Rapid assessment of correlated amino acids from pair-to-pair (P2P) substitution matrices

Eran Eyal1,*, Shmuel Pietrokovski2 and Ivet Bahar1

1Department of Computational Biology, School of Medicine, University of Pittsburgh, 3501 Fifth Avenue, Pittsburgh, PA 15213, USA and 2Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, 76100, Israel

Received on February 2, 2007; revised on April 18, 2007; accepted on May 7, 2007

Advance Access publication May 12, 2007

Associate Editor: Burkhard Rost

ABSTRACT

Identification of correlated amino acids in proteins has been a topic of broad interest in view of its functional implications and importance in protein design. A new set of pair-to-pair (P2P) substitution matrices for amino acids was recently introduced as a useful tool for inferring information on such correlated sites. We present a website developed for automated application of these matrices for analysis of query sequences. The site offers options for graphical analysis of correlations, as well as visualization of correlated amino acids on representative, structurally characterized, members of the examined family of sequences.

Availability: http://www.ccb.pitt.edu/p2p

Contact: eyal@ccbb.pitt.edu

1 INTRODUCTION

Multiple sequence alignments (MSAs) are primarily applied for identifying residues conserved among the members of a given family of proteins. In addition, MSAs may provide information on correlated pairs of residues. Such correlations usually arise from direct interactions (spatial contacts) between residues, although in some cases allosteric effects may result in correlations between distant residues. While the utility of MSA-based approaches for detecting correlated amino acids has been known for almost two decades (Altschuh et al., 1987; Gobel et al., 1994; Neher, 1994), and improved methods are being developed (Halperin, et al., 2006), their broader usage by the biological community has been limited by a few practical issues. Many theoretical and computational approaches are not available as open source tools, or accessible through user friendly interfaces. Only a few programs have been implemented to date on the web (Fleishman et al., 2004; Kass and Horovitz, 2002; Kundrotas and Alexov, 2006). Some others PlotCor (Pazos et al., 1997), CorrMut (Fleishman et al., 2004) and CRASP (Afonnikov and Kolchanov, 2004) are available for download and local usage. Fodor and Aldrich, 2004 implemented several algorithms in a Java code, available at http://www.afodor.net. The few accessible servers do not provide, however, graphical analysis tools, making it difficult to analyze the correlations and examine them with respect to structural data.

We recently introduced $20^2 \times 20^2$ matrices for simultaneous substitutions of pairs of amino acids, referred to as pair-to-pair (P2P) substitution matrices. These matrices were shown to be sensitive and useful for predicting potentially interacting residues and for rank-ordering decoy sets based on sequence information alone (Eyal et al., 2007). Here we present a web-based tool for online calculation and interactive analysis of residue correlations obtained using the P2P matrices, as well as their visualization on representative protein structures.

2 PAIR TO PAIR SUBSTITUTION MATRICES

The simultaneous substitution of the amino acids at the respective sequence positions $i$ and $j$ of a given sequence is assigned in our server a correlation score of the form

$$S(i,j) = \sum_{x,y,u,v} w_x w_y M(xy; uv) \delta_{ix} \delta_{iy} \delta_{ju} \delta_{jv}$$

Where $w_x$ and $w_y$ refer to the weights of the sequences $s$ and $t$ of the MSA (Henikoff and Henikoff, 1994), respectively, $M(xy; uv)$ is the particular element of the P2P matrix corresponding to the correlated substitutions of amino acids of type $x \leftrightarrow u$ and $y \leftrightarrow v$ at the respective $i$th and $j$th positions of the two sequences, and $\delta_{ix} = 1$ if the amino acid type of the $i$th residue in the $s$th sequence is $x$. We have developed different forms of P2P matrices, based on different versions of BLOCKS database (Henikoff et al., 1999), different types of residue pairs (intra- and inter-domain contacts, or all contacts), and either incorporating or excluding structural information. The elements of the P2P matrices that incorporate structural data are evaluated from

$$M(xy; uv) = \ln \frac{p^+(xy; uv)}{p^+(x; u)p^+(y; v)} = \ln \frac{p^-(xy; uv)}{p^-(x; u)p^-(y; v)}$$

where $p(xy; uv)$ is the probability of occurrence of the double substitution $xy \leftrightarrow uv$ in the space of all possible $400^2$ substitutions of amino acid pairs, $p(x; u)$ is the probability of the singlet substitution $x \leftrightarrow u$ and the superscripts $+$ and $-$ refer to the subsets of pairs that make ($+$), or do not make ($-$) contacts in the folded state (Eyal et al., 2007).
3 CORRELATION ANALYSIS

The P2P website permits to estimate the correlations between any pair of amino acids, for any given query sequence of amino acids and for a given MSA submitted in suitable format. Correlation scores are computed online, based on the selected P2P matrices and released with graphical options for analysis of the results.

3.1 Input options

Even though the core calculations are based on MSAs, the user has a considerable flexibility regarding the format of the input data. User-defined MSAs, Pfam accession number for pre-calculated alignments, single sequences in FASTA format or PDB structures, are all acceptable as inputs. In the last two cases, the server automatically makes a Blast search (Altschul et al., 1997) and aligns identified family members using ClustalW (Thompson et al., 1994).

3.2 Output

The results can be obtained both interactively and by email. The email option is useful when using large or long multiple alignments, as the computation time scales as $O(N^2L^2)$, where $N$ is the number of sequences in the MSA and $L$ is the sequence length. Apart from the interactive analysis tools described subsequently, easily parsed plain text output files with correlations given in matrix or list forms are available. For an assessment of the meaning of the released correlations (scores), we provide a graph showing the distribution of scores for the family, along with statistical data on the dependence of accurately predicted native contacts on the individual scores.

3.3 Graphical analysis environment

Results for correlations between pairs of residues are conveniently shown by 2D correlation maps as a function of residue index/types (Fig. 1). Red colors indicate weak association and blue colors indicate strong association. Our method, in contrast to traditional correlated mutation methods, does not suggest residues that are anti-correlated, but provides a measure of the strength of correlation, with the lowest scores corresponding to the weakest associations. Portions of these maps can be enlarged upon clicking on selected regions (Fig. 1 inset). Upon further clicking on the enlarged maps, particular pairs of residues are highlighted in the frame on the left. This frame shows the 3D structure of a representative member from the family of the query sequence, if such a structure exists, using the Jmol molecular graphics program (http://jmol.sourceforge.net/).

A rank-ordered list of correlated pairs of amino acids is provided in the lower left frame, with the amino acid identities and numbers corresponding to the first sequence in the multiple alignment on which the calculations are based. Pairs of residues in the alignment can be selected, either from the list of top correlation scores, or from the correlation map. Lines connecting the selected residue pairs are displayed with a color scheme coded after the correlation score.

Fig. 1. Analysis of correlation scores predicted by P2P matrices, illustrated for the catalytic domain of protein kinases, taken from the P2P website. The correlation map, based, in this case, on the PFAM seed alignment of the kinases catalytic domain (PF00069) is shown on the right panel. Blue colors indicate high scores and red indicate low scores. The user can zoom into any sub-region of interest. Black dots indicate amino acids with direct contact. By clicking on the cells of this matrix, the corresponding residue pairs are displayed on the structure in the Jmol applet. In this example, the residues that participate in the 10 top correlation scores, all located in the central region of the domain, are mapped on the structure of PKC (PDB 1zrz).

3.4 Local run

Our core program (termed P2PConPred) is also available for downloading, and has been tested on various Unix/Linux systems (including Linux Red Hat and Suse, Sun and SG) and on Windows (XP and Cygwin). In practice, for alignments of intermediate sizes, calculations are done within seconds on current hardware.

3.5 Programming

The program is written in C++. The website is written in Perl using the CGI and the GD modules.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges was provided by NIH R01-LM007994-01A1.

Conflict of Interest: none declared.

REFERENCES


