SDPPRED: A METHOD FOR PREDICTION OF AMINO ACID RESIDUES THAT DETERMINE DIFFERENCES IN FUNCTIONAL SPECIFICITY OF HOMOLOGOUS PROTEINS AND ITS APPLICATION TO THE MIP FAMILY OF MEMBRANE TRANSPORTERS

Kalinina O.V.*1, Novichkov P.S.1, Mironov A.A.1,2, Gelfand M.S.2,3, Rakhmaninova A.B.1

1 Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia; 2 State Scientific Center GosNIIGenetika, Moscow, Russia; 3 Institute for Problems of Information Transmission RAS, Moscow, Russia

* Corresponding author: e-mail: ok81@yandex.ru

Keywords: protein specificity prediction, mutual information, Bernoulli cutoff

Summary

Motivation: The increasing volume of genomic data opens new possibilities for the analysis of protein function.

Results: SDPpred (Specificity Determining Position prediction) is a tool for prediction of residues in protein sequences that determine the proteins’ functional specificity. It is designed for the analysis of protein families, whose members have biochemically similar but not identical interaction partners (e.g., different substrates for a family of transporters). SDPpred predicts residues that could be responsible for the proteins’ choice of their correct interaction partners. SDPpred does not require information about secondary or 3D structure of proteins.

Availability: SDPpred is available at http://math.belozersky.msu.ru/~psn/.

Introduction

Many protein families contain homologous proteins that have a common biological function, but different specificity towards substrates, ligands, effectors, DNA, proteins and other interacting molecules, including other monomers of the same protein. All these interactions must be highly specific. The proteins can be assigned to specificity groups based on experimental data or comparative genomic analysis.

Identification of residues that account for the protein specificity might be useful in many biological studies. One obvious application of SDPpred is to minimize the number of point mutations required to switch the specificity of an enzyme, regulator or transporter. Analysis of the predicted residues can also provide a deeper insight into the nature of functional specificity. Prediction of SDPs is also reasonable for families containing specificity groups of any nature (e.g., proteins of different thermostability).

Amino acid residues that determine differences in the protein specificity and account for correct recognition of interaction partners, are usually thought to correspond to those positions of a protein multiple alignment, where the distribution of amino acids is closely associated with grouping of proteins by specificity. SDPpred searches for positions that are conserved within specificity groups but differ between them. These positions are called SDP's (Specificity-Determining Positions). Such positions, though obvious in alignments containing a small number of proteins and specificity groups, become a challenge to find in large protein families with a variety of specificities.

Recently, a number of algorithms addressing the described problem have been developed (for a review of some of these methods see Hannenhalli, Russell, 2000; Mirny, Gelfand, 2002; Kalinina
et al., 2004). SDPpred implements the algorithm described in (Kalinina et al., 2004). Compared with other methods, this algorithm has several advantages. Firstly, it does not use any information about the protein structure. The procedure is based solely on statistical analysis of an alignment. Secondly, it automatically calculates the number of SDPs and the probability of occurrence of these positions by chance. It does not incorporate any ad hoc cutoff setting. Thirdly, substitutions are weighted according to physical properties of amino acids, using a substitution matrix, so that substitutions to amino acids with similar properties are only weakly penalized. And finally, SDPpred incorporates information about evolutionary distance within and between groups by using different amino acid substitution matrices.

**Algorithm and Web interface description**

The only information needed for prediction of SDPs is a multiple alignment of protein sequences divided into specificity groups. We search for positions of a multiple protein alignment, for which the distribution of amino acid frequencies is closely associated with grouping by specificity. As a measure of such association we calculate the **mutual information** of each column $p$ of the alignment:

$$I_p = \sum_{i=1}^{20} \sum_{\alpha=1}^{20} f_p(\alpha,i) \log \frac{f_p(\alpha,i)}{f_p(\alpha)f(i)},$$

where $\alpha = 1,\ldots,20$ is a residue type, $f_p(\alpha,i)$ is the ratio of the number of occurrences of residue $\alpha$ in group $i$ at position $p$ to the length of the alignment column, $f_p(\alpha)$ is the frequency of residue $\alpha$ in the alignment column, $f(i)$ is the fraction of proteins belonging to group $i$. High values of the mutual information indicate columns with high correlation between amino acid frequencies and the grouping by specificity.

To address the fact that the frequencies are calculated based on a small sample, and that substitutions to amino acids with similar physical properties should be weakly penalized, the observed amino acid frequencies are modified. Instead of using $f(a,i) = n(a,i) / n(i)$, where $n(a,i)$ is the number of occurrences of residue $\alpha$ in group $i$, $n(i)$ is the size of group $i$ (here $i$ is a single group or the whole alignment), SDPpred uses **smoothed frequencies**

$$f^\beta(\alpha,i) = \left( \sum_{\beta=1}^{20} m(\beta \rightarrow \alpha) \right) \left( \frac{\sum_{\beta=1}^{20} m(\beta,i) m(\beta \rightarrow \alpha)}{n(i) + \kappa m(i)} \right)^{-1},$$

where $m(\beta \rightarrow \alpha)$ is the probability of amino acid substitution $\beta \rightarrow \alpha$ according to the matrix corresponding to the average identity in group $i$, $0 \leq \kappa \leq 1$ is a smoothing parameter. Additionally, the necessary pseudocounts are introduced in a natural way.

Then statistically significant values of the mutual information are selected using a novel procedure, called the **Bernoulli estimator** and is described in detail in (Kalinina et al., 2004). Briefly, we search for those positions, which are least probable to be obtained by chance. These positions we call SDPs (Specificity Determining Positions).

SDPpred outputs the set of SDPs, i.e. positions of the alignment, which are likely to determine differences in the functional specificity among the given groups. This set can be visualized as an alignment of the family with the SDPs highlighted, a detailed description of each SDP, or a plot of probabilities, from which the minimum is chosen to set the cutoff. SDPpred is publicly available at http://math.belozersky.msu.ru/~psn/.
Results and Discussion

The results of testing, which agree well with available structural and experimental data, are described in (Kalinina et al., 2004). In that study, we analyzed two protein families: the LacI family of bacterial transcription factors and the MIP family of membrane channels in bacteria. Both these families include proteins with the resolved 3D structure, which was used to evaluate predictions. In both cases the fraction of contacting residues among SDPs is much larger than in the whole alignment (Table).

In both cases the proteins function as oligomers and a substantial fraction of SDPs lies on the surface of contact between subunits. For example, in the case of the MIP family we predicted 21 SDPs, which were mapped onto the 3D structure of the tetramer of GlpF from *E. coli* (Fu et al., 2000). Sixteen of them (22Ile, 48Trp, 135Phe, 136Ser, 137Thr, 159Leu, 187Ile, 191Gly, 194Met, 195Gly, 199Gly, 200Phe, 201Ala, 207Asp, 211Lys, 236Pro) either contact the GlpF substrate glycerol or lie in the channel-forming helices on the side exposed into the channel, whereas the remaining five (20Leu, 24Ile, 43Glu, 108Tyr, 193Ser) lie on the outer surface of the monomer (Kalinina et al., 2004) (Fig. 1) and contact other subunits (Fig. 2). These five SDPs form two types of spatial clusters: one A-type cluster formed by 43Glu of all four monomers of the GlpF tetramer, and four B-type clusters formed by 20Leu, 24Ile and 108Tyr of one subunit and 193Ser of another subunit (for details see (Kalinina et al., in press).

This spatial arrangement of amino acid residues corresponding to SDPs suggests that evolutionary pressure on amino acids that establish intersubunit contacts is correlated with the evolutionary pressure on amino acids that account for the correct recognition of interaction partners. In the case of the MIP family, the residues lying on the surface of contact between subunits cluster together, possibly forming “structural clasps” that prevent formation of chimeric aquaporin-glyceroporin tetramers.

**Table.** Residues of different contact types among SDPs and in the whole alignment of the MIP and LacI protein families

<table>
<thead>
<tr>
<th>Contact (distance to an interaction partner &lt;5Å)</th>
<th>SDPs for the MIP family</th>
<th>Whole alignment of the MIP</th>
<th>SDPs for the LacI family</th>
<th>Whole alignment of the LacI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Possible contact (distance to an interaction partner 5-10Å)</td>
<td>8</td>
<td>73</td>
<td>19</td>
<td>89</td>
</tr>
<tr>
<td>Not contact (distance to an interaction partner &gt;10Å)</td>
<td>0</td>
<td>113</td>
<td>3</td>
<td>177</td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>281</td>
<td>44</td>
<td>348</td>
</tr>
</tbody>
</table>

SDPpred can be applied to any protein family that includes proteins of different specificity. It produces results that agree with available structural and experimental data. It proved to be useful not only for identification of candidate sites for protein functional redesign or prediction of specificity of family members, for which the latter is unknown. It also provided deeper insight into the nature of protein-protein interactions and the mechanism of molecular recognition.

**Acknowledgements**

This study was partially supported by grants from HHMI (55000309), LICR (CRDF RB0-1268), RFBR (04-04-49440), the Fund for Support of Russian Science, and the Program in Molecular and Cellular Biology of the Russian Academy of Sciences.
Fig. 1. Predicted SDPs mapped onto the structure of GlpF from *E. coli* (1fx8). Channel-forming SDPs are shown as gray spheres. SDPs located on the outer surface of the monomer are shown as white spheres.

Fig. 2. Residues making "structural clasps" in the structure of the tetramer of the GlpF of *E. coli* (1fx8, biological subunit). SDPs lying on the surface of contact between subunits are shown as white spheres.

References


