In Silico Biology 8, 0011 (2008); ©2008, Bioinformation Systems e.V.

COPid: Composition based protein identification

Manish Kumar, Varun Thakur and Gajendra P. S. Raghava*

Institute of Microbial Technology, Sector 39-A, Chandigarh, India

* Corresponding author

Email: raghava@imtech.res.in
Phone: +91-172-2690557 or 2690225; Fax: +91-172-2690632 or 2690585
Web: http://www.imtech.res.in/raghava/

Edited by H. Michael; received October 29, 2007; revised January 12, 2008; accepted January 28, 2008; published April 05, 2008

Abstract

In the past, a large number of methods have been developed for predicting various characteristics of a protein from its composition. In order to exploit the full potential of protein composition, we developed the web-server COPid to assist the researchers in annotating the function of a protein from its composition using whole or part of the protein. COPid has three modules called search, composition and analysis. The search module allows searching of protein sequences in six different databases. Search results list database proteins in ascending order of Euclidean distance or descending order of compositional similarity with the query sequence. The composition module allows calculation of the composition of a sequence and average composition of a group of sequences. The composition module also allows computing composition of various types of amino acids like (e. g. charge, polar, hydrophobic residues). The analysis module provides the following options; i) comparing composition of two classes of proteins, ii) creating a phylogenetic tree based on the composition and iii) generating input patterns for machine learning techniques. We have evaluated the performance of composition-based (or alignment-free) similarity search in the subcellular localization of proteins. It was found that the alignment free method performs reasonably well in predicting certain classes of proteins. The COPid web-server is available at http://www.imtech.res.in/raghava/copid/.

Keywords: compositional biasness, alignment-free similarity search, composition-based phylogenetic tree, residue composition, dipeptide composition

Introduction

The similarity search is one of the oldest and most powerful techniques that are commonly used for predicting the function of a protein. Over the years, there has been tremendous progress in this field, which resulted in

the development of number of techniques that include DOTPLOT; dynamic programming [1]; approximation techniques [2, 3] and profile based searching [4]. These programs are based on alignment where they compute identity/similarity between two sequences by maintaining the order of residues. The significance of similarity depends on alignment score and length of alignment. The above-mentioned similarity-based methods have their own limitations; they fail in case of far divergent evolution where sequence identities are below 25%. For example, in the large globin family, sequence similarity between the most distant relatives is not detectable using standard alignment techniques [5, 6].

In the past, an alternate approach for predicting various functions of a protein from its simple amino acid or dipeptide composition has been used successfully. Few notable examples where composition of a protein has been used for predicting its function are: i) structural class (e. g. α, β, α / β, α + β and irregular) prediction of a protein [7]; ii) classification of proteins into intracellular and extracellular enzymes and intra- and extracellular non-enzymes [8]; iii) subcellular localization of proteins [9-12]; iv) mitochondrial proteins [13]; v) prediction of secondary structure content and structural class [14, 15]. A slightly different notation of composition is also popular which is termed as "pseudo amino acid composition". It has also been widely used in different classification problems such as protein subcellular localization [16, 17], enzyme functional class [18], protein structural class[19], protein fold pattern [20], conotoxin superfamily classification [21]. These studies have firmly established the influence of amino acid and dipeptide composition of proteins on its behavior and function. In order to understand the relation between composition and function of a protein, it is important to study the composition of a protein or group of proteins having specific function, however, the existing methods/web-servers provide limited operations on proteins like computation of amino acid composition.

In this study, an attempt has been made to exploit the potential of protein composition. We have developed a comprehensive web-server COPid for performing various tasks based on the protein composition that includes similarity search, phylogenic tree, patterns for machine learning techniques etc. In order to demonstrate the utility of composition-based similarity search, we have evaluated the performance of composition-based similarity search in subcellular localization of proteins.

Materials and methods

Amino acid and dipeptide composition: The amino acid composition is a quantitative measure of fraction of each amino acid type within a protein (eq. 1). Dipeptide composition encapsulates information about the fraction of amino acids as well as their local order by a fixed length pattern of 400 (20×20) possible dipeptides (eq. 2).

$$comp(i) = \frac{R_i}{N} \times 100 \quad (1)$$

where $comp(i)$ is the percent composition of a residue of type i. R_i and N are number of residues of type i, and total the number of residues in protein i (length of protein) respectively.

$$dpep(i) = \frac{D_i}{N} \times 100 \quad (2)$$

where $dpep(i)$ is the fraction or composition of dipeptide type i. D_i and N are the number of dipeptide of type i and the number of overlapping peptides in protein i, respectively.

Compositional distance: The degree of similarity is inferred from the Euclidian distance between composition (amino acid/dipeptide) of query and target sequence. Following equations were used to calculate Euclidean distances:

$$ED_r\,(x1,x2) = \sqrt{\sum_{i=1}^{20}\left(comp_{x1}\,(i) - comp_{x2}\,(i)\right)^2} \quad (3)$$

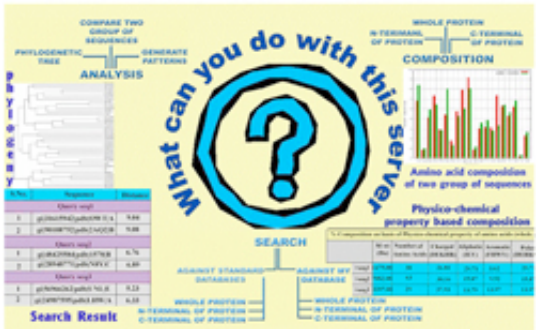
$$ED_d\,(x1,x2) = \sqrt{\sum_{i=1}^{400}\left(dpep_{x1}\,(i) - dpep_{x2}\,(i)\right)^2} \quad (4)$$

where $ED_r(x1, x2)$ and $ED_d(x1, x2)$ are the Euclidian distances between residue or dipeptide composition of protein $x1$ and $x2$.

Description of Web Server

COPid is hosted on a SUN server under Solaris Environment and is accessible from http://www.imtech.res.in/raghava/copid/. Most of the web pages are written in HTML and JavaScript, whereas the common gateway interface is written using Perl. The backend programs are written in Perl or C language depending on the complexity of arithmetic calculations (simple programs in Perl and CPU intensive programs in C). In order to create on-line plots, we have used Gnuplot. In addition, a mirror site has been established which is accessible from http://bic.uams.edu/raghava/copid/. The mirror site is hosted on an Apple Macintosh's bioinformatics cluster. COPid is a simple web-server that can be used even by users with very little knowledge of computers.

Basic user interface: The user will interact with COPid via a central web portal, which displays all the functionalities (Fig. 1). On clicking, users will be directed to the specific web form prompting for input data and parameters. The user can navigate between different application/modules by the collapsible menu bar. Input data can be submitted in single letter amino acid code in standard FASTA format. Most parameters of COPid have online help to assist users in choosing appropriate parameters. It can be viewed as a small pop-up window just by clicking the '?' tag present right besides the parameters. The help page also contains a manual, which comprises a step-by-step description to use COPid.



Click on the thumbnail to enlarge the picture

Figure 1: Screen-shot of home page of COPid with screen-shoots of output of various modules of COPid.

Modules of COPid

Search Module: The search module allows the user to perform similarity searches using amino acid or dipeptide composition of a protein, where user can search compositionally similar proteins. In addition to the composition of an entire protein, it also allows the user to search proteins that have a similar N-terminal or C-terminal composition. In the search module, similarity is measured as inverse of Euclidian distance between the composition of proteins. This module also allows the user to submit more than one sequence for

performing similarity search using BATCH or MEAN mode. In case of BATCH mode, the server searches compositionally similar proteins in the database corresponding to each protein submitted by the user. In case of MEAN mode, the server first computes the mean composition of submitted proteins. Subsequently, it searches proteins in a database that have composition similar to the mean composition. This module allows the user to perform a similarity search against three types of database: i) standard, ii) subcellular localization and iii) functional class. In case of search against standard type databases, the server allows the user to search their sequence against sequences of the manually annotated database Swiss-Prot and the structure database Protein Data Bank (PDB). In case of search against subcellular localization type database server performs a search against ePSORTdb v.2.0. [22], a subcellular localization database of prokaryotic proteins, and eSLDB [23] a subcellular localization database of eukaryotic proteins. Along with each hit the corresponding subcellular location is also displayed (Fig. 2). The topmost hit has maximum similarity with query protein, hence the subcellular location of the first hit will be the most probable location of the query protein. In COPid only experimentally annotated proteins from ePSORTdb and eSLDB have been kept. This search can help the user in predicting subcellular location of their protein. In case of searching Functional class type databases, the server performs a similarity search against the EXPASY ENZYME database (http://www.expasy.org/enzyme/) and the PIRSF protein family database (http://pir.georgetown.edu/pirwww/dbinfo/pirsf.shtml). Search against EXPASY ENZYME will allow the user to know whether their sequence have compositional similarity with any enzyme. PIRSF is a database which classifies proteins into distinct families on the basis of evolutionary relationship. Hence search against PIRSF can be used to find out the classes which are compositionally similar to the query protein. In order to extend this facility beyond the above described databases the server allows the user to search the query against their own protein database. The server provides hits in ascending order of Euclidean distance or descending order of composition similarity. It means that top-ranking proteins have maximum similarity with the query proteins. In order to save searching time we removed identical or highly similar sequences having similarity more than 80% from above databases using CD-HIT [24]. CD-HIT generates clusters of similar sequences, a cluster contains sequences having identity more than a specified threshold. In each cluster the largest sequence is treated as the cluster representative. COPid database contains only the representative sequence of each cluster. Sequences of the search result were hyperlinked with the corresponding database to facilitate quick overview.

Click on the thumbnail to enlarge the picture

Figure 2: Search result of A1AF_CAVPO against eukaryotic subcellular localization database (eSLDB). Human secretory protein A1AT_HUMAN is the topmost hit; it means COPid predicted A1AF_CAVP0 as secretory protein. In reality A1AF_CAVP0 is a Guinea pig secretory protein (antitrypsin).

Composition Module: This module allows computing the composition of one or more than one group of sequences. In case of one group of sequences, the server computes the composition of each sequence as well as average composition. It also allows the user to compute the composition of N-terminal or C-terminal parts of proteins. This module also contains an option 'physico-chemical property composition' which allows to compute composition of various type of residues that includes composition of charged (DEKHR), aliphatic (ILV), aromatic (FHWY), polar (DERKQN), neutral (AGHPSTY), hydrophobic (CVLIMFW), positively charged (HKR), negatively charged (DE), tiny (ACDGST), small (EHILKMNPQV) and large (FRWY) residues [25].

Analysis Module: This module has three main options: The first option was developed for creating distance

matrices required for generating phylogenetic trees or clusters using the programs OC (http://bic.uams.edu/raghava/oc/) or PHYLIP (top left corner of Fig. 1). This module calculates the amino acid or dipeptide composition of a protein and then calculates the Euclidian distance between each pair of sequences using equation 3 or 4, respectively. Finally it arranges the pair-wise composition distance in a format required by PHYLIP or OC. The second option allows one to compare the composition of two groups of sequences. It calculates the amino acid composition of each sequence in a group of proteins and the average composition of proteins in the group. Finally it compares two groups of proteins in form of a bar graph (top right corner of Fig. 1). The third option allows computing input patterns required to train/test machine learning techniques like the Support Vector Machine program (SVM^{light}; http://svmlight.joachims.org/), the Artificial Neural Network program SNNS (http://www-ra.informatik.uni-tuebingen.de/SNNS/) and the nearest neighbor based method TIMBL (http://ilk.uvt.nl/software.html). The patterns for above-mentioned tools can be created on the basis of amino acid or dipeptide composition. The module also allows composition calculation of a whole protein or a specified region of a protein (N- or C-terminal sequences).

Results and discussion

In order to test the practical application of the composition based similarity approach used in the search module of COPid, we have evaluated its performance in determining the subcellular localization of eukaryotic proteins in the datasets used for the development of ESLpred [11], NNPSL [26] and SubLoc [27]. It contains a total of 2427 eukaryotic proteins (1097 nuclear, 684 cytoplasmic, 321 mitochondrial and 325 extra-cellular proteins). The composition based similarity search method was evaluated using Leave One Out Cross-Validation (LOOCV) technique, where each protein was searched against a database that didn't contain the query protein; the procedure was repeated N times (N = total number of sequences to be searched) so that each protein was used once for testing. As shown in Tab. 1, we have achieved an accuracy of 79.63% using the amino acid composition based similarity search. The performance of the dipeptide composition based similarity search has decreased significantly to 68.76%. We have achieved an average accuracy of 70.57% and 55.27% using the amino acid composition of the N- and C-terminal 20 amino acids, respectively. On the same dataset Bhasin and Raghava [11] have reported 73.60% accuracy by BLAST. These results, thus, demonstrated that simple amino acid composition might work better than the highly sophisticated similarity search methods in certain cases like subcellular localization of eukaryotic protein.

Table 1: Performance of BLAST and composition-based similarity search with composition of whole sequence and of the N- and C-terminal 20 amino acids.

Class	BLAST[a]	Composition-based similarity search			
		Whole Protein		Amino Acid Composition	
		AAC[b]	DPC[c]	N-Terminal	C-Terminal
Cytoplasmic (684)	78.07	88.99	86.84	65.94	63.74
Mitochondrial (321)	57.00	59.50	39.56	60.75	37.69
Nuclear (1097)	76.57	86.96	77.85	71.56	66.27
Extracellular (325)	82.76	83.08	70.77	84.00	53.38
Average Accuracy	**73.60**	**79.63**	**68.76**	**70.57**	**55.27**

The performance was evaluated using LOOCV on ESLpred dataset.
[a] Performance of BLAST on ESLpred dataset
[b] AAC: Amino acid composition
[c] DPC: Dipeptide composition

The similarity search methods are playing a pivotal role in the classification of proteins in the post-genomic era. These methods facilitate predicting the function or structure of a protein, provided the query sequence have high sequence similarity with any annotated protein or known protein structure. However, in absence of significant sequence similarity, these methods fail. It has been shown in previous studies [7-10] that the proteins belonging to the same cellular location or fold or structure class, exhibit a similar type of amino acid composition despite having poor sequence similarity. One of the major advantages of using protein composition is that the protein can be represented by fixed length pattern (20 for amino acid composition and 400 for dipeptide composition). Thus it is easy to develop machine-learning techniques based on the composition, as they need fixed length patterns. This encourages researchers to develop the composition-based methods for predicting the function of a protein, using pattern recognition techniques. Despite the importance of protein composition, there is no comprehensive tool available for analyzing/comparing the composition of proteins or for searching similar proteins based on composition.

The web-server COPid described in this paper is an attempt to develop a tool to perform multiple tasks. It will help researchers in understanding the function and nature of proteins from their composition. The main utility of COPid is searching of compositionally similar proteins in protein molecular database. Our study has also demonstrated that a simple alignment free similarity approach may predict subcellular localization with high accuracy, in some cases it was as good as BLAST. It is known that the proteins belonging to different cellular locations have different compositions. Thus comparison of BLAST with our composition-based method is not fair. In any case, the aim of this study is not to compare these two approaches as BLAST is much more sophisticated in comparison to composition based similarity search. The aim was to demonstrate that the composition based similarity search might be used as an alternate to sequence based similarity search in predicting certain type of proteins. The analysis module of COPid allows a user to compare two sequences/groups of sequences. This may help in understanding whether a class of proteins has similar or dissimilar composition with other class of proteins. The researchers may also check whether a given class of proteins can be predicted using the compositional based methods or not. In addition, the options like the phylogenetic tree will assist the users in creating phylogenetic trees based on the composition of proteins. This will complement the alignment based phylogenetic tree creation methods.

Acknowledgement

Authors are thankful to Dr. Alok Mondal for reading the manuscript critically. We wish to thank the Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), Govt. of India, for financial support. Manish Kumar is supported by CSIR as a senior research fellow. This manuscript has IMTech communication number 15/2006.

References

1. Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48**, 443-453.
2. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. **215**, 403-410.
3. Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA **85**, 2444-2448.
4. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J.

(1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**, 3389-3402.

5. Hobohm, U. and Sander, C. (1995). A sequence property approach to searching protein databases. J. Mol. Biol. **251**, 390-399.
6. Devos, D. and Valencia, A. (2000). Practical limits of function prediction. Proteins **41**, 98-107.
7. Nishikawa, K., Kubota, Y. and Ooi, T. (1983). Classification of proteins into groups based on amino acid composition and other characters. I. Angular distribution. J. Biochem. (Tokyo) **94**, 981-995.
8. Nishikawa, K., Kubota, Y. and Ooi, T. (1983). Classification of proteins into groups based on amino acid composition and other characters. II. Grouping into four types. J. Biochem. (Tokyo) **94**, 997-1007.
9. Nakashima, H. and Nishikawa, K. (1994). Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J. Mol. Biol. **238**, 54-61.
10. Cedano, J., Aloy, P., Pérez-Pons, J. A. and Querol, E. (1997). Relation between amino acid composition and cellular location of proteins. J. Mol. Biol. **266**, 594-600.
11. Bhasin, M. and Raghava, G. P. S. (2004). ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res. **32**, W414-419.
12. Chou, K.-C. and Elrod, D. W. (1999). Protein subcellular location prediction. Protein Eng. **12**, 107-118.
13. Kumar, M., Verma, R. and Raghava, G. P. S. (2006). Prediction of mitochondrial proteins using support vector machine and hidden Markov model. J. Biol. Chem. **281**, 5357-5363.
14. Chou, K.-C. and Cai, Y.-D. (2004). Predicting protein structural class by functional domain composition. Biochem. Biophys. Res. Commun. **321**, 1007-1009.
15. Liu, W.-m. and Chou, K.-C. (1999). Prediction of protein secondary structure content. Protein Eng. **12**, 1041-1050.
16. Chou, K.-C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins **43**, 246-255.
17. Gao, Y., Shao, S., Xiao, X., Ding, Y., Huang, Y., Huang, Z. and Chou, K.-C. (2005). Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids **28**, 373-376.
18. Chou, K.-C. (2005). Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics **21**, 10-19.
19. Chen, C., Zhou, X., Tian, Y., Zou, X. and Cai, P. (2006). Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal. Biochem. **357**, 116-121.
20. Shen, H.-B. and Chou, K.-C. (2006). Ensemble classifier for protein fold pattern recognition. Bioinformatics **22**, 1717-1722.
21. Mondal, S., Bhavna, R., Mohan Babu, R. and Ramakumar, S. (2006). Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J. Theor. Biol. **243**, 252-260.
22. Rey, S., Acab, M., Gardy, J. L., Laird, M. R., deFays, K., Lambert, C. and Brinkman, F. S. L. (2005). PSORTdb: a protein subcellular localization database for bacteria. Nucleic Acids Res. **33**, D164-168.
23. Pierleoni, A., Martelli, P. L., Fariselli, P. and Casadio, R. (2007). eSLDB: eukaryotic subcellular localization database. Nucleic Acids Res. **35**, D208-212.
24. Li, W., Jaroszewski, L. and Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics **17**, 282-283.

25. Yu, X., Cao, J., Cai, Y., Shi, T. and Li, Y. (2006). Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. J. Theor. Biol. **240**, 175-184.

26. Hua, S. and Sun, Z. (2001). Support vector machine approach for protein subcellular localization prediction. Bioinformatics **17**, 721-728.

27. Reinhardt, A. and Hubbard, T. (1998). Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res. **26**, 2230-2236.