W424–W428 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh391

SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins

Olga V. Kalinina[1], Pavel S. Novichkov[1], Andrey A. Mironov[1,2], Mikhail S. Gelfand[2,3] and Aleksandra B. Rakhmaninova[1,*]

[1]Department of Bioengineering and Bioinformatics, Moscow State University, Vorob'evy gory, 1-73, Moscow, 119992, Russia, [2]State Scintific Center GosNIIGenetika, 1st Dorozhny pr., 1, Moscow, 113545, Russia and [3]Institute for Problems of Information Transmission RAS, Bolshoi Karetny per., 19, Moscow, 127994, Russia

Received February 12, 2004; Revised and Accepted March 24, 2004

ABSTRACT

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 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. The input of SDPpred is a multiple alignment of a protein family divided into a number of specificity groups, within which the interaction partner is believed to be the same. SDPpred does not require information about the secondary or three-dimensional structure of proteins. It produces a set of the alignment positions (specificity determining positions) that determine differences in functional specificity. SDPpred is available at http://math.genebee.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 protein specificity might be useful in many biological studies. For instance, these residues can be used for planning experiments on functional analysis or protein redesign. One obvious application of SDPpred (Specificity Determining Position Prediction) 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. Our experience is that specificity determining positions (SDPs) include not only residues located in active sites of proteins, but also residues involved in establishing contact between subunits.

Construction of phylogenetic trees does not always allow one to assign specificity to all members of a family. An algorithm that extends the idea of SDP analysis and addresses this problem will be put on the web in the near future. It will predict specificity of the unclassified family members. This will provide a possibility to use SDPpred as a tool for detailed protein annotation.

Amino acid residues that determine differences in 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 well conserved within specificity groups but differ between them. These positions are called SDPs. Such positions, though obvious in alignments containing a small number of proteins and specificity groups, become challenging to find in large protein families with a variety of specificities. Prediction of SDPs is reasonable not only for protein families whose members have different interaction partners, but also for any family containing specificity groups of any nature (e.g. proteins of different thermostability).

Recently, a number of algorithms addressing the problem described above have been developed. Several approaches exploit information about protein structure or functional sites (1,2). Some methods use only protein sequences (3–8).

*To whom correspondence should be addressed. Tel: +7 095 9394331; Fax: +7 095 2090579; Email: abr@belozersky.msu.ru

SDPpred implements the algorithm described in (8). Compared with other methods, the algorithm implemented in SDPpred and described in detail in (8) has several advantages. First, it does not use any information about protein structure. The procedure is based solely on statistical analysis of an alignment, and thus it can be applied to protein families that do not contain any members with resolved three-dimensional (3D) structure. Second, 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 and thus does not require any prior knowledge about special properties of the analyzed family. Third, substitutions within specificity groups 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. Finally, SDPpred incorporates information about evolutionary distance within and between groups by using different amino acid substitution matrices. To the best of our knowledge, currently there is no publicly available server, which addresses the problem of identification of SDPs.

ALGORITHM DESCRIPTION

The algorithm implemented in SDPpred is described in detail in (8). Some of its features are inherited from the method of (7).

Briefly, consider a multiple protein sequence alignment. The proteins are divided into N specificity groups, numbered by $i = 1, \ldots, N$. The goal is to identify columns (positions) in the alignment in which the amino acid distribution is closely associated with the grouping by specificity. This association in column p of the alignment is measured by the *mutual information*

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

where $\alpha = 1, \ldots, 20$ is a residue type, $f_p(\alpha, i)$ is the ratio of the number of occurrences of residue α in group i at position p to the length of the whole alignment column, $f_p(\alpha)$ is the frequency of residue α in the whole alignment column, $f(i)$ is the fraction of proteins belonging to group i. The mutual information reflects the statistical association between two discrete random variables α and i.

To address the facts 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 $f(\alpha, i) = n(\alpha, i)/n(i)$, where $n(\alpha, i)$ is the number of occurrences of residue α 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*

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

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, and $0 \leqslant \kappa \leqslant 1$ is a smoothing parameter. SDPpred uses matrices of the BLOSUM series (9) for groups with average identity $\leqslant 60\%$ and their analogs calculated as described in (10) for groups with larger average identity. Additionally, zero frequencies are avoided automatically, and thus the necessary pseudocounts are introduced in a natural way.

To calculate the statistical significance of the obtained values of I_p, each column is shuffled, yielding the distribution $F(I^{\mathrm{sh}})$. To offset the background similarity of proteins, which is higher within groups than between groups, SDPpred calculates I^{exp}, the expected mutual information for column p, as a linear transform of I^{sh}, as described in (7).

Then, Z-scores are calculated:

$$Z_i^I = \frac{I_i - \langle I_i^{\mathrm{exp}}\rangle}{\sigma\left(I_i^{\mathrm{exp}}\right)}.$$

A high Z-score value indicates a position where the amino acid distribution is much more closely associated with grouping by specificity than for an average position of the alignment, and which is thus likely to be an SDP.

Given a series of Z-scores corresponding to every position of the multiple alignment, one needs to evaluate the significance of the Z-scores in order to tell whether the observed Z-score is sufficiently high to indicate an SDP. SDPpred uses an automated procedure for setting the thresholds based on the computation of the Bernoulli estimator. The observed Z-scores are arranging in decreasing order: $Z_1, Z_2, \ldots$. The threshold is defined as

$$k^* = \arg\min_k P(\text{there are at least } k \text{ observed } \mathrm{Z} - \text{scores } \mathrm{Z} \geqslant \mathrm{Z}_k)$$
$$= \arg\min_k \left(1 - \sum_{i=n-k+1}^{n} C_n^i q^i p^{n-i}\right),$$

where n is the total number of considered positions,

$$p = P(\mathrm{Z} \geqslant \mathrm{Z}_k) = \int_{Z_k}^{\infty} \frac{1}{\sqrt{2\pi}}\exp(-\mathrm{Z}^2)d\mathrm{Z}, \quad q = 1 - p.$$

k^* positions having highest Z-scores are designated SDPs, as they are the least probable to constitute a tail of the Gaussian distribution, and thus are non-randomly generated positions.

The described procedure depends on the distribution of Z-scores. It can be proved that the distribution of the mutual information lies asymptotically between the Gaussian and exponential distributions. On real data the procedure is robust relative to the distribution, and the set of SDPs is almost the same assuming Gaussian and exponentially distributed Z-scores.

The results of testing, which agree well with available structural and experimental data, are described in (8). 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 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 1). Interestingly, in the case of the MIP family we not only described the channel very well (all residues known to interact with the substrate are either conserved or belong to the predicted set of SDPs), but also identified some residues that lie on the surface

Table 1. Residues of different contact type among SDPs and in the whole alignment of the MIP and LacI protein families

	SDPs for the MIP family	Whole alignment of the MIP family	SDPs for the LacI family	Whole alignment of the LacI family
Contact (distance to an interaction partner <5 Å)	13	95	22	82
Possible contact (distance to an interaction partner 5–10 Å)	8	73	19	89
Not contact (distance to an interaction partner >10 Å)	0	113	3	177
Total	21	281	44	348

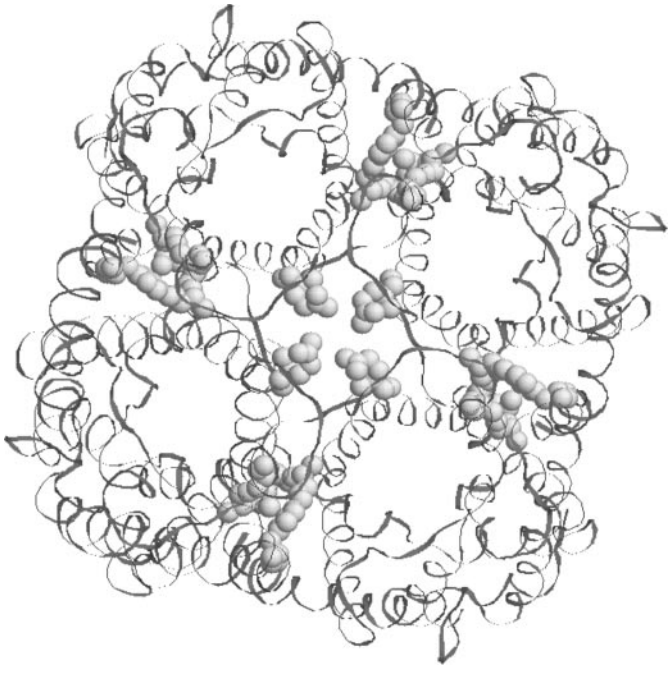

Figure 1. Residues making 'structural clasps' in the structure of the tetramer of the GlpF of *Escherichia coli* (1fx8, biological subunit). SDPs lying on the surface of contact between subunits are shown by white spheres.

of contact between the subunits and cluster together, possibly forming 'structural clasps' (Figure 1) (11).

DESCRIPTION OF THE WEB INTERFACE

The only information needed for prediction of SDPs is a multiple alignment of protein sequences divided into specificity groups. SDPpred does not require any information about protein 3D structure.

SDPpred can analyze alignments of length up to 2000 positions, containing at most 1000 proteins. There can be up to 1000 specificity groups; however, it is recommended that each group contain at least three sufficiently divergent sequences. On the other hand, the average identity in each group should not be <25%. Having more than two groups also strongly improves the quality of prediction due to more efficient elimination of the background evolutionary similarity.

A typical SDPpred query of is shown in Figure 2. The aligned sequences should be in FASTA, GDE or Pfam plain text (with gaps as dashes and all characters in upper case) alignment format. Alignments in one of these formats

```
=== AQP ===
%sp|Q9L772|AQPZ_BRUME
-------------------------------------mlnklsaeffgtfwlvfggcgsa
ilaa--afp-------elgigflgvalafgltvltmayavggisg--ghfnpavslgltv
ag---------------rlpakdlipywvaqvlgaiaaaailyviasgkd----------
-----g------fsagg---lasngygelsp---------ggysmmagllieiiltaffi
iiilgsts--------------------------------slap------------------
agfapiaigfgltlihlvsipvtntwvnparstgvalfadtaals----------------
qlwlfwvaplvgavigaiiwkgllgrd-----------------------------------
------
%sp|P48838|AQPZ_ECOLI
-------------------------------------mfrklaaecfgtfwlvfggcgsa
vlaa--gfp-------elgigfagvalafgltvltmafavghisg--ghfnpavtiglwa
gg---------------rfpakevvgyviaqvvggivaaallyliasgkt----------
-----g------fdaaasg-fasngygehsp---------ggysmlsalvvelvlsagfl
lvihgatd---------------------------------kfap------------------
agfapiaiglaltlihlisipvtntsvnparstavaifqggwale----------------
qlwffwvvpivggiiggliyrtllekrd-----------------------------------
------
%tr|Q8UJW4
-------------------------------------mgrkllaeffgtfwlvfggcgsa
vfaa--afp-------elgigftgvalafgltvltmayavggisg--ghfnpavsvgltv
ag---------------rfpasslvpyviaqvagaivaaaalyviatgka----------
-----g------idlgg---fasngygehsp---------ggyslvsallieiiltaffl
ivilgsth---------------------------------grvp------------------
agfapiaiglaltlihlisipvtntsvnparstgqalfvggwalq----------------
qlwlfwlapivggaagaviwklfgekd-------------------------------------
------
%tr|Q92ZW9
-------------------------------------mfkklcaeflgtcwlvlggcgsa
vlas--afp-------qvgigllgvsfafgltvltmaytvggisg--ghfnpavslglav
ag---------------rvpaaslvsyviaqvagaiiaaavlyviatgka----------
-----d------fqlgs---faangygehsp---------ggysltaalvtevvmtfffl
iiilgsth---------------------------------rrvp------------------
agfapiaiglaltlihlvsipvtntsvnparstgqalfvggwals----------------
qlwlfwiaplfgaaiagivwksvgeefrpvd---------------------------------
------
=== GLP ===
%sp|P11244|GLPF_ECOLI
-----------------------------msqt---stlkgqciaeflgtglliffgvgcv
aalkvag---------a-sfgqweisviwglgvamaiyltagvsg--ahlnpavtialwl
fa---------------cfdkrkvipfivsqvagafcaaalvyglyynlfffdfe------
-----q--t-hhivrgsvesvdlagtfstypn--------phinfvqafavemvitailm
glilaltd----------------------------dgn--------------g-vpr
gplaplligliavigasmgpltgfamnpardfgpkvfawlagw--gnvaftggrd-ipy
-flvplfgpivgaivgafayrkligrhlpcdicvveek--etttpseqkasl--------
------
%sp|P44826|GLPF_HAEIN
---------------------------mdks-----lkancigeflgtalliffgvgcv
aalkvag---------a-sfglweisimwgmgvalavyataglsg--ahlnpavtialwk
fa---------------cfdgkkvipyiisqmlgaffaaalvyalyrnvfidye------
-----t--v-hnivrgtqeslslagtfstyph--------pslsiggafavefvitailm
alimaltd----------------------------dgn--------------g-vpr
gplaplligiliaviggamgpltgfamnpardfgpkffaylagw--gelaltggre-ipy
-fivpmvapvlgalagawlykkaiggnlpcncgce-------------------------
------
```

Figure 2. A typical SDPpred query.

can be easily obtained from databases (e.g. Pfam) or alignment programs (e.g. ClustalW). The alignment should be manually edited according to the specificity group assignments. The specificity groups should be separated by lines beginning with the equals sign and containing the name of the following group (e.g. '=Group1').

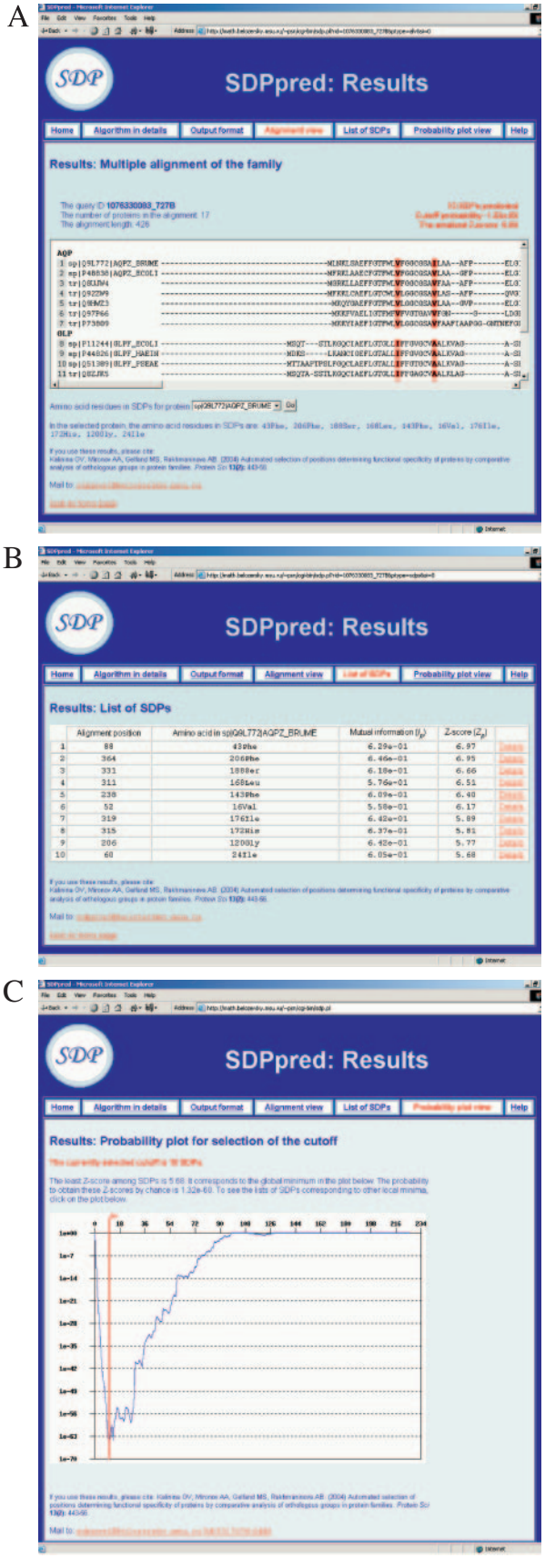

Figure 3. SDPpred output: (**A**) alignment of the query protein family with SDPs highlighted; (**B**) detailed description of the SDPs; (**C**) probability plot.

The user has to select the number of shuffles for computation of the statistical significance (between 1000 and 10 000). An alignment of a thousand sequences divided into several hundreds of specificity groups is analyzed in a couple of hours if each column is shuffled 10 000 times. Using fewer shuffles reduces the required time proportionally, but makes the results less reliable. Typically, the top of the SDP list remains the same, but minor variations may appear near the cutoff. An average query of 72 sequences divided into 12 specificity groups, where the average protein length is 400 amino acids and the alignment length is 587 positions, is analyzed in 4 min.

The last parameter is the maximum percentage of gaps allowed in a column to be analyzed. Columns with a greater fraction of gaps are excluded from the analysis. Typically, this number should not exceed 30%, but in some cases (e.g. when the user is analyzing group-specific loops) it may be reasonable to set this parameter to a higher value. However, a large percentage of allowed gaps produces many SDPs at the termini of the alignment, which are likely to be incorrect.

OUTPUT

SDPpred outputs the set of SDPs, i.e. positions of the alignment, which are likely to determine differences in functional specificity among the given groups. This set can be visualized in several ways. The user can switch between the alignment of the family with the SDPs highlighted, the detailed description of each SDP, and the plot of probabilities, from which the minimum is chosen to set the cutoff (Figure 3). The latter is particularly useful in the case when there are several local minima of close significance. Then, it might be useful to consider them all.

The predicted SDPs can be mapped on to any protein of the alignment. Amino acid residues corresponding to the SDPs in the selected protein are listed below the alignment on the alignment page and in the tables describing SDPs in detail.

PROSPECTS

We plan to implement the algorithm that predicts specificity for those members of the family whose specificity is unknown. The algorithm is described in detail in (8).

ACKNOWLEDGEMENTS

This study was partially supported by grants from the Howard Hughes Medical Institute (55000309) and the Ludwig Institute for Cancer Research (CRDF RB0-1268).

REFERENCES

1. Lichtarge,O., Bourne,H,R. and Cohen,F,E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.
2. Johnson,J.M. and Church,G.M. (2000) Predicting ligand-binding function in families of bacterial receptors. *Proc. Natl Acad. Sci. USA*, **97**, 3965–3970.

3. Livingstone,C. and Barton,G. (1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.*, **9**, 745–756.
4. Casari,G., Sander,C. and Valencia,A. (1995). A method to predict functional residues in proteins. *Nat. Struct. Biol.*, **2**, 171–178.
5. Gaucher,E.A., Gu,X., Miyamoto,M.M. and Benner,S.A. (2002). Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem. Sci.*, **27**, 315–321.
6. Hannenhalli,S.S. and Russell,R.B. ((2000). Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.*, **303**, 61–76.
7. Mirny,L.A. and Gelfand,M.S. (2002). Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J. Mol. Biol.*, **321**, 7–20.
8. Kalinina,O.V., Mironov,A.A., Gelfand,M.S. and Rakhmaninova,A.B. (2004) Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci.*, **13**, 443–456.
9. Henikoff,S. and Hennikoff,J. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
10. Sutormin,R.A., Rakhmaninova,A.B. and Gelfand,M.S. (2003). BATMAS30—the amino acid substitution matrix for alignment of bacterial transporters. *Proteins*, **51**, 85–95.
11. Kalinina,O.V., Gelfand,M.S., Mironov,A.A. and Rakhmaninova,A.B. Amino acid residues forming specific contacts between subunits in tetramers of the membrane channel GlpF. *Biofizika*, in press.