SitCon: Binding Site Residue Conservation Visualization and Protein Sequence-to-Function Tool

VISVALDAS KAIRYS, MIGUEL X. FERNANDES

Centro de Química da Madeira, Universidade da Madeira, Campus da Penteada, 9000–390 Funchal, Portugal

Received 10 October 2006; accepted 8 January 2007
Published online 7 May 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI 10.1002/qua.21396

ABSTRACT: We introduce SitCon (SITe CONservation), a program designed to explore conservation of functionally important sites in a series of hypothetically homologous candidate protein structures, given amino acid sequence as an input. This can especially be useful when looking for an unknown function of a protein. SitCon exploits the fact that binding sites of proteins are preserved better than the overall residue sequence conservation. To test the capability of unknown function prediction, we randomly chose known function proteins from *Caenorhabditis elegans* genome. To imitate a behavior of an unknown function target, only the low homology proteins with $0.01 < E\text{-score} \leq 100$ were analyzed as templates. Out of 29 enzyme targets, SitCon was able to provide various hints about their function in at least 69% of the cases. For the eight nonenzyme targets, the predictions matched in only 25% of the cases. SitCon was also tested for a capability to predict presence or absence of metal-containing heterogroups in the target enzymes with ~80% success rate. Because this algorithm is not based on specific protein signatures, it may allow detection of overlooked relationships between proteins. SitCon is also very effective as a tool allowing visual comparison of binding site residue conservation between the target and homologous templates side-by-side. © 2007 Wiley Periodicals, Inc. Int J Quantum Chem 107: 2100–2110, 2007

Key words: genome; sequence-to-function; protein function prediction; binding site; amino acid residue conservation

1. Introduction

One of formidable tasks that scientists face with the advance of number of solved genomes is the correct assignment of the functions of proteins encoded by Open Reading Frames. For example, up to 60% of the malaria parasite *Plasmodium falciparum* genome did not have sufficient similarity to proteins in other organisms to justify provision of functional assignments [1].

Currently there exist a variety of tools or web servers, which search and/or catalog signatures of characteristic domains or functionally important regions in proteins, which take protein sequence as an input. These methods have been reviewed elsewhere [2, 3], but we will briefly mention some of

This article contains supplementary material available via the Internet at http://www.interscience.wiley.com/jpages/0020–7608/suppmat.

Correspondence to: M. X. Fernandes; e-mail: mxf@uma.pt
Contract grant sponsor: Fundação para a Ciência e a Tecnologia.
Contract grant number: SFRH/BPD/20716/2004.

them. Probably one of the best known tools is Inter Pro [4]/InterProScan [5], which puts together nine different protein signature databases. ConSeq [6] identifies structurally and functionally important residues in protein sequences by analyzing multiple sequence alignments. SAS [7] uses FASTA to scan a given amino acid sequence against all proteins in PDB (Protein Data Bank [8]), yielding multiple alignment annotated by different structural features.

Because of substrate recognition, it is well known that active site residues are preserved better compared to the overall protein conservation. The idea behind the proposed method is to use structural data from PDB to enrich target protein sequence alignment results with proteins for which residues in the functional sites are preserved better. The approach does not require availability of a 3-D structure for the target. The conservation of the binding site residues has been utilized before, but most existing methods deal with 3-D descriptors or templates around the binding site, and hence are generally applicable to structure-to-function predictions. In a method closest in spirit to the discussed in this article, Das and Gerstein used active site sequence conservation ratio ("ASC ratio") to detect functional shifts in isocitrate dehydrogenase protein family [9]. The methods that connect 3-D structure of the binding site and the function of protein are much more numerous, notably in works by Thornton and coworkers, for example [10] and [11]. Panchenko et al. described a method to discover functional sites of the proteins, using functional site residue conservation and spatial clustering, given 3-D structure as input [12]. Fetrow and Skolnick developed "fuzzy functional forms," or FFF, approach [13], which used sequence–structure–function paradigm, based on a related approach, to mine *E. coli* genome for proteins with glutaredoxin/thioredoxin disulfide oxidoreductase activity [14]. Our approach is different from Fetrow and Skolnick's, since essentially they compare structure under investigation (or homology model built by threading if the structure is not known) to a precomputed three-dimensional descriptor of the functional site. In contrast to many existing variants of the method, our approach uses no previous knowledge of precomputed motifs or patterns, nor it explicitly uses 3-D functional site descriptors, which in essence are 3-D version of motifs. In this sense the proposed procedure can be considered "first principles" method. Our method is quite simple and could be used by researchers to answer the following questions: Are there structures in PDB, which have binding sites that potentially can be present in the test sequence? If the answer is "yes," what are the matching residues between the target sequence and template PDB structure?

The proposed method may be especially useful in the "twilight zone" (20–35%) of the sequence identities between proteins in which explosion of false positives is observed [15]. We named this approach SitCon (SITe CONservation). An advantage of the proposed method is that it is not based on a precompiled list of functional-site specific patterns, and therefore can detect previously undiscovered trends, and hence may complement existing tools. In addition, this method allows simple side-by-side comparison of the binding sites between the target protein and series of homologous templates. Presently, SitCon targets cavities and heterogroups (including nucleotides and short peptidic substrates) inside the PDB files, but this approach could be extended to include some other sequence enrichment "hot spots," for example, protein–protein or subunit interfaces.

This article is organized as follows: in the Methods section the algorithm is introduced. In Section 3 we discuss the ability of SitCon to find among the list of low homology proteins those with Enzyme Commission (E.C.) number similar to the target, as well as to discover co-factors and specialized domains of the target. In another subsection, SitCon prediction of metal presence in the target protein is analyzed. Finally, in Section 4 we present advantages and disadvantages of the method, and some possible venues for improvement.

2. Methods

Figure 1 shows an outline of the algorithm. SitCon needs externally generated list of homologous proteins and their alignments against the target gene. The alignments were generated using FASTA [16] by aligning the target gene sequence against the sequences in the Protein Data Bank (as of March 31, 2006), with upper *E*-score threshold of 100. The resulting FASTA report file was used as an input into a SitCon script, written in Perl. The method therefore does not produce any other proteins rather than those generated by FASTA, but SitCon adjusts the ranking of the proteins. The unknown function gene behavior was imitated by removing all highly homologous proteins with *E*-score ≤ 0.01, leaving only the

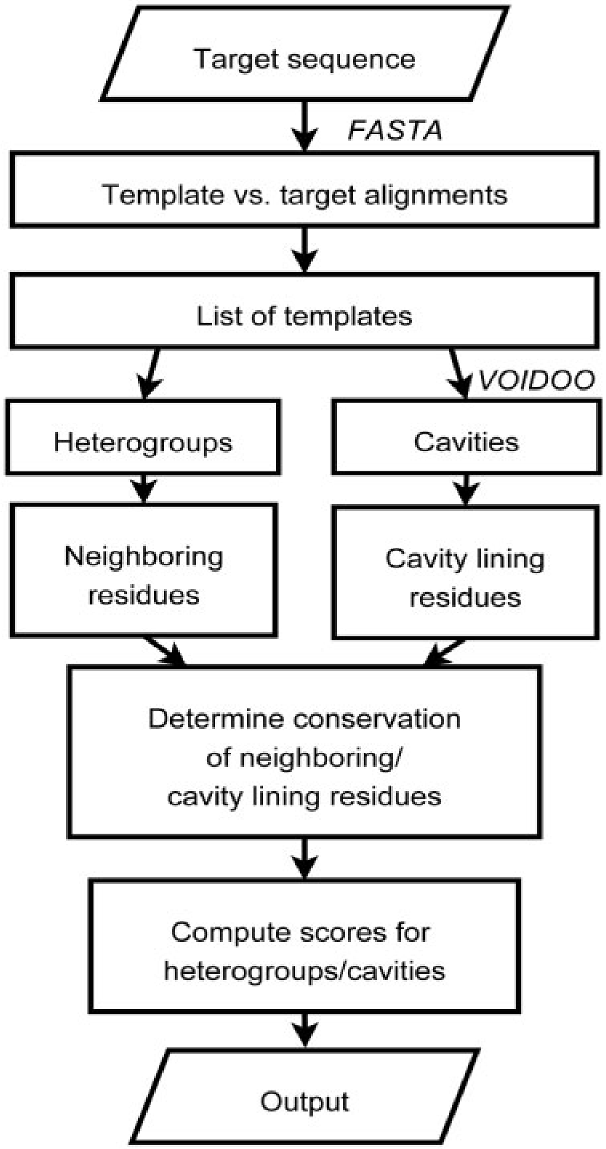

Target sequence
FASTA
Template vs. target alignments
List of templates
VOIDOO
Heterogroups
Cavities
Neighboring residues
Cavity lining residues
Determine conservation of neighboring/ cavity lining residues
Compute scores for heterogroups/cavities
Output

FIGURE 1. SitCon flowchart. External programs used to execute some of the steps are italicized. The FASTA alignments are used to obtain list of template proteins from the PDB. The SitCon script then finds heterogroups in those templates and determines their neighboring residues. VOIDOO program is used to generate list of residues around the cavities. The script then determines conservation of heterogroups and cavities from the FASTA alignments and computes their scores.

weakly homologous ones. *E*-score or *E*-value is the most frequently used statistical estimator of the validity of alignment scores. The templates with *E*-score values between 0.01 and 100 used in this paper usually corresponded to the 20–35% identity between the target and the template, and therefore signified the protein similarity "twilight" zone with many false positives [15].

An essential part of SitCon is determination of neighbors of heterogroups (nonpeptide molecules and fragments) and cavities. The neighbors of heterogroups were defined as all amino residues having at least one heavy atom within 6 Å from heavy atoms of a heterogroup (water molecules were excluded). The cavity neighboring residues were found using VOIDOO program [17]. In principle, the neighboring residue lists could be pre-computed for each PDB entry, which would dramatically reduce SitCon running time. In many instances, cavities and heterogroups overlapped, but we did not remove the duplicates. Next, conservation of the neighboring residues between the homolog and the target was examined for each heterogroup/cavity using the initially generated alignment, and scored. For this preliminary study we used raw score computed using BLOSUM62 amino acid substitution matrix [18] (Fig. 2). For the sake of convenience, the BLOSUM62 matrix scores are further multiplied by 25, so that the conserved residues scored at least +100 and the dissimilar residues scored up to −100, and missing residues are scored −100. For example, conserved alanine and tryptophane have +100 and +425 score, respectively. Differently from the conventional BLOSUM scores, gaps were not penalized because neighbors of heterogroups generally consist of several patches of amino residue chain. The modified raw BLOSUM62 score will be referred to as the SitCon score. Among our future plans is the computation of a substitution matrix better represent-

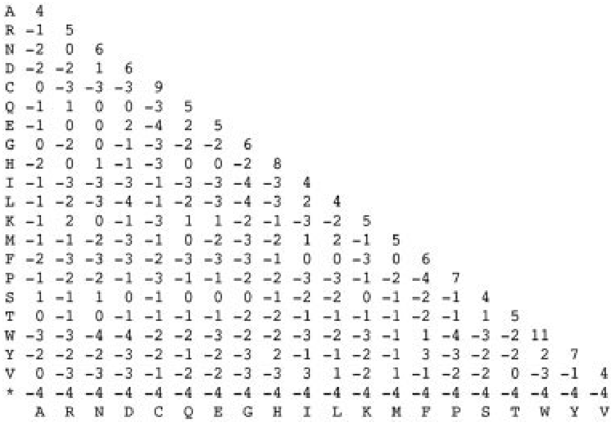
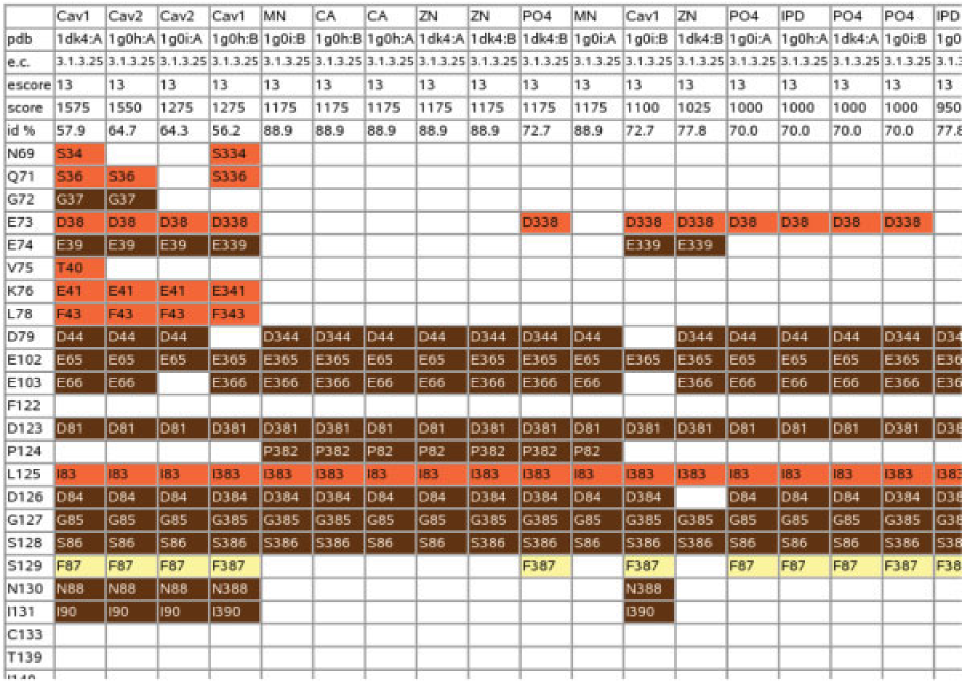

```
A  4
R -1  5
N -2  0  6
D -2 -2  1  6
C  0 -3 -3 -3  9
Q -1  1  0  0 -3  5
E -1  0  0  2 -4  2  5
G  0 -2  0 -1 -3 -2 -2  6
H -2  0  1 -1 -3  0  0 -2  8
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
```

FIGURE 2. BLOSUM62 amino acid substitution matrix [18].



	Cav1	Cav2	Cav2	Cav1	MN	CA	CA	ZN	ZN	PO4	MN	Cav1	ZN	PO4	IPD	PO4	PO4	IPD
pdb	1dk4:A	1g0h:A	1g0i:A	1g0h:B	1g0i:B	1g0h:B	1g0h:A	1dk4:A	1dk4:B	1dk4:B	1g0i:A	1g0i:B	1dk4:B	1g0i:A	1g0h:A	1dk4:A	1g0i:B	1g0
e.c.	3.1.3.25	3.1.3.25	3.1.3.25	3.1.3.25	3.1.3.25	3.1.3.25	3.1.3.25	3.1.3.25	3.1.3.25	3.1.3.25	3.1.3.25	3.1.3.25	3.1.3.25	3.1.3.25	3.1.3.25	3.1.3.25	3.1.3.25	3.1.3
escore	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13
score	1575	1550	1275	1275	1175	1175	1175	1175	1175	1175	1175	1100	1025	1000	1000	1000	1000	950
id %	57.9	64.7	64.3	56.2	88.9	88.9	88.9	88.9	88.9	72.7	88.9	72.7	77.8	70.0	70.0	70.0	70.0	77.8
N69	S34			S334														
Q71	S36	S36		S336														
G72	G37	G37																
E73	D38	D38	D38	D338						D338		D338	D338	D38	D38	D38	D338	
E74	E39	E39	E39	E339								E339	E339					
V75	T40																	
K76	E41	E41	E41	E341														
L78	F43	F43	F43	F343														
D79	D44	D44	D44		D344	D344	D44	D44	D344	D344	D44		D344	D44	D44	D44	D344	D34
E102	E65	E65	E65	E365	E365	E365	E65	E65	E365	E365	E65	E365	E365	E65	E65	E65	E365	E36
E103	E66	E66		E366	E366	E366	E66	E66	E366	E366	E66		E366	E66	E66	E66	E366	E36
F122																		
D123	D81	D81	D81	D381	D381	D381	D81	D81	D381	D381	D81	D381	D381	D81	D81	D81	D381	D38
P124					P382	P382	P82	P82	P382	P382	P82							
L125	I83	I83	I83	I383	I383	I383	I83	I83	I383	I383	I83	I383	I383	I83	I83	I83	I383	I38
D126	D84	D84	D84	D384	D384	D384	D84	D84	D384	D384	D84	D384		D84	D84	D84	D384	D38
G127	G85	G85	G85	G385	G385	G385	G85	G85	G385	G385	G85	G385	G385	G85	G85	G85	G385	G38
S128	S86	S86	S86	S386	S386	S386	S86	S86	S386	S386	S86	S386	S386	S86	S86	S86	S386	S38
S129	F87	F87	F87	F387						F387		F387		F87	F87	F87	F387	F38
N130	N88	N88	N88	N388								N388						
I131	I90	I90	I90	I390								I390						
C133																		
T139																		

FIGURE 3. An example of SitCon output in HTML format for *C. elegans* gene Q9N2M2, fructose-1,6-biphosphatase protein 1. The columns contain the SitCon hits (cavities and heterogroups), rows contain information about protein (E.C. number and *E*-score), followed by SitCon score and identity percent for the hit. Starting from the seventh row, aligned amino acid residues are highlighted: dark cells signify identity between target and template, lighter cells – similarity (conservative replacement), and light cells – dissimilar residues. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

ing possible amino acid mutations around cofactors and substrates, compared to BLOSUM.

Selenomethionines and other modified amino acid residues were converted to the original amino acids by batch editing of the PDB file using a script. Each nucleotide or residue in PDB chains consisting of nucleic acids and short amino acid sequences (<21 amino acid residues) were considered as heterogroups. If the neighboring residue was outside FASTA aligned region, it was not included in the score.

SitCon was executed on a Linux box equipped with 1 GHz AMD Athlon 64-bit processor. Depending on the number of homologous proteins and their nature, it took from several minutes to few hours for an analysis to complete per target, with cavity detection as the most time consuming step. The final results were generated as a table in an HTML format for a convenient browsing (Fig. 3).

For testing of SitCon, we selected genes from the relatively well investigated *Caenorhabditis elegans* genome. Out of nearly 23,000 genes available, for example, in UniProt, a subset of about 4,000 that had an unambiguously annotated function (the annotation did not have words "unknown," "hypothetical," "putative," etc.) was selected. Initially, 14 genes with annotated E.C. numbers were selected at random from the latter set. The number of test proteins was fairly low because at this method development stage analysis of the best scorers was fairly laborious because of the need to inspect the hits thoroughly in order to compile results presented here. Since some of the previously selected enzymes had identical or similar (i.e. having same level three E.C. subclass) E.C. numbers, additional 15 genes were randomly chosen so that none of the newly selected genes had the same level three subclass with previously selected genes. Both selections were added together, resulting in a test set consisting of 29 gene-encoded enzymes. For detection of domains and cofactors (Subsection 3.2), eight random genes without an E.C. number (nonenzymes) were additionally selected.

3. Results and Discussion

To investigate the behavior of SitCon hits, we tested if SitCon hits can provide information about the E.C. number of the "unknown" target gene, or give some hints about their cofactors, substrates, and specialized domains. Depending on the behavior of the E.C. numbers of the homologous proteins, all targets were divided into two groups. The first group contained all targets for which low homology proteins (i.e. with *E*-scores between 0.01 and 100) had at least one protein with a similar E.C. number to the target. Sixteen targets were in this group. The remaining 21 targets were included in the second group. For this group matching of correct cofactors and/or domains between target and SitCon top hits was analyzed in Subsection 3.2. In Subsection 3.3 prediction of metal-containing species in the target enzymes is analyzed. The FASTA alignments and SitCon outputs for the investigated proteins are available as Supplementary Materials from this Journal's web site.

3.1. PREDICTION OF THE E.C. NUMBERS

We considered the E.C. numbers to be similar between two enzymes if they belonged to the same level 3 or 4 subclass, for example, numbers 2.1.1.98, 2.1.1.1, and 2.1.1.- are similar to 2.1.1.98 for our purposes. For batch processing of the proteins, E.C. numbers for protein chains were taken from the PDBSProtEC database [19]. For further analysis of SitCon output, PDBsum [20] or OCA (J. Prilusky, http://bip.weizmann.ac.il/oca-bin/ocamain) annotations were examined if PDBSProtEC did not report the E.C. number. Since some proteins closely related to the target did not have an assigned E.C. number, hits originating from correct cofactors or bound to correct domains, as annotated by InterPro [4], were considered as successful hits.

Table I shows comparison of SitCon and FASTA rankings for the 16 target enzymes from this group. Heterogroups/cavities originated from very similar protein chains, for example, having simultaneously identical E.C. and similar *E*-scores, were considered as one hit. Based on evaluation of results of Subsections 3.1–3.3 we chose SitCon score threshold of 600, above which hits we considered good, even though, as it can be seen in analysis later, there are some meaningful hits below 600, and some false hits above 600. In 14 out of 16 cases a correct hit was found by SitCon among the topmost two hits, which was better compared to 12 cases of analogous prediction by FASTA (the ranking in the latter was based on the *E*-score of the protein chain). We used the top two hits because of the necessity to take into account possible false positives (see examples in the next subsection). A high rate of good hits found by FASTA shows that among proteins with *E*-scores just above 0.01 relevant proteins can often be found. However, SitCon was able to highlight relevant proteins among the low homologous proteins. SitCon was clearly superior to FASTA for two targets, ubiquitin ligase protein

TABLE I
Ranks of best SitCon hits with E.C. numbers (with three or four positions) coinciding with the target gene, among low homology proteins (0.01 < *E*-score ≤ 100).

		Rank of best SitCon and FASTA hit with close E.C. (SitCon score)	
Gene	E.C.	SitCon	FASTA
G3P4_CAEEL	1.2.1.12	1 (1675)	1
CATA2_CAEEL	1.11.1.6	low (300)	low
		1 (900)[a]	*1*
Q17514	2.1.1.98	2 (900)	1
		1 (925)[b]	
FLR4_CAEEL	2.7.11.1	2 (1450)	3
		1 (2050)[c]	*1*
O61371	3.1.1.7	2–3 (1125)	1
Q9N2M2	3.1.3.11	1 (1575)	1
O62272	3.1.3.16	1 (1725)	1
Q27501	3.1.3.16	1 (1800)	1
CPR3_CAEEL	3.4.22.-	6 (525)	2
NAS38_CAEEL	3.4.24.21	2 (950)	low
PSA2_CAEEL	3.4.25.1	2 (850)	1
Q9XUV0	3.4.25.1	low (275)	1
BLM_CAEEL	3.6.1.-	low (325)	2
		1 (775)[d]	*1*
O16880	4.6.1.1	2 (850)	3
Q7K707	5.3.1.9	1 (1150)	4
RPM1_CAEEL	6.3.2.-	2 (1400)	low
		1 (1725)[e]	*low*

The dash in column 2 indicates absence of chains with close E.C. numbers among FASTA hits in the *E*-score ranges specified. The corresponding SitCon scores are in parentheses. In italics: hits from proteins that do not have a similar E.C. to the target but are otherwise closely related, either through cofactor, or a common domain.

[a] Target and template shares heme cofactor with the target.

[b] SAH cofactor and methyltransferase domain.

[c] Kinase domain.

[d] DEAD/DEAH helicase domain, same as in target.

[e] RING-type zinc finger domain.

TABLE II
The SitCon results for genes not included in Table I.

Target	E.C. (enzymes) or annotation (nonenzymes)	SitCon true positives
DHSA_CAEEL	1.3.5.1	Ranks 1–3: oxidoreductases, FAD cofactor
NU5M_CAEEL	1.6.5.3	*Ranks 1–2: other oxidoreductases*
GALT5_CAEEL	2.4.1.41	—
SQV2_CAEEL	2.4.1.134	—
KICB2_CAEEL	2.7.1.32	Rank 2: kinase (level 2 E.C. similarity)
RPB2_CAEEL	2.7.7.6	Rank 1: Zn; rank 2: level 2 E.C. similarity
VATC_CAEEL	3.6.3.14	*High hits related to nucleic acid binding*
FZO1_CAEEL	3.6.5.-	Ranks 1, 2: similar substrates (ADP vs. GTP)
APN1_CAEEL	4.2.99.18	Rank 1: TIM barrel domain
CYP1_CAEEL	5.2.1.8	—
PDI1_CAEEL	5.3.4.1	Ranks 1–3 hits: thioredoxin domain
SYK_CAEEL	6.1.1.6	*High hits related to nucleic acid binding*
DNLI_CAEEL	6.5.1.1	*Rank 1 hit: nucleic acid*
ADF2_CAEEL	Actin-depolymerizing factor 2	Rank 1: ADF domain
O02144	Prion-like protein 22, isoform c	—
O44760	Prion-like protein 64, isoform a	—
O61883	Seven TM receptor protein 114	*Single- or multi-pass membrane protein hits*
O61947	Ground-like protein 28	—
Q6AHQ3	Troponin T protein 3, isoform c	—
Q7JNG6	Uncoordinated protein 73, isoform c	Ranks 1–4: DH, SH3, FN III, PH domains
SRA18_CAEEL	Serpentine receptor class α-18	—

On the right listed are the high SitCon hits, which provide an insight into the function, domains, or cofactors of the target gene among the low homology proteins ($0.01 < E$-score ≤ 100).

RPM-1 (RPM1_CAEEL), and zinc metalloproteinase NAS-38 precursor (NAS38_CAEEL). For the latter, the relevant (rank 2) hit originated from an inhibitor in matrix metalloproteinase-3 1CAQ (*E*-score: 84; SitCon score: 950). The false positive rank 1 metallothionein hit for this target is analyzed in the next subsection. For RPM1_CAEEL target, the multiple topmost SitCon hits originate from zinc atoms in RING zinc finger domains, with *E*-scores for the templates ranging from 8 to 60.

3.2. PREDICTION OF COFACTORS AND DOMAINS

Table II contains list of correct SitCon predictions for 13 enzymes not included in Table I, i.e. enzymes for which among low homology proteins according to PDBSProtEC there are no proteins with similar E.C. number. For this group of targets, we looked for correspondences between the cofactors and domains of the target and the SitCon high hits, and also for hints about the type of substrate. Information about cofactors and domains was taken from PDBsum [20] and InterPro [4]. In addition, eight randomly chosen nonenzyme encoding genes from *C. elegans* genome (i.e., without an E.C. number) were added to the list.

Out of 29 and 8 enzyme and nonenzyme targets from both tables, in about 20 + 4 and 2 + 1 cases, correspondingly, SitCon gives correct predictions of similar E.C. number, or similar substrate, cofactor or correct specialized domain. "+4" denotes the number of less strong predictions; in Table II the latter are specified in italics. A better rate of predictions for enzymes compared to nonenzymes can be explained by several reasons. Firstly, catalytic sites are under stronger selection pressure which causes good conservation of catalytic sites among proteins with large sequence diversity [12]. Secondly, some domains in nonenzymes do not have associated specific heterogroups/cavities associated with them. Thirdly, it is likely that nonenzymes are less well represented in PDB database, because enzymes are the most common targets of pharmaceutical research which is reflected in numbers of structures in PDB, or could be harder to crystallize (i.e., transmembrane proteins). For example, the median number of highly homologous (E-score ≥ 0.01) protein chains in the PDB database for enzymes used in this study is 30, while for nonenzymes it is 11. For these reasons, existing motif databases in many cases, especially for non-enzymes, are better suited to perform the search for domains.

An example of successful prediction is uncoordinated protein 73, isoform c (UNC-73), UniProt code Q7JNG6. UNC-73 is required for cell migrations and axon guidance in *C. elegans* [21]. According to InterPro, this protein contains several domains. Many of them are discovered by the algorithm: a cavity in 1XCG, chain A (*E*-score: 0.02), has score 875 and corresponds to Dbl homology (DH) domain. A cavity in 1UEC (*E*-score: 19) has score 850 and corresponds to Src homology-3 (SH3) domain. N-acetyl-d-glucosamine (NAG) molecule in 1CFB (*E*-score: 78) has score 725 and corresponds to fibronectin, type III (FN III) domain. A cavity in 1AWE (*E*-score: 11) has score 625 and corresponds to pleckstrin homology (PH) domain.

It is interesting to analyze cases where target-related proteins had low Sitcon scores while FASTA ranked them high based on the *E*-score (false negatives). Examples are cathepsin B-like cysteine proteinase three precursor (CPR3_CAEEL) and proteasome subunit Q9XUV0 (Table I). For CPR3_CAEEL, the highest hit with similar E.C. scored only 525. The low score can be explained by the fact that it originates from cathepsin L light chain 1MHW with *E*-score 1.9, which is only 42 residues long, and this chain does not have enough residues close to the binding site to accumulate high score. The other, heavy chain of the same template, has *E*-score 0.001 and SitCon score 1575 but it falls outside of the list of low homologs.

Related problems can arise when several distinct subunits are assigned the same E.C. number. For example, 14 subunits of proteasome 20S share the same E.C. number (3.4.25.1), while each of them are functionally specialized within the proteasome. Some of them have a weak homology with each other. Under these circumstances, it is not surprising that SitCon does not find good scores among other weakly homologous proteasome subunits for proteasome subunit Q9XUV0 (homologous to subunit $\beta 5$ in some other organisms [22]).[1]

Because SitCon uses a different approach compared to most of sequence-to-function programs, it may highlight some possible relationships between target and the template, which could be overlooked using other methods, or it may point to possibly interesting domains or regions in the target. The latter can be illustrated using proteasome subunit Q9XUV0 target as an example. Alignment against PDB database gives many highly homologous proteasome subunits in other organisms, but Q9XUV0 residues 1–64 are outside of these alignments. Running alignment of these residues against all other proteins in UniProt database finds only one close homolog, subunit in a proteasome of related organism, *C. briggsae*. This fact may signify existence of some special function for these residues in *Caenorhabditis* genus. SitCon may provide additional hints where other methods fail. For example, SitCon finds acetyl in 1fu1 (DNA repair protein). Its score is only 475, but the overall conservation for the neighbors is 56%, and the fragment is fairly short (compare with CPR3_CAEEL light chain example above).

Interestingly, in three out of four investigated oxidoreductase (E.C. 1.-.-.-) targets hemes (porphyrins) have high scores. However, in only one case, CATA2_CAEEL (catalase-2), heme is the actual cofactor. In two other targets, G3P4_CAEEL (glyceraldehydes-3-phosphate dehydrogenase) and NU5M_CAEEL (NADH:ubiquinone oxidoreductase chain 5), it is apparently a false hit since both enzymes do not have heme cofactors: the first one is NAD-dependent, and the second one is a subunit in a respiratory complex I, which has a very complicated and relatively little investigated structure with flavine nucleotides and iron-sulfur clusters as cofactors [23, 24]. While the reason of this high score is unclear, it may signify some relationship (for example, evolutionary) between oxidoreductases with different cofactors. Interestingly, hemes did not score high in other investigated targets. However, the false or true positives of the high heme hits can often be recognized by analyzing conservation of residues responsible crucial interactions of heme with the protein, for example, axial ligands of the heme metal atom.

Several highest hits for NAS38_CAEEL target are metal atoms in metallothioneins (highest SitCon score 1225 in 1AQS). The *E*-score of 1AQS (Cu-metallothionein) is 56, and the template and the target have an overall 33% sequence identity. However, if only the neighbors of the copper metal atoms inside the protein are considered (marked with "+" in the fifth row of a table within Fig. 4), the identity increases to 55%. The region of NAS38_CAEEL corresponding to this region is an-

[1] At the time of preparation of manuscript, the UniProt annotation of Q9XUV0 was changed to "hypothetical protein pbs-5," but other tools (sequence alignment, InterPro) indicate a close relation to proteasome subunits, and hence this does affect our conclusions about this target.

```
TSP1 consensus     W  WS  CS  CG G    R  R       C       C    C
NAS38_CAEEL       AWSEWQGECSQQCGGCGHRLR-KRECK-KEACRKEEKRPCNFSACP
Similarity               ::. :::.: .  . .. :.   .: ...: ::
MTCU_YEAST (1aqs) NFQNEGHECQCQCGSCKNNEQCQKSCSCPTGCNSDDKCPCGNKSEE
Cu Neighbors             +++++++++      +  ++    + +++++++
```

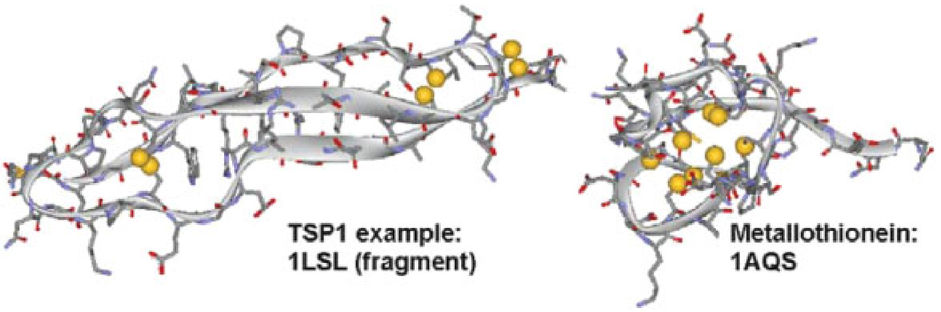


FIGURE 4. An example of a false SitCon positive. At the top: alignment between Zinc metalloproteinase-38 precursor (Swiss-Prot entry NAS38_CAEEL) and Cu-metallothionein 1AQS (MTCU_YEAST), with identical and similar residues shown in the middle. The top line shows 60% consensus (according to SMART) sequence of thrombospondin type 1 (TSP1) repeat, which is found in NAS38_CAEEL. In the bottom line, residues that are neighbors of the copper atoms in 1AQS are shown. Copper neighboring residues have fairly high identity/similarity with NAS38_CAEEL, but apparently it is a false hit since many key TSP1 residues are missing. Below: structures of TSP1 domain in 1LSL and metallothionein 1AQS with different folds. The cysteine sulfur atoms are displayed as spheres. Copper atoms in copper-sulfur cluster in 1aqs are not shown for clarity. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

notated in InterPro database as thrombospondin type 1 (TSP1) domain. Nevertheless, the available experimental structures of this domain are very different from the methallothionein structure: while in 1AQS nine cysteine residues are coordinated with metal ions, in TSP1 six cysteine residues form three S—S bonds [25]. While the similarity between methallothionein active site and TSP1 is interesting and could be hypothesized to have some relationship, for practical purposes it is an accidental match. Notably, one of the reasons the methallothionein score has such a large score is the fact that BLOSUM62 substitution matrix element for Cys identity is high (9). The following, second ranked SitCon hit is a true positive (Table I).

3.3. PREDICTION OF METAL-CONTAINING COFACTORS

We also explored if SitCon is capable of predicting a presence of metals or metal-containing heterogroups essential for the function or structure of the target protein. A set of 29 enzymes was used for this test. Nonenzymes were not chosen because metal ions are important for the enzymatic reaction, and enzymes often are better investigated (for example, BRENDA database [26]). All proteins homologous to a target enzyme were divided into two groups: high (*E*-score ≤ 0.01) and low homology (*E*-score >0.01). The data for the high homology proteins were used as the main criterion for deciding if a metal is essential for the target gene. The SitCon scores for metal-containing species (ions and molecules) for high and low homologs of these targets are summarized in Table III.

To somewhat simplify the analysis, we will split the 29 target genes into four groups based on the SitCon scores of metal-containing hits in the high and low homology templates: (a) both high and low homology proteins contain metallic species (12 instances); (b) high homology proteins contain metals but they are absent from the low homology proteins (5 instances); (c) low homology proteins contain metals, but not the high homology proteins (5 instances); (d) neither low nor high homology proteins contain metals (7 instances). To be able to divide results into groups (a)–(d), a score threshold for the presence of the metals had to be applied. The threshold had to be large enough to filter out cases in group (d), which are likely to have no metals, but small enough to maximize group (a). For the inves-

TABLE III

The prediction of metal-containing cofactors or metal ions by SitCon.

		High homology proteins		Low homology proteins			
Gene	E.C.	Number of metal hits	Max metal score	Number of metal hits	Max metal score	Group	Prediction or metal, and hit (+) or miss (−)
G3P4_CAEEL	1.2.1.12	0	0	4	875^b (325)	C(D)	N(+)^b
DHSA_CAEEL	1.3.5.1	39	1400	0	525	B	N(−)
NU5M_CAEEL	1.6.5.3	0	−200	20	875^b (400)	C(D)	N(+)^b
CATA2_CAEEL	1.11.1.6	151	5300	2	900	A	Y(+)
Q17514	2.1.1.98	0	—	0	500	D	N(+)
GALT5_CAEEL	2.4.1.41	5	1025	2	1650	A	Y(+)
SQV2_CAEEL	2.4.1.134	0	−200	0	400	D	N(+)
KICB2_CAEEL	2.7.1.32	0	475	0	425	D	N(+)
FLR4_CAEEL	2.7.1.37	46	1675	3	900	A	Y(+)
RPB2_CAEEL	2.7.7.6	40	1325	2	1000	A	Y(+)
O61371	3.1.1.7	0	575	0	550	D	N(+)
Q9N2M2	3.1.3.11	164	1550	16	1175	A	Y(+)
O62272	3.1.3.16	44	1875	10	1625	A	Y(+)
Q27501	3.1.3.16	45	1875	*2*	*1600*	A	Y(+)
CPR3_CAEEL	3.4.22.-	1	1450^c (250)	2	750	A(C)	Y(?)
NAS38_CAEEL	3.4.24.21	7	1250	93	1225	A	Y(+)
PSA2_CAEEL	3.4.25.1	14	1200	0	375	B	N(−)
Q9XUV0	3.4.25.1	22	1000	0	500	B	N(−)
BLM_CAEEL	3.6.1.-	3	1425	1	625	A	Y(+)
VATC_CAEEL	3.6.3.14	0	−200	0	475	D	N(+)
FZO1_CAEEL	3.6.5.-	0	−200	0	425	D	N(+)
APN1_CAEEL	4.2.99.18	9	1925	0	525	B	N(−)
O16880	4.6.1.1	36	1100	3	675	A	Y(+)
CYP1_CAEEL	5.2.1.8	2	1200	54	825	A	Y(+)
Q7K707	5.3.1.9	0	—	3	800	C	Y(+^{a})
PDI1_CAEEL	5.3.4.1	0	525	3	825	C	Y(+^{a})
SYK_CAEEL	6.1.1.6	2	725	*0*	*475*	B	N(−)
RPM1_CAEEL	6.3.2.-	*0*	—	22	1725	C	Y(+^{a})
DNLI_CAEEL	6.5.1.1	0	450	0	425	D	N(+)

Total number of SitCon metal hits with ≥600 and maximum SitCon score for metal-containing heterogroups are presented for high homology (E-score ≤ 0.01) and low homology (0.01 < E-score ≤ 100) proteins. The results for which the number of SitCon hits are less than 5 (less statistically significant) are italicized. For explanations and analysis, see text.

[a] Information about metal requirement taken from BRENDA.

[b] Heme and heme-like false hits that can be discarded (see Section 3.2).

[c] Top hit metal that is not a part of the protein.

tigated set of tests, score of 600 was the most appropriate choice. Each of the four groups is briefly analyzed below.

Group A. In this case, SitCon predicts presence of the metallic heterogroups. For the majority of targets the total number of metal hits is high which makes the conclusion statistically more reliable. However, if the hits are very few and/or they have low scores, caution should be exercised. An example is cathepsin B-like cysteine proteinase three precursor (CPR3_CAEEL). A highly homologous protease omega 1PPO contains a mercury atom (score 1450), but it is not a part of a protein (it is covalently bound to active site residue Cys 25), hence it can be ignored. For this reason, in Table III this target is annotated as A(C) in the "Group" column, where C refers to the group after a false positive is discarded. For low homology templates of CPR3_CAEEL target there are only two hits: Zn in zinc finger hit domain containing protein 2 1X4S (score 750) and a Ca ion hit (score 600) in neuraminidase 2AEP, in the first case apparently a false positive. For this particular target, SitCon metal prediction for this gene is questionable, however, for the remaining 11 cases the SitCon metal predictions in this group can be considered reliable (Table III).

Group B. This group of targets can generally be considered a failure to detect metallic species among the low homology proteins. For example, SitCon finds multiple metal hits among the high homology templates of flavoprotein subunit of succinate dehydrogenase (DHSA_CAEEL), but not in the low homology templates. Interestingly, the high homology hits belong to K, Na, Ca ions, which may be crystallization artifacts. Indeed, BRENDA does not report a requirement of metal ions for the succinate dehydrogenase flavoprotein subunit. An interesting advantage of using low homology templates versus high homology for the analysis is that neighboring residues to nonessential heterogroups will not be preserved in the low homology proteins, and hence will have low conservation scores.

For lysyl-tRNA synthetase (SYK_CAEEL) there is only one hit among the low homology templates. This suggests that this gene may have very few, if any, related proteins among the proteins selected by FASTA, and hence the SitCon prediction regarding the absence of metallic species is inconclusive. In Table III, this is reflected by italicizing the metal score for low homology templates for SYK_CAEEL.

Group C. For this group, since metals did not score high among high homology templates, additional information from external sources was necessary for verification. Targets for which very few homologous proteins exist in the PDB also generally fall into the same category. For protein disulfide isomerase-1 precursor (PDI1_CAEEL) and glucose-6-phosphate isomerase (Q7K707), BRENDA indicates metal ions being essential for the function of the protein, and SitCon does find conserved metal ion binding sites in the low homology proteins. For ubiquitin ligase protein RPM-1 (RPM1_CAEEL) there is only one homologous protein (regulator of chromosome condensation) above the 0.01 *E*-score threshold, and it does not contain metals, however, SitCon does find numerous Zn ions in the low homology templates, in agreement with BRENDA.

In G3P4_CAEEL and NU5M_CAEEL, all low homology metal hits originate from heme-like groups, which turned out to be false hits. In Subsection 3.2 we analyzed these cases and the possibility to verify these hits by analyzing conservation of crucial residues.

Group D. In this case, there is a match between the absence of the metal in the high and low homology proteins, hence we considered this to be a reliable indication of the absence of metal in the target. In abnormal acetylcholinesterase protein 2 (O61371), SitCon does not find high scoring metal heterogroups, in spite of the fact that metal ions can bind to the external binding site and are essential for their activity [27], but, of course, SitCon cannot detect that, since structures available in the PDB do not have metals in the external site.

With careful examination of the results, SitCon helps predict presence or absence of metal in about 80% of investigated cases. It should be noted that there exist tools of metal prediction available (for example, neural networks-based tools [28, 29]) with success rates around 90%. However, in many cases they use structure, not sequence as an input. Having in mind that we use a fairly smaller subset of the total Protein Data Bank to “train” our method, compared to the methods mentioned earlier, we consider this result a fair success.

4. Conclusions and Future Work Directions

The proposed method, which is based on binding site residue conservation analysis, is capable to provide hints to the nature of target gene, without a previous knowledge of domain-specific patterns. For example, we demonstrated that SitCon was able in about 80% of analyzed cases to correctly predict presence of metal-containing species in the target gene. Because of a different approach the method can be used as an alternative to pattern-based protein domain and function prediction tools. It also can be used to find not obvious similarities between the binding sites of different proteins, although such relationship probably needs to be further investigated using other approaches. SitCon in principle could be used to correct existing sequence alignments, or to help building a 3-D structure of a target protein. SitCon can also be useful in discovering relationships between a sequence and unknown function protein structure. In its simplest mode, SitCon scripts can be utilized to visualize conservation of residues in the binding sites in a series of homologous proteins.

Because of several steps existing in SitCon algorithm, there are many directions of possible development of the method. The scoring scheme, which now is based on BLOSUM62 amino acid substitution matrix, can be improved. During the initial alignment of a target sequence against the protein data base, amino acids neighboring the heterogroups or cavities in the protein structures can be given a larger weight, to give more useful list of homologous proteins. Multiple alignment regions could be used during the alignment instead of one contiguous stretch of residues. In addition to heterogroups and cavities, other protein hotspots characterized by increased residue conservation, e.g. some subunit interfaces, could be used in SitCon. To increase speed of the method, lists of neighbors for each PDB entry can be precomputed. Finally, it is relatively easy to extend the method to perform structure-to-function search. Our eventual goal is to present the proposed method to the scientific community as an easily available tool.[2]

[2] At the present stage, SitCon analysis is available upon request by an e-mail.

References

1. Doolittle, R. F. Nature 2002, 419, 493.
2. Friedberg, I. Brief Bioinform 2006, 7, 225.
3. Ofran, Y.; Punta, M.; Schneider, R.; Rost, B. Drug Discov Today 2005, 10, 1475.
4. Mulder, N. J.; Apweiler, R.; Attwood, T. K.; Bairoch, A.; Bateman, A.; Binns, D.; Bradley, P.; Bork, P.; Bucher, P.; Cerutti, L.; Copley, R.; Courcelle, E.; Das, U.; Durbin, R.; Fleischmann, W.; Gough, J.; Haft, D.; Harte, N.; Hulo, N.; Kahn, D.; Kanapin, A.; Krestyaninova, M.; Lonsdale, D.; Lopez, R.; Letunic, I.; Madera, M.; Maslen, J.; McDowall, J.; Mitchell, A.; Nikolskaya, A. N.; Orchard, S.; Pagni, M.; Ponting, C. P.; Quevillon, E.; Selengut, J.; Sigrist, C. J. A.; Silventoinen, V.; Studholme, D. J.; Vaughan, R.; Wu, C. H. Nucleic Acids Res 2005, 33, D201.
5. Quevillon, E.; Silventoinen, V.; Pillai, S.; Harte, N.; Mulder, N.; Apweiler, R.; Lopez, R. Nucleic Acids Res 2005, 33, W116.
6. Berezin, C.; Glaser, F.; Rosenberg, J.; Paz, I.; Pupko, T.; Fariselli, P.; Casadio, R.; Ben-Tal, N. Bioinformatics 2004, 20, 1322.
7. Milburn, D.; Laskowski, R. A.; Thornton, J. M. Protein Eng 1998, 11, 855.
8. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Res 2000, 28, 235.
9. Das, R.; Gerstein, M. Proteins 2004, 55, 455.
10. Kasuya, A.; Thornton, J. M. J Mol Biol 1999, 286, 1673.
11. Laskowski, R. A.; Watson, J. D.; Thornton, J. M. J Mol Biol 2005, 351, 614.
12. Panchenko, A. R.; Kondrashov, F.; Bryant, S. Protein Sci 2004, 13, 884.
13. Fetrow, J. S.; Skolnick, J. J Mol Biol 1998, 281, 949.
14. Fetrow, J. S.; Godzik, A.; Skolnick, J. J Mol Biol 1998, 282, 703.
15. Rost, B. Protein Eng 1999, 12, 85.
16. Pearson, W. R.; Lipman, D. J. Proc Natl Acad Sci USA 1988, 85, 2444.
17. Kleywegt, G. J.; Jones, T. A. Acta Crystallogr Sect D 1994, 50, 178.
18. Henikoff, S.; Henikoff, J. G. Proc Natl Acad Sci USA 1992, 89, 10915.
19. Martin, A. C. Bioinformatics 2004, 20, 986.
20. Laskowski, R. A.; Chistyakov, V. V.; Thornton, J. M. Nucleic Acids Res 2005, 33, D266.
21. Steven, R.; Kubiseski, T. J.; Zheng, H.; Kulkarni, S.; Mancillas, J.; Ruiz Morales, A.; Hogue, C. W. V.; Pawson, T.; Culotti, J. Cell 1998, 92, 785.
22. Davy, A.; Bello, P.; Thierry-Mieg, N.; Vaglio, P.; Hitti, J.; Doucette-Stamm, L.; Thierry-Mieg, D.; Reboul, J.; Boulton, S.; Walhout, A. J. M.; Coux, O.; Vidal, M. EMBO Rep 2001, 2, 821.
23. Carroll, J.; Fearnley, I. M.; Skehel, J. M.; Shannon, R. J.; Hirst, J.; Walker, J. E. J Biol Chem 2006, 281, 32724.
24. Sazanov, L. A.; Hinchliffe, P. Science 2006, 311, 1430.
25. Tan, K.; Duquette, M.; Liu, J. H.; Dong, Y.; Zhang, R.; Joachimiak, A.; Lawler, J.; Wang, J. H. J Cell Biol 2002, 159, 373.
26. Schomburg, I.; Chang, A.; Ebeling, C.; Gremse, M.; Heldt, C.; Huhn, G.; Schomburg, D. Nucleic Acids Res 2004, 32, D431.
27. Tomlinson, G.; Mutus, B.; McLennan, I. Can J Biochem 1981, 59, 728.
28. Sodhi, J. S.; Bryson, K.; McGuffin, L. J.; Ward, J. J.; Wernisch, L.; Jones, D. T. J Mol Biol 2004, 342, 307.
29. Lin, C.-T.; Lin, K.-L.; Yang, C.-H.; Chung, I.-F.; Huang, C.-D.; Yang, Y.-S. Int J Neural Syst 2005, 15, 71.