ISSN 0026-8933, Molecular Biology, 2007, Vol. 41, No. 1, pp. 137–147. © Pleiades Publishing, Inc., 2007.
Original Russian Text © O.V. Kalinina, R.B. Russell, A.B. Rakhmaninova, M.S. Gelfand, 2007, published in Molekulyarnaya Biologiya, 2007, Vol. 41, No. 1, pp. 151–162.

MATHEMATICAL AND SYSTEM BIOLOGY

UDC 577.112:004.02

Priority Publication

Computational Method for Predicting Protein Functional Sites with the Use of Specificity Determinants

O. V. Kalinina[a, b], R. B. Russell[b], A. B. Rakhmaninova[a, c], and M. S. Gelfand[a, c]

[a] *Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, 119992 Russia*

[b] *European Molecular Biology Laboratory, 69117 Heidelberg, Germany*

[c] *Institute of Information Transmission Problems, Russian Academy of Sciences, Moscow, 127994 Russia; e-mail: gelfand@iitp.ru*

Received and accepted for publication September 5, 2006

Presented by A.V. Finkel'shtein

Abstract—The currently available body of decoded amino acid sequences of various proteins exceeds manifold the experimental capabilities of their functional annotation. Therefore, in silico annotation using bioinformatics methods becomes increasingly important. Such annotation is actually a prediction; however, this can be an important starting point for further laboratory research. This work describes a new method for predicting functionally important protein sites, SDPsite, on the basis of identification of specificity determinants. The algorithm proposed utilizes a protein family alignment and a phylogenetic tree to predict the conserved positions and specificity determinants, map them onto the protein structure, and search for clusters of the predicted positions. Comparison of the resulting predictions with experimental data and published predictions of functional sites by other methods demonstrates that the results of SDPsite agree well with experimental data and exceed the results obtained with the majority of previous methods. SDPsite is publicly available at http://bioinf.fbb.msu.ru/SDPsite.

DOI: 10.1134/S0026893307010189

Key words: structural genomics, functional site, prediction, comparative sequence analysis, specificity determinants, specificity-determining positions

INTRODUCTION

The exponential growth in the volumes of databases compiling information about DNA sequences exceeds considerably the capabilities of their experimental functional annotation (description of the functional characteristics). So far, 335 bacterial, 41 eukaryotic, and 27 archaebacterial genomes have been sequenced completely and 1596 similar projects are in progress (according to the Genome OnLine Database, GOLD [1]). Preliminary annotation of the sequences by computational methods is part of the routine procedures in such projects. In addition to the primary sequences, knowledge about protein spatial structures is most important for understanding their functions. In 2000, the international project on structural genomics was launched [2], whose goal is to resolve a representative set of spatial structures for proteins of various organisms. The main stages of this project are (1) grouping all known protein sequences into families, (2) choosing one or several representatives from each family as a target, (3) resolving the spatial structure of the target by X-ray analysis or NMR, and (4) constructing the spatial structure models for other representatives of each family. Implementation of this project will give the structures of many proteins with not only unknown localization of their active centers and/or other functional sites, but frequently with unknown overall function itself; moreover, these proteins have no well-studied homologs. Various computational methods are used in such cases for searching for the functional sites.

Russell et al. [3] have reviewed a number of methods combining the information about a sequence alignment with the data on spatial structures. These methods search for the regions important for the protein function or specificity on the protein surface. For example, analysis of the Mj0577 structure, resolved within the framework of the Structural Genomics Project by the ConSurf method [4], assisted in detecting the ATP-binding site and demonstrated a functional importance of the contact interface between homodimer subunits [5].

Several methods used in searching for functional sites, such as ConSurf [4] and the method developed by Aloy et al. [6], postulate that a position is functionally important when it is conserved in alignments of related sequences. The methods by Ma et al. [7] and

Landgraf et al. [8] use a milder structural conservation instead of sequence conservation; however, the general idea is the same. Del Sol Mesa et al. [9] have introduced the conception of correlated mutations in a sequence (simultaneous mutations at remote positions in one alignment or different alignments) and considered the positions affected by such mutations during evolution to be functionally important. Lichtarge et al. [10, 11] have developed an evolutionary trace method, whose essence is grouping proteins at various similarity levels to detect the conserved positions (CPs) for each group. The sequence of the positions conserved in a group is called its evolutionary trace. The evolutionary traces are compared for different groups, and the positions contained in the evolutionary traces of a large number of groups are considered significant.

Hannenhalli and Russell [12] and Mirny and Gelfand [13] have developed methods searching for specificity determinants (to ligand, DNA, other protein, etc.) in protein sequence alignments: the alignment is considered prepartitioned into groups of proteins with the same specificity, and the positions conserved within the specificity groups but differing between groups are considered the specificity determinants. Based on these methods, we have proposed the algorithm SDPpred [14], which follows the main features of the algorithm described in [13] but is more appropriate technically for analyzing large data massifs, as it has an automated procedure for threshold selection and better takes into account the evolutionary distance between proteins and amino acid similarity.

In this work, we present a new algorithm, SDPsite, for predicting the functional sites in proteins. This algorithm combines the features of many methods mentioned above, namely, the CPs are detected in sequence alignment; specificity determinants are predicted based on the alignment and phylogenetic tree (for this purpose, a specialized procedure of automated search for specificity groups has been developed); and the best cluster of specificity determinants and CPs is found using the spatial structure of one of the proteins. SDPsite is publicly available at http://bioinf.fbb.msu.ru/SDPsite.

SDPsite was tested using the family of bacterial transcription factors LacI, subtilisin-like proteases, and 68 domains from the Conserved Domain Database (CDD).

ALGORITHM

The algorithm for prediction of a functional site consists of three parts: (1) prediction of specificity determinants (using an automated partitioning of an alignment into specificity groups), (2) prediction of CPs, and (3) selection of the best cluster.

Prediction of Specificity Determinants

The algorithm predicting specificity-determining positions (SDPs) is described in [14]. The input data for the prediction is an alignment of amino acid sequences where the proteins are divided into specificity groups. It is assumed that a specificity group contains proteins with the same substrate specificity and the specificities of proteins belonging to different groups are distinct. Briefly, the algorithm is as follows. Each position in the alignment is considered separately. To assess whether a particular position is SDP, the following mutual information is used:

$$I_p = \sum_{\text{over all specificity groups } i} \sum_{\text{over all amino acids } \alpha} f_p(\alpha, i)\log\frac{f_p(\alpha, i)}{f_p(\alpha)f(i)}, \quad (1)$$

where $f_p(\alpha, i)$ is the frequency of amino acid α at position p in group i; $f_p(\alpha)$ is the frequency of amino acid α at position p in the total sample; and $f(i)$ is the size (fraction) of group i.

Here, a number of corrections, discussed in detail in [14], are introduced to take into account the specific features of actual biological data. The mean of a column and the standard deviation of the distribution of its expected informational content, $\langle I_p^{\exp}\rangle$ and $\sigma(I_p^{\exp})$, are calculated by random shuffling followed by computing the statistical significance for each position:

$$Z_p = \frac{I_p - \langle I_p^{\exp}\rangle}{\sigma(I_p^{\exp})}. \quad (2)$$

To determine the number of SDPs among the most significant positions, an original procedure based on Bernoulli's estimate is used [15]. First, all positions are arranged in order of decreasing Z_p. Then, such k^* is chosen that the obtaining of k^* values of Z that are not lower than $Z_{(k^*)}$ is least probable in the case of the normal distribution of Z (i.e., the least probable set of positions for a random situation, a "heavy tail," is chosen; P is the cut-off probability):

$$\begin{aligned} k^* &= \arg\min_k P\{\text{there exist at least} \\ &\quad k \text{ observations: } Z \ge Z_{(k)}\} \\ &= \arg\min_k\left(1 - \sum_{i=L-k+1}^{L} C_L^i q^i p^{L-i}\right), \end{aligned} \quad (3)$$

where

$$p = P(Z \geq Z_{(k)}) = \int_{Z_{(k)}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-Z^2) dZ,$$

$$\text{and } q = 1 - p.$$

Thus, a set of k^* SDPs is obtained. The probability

$$P^* = P\{\text{there exist at least } k \text{ observations: } Z \geq Z_{(k)}\}, \quad (4)$$

providing this minimum is named the statistical significance of the set of k^* positions.

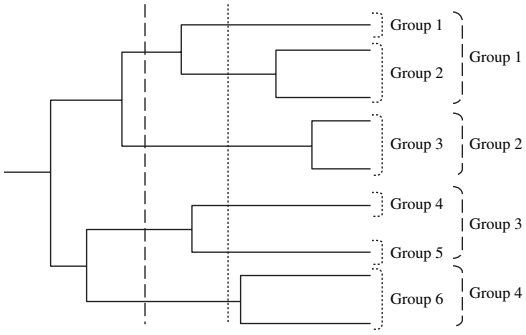

Fig. 1. Grouping in the evolutionary trace method. Two possible groupings are shown with dotted and dashed lines.

Automated Partitioning into Specificity Groups

A wide application of the SDPsite algorithm requires a procedure that provides an automated partitioning of an alignment into specificity groups. We used the technique analogous to that used in the evolutionary trace method [10]. We consider an unrooted initial tree and assume that the root is in the middle of the longest way from one leaf to another. Then, we consider a set of groups generated by dissecting the tree at a certain distance from the root (Fig. 1). In this process, all groups containing less than three sequences are rejected. SDPs are found for each group as described in [14], and the statistical significance for the SDP set P^* is calculated using Eq. (4). The set with the minimal P^*, i.e., the least probable set of SDPs, is considered the best.

However, the Z values, calculated using Eq. (2), need a correction: if sequences are added to the alignment without changing the number of groups but increasing uniformly the number of sequences in each group, the maximal Z value grows, and its growth is well approximated with a logarithmic function (data not shown). This corresponds to the understanding of statistical significance from the standpoint of common sense. However, the increase in statistical significance for the groups with a large number of sequences prevents a correct comparison of the partitions of the same alignment into different numbers of groups: a partition into "thicker" groups always wins. To compensate for this logarithmic growth, we introduce the correction

$$Z := Z/\log(\text{mean group thickness}). \quad (5)$$

Prediction of Conserved Positions

Various approaches to detection of CPs are reviewed by Valdar [16]. In this work, we used the Sander–Schneider conservation measure [17], calculating the conservation of position p as

$$C_p = \left(\sum_{i}^{N} \sum_{j>i}^{N} d(s_i, s_j) M(s_i(p), s_j(p)) \right) / \left(\sum_{i}^{N} \sum_{j>i}^{N} d(s_i, s_j) \right), \quad (6)$$

where N is the number of sequences in an alignment; $d(s_i, s_j)$ is the distance between sequences s_i and s_j, amounting to $1 - \frac{\text{percent identity}}{100}$; $s_k(p)$ is the amino acid at position p of sequence s_k; and $M(\alpha, \beta)$ is the matrix of amino acid substitutions (in this case, we used the matrix BLOSUM62 [18]).

As is indicated in the review [16], this conservation measure is sufficiently satisfactory from the common sense standpoint: its range is continuous and limited (the segment from 0 to 1); it takes into account the frequencies of amino acids in a column as well as the frequencies of amino acid substitutions and their physicochemical properties with the help of the matrices of amino acid substitutions; and it is normalized taking into account the alignment degeneration (i.e., the distances between sequences).

The statistical significance is calculated for each C_p. We introduce the background distribution C_p as conservation of the columns composed of random positions of each sequence in the alignment. Thus, we calculate 10,000 random values of the conservation C_p^{rand} for each C_p and then the statistical significance

$$\tilde{Z}_p = \frac{C_p - \langle C_p^{\text{rand}} \rangle}{\sigma(C_p^{\text{rand}})}. \quad (7)$$

Here, C_p^{rand} accounts for the conservation of columns in the set of unaligned sequences. As the alignment of

two random sequences has a nonzero weight, we center the statistical significance to eventually obtain the following equation:

$$Z_p = \tilde{Z}_p - \langle \tilde{Z}_p \rangle. \quad (8)$$

Then we apply the same procedure for selecting the number of significant positions as when predicting SDPs (Eqs. (3) and (4)).

Segregation of the Best Cluster

To predict the SDPs and CPs, the algorithm requires only an alignment of protein family sequences and the corresponding phylogenetic tree. To select the best cluster, the algorithm additionally requires the 3D structure of a protein from this family. If a family contains several proteins with the determined 3D structures, the resulting cluster may depend on the structure chosen. However, the tests with actual examples demonstrate that the best clusters for different structures overlap sufficiently (data not shown).

On a specified 3D protein structure, our algorithm finds the residues corresponding to the predicted SDPs and CPs and clusters them spatially according to the nested cluster algorithm based on the graph density [19]. Nested clusters are constructed as follows. First, all graph nodes are considered (in our case, they correspond to the set of all SDPs and CPs on the 3D structure), i.e., the cluster H_0. For each node i, its weight is calculated according to the following equation:

$$\mu_i = \lambda_i \sum_j \omega_{ij}, \quad (9)$$

where j runs over the set of the rest H_0 nodes and ω_{ij} is the weight of the edge between the nodes i and j, calculated as

$$\omega_{ij} = \begin{cases} \dfrac{R}{d_{ij}}, & \text{if} \quad d_{ij} < D \\ 0, & \text{if} \quad d_{ij} \geq D, \end{cases} \quad (10)$$

where d_{ij} is the Euclidian distance between the closest atoms of the amino acids that correspond to the nodes i and j; $R = 5$ Å is the mean distance between the atom centers that provide a contact of these atoms; and $D = 15$ Å is the distance over which the effect of an atom extends. R and D are constants; their values were selected from empiric and heuristic considerations. $\lambda_i = 0.5$, if the node i corresponds to CP; otherwise it equals unity. Thus, the CP significance is artificially underrated. This is performed in order to prevent the algorithm from selecting the geometric core (the group of conserved residues necessary for forming a correct 3D protein structure) as a significant cluster.

Then the set of nodes $F_0 \subset H_0$ is found for which μ is minimal and equals $\mu_0^{\min}$. The cluster $H_1 = H_0 \backslash F_0$ is constructed; this procedure is repeated until an empty set is obtained at a certain step. Thus, a family of nested clusters $H_0 \supset H_1 \supset \mathrm{K} \supset H_N \supset \varnothing$ is generated. The cluster n for which $\mu_n^{\min} = \max\{\mu_k^{\min} | k = 0, \ldots, N\}$ is selected as the most significant cluster. Hereinafter, we call this cluster the best cluster.

In this work, we considered two most significant clusters. The second cluster was found according to the same algorithm with a preliminary elimination of all nodes contained in the first cluster from H_0 (the second best cluster).

The algorithm for prediction of the functional site was named SDPsite and realized as a web server, available at http://bioinf.fbb.msu.ru/SDPsite.

RESULTS AND DISCUSSION

The algorithm SDPsite was tested using three examples. SDPsite was applied to the LacI family of bacterial transcription factors, regulating catabolism of various sugars and several other metabolic pathways. Extensive data are available on the specificities of various proteins from this family [20] as well as on the effects of mutations of each residue on the protein function [21]. The results of SDPsite application fit well the experimental data. The operation of SDPsite was compared with that of other methods predicting functional sites [22]. SDPsite displayed better results than the other methods when using the examples considered in this work, namely, LacI and subtilisin-like proteases. SDPsite was applied to a large number of families from the NCBI CDD. This database contains the alignments of protein domains where certain positions are indicated as "features:" the active center, the contact interface with the ligand, a phosphorylation site, etc. We assume that these particular features are functionally important positions. Although we inevitably underestimate our own results when using this approach (since the positions not indicated as features may also be functionally important, whereas the set of features includes certain positions that are beyond the definition of a functional site, for example, phosphorylation sites, glycosylation sites, etc.), SDPsite gives satisfactory results.

Application of SDPsite to the Bacterial LacI Transcription Factor Family

An alignment of regulators from the LacI family containing 125 sequences was considered. The alignment was divided into the following specificity groups differing in the type of effector and DNA operator sequences: PurR, ScrR, RbsR(EC), GntR, RbsR(PP), GalR, MalR, CytR, CcpA, and FruR. This grouping

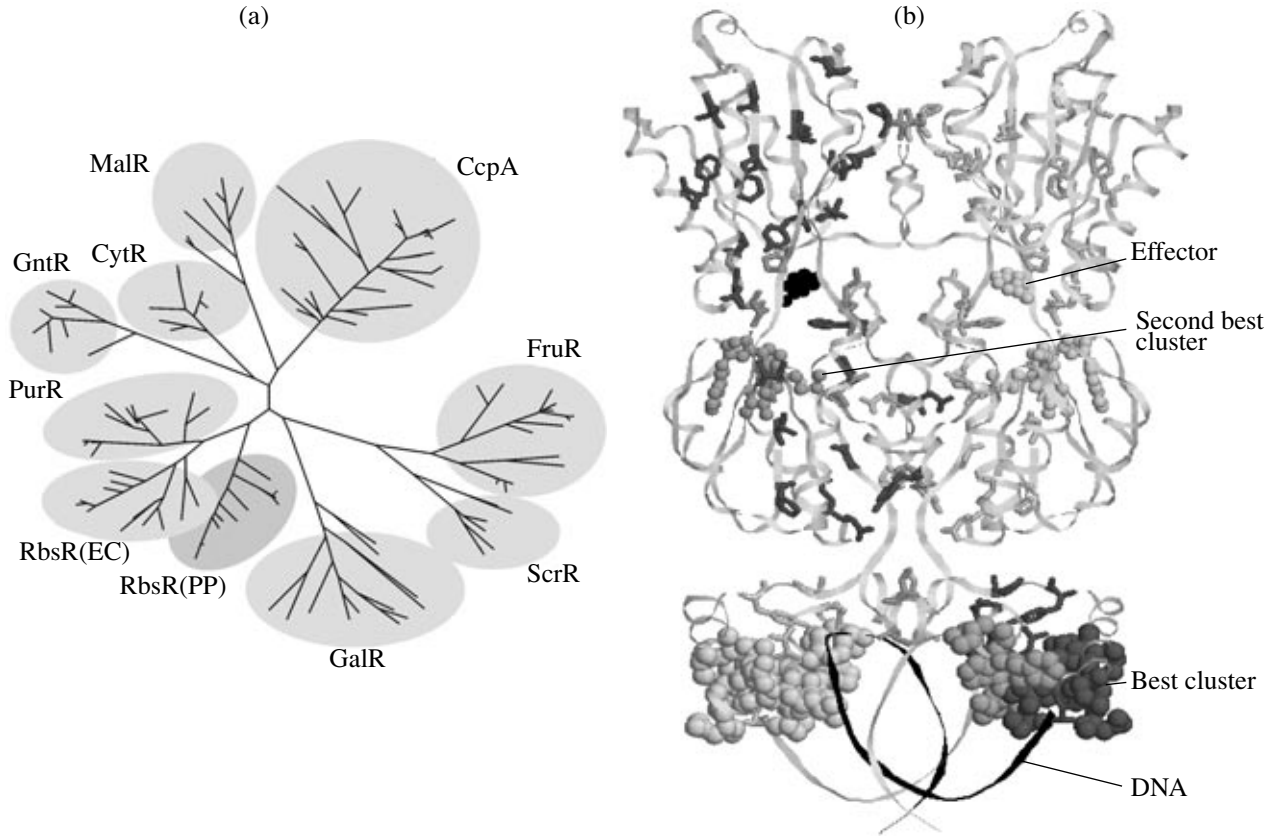

Fig. 2. (a) Phylogenetic tree of the LacI family. (b) SDP (light gray) and CP (dark gray) in the structure of *E. coli* PurR (PDB identifier 1bdh).

was obtained by analysis of the genome context and regulatory sites and by comparative genomics methods [20]. The evolutionary relationships of these proteins and their grouping are shown in a phylogenetic tree (Fig. 2a). We used the structure of *E. coli* PurR (PDB identifier 1bdh) for visualizing the predictions and detecting the clusters.

The above partition into specificity groups was not used for predicting SDPs; on the contrary, the grouping was made automatically. In this process, the groups virtually coincided with the initial groups (Table 1). The predicted positions are shown in Table 2 (numerated as in *E. coli* PurR) and Fig. 2b. As is evident, the majority of predicted SDPs are in the regions of contacts with effector or DNA or the interface between the subunits. This corresponds to an intuitive conception that the amino acid residues responsible for the specific interaction of the protein with its ligand or another subunit must be in these particular regions. CPs also occur in these regions (especially in the interface with DNA); however, they are considerably more numerous inside the protein globule, where they are inaccessible for the solvent and, consequently, unable to be directly involved in the protein function; presumably, they serve for stabilization of the spatial structure. The two best clusters found by SDPsite are in the two most important sites of this protein—the DNA-binding domain and the effector-binding pocket.

Suckow et al. [21] described the effect of a mutation of each amino acid residue in the LacI sequence on its function. All residues are divided into classes depending on whether a particular residue could be substituted and what would be the effect of the possi-

Table 1. Specificity groups separated during automated grouping

Group in automated grouping	Corresponding group separated in [20]
Group 1	CcpA
Group 2	CytR
Group 3	GntR
Group 4	FruR + ScrR
Group 5	MalR
Group 6	GalR
Group 7	RbsR(PP)
Group 8	PurR + RbsR(EC)

Table 2. Positions predicted for the LacI family (numbered as in *E. coli* PurR)

Type of position	Number of predicted positions	Numbering in *E. coli* PurR
SDP	20	5, 15, 16, 20, 25, 27, 53, 55, 91, 96, 123, 144, 145, 146, 147, 160, 162, 284, 294, 323
CP	40	3, 6, 7, 8, 11, 12, 13, 14, 17, 19, 23, 28, 32, 35, 36, 45, 47, 63, 74, 82, 90, 117, 118, 141, 143, 158, 161, 181, 186, 200, 242, 244, 248, 253, 266, 271, 274, 285, 287, 298
Best cluster	19	5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 19, 20, 25, 27, 28, 32, 35, 36
Second best cluster	4	144, 145, 146, 147

ble substitution. Based on these classes, we partitioned all amino acid residues into five groups: (1) residues whose substitution has no effect on the protein function; (2) residues that can be substituted only with small amino acids to retain the protein function; (3) residues that do not contact directly the effector but their substitution impairs its binding or signal transduction; (4) residues that do not contact directly either the effector or DNA and cannot be substituted without a loss of the function; and (5) residues that contact directly the effector or DNA and cannot be substituted without a loss of the function. The distribution of all amino acid residues of the protein and the positions predicted for these groups are shown in Fig. 3. It is evident that the fractions of SDPs and CPs in the groups most important for the protein function (4 and 5) are higher than average and the fraction of clusters is even higher. On the other hand, the more significant the group for the protein function is, the larger the relative number of predicted positions it contains is and the larger the portion of such positions in clusters.

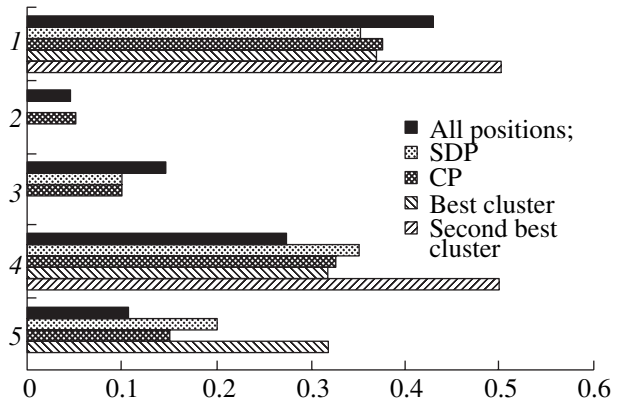

Fig. 3. Distribution of the predicted positions through importance groups for the sample from [20]: (*1*) residues whose substitution has no effect on the protein function; (*2*) residues that can be substituted only with small amino acids; (*3*) residues that do not contact directly the effector but whose substitution impairs its binding or signal transduction; (*4*) residues that do not contact directly the effector or DNA and cannot be substituted; and (*5*) residues that directly contact the effector or DNA and cannot be substituted. The fraction of groups *4* and *5* is higher among the clustering residues.

Comparison of SDPsite with Other Methods

Soyer and Goldstein [22] have compared several methods searching for functional sites with the example of the LacI and subtilisin-like protease families, namely, computing the frequency of the most common amino acid [16], the conservation index based on entropy [16], the Valdar–Thornton conservation index [16], the evolutionary trace method [10], ConSurf [4], the likelihood logarithm calculated using PAML [23], and the evolutionary model of site classes [22]. We compared the results of SDPsite with those reported in [22].

As is mentioned above, the complete data on the effect of a substitution of each residue on the protein function are available for LacI [21]. In the case of subtilisin, there are also extensive data (for approximately half residues of the protein) on the effects of mutations at various positions on the overall function [24]. The rest positions either may be unimportant for the function or have never been studied. We assume further that they are unimportant and, presumably, artificially decrease the quality of our predictions.

Alignments for these tests were constructed as described in [22]. BLAST was applied to search the SwissProt database for the proteins similar to *E. coli* (P03023) LacI with the condition E-value > 0.001; 75 sequences were extracted. Upon discarding the sequences with large terminal deletions, the set contained 70 sequences from 24 bacterial genomes with an average similarity to *E. coli* LacI of 23.5%. In the case of subtilisin-like proteases, we took the family HBG020722 from the database HOBACGEN release 10 [25]. This family contains 80 sequences with a 35% average similarity to *Bacillus amyloliquefaciens* subtilisin. Ten sequences were discarded from the alignment because of large terminal deletions; only the sequence fragment corresponding to the active enzyme was analyzed.

The predicted positions superimposed on the structure of *E. coli* PurR in the case of LacI and *B. amyloliquefaciens* subtilisin in the case of subtilisin-like proteases are shown in Fig. 4. It is evident that the

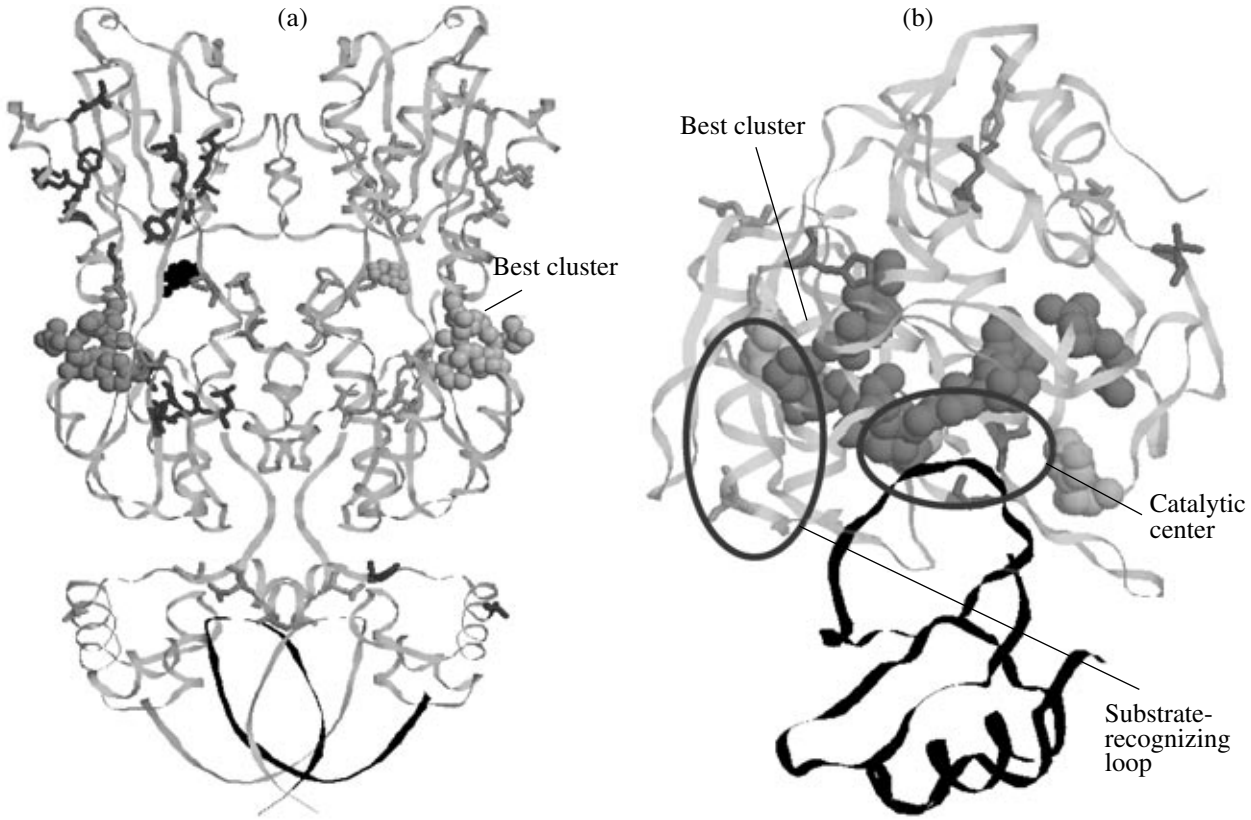

Fig. 4. Predicted positions in the structures of (a) *E. coli* PurR (PDB identifier 1bdh) and (b) *B. amyloliquefaciens* subtilisin (PDB identifier 1to2): SDPs are light gray and CPs, dark gray.

cluster of the DNA-binding domain is lost in LacI; however, the overall prediction is good. In the case of subtilisin, the SDP prediction is weak; therefore, the best cluster is mainly composed of CPs.

The ratio of sensitivity to overprediction is used in [22] as a measure for evaluating the prediction quality. The sensitivity is calculated as $TP/(TP + FN)$, where *TP* (true positives) are the residues predicted by the method and are actually important, and *FN* (false negatives) are the residues unpredicted by the method but are actually important. The overprediction is calculated as $FP/(FP + TN)$, where *FP* (false positives) are the residues predicted by the method but are unimportant, and *TN* (true negatives) are the residues unpredicted by the method and are unimportant. All the methods compared do not predict clusters of important residues but give a continuous list of residues sorted according to the degree of their predicted importance. Therefore, the ratio $TP/(TP + FN)$ versus $FP/(FP + TN)$ is not a point but a plot, called an ROC (receiver operating characteristic) curve. All methods give satisfactory predictions for LacI: the sensitivity exceeds the overprediction over a wide range; however, all predictions are no better than random ones in the case of subtilisin. It has been assumed that such unsatisfactory results for subtilisin are connected with an improper alignment or an overall low similarity between proteases of this type [22].

The analysis using SDPsite predicts not only the relative significance of every position, but also the optimal number of positions; therefore, the SDPsite prediction is a point on the ROC plot. The values of sensitivity, overprediction, and specificity for both families are summarized in Table 3. In all four cases, the SDPsite predictions fall into the lower left corner of the ROC curve and are, at least, not lower than the

Table 3. Results of SDPsite operation in tests [22]

	LacI			Subtilisin		
	overprediction	sensitivity	specificity	overprediction	sensitivity	specificity
A	0.007	0.06	0.75	0.024	0.17	0.57
B	0.05	0.07	0.5	0.043	0.128	0.43

Note: A, a wide range of significant positions (LacI, groups 2–5; subtilisin, all positions with observed changes in activity) and B, a narrow range of significant positions (LacI, group 5; subtilisin, the positions involved in catalytic activity or substrate recognition).

Table 4. Test sample from CDD

Domain name	Alignment length	Total sequences	PDB identifier	Chain	Domain name	Alignment length	Total sequences	PDB identifier	Chain
35EXOc	112	55	2KZM	A	LIGANc	253	44	1DGS	A
53EXOc	194	56	1EXN	B	LMWPc	94	72	1D1P	B
ACTIN	296	45	1NM1	A	MADS	59	91	1MNM	B
ADF	108	51	1COF	A	MYSc	472	43	2MYS	A
aklPPc	301	29	1ELZ	A	PI3Kc	299	21	1E8X	A
Aminopeptidase	59	18	1B65	A	PIPKc	264	16	1BO1	A
AP2	59	23	1GCC	A	PLCc	184	14	1GYM	A
AP2Ec	188	40	1QUM	A	PNPsynthase	230	18	1HO4	A
Arfaptin	194	9	1I4D	A	POLXc	294	10	2BPF	A
BPI	123	31	1BP1	A	PP2Ac	235	19	1AUI	A
C2	72	100	1DQV	A	PP2Cc	158	100	1A6Q	A
CAP_ED	83	100	1RGS	A	PRCH	224	11	1PRC	H
CASc	197	52	1CP3	A	PROF	107	26	1D1J	D
CBM9	144	19	1I82	A	PTB	90	51	2NMB	A
CH	82	53	1AOA	A	PTPc	180	99	2SHP	A
DED	62	32	1A1Z	A	PTS_IIA_fru	107	99	1A6J	B
DEXDc	82	100	1D9X	A	PTS_IIA_lac	97	27	1E2A	A
DSPc	112	51	1VHR	A	PTS_IIA_man	97	43	1PDO	A
DSRM	52	100	1DI2	A	PTS_IIB_glc	74	88	1IBA	A
ENDO3c	115	100	1MUY	A	RA	74	48	1EF5	A
eu-GS	442	10	2HGS	A	RhoGAP	138	75	1AM4	A
fer2	60	100	1B9R	A	S4	51	100	1DM9	B
FGF	107	31	1QQK	A	SAM	53	99	1B0X	A
FH	59	48	1E17	A	Sec7	165	34	1PBV	A
G-α	302	63	1AZT	B	SEC14	123	100	1AUA	A
GMPK	93	57	1GKY	A	SERPIN	239	91	1OVA	A
GYF	55	21	1GYF	A	SH2	54	100	1AYA	A
H15	79	70	1HST	A	SNc	91	30	2SNS	A
HDc	85	100	1F0J	A	TBOX	174	32	1XBR	A
HECTc	313	48	1C4Z	A	TNF	96	33	1A8M	A
HELICc	104	100	1D2M	A	Topo6_Spo	245	25	1D3Y	B
HPT	87	71	1QSP	A	UBCc	129	70	2UCZ	A
HTH_ASER	66	100	1SMT	B	vWFA	83	100	1DZI	A
KISc	224	52	3KAR	A	XPG	249	34	1A76	A

predictions by the other methods tested. SDPsite demonstrates a rather good specificity (the ratio of *TP* to the total number of predicted positions). The position in the lower left corner means that SDPsite displays a rather low sensitivity according to these tests. This may stem from the fact that large enough sets of positions were considered important in these tests, whereas not all of these positions are directly involved in the protein function. For example, all conserved positions responsible for stabilization of the overall protein structure are considered important for LacI. Note that the majority of methods compared in [22] display not only a good sensitivity, but also a high overprediction rate. On the contrary, SDPsite is purposefully designed to decrease the overprediction rate as low as possible.

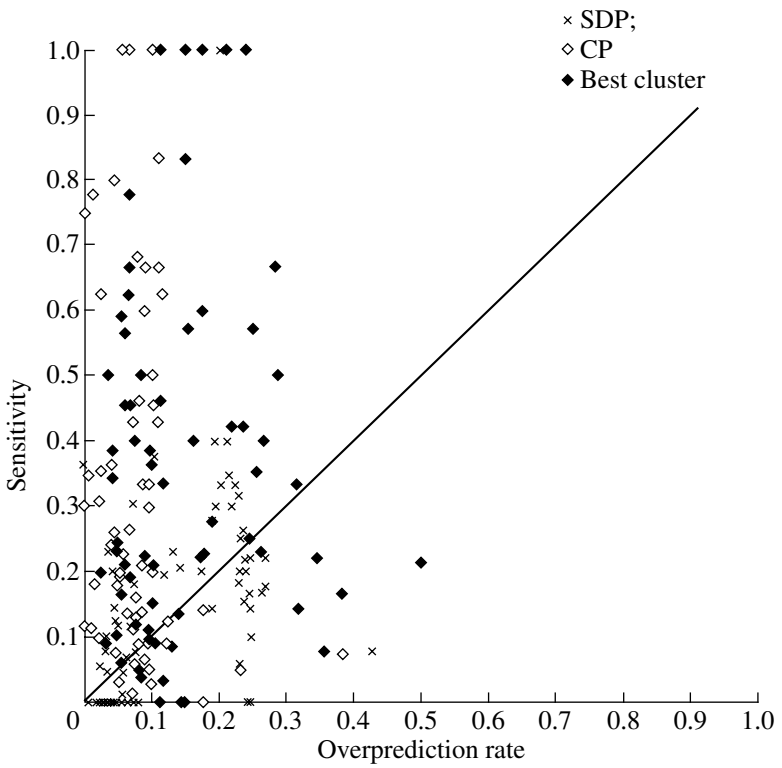

Fig. 5. Ratio of sensitivity to overprediction rate for SDP, CP, and the best cluster for the considered domains from CDD. The diagonal corresponds to a random choice of positions.

Application of SDPsite to a Set from CDD

When analyzing CDD, we considered the same set of domains as Panchenko et al. [26]. These domains have one or several features, and the corresponding alignments contain at least one protein with a known 3D structure. In total, 68 domains were left upon discarding the alignments with a length of less than 50 amino acid residues or with a tree whose structure prevented separation of at least two groups of three or more sequences (Table 4). Only the positions marked with a feature were considered functionally significant. This gives a lower level of quality estimate for the method, as certain residues that were not marked with a feature might also be important, whereas some features might not satisfy an intuitive definition of a functional site, for example, the sites for amino acid modification (phosphorylation, glycosylation, etc.). Thus, the actual sensitivity of the method is not lower and the overprediction rate is not higher than the values determined in such a way.

The ratio of sensitivity $TP/(TP + FN)$ (the ordinate) to overprediction $FP/(FP + TN)$ (the abscissa) is shown in Fig. 5. As is evident, this ratio for CPs and clusters is, on average, better than the random choice of positions (the diagonal) despite the above shortcomings of such evaluation. However, it is not so evident for SDPs. Presumably, this is connected with the fact that only some of the families considered actually contain groups with different specificities or that the majority of annotated features must be implicitly conservative in the total family. The CPs demonstrate a rather good ratio of sensitivity to overprediction (the majority of points fall into the upper triangle); however, the clusters display the best sensitivity on average. Figure 6 shows the number of domains for which the SDPsite predictions exceed a certain sensitivity threshold. It is evident that the clusters surpass the

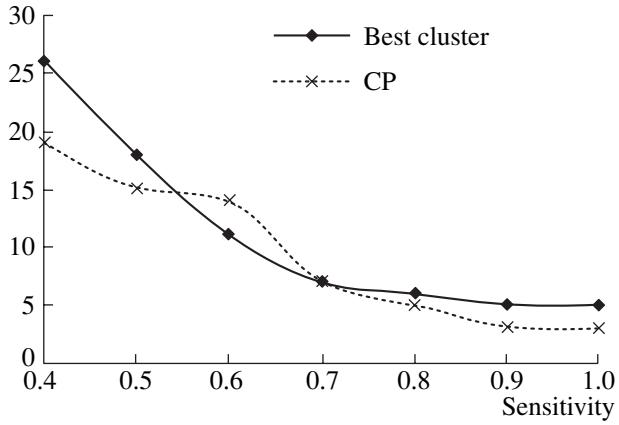

Fig. 6. Number of CDD domains for which SDPsite predictions exceed a specified sensitivity level.

CPs virtually over the entire range of interesting predictions (the sensitivity exceeds or equals 0.4). Moreover, the overprediction rate, in no case, exceeds 0.3. The mean sensitivity for clusters amounts to 0.35; for five families, the sensitivity is 1; the mean sensitivity for CPs is 0.30; and the number of families with a sensitivity of 1 is three.

CONCLUSIONS

This work describes a new method searching for the functional sites, SDPsite. This method combines many features of earlier methods searching for the functional sites, such as a search for conservative positions and selection of the best cluster on the protein structure. The specific feature of our method is the prediction of SDPs and the related automated partition of an alignment into specificity groups.

We compared this method, first, with experimental data (with the example of LacI) and, second, with other analogous methods (with the example of LacI and subtilisin) and performed a mass test with a large number of families from CDD.

The main difficulty when assessing the methods for prediction of functional sites is the absence of reliable controls. In the case of the LacI family, there are virtually complete data on the effect of mutations of various residues on the protein function, while the situation is completely different for other families. In this work, we assumed that all residues not described in the input data are insignificant for the function and decided not to separate the residues according to their functional types, which may essentially underrate the prediction quality. Nonetheless, the prediction results obtained using SDPsite fit well the control data.

Analysis of the predictions obtained for the LacI family and subtilisin-like proteases suggests that, when a family contains distinct specificity groups (as for LacI), SDPs are predicted well and play the primary role when selecting the best cluster. Correspondingly, the predicted functional site was localized to the region of specific interaction. When the specificity groups are indistinct (as for proteases), CPs play the main role in selecting the best cluster.

Comparison of SDPsite with other methods demonstrates that SDPsite operates similar to the best methods and, likely, even better with respect to the ratio of sensitivity to overprediction. However, SDPsite demonstrates a sufficiently low sensitivity. In part, this may be connected with the ideology of SDPsite: a considerably large number of SDPs and CPs predicted at the first stage are discarded when forming the best cluster with the aim to minimize overprediction. However, another explanation is possible: when studying the effect of mutations on the function, a large number of positions that have no direct effect on the function are considered important. It is directly confirmed by the fact that, when the class of significant positions for LacI was narrowed to the essential positions directly involved in the interaction with effectors or DNA, the fraction of predicted positions increased to one-third.

When analyzing the data obtained for the domains from CDD, we see a rather large number of results at the level of a random noise (left lower quarter). Especially bad results were obtained when considering SDP only. This may be connected with the fact that many alignments considered contained a small number of sequences and did not contain proteins with different specificities. In this case, the prediction of SDPs has no sense. The poor results for CPs and clusters may be explained by the properties of certain annotated features; for example, phosphorylation sites are weakly conserved in related proteins. In the case when this is the only annotated specific feature, SDPsite will most likely find the best cluster in some other region of the protein, thereby leading to a very poor prediction (a zero sensitivity). On the other hand, a sufficiently large number of predictions have a sensitivity exceeding 0.4 and an overprediction rate below 0.3, which may be considered a good result.

The goal of structural genomics is identification and functional description of as large a number of proteins from various organisms as possible. As proteins belonging to poorly studied families and having no close homologs with the known structures are frequently chosen for structural analysis, their functional annotation by the available methods (a search for similar well-studied sequences or structures) is difficult. We believe that SDPsite can be successfully applied to search for functional sites in such structures and, consequently, useful for their annotation.

ACKNOWLEDGMENTS

We are grateful to A.A. Mironov for valuable criticism during the work on this project and A.A. Finkel'shtein for critical perusal of the manuscript and beneficial comments.

The work was supported by the Russian Foundation for Basic Research (project no. 04-04-49438), Howard Hughes Medical Institute (grant no. 55005610), INTAS (grant nos. 05-1000008-8028 and 04-83-3704), and the program Molecular and Cell Biology of the Russian Academy of Sciences.

REFERENCES

1. Liolios K., Tavernarakis N., Hugenholtz P., Kyprides, N.C. 2006. The Genome OnLine Database (GOLD) v. 2: A monitor of genome projects worldwide. *Nucleic Acids Res*. **34**, D332–D334.
2. Chandonia J.-M., Brenner S.E. 2006. The impact of structural genomics: Expectations and outcomes. *Science*. **311**, 347–351.

3. Russell R.B., Alber F., Aloy P., Davis F.P., Korkin D., Pichaud M., Topf M., Sali A. 2004. A structural perspective on protein–protein interactions. *Curr. Opin. Struct. Biol*. **14**, 313–324.
4. Glaser F., Pupko T., Paz I., Bell R.E., Bechor-Shental D., Martz E., Ben-Tal N. 2003. ConSurf: Indentification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*. **19**, 163–164.
5. Bell R.E., Ben-Tal N. 2003. In silico identification of functional protein interfaces. *Comp. Funct. Genomics*. **4**, 420–423.
6. Aloy P., Querol E., Aviles F.X., Sternberg M.J.E. 2001. Automated structure-based prediction of functional sites in proteins: Application to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol*. **331**, 395–408.
7. Ma B., Elkayam T., Wolfson H., Nussinov R. 2003. Protein–protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl. Acad. Sci. USA*. **100**, 5772–5777.
8. Landgraf R., Xenarios I., Eisenberg D. 2001. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol*. **307**, 1487–1502.
9. del Sol Mesa A., Pazos F., Valencia A. 2003. Automatic methods for predicting functionally important residues. *J. Mol. Biol*. **326**, 1289–1302.
10. Lichtarge O., Bourne H.R., Cohen F.E. 1996. An evolutionary trace method defined binding surfaces common to protein families. *J. Mol. Biol*. **257**, 342–358.
11. Yao H., Kristensen D.M., Mihalek I., Sowa M.E., Shaw C., Kimmel M., Karvaki L., Lichtarge O. 2003. An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol*. **326**, 255–261.
12. Hannenhalli S.S., Russell R.B. 2000. Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol*. **303**, 61–76.
13. Mirny L.A., Gelfand M.S. 2002. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J. Mol. Biol*. **321**, 7–20.
14. Kalinina O.V., Mironov A.A., Gelfand M.S., 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.
15. Vinogradov D.V., Mironov A.A. 2002. SiteProb: Yet another algorithm to find regulatory signals in nucleotide sequences. *Proc. 3rd Int. Conf. on Bioinformatics of Genome Regulation and Structure BGRS'2002, Novosibirsk, Russia*, pp. 28–30.
16. Valdar W.S.J. 2002. Scoring residue conservation. *Proteins*. **48**, 227–241.
17. Casari G., Sander C., Valencia A. 1995. A method to predict functional residues in proteins. *Nature Struct. Biol*. **2**, 171–178.
18. Henikoff S., Henikoff J. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci*. **89**, 10,915–10,919.
19. Mirkin B., Muchnik I. 2002. Layered clusters of tightness set functions. *Appl. Math. Lett*. **15**, 147–151.
20. Laikova O.N. 2003. The LacI family of bacterial transcriptional regulators and the evolution of sugar utilization regulons in bacteria. In: *Proc. 1st Int. Moscow Conference on Computational Molecular Biology MCCMB'03, Moscow, Russia*, pp. 121–122.
21. Suckow J., Markiewicz P., Kleina L.G., Miller J., Kisters-Woike B., Muller-Hill B. 1996. Genetic studies of the Lac Repressor XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. *J. Mol. Biol*. **261**, 509–523.
22. Soyer O.S., Goldstein R.A. 2004. Predicting functional sites in proteins: site-specific evolutionary models and their application to neurotransmitter transporters. *J. Mol. Biol*. **339**, 227–242.
23. Ko J., Murga L.F., Andre P., Yang H., Ondrechen M.J., Williams R.J., Agunwamba A., Budil D.E. 2005. Statistical criteria for the identification of protein active sites using theoretical microscopic titration curves. *Proteins*. **59**, 183–195.
24. Bryan P.N. 2000. Protein engineering of subtilisin. *Biochim. Biophys. Acta*. **1543**, 203–222.
25. Perriere G., Duret L., Gouy M. 2000. HOBACGEN: Database system for comparative genomics in bacteria. *Genome Res*. **10**, 379–385.
26. Panchenko A.R., Kondrashov F., Bryant S. 2004. Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci*. **13**, 884–892.