Effective detection of remote homologues by searching in sequence dataset of a protein domain fold

S. Sandhya[a,b,1], S. Kishore[b,1,2], R. Sowdhamini[b], N. Srinivasan[a,*]

[a] *Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India*
[b] *National Center for Biological Sciences, UAS, GKVK Campus, GKVK Post, P.B. No. 6501, Bangalore 560 065, India*

Accepted 14 August 2003

First published online 27 August 2003

Edited by Robert B. Russell

Abstract Profile matching methods are commonly used in searches in protein sequence databases to detect evolutionary relationships. We describe here a sensitive protocol, which detects remote similarities by searching in a specialized database of sequences belonging to a fold. We have assessed this protocol by exploring the relationships we detect among sequences known to belong to specific folds. We find that searches within sequences adopting a fold are more effective in detecting remote similarities and evolutionary connections than searches in a database of all sequences. We also discuss the implications of using this strategy to link sequence and structure space.

Key words: Evolution; Profile-based method; Remote homology; Superfamily

1. Introduction

The determination of a large number of protein structures in the last three decades reveals that protein sequences with no obvious sequence similarity can adopt the same fold. Whether such sequentially dissimilar but structurally similar proteins have evolved from a common ancestor is not a trivial problem to address. In the absence of sequence similarity, functional similarity may suggest a common ancestor for proteins adopting the same fold. A common ancestry is harder to discern for structurally similar but sequentially and functionally dissimilar proteins. The organization of the protein structure classification database, Structural Classification of Proteins (SCOP), [1] is based on this understanding of protein structure and evolution. Proteins with clear sequence similarity are classified as belonging to the same family, those with structural and functional similarity but lacking significant sequence similarity are classified under the same superfamily, while structurally similar but otherwise dissimilar proteins are classified under different superfamilies of the same fold.

Genome sequencing initiatives have phenomenally increased the size and coverage of protein sequence databases. Viewing protein sequence space as continuous rather than discrete implies that a number of uncharacterized protein sequence families in the sequence database may provide evolutionary connections between previously unrelated protein structure superfamilies [2,3]. Recent work from different groups [4–6] attempts to connect different families and superfamilies by combining sensitive sequence-based search methods such as PSI-BLAST [7] and information from annotated sequence and structure databases such as PFAM [8] and SCOP [9], respectively.

We describe here a protocol to detect distant relationships within a fold based on searches using PSI-BLAST against a fold library enriched with sequences known to adopt the same fold. This approach has been validated for a number of folds. We first generate a dataset of sequences likely to adopt a particular fold, say, the TIM barrel fold. This is performed by searching the non-redundant database (NRDB) using PSI-BLAST with representative sequences from each SCOP TIM barrel family as queries. Distinct hits from all PSI-BLAST runs are pooled together to form an enriched database of TIM barrel sequences. We then repeat a PSI-BLAST search against this enriched database using the same representative sequences, from the families within TIM fold in SCOP used to form the enriched database, and analyze connections made at the superfamily and fold level. We also validate the approach by assessing it on other folds from the α, β, α/β and α+β classes. We discuss the implications of using this approach to link sequence and structure space in proteins, its application in functional annotation of hypothetical proteins and in fold recognition.

The main difference between the proposed approach and approaches such as those of Chothia and coworkers [10] and Sternberg and coworkers [11] is that the sequence-enriched structure databases used in their work culled sequences known to adopt different folds into a single dataset. Our protocol suggests higher sensitivity of homology detection if the database is composed of sequences proposed to adopt a given fold. We propose that higher sensitivity can be achieved in PSI-BLAST-based fold recognition methods if the search for a query is made against one database at a time with each database corresponding to a sequence-enriched database of proteins adopting a specific fold integrated with sequences of homologues of the query.

2. Materials and methods

The protocol and assessment are described below in Sections 2.1–2.5 for the TIM fold. The same approach has been taken for several other folds as well.

*Corresponding author. Fax: (91)-80-360 0535.
E-mail address: ns@mbu.iisc.ernet.in (N. Srinivasan).

[1] These authors contributed equally to this work.
[2] Present address: Department of Neurobiology and Behavior, State University of New York, Stony Brook, NY 11794-5230, USA.

0014-5793/03/$22.00 © 2003 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(03)00929-3

2.1. Building the fold library of sequences

One representative from each of the 55 families in SCOP (release 1.57) that adopt the TIM barrel fold is chosen as query. PSI-BLAST searches [7] were initiated against the non-redundant protein database at NCBI (version 2.2.4 (August 2002), 841 211 sequences) for each of the 55 query sequences from as many families within the TIM fold. PSI-BLAST uses an iterative profile-based procedure and constructs position-dependent weight matrices on the basis of alignments generated by a BLAST database search. Sequences whose expectation (E) value is more significant than a given inclusion threshold are used for the construction of a weight matrix. These weight matrices are then used to score the next iteration of database searching until no more sequences satisfying the inclusion thresholds are detected and can be included into the profile. PSI-BLAST searches were performed for 20 iterations with an inclusion threshold of $E < 0.0001$. For most queries PSI-BLAST reached convergence before the 20th round. In case convergence is not reached the results of the various rounds of PSI-BLAST were manually analyzed to pick up the homologues of the query at the largest round with query remaining as one of the top ranking hits. We have also ensured that, in general, the profile does not drift from the query by ensuring that the query is always present in the iteration considered for analysis. In order to minimize the chances of picking up false positives, hits with the number of residues shorter than 75% of the query length were not included in the TIM fold dataset. The 55 PSI-BLAST output files were, further, manually scrutinized for any obvious false positives, based on the features such as annotation and extent of low complexity region involved in the alignment.

Homologues thus identified for each of the 55 family representatives against NRDB were pooled together to form a sequence-enriched structure database or a TIM fold library. To avoid including unrelated extra domains in the hits, only the stretch of alignments reported in PSI-BLAST was added to the database. Sequences that were more than 90% identical to the query or to any other hit were removed to reduce the redundancy in the database. To aid subsequent analysis, the hits that were pooled into the fold library were annotated with SCOP superfamily and family codes of the query as obtained from SCOP parseable files.

2.2. Search within the sequence-enriched fold library

PSI-BLAST runs were reinitiated against the sequence-enriched TIM fold library using each of the 55 representative sequences as a query. The same inclusion threshold of 0.0001 as in NRDB searches was used to identify the homologues. A comparison of the superfamily and family codes of the hits with those of the query allowed the classification of the results as belonging to the same family, superfamily, or a different superfamily with respect to the query.

2.3. Search within sequence-enriched TIM fold library with non-TIM sequences as queries

PSI-BLAST runs were also initiated against the sequence-enriched TIM fold library using 430 sequences known not to adopt the TIM fold. The results were analyzed using the same criteria and filters as in the previous analysis. These runs were performed to serve as a control to check if searching in a specialized database of sequences of the TIM fold randomly assigned statistically significant scores to false positives.

2.4. Assessing the sensitivity of the profiles

We checked if the profiles generated in searches within the enriched TIM dataset in the final connecting steps are sensitive and do not assign high E-values to any protein irrespective of the fold. For this purpose we aligned the profiles obtained by searching in the fold library of sequences with all proteins from the 1.57 release of the SCOP domain database. The SCOP domain database of 34 314 sequences was obtained from the ASTRAL compendium [12] for sequence and structural analysis and contains sequences from all the folds in SCOP. In addition we performed a jack-knife test by assessing the sensitivity of detection of remote similarities between the profiles corresponding to the 55 families in the TIM fold dataset and 15 new families within the TIM fold which have appeared in an updated version of SCOP (1.63 release).

2.5. Assessing the role of intermediate sequences in the connections

We also examined if the connections obtained in the final connecting step are due to common intermediates detected in the search against the NRDB by comparing the homologues identified in the NRDB search with each of the 'new' connections made in searches within the enriched dataset. The presence of such common homologues would obviate the need for a search within the fold library.

2.6. Performance of other folds

In order to validate the approach we also performed a similar assessment for seven other folds such as the globin-like, ATC-like, THDP (thiamin diphosphate binding fold-like), ribonuclease H-like motif fold, flavodoxin-like, cystatin-like, and reductase (isomerase/elongation common factor domain) fold and assessed the connections obtained by searching within each of the respective fold libraries.

3. Results

Much of the discussion below pertains to the results of search within the dataset of sequences belonging to the TIM fold. A brief account of the results for other folds is covered in a subsequent section.

The enriched TIM fold library (see Section 2) of 55 families contains 8340 sequences. A few connections between superfamilies are established while searching the NRDB even prior to searching against the enriched database. With the exception of enolase, all the superfamilies connected at this stage share a common phosphate binding motif at the end of the seventh β-strand. The homology between these phosphate binding barrels has been investigated in detail [4,13] and clearly suggests a common ancestry for these phosphate binding superfamilies. We have also detected all the relationships across the superfamilies detected by Copley and Bork [4] in our search against NRDB (data not shown).

3.1. Inferring homology from iterative searches against a sequence-enriched database

The enrichment process allows us to map sequences with unknown structure to families with known structure (see Section 2). A sequence mapped to a certain structural family A serves as a link between two families A and B, if a search against the enriched database using a sequence representing the family B picks up the sequence mapped to family A. The two families could belong to the same superfamily or belong to two different superfamilies. Using this strategy, we find a number of connections at significant E-values between and within TIM barrel superfamilies (Table 1). However, searches against a small database with relatively relaxed E-values might make connections that may not have evolutionary bearing [14]. The construction of the enriched fold library eliminates the possibility of obvious false positives, and any connection obtained by searching the enriched database is a true positive. We analyze the structure and function of proteins connected by searching against this enriched database to justify the connections made by PSI-BLAST.

3.2. Detecting relationships across families

Proteins that belong to different families within the same superfamily lack significant similarity, and the basis of their classification in the superfamily is a similarity in function. PSI-BLAST is generally accepted as a tool sensitive enough to detect relationships between proteins belonging to the same superfamily [11,15,16]. Twelve of the 24 TIM barrel superfamilies are multi-member superfamilies.

In a number of multi-member superfamilies, searching within the enriched fold library establishes connections between families within the superfamily (Table 1). There are at least

Table 1
Detection of distant relationships using searches within the fold library of sequences

S. No. and fold name	Query	Family name	Superfamily	No. of other families in superfamily	Round	Hit	Family name of hit	*E*-value
TIM								
1a	1dxi	Xylose isomerase	Xylose isomerase-like	4	12	1d8w	L-Rhamnose isomerase	10^{-94}
						1qtw	Endonuclease IV	10^{-60}
1b	1d8w	L-Rhamnose isomerase	Xylose isomerase-like	4	15	1qtw	Endonuclease IV	10^{-50}
2a	1a49	Pyruvate kinase	Phosphoenolpyruvate/pyruvate domain	5	12	1dik	Pyruvate phosphate dikinase, C-terminal domain	10^{-30}
						1dxe	2-Dehydro-3-deoxy-galactarate aldolase	10^{-87}
2b	1dxe	2-Dehydro-3-deoxy-galactarate aldolase	Phosphoenolpyruvate/pyruvate domain	5	6	1dik	Pyruvate phosphate dikinase, C-terminal domain	10^{-45}
3	1b57	Class II aldolase	Aldolase	3	2	1qo2	Histidine biosynthesis enzymes	10^{-51}
4	1ct5	Hypothetical protein ybl036c	PLP binding barrel	1	2	1bd0	Alanine racemase-like, N-terminal domain	10^{-75}
5	1dbt	Decarboxylase	Ribulose-phosphate binding barrel	3	15	1qo2	Histidine biosynthesis enzymes	10^{-16}
6a	1j79	Dihydroorotase	Metallo-dependent hydrolases	7	16	2ubp	α subunit of urease, catalytic domain	10^{-61}
						1pta	Phosphotriesterase-like	10^{-26}
						1a4m	Adenosine deaminase	10^{-57}
6b	1pta	Phosphotriesterase-like	Metallo-dependent hydrolases	7	15	1j79	Dihydroorotase	10^{-46}
						1a4m	Adenosine deaminase	10^{-48}
Reductase								
7	1qfz	Ferredoxin reductase FAD binding domain-like	Riboflavin synthase domain-like	2	15	1ddg	NADPH cytochrome P450 reductase FAD binding domain-like	$2e^{-27}$
Ribonuclease H-like motif								
8	1hux	Hydroxyglutaryl-CoA dehydratase component A	Actin-like ATPase domain	5	15	1e4f	Actin/HSP70 (cell division protein)	$2e^{-42}$
						1bu6	Glycerol kinase	$1e^{-47}$
Flavodoxin-like								
9	1dbw	CheY-related (transcriptional regulatory protein)	CheY-like	4	5	1a9x	Class I glutamine amidotransferase (carbamoyl phosphate synthetase, small subunit C-ter domain)	$3e^{-05}$
Globin-like								
10	1h7w	Dihydropyrimidine dehydrogenase, N terminal domain	α helical ferredoxin	1	3	1qla	Fumarate reductase/succinate dehydrogenase Fe/S protein C-ter domain	$2e^{-31}$

Relationships detected between families in searches within the fold library but not made in searches against the NRDB.

six multi-member superfamilies where relationships not detected between families in searches against the NRDB are detected by searching in the database of sequences adopting TIM fold (Table 1).

An interesting connection we detect through this sequence search approach is between α subunit of urease, which is the catalytic domain of urease and adenosine deaminase. The pairwise identity between these sequences is only 20%. The relationship between the two is not identified even in relaxed *E*-value cut-off (0.0001) searches against NRDB. However, a search within a database of known relationships like the TIM fold library allows such detection at statistically significant *E*-values (10^{-173}). The availability of structural information of both domains confirms a common evolutionary origin between the two sequences, which are grouped under the metallo-dependent hydrolase superfamily [9]. Functional similarities between the two include features such as the requirement for a binuclear metal center and also a water-ion-mediated nucleophilic attack of the substrate [17]. Another such connection in the same superfamily is between phosphotriesterase and dihydroorotase. This connection is characterized by the same functional similarities as seen above. Of interest is the fact that dihydroorotase was predicted to have a TIM barrel fold only after a detailed sequence and structure analysis [18].

PSI-BLAST searches, against the database of TIM fold adopting sequences, using xylose isomerase as the query detects relationships with rhamnose isomerase and endonuclease IV, other families of the same superfamily. PSI-BLAST searches against NRDB, however, fail to detect these relationships. While all three proteins require zinc for catalysis, the functional relationship between xylose isomerase and rhamnose isomerase is more explicit than the functional relationship of endonuclease to the other two enzymes [19,20]. The sequence identity between rhamnose isomerase and xylose isomerase is only 13%, and the functional similarity between the two proteins became evident only after the structure of rhamnose isomerase was determined [19].

Thus, for a number of cases (Table 1), searching against the fold library enables detection of relationships between members of a superfamily not observed in relaxed searches against NRDB. These relationships could have been picked up even before the structure and function of these proteins became available if the sequence search had been performed in a specialized dataset of sequences adopting the TIM fold integrated with the homologues of the query. The observation that PSI-BLAST is unable, in some cases (Table 1), to detect relationships within superfamilies suggests that some relationships at the superfamily level still require human knowledge and expertise subsequent to structure determination.

3.3. Searching the TIM fold library of sequences with sequences adopting other folds

Of the 430 'non-TIM fold' sequences used to query the database, 39 queries belonged to all α class of proteins, 55 to all β class of proteins, 177 to the α/β class of proteins and 159 to the α+β class of proteins in SCOP. We used more controls from the α+β and α/β class of proteins primarily because there may be false positives appearing due to shared α/β motifs. Interestingly, searches initiated within the TIM fold library did not identify any hits for such sequences. Thus, the control experiment worked extremely well with no false positive identified as a hit in 430 PSI-BLAST searches. A problem sometimes faced in using sensitive procedures such as PSI-BLAST is the possibility of false positives qualifying the inclusion threshold and influencing the profile in subsequent iterations of database searching. The generation of such fold libraries comprising solely of all sequences likely to adopt a fold may therefore minimize the problems of profile drift when using a profile-based method such as PSI-BLAST.

3.4. Assessing the sensitivity of the profiles

We have aligned the profiles we obtain in the connecting steps (obtained when searching the TIM fold dataset) with every sequence available in ASTRAL, which has a collection of sequences of all protein domains available in SCOP. None of the profiles we generated in the connecting step in searches within the TIM fold dataset matched with a non-TIM sequence with a significant *E*-value. All the homologues identified were sequences belonging to the TIM fold. Of particular interest was the fact that the connections we make by performing searches within the enriched TIM dataset are not all made by performing a single iteration of PSI-BLAST against the ASTRAL sequences. This suggests that more iterations of PSI-BLAST that enable and allow profile drift within the fold library are required to detect these connections. For instance, the relationship between Class II aldolase and histidine biosynthesis enzymes is not made in a single iteration against the database of all SCOP domains. This relationship, however, is easily established when searches are made within the enriched TIM dataset. Clearly, a database specific for a fold can direct the searches within the fold and enable the detection of such connections.

We have also performed a jack-knife test for the detection of TIM fold. Subsequent to the completion of much of the work we found 15 new families within the TIM fold in an updated version of SCOP to result in a total of 70 families. Homologues of the 15 new families, detected from NRDB, result in a database of TIM fold sequences which is independent of the original set corresponding to 55 families and bears no clear similarity to the original dataset. We have aligned the profiles we obtained in the connecting steps (obtained when searching the original 55-family TIM fold dataset) with the sequences present in the dataset corresponding to 15 new families. Nineteen connections between proteins of known structure in the new dataset of 15 families have been made with some of the profiles of original 55 families with the *E*-values ranging between 7×10^{-61} and 10^{-7}. These connections included the relationships between the families of alkane sulfonate monoxygenase and non-fluorescent flavoprotein, both belonging to the superfamily of bacterial luciferase-like, cytosine deaminase catalytic domain and dihydroorotase, both belonging to the superfamily of metallo-dependent hydrolase. Interestingly some of the connections made are across the superfamilies within the TIM fold. For example, cytosine deaminase catalytic domain could also be related to endonuclease IV (*E*-value: 8×10^{-13}), which belongs to the different xylose isomerase-like superfamily. Thus it is very clear that two different datasets of sequences of TIM fold generated independent of each other could result in identification of connections, with significant *E*-values, across the datasets.

3.5. Searches within the fold library identifies more intermediates

The profiles generated by searching within the enriched da-

taset of 55 families are clearly very specific for each family in the TIM fold. Our contention that pooling together the sequences generated in searches against the NRDB performs better in relating families than individual profiles made in the NRDB searches is reaffirmed on examining the homologues detected for every family in the first set of searches against the NRDB. Whilst common homologues are obtained between many of the families that we relate finally (Table 1) while searching the NRDB itself, no direct relationship to a structural family is made in NRDB search. It is interesting that in at least five of the pairs of structural families (Table 1) there are no common hits between the homologues of either family in the searches against the NRDB. Further these queries did not match significantly with the profiles of the individual families composed solely of homologues of each family picked up from NRDB. This clearly suggests that it is important to first pool together the homologues identified by searching against the NRDB for every query representing the family and then to search within such an enriched dataset that is a 'sum of truly TIM relatives'.

3.6. Assessment with other folds

As seen in Table 1 our approach shows positive results and detects relationships in the other folds for which the assessment has been made. For example, we make additional connections in the flavodoxin-like, reductase, ribonuclease H-like motif and globin-like folds. We detect the relationship between ferredoxin reductase FAD binding domain-like family and NADPH cytochrome P450 reductase FAD binding domain-like family. SCOP already classifies these two families in the same superfamily suggesting similarities in structural features. Yet, in searches against the NRDB, these relationships are not detectable. PSI-BLAST searches using hydroxyglutaryl-CoA dehydratase component A as a query detects actin/HSP70 and glycerol kinase as homologues. All three sequences are members of the same superfamily and it is known that they share a common mechanism of ATP hydrolysis and also display similar conformational changes following hydrolysis [21]. It is also known that these proteins share good structural similarity with hydroxyglutaryl-CoA dehydratase component A appearing to be a core version for this superfamily.

The ability to detect relationships between the families and superfamilies in a fold seems to depend on the diversity of the fold. Well-populated folds with many families and superfamilies fair better in such approaches than conservative folds with poor representation. The success of this approach relies on the availability of large evolutionary distances between the families and the availability of large numbers of intermediates between the families. In the course of time, with the availability of more genome data, we expect that such problems will be overcome.

4. Discussion

The post-genomic era and the rapid pace of protein structure determination have resulted in the development of a number of new methods that aid functional annotation of uncharacterized sequences. The current paradigm for assigning function to uncharacterized protein sequences is to query the NRDB using PSI-BLAST as a first step, and in the absence of a clear result, the next step is resorting to fold recognition for a less direct inference of function. While fold recognition provides useful information about the overall structure of the protein, it is often difficult to assign any function to that protein. In the approach discussed in this paper, we attempt to bridge the gap between PSI-BLAST and fold recognition for functional annotation of proteins, by demonstrating that functional similarities that were originally inferred only on the basis of structural information can in fact be made with significantly less human intervention.

The idea of identifying intermediate sequences to link distant protein structure superfamilies is not new [2,3]. However, this work differs from earlier approaches in systematically mapping the results from sequence searches against the NRDB to SCOP classification, thus facilitating an easy and reliable way, with minimal manual intervention, to find relations at different levels in the protein sequence space. Our approach differs fundamentally from other implementations of fold detection using intermediate sequences [22–26]. There is a key difference between these methods and our method in terms of the implementation of the basic principle. We suggest forming a number of datasets of sequences each corresponding to a fold and a query sequence is searched in each of these sequence datasets after integrating each sequence dataset with homologues of the query. We are not aware of any other method in which such an approach is employed.

As a given dataset is confined to sequences known to belong to a fold and homologues of the query the chances of drift of the profile in PSI-BLAST searches is minimized. Consequently, the elimination of false positives is more effective in our approach compared to other approaches. As seen from Table 1, our approach enables picking up new connections that are not obtained when a search is made in the consolidated dataset of sequences (i.e. in procedures such as PDB-ISL).

Most of the connections we detect are from within the superfamily. The inability of the current approach to detect many connections across the superfamilies does not conclusively establish that these superfamilies are evolutionarily unrelated. The growth of sequence databases coupled with a more exhaustive search procedure, which is an extension of the current approach, could possibly yield a greater number of connections between superfamilies within a fold. Using all characterized members of a superfamily in place of one representative sequence for every family, in turn, might enable better representation of the sequence space of a fold and connect more superfamilies within the fold when using a fold space-based strategy such as the one described in this paper. It is interesting to note that most of the superfamilies that remain unconnected in the current work have fairly specialized functions and, as such, do not appear to be involved in central metabolism. As far as proteins adopting TIM fold are concerned, given the abundance of genes encoding for TIM barrel proteins in all organisms, and the adaptability of its active site, the evolution of a number of apparently unrelated superfamilies could be the result of a lineage specific gene expansion due to gene duplication and subsequent divergence [27,28].

A sequence-enriched structure database, with a clear mapping between sequence and structure, for querying has wider applications beyond what is described in this paper. All genomes have a large number of hypothetical sequences, and using these as queries against enriched databases of protein superfamilies might provide some clues about the putative

functions of these proteins. Searching within a database of individual fold libraries, by supplementing sequences known to adopt a fold with additional sequences derived from protein databases, may be used in fold recognition. In order to implement this idea effectively one could integrate the sequences of the homologues of the query (obtained from the NRDB search) with a specialized fold dataset of sequences and a further search for the query could be made in the integrated dataset of sequences. Such a search could be done for datasets of sequences of every known fold. The effectiveness of this protocol in detecting similarities within superfamilies indicates that similarities between proteins can be effectively established by querying a library of protein folds instead of a comprehensive database such as the NRDB. Finally, this approach could be applied in a systematic manner to build robust phylogenies for protein sequence and structure families.

Acknowledgements: This work was supported by the award of Senior Research Fellowships in biomedical sciences, by the Wellcome Trust, London, to R.S. and N.S. as well as by the Wellcome Trust supported Computer Farm project jointly between R.S. and N.S. The Wellcome Trust, London, supported S.S. and S.K. N.S. also thanks the Department of Biotechnology, New Delhi, for the support in the computational genomics initiative. We thank the anonymous referees for their very valuable comments and suggestions regarding the validation and assessment of the approach.

References

[1] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) J. Mol. Biol. 247, 536–540.
[2] Park, J., Teichmann, S.A., Hubbard, T. and Chothia, C. (1997) J. Mol. Biol. 273, 349–354.
[3] Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. and Chothia, C. (1998) J. Mol. Biol. 284, 1201–1210.
[4] Copley, R.R. and Bork, P.P. (2000) J. Mol. Biol. 303, 627–641.
[5] Pandit, S.B., Gosar, D., Abhiman, S., Sujatha, S., Dixit, S.S., Mhatre, N.S., Sowdhamini, R. and Srinivasan, N. (2002) Nucleic Acids Res. 30, 289–293.
[6] Aloy, P., Oliva, B., Querol, E., Aviles, F.X. and Russell, R.B. (2002) Protein Sci. 11, 1101–1116.
[7] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997) Nucleic Acids Res. 25, 3389–3402.
[8] Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A. and Durbin, R. (1998) Nucleic Acids Res. 26, 320–322.
[9] Lo Conte, L., Brenner, S.E., Hubbard, T.J.P., Chothia, C. and Murzin, A. (2002) Nucleic Acids Res. 30, 264–267.
[10] Lesk, A.M., Branden, C.I. and Chothia, C. (1989) Proteins 5, 139–148.
[11] Muller, A., MacCallum, R.M. and Sternberg, M.J. (1999) J. Mol. Biol. 293, 1257–1271.
[12] Chandonia, J.M., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M. and Brenner, S.E. (2002) Nucleic Acids Res. 30, 260–263.
[13] Bork, P., Gellerich, J., Groth, H., Hooft, R. and Martin, F. (1995) Protein Sci. 4, 268–274.
[14] Jones, D.T. and Swindells, M.B. (2002) Trends Biochem. Sci. 27, 161–164.
[15] Aravind, L. and Koonin, E.V. (1999) J. Mol. Biol. 287, 1023–1040.
[16] Lindahl, E. and Elofsson, A. (2000) J. Mol. Biol. 295, 613–625.
[17] Benini, S., Rypniewski, W.R., Wilson, K.S., Miletti, S., Ciurli, S. and Mangani, S. (1999) Struct. Fold. Des. 7, 205–216.
[18] Holm, L. and Sander, C. (1997) Proteins 28, 72–82.
[19] Kornörfer, I.P., Fessner, W. and Matthews, B.W. (2000) J. Mol. Biol. 300, 917–933.
[20] Hosfield, D.J., Guan, Y., Haas, B.J., Cunningham, R.P. and Tainer, J.A. (1999) Cell 98, 397–408.
[21] Locher, K.P., Haans, M., Yeh, A.P., Buckel, W. and Rees, D.C. (2001) J. Mol. Biol. 307, 297–308.
[22] Teichmann, S.A., Park, J. and Chothia, C. (1998) Proc. Natl. Acad. Sci. USA 95, 14658–14663.
[23] Teichmann, S.A., Chothia, C., Church, G.M. and Park, J. (2000) Bioinformatics 16, 117–124.
[24] Koretke, K.K., Russell, R.B. and Lupas, A.N. (2002) Protein Sci. 11, 1575–1579.
[25] Koretke, K.K., Russell, R.B. and Lupas, A.N. (2001) Proteins Suppl. 5, 68–75.
[26] Koretke, K.K., Russell, R.B., Copley, R.R. and Lupas, A.N. (1999) Proteins Suppl. 3, 141–148.
[27] Jordan, I.K., Makarova, K.S., Spouge, J.L., Wolf, Y.I. and Koonin, E.V. (2001) Genome Res. 11, 555–565.
[28] Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I. and Koonin, E.V. (2002) Genome Biol. 3, research0008.1-0008.9.