Genome analysis: Assigning protein coding regions to three-dimensional structures

ASAF A. SALAMOV,¹ MAKIKO SUWA,¹ CHRISTINE A. ORENGO,²
AND MARK B. SWINDELLS^{1,3}

¹Helix Research Institute, 1532-3 Yana, Kisarazu-shi, Chiba, 292, Japan

²Biomolecular Structure and Modelling Unit, Department of Biochemistry, University College London,
Gower Street, London, United Kingdom

³Inpharmatica Ltd., 60 Charlotte Street, London W1P 2AX, United Kingdom

(RECEIVED August 31, 1998; ACCEPTED January 6, 1999)

Abstract

We describe the results of a procedure for maximizing the number of sequences that can be reliably linked to a protein of known three-dimensional structure. Unlike other methods, which try to increase sensitivity through the use of fold recognition software, we only use conventional sequence alignment tools, but apply them in a manner that significantly increases the number of relationships detected. We analyzed 11 genomes and found that, depending on the genome, between 23 and 32% of the ORFs had significant matches to proteins of known structure. In all cases, the aligned region consisted of either >100 residues or >50% of the smaller sequence. Slightly higher percentages could be attained if smaller motifs were also included. This is significantly higher than most previously reported methods, even those that have a fold-recognition component. We survey the biochemical and structural characteristics of the most frequently occurring proteins, and discuss the extent to which alignment methods can realistically assign function to gene products.

Keywords: CATH; genome analysis; sequence search; structure analysis

The advent of complete genome sequencing has emphasized the need for comparable developments in automated analysis. Currently, methods for analyzing large quantities of data limit their analysis to the proteins encoded by each open reading frame (ORF), as these are of primary interest to the majority of researchers. Each encoded protein may consist of any combination of the following: globular domains, transmembrane regions, or repetitive structures such as a leucine rich repeats. In principle, all of these regions are of interest, but in practice most detailed knowledge is limited to globular domains. To obtain the most appropriate (though still tentative) functional assignment for each region, a series of database searches are applied so that when a well-annotated homologue is identified, the functional annotation can be transferred.

The first attempt to provide a completely automated procedure for genome analysis was GeneQuiz (Casari et al., 1995). More recently a competitive system called Pedant has also been released (Frishman & Mewes, 1997). Both systems categorize the hits to globular domains into those matching (with decreasing utility); a sequence of known three-dimensional (3D) structure, a homologue with a clear functional annotation, or a sequence of unknown function. Knowing the approximate 3D structure of a protein is

obviously attractive, particularly when combined with the functional data from all homologues. As a result, extra effort is often given to maximizing the number of such hits.

In GeneQuiz, five genomes have currently been analyzed and 8–12% of the ORFs have at least one region of their sequence linked to a domain of known structure. Pedant now has 10 genomes analyzed in which 3D structures are available for 10–16% of the sequences. In both cases, the best results come from the *Mycoplasma genitalium* genome and the improvement shown by Pedant may result in part from increases in database size over the two years that separate these analyses. Both of these systems use only standard sequence based procedures such as fasta (Pearson & Lipman, 1988) and blast (Altschul et al., 1990).

Two further methods were limited to *M. genitalium*. Fischer and Eisenberg used standard sequence based searches to assign 16% of the ORFs to a known structure, followed by fold recognition to assign a further 6% (Fischer & Eisenberg, 1997). More recently, a new method that compares PSI-blast created profiles has been reported to assign 38% of the *M. genitalium* ORFs to known structures (Rychlewski et al., 1998). These results will be discussed in our paper.

We are interested in maximizing the number of relationships that can be identified on the basis of sequence analysis alone because these always precede, and optimally preclude, the use of slower fold-recognition procedures (Bowie et al., 1991; Jones et al., 1992).

Reprint requests to: Mark Swindells, Inpharmatica Ltd., 60 Charlotte Street, London W1P 2AX, United Kingdom; e-mail: m.swindells@inpharmatica.co.uk.

Furthermore, we want to provide a method that can be immediately applied to all finished genomes, rather than a proof of concept on the smallest available genome.

Our paper describes an automated system that uses developments in protein structure classification such as the CATH database (Orengo et al., 1997), as well as more sophisticated sequence search strategies that use intermediate sequences (Gribskov, 1992; Koonin & Tatsuov, 1994; Brenner et al., 1998) and profile based methods (Altschul et al., 1997) to maximize the number of sequences that are linked to a known structure. We first describe how safe cutoffs were determined for the selected sequence alignment methods. Then we show how these methods were applied to 11 microbial genomes, and finally we discuss the structural and functional characteristics of the most frequently occurring proteins.

Results and discussion

Application to 11 genomes

In Table 1 we show the results of applying our system to 11 genomes. For each genome we show how many ORFs could be linked to a Protein Data Bank (PDB) file by searching either the PDB, gp_PDB, or psi_PDB databases. Because of the vast amount of data generated, we limit the discussion here to some of the more significant trends. However, all detailed data are available as an internet solution (<http://www.hri.co.jp/genius>) in the manner summarized by Figure 1. In this manner, information is available for each hit, when both the PDB/intermediate and intermediate/gene scores are beyond the required thresholds. We place particular emphasis on the subset of results in which the aligned region is either >50% of the smaller sequence (gene or PDB sequence) or >100 residues long, as these are most likely to represent evolutionary relationships.

Our best results always come from using the psi_PDB database, where the results are more than double those of GeneQuiz and Pedant. For example, we can assign nearly 32% of the *M. genitalium* ORFs to a region of known structure, while still requiring

that >100 residues or >50% of the smaller protein are aligned. We use *M. genitalium* here for comparison as all groups have performed calculations on this genome. This compares favorably with the 1996 results from GeneQuiz of 12%, the most recent results from Pedant server (generated in November 1997) of 16% and the frsvr server (Fischer & Eisenberg, 1997), where 16% were detected by sequence alignment with an additional 6% from fold recognition. Although Rychlewski et al. report a higher sensitivity (38%), one should note that they ran PSI-Blast with a threshold that we feel cannot be universally applied ($E < 0.1$) without additional checks. Furthermore they did not require a large aligned region and did not mask membrane sequences.

The worst performing genomes are perhaps not surprisingly from archaea: *Methanococcus jannaschii* and the gram negative bacterium; *Helicobacter pylori*, which has a large number of membrane proteins and has adapted to living in a low pH environment. Both have only 22% of their ORFs assigned, using the same criteria as above. Nevertheless, even with these less successful results included, the average assignment level over 11 genomes is 25.6%. This again is more than twice the average result from Pedant (currently the only competitive system for which more than five analyzed genomes are available).

What proteins are hit the most?

Although it is not practical to look at the results on a case by case basis, it is valuable to look at the structures that are hit most frequently. Twenty-one proteins are hit by more than 50 open reading frames each; nine of these are single domain entries, whereas 12 are multidomain (Table 2). When we look at the functions of these domains, taking into account which were hit in a multidomain structure, we find that half bind to phosphate containing coenzymes (such as ATP, GTP, PLP, and FAD). Reasons for this bias can be gleaned from an analysis of known structures, where despite the vast array of enzymes, coenzyme functionality is most frequently provided by one of only a small set of distinct domains. Therefore, it would seem likely that these domains that have been

Table 1. Results for searches on 11 genomes

Genome	Total ORFs	Average sequence length	PDB database				gp_PDB database				psi_PDB database			
			Any hit above threshold		>50% or >100 residue overlap		Any hit above threshold		>50% or 100 residue overlap		Any hit above threshold		>50% or 100 residue overlap	
			Number	%	Number	%	Number	%	Number	%	Number	%	Number	%
<i>M. genitalium</i>	467	364	94	20.1	91	19.5	108	23.1	99	21.2	162	34.7	148	31.7
<i>Mycoplasma pneumoniae</i>	677	351	102	15.1	97	14.3	116	17.1	108	16.0	174	25.7	157	23.2
<i>Borrelia burgdorferi</i>	850	333	138	16.2	132	15.5	155	18.2	143	16.8	224	26.4	203	23.9
<i>H. pylori</i>	1,565	317	218	13.9	206	13.2	262	16.7	240	15.3	398	25.4	353	22.6
<i>M. jannaschii</i>	1,715	282	184	10.7	160	9.3	273	15.9	227	13.2	464	27.1	384	22.4
<i>Haemophilus influenzae</i>	1,717	303	330	19.2	308	17.9	382	22.2	348	20.3	573	33.4	503	29.3
<i>Methanobacterium thermoautotrophicum</i>	1,869	281	244	13.1	215	11.5	325	17.4	275	14.7	529	28.3	445	23.8
<i>Archaeoglobus fulgidus</i>	2,407	275	301	12.5	286	11.9	408	17.0	359	14.9	701	29.1	601	25.0
<i>Synechocystis</i>	3,169	326	475	15.0	441	13.9	606	19.1	534	16.9	935	29.5	810	25.6
<i>Bacillus subtilis</i>	4,099	297	677	16.5	643	15.7	805	19.6	742	18.1	1,199	29.3	1,069	26.1
<i>E. coli</i>	4,289	317	755	17.6	718	16.7	936	21.8	847	19.7	1,332	31.1	1,178	27.5
Total	22,824	306	3,518	15.4	3,297	14.4	4,376	19.2	3,922	17.2	6,691	29.3	5,851	25.6

Fig. 1. Examples of the results and how they can be viewed at <http://www.hri.co.jp/genius>. In the Netscape window, each ORF from a genome is listed together with the intermediate sequence, PDB target, the region hit in the PDB target, the hit length, and percent overlap. In the RasMol window, the protein chain is shown originally in white and then masked successively with aligned portions of the intermediate in green and ORF in red. The amount of intermediate colored green, therefore, depends on how much of the ORF is subsequently aligned. In the example, regions of the ORF and intermediate are similar, so most of the intermediate ends up being colored red. In contrast the PDB structure has an extra domain (white) that is not covered at all by the search.

continually reused during evolution, through repeated genetic rearrangement. With the exception of the most frequently hit protein CheY (a signal transducer phosphorylated by CheA), the remaining proteins are predominantly enzymes.

In structural terms, the results are unusual. Our analysis reveals that not only do all the single domain proteins belong to the $\alpha\beta$ class of structures, but so do nearly all of the domains hit within multidomain proteins (Table 2). This was not anticipated, as typically a nonhomologous database contains only 50% in the $\alpha\beta$ class, with the remainder divided equally between mainly α and mainly β structures (Orengo et al., 1997). Although an elevated fraction of $\alpha\beta$ structures might be anticipated due to the high frequency of coenzyme binding domains, which themselves have a high preference (about 80%) for the $\alpha\beta$ class (Martin et al., 1998), this still does not explain the prevalence observed here.

Hits that do not pass the requirement of covering either >50% of the smaller sequence or >100 residues tend to correspond to motifs. In *Escherichia coli* for example, some of the ABC transporter sequences (for which a structure is not yet known) hit part of the mitochondrial ATPase (PDB code: 1bmfE) but the overlap is less than 100 residues, as well as being less than 50% of the smaller sequence. The ABC transporters are of wide interest, prin-

cipally due to the connection between mutations in this gene and cystic fibrosis. Four structure predictions (Hyde et al., 1990; Mimura et al., 1991; Anneraeu et al., 1997; Hoedemaeker et al., 1998) have been published for the ATP binding domain, but significantly each one is different. Analysis of our data shows that the aligned region surrounds the P-loop, which is responsible for binding mononucleotide phosphate. While it is almost certainly true that the ABC transporters use a P-loop to bind ATP, many of the P-loops in known structures may have arisen from convergent evolution, as their overall topologies can be quite different (e.g., adenylate kinase, elongation factor-Tu, and RecA). Therefore, in our opinion, there is currently no clear evidence for a homologous relationship between ABC transporters and a protein of known structure.

Transferring functional annotations

Accuracy of the function assigned to a sequence ultimately depends on the similarity it has with well-annotated sequences. While close homologues can safely have their annotations transferred (assuming that the primary annotation was correct), this becomes more problematic as the evolutionary distance between the proteins increases. For example, while the relationship between ATP

Table 2. Most frequently hit protein structures

PDB code	Number of hits	Description	Bound Hetatm/ability to bind	CATH number(s) ^a from version 1.4
3chy0	154	Signal transduction, phosphorylated by cheA	PO4	3.40.50.300
1dar0	136	Elongation factor G	GTP (Walker type motif)	3.40.50.300 2.40.90.70 ^b 2.20.29.10 ^b 3.30.230.10 3.30.70.240
1ak50	119	Inosine monophosphate dehydrogenase	SO4	Not processed, $\alpha\beta$
1vid0	109	Catechol o-methyltransferase		3.40.50.500
1gpmD	108	GMP synthetase	AMP	3.40.50.880 3.40.50.610 3.30.300.10
1fds0	111	Hydroxysteroid dehydrogenase		3.40.50.300
2recE	85	RecA	ATP (Walker)	3.40.560.10
1ajsA	97	Aspartate aminotransferase	PLP derivative	3.30.70.160 3.40.640.10
2admA	79	Adenine methyltransferase		3.40.50.150 3.90.220.10
1bme0	84	Beta lactamase		3.30.449.10 1.10.500.10 ^c
1al30	86	CysB cofactor binding fragment	SO4	Not processed, $\alpha\beta$
1lci0	78	Luciferase (Acyl-coA ligase superfamily)		3.40.50.980 3.40.50.980 2.30.35.10 3.40.31.10
1broA	76	Haloperoxidase		3.40.50.950
1gdhA	66	Glycerate dehydrogenase	SO4	3.40.50.300 3.40.50.300
1bncA	67	Biotin carboxylase	ATP	3.40.50.20 3.30.150.10 3.10.210.10
1aq6A	64	Haloacid dehalogenase		3.40.50.1000 1.10.164.10
1fcdA	66	Flavocytochrome C sulfide dehydrogenase	FAD	3.40.370.20 3.40.370.40
1dad0	64	Detheobiotin synthase	ATP (Walker)	3.40.50.470
1pir0	65	DNA helicase (homology to/structure like, RecA)	ATP (Walker)	Not processed, $\alpha\beta$
1poxA	56	Pyruvate oxidase	FAD + thiamine diphosphate	3.40.50.970 3.40.50.70 3.40.50.970
1gln0	51	Glutamyl tRNA synthase	ATP	3.40.50.990 3.90.40.10 1.10.150.10 1.10.10.100

^aCATH numbers have four components for class, architecture, topology, and homologous superfamily. Class 1 = mainly α , Class 2 = mainly β , Class 3 = $\alpha\beta$.

^bAligned regions do not cover these domains.

^cMay be more suitably described as a single $\alpha\beta$ domain.

binding domains of DNA gyrase and Heat Shock Protein-90 can be detected by PSI-Blast, transferring their annotations would not be advantageous. Because so many of the relationships described in this paper rely on the detection of distant similarities, we do not describe our results in terms of predicted function.

Future enhancements

Our automatically generated results are extremely competitive, even without requiring fold recognition techniques. Genomes can be calculated quickly and updated at regular intervals. In our future

versions, we hope to increase the sensitivity by taking full advantage of all the sequences available from sequencing projects, rather than the current approach which only uses sequences from OWL, during the PSI-Blast step. We will also try to derive extra sensitivity by running the described process in the “reverse” direction, so that PSI-blast profiles are built for all genome sequences and then each PDB sequence is compared against the resulting set of intermediates.

Materials and methods

Assessing sensitivity and specificity with the CATH classification

To determine optimal combinations of algorithm, matrix, and threshold, it is necessary to have a suitable list of true and false relationships with which to perform the assessment. We use CATH (Orengo et al., 1997), a comprehensive, hierarchical classification of structures from the PDB (Bernstein et al., 1977). Each homologous superfamily (H-level) in CATH contains proteins whose relationships can either be identified by sequence searching, or when the sequence similarity is too low, structure comparison coupled with functional analysis. Eight hundred and forty-one single domain proteins with less than 95% sequence identity were taken from version 1.1. These are distributed over 241 homologous superfamilies, but 157 of them have only a single entry and play their role by increasing the chance of false relationships being predicted. There are 3,442 true relationships contained within 84 homologous superfamilies with which to assess sensitivity and 349,778 unrelated pairs with which to establish thresholds for reliable database searching. A file containing these relationships is available from <http://www.hri.co.jp/~swintech/841.list>

Sequence masking

All sequences were masked to take account of membrane regions, coiled coils, and low complexity regions by using memsat (Jones et al., 1994), paircoil (Berger et al., 1995), and seg (Wooton, 1994), respectively. Parameters, where required, were set to default values.

Pairwise alignment programs and matrices

We assessed fasta (Pearson & Lipman, 1988) and gapped-blast (Altschul et al., 1997), the only two algorithms sufficiently fast for database searching with two matrices, blosum62 and blosum50 (Henikoff & Henikoff, 1992), in this work. Parameters, where required, were set to the default values. For gapped-blast the gap opening and extension penalties were -10 and -1 , respectively. For fasta these were -12 and -2 .

We tried various ways to normalize the raw score, but the most effective (irrespective of the algorithm/matrix combination) was one that took into account the lengths of both sequences, and the extreme distribution of alignment scores identified by Karlin and Altschul (1990). The normalized score S is calculated as

$$S = s - (\ln(Knm)/\lambda) \quad (1)$$

where s is the raw score, n and m are the lengths of the sequences being aligned, and λ and K are constants taken from Altschul and Gish (1996). Specifically these were $\lambda = 0.216$ and $K = 0.014$ for blosum62, $\lambda = 0.158$ and $K = 0.019$ for blosum50. This scheme is

independent of database size but is different to the bit score in gapped-blast.

Assessments were carried out by first searching all 841 CATH entries against the Owl sequence database (which contains over 200,000 entries) and generating a list of hits for each domain. Then for each scoring system described above, we took the 349,778 *unrelated* pairs, compared the lists of Owl hits for each pair, and when a sequence occurred in both lists, recorded the lower of the two alignment scores (as this was the score at which the query and target sequence would be incorrectly linked by an intermediate). After repeating this task for all 349,778 unrelated pairs, we found the highest recorded score that would simultaneously link two unrelated proteins via an intermediate. For fasta, the best performing matrix was blosum62, whereas for gapped-blast it was blosum50 (though the difference with other combinations was only marginal). For these combinations, the thresholds beyond which no false positives occurred were 88 for fasta and 18 for gapped-blast. Use of an Owl intermediate is a vital part of the test, as it provides a much better estimate of the appropriate thresholds for safe database searching. Without OWL, the thresholds would have appeared to be 77 and 14, respectively.

Sensitivity was assessed by determining how many of the true relationships occur above the threshold. Again we allowed sequences to be linked via an intermediate sequence from Owl, in an analogous manner to the procedure for setting the threshold. This use of an intermediate to link related proteins is essentially the same as the test devised by Park et al. (1997), except that we added the criterion that alignment overlap must be greater than 50% of the smallest protein. To quantify the merits of this approach, we also calculated the sensitivity without using an intermediate. Note, however, that the stricter set of thresholds were still used. These results are shown in Figure 2. Although the actual percentages are specific to the set of proteins that we used, the important points to note are that (1) gapped-blast performs essentially as well as fasta with the added advantage of being about four times quicker and (2) both these methods benefit from the inclusion of intermediate sequences.

Assessing PSI-blast

PSI-Blast comparisons were run twice, first with the E-value threshold set to 0.001 and then using 0.01. Other parameters were identical to those used for gapped-blast. We took each sequence from

Fig. 2. Increases in sensitivity, assessed with our 841 protein dataset. Use of fasta and gapped-blast in direct comparisons (red) is compared with the inclusion of an OWL intermediate (green) and finally the execution of PSI-blast on a database that includes the 841 dataset as well as the OWL database.

our dataset and ran it against a database consisting of both OWL and our 841 sequences. A maximum of 10 iterations was assigned for 815 sequences (all but 26 immunoglobulins) and 665 of these converged within that limit. Because of the enormous number of sequences identified at each step by the 26 immunoglobulins, each iteration was taking up to 2 h (and therefore almost a whole day if 10 iterations were allowed), and so we limited these to only four iterations. The problem appears to occur not at the search step but during matrix construction; PSI-blast removes proteins with more than 98% sequence identity from consideration, but we assume that this approach does not sufficiently reduce the number of immunoglobulin sequences.

We found that E-value thresholds of 0.01 and 0.001 both gave a specificity of over 99%. Some of the remaining errors resulted from the alignment of short motifs and could be removed by requiring the aligned region to be either >50% of the smaller sequence or >100 residues. These criteria were used during genome analysis as described.

Effect of single and multidomain proteins in threshold assessment

It is important to describe here our principles for dealing with domains. The thresholds for each algorithm were originally determined using a subset of single domain proteins. This was for two reasons: first, it made the process of separating related and unrelated proteins easier, and did not rely on our subjective division of protein structures into domains that had not been experimentally verified. Second, it helps us to determine a conservative threshold, above which unrelated sequences are unlikely to occur. This is because of the way it considers sequence length when modifying the raw score and is best explained by way of example. Consider two sequences, one containing only a single SH3 domain and another containing an SH3 domain linked to an SH2. Although the raw score would be identical, the normalized score would be lower than if the multidomain protein were split up and the two SH3 regions aligned by themselves. Now consider two unrelated pro-

teins. In a similar manner, the normalized scores of chance alignments will also be lower when the sequences used become longer, thus making the threshold for real database searching safer.

Making a flexible system for genome analysis

Our procedure combines fast, pairwise sequence alignment techniques, with the benefits of intermediate sequences and profiles. To combine these into an effective method for genome analysis, we use the following steps (also see Fig. 3):

1. All sequences from PDB were compared with one another in a pairwise manner using fasta (gapped-blast could have been used but fasta program was operationally more convenient). Sequences having scores above a normalized score of 88 and at least 80% overlapping with one another, were clustered into families and a representative selected. This is the nonredundant or representative PDB dataset.
2. Using PSI-blast, each representative PDB sequence was run against a version of the OWL database. All aligned regions (i.e., not the complete sequence) from sequences below $E = 0.001$ threshold were stored together with information about which PDB sequence was associated with each aligned region. In practice, all proteins removed in (1) reappear here, but this approach is more efficient as it reduces the number of time consuming PSI-Blast runs required. We will refer to this database as psi_PDB. To quantify the benefits of using psi_PDB later on we also made an equivalent gp_PDB database using the results of a gapped-blast search.
3. For each genome we took the sequence of every open reading frame and using the gapped blast algorithm searched the resulting sequence against three different databases: PDB, gp_PDB, and psi_PDB. Because of the way that gp_PDB and psi_PDB were made, whatever sequence is hit, a link can be immediately made to a region of a known structure. A normalized score of 20 was used for additional safety. The increase is equivalent to

Fig. 3. Summary of method used for genome analysis. First, a database containing sequences with homology to each PDB sequence is constructed (procedure from left to right). Then each ORF is scanned against this enlarged database, which we call psi_PDB.

50% of the difference between the two thresholds 14 and 18, derived in the absence, and then presence of intermediates.

Effects of single and multidomain proteins in genome analysis

Although our initial assessments of alignment tools used single domain proteins, we do not split multidomain proteins into their constituents for genome analysis. Our reasons for this are: (1) Despite significant work in the field there are still no computational techniques that reliably identify domains in all structures (Jones et al., 1998). (2) Verified domains from CATH could theoretically be used, but they often split a protein on structural rather than evolutionary considerations and this may be disadvantageous when searching for distant homologues. (3) The database searches used here are based on local alignments and therefore are not adversely affected by the presence of multiple domains. (4) As explained previously, the thresholds are equally applicable to multidomain proteins and single domain proteins.

References

- Altschul SF, Gish W. 1996. Local alignment statistics. *Methods Enzymol* 266:460–480.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
- Anneraeu J-P, Wulbrand U, Vankeerberghen A, Cuppens H, Bontems F, Tümmeler B, Cassiman J-J, Stoven V. 1997. A novel model for the first nucleotide binding domain of the cystic fibrosis transmembrane conductance regulator. *FEBS Lett* 407:303–308.
- Berger B, Wilson DB, Wolf E, Tonchev T, Milla M, Kim PS. 1995. Predicting coiled coils by use of pairwise residue correlations. *Proc Natl Acad Sci USA* 92:8259–8263.
- Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer based archival file for macromolecular structures. *J Mol Biol* 122:535–542.
- Bowie JU, Lüthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three dimensional structure. *Science* 253:164–169.
- Brenner SE, Chothia C, Hubbard TJE. 1998. Assessing sequence comparison methods with reliable structurally-identified distant evolutionary relationships. *Proc Natl Acad Sci USA* 95:6073–6078.
- Casari G, Andrade MA, Bork P, Boyle J, Daruvar A, Ouzounis C, Schneider R, Tamames J, Valencia A, Sander C. 1995. Challenging times for bioinformatics. *Nature* 376:647–648.
- Fischer D, Eisenberg D. 1997. Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proc Natl Acad Sci USA* 94:11929–11934.
- Frishman D, Mewes H-W. 1997. PEDANTIC genome analysis. *Trends Genet* 13:415–416.
- Gribskov M. 1992. Translational initiation factors IF-1 and eIF- α share an RNA binding motif with prokaryotic ribosomal protein S1 and polynucleotide phosphorylase. *Gene* 119:107–111.
- Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 89:10915–10919.
- Hoedemaeker FJ, Davidson AR, Rose DR. 1998. A model for the nucleotide binding domains of ABC transporters based on the large domain of aspartate aminotransferase. *Proteins* 30:275–286.
- Hyde SC, Emsley P, Hartshorn MJ, Mimmack MM, Gileadi U, Pearce SR, Gallagher MP, Gil DR, Hubbard RE, Higgins CF. 1990. Structural model of ATP binding proteins associated with cystic fibrosis, multi-drug resistance and bacterial transport. *Nature* 346:366–368.
- Jones DT, Taylor WR, Thornton JM. 1992. A new approach to fold recognition. *Nature* 358:86–89.
- Jones DT, Taylor WR, Thornton JM. 1994. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* 33:3038–3049.
- Jones S, Stewart M, Michie A, Swindells MB, Orengo C, Thornton JM. 1998. Domain assignment for protein structures using a consensus approach: Characterization and analysis. *Protein Sci* 7:233–242.
- Karlin S, Altschul SF. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA* 87:2264–2268.
- Koonin E, Tatuov RL. 1994. Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity. *J Mol Biol* 244:125–132.
- Martin ACR, Orengo CA, Hutchinson EG, Jones S, Taroni C, Thornton JM. 1998. Protein folds and functions. *Structure* 6:875–884.
- Mimura CS, Holbrook SR, Ames GF-L. 1991. Structural model of the nucleotide binding conserved component of periplasmic permeases. *Proc Natl Acad Sci USA* 88:84–88.
- Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH—A hierarchic classification of protein domain structures. *Structure* 5:1093–1108.
- Park J, Teichmann SA, Hubbard T, Chothia C. 1997. Intermediate sequences increase the detection of homology between sequences. *J Mol Biol* 273:349–354.
- Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 85:2444–2448.
- Rychlewski L, Zhang B, Godzik B. 1998. Fold and function predictions for *Mycoplasma genitalium* proteins. *Folding Design* 3:229–238.
- Wootton JC. 1994. Non-globular domains in protein sequences: Automated segmentation using complexity measures. *Comput Chem* 18:269–285.