Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Software**

A tool for the prediction of functionally important sites in proteins using a library of functional templates

Christopher J. Lanczycki[1] and Saikat Chakrabarti[1,*]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA; Saikat Chakrabarti * - E-mail: chakraba@ncbi.nlm.nih.gov; * Corresponding author

received February 05, 2008; accepted February 11, 2008; published February 22, 2008

Abstract:
Understanding and characterizing the biochemical and evolutionary information within the wealth of protein sequence and structural data, particularly at functionally important sites, is very important. A comprehensive analysis of physico-chemical properties and evolutionary conservation patterns at the molecular and biological function level is expected to yield important clues for identifying similar sites in as-yet uncharacterized proteins. We present a library of protein functional templates (PFTs) designed to represent the compositional and evolutionary conservation patterns of functional sites at the molecular and biological function level. Subsequently we developed LIMACS (LInear MAtching of Conservation Scores), a software tool that uses the template library for the prediction of functionally important sites in a multiple sequence alignment, transferring the molecular function annotation from the most-similar functional site in the template library to a predicted site.

Availability: The PFT library, the LIMACS program and source code are available for PC, Mac and Linux operating systems from ftp://ftp.ncbi.nih.gov/pub/lanczyck/limacs.

Keywords: prediction; proteins; functional templates; library

Background:
Determining the functional importance of proteins remains a challenging task despite it being several years into the post-genomic era. Several computational efforts have been undertaken to derive functional insights using three-dimensional structural information **[1, 2]**, sequentially conserved residues **[3-5]**, multiple sequence alignments (MSAs) **[6-10]** and information extracted from experimental studies and literature searches. Despite these varied efforts, the accuracy of functional site prediction remains low. Prediction accuracy can suffer in part from the lack of comprehensive knowledge of evolutionary conservation patterns of amino acid residues for a wide range of known functionally important sites. Additional factors such as the availability of high quality MSAs, accurate identification of domain structure and distant homologous sequences may also affect the success of protein function prediction.

Here, we present a library of protein functional templates (PFTs) along with the companion software tool LIMACS (LInear MAtching of Conservation Scores), which uses the PFT library to predict functional sites for a query MSA. Each PFT is derived from a specific column of a high-quality MSA where a known functionally important site exists, and summarizes the compositional and evolutionary conservation patterns present at that location. The Conserved Domain Database (CDD) **[11]** maintains such functionally-annotated MSAs for a wide range of protein families. The curated CDD alignments and their linked functional annotations have been used to compute the various quantitative measures of conservation that define a PFT at each known functionally important site. The LIMACS software implements an algorithm designed to predict functionally important sites in a query multiple alignment, transferring the molecular function annotation from the most-similar PTF to the predicted query sites. This simple prediction approach is solely based on information derived from homologous sequences and no structural information is required. As a result we envisage this method as being extremely useful in the context of large scale functional annotation.

Methodology:
PFT library
A library of protein functional templates has been assembled by creating a PFT for each of the 7108 functionally important sites specified in a set of 340 curated alignments taken from the CDD (version 2.09).

The CDD alignments for these families contain functionally important sites (*e.g.,* active sites, ligand binding sites, protein binding sites) identified in an extensive manual curation effort, and are based on evidence available in the literature and other relevant scientific sources **[11]**.

Each PFT has a quantitative portion designed to represent the compositional and evolutionary characteristics of the functional site from which it was derived. The former is given by its functional group composition pattern expressed in terms of a ten-letter reduced amino acid alphabet that organizes the twenty standard amino acids based on their physico-chemical properties (see Table SM1 in Supplementary material). A composition pattern vector is defined, the elements of which represent the fraction of a column's residues in each of the ten reduced functional groups. To quantify the degree of evolutionary conservation we include the information content, median PSSM (Position Specific Scoring Matrix) score, frequency of negative PSSM scores and relative weight of negative PSSM scores in the PFT **[12]**.

Each template also has a qualitative portion comprising its molecular and biological functional assignment. Assignments of molecular and biological function have been performed by the authors through an extensive survey of the available literature and experimental references. Tables SM2 and SM3 (see supplementary material) specify the six molecular and sixteen different biological functional categories used, and Figure 1 provides the representation of each within the PFT template library.

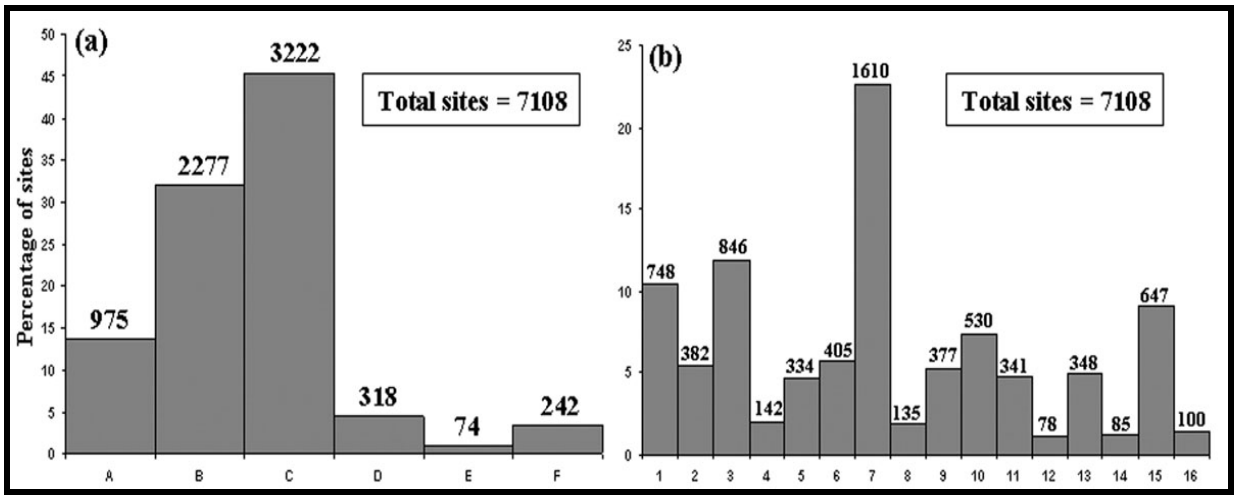

Figure 1: Percentage of sites belonging to six molecular and sixteen biological functional categories within the PFT library. Percentage of sites is plotted for each molecular (a) and biological (b) functional category. Actual numbers of the sites are shown on top of each bar. Codes and numbering of each functional category are same as described in Table SM2 and Table SM3 in Supplementary material.

LIMACS program

LIMACS employs the same scoring function we introduced previously to examine the feasibility of using a template-based approach to functional site prediction **[13]**. Given a multiple sequence alignment, LIMACS utilizes its heuristic match score to find those alignment columns that are significantly similar to a PFT from the library. The alignment columns so identified are the predicted functionally important sites, and the molecular function annotation of the best-scoring PFT is transferred to each predicted functional site. Only gapless columns of the query MSA are compared to the PFT library, although LIMACS can be extended to deal with gapped alignment columns. Additional details about the LIMACS scoring scheme can be found at Supplementary material.

Features of LIMACS

LIMACS accepts input MSAs in FASTA format. To help avoid false-positives, hits between query columns and PFTs are by default screened against a set of 2000 randomly aligned sequences of 500 residues to estimate statistical noise and deal with cases where the signal-to-noise ratio is low. Various scores are provided for each reported query/template hit to aid the user in interpreting the significance of individual results (see the distribution's 'README.txt' file for details.).

Prediction of functional important sites

The performance of LIMACS was measured by computing the average fraction of predicted true positives in a 5-fold cross-validation analysis, where the full template library was divided into five parts: four parts together were used as the database template library with the remaining part as test set. This procedure was

repeated five times by randomly generating different subsets. The accuracy of the prediction was calculated as the number of correctly matched functional sites divided by the total number of predictions at given match score threshold. This cross-validation analysis suggests a high accuracy (~73%) in functional site attribution.

Next, the sensitivity and specificity of the prediction algorithm were evaluated using a slightly modified 5-fold cross-validation analysis, in which we examine the ability of LIMACS to pick out functional sites among a mixture of functional and non-functional sites in a MSA. As before, the full template library was randomly divided by removing a test set containing 20% of the PFTs. But in this case we do not independently extract individual PFTs; rather, all of the PFTs in the library derived from the same CDD alignment are extracted together. This results in a test set of MSAs no longer represented among the remaining PFTs in the template library, and each of which has a set of well-defined true-positives (i.e., the 20% of PFTs extracted from the full template library). The MSAs containing PFTs in this test set are used as the input to LIMACS, which attempts to match each column (*not* just those columns corresponding to the PFTs) against a PFT from the template library. This procedure was repeated five times by randomly generating different subsets. The sensitivity of predicting correct functional sites at 15% false positive rate (error rate) is ~67% (unpublished data).

Caveats and future development:

LIMACS is written in C++ and the source code is available for download. The executables are available for Windows, Macintosh and Linux operating systems. NCBI C++ Toolkit and C++ development tools are required to build LIMACS from source code (instructions provided with the package). We have plans to develop a web version of the LIMACS program.

Acknowledgement:

This work was supported by the Intramural Research Program of the National Library of Medicine at National Institutes of Health/DHHS.

References:

[01] J. S. Fetrow & J. Skolnick, *J. Mol. Biol.,* 281: 949 (1998) [PMID: 9719646]

[02] B. Zhang, *et al*., *Protein Sci.,* 8: 1104 (1999) [PMID: 10338021]

[03] K. Hofmann, *et al*., *Nucleic Acids Res.,* 27: 215 (1999) [PMID: 9847184]

[04] M. Ashburner, *et al*., *Nature Genetics* 25: 25 (2000) [PMID: 10802651]

[05] G. Casari, *et al.*, *Nat. Struct. Biol.,* 2: 171 (1995) [PMID: 7749921]

[06] S. Hannenhalli & R. B. Russell, *J. Mol. Biol.,* 303: 61 (2000) [PMID: 11021970]

[07] L. Li, *et al*., *Proc. Natl. Acad. Sci. USA.,* 100: 4463 (2003) [PMID: 12679523]

[08] J. Pei & N. V. Grishin, *Bioinformatics*, 17: 700 (2001) [PMID: 11524371]

[09] O. Lichtarge, *et al*., *J Mol Biol*., 257: 342 (1996) [PMID: 8609628]

[10] C. Berezin, *et al*., *Bioinformatics.,* 20: 1322 (2004) [PMID: 14871869]

[11] A. Marchler-Bauer, *et al*., *Nucleic Acids Res.,* 30: 281 (2002) [PMID: 11752315]

[12] S. Chakrabarti, *et al*., *Nucleic Acids Res.,* 34: 2598 (2006) [PMID: 16707662]

[13] S. Chakrabarti & C. J. Lanczycki, *Protein Sci*., 16: 4 (2006) [PMID: 17192586]

Edited by P. Kangueane

Citation: Lanczycki and Chakrabarti, Bioinformation 2(7): 279-283 (2008)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Software**

Supplementary material:

LIMACS scoring scheme:

To compare a PFT p from the template library with a column c of an input alignment, the PFT match score $M_{p,c}$ is computed as a weighted linear combination of the differences of corresponding quantitative properties. $M_{p,c}$ is defined as:

$$M_{p,c} = W_g D_{p,c} + \sum_{x \in \{I,m,f,w\}} W_x \ \sigma^x_{p,c}. \qquad \rightarrow \qquad (1)$$

The first term $D_{p,c} = 1 - \left\| \mathbf{g}_p / \left\| \mathbf{g}_p \right\| - \mathbf{g}_c / \left\| \mathbf{g}_c \right\| \right\|$ is a measure of difference in composition pattern between p and c. A composition pattern vector **g** is defined, the elements of which represent the fraction of a column's residues in each of the ten reduced functional groups. It gives a high score when the Euclidean distance between the normalized compositional pattern vectors $\mathbf{g}_p$ and $\mathbf{g}_c$ is small. The four measures of evolutionary conservation (information content, median PSSM score, frequency of negative PSSM scores and relative weight of negative PSSM scores), represented by the scalar quantities I, m, f and w, respectively, make up the second term of the match score. Each scalar quantity x makes a contribution proportional to the relative similarity between p and c, defined as $\sigma^x_{p,c} = 1 - \left| x_p - x_c \right| / \max(x_p, x_c)$. The various W coefficients are weights constrained to sum to one, and individually must be between zero and one. Therefore a perfect match to a functional template has the maximal score of one.

Tables:

Functional group	Amino acid
Amido (AMDO)	Gln and Asn
Primary amine (AMNP)	Lys
Carboxyl (CBXL)	Asp and Glu
Guanidino (GNDO)	Arg
Hydroxyl (HDXL)	Ser, and Thr
Imidazole (IMZL)	His
Nonpolar (NONP)	Ala, Gly, Ile, Leu, Val and Pro
Phenyl (PHEN)	Phe, Trp and Tyr
Sulfur (SULF)	Met
Thiol (THIO)	Cys

Table SM1: List of functional groups. Note: Standard 3 letter code has been used to depict amino acids.

Category code	Molecular function category	Subdivision
A	Active site	All active sites or catalytic sites
B	Ligand binding site	Nucleotide binding sites, lipid binding sites, carbohydrate binding sites, small organic ligand binding sites and inorganic ligand binding sites
C	Protein binding site	Protein-protein interaction and peptide binding sites.
D	Metal binding site	All metal binding sites
E	Post translational modification site	Acetylation, cleavage, glycosylation, lipoylation and phosphorylation sites
F	Miscellaneous site	Sites involved in structural changes, disulfide bonding, hinge regions etc

Table SM2: Categories of molecular functions within the protein functional template (PFT) library.

Category ID	Biological function category
1	DNA synthesis and processing
2	Cellular transport and transport mechanism
3	Intra cellular communication/signal transduction
4	RNA processing
5	Protein fate
6	Energy
7	Metabolism
8	Cellular structural organization
9	Inter cellular communication/cellular environment
10	Transcription
11	Protein synthesis
12	Development
13	Defense stress and detoxification
14	Cell death
15	Miscellaneous (mixed functions)
16	Unknown

Table SM3: Categories of biological functions within the protein functional template (PFT) library.