David H. Kitson, Azat Badretdinov, Mikhail Velikanov, David J. Edwards, Krzysztof Olszewski, Sándor Szalma and Lisa Yan
are members of either the Computational Molecular and Structural Biology R&D Group or the Life Sciences Marketing Group of Accelrys Inc

Zhan-yang Zhu
is currently a Senior Bioinformatics Scientist at Structural GenomiX.

Functional annotation of proteomic sequences based on consensus of sequence and structural analysis

David H. Kitson, Azat Badretdinov, Zhan-yang Zhu, Mikhail Velikanov, David J. Edwards, Krzysztof Olszewski, Sándor Szalma and Lisa Yan

Date received (in revised form): 11th November 2001

Abstract

To maximise the assignment of function of the proteins encoded by a genome and to aid the search for novel drug targets, there is an emerging need for sensitive methods of predicting protein function on a genome-wide basis. GeneAtlas™ is an automated, high-throughput pipeline for the prediction of protein structure and function using sequence similarity detection, homology modelling and fold recognition methods. GeneAtlas is described in detail here. To test GeneAtlas, a 'virtual' genome was used, a subset of PDB structures from the SCOP database, in which the functional relationships are known. GeneAtlas detects additional relationships by building 3D models in comparison with the sequence searching method PSI-BLAST. Functionally related proteins with sequence identity below the twilight zone can be recognised correctly.

Keywords: *high-throughput modelling, homology modelling, remote homology detection, function prediction*

INTRODUCTION

The genomes of many organisms have been or are now being sequenced. A critical step in exploiting the data is to determine the function of the encoded proteins.[1] Assignment of putative function is often done by attempting to find a relationship between the sequence of a new protein and the sequences of existing proteins of known function.[2] Protein function, however, is more directly related to the three-dimensional structure of the protein than to its amino acid sequence. Structure, therefore, tends to be more highly conserved than sequence. Thus, prediction of the three-dimensional structure of a protein and comparison with proteins of known structure and function might enable the elucidation of function in cases where searching for sequence homology fails.[3–9]

Structure and function prediction efforts are closely allied with structural genomics initiatives.[10] These are oriented towards the large-scale experimental determination of protein structures, with the goal of expanding our knowledge of the range of folds accessible to proteins and of placing all proteins within 'modelling distance' of an experimental structure. The success of structure prediction methods depends on the existence of a structure that adopts a similar fold to the sequence for which a relationship is being sought. Structural genomics initiatives will, therefore, enhance the applicability of structure prediction methods, which in turn will provide the tools to fully utilise the experimental data.

An automated 'pipeline' (GeneAtlas™)[11] for the high-throughput annotation of protein sequences has been constructed. The methodology used in GeneAtlas is described below, followed by the results of a validation study and the results of the application of the pipeline to the proteins encoded by complete genomes.

David H. Kitson,
Accelrys Ltd,
230/250 The Quorum,
Barnwell Road,
Cambridge,
CB5 8RE, UK

Tel: +44 1256 422918
Fax: +44 1256 429035
E-mail: dkitson@accelrys.com

METHODOLOGY

The GeneAtlas pipeline

Use of a combination of methods enhances homology recognition

GeneAtlas is designed to predict structures of proteins for large sets of protein sequences and to assign function based on sequence and structural analysis.[3] The major components of GeneAtlas (Figure 1) include homology searching using PSI-BLAST, fold recognition using SeqFold, high-throughput homology modelling using MODELER, and functional annotation based on the sequence and predicted structure. Park *et al.*[12] observed that different sequence similarity searching methods give both common hits and hits that are unique to each method. GeneAtlas, therefore, uses the combination of several methods to enhance homology recognition. The confidence of the assignment is also increased when consensus hits are obtained by different methods.

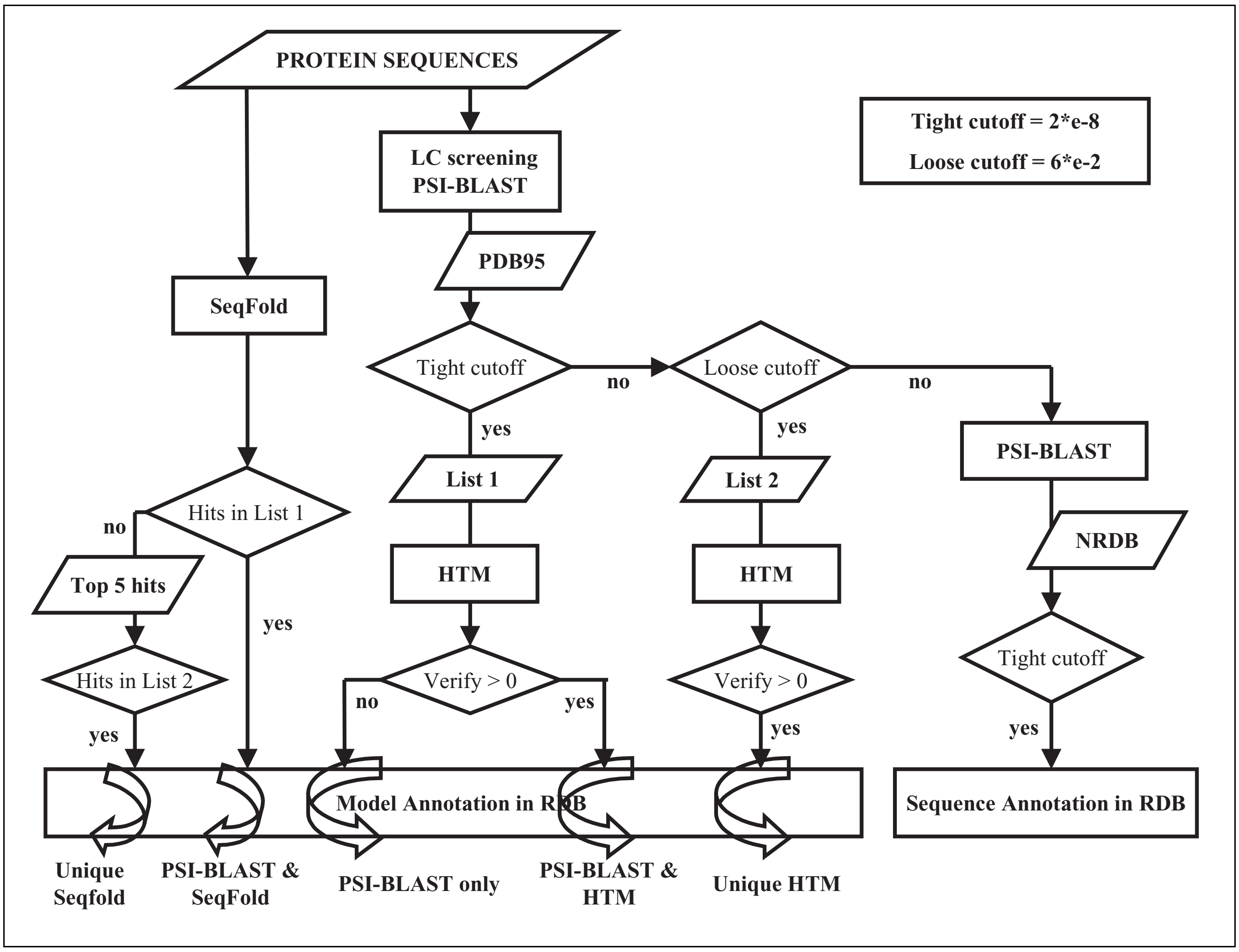


Figure 1: GeneAtlas flowchart. Hits obtained with GeneAtlas are divided into several categories. PSI-BLAST and HTM hits are the hits with a PSI-BLAST score better than tight cutoff and a positive model verification score. PSI-BLAST and SeqFold hits are hits with a PSI-BLAST score better than tight cutoff that also occur in the SeqFold hit list. PSI-BLAST-only hits are hits with a PSI-BLAST score better than tight cutoff that are not found in SeqFold and that have a negative model verification score. Unique HTM hits are hits with a PSI-BLAST score between tight cutoff and loose cutoff that have a positive model verification score. After removing PSI-BLAST and SeqFold consensus hits, the remaining SeqFold hits are sorted using the SeqFold raw score and the top five hits are selected. Any of these top five hits that are found in the PSI-BLAST hit list with a score between tight cutoff and loose cutoff are classified as SeqFold unique hits. If a sequence does not have any homologue in the structure database, the hits found by searching the nrdb database are reported as sequence annotations

PSI-BLAST search

PSI-BLAST[13] is used to identify homology relationships between a query sequence and sequences of proteins with known structure in the PDB95 database (see below). Using PSI-BLAST, each query sequence is used to search a non-redundant sequence database, nrdb90,[14] to build a position-specific scoring matrix (a 'profile') for the query. The profile is then used to search the PDB95 database to identify possible structural templates. To maximise homology recognition, a reverse search starting from template sequences is also carried out (Figure 2). The hits from the direct and reverse searches are then combined.

A 'reverse' PSI-BLAST search enhances homology recognition

The profiles are built by running 20 iterations of PSI-BLAST against the nrdb90 database, unless convergence occurs earlier. The *E*-value threshold used for selecting the hits that are used for constructing the profile is set at 0.0005. The BLOSUM-62 scoring matrix is used. Profile divergence is one of the main problems that can lead to false positive hits. To test for divergence, checkpoints are set at iterations 3, 5, 10, 15 and 20. If the top hit in the initial iteration disappears at a checkpoint, the profile is considered to have diverged. If the sequence identity of the top hit in the initial iteration is much higher (difference larger than 50 per cent) than the top hit at a checkpoint and the length of the former is at least 80 per cent of the latter, the profile is also considered to have diverged. When divergence occurs, the PSI-BLAST search is stopped and the profile of the previous checkpoint is used. Since sequences with low-complexity regions often result in diverged profiles, the low-complexity screening option in PSI-BLAST is used to remove these regions during a search.

An algorithm is used to reduce the effects of PSI-BLAST profile divergence

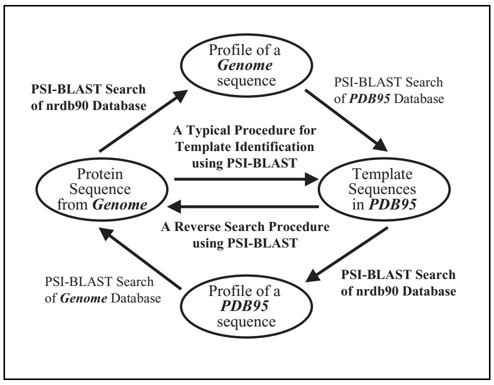

Profile of a *Genome* sequence
PSI-BLAST Search of nrdb90 Database
PSI-BLAST Search of *PDB95* Database
A Typical Procedure for Template Identification using PSI-BLAST
Protein Sequence from *Genome*
Template Sequences in *PDB95*
A Reverse Search Procedure using PSI-BLAST
PSI-BLAST Search of *Genome* Database
Profile of a *PDB95* sequence
PSI-BLAST Search of nrdb90 Database

Figure 2: PSI-BLAST search protocol. Homology relationships between genome sequences and structures in the PDB95 database are found using both a direct search and a reverse search. In the direct search, a PSI-BLAST profile is generated for a genome sequence and this profile is used to search the sequences of the PDB95 proteins. In the reverse search, a profile is generated for a PDB95 protein and this is used to search the genome sequence database. Profiles are generated by searching the nrdb90 database

High-throughput modelling (HTM)

Relationships between query sequences and PDB95 structures found in both the direct and reverse PSI-BLAST searches are used to build homology models of the query sequences.

A number of protein families include many structures that are close homologues. When a query sequence is homologous to this type of family, many matching templates will cover the same region of the query sequence. Building models for all of these hits will result in little additional annotation of the query sequence, while requiring extensive computational time. For each region of the query sequence, a maximum of 10 top-ranking hits is selected for building models (see the supplementary material[15] for details).

MODELER is used to build models

The models are built using the homology modelling package MODELER,[16] with the alignments taken directly from PSI-BLAST. They are then evaluated using the Profiles-3D[17] and PMF (A. Šali, personal communication) methods. Profiles-3D tests a structure by examining the environment of each residue and assigning a score to reflect the suitability of the environment. The total score for all residues is then compared to expected upper and lower bounds for a protein of the same length. The upper bound is derived by analysis of protein structures known to be correct or incorrect. In the present work, the lower bound was set to 50 per cent of the upper bound. The score is normalised using:

$$\text{Normalised score} = \frac{\text{raw score} - \frac{1}{2}\text{high score}}{\frac{1}{2}\text{ high score}}$$

Models with a normalised score above 0 are accepted. The PMF score is similar to the Profiles-3D score, but uses a pairwise potential that accounts for the pairwise residue interaction energy.

Any hits with an *E*-value better than the tight cutoff (Figure 1) are considered good hits regardless of their model scores, and are retained. The *E*-value cutoff that is used for selecting hits for model building (the loose cutoff; Figure 1) is much looser than the cutoff that would normally be used for PSI-BLAST searches. The large number of false positives included in the PSI-BLAST hits between the tight and loose cutoffs are weeded out based on the quality of the three-dimensional models, as described above. By using HTM and SeqFold (see below) as an extension of the PSI-BLAST search, low-confidence hits found in the sequence search are, therefore, confirmed or rejected by structural information.

Low-confidence PSI-BLAST hits are judged by evaluation of 3-D models

SeqFold

SeqFold[11,18,19] is a secondary structure-enhanced, sequence-similarity searching program. The query sequence is searched against a library of known structures to find potential folds. In GeneAtlas, the fold library is constructed based on the PDB95 database and contains both the sequence and the secondary structure derived using the Kabsch and Sander method.[20] The secondary structure for the query sequence, which is compared with the actual secondary structure of the members of the fold library, is predicted using Discrimination of protein Secondary structure Class (DSC).[21] SeqFold hits are selected in combination with the PSI-BLAST hits as shown in Figure 1.

Searching the nrdb sequence database

If no homologues of a query sequence are found in the PDB95 structure database, no model can be built for this sequence. However, there is a greater chance of finding homologues in the nrdb sequence database since this contains many more sequences, which often have functional annotation. GeneAtlas uses the PSI-BLAST profile generated in the previous step and runs one iteration of PSI-BLAST to identify hits in nrdb.

Databases

PDB95

This database, in which any pair of sequences has less than 95 per cent sequence identity, is derived from the Protein Data Bank (PDB).[22] Theoretical models in the PDB are not included. The sequences for multi-chain proteins are divided into separate sequences for each chain. A filtering mechanism is used that removes chains with greater than 95 per cent sequence identity to any other member of the database (see the supplementary material[15] for details).

The current PDB95 database was created based on the PDB release of 14th December, 2000, and consists of 5,460 PDB chains extracted from 4,658 PDB structures.

nrdb and nrdb90

The nrdb sequence database[23] contains non-redundant protein sequences from GenBank CDS translations, PDB, SwissProt, PIR and PRF databases. In the current GeneAtlas release, the nrdb version of 5th February, 2001, which consists of 616,766 sequences, is used.

The nrdb90 database[14] is a subset of nrdb and included 359,584 sequences on 1st February, 2001. Any two sequences in this database have less than 90 per cent sequence identity.

Validation of high-throughput modelling

The premise behind the GeneAtlas pipeline is that, by using structural methods in addition to sequence-based methods, we can assign additional relationships between remote homologues, over and above those that can be found by sequence-based methods alone. The ability of the high-throughput modelling component of GeneAtlas to assign relationships has been evaluated using a 'virtual genome' in which the relationships are already known.

A 'virtual genome' (PDBD40-JA) is used for validation

Validation database

The goal of the validation study is to selectively identify 1,980 superfamily-level pairs

The Structural Classification Of Proteins (SCOP) database[24] contains the sequences and structures of protein domains. These are classified according to their structures and evolutionary relationships. Families contain domains with a close evolutionary relationship. Superfamilies contain families where the structures and other evidence indicate a common origin. The sequence similarity at this level, however, may be low. The fold level includes superfamilies that have the same overall structural fold, but which are believed not to have an evolutionary relationship. Finally, folds with the same gross overall structure (all-alpha, all-beta, etc.) are brought together at the class level.

This validation study focuses on the HTM component of GeneAtlas

Park *et al.*[12] derived a subset of the SCOP database (PDBD40-J) that included only domains that have a pairwise sequence identity of 40 per cent or less. This was used in their evaluation of the ability of various sequence-based methods to find relationships between remote homologues. Our database was based on PDBD40-J, with changes made to reflect updates in SCOP version 1.37 and to remove domains that could be unsuitable for structural modelling (for example, those with only alpha-carbon coordinates). This left 912 domains, which are listed in the supplementary material.[15] We shall refer to this test set as PDBD40-JA.

PDB40-JA includes 912 SCOP domains

Pairs of functionally related proteins will fall within the same SCOP family or in different families within the same superfamily. Thus, the task of identifying pairs of homologous domains equates to identifying, out of all possible pairwise relationships, those pairs of domains that are within the same SCOP superfamily. In our test set of 912 domains, there are 1,980 pairs that are in the same superfamily. Owing to the uncertainty in the relationship between domains that are in the same fold class, but different superfamilies,[12] pairs that have this relationship were excluded from the analysis – there were 1,975 such pairs. This leaves 411,461 non-homologous pairs. The challenge of identifying homologous pairs is, therefore, to find as many as possible of the 1,980 superfamily-level pairs, while minimising the incorrect selection of false positives from the 411,461 non-homologous pairs.

This is a very challenging test set, with 72 per cent (1,426) of the relationships having a pairwise sequence identity of less than 20 per cent and 41 per cent (811) having a pairwise identity of less than 16 per cent (Figure 3).

Protocol used for validation study

In this validation study, we focused on assessing the HTM component of GeneAtlas, using a subset of the pipeline (Figure 4). The two most significant changes to the pipeline were: (a) SeqFold was not used; and (b) since the goal was to search for pairwise relationships within the PDBD40-JA test set, this set of structures was used in place of the PDB95 template database. This means that the same set of 912 protein domains (PDBD40-JA) constitutes both the query (target) sequences and the template structures. A PSI-BLAST profile was constructed for each query sequence as described above and this was used to search a database containing the sequences of the 912 domains. All hits with an *E*-value of less than or equal to 10.0 were accepted, as long as the alignment included at least 25 residues of the target sequence.

To test each hit, a model was built of the query sequence using MODELER, using the three-dimensional structure of the hit as a template. The models were evaluated using Profiles-3D and models with a positive normalised score were

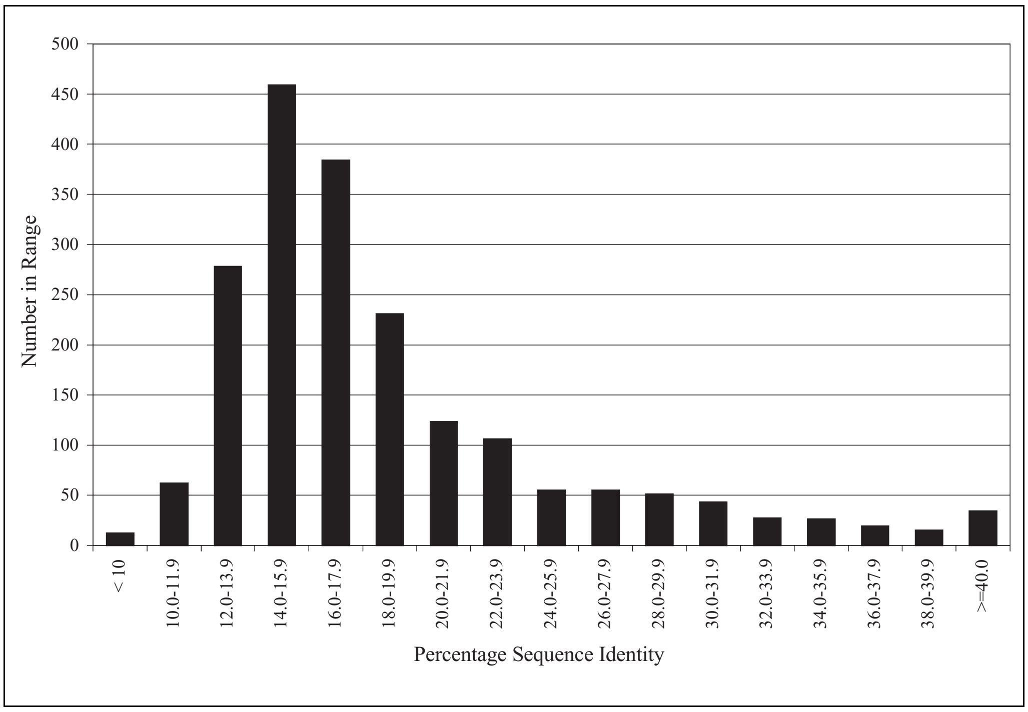


Figure 3: Sequence identities within the test set. The distribution of pairwise sequence identities for the 1,980 superfamily level relationships within the test set is shown. The sequences were aligned using a global alignment algorithm and the percentage identity was calculated for the alignment. Many of the pairs have a very low percentage identity. Given the low sequence identity, it is probable that some of the sequence alignments are in error (see also the results discussed below relating to accuracy of sequence alignments). Thus, these percentage identities should be considered an upper bound on the true values (those that could be derived from a structural alignment of the domains)

Many true pairs have very low sequence identity

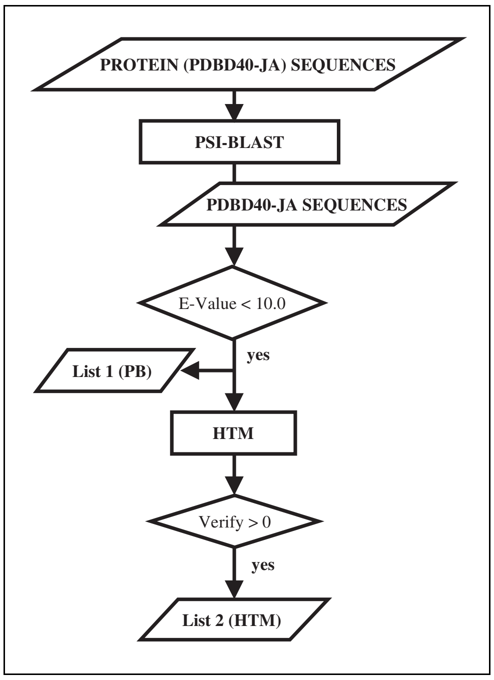


Figure 4: Subset of pipeline used for validation study. PSI-BLAST was used to search the PDBD40-JA database using each sequence in the same database as a query sequence in turn. All hits with an *E*-value of less than 10.0 were saved (List 1). Models were built for each hit and evaluated with Profiles-3D. All models with a positive score were accepted (List 2)

accepted. In this experiment the use of a tight cutoff (Figure 1) was not applied – all hits were modelled, and accepted or rejected based on their Profiles-3D score.

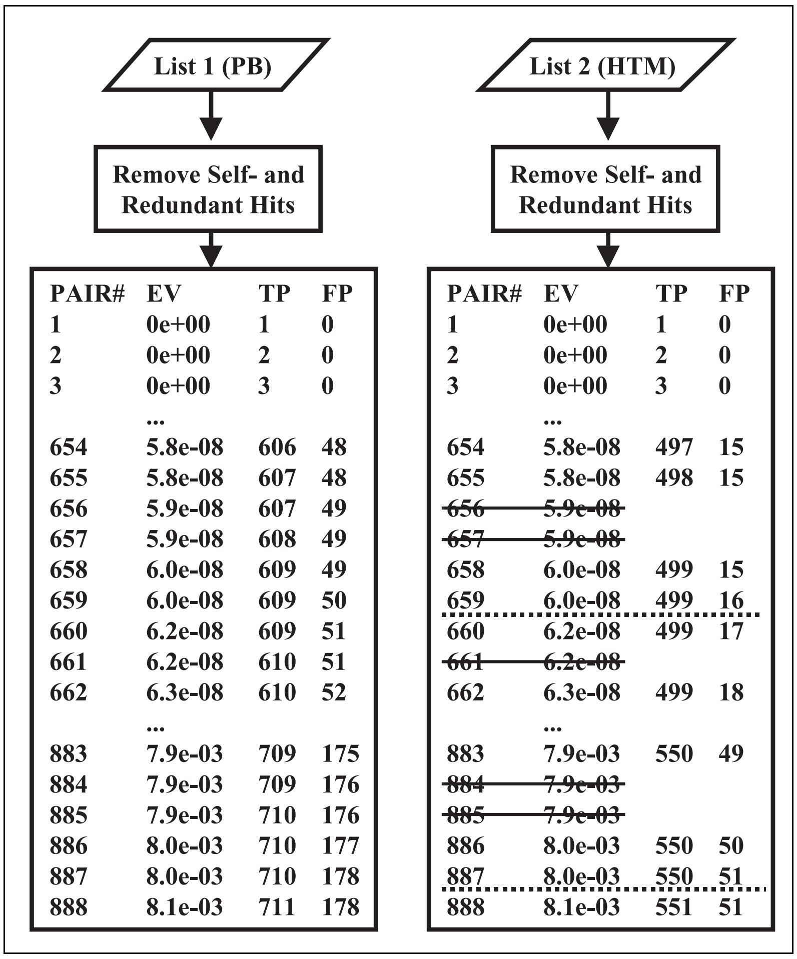
Assessment of hit lists

Since the relationships between the 912 domains are known, we can evaluate whether the relationships (hits) that we find are correct matches (true positives; both domains are in the same SCOP superfamily) or false positives (in different superfamilies). The results from the PSI-BLAST searches were analysed first. The hit list was first processed to remove self-hits and redundant hits (this is described in detail in the supplementary material[15]). The resulting hit list was ranked according to the *E*-values, with the hits with the lowest (most favourable) *E*-value at the start of the list (Figure 5). The list could then be scanned to count, for a given *E*-value cutoff, the numbers of true and false positive relationships that were found. Alternatively, the number of true positives found for a given number of false positives could be counted.[25]

As described above, for each PSI-BLAST hit, a model was built. In assessing the results of the modelling, pairs were first rejected for which the Profiles-3D score indicated an incorrect model. Again, the hit list was processed to remove redundant hits (see the supplementary material[15]) and then all pairs were ranked according to the PSI-BLAST *E*-value (Figure 5). The hit list was then analysed as described above.

Since both true and false positives are eliminated from the HTM hit list when models are rejected, the HTM hit list must be a subset of the PSI-BLAST hit list. This means that the cutoff for a given number of false positives will occur ‘further down’ the HTM hit list (at higher, less favourable *E*-values). In comparing the lists, the hits that fall between the cutoff for a given number of false positives in the PSI-BLAST hit list and the cutoff for the same number of false positives in the HTM hit list are considered to be additional hits (true or false) found by HTM. This is illustrated in Figure 5. The 52 true positives and 35 false positives that fall between the dotted lines on the HTM hit list in this figure are considered to be additional hits at the 50 false positive cutoff.

To allow for a comparison between BLAST[26] and PSI-BLAST for our test set, a straightforward BLAST search was also run for each of the 912 sequences. This was run and analysed in the same way as the PSI-BLAST search described above,


List 1 (PB)
Remove Self- and Redundant Hits
List 2 (HTM)
Remove Self- and Redundant Hits

PAIR#	EV	TP	FP
1	0e+00	1	0
2	0e+00	2	0
3	0e+00	3	0
	...		
654	5.8e-08	606	48
655	5.8e-08	607	48
656	5.9e-08	607	49
657	5.9e-08	608	49
658	6.0e-08	609	49
659	6.0e-08	609	50
660	6.2e-08	609	51
661	6.2e-08	610	51
662	6.3e-08	610	52
	...		
883	7.9e-03	709	175
884	7.9e-03	709	176
885	7.9e-03	710	176
886	8.0e-03	710	177
887	8.0e-03	710	178
888	8.1e-03	711	178

PAIR#	EV	TP	FP
1	0e+00	1	0
2	0e+00	2	0
3	0e+00	3	0
	...		
654	5.8e-08	497	15
655	5.8e-08	498	15
~~656~~	~~5.9e-08~~		
~~657~~	~~5.9e-08~~		
658	6.0e-08	499	15
659	6.0e-08	499	16
660	6.2e-08	499	17
~~661~~	~~6.2e-08~~		
662	6.3e-08	499	18
	...		
883	7.9e-03	550	49
~~884~~	~~7.9e-03~~		
~~885~~	~~7.9e-03~~		
886	8.0e-03	550	50
887	8.0e-03	550	51
888	8.1e-03	551	51

Figure 5: Protocol for assessing hit lists. After removing self-hits and redundant hits, the remaining pairs were ranked according to their PSI-BLAST *E*-value (EV). For any given pair, the total number of true positives (TP) and false positives (FP) up to that point is listed. For the HTM hit list, pairs that had an unfavourable model score were rejected (shown struck through). The dotted lines on the HTM hit list enclose the additional hits (52 true and 35 false) that are found between the cutoff at which 50 false positives are found in the PSI-BLAST list (6.0e−08) and the cutoff at which 50 false positives are found in the HTM list (8.0e−03). Hits that have been eliminated above the upper dotted line (including 656 and 657 in this illustration) are unique to the PSI-BLAST hit list at the 50 false positive rate

except that the initial step of building sequence profiles by searching nrdb90 was omitted.

RESULTS

Remote homology detection rates of PSI-BLAST and HTM

Using BLAST, the total numbers of true and false positives found were 462 and 2,702, respectively, if we do not apply an *E*-value cutoff to the hit list (ie we include all hits up to the limit of 10.0 that was used in the search). For PSI-BLAST the corresponding numbers were 899 and 2,750. After eliminating the pairs for which HTM indicated an incorrect model, 611 true and 284 false positives remained. These are out of a total of 1,980 true positive pairs present in the test set, and 411,461 false positives. Thus, PSI-BLAST finds more true positives than HTM, but at the expense of almost ten times more false positives.

PSI-BLAST identifies more true positives than HTM, but also many more false positives

To compare the methods, we can use a plot that indicates the numbers of true positives found for a fixed number of false positives[25] – Figure 6. This shows that the first several hundred true positives are readily found by all methods, at a cost of few false positives. After around five false positives have been found, however, the number of false positives found per true positive increases significantly. Table 1 lists the number of true positives found at rates of 5, 10, 50, 100, 200 and 400 false positives, and shows the *E*-value thresholds corresponding to these false positive rates.

Comparison of homologues found using PSI-BLAST and HTM

As discussed above, we can characterise hit lists by counting the numbers of true positives found for a given rate of false positives. At a rate of 100 false positives, we find 681 true positives with PSI-BLAST and 590 with HTM, at *E*-value

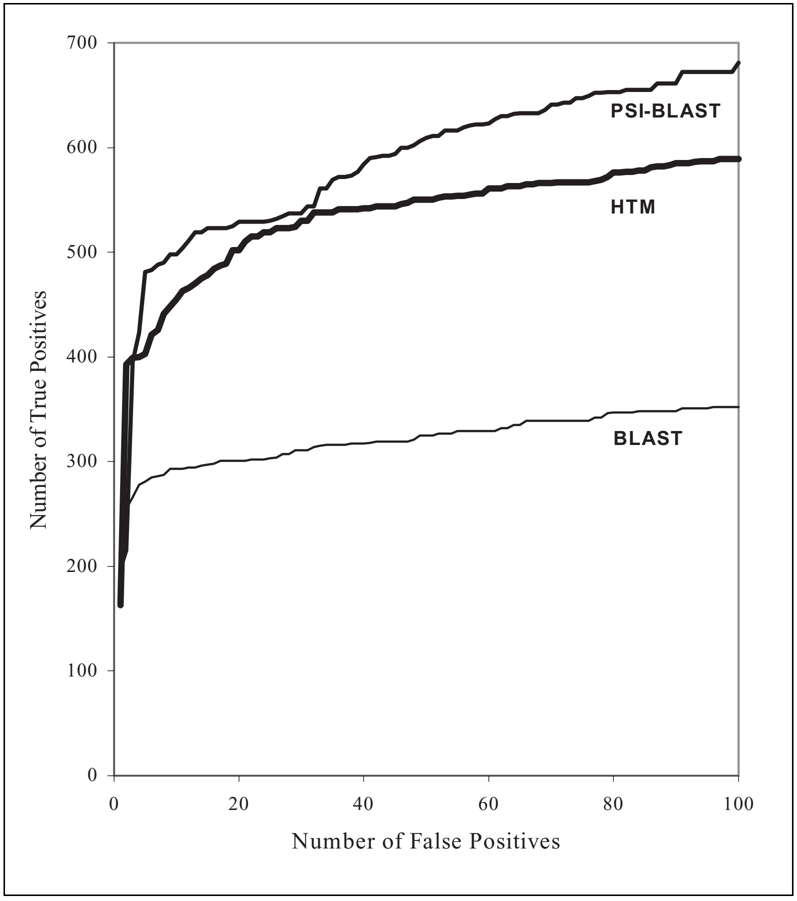


Figure 6: Hit rates for homology search methods. The number of correct matches (true positives) found by different search methods, for between 1 and 100 false positive relationships is shown. Hit lists were ranked according to the *E*-values of the hits. The hit lists were then processed, starting at the lowest (most favourable) *E*-value, counting the numbers of true and false positives found. Comparison of this plot and Figure 1 in Park *et al.*[12] shows that, for 100 false positives, approximately 100 fewer true positives were found using PSI-BLAST in the present study than in the work of Park *et al.* Detailed comparison of the hit lists from this work and from Park *et al.* (J. Park, personal communication) indicated that this was due to differences between SCOP version 1.37, used in the present study, and version 1.35, used by Park *et al.* (ie some pairs of domains that were identified as superfamily-level homologues in version 1.35 were placed in different superfamilies in version 1.37)

Table 1: Numbers of correct matches (true positives) found for different rates of false positives using sequence-based methods and HTM

No. false Positives	BLAST		PSI-BLAST		HTM	
	E-value threshold	No. true positives[a]	*E*-value threshold	No. true positives[a]	*E*-value threshold	No. true positives[a]
5	1.5e−02	281 (14.2)	3.0e−15	481 (24.3)	5.0e−13	403 (20.4)
10	2.7e−02	293 (14.8)	2.0e−14	498 (25.2)	7.0e−09	455 (23.0)
50	1.1e−01	325 (16.4)	6.0e−08	609 (30.8)	8.0e−03	550 (27.8)
100	2.4e−01	352 (17.8)	3.0e−04	681 (34.4)	1.5e−01	590 (29.8)
200	5.0e−01	371 (18.7)	1.0e−02	744 (37.6)	1.9e+00	607 (30.7)
400	1.1e+01	396 (20.0)	1.2e−01	814 (41.1)	–[b]	– –

[a] The number of true positives is given, followed by the percentage of the total possible number of true positives (1,980) in parentheses.
[b] A total of only 284 false positives was found using HTM.

HTM finds additional matches

thresholds of 3.0e−04 and 1.5e−01, respectively (Table 1). However, while PSI-BLAST finds a larger number of correct matches than HTM, our primary interest is to discover whether or not we find **additional** matches (as defined above – 'Assessment of hit lists') using HTM, that are not found by PSI-BLAST. To investigate this, we can compare the specific hits found by each method. Table 2 lists the number of matches that were found by both methods, and the number found by only one, for various rates of false positives. This indicates that we do indeed find additional matches using HTM. For example, at a false positive rate of 50 for either PSI-BLAST or HTM, PSI-BLAST finds 609 true positives (498 + 111 in Table 2). HTM finds an additional 52 true positives, to give a total of 661 (8.5 per cent more than PSI-BLAST alone). Combining the hit lists in this way also leads, of course, to additional false positives.

The *E*-value thresholds for 100 false positives are 3.0e−04 and 1.5e−01 for PSI-BLAST and HTM, respectively. All hits in the HTM list with an *E*-value of less than or equal to 3.0e−04 must also be in the PSI-BLAST list. Thus, the additional hits in the HTM list are those with an *E*-value of between 3.0e−04 and 1.5e−01 (see Figure 5). This illustrates the function of the HTM process – to evaluate the pairs for which the PSI-BLAST *E*-value is too high for the relationship to be clear. If the threshold for the PSI-BLAST list had been set to

Table 2: Numbers of unique and common hits found using PSI-BLAST and HTM for different rates of false positives

No. false positives	*E*-value thresholds		Common hits		PSI-BLAST unique hits		HTM unique hits	
	PSI-BLAST	HTM	TP[a]	FP[a]	TP[a]	FP[a]	TP[a]	FP[a]
5	3.0e−15	5.0e−13	391	1	90	4	12	4
10	2.0e−14	7.0e−09	411	3	87	7	44	7
50	6.0e−08	8.0e−03	498	16	111	34	52	35
100	3.0e−04	1.5e−01	536	33	145	78	54	68
200	1.0e−02	1.9e+00	567	77	177	127	40	123

[a] The number of true positives (TP) and false positives (FP) found in common and by each method uniquely. The thresholds chosen were those at which the 5th, 10th, etc. false positive occurred. In some cases, several false positives were found with this threshold *E*-value. For this reason, the total number of false positives may be slightly higher than the specified number (for example, 33 + 78 = 111 for PSI-BLAST for the 100 false positive rate). The total number of true or false positives found by both methods **combined** can be calculated for any false positive rate by adding together the number of common hits and the unique hits – for example, 661 true positives and 85 false positives at the 50 false positive rate.

1.5e−01, rather than 3.0e−04, we would have found an extra 465 hits, of which 140 are true positives and 325 are false positives. After building and evaluating models of these 465 hits, using HTM, 54 that are true positives and 68 that are false positives were retained (Table 2). Therefore, 79 per cent of the additional false positives were eliminated, but also 61 per cent of the true positives. We are investigating ways in which the process can be improved so that even more of the false positives can be eliminated, while reducing the number of true positives that are eliminated (see below).

There is scope for improvement

Characteristics of homologues found using PSI-BLAST and HTM

The results of the PSI-BLAST and HTM calculations can be analysed to determine whether particular SCOP structural classes are favoured in the hits that are obtained (Table 3). For most classes approximately 40–60 per cent of the pairs in the test set are found by the combination of PSI-BLAST and HTM (see the 'All hits' column). A much smaller proportion of the pairs in the all-beta class is found, however. PSI-BLAST finds only 94 out of the 547 pairs (17 per cent). While a significant proportion of the unique HTM hits are in this class (16 out of 54), the total number of hits for the all-beta class is still only 110 (20 per cent). Potential reasons for this very low hit rate for all-beta proteins are being investigated and will be reported in a subsequent publication. A low hit rate for this fold class was also found by Müller *et al.*[27]

Only 17 per cent of all-beta-class pairs are found

A large proportion of the unique hits in the PSI-BLAST hit list – 98 out of 145 (68 per cent) – is for the SCOP class 1.003. This is also the most significant class for the unique hits in the HTM hit list – 23 out of 54 hits (43 per cent). Within this class, the 1.003.004.001 superfamily (SCOP version 1.37; the 'ferredoxin reductase-like, C-terminal NADP-linked domain') is significantly represented in the test set, with a total of 120 pairs between members of the superfamily. Of these 120 pairs, 47 are found by both PSI-BLAST and HTM, 44 are found by only PSI-BLAST and 1 by only HTM. It is instructive to consider the pairs that are found only by PSI-BLAST. Models are being constructed for these 44 correct pairs during HTM, but these models are being rejected – consideration of the reasons for this could suggest enhancements to the procedure.

One potential problem with HTM is that the sequence alignment used for building the model could be significantly incorrect so that, even though an

Table 3: Number of true positives found using PSI-BLAST and HTM for different SCOP classes, for the hit lists containing 100 false positives for each method

Class description	SCOP class	Total pairs in test set	Common hits[a]	PSI-BLAST unique hits	HTM unique hits	All hits[b]
All alpha	1.001	342	139	11	3	153 (45)
All beta	1.002	547	84	10	16	110 (20)
Alpha and beta (a/b)[c]	1.003	757	193	98	23	314 (41)
Alpha and beta (a + b)[d]	1.004	136	56	13	3	72 (53)
Multi domain (alpha and beta)	1.005	14	7	1	0	8 (57)
Membrane and cell surface	1.006	9	0	0	0	0 (0)
Small proteins	1.007	175	57	12	9	78 (45)
All classes		1980	536	145	54	735 (37)

[a] The number of hits found by both PSI-BLAST and HTM.
[b] The total number of hits found by PSI-BLAST or HTM or both. The percentage of the total number of pairs that are in the test set is given in parentheses.
[c] Mainly parallel beta sheets (beta–alpha–beta units).
[d] Mainly antiparallel beta sheets (segregated alpha and beta regions).

appropriate template is being used for constructing the model, the model will be inaccurate. Here, the alignment produced by PSI-BLAST is used.[28,29] Given the very low sequence similarity of many of the pairs in the test set, there is a significant chance that the alignment could be inaccurate. It could also be the case that, while the residue alignment could be correct, there could be large insertions or deletions in the sequence of the model relative to the template, or there could be other significant differences between the structure of the target and template. These problems could also give rise to an inaccurate model. Since, in this study, we know the correct (X-ray) structure of each target domain, we can estimate the accuracy of a model by performing a least squares superposition of the model onto its X-ray structure, followed by calculating the root mean square (RMS) deviation between the alpha carbon atoms of the model and the X-ray structure. A low deviation would indicate a model built using an accurate alignment. Calculated in this way, the RMS deviations for the models that are found only by PSI-BLAST range from 0.6 to 20.1 Å. Of the 44 models, 27 have a deviation of greater than 4 Å. Thus, many of the models are clearly being built from inaccurate alignments, or from alignments in regions where there are significant differences between the target and template structures. Methods that make use of the structure of the template in addition to its sequence in producing the alignment are available (ALIGN2D – Šali *et al.*, in preparation) and will be evaluated in the pipeline.

GeneAtlas can be applied to whole genomes

Inaccurate alignments are a problem

While 27 models have high RMS values when compared to their X-ray structures, 17 have an RMS of less than 4 Å. Why are these models being rejected? The models are built for the portion of the target protein that is covered by the PSI-BLAST alignment. These 17 models are all short – ranging from 27 to 57 residues, with an average of 33 residues. The X-ray structures of the targets, however, range in length from 126 to 322 residues, with an average of 261 residues. Thus the models, although accurate, represent only a small portion of the complete structure, and residues that should be buried in the core of the protein will be exposed on the surface of the model. These residues will be assigned a low score by Profiles-3D during the validation procedure, and this leads to the rejection of the models.

More detailed structural analysis gives additional functional information

Whole genome analysis using the GeneAtlas pipeline

The proteins encoded by a number of fully sequenced genomes have been processed using GeneAtlas (the complete pipeline, illustrated in Figure 1, was used for this work). The percentage of the genome that is annotated is given for several examples in Table 4. This table also illustrates how the percentage that can be annotated structurally increases as the size of the PDB database increases (see the results for *Mycoplasma genitalium* obtained with different versions of the PDB database).

DISCUSSION

This study has shown that the use of high-throughput modelling enables us to assign additional relationships beyond those that can be assigned using a purely sequence-based method. HTM, however, also leads to the introduction of additional false positives, and causes the rejection of valid relationships. Some of these rejected models are due to inaccuracies in the sequence alignments that are used to generate the models. This points to one possible area of improvement.

The methods that we are exploring are designed to identify proteins that have related structures, even at low levels of sequence similarity. In most cases, similarity of structure is likely to lead to clues about function. Proteins with the same fold can, however, have quite different functions.[30] In these cases, a more detailed analysis of predicted structures might lead to more conclusive information on function. This type of analysis might include the identification of

Table 4: Genome sequences annotated using GeneAtlas

Genome	No. of protein sequences	Structural annotations[a]	SCOP coverage[b]	Active site annotations[c]	Structural template annotations[d]	Sequence-only annotations[e]
Homo sapiens	29,304	14,979 (51.1%)	652	5,515	15,822	11,229 (38.3%)
Arabidopsis thaliana	25,571	15,334 (60.0%)	654	6,676	11,365	8,972 (35.1%)
Caenorhabditis elegans	19,835	10,962 (55.3%)	648	4,228	8,473	5,207 (26.2%)
Drosophila melanogaster	14,332	8,384 (58.5%)	646	3,665	7,466	5,854 (40.8%)
Mycoplasma genitalium[g]	468	328 (70.1%)	N/A[f]	N/A[f]	N/A[f]	109 (23.3%)
Mycoplasma genitalium[h]	468	273 (58.3%)	N/A[f]	N/A[f]	N/A[f]	163 (34.8%)

[a] Sequences for which a match to at least one structure in PDB95 was found.
[b] The total number of SCOP superfamilies represented in all models built for the genome.
[c] Sequences for which at least one functional assignment can be made as a result of a template (used to build a model for a region of the sequence) containing an active site definition (in the PDB file).
[d] For each sequence, the total number of occurrences of structural templates (3D patterns) in all models built for that sequence was counted. Where the same template was found more than once in the models for a sequence, it was counted only once. The number given is the sum over all sequences in the genome.
[e] Sequences for which no match to a structure was found, but for which a match to at least one sequence in nrdb was found.
[f] These data were not calculated.
[g] These results were obtained using the PDB95 database that was created based on the PDB release of 14th December, 2000, and that consists of 5,460 PDB chains extracted from 4,658 PDB structures.
[h] These results were obtained using the PDB95 database that was created based on the PDB release of November 1999 and that consists of 4,250 PDB chains.

patterns of residues that are known to confer a particular function or functions.[31] The PDB95 template structures have been annotated both to indicate the presence of active-site definitions from the original PDB files and to indicate the presence of patterns of residues, derived by the functional annotation using structural templates method (Milik *et al.*, in preparation), that are associated with a particular type of function. Models can be analysed to see if either of these types of features, where present in the template, are conserved in the model (Table 4).

In other cases, proteins of different folds may have the same function.[30] In this case a comparison of the overall structures would not permit a structural and functional relationship to be established. When it is possible to build models based on structurally similar templates, however, detailed analysis of the active sites might again permit functional relationships to be derived.

Acknowledgments
We thank Dr Andrej Šali for his input and his pioneering work in the development of automated modelling pipelines. We are grateful to Dr Scott Kahn and Prof. Rod Hubbard for their advice and to Dr Jong Park for helpful discussions and the provision of data. We thank SGI and Sun Microsystems for providing us with access to the large-scale computational facilities that were required for this work.

References

1. Bork, P., Dandekar, T., Diaz-Lazcoz, Y. *et al.* (1998), 'Predicting function: from genes to genomes and back', *J. Mol. Biol.*, Vol. 283, pp. 707–725.
2. Bork, P. and Koonin, E. V. (1998), 'Predicting functions from protein sequences – what are the bottlenecks?', *Nature Genet.,* Vol. 18, pp. 313–318.
3. Sánchez, R. and Šali, A. (1998), 'Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome', *Proc. Natl Acad. Sci. USA,* Vol. 95, pp. 13597–13602.
4. Fetrow, J. S., Godzik, A. and Skolnick, J. (1998), 'Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: Identification of proteins exhibiting the glutaredoxin/ thioredoxin disulfide oxidoreductase activity', *J. Mol. Biol.,* Vol. 282, pp. 703–711.
5. Pawłowski, K., Zhang, B., Rychlewski, L. and Godzik, A. (1999), 'The *Helicobacter pylori* genome: from sequence analysis to structural and functional predictions', *Proteins: Struct. Funct. Genet.,* Vol. 36, pp. 20–30.
6. Huynen, M., Doerks, T., Eisenhaber, F. *et al.* (1998), 'Homology-based fold predictions for

Mycoplasma genitalium proteins', *J. Mol. Biol.,* Vol. 280, pp. 323–326.

7. Fischer, D. and Eisenberg, D. (1997), 'Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*', *Proc. Natl Acad. Sci. USA,* Vol. 94, pp. 11929–11934.
8. Teichmann, S. A., Park, J. and Chothia, C. (1998), 'Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements', *Proc. Natl Acad. Sci. USA,* Vol. 95, pp. 14658–14663.
9. Jones, D. T. (1999), 'GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences', *J. Mol. Biol.,* Vol. 287, pp. 797–815.
10. Pennisi, E. (1998), 'Taking a structured approach to understanding proteins', *Science*, Vol. 279, pp. 978–979.
11. Accelrys Inc., San Diego, CA 92121, USA; URL: http://www.accelrys.com/
12. Park, J., Karplus, K., Barrett, C. *et al.* (1998), 'Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods', *J. Mol. Biol.*, Vol. 284, pp. 1201–1210.
13. Altschul, S. F., Madden, T. L., Schaffer, A. A. *et al.* (1997), 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic Acid Res.*, Vol. 25, pp. 3389–3402.
14. Holm, L. and Sander, C. (1998), 'Removing near-neighbour redundancy from large protein sequence collections', *Bioinformatics*, Vol. 14, pp. 423–429.
15. URL: http://www.accelrys.com/references/supplemental/GeneAtlas_2001paper_supp.html
16. Šali, A. and Blundell, T. L. (1993), 'Comparative protein modelling by satisfaction of spatial restraints', *J. Mol. Biol.,* Vol. 234, pp. 779–815.
17. Lüthy, R., Bowie, J. U. and Eisenberg, D. (1992), 'Assessment of protein models with three-dimensional profiles', *Nature*, Vol. 356, pp. 83–85.
18. Fischer, D. and Eisenberg, D. (1996), 'Protein fold recognition using sequence-derived predictions', *Protein Sci.,* Vol. 5, pp. 947–955.
19. Olszewski, K. A., Yan, L. and Edwards, D. J. (1999), 'SeqFold – fully automated fold recognition and modeling software – evaluation and application', *Theor. Chem. Acc.,* Vol. 101, pp. 57–61.
20. Kabsch, W. and Sander, C. (1983), 'Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features', *Biopolymers*, Vol. 22, pp. 2577–2637.
21. King, R. D., Saqi, M., Sayle, R. and Sternberg, M. J. E. (1997), 'DSC: public domain protein secondary structure prediction', *CABIOS*, Vol. 13, pp. 473–474.
22. Berman, H. M., Westbrook, J., Feng, Z. *et al.* (2000), 'The Protein Data Bank', *Nucleic Acids Res.,* Vol. 28, pp. 235–242.
23. National Center for Biotechnology Information; URL: http://www.ncbi.nlm.nih.gov/
24. 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.*, Vol. 247, pp. 536–540.
25. Brenner, S. E., Chothia, C. and Hubbard, T. (1998), 'Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships', *Proc. Natl Acad. Sci. USA,* Vol. 95, pp. 6073–6078.
26. Altschul, S. F., Gish, W., Miller, W. *et al.* (1990), 'Basic local alignment search tool', *J. Mol. Biol.*, Vol. 215, pp. 403–410.
27. Müller, A., MacCallum, R. M. and Sternberg, M. J. E. (1999), 'Benchmarking PSI-BLAST in genome annotation', *J. Mol. Biol.*, Vol. 293, pp. 1257–1271.
28. Sauder, J. M., Arthur, J. W. and Dunbrack, R. L. (2000), 'Large-scale comparison of protein sequence alignment algorithms with structure alignments', *Proteins: Struct. Funct. Genet.,* Vol. 40, pp. 6–22.
29. Friedberg, I., Kaplan, T. and Margalit, H. (2000), 'Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments', *Protein Sci.,* Vol. 9, pp. 2278–2284.
30. Hegyi, H. and Gerstein, M. (1999), 'The relationship between protein structure and function: a comprehensive survey with application to the yeast genome', *J. Mol. Biol.,* Vol. 288, pp. 147–164.
31. Russell, R. B., Sasieni, P. D. and Sternberg, M. J. E. (1998), 'Supersites within superfolds. Binding site similarity in the absence of homology', *J. Mol. Biol.,* Vol. 282, pp. 903–918.