2003. 19: p. 2122-2130

40. Shihyen Chen; Bin Ma; Kaizhong Zhang; On the similarity metric and the distance metric. *Theoretical Computer Science* **2009**, *410*, 2365-2376, 10.1016/j.tcs.2009.02.023.

41. András Kocsor; Attila Kertész-Farkas; László Kaján; Sándor Pongor; Application of compression-based distance measures to protein sequence classification: a methodological study. *Bioinformatics* **2005**, *22*, 407-412, 10.1093/bioinformatics/bti806.

42. Paolo Ferragina; Raffaele Giancarlo; Valentina Greco; Giovanni Manzini; Gabriel Valiente; Compression-based classification of biological sequences and structures via the Universal Similarity Metric: experimental assessment. *BMC Bioinformatics* **2007**, *8*, 252-252, 10.1186/1471-2105-8-252.

43. Ming Li; Jonathan H. Badger; Xin Chen; Sam Kwong; Paul Kearney; Haoyong Zhang; An information-based sequence distance and its application to whole mitochondrial genome phylogeny.. *Bioinformatics* **2001**, *17*, 149-154, 10.1093/bioinformatics/17.2.149.

44. N. Krasnogor; D. A. Pelta; Measuring the similarity of protein structures by means of the universal similarity metric. *Bioinformatics* **2004**, *20*, 1015-1021, 10.1093/bioinformatics/bth031.

45. B.J. Strait; T.G. Dewey; The Shannon information entropy of protein sequences.. *Biophysical Journal* **1996**, *71*, 148-155, 10.1016/s0006-3495(96)79210-x.

46. S. Kullback; R. A. Leibler; On Information and Sufficiency. *The Annals of Mathematical Statistics* **1951**, *22*, 79-86, 10.1214/aoms/1177729694.

47. Fei Nan; Nald Adjeroh; On complexity measures for biological sequences. *Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004.* **2004**, *5*, 501-505, 10.1109/csb.2004.1332483.

48. Mehul Jani; Rajeev K. Azad; Information entropy based methods for genome comparison. *ACM SIGBioinformatics Record* **2013**, *3*, 1-4, 10.1145/2500124.2500126.

49. Dianhui Wang; Sarwar Tapan; MISCORE: a new scoring function for characterizing DNA regulatory motifs in promoter sequences. *BMC Systems Biology* **2012**, *6*, S4-S4, 10.1186/1752-0509-6-S2-S4.

50. Matteo Comin; Morris Antonelli; Fast Alignment-free Comparison for Regulatory Sequences using Multiple Resolution Entropic Profiles. *International Conference on Bioinformatics Models, Methods and Algorithms* **2015**, *2*, 171-177, 10.5220/0005251001710177.

51. Ivan Erill; Michael C O'neill; A reexamination of information theory-based methods for DNA-binding site identification. *BMC Bioinformatics* **2009**, *10*, 57-57, 10.1186/1471-2105-10-57.

52. Minli Xu; Zhengchang Su; A Novel Alignment-Free Method for Comparing Transcription Factor Binding Site Motifs. *PLOS ONE* **2010**, *5*, e8797, 10.1371/journal.pone.0008797.

53. Vinga, S., Information theory applications for biological sequence analysis. Brief Bioinform, 2014. 15(3): p. 376-89.

54. Liu, B.; Wang, X.; Zou, Q.; Dong, Q.; Chen, Q. Protein Remote Homology Detection by Combining Chou’s Pseudo Amino Acid Composition and Profile-Based Protein Representation. Mol. Inf. 2013, 32, 775–782.

55. Compeau, P.E.C.; Pevzner, P.A.; Tesler, G. How to apply de Bruijn graphs to genome assembly. Nat. Biotechnol. 2011, 29, 987.

56. Ames, S.K.; Hysom, D.A.; Gardner, S.N.; Lloyd, G.S.; Gokhale, M.B.; Allen, J.E. Scalable metagenomic taxonomy classification using a reference genome database. Bioinformatics 2013, 29, 2253–2260.

57. Ounit, R.; Wanamaker, S.; Close, T.J.; Lonardi, S. CLARK: Fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genom. 2015, 16, 236.

58. Gustafsson, C.; Govindarajan, S.; Minshull, J. Codon bias and heterologous protein expression. Trends Biotechnol 2004, 22, 346–353.
