Detecting the features of functional specificity in protein families based on the local sequence similarity

Boris Sobolev, D.A. Karasev, A.V. Veselovsky, D.A. Filimonov, V.V. Poroikov

Institute of Biomedical Chemistry (IBMC), Pogodinskaya street 10/8, 119121, Moscow, Russia, boris.sobolev@ibmc.msk.ru

Oparina N.Yu.

Engelhardt Institute of Molecular Biology (EIMB), 119991, Vavilova, 32, Moscow, Russia

Department of Medical Biochemistry and Microbiology, Uppsala University, Husargatan 3, PO Box 582, SE-75123 Uppsala, Sweden

One of the main tasks of protein sequence analysis is the recognition of sites related to the certain functional features. Detection of residues responsible for the certain functions is applied in protein engineering manipulations, discovery of new drug targets, study of disease-associated mutations etc. In studying the functionally diverged protein family, the sites conserved within the separate groups are found. These group-specific positions can be important for substrate or inhibitor binding, protein-protein interactions and other diverged functions.

Commonly, tools applied for recognition of the group-specific positions in homologous proteins use the MSA as software input; the output data usually contain statistical scoring for estimation of residue positions. These methods allow obtaining the acceptable result if the MSA provides precise superposition of residues significant for the studied function. Unfortunately, the required representative set of sequences providing the "smooth" transition from the homologue to homologue is frequently not accessible, casing the ambiguity in matching the significant residues. It probably explains the fact, that aligned-base methods recognize the group-specific residues with lower efficiency comparing with the positions related to the properties common for all proteins of the family [1].

We modified earlier developed original method of sequence comparison [2] to estimate the specificity of sequence position by the method differed from the alignment-based procedures but based on independent comparisons of segment pairs from all studied sequences. We suppose that our method SPrOS (Specificity Projection On Sequence) allows matching the

significant positions, which are shifted in MSA and can be related with alignment errors or mutations compensating the exchanges in neighboring positions. The developed software allows handling the intersected class, every time dividing the training set into the target and complement classes. It is better suited to screen the large sets representing the proteins and their ligands. When operating a program SPrOS, the tested sequence is compared with training sequences. The similarity scores for tested sequence positions are calculated based on the close surroundings of matched residues and then used to statistics B_{ic}, which allow estimating the specificity of position i for class C. The *p-values* of obtained B_{ic} are used as probabilistic measures of class belonging.

At first, our method was tested on the artificial sequences, which were obtained by simulating the evolutionary processes. Two modeled "families" were generated and each family was divided into the three classes by introducing the group-specific amino acid exchanges into sequences. Various exchanges simulated different types of class-specific markers: from "mutations" occupying the same matched positions in "positionally-sliding" ones. The method SPrOS displayed the high prediction efficiency for both artificial families. The slight deviation out of maximum possible accuracy is explained by the more complicated configuration of certain exchanges. Comparative trial with popular method SDPfox using default parameters [3] showed that our method is the more successful in recognition in positionally-sliding "mutations".

For testing our methods with natural proteins we used three well-studied families. Two of them were bacterial releasing factors (RF) and LacI/GalR, which divided into clearly distinguishing classes, related to different ligand types. The third family was presented by protein kinases partitioned into the classes of inhibitor specificity.

The testing with two first families showed the recognition of class specific positions with highly significant statistical estimates. By mapping the results on the 3D structures, the certain ligand contact residues were matched with predicted positions. Other revealed positions were coincided with mutations altering the protein function [4] or sites responsible for allosteric influence on ligand-binding [5]. In addition, we do not exclude the presence evolutionary coupled mutations, which may not be related with the studied functional specificity.

The more difficult task arises when the proteins of the same large family have the wide spectra of activities (e.g., bound ligands) so that each protein can belong to several specificity classes and these classes are significantly intersected. The protein kinases characterized by their interaction with inhibitors are the typical example of such intricate classification. Most of them compete with ATP for specific binding sites. However, many inhibitors show the pronounced selectivity to the separate kinases or their subfamilies.

The classification of kinases was based on their experimentally detected interaction with inhibitors [6]. Belonging to class of inhibitor specificity was established in accordance with the given thresholds of interaction index.

We predicted positions related to spectra of inhibitor specificity of protein kinases, using the training set classified by inhibition with different ligands. By verifying the prediction by mapping results in the 3D structures, the predicted positions specific for certain inhibitory classes were in the ATP-binding cleft, which is the target for most known kinase inhibitors. Somewhat difficulties in interpretation of results seem to be related with structural similarity of different inhibitors, which may not be always coincided to ligand specificity spectra.

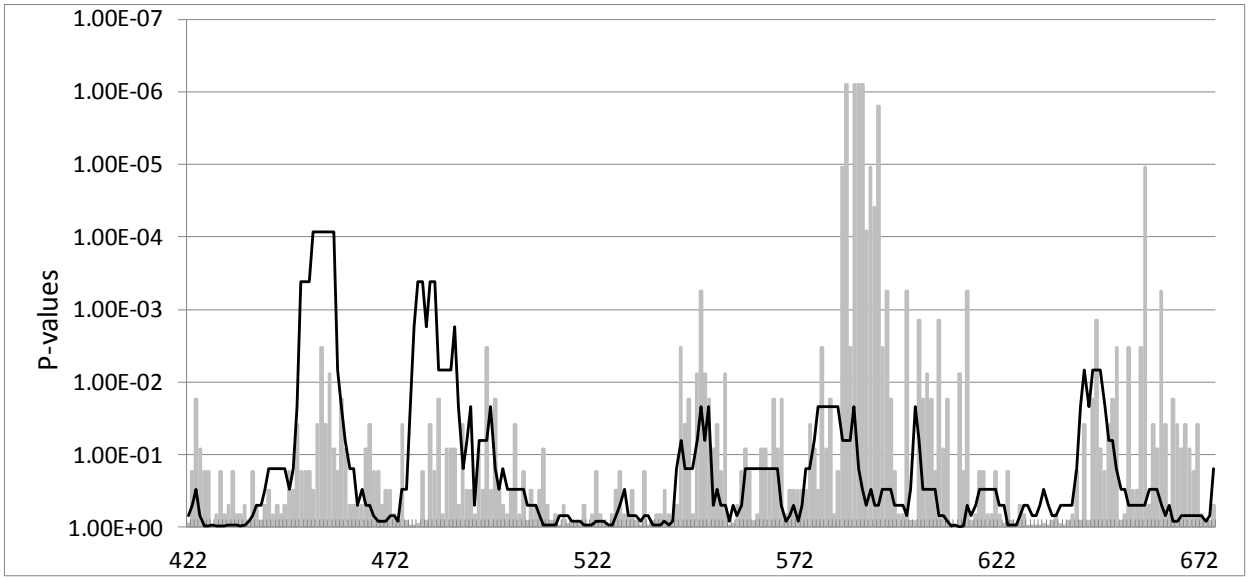

Figure 1. Prediction of ligand-position specificity for FAK kinase before (gray columns) and after (black line) excluding the most close homologues from training set.

The serious problem arises from the evolutionary coupled mutations. To overcome the problem

of evolutionary trace we suggest the approach illustrated in fig. 1. When the training set included very close homologues of the tested FAK kinase, the “additional” positions can be recognized, displaying rather evolutionary trace than functional specificity. After excluding these proteins, the most significant positions were located in the area of ATP-binding pocket.

Tested on the simulated sequences, the SPrOS method recognized the various types of position exchanges including the more complicated variants, which were not found by the method based on MSA. We successfully tested our programs on natural protein sets, clear divided into functional groups. It is particular interesting to apply our approach to detect functionally significant residues in protein families displaying wide and intersected specificities. We demonstrated the applicability of the SPrOS methods for solving the more complicated task such as recognizing the positions of protein kinases associated with inhibitor binding.

The work is done in the framework of the Russian State Academies of Sciences fundamental research program for 2013-2020.

1. E.Teppa et al. (2012) *BMC Bioinformatics,* **13**:235.

2. B.Sobolev et al. (2010) *BMC Bioinformatics*, **11**:313.

3. P.V.Mazin et al. (2010) *Algorithms Mol. Biol.*, **5**:29.

4. J.Suckow et al. (1996) *J. Mol. Biol.*, **261**:509–523.

5. J.L.Huffman et al. (2002) *Biochemistry*, **41**:511-20.

6. Y.Gao et al. (2013) *Biochem J.*, **451**:313-28.