BIOINFORMATICS

Vol. 16 no. 2 2000
Pages 117–124

Fast assignment of protein structures to sequences using the Intermediate Sequence Library PDB-ISL

Sarah A. Teichmann[1], Cyrus Chothia[1], George M. Church[2] and Jong Park[2]

[1]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK and [2]Department of Genetics, Harvard Medical School, Warren Alpert Building, 200 Longwood Avenue, Boston, MA 02115, USA

Received on July 20, 1999; revised on September 17, 1999; accepted on September 23, 1999

Abstract

Motivation: *For large-scale structural assignment to sequences, as in computational structural genomics, a fast yet sensitive sequence search procedure is essential. A new approach using intermediate sequences was tested as a shortcut to iterative multiple sequence search methods such as PSI-BLAST.*

Results: *A library containing potential intermediate sequences for proteins of known structure (PDB-ISL) was constructed. The sequences in the library were collected from a large sequence database using the sequences of the domains of proteins of known structure as the query sequences and the program PSI-BLAST. Sequences of proteins of unknown structure can be matched to distantly related proteins of known structure by using pairwise sequence comparison methods to find homologues in PDB-ISL. Searches of PDB-ISL were calibrated, and the number of correct matches found at a given error rate was the same as that found by PSI-BLAST. The advantage of this library is that it uses pairwise sequence comparison methods, such as FASTA or BLAST2, and can, therefore, be searched easily and, in many cases, much more quickly than an iterative multiple sequence comparison method. The procedure is roughly 20 times faster than PSI-BLAST for small genomes and several hundred times for large genomes.*

Availability: *Sequences can be submitted to the PDB-ISL servers at http://stash.mrc-lmb.cam.ac.uk/PDB_ISL/ or http://cyrah.ebi.ac.uk:1111/Serv/PDB_ISL/ and can be downloaded from ftp://stash.mrc-lmb.cam.ac.uk/pub/PDB_ISL/ or ftp://ftp.ebi.ac.uk/pub/contrib/jong/PDB_ISL/*

Contact: *sat@mrc-lmb.cam.ac.uk and jong@ebi.ac.uk*

Introduction

The fastest and easiest way of matching two sequences is by pairwise sequence comparison methods. However, protein sequences can diverge to such an extent that these types of comparisons fail to detect evolutionary relationships. For related proteins that have sequence identities of 20–30%, only one-half of the relationships can be detected by pairwise sequence comparisons, and for related proteins with lower identities, the proportion is much smaller (Brenner *et al.*, 1998). The limitation of pairwise sequence comparison methods can be overcome, at least in part, by using multiple sequences to build family-specific profiles as in hidden Markov models (Krogh *et al.*, 1994; Eddy, 1996) and PSI-BLAST (Altschul *et al.*, 1997). These models can be built in two different ways. One iterates searches over a big sequence database to utilise intermediate homologues of the query sequence to find more distant ones (Tatusov *et al.*, 1994). The other builds models from sets of prealigned sequences (Bateman *et al.*, 1999). Both the model and the non-model-based algorithms can benefit from iteration. Examples of iterative sequence search procedures are PSI-BLAST (Altschul *et al.*, 1997) and the SAM-T98 iterative HMMs (Karplus *et al.*, 1998). From an assessment of these methods (Park *et al.*, 1998) it is evident that PSI-BLAST and SAM-T98 do about three times better than pairwise methods at detecting relationships between homologous sequences whose identities are less than 30%. The disadvantage of these methods, especially with model-based ones, is that they can be slow (due to the large number of matches that are made for each query, especially with large sequence families) to an extent that makes large-scale searches prohibitive.

Intermediate sequence search (ISS) procedures for proteins of known structure

Here we describe a procedure that circumvents the speed problem associated with iterative methods when matching genome sequences to proteins of known structure. From

a large sequence database, PSI-BLAST first collects homologues of the sequences of proteins of known structure. We call this collection the PDB-intermediate sequence library (PDB-ISL). Matching genome sequences to PDB-ISL sequences by implication matches them to proteins of known structure (Park *et al.*, 1998). We then show that by using FASTA (Pearson and Lipman, 1988) to carry out these matches, the same number of hits as PSI-BLAST are produced, and it is much faster for large-scale searches.

Knowing the fold of a sequence can be desirable for several reasons, the most common being that the structure gives clues as to the function of the sequence. These clues can be either general, such as that Rossmann folds bind NAD(P), or specific, such as the function of a particular residue. Having an indication as to the structure of a protein also gives information on its evolutionary relationships and hence on the development of the protein repertoire. In the light of structural genomics projects (Rost, 1998; Shapiro and Lima, 1998), assigning structures to sequences is essential, and several attempts to do this for the small bacterium *Mycoplasma genitalium* (MG) (Teichmann *et al.*, 1998; Fischer and Eisenberg, 1997; Huynen *et al.*, 1998; Rychlewski *et al.*, 1998) as well as for yeast (Sanchez and Sali, 1998) and other completely sequenced genomes (Gerstein, 1997; Wolf *et al.*, 1999) have been made.

Previously we have described an intermediate sequence search method (ISS) (Park *et al.*, 1997). It is based on the fact that two related sequences, which have diverged beyond the point where their homology can be recognized by pairwise sequence comparisons, can both be matched by a third sequence that is suitably intermediate between the two. This procedure simply consists of two pairwise sequence searches, and for distantly related proteins it is one and a half times more sensitive than a single pairwise sequence search (Park *et al.*, 1998). A similar approach called FPS (Family Pairwise Search: Grundy, 1998; Grundy and Bailey, 1999) also showed the effectiveness of utilising related family members (intermediates) in homology searches.

The use of only one intermediate is not as sensitive as the multiple sequence methods mentioned above. Here the idea of an intermediate sequence library for PDB domain sequences is extended by collecting sequences homologous to the PDB sequences using the iterative multiple sequence features of PSI-BLAST rather than a pairwise algorithm like FASTA. In other words, instead of two pairwise sequence searches, a pairwise sequence search is done on a pre-computed library of sequences found by PSI-BLAST. For structural assignment, this library enables a single pairwise sequence search to be as sensitive as PSI-BLAST searches with each of the PDB domain sequences. The advantages over using PSI-BLAST directly over a sequence database are

1. Speed: it searches a much smaller pre-selected library of sequences with structural homologues.
2. Reliability: it is more reliable, because the intermediate sequence library can be calibrated and automatically edited to remove sequence with characteristics that give mismatches, i.e. matches to two PDB sequences of unrelated structure.

The library is available by ftp or can be searched against via the PDB-ISL server over the web (URLs as below).

Creation of the PDB intermediate sequence library: PDB-ISL

The structural, functional and evolutionary unit in proteins is called a domain. Since domains are combined in different ways in proteins, PDB-ISL must consist of separate domains to avoid relating two sequences which may only share one of several domains. The domain is the unit of classification in the Structural Classification of Proteins (SCOP) database (Murzin *et al.*, 1995), and the first version of PDB-ISL uses the library of domains and their definition in version 1.38 of SCOP. The sequences corresponding to SCOP domains will be denoted as PDBD sequences. (Multiple domains that are always found in the same arrangement in association with one another in the set of currently solved structures are considered as a single unit in SCOP: these constitute about 9% of the superfamilies in SCOP, hence only a small part of the PDB95D and PDB40D databases mentioned below.)

Since PDB is highly redundant, containing many variants of proteins that differ only in the presence or absence of a substrate or by an engineered mutation, a set of the PDBD sequences filtered to 95% sequence identity was used here. The resulting database, PDB95D, contains 3206 sequences and is available from the SCOP website (http://scop.mrc-lmb.cam.ac.uk/). In the work described here, we used an edited version of PDB95D from which we removed those sequences that have characteristics that are likely to produce false matches with good expectation values (E-values). These include membrane proteins, proteins with high cysteine or leucine contents and coiled coils (459 sequences have been removed). This version of PDB95D, called PDB95D-I, is available at the PDB-ISL server URL given above.

The PSI-BLAST program (version 2.0.6) was used to match the PDB95D-I sequences to sequences in a large, non-redundant representative sequence database filtered to 90% sequence identity, called NRDB90 (Holm and Sander, 1998, July 1998 version). PSI-BLAST begins with an initial BLAST2 search that collects homologues from the database that match the query sequence with E-values below a threshold defined by the user. A position-specific

matrix is built from the alignment of these sequences. The sequence database is then searched with this profile, and sequences that match with a score below the threshold are used to build a new matrix for the next round of searching. This process is continued to convergence, i.e. to the point at which no new sequences are found by the profile, or until the iteration number specified by the user is reached. Here, we used the PSI-BLAST parameters that had been found to be effective in giving a good coverage-to-error ratio (Park *et al.*, 1998): an E-value threshold of 0.0005 for the selection of homologous sequences for the PSI-BLAST profile, allowing up to 20 iterative searches, and an E-value threshold of 10^{-5} to be considered significant for a match between the query profile and a target sequence. The error per query using these values was estimated to be about 1%.

The names of the regions of the NRDB90 sequences that were found to match PDB95D-I sequences were 'attached' to the PDB95D-I sequences, so that the PDB-ISL library has the following FASTA format:

```
>ADH2_HORVU_178-320_d1agna2_3.22.1
GISTGLGATLNVTKPKKGMTVAIFGLGAVGLAAMEGA
RMSGASRIIGVDLNPAKHEQAKKFGCTDFVNPKDHTK
```

ADH2_HORVU_178-320 is an intermediate region of a domain structure d1agna2 in SCOP superfamily 3.22.1. This large set of sequences was filtered to remove 100% identical intermediates using the program nrdb (nrdb2: http://blast.wustl.edu/pub/nrdb/). The redundancy in this initial database is due to related PDB sequences matching the same sequence in NRDB90. There were 739 820 sequences removed from 1 162 385, resulting in 422 550. Low-complexity regions were removed using the SEG program (Wootton and Federhen, 1993). To reduce the 'edge effect', the N- and C-terminal 3 residues of each ISL sequence were truncated. By edge effect, we mean the fact that search methods sometimes extend a match from a high-scoring into a low-scoring region as discussed in Park *et al.* (1997). As mentioned above, this effect can potentially cause problems in multi-domain proteins. A total of 32 614 sequences, which contained regions of overlap due to the edge effect, were removed (a threshold of 30 residues overlap was used for an intermediate sequence when attached to two unrelated PDB domains, so the erroneous intermediates were those attached to two folds in the same sequence region.) The final database contains 389 936 sequence regions, about one-quarter of the size of NRDB90 in amino acids, and is therefore much faster to search.

Assessing pairwise sequence search procedures for PDB-ISL

To assess the optimal parameters for pairwise sequence search programs on PDB-ISL, a database with sequences of low similarity and known evolutionary relationships was used. This database is derived from sequences of known structures in SCOP. This is because the SCOP database contains a description of the evolutionary relations of proteins that are apparent not just from sequence similarities, but also from structural and functional features (Murzin *et al.*, 1995). The complete assessment package (SAT; Park *et al.*, 2000) is available from ftp://ftp.ebi.ac.uk/pub/contrib/jong/SAT.

The unit of classification in the database is the protein domain. Domains are clustered together into Families if they have close evolutionary relationships. Superfamilies bring together Families whose proteins have low sequence identities but whose structural details and, in many cases, functional features suggest that a common evolutionary origin is very probable, for example, the variable and constant domains of immunoglobulins.

To test the search programs on PDB-ISL, we measured the extent to which they could detect the evolutionary relationships described in the SCOP database. As there are few problems in finding relationships between proteins that have 40% or more sequence identity, we used a set of sequences that have pairwise identities of 40% or less. We call this set PDB40D. As for PDB95D, membrane proteins and the other types of proteins prone to errors in sequence comparison were omitted (189 sequences removed), such that the resulting database, called PDB40D-I, contains 1567 sequences. It has 283 PDB domains that are unique representatives of their superfamily and 1284 PDB sequences in 261 superfamilies which have two or more members. This version of PDB40D does not in fact perform better compared with the previous versions that contained membrane proteins etc., because the low-complexity regions of the query sequences are always masked (J. Park *et al.*, 2000, unpublished results).

The 1567 sequences can form 1 226 961 different pairs. According to the SCOP classification version 1.38, 6964 of these pairs are between proteins that have an evolutionary relationship at the family or superfamily level (848 pairs between proteins in the all-α class, 3025 in the all-β class and 3006 in the α/β or $\alpha + \beta$ classes and 90 in the multi-domain class.)

Pairs of sequences in the same superfamily are defined as homologous for the purposes of testing sequence comparison procedures on PDB-ISL. The SCOP database tends to be somewhat conservative in that it requires what its authors regard as good evidence to put structures into the same superfamilies rather than just the same fold category. Therefore, the 6340 pairs of sequences that have the same fold, but are not in the same superfamily, are defined as being of uncertain relationship (Park *et al.*, 1998). Hence such relationships are scored as neither true positives nor false positives, but simply as neutral. Non-homologous pairs are formed by any two sequences

that have different folds. (There are 1 213 652 non-homologous pairs in PDB40D-I.)

In the assessment of pairwise sequence comparison methods carried out (Brenner *et al.*, 1998), the error rate was given as the error-per-query rate (EPQ), which is the number of non-homologous matches divided by the number of query sequences. For an EPQ rate of 1% or less, the program FASTA (Pearson and Lipman, 1988) version 3.0 with k-tuple $=$ 1 is able to detect 18% of the evolutionary relationships described in SCOP. The conventional BLASTP algorithm is able to detect 11% of the relationships. In an assessment of the multiple sequence comparison methods SAM-T98 (Karplus *et al.*, 1998), PSI-BLAST and ISS (Park *et al.*, 1998), the error rate is reported as the rate of false positives (RFP) because there is no method for calculating statistical scores for the first of these methods. The RFP is the fraction of false positives made in the search divided by the total number of possible false relationships in the database. For the database used by Park *et al.* (1998), an EPQ of 1% corresponded to an RFP of 1/50 000. At this RFP, SAM-T98, PSI-BLAST and ISS find 35%, 30% and 25% of the evolutionary relationships described in SCOP. We are assuming roughly the same proportion of coverage at the same error rates was obtained when creating PDB-ISL, as the PDB40D-I database and the NRDB90 database used would be very similar (though updated) to those in the assessment of Park *et al.* (1998).

To calibrate the PDB-ISL for different sequence comparison methods, the 1567 sequences in PDB40D-I were searched against PDB-ISL. The three sequence comparison methods used were FASTA (k-tuple $=$ 1), SSEARCH, and BLAST2 (PSI-BLAST without iterations). As in Park *et al.* (1998), coverage is plotted as the fraction of homologues detected out of all the pairs of homologues in PDB40D-I (6964). An EPQ of 1% (16 false positives) was used to compare the coverage obtained with that in previous assessments (see Figure 1).

The coverage for FASTA and SSEARCH against PDB-ISL is 28.8% at a 1% error rate, and the E-value thresholds are 0.085 and 0.016 respectively. This is nearly the same as the performance of PSI-BLAST (30%) and higher than ISS (25%) as found in the assessment by Park *et al.* (1998) using an older PDB40D database. Coverage of BLAST2 with PDB-ISL at a 1% error rate was only 23.8%. The results are plotted in Figure 1. Going beyond a single iteration of PSI-BLAST, where a single iteration corresponds to BLAST2, affected the performance negatively, because errors accumulated without a significant increase in coverage. An E-value threshold of 0.01 is recommended for searching PDB-ISL with SSEARCH and 0.05 with FASTA to obtain an optimum coverage with a low error rate.

This means that FASTA on PDB-ISL is attaining

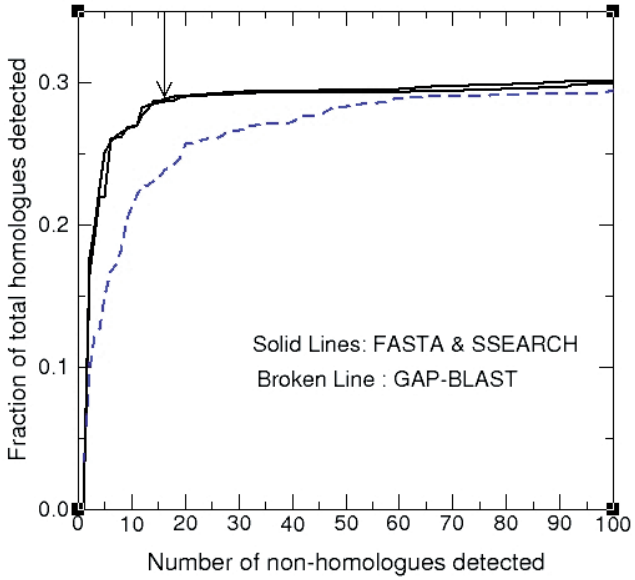

Fig. 1. The performance of PDB-ISL with different pairwise search methods. The coverage is shown on the y-axis as a fraction of the 6964 SCOP evolutionary relationships in the PDB40D-I database. The number of errors accumulated at increasing E-values is shown on the x-axis. The values at the arrow represent the coverage at 16 errors, a 1% error rate. The two solid curves are for FASTA and SSEARCH and the dotted curve is for BLAST2.

essentially the same coverage-to-error ratio as a full PSI-BLAST search on a large-sequence database such as NRDB90. Hence the profile is not providing much additional information above and beyond that in the collection of intermediate sequences. However, 32 614 out of 422 550 intermediate sequences in PDB-ISL that match PDB40D-I sequences in different folds were automatically removed in creating the PDB-ISL. It appears that such sequences are down-weighted when within a profile, but lead to an accumulation of errors in an intermediate sequence search. This is a general problem of a repetitive intermediate sequence search, also known as walking. One single wrong entry in PDB-ISL will result in wrong matches. This is one of the reasons why it is critical to remove all the mismatching intermediates in PDB-ISL through manual and automatic editing at the stage of construction.

Time required to search PDB-ISL

To illustrate the gain in speed of using FASTA with PDB-ISL over PSI-BLAST with a large-sequence database, searches with PDB sequences of different lengths were carried out. The speed of the FASTA search depends only on the query sequence length. The speed of the PSI-

Table 1. Times taken for searches. Sequences of different lengths using FASTA with PDB-ISL, and PSI-BLAST searches with NRDB90. The elapsed times are minutes taken on a DEC UNIX Alpha 21 164 500 MHz processor with 1 Gb of memory

Sequence Length	PDB-ISL and FASTA	PSI-BLAST and NRDB90			
		Iteration	Time	Iteration	Time
124	0.50	1	0.23	18	28.3
131	0.50	1	0.28	3	0.36
356	1.19			8	2.47
609	1.41			2	1.09
792	1.51			10	24.4

BLAST search depends on the query sequence length, number of homologues in the database and the number of iterations before the search converges. The platform used was a DEC Alpha 21164, 500 MHz processor, with 1 Gb memory and a UNIX operating system. Table 1 shows the elapsed times for the searches carried out. For short sequences less than 200 residues in length which converge after five or fewer iterations, PSI-BLAST is as fast or 10–30 s faster than the FASTA search. For longer sequences, or sequences which have many homologues in the database, FASTA is at least 1 min faster and can be up to 23 min faster. (PSI-BLAST becomes extremely slow for sequences that collect a lot of homologues and stretch the limits of available memory.)

For sequence searches involving large databases, such as entire genomes, computational efficiency is very important. Searching PDB-ISL with the MG genome sequences (468 sequences with an average length of 364 residues) is at least 20 times faster than doing a PSI-BLAST search with the parameters described above. With a very large genome such as *Caenorhabditis elegans* (CE) (19 099 proteins), searching PDB-ISL will be several hundred times faster than using PSI-BLAST. The two methods can be combined to obtain an optimal coverage: a PSI-BLAST search with the 2700 PDB95D sequences as queries can be carried out, and the remaining sequences can be used as queries for PDB-ISL.

In Wolf *et al.* (1999), PSI-BLAST matrices are built for each PDB domain and then saved. These are scanned against the genome sequences to make structural assignments. Figures for the time taken to make these matches are not available, but they will be similar to PDB-ISL. However, use of the final matrix only does not find as many homologues as if matches are collected at each iteration (Park *et al.*, 1998), as is done in creating PDB-ISL. Hence, this method may not be as sensitive as PDB-ISL, and is of course only equivalent to a one-way PSI-BLAST search starting from the PDB sequences as queries.

Table 2. Superfamilies which are found to be related by PDB-ISL. Sequences of the five TIM-barrel superfamilies and two α/α toroid and right-handed beta-helix superfamilies, as well as the last three superfamilies in the table are found to be significantly similar with PDB-ISL. The SCOP superfamily numbers and names are as in SCOP version 1.38

FOLD	SCOP id	SCOP superfamily name
ρ/α TIM barrel	3.1.5	NAD(P)-linked reductase
	3.1.7	FMN-linked oxidoreductase
	3.1.8	Tryptophan biosynthesis enzymes
	3.1.10	RubisCo C-terminal domains
	3.1.11	Tryphosphate isomerases
α/α Toroid	1.81.2	Cyclases
	1.81.3	Protein farnesyltransferases, β subunit
Right-handed β-helix	2.62.1	Pectin/pectate lyase
	2.62.3	Rhamnogalacturonase A
Rossman domains	3.4.1	FAD/NAD(P)-binding domain
	3.21.1	A nucleotide-binding domain
	3.22.1	NAD(P)-binding Rossman-fold domain

Distant evolutionary relationships detected with PDB-ISL

A superfamily brings together those families that have low sequence identities with each other but whose structural details and, in many cases, functional features suggest that a common evolutionary origin is very probable, for example, the variable and constant domains of immunoglobulins. The fold classification brings together superfamilies that have the same secondary structures in the same arrangement. For most superfamilies that share a common fold, there is good evidence that they do not have evolutionary relationships. In a few cases, the situation is less clear in that current evidence weakly supports the existence of evolutionary relationships. In these cases, the superfamilies are kept separate in SCOP until the subsequent discovery of intermediate structures provides stronger support for their merger. During the calibration of PSI-BLAST and FASTA on PDB-ISL with the PDB40D-I database, several relationships of sequences between SCOP superfamilies and even folds were detected with low, so very reliable E-values (see Table 2). There is good reason to believe that some of these are in fact distantly related rather than being false matches by the sequence comparison program.

Five superfamilies in the β/α (TIM) barrel fold are matched with E-values between 0 and $2e{-}36$. As pointed out in Wilmanns *et al.* (1991) and Janecek and Bateman (1996), these superfamilies are all enzymes with a common phosphate-binding site covering mainly the last two β-strands in the barrel, $\beta 7$ and $\beta 8$. Therefore, although the entries in SCOP do not share a high level of

structural similarity, there is evidence for divergence of these proteins based on structural as well as functional similarities.

In the same way, three nucleotide-binding superfamilies in separate folds are found related: the FAD/NAD(P)-binding domain, NAD(P)-binding Rossmann-fold domains and a nucleotide-binding domain. The nucleotide-binding domain superfamily is annotated in SCOP as 'sharing the common nucleotide-binding site with and providing a link between the Rossmann-fold NAD(P)-binding and FAD/NAD(P)-binding domains'. Since PDB-ISL detects these relationships, it is clear that there is a signal at the sequence as well as at the structural level.

Two pairs of superfamilies within folds were also detected as related with very good E-values. Within the right-handed β helix fold, the rhamnogalacturonase A sequence matches all the pectin and pectate lyase sequences with very low E-values, even though these are two different superfamilies. As both of these superfamilies consist of superhelices built from turns of three strands connected by short links, they may be divergently related. Finally, two sequences from the cyclase and protein farnesyltransferase (β subunit) superfamilies in the alpha/alpha toroid fold match with an E value of $1e-157$.

Structural annotation of genomes with PDB-ISL

In the preceding sections, the speed and sensitivity of PDB-ISL have been demonstrated. These are the two characteristics which are important in a method used to assign structures to genome sequences, in other words computational structural genomics. The genome that has been used most widely as a test of different structural assignment methods is that of the small, parasitic bacterium MG. Here, the MG protein sequences were matched to PDB-ISL sequences with FASTA and an expectation value threshold of 0.05. Using this method, 44% of the MG proteins and 29% of the amino acids obtained a structural annotation. By using PDB domains as queries and PSI-BLAST as the search procedure to match MG sequences embedded in a large database, we found matches to 41% of the proteins and 28% of the amino acids. Doing an additional PSI-BLAST search with the unmatched MG sequences as queries finds an additional 4% proteins and 4% amino acids. Therefore, the 'two-way' PSI-BLAST search is the best method, but is only a little better than PDB-ISL assignment.

With a minuscule genome such as MG, it is feasible to do PSI-BLAST searches for all the PDB domains and all the MG proteins. However, for larger genomes, this becomes computationally prohibitive. The *Saccharomyces cerevisiae* (SC) genome, for instance, is over 15 times larger than MG and CE is over 40 times larger. Since the number of PDB domain sequences is the same, it makes sense to do a 'one-way' PSI-BLAST search with these sequences as queries. The remaining unmatched genome sequences can then be used as queries for a FASTA search against PDB-ISL. With this approach, 15% of the CE sequences amino acids were matched using PSI-BLAST and an additional 3% found with PDB-ISL, a gain of 20%. In SC, 17% of the amino acids were matched with PSI-BLAST and PDB-ISL found 3% more, a gain of 18%. This means that PDB-ISL can make a considerable contribution to the computational side of structural genomics.

Many of the additional matches found by searching the remaining unmatched CE and SC genome sequences against PDB-ISL after an initial PSI-BLAST search starting from the PDB domains are due to repetition of a single domain within the proteins. For instance, using a one-way PSI-BLAST search starting from all the immunoglobulin superfamily domains in SCOP, 334 such domains are identified in CE. Taking the remaining unmatched regions of the CE proteins, another 53 immunoglobulin domains are found in the worm genome. This number is corroborated by HMM searches for immunoglobulin domains, where as many as 462 such domains are identified.

The PDB-ISL server

Locally installed PDB-ISL as a library or database with a pairwise search algorithm is useful for large-scale structural assignment, for instance of genomic sequences. This is due to the increase in speed as compared with multiple sequence comparison methods that attain the same coverage. The database can be obtained by ftp from ftp://stash.mrc-lmb.cam.ac.uk/pub/PDB_ISL/. In addition to this, a Web-based user interface is available to query individual protein sequences of interest at http://stash.mrc-lmb.cam.ac.uk/PDB_ISL/ and http://cyrah.ebi.ac.uk/Serv/PDB_ISL/. A variety of sequence search methods and parameters are available on the PDB-ISL server. The page of PDB-ISL search results has SCOP classification and PDB name links, so that users can see the matched structures immediately.

The server, all the PDB-ISL related files and programs are available on request on a CD-ROM for Linux platform, so that a faster in-house PDB-ISL server can be installed.

Discussion and conclusion

We have constructed a library of intermediate protein sequences for the sequences of proteins of known structure (PDB-ISL). It can be used for fast and reliable assignment of structures to sequences.

After the completion of this work, the efficiency of this approach has been demonstrated by Salamov *et al.* (1999) using a somewhat similar intermediate sequence library for PDB sequences with PSI-BLAST. However,

they do not compare the performance of their intermediate sequence library with the performance of PSI-BLAST. This is essential to demonstrate objectively that PDB-ISL is a useful tool. Equally important is the fact that PDB-ISL is accessible on the web as a server and the database and the server package can be downloaded by ftp for critical evaluation, while the intermediate sequence library of Salamov *et al.* (1999) is described as being not freely available.

Using pairwise sequence comparison methods on this database achieves a similar coverage-to-error ratio for detection of structural homologues as a normal PSI-BLAST search against a very large non-redundant database with a maximum of 20 iterations. At the same time, FASTA with PDB-ISL is easier to use and is much faster than PSI-BLAST for many sequence searches. Libraries can also be made for families of interest that do not have structures, for instance the library of families in the Pfam HMM database (Bateman *et al.*, 1999), which in 1999 contained, among other things, 70% of SCOP families (Elofsson and Sonnhammer, 1999).

Acknowledgements

S.A.T. thanks the Boehringer Ingelheim Fonds. J.P. is supported by a grant from Hoechst Marion Roussel. We thank many of the Church laboratory members for criticism, especially Jason Johnson.

References

Altschul,S.F., Madden,T.L., Schaffer,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. *Nucl. Acids Res.*, **25**, 3389–3402.

Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Finn,R.D. and Sonnhammer,E.L.L. (1999) Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucl. Acids Res.*, **27**, 260–262.

Brenner,S.E., Chothia,C. and Hubbard,T.J.P. (1998) Assessing sequence comparison methods with reliable structurally identified evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.

Eddy,S.R. (1996) Hidden Markov models. *Curr. Op. Struc. Biol.*, **3**, 361–365.

Elofsson,A. and Sonnhammer,E.L. (1999) A comparison of sequence and structure protein domain families as a basis for structural genomics. *Bioinformatics*, **15**, 480–500.

Fischer,D. and Eisenberg,D. (1997) Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proc. Natl Acad. Sci. USA*, **94**, 11929–11934.

Gerstein,M. (1997) A structural census of genomes: comparing bacterial, eukaryotic and archaeal genomes in terms of protein structure. *J. Mol. Biol.*, **274**, 562–576.

Grundy,W.N. (1998) Homology detection via family pairwise search. *J. Comput. Biol.*, **5**, 479–491.

Grundy,W.N. and Bailey,T.L. (1999) Family pairwise search with embedded motif models. *Bioinformatics*, **15**, 463–470.

Holm,L. and Sander,C. (1998) Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, **14**, 423–429.

Huynen,M., Doerks,T., Eisenhaber,F., Orengo,C., Sunyaev,S., Yuan,Y. and Bork,P. (1998) Homology-based fold predictions for *Mycoplasma genitalium* proteins. *J. Mol. Biol.*, **280**, 323–326.

Janecek,S. and Bateman,A. (1996) The parallel (a/b)8 barrel: perhaps the most universal and the most puzzling fold. *Biologia*, **51**, 613–628.

Karplus,K., Barrett,C. and Hughey,R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.

Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) Hidden Markov-models in computational biology—Applications in protein modelling. *J. Mol. Biol.*, **235**, 1501–1531.

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.

Park,J., Teichmann,S.A., Hubbard,T. and Chothia,C. (1997) Intermediate sequences increase the detection of distant sequence homologies. *J. Mol. Biol.*, **273**, 349–354.

Park,J., Karplus,K., Barrett,C., Hughey,R., Haussler,D., Hubbard,T. and Chothia,C. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.*, **284**, 1201–1210.

Park,J., Holm,L. and Chothia,C. (2000) Sequence search algorithm assessment and testing toolkit (SAT). *Bioinformatics*, **16**, 104–110.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Rost,B. (1998) Marrying structure and genomics. *Structure*, **6**, 259–263.

Rychlewski,L., Zhang,B. and Godzik,A. (1998) Fold and function predictions for *Mycoplasma genitalium* proteins. *Folding and Design*, **3**, 229–238.

Salamov,A.A., Suwa,M., Orengo,C.A. and Swindells,M.B. (1999) Genome analysis: assigning protein coding regions to three-dimensional structures. *Protein Sci.*, **8**, 771–777.

Sanchez,R. and Sali,A. (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl Acad. Sci. USA*, **95**, 13597–13602.

Shapiro,L. and Lima,C.D. (1998) The Argonne Structural Genomics Workshop: Lamaze class for the birth of a new science. *Structure*, **6**, 265–267.

Tatusov,R.L., Altschul,S.F. and Koonin,E.V. (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Natl Acad. Sci. USA*, **91**, 12091–12095.

Teichmann,S.A., Park,J. and Chothia,C. (1998) Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplication and domain rearrangement. *Proc. Natl Acad. Sci. USA*, **95**, 14658–14663.

Wolf,Y.I., Brenner,S.E., Bash,P.A. and Koonin,E.V. (1999) Distribution of protein folds in the three superkingdoms of life. *Genome*

Res., **9**, 17–26.

Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, **17**, 149–163.

Wilmanns,M., Hyde,C.C., Davies,D.R., Kirschner,K. and Jansonius,J.N. (1991) Structural conservation in parallel beta/alpha-barrel enzymes that catlyze three sequential reaction in the pathway of tryptophan biosynthesis. *Biochemistry*, **30**, 9161–9169.