The Method SPrOS

The method SPrOS (Specificity Projection On Sequence) is developed to analyze the amino acid sequences related to the same protein family in order to recognize the sites responsible for the specificity of separated subclasses within this family.

Comparing the sequences within a protein family, one can observe positions conserved across the all studied proteins while other positions display different levels of variability. Given the reasonable partition of the family sequences into the several groups, the sites can be found, where the amino acid composition is noticeably restricted in the separate group relative to the entire family. These group discriminating sequence markers are considered as the patterns could be associated with the functional divergence of the established groups.

The algorithm SPrOS requires the training set of preliminary classified amino acid sequences. A user should choose the sequence(s) to be mapped with specificity sites. The test sequence is excluded from the training set and compared with all the rest sequences. The obtained similarity scores are used as input to the procedure, which estimates the specificity to each given class for each position of a test sequence.

Positional similarity scores

The smooth positional scores are calculated as follows. For each fragment of the sequence *Q*, the most similar fragment of sequence *D* is detected and the similarity score of these two fragments is assigned to the fragment from *Q*. Each position *i* is included into the *F* fragments, and the positional S_i score is equal to the maximal value of similarity scores obtained for these fragments:

$$A_{ih} = \sum_{k=1}^{i} sim(q_k, d_{k+h})$$

$$R_i = \max_h \left(A_{ih} - A_{i-F,\,h} \right)$$

$$S_i = \max_j R_{i+j}, 0 \le j < F$$

where

sim(*q*, *d*) is the similarity estimation of the superposed amino acid residues according to the given measure (e.g., the residue identity or a certain substitution matrix);

q_x and d_y are residues in the indexed positions of *Q* and *D*, respectively;

h is the shift value in an acceptable range;

F is the parameter defining the length (odd value) of the comparable segment;

R_i is a maximal similarity score of the segment $[q_i\text{-}F+1;\ q_i]$ with the sequence *D*.

The focused positional scores are calculated as follows. When the fragments of length *F* from the test sequence (*Q)* and the training sequence (*D)* are superposed, the similarity score of these two fragments are assigned to all positions of the fragment from *Q*. If the matched residues q_k and d_{k+h} are identical the positional score is multiplied by the correction factor *w*. Finally, the maximal value of scores obtained for the position *i* in all fragmental comparisons is assigned to the score S_i.

The smooth scores allow accounting for the similarity of the position-surrounding regions within a separate group, as well as their intergroup differences. The focused scores better reveal the positions differing in separated groups, even if the surrounding regions are conserved in the whole family.

In this program version:

the residue identity is used as a measure for similarity scoring;

the focused scores are calculated with $w = 2.0$.

Estimation of the position specificity for classes

The test sequence *Q* is compared with each sequence of the training set. The obtained positional scores (S_i) represent the similarity of sequence *Q* with training sequences assigned to predefined classes. The estimate E_{ia}, is introduced, which evaluates the relation of the position *i* to the class *A*.

Let S_{ik} be the score (smooth or focused) of position *i* obtaining by projection of the *k-th* training sequence onto the test sequence. Let a_k and b_k be the coefficients of the *k-th* sequence belonging to non-intersected classes *A* and *B*, respectively. In this program version, class *B* is treated as a complement of class *A*. By projecting all training sequences on the test sequence, E_{ia} is calculated:

$$E_{ia} = \frac{u_i - v_i}{u_i + v_i},$$

$$u_i = \frac{\sum_k S_{ik} a_k}{\sum_k a_k}, \qquad v_i = \frac{\sum_k S_{ik} b_k}{\sum_k b_k}$$

E_{ia} adopts the value in the range of [-1; +1] i.e. from the maximal specificity for class *B* to the maximal specificity for class *A*. If $E_{ia} \approx 0$ then the position *i* is not significant for distinguishing the classes *A* and *B*.

Unlike other tools developed for the prediction of specificity determining positions, the suggested method estimates positions of a single sequence rather than positions in a group of aligned sequences. Note, that coefficients of belonging can define the fuzzy classification of the training sequences. It is more suitable for the enzyme subgroups, which can overlap in substrate specificity. This program version provides the binary coefficients adopting the values 0 or 1; so that $a_k + b_k = 1$ and $a_k \times b_k = 0$.

Probabilistic specificity estimates

The p-values are applied to evaluate the probability of obtaining a specificity estimate *X* not lower than the given E_{ia}, provided that the studied proteins were randomly distributed in classes. The lower the *p-value*, the more specific position *i* for the class *A*. Thus, the *p-values* are used as probabilistic estimates of the intergroup differences in the studied sequences.

The p-values are obtained by "shuffling" the classes of training sets. The belonging coefficients are randomly assigned to sequences on assumption that class sizes remained unchanged. Generating a given number of randomized classifications, the E_{ia} distribution is calculated to obtain the *p-values* of E_{ia} for real classes.

In this program version, 5000 randomized classifications are provided for calculation of *p-values*.

Running the calculation on the web-server

Use ***Sequence*** to upload a file containing the set of protein sequences in a *Fasta* format. The sequences should not contain the gap ('-') or any other letters besides of twenty canonical amino acid symbols.

```
>S1m2
MEALALLALMASCLILISVWRNSSGRGKVPPGPAAAPVLGQLFNISIRDTSWSLTQVGRI
FGPVFTLYWPVAGVIILHGYEAVRDAIVEVGQEFSGRRIFPWGDKGQRAYPVLFSQGKRW
KDVRRFNVITIRNFRMKKQSLEDRMDEEARCIIEEVRKTKAAPSDGTFAIPSAPTNILCS
IIFHKRFEYRDNQFIQAPSDGMFLFYSAPTNILDNQFILISVWRT
. . . . . .
>S15m2
MDAVIVILLMVTCLLLISLWRQSSGRGRLPPGPTGLPVVGTLFELGMHDISLSVTNIGEL
YGPVFTLYFPLRPLVVLHGYEAVRDGIVDVAQDFGGRLITPLADRAQKGYGVLYTNGKRW
KEIKRFTLLSVKQWPMGKCSAEDRMNEDSEVIADDLRKTKATPCEGSFAIGSAPSNIVCS
IIFHKRFEYRENQYFNTPCEGSFYIRSAPSNIVENQYLLISLWRT
```

Use ***Sequence Classification*** to upload a text file containing tab-delimited pairs of protein IDs and identifiers of classes, to which the proteins are assigned. One and the same protein can be related to more than one class, allowing the intersected classes. The first row should contain headers. The last raw should contain the single symbol '#'. At least two classes must be provided. The specificity classes are automatically selected by size. The class accepted for further treatment will contain at least five proteins and no more than two-thirds of the total number of proteins included into the training set. The sequences not included into the given class are processed as the class complement, which also must contain at list five proteins.

PID	**Group**
S1m2	m2g1
S2m2	m2g1
. . . . .	
S14m2	m2g2
S15m2	m2g2
. . . . .	
#	

Press ***Test Sequences*** to select test sequences form the uploaded sequence set

Pressing ***Mode***, you can define the *Smooth* or *Focused* type of positional scores.

Pressing ***Frame***, you can define the length of compared sequence segments.

Pressing ***Cutoff p-value,*** you can define the upper limit of *p-value* for output results.

Press ***Get Results*** to run calculation and obtain results

Result output

The tab-delimited output file contains the rows, presenting the following items:

Prot. ID	**Position**	**AA**	**Group**	**Sp. Estim.**	**p-value**	**Belong**
Protein Identifier	*Amino acid position number*	*Amino acid type*	*Class Identifier*	*Estimates of Specificity of a position to a given class (E_{ia})*	*p-value calculated for the obtained E_{ia}*	*Coefficient of belonging of the protein to the given class (predefined in training data)*
S1m2	48	R	m2g1	0.300	0.0005	1
S1m2	87	I	m2g3	0.100	0.0880	0
S15m2	95	G	m2g3	0.452	<8.89E-07	1

If the calculated *p-value* is equal to 0, then the minimal non-zero value, which can be obtained using randomized classifications, is output with the prefix symbol **'<'**.

For publication of results please cite the following article:

Karasev D.A., Veselovsky A.V., Oparina N.Y., Filimonov D.A., Sobolev B.N. (2016) Prediction of amino acid positions specific for functional groups in a protein family based on local sequence similarity. *J. Mol. Recognit.*, **29(4)**, 159-169.