Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W411–W415***
doi:10.1093/nar/gkm264

QSCOP-BLAST—fast retrieval of quantified structural information for protein sequences of unknown structure

Stefan J. Suhrer, Markus Gruber and Manfred J. Sippl*

Center of Applied Molecular Engineering, Department of Bioinformatics, University of Salzburg, Hellbrunnerstrasse 34, 5020 Salzburg, Austria

Received January 31, 2007; Revised March 30, 2007; Accepted April 8, 2007

ABSTRACT

QSCOP is a quantitative structural classification of proteins which distinguishes itself from other classifications by two essential properties: (i) QSCOP is concurrent with the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank and (ii) QSCOP covers the widely used SCOP classification with layers of quantitative structural information. The QSCOP-BLAST web server presented here combines the BLAST sequence search engine with QSCOP to retrieve, for a given query sequence, all structural information currently available. The resulting search engine is reliable in terms of the quality of results obtained, and it is efficient in that results are displayed instantaneously. The hierarchical organization of QSCOP is used to control the redundancy and diversity of the retrieved hits with the benefit that the often cumbersome and difficult interpretation of search results is an intuitive and straightforward exercise. We demonstrate the use of QSCOP-BLAST by example. The server is accessible at http://qscop-blast.services.came.sbg.ac.at/

INTRODUCTION

The retrieval of structural information from current databases for the annotation of protein sequences with unknown structure is a fundamental challenge of structural and molecular biology. The task faces numerous problems. The available structural classifications are incomplete having a large backlog of unclassified structures and they lack clear quantitative rules that can be used to quantify and judge family membership. Many complex proteins are available only as complete chains rather than individual domains so that the scanning of hit lists and the analysis of single hits is cumbersome and time consuming. In addition, the reliability of the various sequence and structure classification schemes is difficult to judge in general, and accuracy of annotations and classifications may vary widely depending on the protein family of interest.

The protein structure classification QSCOP (1) addresses some of these problems. It endows classic SCOP [Structural Classification Of Proteins (2,3)] with quantified structural information and it is concurrent with Protein Data Bank (PDB) (4), containing all available structures in the public domain. To build QSCOP, the protein chains not contained in SCOP are cut into domains and the resulting domains are classified against the domains contained in the SCOP database. QSCOP is updated every week with the newly released PDB entries.

The intention of the QSCOP-BLAST server is to provide access to all available protein structures through a search engine which retrieves structural information for a given query sequence. Since QSCOP is organized in hierarchical layers defined by quantitative structural relationships, the redundancy and structural diversity of the result obtained is conveniently controlled by the user.

In the annotation and characterization of protein sequences of unknown structure frequent questions are (i) is there a known structure for a related protein, (ii) how many related structures are available that may serve as a model for the unknown structure of the query sequence and (iii) what is the domain structure of the query sequence and for which domains is structural information available. These and related questions are critical in many areas of protein structure research. Reliable answers are particularly important for large-scale initiatives like structural genomics projects, where the decision of whether or not a particular protein target should be channeled into the structure determination pipeline critically depends on the effective and reliable retrieval of all structural information available for that target.

*To whom correspondence should be addressed. Tel: 0043-662-8044-5796; Fax: 0043-662-8044-176; Email: sippl@came.sbg.ac.at

The QSCOP-BLAST server is specifically designed to address such questions and to make the interpretation of the retrieved results intuitive and straightforward. In the following sections, we review the components of QSCOP-BLAST and demonstrate its use by a worked out example.

METHODS

QSCOP

SCOP is one of the major protein structure classification schemes used in genome and protein research. In applications, it is generally assumed that the hierarchical organization of domains in SCOP families, superfamilies, folds and classes reflects quantitative structural relationships. This is not the case. Many SCOP families are structurally diverse containing folds that are quite dissimilar, and the extent of diversity varies strongly among the various SCOP families (1). However, for the implementation of efficient search engines and the straightforward interpretation of hit lists clearly defined quantitative relationships among protein domains are indispensable. QSCOP endows classic SCOP families with quantitative structural relationships (1) which are essential in protein structure research.

QSCOP consists of hierarchical layers reflecting decreasing structural similarity of protein domains. These layers are defined by the number of structurally equivalent residues (5) shared among two domains. The first layer covers all structures that have at least 99% equivalent residues in common. This basic layer combines identical and very similar structures of a SCOP family into a single group. The successive layers are defined by progressively smaller numbers of equivalent residues. The current version of QSCOP computes these layers in steps of 10% down to 30%. The structural diversity of a particular SCOP family is quantified by its granularity which is defined as the number of distinct groups on a given layer (1). The hierarchical organization of the classification layers obtained in this way provides a convenient data structure for the classification of new domains and for searches against the QSCOP database.

Concurrent QSCOP

The latest SCOP version 1.71, released at the end of 2006, contains 75 930 domains derived from 59 719 protein chains found in 27 600 PDB files. On the other hand, in January 2007 PDB contains over 41 200 files, where ~2000 of these files contain only non-protein chains. Hence, although recently updated, SCOP has a backlog of 13 000 files corresponding to more than a quarter of currently available PDB files.

Concurrent QSCOP contains all protein chains found in all available PDB files. The PDB files not represented in SCOP are classified against the 75 930 domains contained in the most recent SCOP release. The update yields 45 045 new domains so that the total number of domains in QSCOP is 120 975 (75 930 SCOP domains + 45 045 new domains). The QSCOP classification is updated with every new PDB release and therefore, it stays concurrent with PDB. Consequently, the QSCOP-BLAST service always matches a protein sequence against the complete volume of available knowledge on protein structures.

QSCOP-BLAST search

As the name implies QSCOP-BLAST uses the BLAST program (6,7) to search the QSCOP classification. The sequences of all QSCOP domains are extracted from the respective PDB files and the standard BLAST database files required by the BLAST engine are constructed using the BLAST suite of programs. The behavior of the BLAST program can be controlled by several parameters which affect the search results. The QSCOP-BLAST web service uses the recommended default parameters (the score matrix is BLOSUM62, gap open and extension penalties are set to 11 and 1, respectively, and the e-value cutoff is 10).

Processing of QSCOP-BLAST hits

A major problem in the interpretation of hit lists is the redundancy of protein families. Some SCOP families contain several hundred domains of varying degree of similarity and frequently subsets of families have identical or very similar sequences (8–11). On the other hand, there are proteins that have identical sequences but quite dissimilar structures. Examples are domain-swapped proteins or proteins having multiple conformations in active and inactive states. Although in such cases the sequences are identical, the corresponding structures are generally found in distinct QSCOP groups, which is a consequence of the fact that QSCOP classifies structures as opposed to sequences.

The QSCOP-BLAST engine scans a query sequence against all available protein domains, but the resulting hit list can be manipulated so that only the hits corresponding to groups on specified layers are reported. The user controls the desired granularity or redundancy of the reported hit list by choosing the appropriate layer in the QSCOP hierarchy. The advantage is 2-fold. On the one hand, the redundancy of families having a large number of members of similar sequence and structure is reduced to the desired level and on the other hand, hits that are scattered over several SCOP families, which frequently happens for sequences corresponding to multi-domain proteins, are easily recognized. In addition, proteins having similar or identical sequences but multiple conformations are easily spotted in the reduced hit lists.

WEB SERVER USAGE

Submission of queries and display of results

The QSCOP-BLAST server accepts query sequences in any format compatible with BLAST or FASTA (12). The query sequence is pasted into the sequence entry widget, and the desired QSCOP layer is chosen from a drop down menu. Submission triggers the QSCOP-BLAST engine and the resulting hit list is returned immediately. The hit list summarizes BLAST and QSCOP information on the domains found in the search, including the sequence location of domains, their SCOP classification string, alignment length and sequence identity and the BLAST e-value. BLAST alignments of query sequence and domains are displayed in the familiar BLAST format. The domain identifier used in SCOP starts with the letter 'd'. In contrast, domain names of domains which are classified in QSCOP but not in SCOP start with the letter 'c'.

Query example

The typical application of QSCOP-BLAST is the retrieval of structural information for a given protein sequence of unknown structure. In the following example we study the sequence of the α subunit of methylmalonyl-CoA-decarboxylase of *Pyrococcus furiosus*, which has been elected as a structural genomics target, code name Pfu-683389-001 of the Southeast Collaboratory of Structural Genomics. The status of this target is found to be selected and cloned.

When submitted to QSCOP-BLAST the server returns a hit list, sorted by BLAST e-values, where the first 95 domains have BLAST e-values smaller than 1.0×10^{-5}, a conservative threshold to indicate significant hits. To reduce the redundant information among the domains, we apply the 'Related' filter which removes all domains which have >75% structurally equivalent residues in common with some other entry in the hit list. The reduced list still contains four domains with e-values below 1.0×10^{-5} (Figure 1).

Note that this reduction of redundancy is not a trivial step since it requires quantitative information on the structural similarity among the domains contained in the complete hit list. All four remaining domains are classified as members of SCOP family c.14.1.4, called biotin-dependent carboxylase carboxyltransferase domain. The top hits with the most significant BLAST e-values are two domains of the A chain of 1vrg (http://dx.doi.org/10.2210/pdb1vrg/pdb), the β subunit of propionyl-CoA carboxylase of *Thermotoga maritima* at 2.30 Å resolution. Incidentally 1vrg is the structural genomics target TM0716 of the Joint Center of Structural Genomics (JCSG) with the PDB release date 22 February 2005. The chain is not classified in SCOP.

The top hit, c1vrgA2, corresponding to the C-terminal domain of the A chain matches residues 269–522 of the query sequence. The second domain, c1vrgA1, (N-terminal domain) matches the N-terminal residues 1–257 of the query. Hence, it is immediately clear, that the query consists of two domains. The respective e-values of 1.59×10^{-98} and 1.37×10^{-92} correspond to sequence identities of 71 and 66%, respectively. Hence, the hits are highly significant. A look up in the QSCOP classification shows that the two domains c1vrgA1 and c1vrgA2 have similar structures although their sequence similarity of 18% is comparatively low (Figure 2 c).

The domain ranked at position three, classified in SCOP as domain d1pixa1, matches residues 4–504 of the query sequence. The respective protein, the carboxyltransferase subunit of the bacterial ion pump glutaconyl-coenzyme A decarboxylase (*Acidaminococcus fermentans*) was solved to 2.20 Å resolution. The PDB entry release date is 5 August 2003. The respective domain is twice as long as the top ranking domains. Although the e-value of 2.55×10^{-25} is considerably higher as compared to the top hits, it may be regarded as significant and the corresponding sequence identity is 24%. On this level of sequence identity, it is likely that the query and the hit have similar structures but one has to expect a considerable variation of structural details.

The definition of the SCOP domain d1pixa1 is confusing in several aspects. First, the terminal letter, '1', of the domain name d1pixa1 corresponding to the domain number indicates that the chain contains more than one domain. However, d1pixa1 corresponds to the complete chain. Second, d1pixa1, i.e. the complete A chain of 1pix, in fact consists of two structural domains. This is rather difficult to see, and this difficulty may be the reason why the chain is not chopped into domains in SCOP, although the two domains are clearly identified in the original determination report (13). But the domain pattern is clearly recognized when the structure of the A chain of 1pix is superimposed with the QSCOP domains c1vrgA1 and c1vrgA2 (Figure 2 a and b).

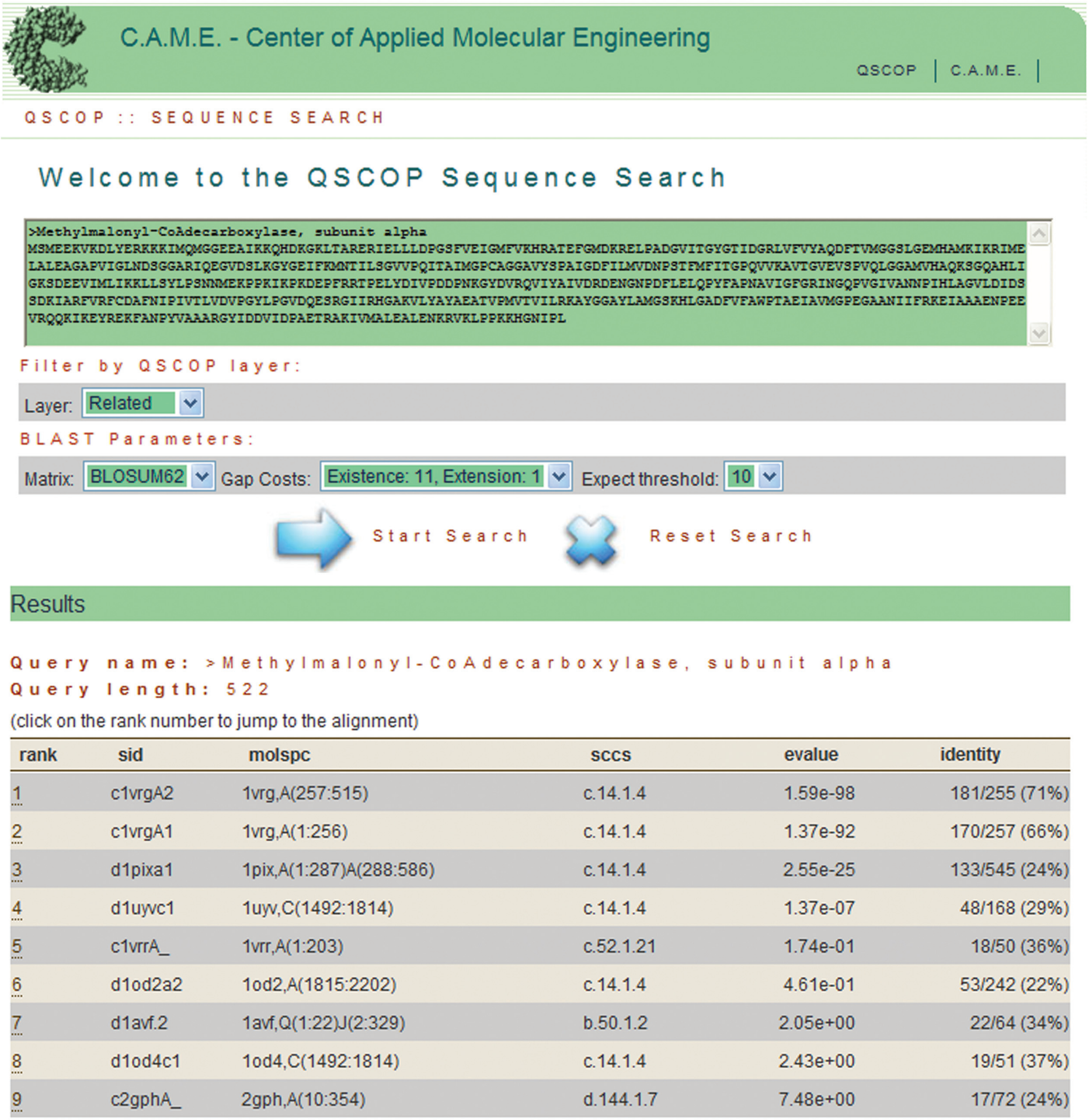
To summarize the results obtained in this example, we find that the QSCOP-BLAST clearly indicates that the query sequence consists of two structural domains that have considerable sequence and structure similarity to the β subunit of propionyl-CoA carboxylase of *Thermotoga maritima* (1vrg). Moreover, we find that the two domains are related in structure although the corresponding sequences have a low percentage of sequence identity (19%). The result indicates that the structure determination of this target most likely will reveal a fold consisting of two domains that are closely related in structure to the corresponding domains of the A chain of 1vrg.

CONCLUSION

The QSCOP-BLAST service retrieves structural information on a given target sequence reliably and fast. The amount of information contained in the hit lists returned by QSCOP-BLAST is, in fact, remarkable. Provided that BLAST is able to detect sequence similarities the entries in the hit list carry information on the domain structure, the structural similarity, and the diversity of known folds related to the query sequence. Database searches involving structure comparison and domain decomposition are in general time-consuming and require considerable computing resources. In contrast, the QSCOP search engine is efficient due to the hierarchical organization of domains in the QSCOP classification which is based on quantitative structural relationships.

ACKNOWLEDGEMENTS

The structure superposition programs ProHit/ProSup and TopMatch used to construct QSCOP and the QSCOP database used by the QSCOP-BLAST server are provided by Proceryon Science for Life GmbH (http://www.proceryon.com) under an academic license



Welcome to the QSCOP Sequence Search

```
>Methylmalonyl-CoAdecarboxylase, subunit alpha
MSMEEKVKDLYERKKKIMQMGGEEAIKKQHDKGKLTARERIELLLDPGSFVEIGMFVKHRATEFGMDKRELPADGVITGYGTIDGRLVFVYAQDFTVMGGSLGEMHAMKIKRIME
LALEAGAPVIGLNDSGGARIQEGVDSLKGYGEIFKMNTILSGVVPQITAIMGPCAGGAVYSPAIGDFILMVDNPSTFMFITGPQVVKAVTGVEVSPVQLGGAMVHAQKSGQAHLI
GKSDEEVIMLIKKLLSYLPSNNMEKPPKIKPKDEPFRRTPELYDIVPDDPNKGYDVRQVIYAIVDRDENGNPDFLELQPYFAPNAVIGFGRINGQPVGIVANNPIHLAGVLDIDS
SDKIARFVRFCDAFNIPIVTLVDVPGYLPGVDQESRGIIRHGAKVLYAYAEATVPMVTVILRKAYGGAYLAMGSKHLGADFVFAWPTAEIAVMGPEGAANIIFRKEIAAAENPEE
VRQQKIKEYREKFANPYVAAARGYIDDVIDPAETRAKIVMALEALENKRVKLPPKKHGNIPL
```

Filter by QSCOP layer:

Layer: Related

BLAST Parameters:

Matrix: BLOSUM62 Gap Costs: Existence: 11, Extension: 1 Expect threshold: 10

Results

Query name: >Methylmalonyl-CoAdecarboxylase, subunit alpha

Query length: 522

(click on the rank number to jump to the alignment)

rank	sid	molspc	sccs	evalue	identity
1	c1vrgA2	1vrg,A(257:515)	c.14.1.4	1.59e-98	181/255 (71%)
2	c1vrgA1	1vrg,A(1:256)	c.14.1.4	1.37e-92	170/257 (66%)
3	d1pixa1	1pix,A(1:287)A(288:586)	c.14.1.4	2.55e-25	133/545 (24%)
4	d1uyvc1	1uyv,C(1492:1814)	c.14.1.4	1.37e-07	48/168 (29%)
5	c1vrrA_	1vrr,A(1:203)	c.52.1.21	1.74e-01	18/50 (36%)
6	d1od2a2	1od2,A(1815:2202)	c.14.1.4	4.61e-01	53/242 (22%)
7	d1avf.2	1avf,Q(1:22)J(2:329)	b.50.1.2	2.05e+00	22/64 (34%)
8	d1od4c1	1od4,C(1492:1814)	c.14.1.4	2.43e+00	19/51 (37%)
9	c2gphA_	2gph,A(10:354)	d.144.1.7	7.48e+00	17/72 (24%)

Figure 1. QSCOP-BLAST result obtained for the structural genomics target Pfu-683389-001. The figure shows part of the web page returned by a QSCOP-BLAST search. The sequence is pasted into the widget on top of the figure. The QSCOP-BLAST server returns the respective hit list, whose redundancy in terms of structural similarity among the hits may be controlled by selecting the appropriate QSCOP layer. In addition, the BLAST alignment for individual hits may be displayed (not shown).

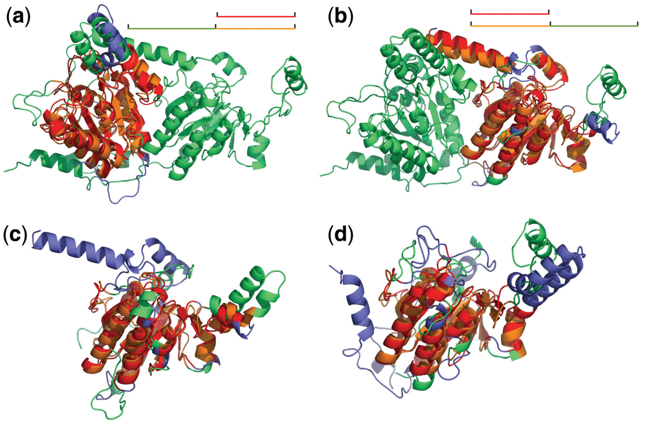


Figure 2. Superposition of various structures found in the hit list shown in Figure 1. For any pair of superimposed structures, the first structure is shown in blue and the second in green. In regions where the structures are equivalent the first structure is shown in red and the second structure in orange. (**a**) d1pixa1 (green/orange) superimposed on c1vrgA2 (blue/red). The structures share 201 residues which occupy equivalent positions in the two structures (red and orange). The C^{α} atoms of these residues superimpose to an root mean square (rms) error of 1.4 Å. The sequence identity in this region is 27%. (**b**) d1pixa1 superimposed on c1vrgA1 (211 equivalent residues, 1.6 Å rms, 23% sequence identity), (**c**) c1vrgA1 superimposed on c1vrgA2 (181 equivalent residues, 1.9 Å rms, 18% sequence identity) (**d**) the structural domains 61–285 and 321–558 of d1pixa1 (168 equivalent residues, 2.3 Å rms, 15% sequence identity).

agreement which is gratefully acknowledged. All superposition figures were prepared using PyMol (http://www.pymol.org). This work was supported by FWF Austria, grant number P13710-MOB. Funding to pay the Open Access publication charges for this article was provided by the University of Salzburg.

Conflict of interest statement. None declared.

REFERENCES

1. Suhrer,S.J., Wiederstein,M. and Sippl,M.J. (2006) QSCOP – SCOP quantified by structural relationships. *Bioinformatics*., **23**, 513–514.
2. Andreeva,A. and Murzin,A.G. (2006) Evolution of protein fold in the presence of functional constraints. *Curr. Opin. Struct. Biol.*, **16**, 399–408.
3. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
4. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
5. Feng,Z.K. and Sippl,M.J. (1996) Optimum superimposition of protein structures: ambiguities and implications. *Fold Des.*, **1**, 123–132.
6. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
7. Schäffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of psi-blast protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
8. Xie,L. and Bourne,P.E. (2005) Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. *PLoS Comput. Biol.*, **1**, e31.
9. Peng,K., Obradovic,Z. and Vucetic,S. (2004) Exploring bias in the protein data bank using contrast classifiers. *Pac. Symp. Biocomput.*, **9**, 435–446.
10. Liu,J. and Rost,B. (2002) Target space for structural genomics revisited. *Bioinformatics*, **18**, 922–933.
11. Brenner,S.E., Chothia,C. and Hubbard,T.J. (1997) Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol.*, **7**, 369–376.
12. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA.*, **85**, 2444–2448.
13. Wendt,K.S., Schall,I., Huber,R., Buckel,W. and Jacob,U. (2003) Crystal structure of the carboxyltransferase subunit of the bacterial sodium ion pump glutaconyl-coenzyme a decarboxylase. *EMBO J.*, **22**, 3493–3502.