

GenTHREADER: An Efficient and Reliable Protein Fold Recognition Method for Genomic Sequences

David T. Jones*

Department of Biological
Sciences, University of
Warwick, Coventry CV4 7AL
UK

A new protein fold recognition method is described which is both fast and reliable. The method uses a traditional sequence alignment algorithm to generate alignments which are then evaluated by a method derived from threading techniques. As a final step, each threaded model is evaluated by a neural network in order to produce a single measure of confidence in the proposed prediction. The speed of the method, along with its sensitivity and very low false-positive rate makes it ideal for automatically predicting the structure of all the proteins in a translated bacterial genome (proteome). The method has been applied to the genome of *Mycoplasma genitalium*, and analysis of the results shows that as many as 46% of the proteins derived from the predicted protein coding regions have a significant relationship to a protein of known structure. In some cases, however, only one domain of the protein can be predicted, giving a total coverage of 30% when calculated as a fraction of the number of amino acid residues in the whole proteome.

© 1999 Academic Press

Keywords: genome; protein structure prediction; fold recognition; threading; sequence alignment

The recent complete sequencing of a number of microbial genomes has highlighted the gap between the number of known protein sequences and the number of experimentally determined protein structures. Methods for protein structure prediction offer some hope for narrowing this gap, and over recent years fold recognition methods have become particularly popular (Bowie *et al.*, 1991; Jones *et al.*, 1992, 1995; Godzik *et al.*, 1992; Bryant & Lawrence, 1993; Ouzounis *et al.*, 1993; Abagyan *et al.*, 1994; Nishikawa & Matsuo, 1994; Flöckner *et al.*, 1995; Lathrop & Smith, 1996; Madej *et al.*, 1995; Fischer Eisenberg, 1996; Defay & Cohen, 1996; Russell *et al.*, 1996). Blind testing has shown that fold recognition methods can be very effective (Shortle, 1997), and so it is surprising that they are not being more widely applied to genome analysis. Three problems with fold recognition methods probably contribute to their lack of use: their slowness, the requirement for human intervention to interpret the results and the inaccuracy of sequence-structure alignments produced. Differ-

ent methods suffer from each of these problems to differing degrees. Of the three problems, the lack of automation in the fold recognition process is perhaps the biggest problem in the application of threading methods to genomic sequence analysis. Whilst it is reasonable to require some human intervention when predicting the structure of just a few sequences, this is clearly not practical when trying to analyse many thousands of genomic sequences.

Here, a new method for fold recognition is described, which has the advantage that it is both very fast and requires no human intervention in the prediction process. Surprisingly, despite its simplicity, the method also produces relatively accurate sequence-structure alignments.

Recognizing Distant Evolutionary Relationships

Threading methods were originally developed to recognise pairs of proteins which have no obvious similarities in sequence yet have similar folds. Recently it has become clear that in fold recognition it is useful to distinguish between pairs of proteins which are homologous (i.e. have obvious common ancestry) and those which are analogous

Abbreviations used: ORF, open reading frame; 5GP, guanosine-5'-monophosphate.

E-mail address of the corresponding author:
jones@globin.bio.warwick.ac.uk

(have no obvious common ancestry; Orengo *et al.*, 1994; Jones, 1997). Where no evolutionary relationship is believed to exist between two structurally similar proteins, clearly threading is going to be the only applicable method for identifying this type of relationship, but for pairs of proteins that do share common ancestry, it might be supposed that sensitive sequence comparison methods might be applicable. Pairwise sequence comparison methods are generally assumed only to be able to recognise closely related sequences, and to overcome this limitation, matching methods have been developed that use information from multiply aligned sequences of protein families (Taylor, 1986; Gribskov *et al.*, 1987; Krogh *et al.*, 1994; Lüthy *et al.*, 1994; Yi & Lander, 1994).

The limits of pairwise sequence alignment in identifying very distant sequence relationships is well known, and has been quantified in various studies (Sander & Schneider, 1991; Abagyan & Batalov, 1997). However, what is surprising is that even though these methods are not useful for identifying very remote sequence similarities, they are often capable of producing accurate sequence-structure alignments.

The GenTHREADER Protocol

The method for fold recognition described here can be divided into three stages: alignment of sequences, calculation of pair potential and solvation terms and, finally, evaluation of the alignment using a neural network. A program implementing the following method has been developed (called GenTHREADER) and can be accessed from the following Web page: <http://globin.bio.warwick.ac.uk/psipred>.

Sequence alignment

Given a target sequence and a template protein structure, a sequence-structure alignment can be made using a wide variety of techniques. Here, alignments are generated using a sequence profile method, though in principle almost any sequence alignment method could be used to generate the initial alignments. To generate profiles for each template structure in the fold library, related sequences were collected by scanning the template sequence against the current OWL non-redundant protein sequence data bank (Bleasby *et al.*, 1994) using the program BLASTP (Altschul *et al.*, 1990). Sequences matching the template sequence with an E -value < 0.01 were extracted from the data bank and aligned using a simplified version of the MULTAL multiple sequence alignment method (Taylor, 1988). Insertions relative to the template protein sequence were skipped, and a sequence profile constructed (Gribskov *et al.*, 1987) using the BLOSUM 50 matrix (Henikoff & Henikoff, 1992).

The fold library used was based on the set of unique protein chains found in the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977) as of

January 1998, though excluding theoretical models and short peptides. Target sequences are aligned with a given template structure using a global-local dynamic programming alignment algorithm (i.e. a global alignment algorithm with no end gap penalties), with an initiation gap penalty of 11 and an extension penalty of 1.

As an alternative to this, it is possible to build a sequence profile from the target protein sequence and to scan this profile against the fold library sequences. In some cases this produces a better result than the first approach, and in practice both approaches are used with the highest scoring alignment being taken as the preferred one.

Threading potentials

Given a sequence alignment, the next step is to evaluate this alignment with reference to the implied structural model. Several methods have been described for evaluating structural models by using statistical potentials of some description (e.g. Jones & Thornton, 1996; Hendlich *et al.*, 1990; Kocher *et al.*, 1994; Park *et al.*, 1997; Miyazawa & Jernigan, 1996). The evaluation function used here is principally based on a set of pairwise potentials of mean force (Hendlich *et al.*, 1990), determined by a statistical analysis of highly resolved protein X-ray crystal structures and the application of the inverse Boltzmann equation as described for the original THREADER program (Jones *et al.*, 1992). In addition to the pairwise potentials, a solvation potential is also used (Jones *et al.*, 1992).

For specified atoms ($C^{\beta} \rightarrow C^{\beta}$ for example) in a pair of residues ab , sequence separation k and distance interval s , the potential is given by the expression:

$$\Delta E_k^{ab} = RT \ln(1 + m_{ab}\sigma) - RT \ln\left(1 + m_{ab}\sigma \frac{f_k^{ab}(s)}{f_k(s)}\right)$$

where m_{ab} is the number of pairs ab observed with sequence separation k , σ is the weight given to each observation, $f_k(s)$ is the frequency of occurrence of all residue pairs at topological level k and separation distance s , $f_k^{ab}(s)$ is the equivalent frequency of occurrence of residue pair ab , and RT is taken to be 0.582 kcal/mol. Here, short (sequence separation, $k < 11$), medium ($11 \leq k \leq 22$) and long ($k > 22$) range potentials have been calculated between the following atom pairs: $C^{\beta} \rightarrow C^{\beta}$, $C^{\beta} \rightarrow N$, $N \rightarrow C^{\beta}$, $C^{\beta} \rightarrow O$, $O \rightarrow C^{\beta}$.

In addition to the pairwise potentials, a solvation potential for an amino acid residue a is defined as:

$$\Delta E_{solv.}^a(r) = -RT \ln\left(\frac{f^a(r)}{f(r)}\right)$$

where r is the degree of residue burial, $f^a(r)$ is the frequency of occurrence of residue a with burial r , and $f(r)$ is the frequency of occurrence of all residues with burial r . The degree of burial for a residue is defined as the number of other C^{β} atoms

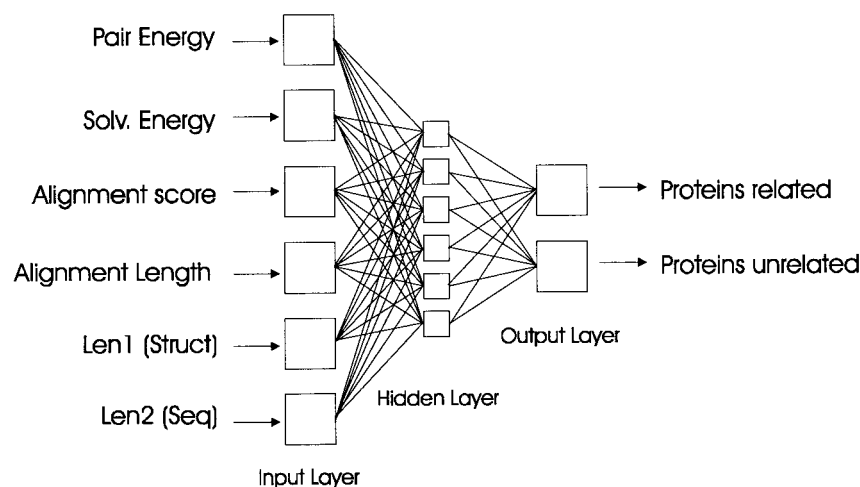


Figure 1. Diagrammatic representation of the neural network architecture.

located within 10 Å of the residue's C^β atom. This parameter was found to inversely correlate well (correlation coefficient >0.85) with the relative solvent accessibility of a residue in a folded protein.

Neural network

Rather than producing just a single score, methods for fold recognition results commonly produce a number of scores which relate to different aspects of the sequence-structure alignment. In the case of the method described here, six scores were found to be important in different circumstances: initial sequence profile alignment score, number of aligned residues, length of target sequence, length of template protein sequence, pairwise energy sum, and solvation energy sum. A poor value for one or more of these scores can indicate that the sequence-structure match is a false-positive. The problem that needs to be addressed is how to reduce the vector of six (or possibly more in future developments of the method) different scores with quite different value ranges into just a single score. This is a fairly common problem in multivariate statistics, of course, with many methods available for achieving this goal. However, most of these methods are not capable of taking account of cross-variable relationships and are limited to producing a scoring function which is a weighted linear combination of the original variables. A simple and popular technique which is not restricted in this way is to make use of a multilayer feedforward neural network. Neural networks are capable of "learning" quite complex interrelationships between multiple variables and reducing this complexity to a single output value (typically between 0 and 1).

For this reason, a simple feedforward neural network was employed here to evaluate sequence-structure alignment quality. A single network input was used for each of six parameters (sequence profile score, pairwise energy, solvation energy, number of aligned residues, length of tem-

plate structure and length of target sequence) as shown in Figure 1. The inputs to the network were scaled to the required 0-1 range by using the standard logistic function:

$$\frac{1}{1 + e^{-a(x-b)}}$$

where a and b are arbitrary constants, and x is the raw input value.

Because of the small number of inputs, only a few hidden units were required in the hidden layer, and whilst the performance of the network was slightly better when six hidden units were used, networks with four or five hidden units fared almost as well.

Neural network training

To train the neural network, a set of representative protein chains was extracted from the CATH database (Orengo *et al.*, 1997). A single representative was selected for each T-level of the CATH classification containing more than one S-level (i.e. fold families containing at least two homologues with low sequence similarity). For each pair of representative chains, the CAT (Class Architecture and Topology) numbers of any constituent domains were compared to decide whether any significant structural similarity exists between the chains.

On the then current version of CATH (Release 1.0), the procedure resulted in a training set comprising 9169 chain pairs, of which 383 pairs shared a common domain fold. For each of these chain pairs a sequence-structure alignment was calculated using the previously described profile method, and pairwise and solvation energies calculated based on the equivalent residue pairs. Based on the six sequence-structure alignment variables previously described, the challenge for the neural network is to discriminate between chain pairs which share a common fold and those which do not.

Neural network weights

Neural network methods are frequently criticised for being “black box” solutions. In other words, although they work very well, it is not possible to find out what “rules” they have learned from the input data. This is generally true, but it is nonetheless fairly easy to interrogate a simple network such as the one used here to find out how the variables are combined together to produce the single output value. One way to achieve this is to hold certain inputs at constant values and to vary the remaining variables to explore how these variables affect the output. This is equivalent to cutting a slice through the multidimensional space represented by the weights in the neural network. Figure 2(a) and (c) shows three representative plots made in just this way. Figure 2(b) is particularly interesting, because it shows the learned relationship between the sequence profile score and the pairwise energy term when the other terms are held constant. As expected, the network has learned that with a relatively high sequence similarity score, the value of the pair energy sum is immaterial and can be more or less ignored. However, even at high sequence similarity, where the pair energy is particularly unfavourable (indicating that the implied structural model is very likely to be incorrect) the network output drops way down.

Confidence estimation

Although the output from the neural network is a value between 0 and 1, this does not imply that the result can be directly taken as a probability value. In order to better interpret the output of the neural network, the same CATH data set described above (9169 chain pairs, of which 383 pairs shared a common domain fold) was used to test the performance of the method and to assign levels of confidence to different output levels from the network. The original set of protein chains from CATH was divided into three parts. Each third of the data set was used as a testing set, and the remaining two thirds used as a training set. In this way the performance of the network could be tested without undue bias from the training set.

The three sets of results were pooled and sorted according to the network output value. The frequency of false positives was then calculated for different ranges of network output (Figure 3). From these test results, it was decided that the output of the network could be interpreted according to Table 1. Of course, no statistical measurement can ever truly assign a certain level of confidence, i.e. where the false positive rate is zero. However, in this case it is reasonable to define a certain category as that where the neural network produces a maximum output value (1.0). Out of 126 pairs of protein chains which produce a maximum network value, no false positives were observed. Generally speaking, the results which are assigned as certain are those which exhibit some degree of

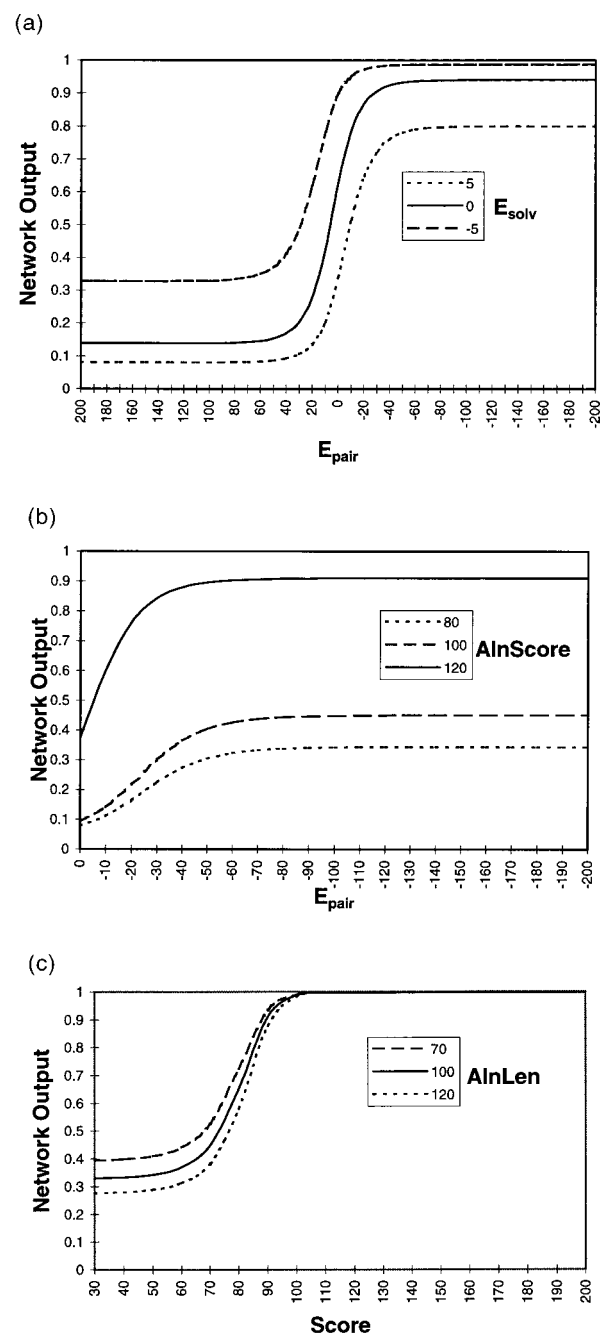


Figure 2. Input/output graphs for the trained neural network shown in Figure 1. (a) Network output is plotted against pairwise energy (E_{pair}) for three different solvation energy sums (E_{solv}). All other network inputs were set to 0.5. (b) Network output is plotted against pairwise energy for three different alignment score values. (c) Network output is plotted against alignment score for three different alignment lengths.

sequence similarity (even though this similarity may be statistically insignificant). The results falling into the lower categories are those which have been recognised not principally by sequence similarity, but by favourable pairwise and solvation energy sums.

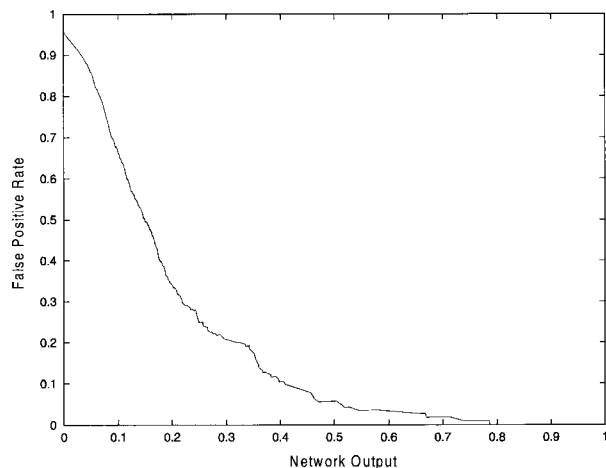


Figure 3. Rate of false positive assignments plotted against neural network output.

The “high” category threshold corresponds to an expected false positive rate of just 1%, which is a common high confidence threshold for genome annotation. Indeed, the type of testing that has been carried out here on GenTHREADER, being based on a large number of protein pairs of known three-dimensional structure, is comparable to the testing that has been carried out elsewhere on sequence comparison methods (Brenner *et al.*, 1998).

Results

Benchmarking

To test the ability of the method to identify distant evolutionary relationships it was applied to a test-set suggested by Fischer *et al.* (1996). This benchmark comprises a set of 68 pairs of proteins with very low sequence similarity, but with highly similar folds. Success for a particular target sequence is achieved if a fold recognition method ranks the correct matching template protein at rank 1. Note that matches to protein chains in the library which have a similar fold to the correct match are not considered false positives (in this work the CATH classification is used to make this determination). Table 2 shows the results obtained from the benchmark test. Note that the target protein (plus any proteins with the same CATH numbers) was removed from the set of proteins used to

construct the statistical potentials and the training set for the neural network. For a sequence-based method, the results in Table 2 are impressive. Using the criteria described by Fischer & Eisenberg (1997), 73.5% of the test cases are correctly recognised (i.e. no incorrect folds have a better score than the expected match). The best method that Fischer & Eisenberg (1997) have tested to date has scored 76.5% on this benchmark, but this method directly incorporates predicted structural information in its formulation. As comparison points, a pairwise sequence alignment algorithm using the standard 250 PAM mutation data matrix achieves a true-positive rate of 51% (other scoring matrices achieve rates of up to 62%), and the best sequence profile method, 63% (results taken from the Web pages of D. Fischer, <http://www.doe-mbi.ucla.edu/people/fischer/BENCH/benchmark1.html>).

Although the true-positive rate achieved by GenTHREADER is comparable with other fold recognition methods, the main feature of the method is its ability to clearly discriminate between true and false positives. A total of 33 of the benchmark targets (48.5%) are given “certain” scores (i.e. with confidence > 99%), and all of these 33 targets are correctly recognised. Furthermore, for 20 of these targets, more than 50% of the alignment is in agreement with the correct structural alignment. In contrast, 23 targets are given a score < 0.5 which implies that the predictions are more likely to be incorrect than correct. In fact, 15 of these 23 targets are incorrectly predicted. This demonstrates that GenTHREADER has exactly the right combination of properties for reliable automatic genome annotation. It has high sensitivity (i.e. high true-positive rate) and a very low false-positive rate.

Alignment accuracy

Although the recognition of a related fold is a necessary first step towards building a three-dimensional model for a target protein, an accurate model can only be obtained if an accurate sequence-structure alignment can be produced. Given that all the examples shown in Table 2 comprise protein pairs with very low sequence similarity (average of 18.6% sequence identity), and that no structural information is incorporated in the alignment algorithm, it might be expected that the accuracy of the alignments will be very poor. Generally it is found that fold recognition methods produce very inaccurate alignments (e.g. Russell

Table 1. Estimated confidence categories for predictions based on Figure 3

Category	Network score cutoff	Total in test set	False positives	Confidence (%)
Certain	1.000	129	0	100
High	0.790	204	2	99
Medium	0.340	243	60	80

These confidence limits are based on cross-validated trials using 383 pairs of proteins which share a common fold and are believed to have common ancestry.

Table 2. Table of results for the benchmark set of 68 protein pairs

Fold	Target	Expected	% ID	Result	Net	CATH								
						Rank	rank	E_{pair}	E_{solv}	AlnSc	Alen	Len1	Len2	AlnCorr
Open sheet	1mioc	1minb	16	1minb	1.000	1	1	-238.2	10.6	952	379	522	525	24.0
IG-fold	2fbjl	8fabb	22	8fabb	1.000	1	1	-185.5	-13.8	836	206	223	213	76.3
Open sheet	2cmd	6ldh	23	6ldh	1.000	1	1	-394.9	-14.3	707	293	329	312	68.4
TIM barrel	1chra	2mnr	20	2mnr	1.000	1	1	-344.3	-8.8	701	333	357	370	51.0
Peroxidase	2hpda	2cpp	18	2cpp	1.000	1	1	-403.9	-14.1	628	375	405	457	46.2
Peroxidase	1lgaa	2cyp	16	2cyp	1.000	1	1	-190.8	-6.7	607	272	293	343	51.2
Globin-like	1dxtb	1hbg	19	1hbg	1.000	1	1	-197.7	-7.2	606	137	147	147	78.9
Open sheet	1gal	3cox	18	3cox	1.000	1	1	-156.3	1.6	435	440	502	581	16.1
EF-hand	1osa	4cpv	24	4cpv	1.000	1	1	-195.6	-8.3	420	141	174	148	62.6
Open sheet	1npx	3grs	20	3grs	1.000	1	1	-337.5	4.7	403	364	461	447	43.2
Lipocalin	1mup	1rbp	14	1rbp	1.000	1	1	-85.3	-7.3	398	152	175	157	45.3
Ribonuclease-H	1hrha	1rnh	24	1rnh	1.000	1	1	-159.8	-6.6	341	122	151	130	67.4
Jelly roll	1bbt1	2plv1	20	2plv1	1.000	1	1	-30.0	4.1	314	182	297	186	26.9
Lipocalin	1mdc	1lfc	21	1lfc	1.000	1	1	-83.1	-7.2	295	127	131	131	73.1
Open sheet	2pia	1fnr	18	1fnr	1.000	1	1	-170.5	-12.4	270	226	296	321	53.2
Cytochrome	1c2ra	1ycc	23	1ycc	1.000	1	1	-35.5	-3.21	229	97	108	116	79.2
SH2	2pna	1shaa	29	1shaa	1.000	1	1	-70.9	-9.0	218	89	103	104	61.4
EF-hand	2sas	2scpa	17	2scpa	1.000	1	1	-92.6	-2.0	216	128	174	185	37.5
Jelly roll	1caub	1caua	18	1caua	1.000	1	1	-110.1	-1.0	213	151	181	184	44.6
Cupredoxin	1aaj	1paz	31	1paz	1.000	1	1	-84.1	-5.0	210	89	99	105	68.7
Hydrolase	1taha	1tca	16	1ede	1.000	3	1	-211.5	1.0	209	255	317	318	31.6
Ribonuclease	1onc	7rsa	26	7rsa	1.000	1	1	-103.8	-4.7	200	99	124	104	65.9
Small	1hip	2hipa	19	2hipa	1.000	1	1	-4.7	-4.9	186	69	71	85	57.7
Open sheet	1gky	3adk	24	3adk	1.000	1	1	-133.6	-6.9	182	162	194	186	40.5
IG-fold	1fc1a	2fb4h	19	2fb4h	1.000	1	1	-183.8	-2.8	168	193	229	207	45.6
Thioredoxin	1aba	1ego	21	1ego	1.000	1	1	-50.7	0.3	142	71	85	87	50.5
Ferredoxin	5fd1	2fxb	21	2fxb	1.000	1	1	-25.1	-3.8	128	72	81	106	47.3
IG-fold	1pfc	3hlab	22	3hlab	1.000	1	1	-100.8	-7.4	118	94	99	111	70.6
4-Helix bundle	1bbha	2ccya	21	2ccya	1.000	1	1	-177.6	-5.2	87	101	106	131	57.9
Trypsin	1arb	4ptp	20	4ptp	1.000	1	1	-60.5	1.0	11	195	223	263	41.3
Cupredoxin	1afna	1aoza	19	1paz	0.998	6	1	-238.8	8.6	114	294	359	331	62.3
Porin	2omf	2por	17	2por	0.997	1	1	-82.0	8.8	244	293	449	340	43.0
IG-fold	3hlab	2rhe	15	2rhe	0.994	1	1	-55.1	-3.5	54	88	114	99	57.2
UB fold	1fxia	1ubq	18	1ubq	0.958	1	1	-45.0	-3.5	52	71	76	96	22.8
Cytochrome	2mtac	1ycc	15	451c	0.943	2	1	-43.6	-1.4	63	78	82	147	41.1
Small	1isua	2hipa	16	2hipa	0.928	1	1	-9.4	-2.0	44	57	71	62	43.6
IG-fold	3cd4	2rhe	25	8faba	0.850	2	1	-85.3	-4.7	70	94	114	178	61.4
TIM barrel	3rubl	6xia	18	1avha	0.836	8	8	-279.6	-7.0	66	274	318	447	0.0
Helix bundle	1aep	256ba	14	2spca	0.802	4	4	-132.9	-4.9	45	126	129	153	0.0
Open sheet	1eaf	4cla	21	4cla	0.787	1	1	-202.6	6.5	36	222	274	243	51.1
Open sheet	3chy	4fxn	14	1cola	0.742	14	14	-130.9	-3.3	55	82	89	128	0.0
Open sheet	1ak3a	1gky	24	1gky	0.690	1	1	-134.0	-2.5	13	179	186	226	36.4
Trypsin	2sga	4ptp	21	14ptp	0.687	1	1	-77.4	4.9	34	161	223	181	36.1
Cupredoxin	2azaa	1paz	11	1paz	0.677	1	1	-45.0	-0.8	49	91	120	129	9.6
Trefoil fold	8ilb	4fgf	18	4fgf	0.634	1	1	-56.8	-3.4	43	111	124	146	38.5
Actin	1atna	1atr	15	1atr	0.278	1	1	-197.0	11.8	14	321	383	372	12.5
Jelly roll	1saca	1ayh	14	8faba	0.224	98	5	-133.9	2.1	78	166	206	204	0.0
Jelly roll	4sbva	2tbva	19	2tbva	0.218	1	1	-25.1	-3.81	102	72	81	106	43.6
IG-fold	1tlk	2rhe	24	2rhe	0.217	1	1	-40.7	-2.9	25	79	114	103	52.6
Mixed	2hhma	1fbpa	13	1trb	0.162	9	9	-199.2	9.5	110	252	330	272	0.0
Globin-like	1cpcl	1cola	17	1hsla	0.140	7	7	-189.2	-5.4	11	132	147	172	0.0
TIM barrel	2mnr	4enl	18	1gox	0.134	>100	1	-311.6	1.2	46	338	436	357	27.0
Thioredoxin	1dsba	2trxa	13	256ba	0.131	10	10	-61.0	-0.2	32	133	318	188	0.0
OB fold	1ltsd	1bova	19	1lfc	0.130	6	6	-94.8	-1.2	28	96	204	103	0.0
DNA-binding (HTH)	1hom	1lfb	19	1lfb	0.109	1	1	-14.5	0.4	47	61	78	68	23.7
4-helix bundle	1rcb	1gmfa	21	1rpra	0.084	5	5	-89.6	-2.2	5	112	121	129	0.0
Monellin	1stfi	1mola	8	1aak	0.065	>100	>100	-22.9	2.6	41	97	151	98	0.0
IG-fold	1cid	2rhe	13	3cd4	0.061	>100	1	-77.6	-2.8	23	169	206	177	18.7
IG-fold	1ten	3hhrb	18	3b5c	0.060	>100	13	-49.8	-2.8	24	79	131	89	0.0
Hydrolase	1crl	1ede	17	4aaha	0.051	>100	>100	-116.8	6.7	4	469	571	534	0.0
Open sheet	2gbp	2liv	16	1sto	0.045	9	9	-58.7	1.3	61	171	213	309	0.0
Trypsin	2snv	4ptp	15	9rnt	0.040	12	12	-60.5	-4.9	45	100	107	151	0.0
Monellin	1cewi	1mola	10	8atcb	0.023	>100	>100	-45.0	-0.7	11	83	146	108	0.0
4-Helix bundle	1bgeb	1gmfa	12	1gmfa	0.022	1	1	-167.2	1.7	4	143	164	165	17.3
Thioredoxin	1gp1a	2trxa	17	2trxa	0.020	1	1	-107.4	-2.8	4	130	162	185	25.9
Beta propellor	1sim	1nsba	12	saaib	0.019	>100	>100	-93.8	-2.1	120	198	262	381	0.0
Trefoil fold	1tie	4fgf	14	1f3g	0.019	>100	>100	-51.3	-1.0	68	119	150	170	0.0
Alpha + beta	2sara	9rnt	12	2msbb	0.015	5	5	-40.9	-0.1	46	76	113	96	0.0

Fold, fold of target protein; Target, target protein chain; Expected, best possible match on the basis of structural similarity; % ID, percentage identical residues between target protein and expected match; Result, matched chain from fold library; Net, strength of prediction (network output); Rank, rank of expected match; CATH rank, rank of next best match with similar fold (i.e. has identical CATH code to the expected match); E_{pair} , pairwise energy sum for predicted fold; E_{solv} , solvation energy sum for predicted fold; AlnSc, sequence alignment score; Alen, number of aligned residues; Len, length of template protein sequence; Len2, length of target protein sequence; AlnCorr, number of alignment positions in exact agreement with a reference structural alignment calculated using the SSAP method (Orengo *et al.*, 1992).

et al. 1996; Lemer *et al.*, 1995), often with none of the sequence-structure alignment agreeing with an alignment produced from a superposition of the three-dimensional structures. For related pairs of proteins, however, fold recognition can produce accurate alignments (Jones, 1997), and this is just what is observed in this case. Despite the fact that a simple sequence alignment method is used to generate the alignments, for 22 of the 68 test proteins in Table 2, the alignment accuracy is over 50%. In other words, for these 22 proteins, more than half of the alignment is in exact agreement with a structural alignment. The average alignment accuracy is 46.2% for all the cases where the correct fold is recognised. Also, on inspection of the alignments, the regions of the alignments which tend to be most accurately aligned are those which correspond to functionally important segments of the proteins (see, for example, the two examples detailed later).

Genome analysis

Given the high degree of reliability of the GenTHREADER algorithm, it is quite straightforward to apply it to automatic genome annotation. Most fold recognition methods require a great deal of human input in order for reliable predictions to be made. Although these methods have proven to be very effective, this reliance on human interpretation skills makes them unsuitable for processing genomic data. To see how effective GenTHREADER is in automatically annotating a genome, the *Mycoplasma genitalium* genome (Fraser *et al.*, 1995) was taken as an example. This is the smallest bacterial genome, with only 468 open reading frames (ORFs). Using a Silicon Graphics Origin200 server (2 R10 K processors) this analysis took one day to complete, including the time required to build the three-dimensional models of each sequence-structure match and to format the output into HTML tables so that the results could be viewed with a standard Web browser.

An overview of the results is shown in Figure 4 and Table 3, where it can be seen that a total of 42% of the ORFs can be linked to a protein of known structure in the highest levels of confidence (certain and high), with 34% falling into the "certain" category. In addition to the matches listed in Table 3, at least a further 18 ORFs, whilst incorporating transmembrane segments, also include ATP-binding cassette (ABC) domains (Yoshida & Amano, 1995; Annerau *et al.*, 1997) which match several ATP and GTP binding proteins in the fold library. These additional ABC-containing ORFs are listed in Table 4, and brings the total of predicted ORFs to 46%. A further six matches fall into the medium confidence bracket, which would increase the total to 47%, but it is not clear from the functional annotations of these six ORFs how correct these predictions are, despite the expected 80% confidence that is expected from the benchmarking for predictions in this category.

Interestingly, the distribution of architectures in the folds assigned to the genome is very similar to that observed in the Brookhaven PDB as a whole (considering all chains with <35% sequence identity). This suggests that contrary to popular belief, the folds in PDB might well constitute a representative set, and may not be biased towards particular folds or families.

Despite the high coverage of ORFs listed in Table 3, it is also necessary to consider the coverage when considered as a percentage of the total number of residues in each predicted protein. Whereas 46% of ORFs seem to have a significant relationship to a protein domain of known structure, these relationships only account for 30% of the total number of amino acid residues in the translated ORFs. This is easily explained by the observation that some of the significant matches are only to a single domain of the target sequence. For example, ORF MG002 shows a highly significant match to Brookhaven PDB (Abola *et al.*, 1987) entry 1HDJ, but this entry represents only a small (yet functionally well-defined) domain of 77 residues, which accounts for only 25% of the entire translated ORF. Nevertheless, for 166 of the 468 ORFs (35%), the proposed model accounts for more than 50% of the chain. Although some of the matches shown in Table 3 can easily be detected by sequence comparison alone, a significant number of the high confidence hits cannot be detected when the energy components are not used. Table 5 lists the fold assignments which are not ranked in first place when the energy terms (pairwise and solvation terms) are not presented to the neural network.

ORF MG276: an adenine phosphoribosyltransferase

As a straightforward example of a structural annotation for an ORF in *M. genitalium*, Figure 5 shows the best matching fold (PDB entry 1HGX) for ORF MG276. Despite the low overall sequence similarity (10% sequence identity), the annotation in this case is given a very high confidence by GenTHREADER. The data bank annotation for ORF MG276 is that of an adenine phosphoribosyltransferase (based on very high sequence similarity to the *Escherichia coli* protein), and this is in clear agreement with the function of the predicted structure, 1HGX, which is a hypoxanthine-guanine-xanthine phosphoribosyltransferase (Figure 5(a)).

Despite the functional plausibility of a structural resemblance between MG276 and 1HGX, to confidently assign MG276 to the same superfamily as 1HGX it is vital to verify the conservation of functionally important residues. Figure 5(b) shows a diagrammatic representation (ref) of the interactions between residues in 1HGX and the bound guanosine-5'-monophosphate (5GP). Only four residues make side-chain hydrogen bonds to 5GP (Thr107, Thr110, Lys134 and Tyr156), and these are all conserved in the alignment shown in Figure 5(c).

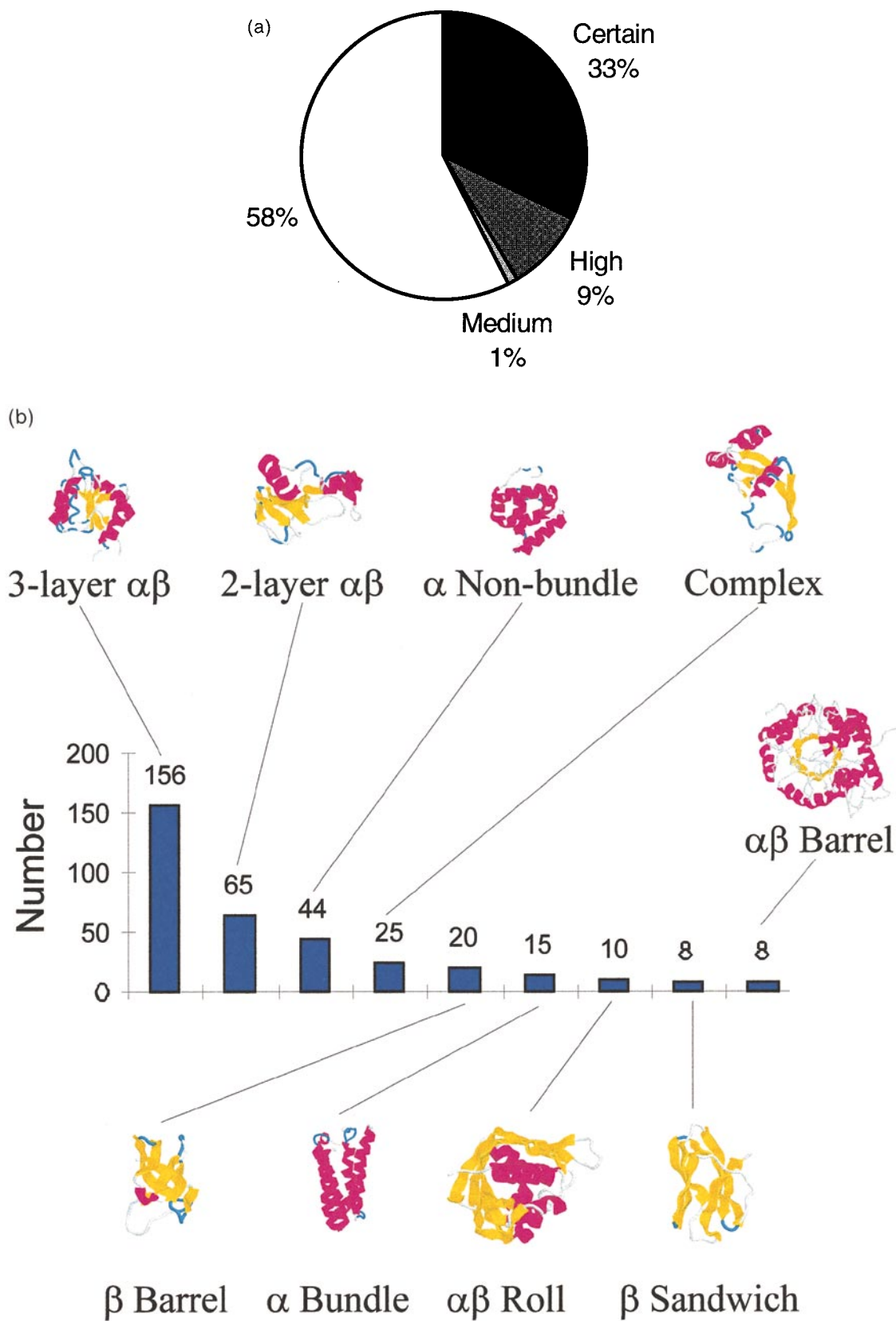


Figure 4. (a) The coverage of high confidence predictions is shown as a pie chart, according to the confidence categories shown in Table 1. The numbers of predictions were as follows: 154 “certain”, 43 “high” and six “medium”. (b) The distribution of protein domain architectures (35) in the predicted protein structures. As expected, the most commonly occurring domain architecture is the three-layer $\alpha\beta$ sandwich, followed by the two-layer $\beta\alpha$ sandwich.

Table 3. Table of high confidence predictions for *M. genitalium*

ORF	Category	Net	E_{pair}	E_{solv}	AlnSc	Alen	Len1	Len2	PDB code	Description
MG001	Cert	1.000	-258.0	-16.2	507	251	366	267	2POL-A	Pol III (beta subunit)
MG002	Cert	1.000	-50.5	-4.5	261	73	77	310	1HDJ	Human hsp40 fragment: j-domain (hdj-1)
MG003	Cert	1.000	-183.0	6.5	608	235	681	650	1BGW	Topoisomerase fragment: residues 410-1202
MG004	Cert	1.000	-359.0	0.6	411	469	681	836	1BGW	Topoisomerase fragment: residues 410-1202
MG005	Cert	1.000	-423.6	-23.5	881	413	421	417	1SES-A	Seryl-tRNA synthetase (serine-tRNA ligase)
MG006	High	0.999	-75.3	3.8	109	186	186	210	1GKY	Guanylate kinase complex with guanosine monophosphate
MG007	High	0.915	-132.1	1.4	82	204	220	254	1AK2	Adenylate kinase isoenzyme-2 (ATP:AMP phosphotransferase)
MG008	Cert	1.000	-187.5	-4.1	144	171	180	442	1HUR-A	Human ADP-ribosylation factor 1 (harf1:arf1:marf1)
MG011	Cert	1.000	-335.1	-10.9	177	264	306	287	1IOV	D-Ala:D-Ala ligase
MG012	Cert	1.000	-244.9	-0.7	286	201	316	225	2GLT	Glutathione biosynthetic ligase (glutathione synthase)
MG013	High	0.801	-177.0	-1.4	71	242	320	273	1GDH-A	D-Glycerate dehydrogenase (apo form)
MG019	Cert	1.000	-71.5	-10.3	302	77	77	389	1HDJ	Human hsp40 fragment: j-domain (hdj-1)
MG020	Cert	1.000	-320.4	-9.7	250	272	277	308	1BRO-A	Bromoperoxidase a2 (haloperoxidase a2: chloroperoxidase a2)
MG021	Cert	1.000	-322.5	-6.0	435	440	468	512	1GLN	Glutamyl-tRNA synthetase
MG023	High	0.884	-224.6	0.6	104	217	265	288	1WSY-A	Tryptophan synthase
MG024	Cert	1.000	-193.1	2.0	284	164	166	367	5P21	c-H-ras p21 protein (amino acid residues 1-166)
MG029	Cert	1.000	-114.4	4.2	229	164	501	186	1GPM-A	GMP synthetase (XMP aminase)
MG030	Cert	1.000	-212.1	-4.0	189	158	173	206	1HGX-A	Hypoxanthine-guanine-xanthine phosphoribosyltransferase (hgxpptase)
MG035	Cert	1.000	-464.7	-17.3	614	354	370	414	1HTT-A	Histidyl-tRNA synthetase (histidine-tRNA ligase) histidyl-adenylate
MG036	Cert	1.000	-427.9	-2.0	813	436	489	550	1LYL-A	Lysyl-tRNA synthetase (lysu)lysine
MG038	Cert	1.000	-601.2	-20.1	993	496	496	508	1GLA-F	Glycerol kinase complex with glycerol and the glucose-specific factor iii
MG039	Cert	1.000	-359.8	-3.8	394	330	340	384	1AA8-A	D-Amino acid oxidase
MG041	Cert	1.000	-116.1	-6.1	287	87	87	88	1PTF	Histidine-containing phosphocarrier protein (hpr)
MG045	Cert	1.000	-263.9	-13.1	273	288	322	483	1POT	Spermidine putrescine-binding protein
MG047	Cert	1.000	-440.0	-5.6	1149	369	383	383	1MXA	S-Adenosylmethionine synthetase
MG048	Cert	1.000	-239.0	0.2	235	196	224	446	1DAD	Dethiobiotin synthetase (dtbs)
MG049	Cert	1.000	-444.6	-14.9	654	233	237	320	1ECP-A	Purine nucleoside phosphorylase
MG050	High	0.840	-184.8	0.5	48	150	154	223	1RVV-A	Riboflavin synthase
MG051	Cert	1.000	-516.4	-16.6	1043	418	440	421	1TPT	Thymidine phosphorylase
MG052	Cert	1.000	-159.0	-5.1	290	119	294	130	1CTT	Cytidine deaminase (cda) complexed with 3:4-dihydrozebularine (dhz)
MG053	Cert	1.000	-570.6	-9.1	880	497	561	550	3PMG-A	Alpha-D-glucose-1:6-bisphosphate (phosphoglucomutase)
MG057	Medium	0.728	-90.6	2.3	43	109	119	178	1SRR-A	Sporulation response regulatory protein (spo0f) Mutant
MG058	Cert	1.000	-275.7	3.5	360	282	465	297	1GPH-1	Glutamine phosphoribosylpyrophosphate (prpp) amidotransferase
MG066	Cert	1.000	-911.4	-37.9	1221	641	678	648	1TRK-A	Transketolase
MG069	Cert	1.000	-131.6	-10.4	385	155	158	908	1GPR	Glucose permease (domain iia)
MG071	Cert	1.000	-211.9	6.0	189	202	220	874	1JUD	L-2-Haloacid dehalogenase
MG081	Cert	1.000	-101.1	-2.3	249	75	76	137	1FOW	L11-C76 fragment: carboxyl-terminal domain of protein L11
MG082	Cert	1.000	-307.7	-16.7	687	221	224	226	1AD2	Ribosomal protein L1 (tL2) mutant
MG083	Cert	1.000	-284.8	-11.4	276	168	193	189	2PTH	Peptidyl-tRNA hydrolase
MG084	Cert	1.000	-247.6	-5.9	447	253	501	290	1GPM-A	GMP synthetase (XMP aminase)
MG088	Cert	1.000	-171.9	-8.8	652	154	155	155	1HUS	Ribosomal protein S7
MG089	Cert	1.000	-671.5	-32.2	1528	628	631	688	1DAR	Elongation factor G (EF-G)
MG090	Cert	1.000	-139.4	-2.9	286	94	97	208	1RIS	Ribosomal protein S6
MG091	Cert	1.000	-90.9	-0.1	146	101	110	160	1KAW-A	Single-stranded DNA binding protein
MG093	Cert	1.000	-212.0	-8.2	405	145	149	150	1DIV	Ribosomal protein L9
MG094	Cert	1.000	-261.6	-0.1	168	253	326	446	2REB	RecA protein
MG097	Cert	1.000	-236.8	-10.8	594	223	228	245	1UDG	Uracil-DNA glycosylase
MG101	High	0.865	-103.6	-3.2	40	131	131	222	1OCT-C	Oct-1 (pou domain)
MG102	Cert	1.000	-355.1	-12.3	357	291	316	315	1TRB	Thioredoxin reductase
MG104	Cert	1.000	-32.1	-4.7	214	74	76	725	1SRO	Pnpase fragment: S1 RNA binding domain
MG106	Cert	1.000	-93.0	-0.9	394	145	147	226	1DEF	Peptide deformylase fragment

continued overleaf

Table 3—Continued

ORF	Category	Net	E_{pair}	E_{solv}	AlnSc	Alen	Len1	Len2	PDB code	Description
MG107	Cert	1.000	-320.1	-15.4	528	184	186	259	1GKY	Guanylate kinase complex with guanosine monophosphate
MG109	Cert	1.000	-235.2	-2.4	313	259	277	362	1PHK	Phosphorylase kinase fragment: gamma subunit
MG110	Cert	1.000	-86.6	2.7	214	122	177	236	1ETU	Elongation factor Tu (domain i) guanosine diphosphate complex
MG112	High	0.988	-178.3	0.2	51	186	247	209	1IGS	Indole-3-glycerol phosphate synthase (igps)
MG113	Cert	1.000	-508.8	-13.7	816	437	489	456	1LYL-A	Lysyl-tRNA synthetase (lysu)lysine
MG118	Cert	1.000	-502.2	-22.6	458	329	338	340	1NAH	UDP-galactose 4-epimerase (epimerase) biological unit: homodimer
MG122	Cert	1.000	-504.6	-26.4	1509	565	566	709	1ECL	<i>Escherichia coli</i> topoisomerase I (<i>Escherichia coli</i> omega protein)
MG124	Cert	1.000	-113.2	-5.7	243	98	105	102	3TRX	Thioredoxin (reduced form)
MG125	High	0.797	-169.6	-0.5	18	180	220	285	1JUD	L-2-Haloacid dehalogenase
MG126	Cert	1.000	-218.5	0.9	382	268	319	347	1TYA-E	Tyrosyl-transfer RNA synthetase
MG127	High	0.916	-59.5	-5.8	40	78	85	145	1EGO	Glutaredoxin (oxidized) (NMR: 20 structures)
MG129	Cert	1.000	-39.4	-8.2	183	74	78	117	1IBA	Glucose permease fragment: domain iib
MG132	Cert	1.000	-170.9	-3.9	311	110	113	141	1KPA-A	Human protein kinase C interacting protein 1
MG134	Medium	0.802	-40.1	-3.0	27	84	86	100	1ACA	Acyl-coenzyme A binding protein (acbp)
MG136	Cert	1.000	-620.2	-16.4	1279	481	489	490	1LYL-A	Lysyl-tRNA synthetase (lysu)lysine
MG137	High	0.972	-122.8	13.8	292	359	487	404	1TYT-A	Trypanothione reductase (oxidized form (e))
MG138	Cert	1.000	-459.5	-11.0	588	444	631	598	1DAR	Elongation factor G (EF-G)
MG139	Cert	1.000	-244.7	-4.4	228	208	230	569	1ZNB-A	Metallo-beta-lactamase (class b beta-lactamase)
MG140	High	0.901	-233.6	25.4	63	618	651	1113	1PJR	PCRA DNA helicase
MG142	Cert	1.000	-289.8	-5.2	268	161	166	619	5P21	c-H-ras p21 protein (amino acid residues 1-166)
MG146	High	0.911	-234.2	-0.5	109	178	343	424	1AK5	Inosine-5'-monophosphate dehydrogenase (impdh)
MG153	Medium	0.676	-68.5	-5.2	19	63	63	106	1R69	434 Repressor (amino-terminal domain) (r1-69)
MG156	High	0.798	-87.9	0.1	43	54	68	144	1CTF	L7/L12 50 S ribosomal protein (C-terminal domain)
MG159	High	0.991	1.6	-4.8	105	188	215	200	1DKZ-A	Substrate binding domain of dnaK
MG160	Cert	1.000	1.4	2.0	320	80	80	81	1RIP	Ribosomal protein s17 (NMR: six structures)
MG161	Cert	1.000	-170.8	-8.6	492	122	122	122	1WHI	Ribosomal protein L14
MG165	Cert	1.000	-191.8	-4.9	416	130	130	141	1SEI-A	Ribosomal protein S8
MG168	Cert	1.000	-237.7	-7.8	564	145	145	211	1PKP	Ribosomal protein S5 (prokaryotic)
MG171	Cert	1.000	-282.7	-8.8	551	211	220	214	1AK2	Adenylate kinase isoenzyme-2
MG172	Cert	1.000	-325.7	-15.4	594	247	263	248	1MAT	Methionine aminopeptidase
MG173	Cert	1.000	-46.2	-2.4	291	70	71	70	1AH9	Initiation factor 1 (IF1)
MG177	Cert	1.000	-54.3	-5.0	234	68	81	328	1COO	RNA polymerase alpha subunit fragment: COOH-terminal domain
MG186	High	0.998	-98.6	-6.0	79	138	141	250	2SNS	Staphylococcal nuclease
MG191	High	0.909	-213.6	-0.1	143	633	684	1444	1CGT	Cyclodextrin glycosyltransferase
MG194	High	0.992	-106.7	13.8	354	268	489	341	1LYL-A	Lysyl-tRNA synthetase (lysu)lysine
MG195	Cert	1.000	-88.4	7.5	325	229	370	806	1HTT-A	Histidyl-tRNA synthetase (histidine-tRNA ligase) histidyl-adenylate
MG196	Cert	1.000	-70.0	1.3	160	48	76	141	1TIF	Translation initiation factor 3 (IF3-n)
MG200	Cert	1.000	-70.7	-9.3	283	77	77	601	1HDJ	Human hsp40fragment: j-domain (hdj-1)
MG201	Cert	1.000	-211.2	-1.5	324	162	164	217	1DKG-A	Nucleotide exchange factor grpe
MG203	Cert	1.000	-194.9	8.7	615	236	681	633	1BGW	topoisomerase fragment: residues 410-1202
MG204	Cert	1.000	-397.6	1.2	319	453	681	781	1BGW	topoisomerase fragment: residues 410-1202
MG206	Cert	1.000	-78.9	10.6	265	223	823	432	1TAQ	<i>Taq</i> DNA polymerase (<i>Taq</i>) mutant
MG211	Medium	0.708	-116.4	-2.0	50	125	215	147	1DKZ-A	Substrate binding domain of dnaK
MG213	Cert	1.000	-205.2	-3.3	141	152	159	471	4DFR-A	Dihydrofolate reductase complex with methotrexate
MG215	Cert	1.000	-548.9	-13.6	859	318	320	323	1PFK-A	Phosphofructokinase (R-state)
MG216	Cert	1.000	-658.2	-10.2	1224	460	519	508	1PKM	m1 pyruvate kinase (PK)
MG217	Cert	1.000	-0.2	-5.9	153	122	122	372	1BIP	Bifunctional trypsin alpha-amylase inhibitor (RBI)
MG218	Cert	1.000	-178.8	3.7	382	667	681	1805	1BGW	Topoisomerase fragment: residues 410-1202
MG227	Cert	1.000	-230.2	-18.1	966	287	316	287	1TSY	Thymidylate synthase mutant biological unit: homodimer
MG228	Cert	1.000	-204.1	-8.3	347	151	159	160	4DFR-A	Dihydrofolate reductase complex with methotrexate
MG229	Cert	1.000	-199.7	-7.0	531	283	340	340	1RIB-A	Protein r2 of ribonucleotide reductase
MG231	Cert	1.000	-565.9	-12.3	1666	684	739	721	1RLR	Ribonucleotide reductase protein r1

MG235	Cert	1.000	-204.6	-0.9	596	277	387	291	1XIS	Xylose isomerase complex with MnCl ₂
MG238	Cert	1.000	-70.2	-1.4	163	99	107	444	1FKB	R506 binding protein (fkbp)
MG244	Cert	1.000	-553.0	0.7	174	640	651	703	1PJR	PCRA DNA helicase
MG248	High	0.836	-223.6	-5.0	50	170	213	218	1VID	Catechol O-methyltransferase (comt)
MG249	Cert	1.000	-157.0	-2.7	228	300	314	497	1SIG	RNA polymerase primary sigma factor fragment
MG251	Cert	1.000	-350.8	2.4	383	331	370	446	1HTT-A	Histidyl-tRNA synthetase (histidine-tRNA ligase) histidyl-adenylate
MG253	Cert	1.000	-256.4	5.8	282	339	540	428	1GTR-A	Glutaminyl-tRNA synthetase
MG259	High	0.930	-178.7	9.3	99	252	393	456	2ADM-A	Adenine-N6-DNA-methyltransferase <i>TaqI</i>
MG262	Cert	1.000	-278.4	-1.2	645	270	294	291	1TFR	T4 RNase H (T4 5u to 3u exonuclease)
MG263	Medium	0.771	-127.5	5.6	18	188	220	291	1JUD	L-2-Haloacid dehalogenase
MG264	Cert	1.000	-208.7	1.7	165	165	186	198	1GKY	Guanylate kinase complex with guanosine monophosphate
MG265	High	0.839	-210.0	-4.1	42	208	220	278	1JUD	L-2-Haloacid dehalogenase
MG266	High	0.801	-178.5	10.3	154	441	540	792	1GTR-A	Glutaminyl-tRNA synthetase complexed with tRNA and ATP (-8 deg.C)
MG268	Cert	1.000	-256.1	1.2	208	190	220	228	1AK2	Adenylate kinase isoenzyme-2
MG269	Cert	1.000	-134.6	-7.5	139	261	314	340	1SIG	RNA polymerase primary sigma factor fragment
MG271	Cert	1.000	-585.7	-28.4	720	449	472	457	3LAD-A	Dihydroloipoamide dehydrogenase
MG272	Cert	1.000	-386.8	-12.0	702	243	243	384	1DPB	Dihydroloipoyl transacetylase catalytic domain
MG273	Cert	1.000	-322.3	6.5	652	305	678	326	1TRK-A	Transketolase
MG274	Cert	1.000	-279.6	3.0	491	320	678	358	1TRK-A	Transketolase
MG275	Cert	1.000	-655.7	-23.6	607	447	447	478	1NHR	NADH peroxidase (npx) mutant with Leu 40 replaced by Cys (L40C)
MG276	Cert	1.000	-253.4	3.8	190	146	173	180	1HGX-A	Hypoxanthine-guanine-xanthine phosphoribosyltransferase (hgxprtase)
MG282	Cert	1.000	-207.6	-12.7	475	157	157	161	1GRJ	Grea transcript cleavage factor from <i>Escherichia coli</i>
MG283	Cert	1.000	-277.1	-0.2	326	331	370	483	1HTT-A	Histidyl-tRNA synthetase (histidine-tRNA ligase) histidyl-adenylate
MG287	Cert	1.000	-32.6	-0.1	146	74	77	84	1ACP	Acyl carrier protein (NMR: two structures)
MG288	Medium	0.702	-71.4	-8.4	79	204	215	414	1DKZ-A	Substrate binding domain of dnaK
MG290	High	0.815	-172.1	-2.1	62	149	180	245	1HUR-A	Human ADP-ribosylation factor 1 (harf1: arf1: marf1)
MG295	Cert	1.000	-203.6	6.5	278	270	501	367	1GPM-A	GMP synthetase (XMP aminase)
MG297	Cert	1.000	-256.9	4.4	273	196	224	346	1DAD	Dethiobiotin synthetase (dtbs)
MG298	High	0.926	-220.3	-1.2	105	309	314	982	1SIG	RNA polymerase primary sigma factor fragment
MG300	Cert	1.000	-423.6	-21.2	1200	404	415	416	3PGK	Phosphoglycerate kinase
MG301	Cert	1.000	-568.3	-21.2	1084	327	334	337	1GD1-O	Holo-D-glyceraldehyde-3-phosphate dehydrogenase
MG305	Cert	1.000	-603.7	-26.8	1216	354	378	595	1HPM	44 kDa ATPase fragment (N-terminal) of 70 kDa heat-shock cognate protein
MG308	High	0.917	-192.1	23.6	76	354	651	410	1PJR	PCRA DNA helicase
MG310	Cert	1.000	-248.3	-0.6	267	260	277	268	1BRO-A	Bromoperoxidase a2 (haloperoxidase a2: chloroperoxidase a2)
MG318	High	0.800	-5.5	2.5	173	57	179	280	1HNF	CD2 (human)
MG321	High	0.962	-125.4	-5.8	197	429	517	934	2OLB-A	Oligopeptide binding protein (OPPA)
MG323	High	0.932	-150.5	9.3	126	215	255	227	1FMC-A	7-Alpha-hydroxysteroid dehydrogenase biological unit: tetramer
MG324	Cert	1.000	-429.5	-8.3	786	350	401	354	1CHM-A	Creatine amidinohydrolase
MG327	Cert	1.000	-376.6	-9.4	297	261	277	268	1BRO-A	Bromoperoxidase a2 (haloperoxidase a2: chloroperoxidase a2)
MG328	Cert	1.000	-197.4	-5.4	227	311	314	756	1SIG	RNA polymerase primary sigma factor fragment
MG329	Cert	1.000	-265.9	1.0	183	159	166	448	5P21	c-H-ras p21 protein (amino acid residues 1-166)
MG330	Cert	1.000	-130.1	6.1	216	215	220	217	1AK2	Adenylate kinase isoenzyme-2
MG333	Cert	1.000	-95.2	6.4	385	125	273	126	1QRD-A	Quinone-reductase (dt-diaphorase)
MG334	Cert	1.000	-225.7	5.8	370	409	468	837	1GLN	Glutamyl-tRNA synthetase
MG335	Cert	1.000	-255.2	-5.0	106	152	166	191	5P21	c-H-ras p21 protein (amino acid residues 1-166)
MG336	Cert	1.000	-496.9	-5.1	203	373	431	408	2DKB	2:2-Dialkylglycine decarboxylase (pyruvate) (dgd)
MG339	Cert	1.000	-448.1	-12.8	713	315	326	340	2REB	RecA protein
MG342	Cert	1.000	-150.0	-6.1	54	126	138	168	5NLL	Flavodoxin biological unit: monomer
MG344	Cert	1.000	-338.8	-8.4	257	259	277	273	1BRO-A	Bromoperoxidase a2 (haloperoxidase a2. chloroperoxidase a2)
MG345	Cert	1.000	-185.2	13.3	224	428	468	895	1GLN	Glutamyl-tRNA synthetase
MG347	High	0.810	-165.4	2.0	109	179	292	210	1XVA-A	Glycine N-methyltransferase
MG351	Cert	1.000	-221.1	-6.3	548	161	174	184	2PRD	Pyrophosphate phosphohydrolase
MG353	Cert	1.000	-48.1	-0.5	76	90	90	109	1HUE-A	Hu protein

continued overleaf

Table 3—Continued

ORF	Category	Net	E_{pair}	E_{solv}	AlnSc	Alen	Len1	Len2	PDB code	Description
MG356	Cert	1.000	-236.1	2.3	174	239	299	280	1CKI-A	Casein kinase I delta mutant
MG358	Cert	1.000	-231.3	-5.1	574	185	203	260	1CUK	Ruva protein
MG359	High	0.911	-141.6	5.4	95	182	186	307	1GKY	Guanylate kinase complex with guanosine monophosphate
MG362	Cert	1.000	-142.0	-6.6	211	68	68	122	1CTF	L7/L12 50 S ribosomal protein (C-terminal domain)
MG365	Cert	1.000	-218.0	-10.3	393	198	209	311	1GAR-A	Glycinamide ribonucleotide transformylase
MG372	High	0.816	-249.5	6.5	133	342	501	385	1GPM-A	GMP synthetase (XMP aminase)
MG375	Cert	1.000	-346.0	-4.3	574	359	370	564	1HTT-A	Histidyl-tRNA synthetase (histidine-tRNA ligase) histidyl-adenylate
MG378	Cert	1.000	-202.5	8.7	265	381	468	537	1GLN	Glutamyl-tRNA synthetase
MG379	Cert	1.000	-271.8	15.8	295	444	485	612	1FEC-A	Trypanothione reductase biological unit. homodimer
HG380	High	0.995	-178.5	2.7	109	182	213	192	1VID	Catechol O-methyltransferase (comt)
MG382	Cert	1.000	-205.2	2.2	152	185	186	213	1GKY	Guanylate kinase complex with guanosine monophosphate
MG383	Cert	1.000	-274.8	-2.0	375	222	501	248	1GPM-A	GMP synthetase (XMP aminase).
MG384	High	0.970	-258.5	-1.6	119	164	166	433	5P21	c-H-ras p21 protein (amino acid residues 1-166)
MG387	High	0.925	-256.0	-7.1	58	167	177	290	1ETU	Elongation factor Tu (domain I)-guanosine diphosphate complex
MG391	Cert	1.000	-548.2	-8.6	930	438	484	447	1LAM	Leucine aminopeptidase (cytosolic aminopeptidase)
MG392	Cert	1.000	-810.0	-26.8	1428	524	525	543	1DER-A	Groel mutant
MG393	Cert	1.000	-97.4	-2.7	348	96	97	110	1AON-A	Groel (60 kDa chaperonin: protein cpn60)
MG394	Cert	1.000	-386.1	-3.7	186	349	396	406	1ARS	Aspartate aminotransferase complexed with pyridoxal-5'-phosphate
MG398	Cert	1.000	-195.8	-6.2	261	132	133	135	1AQT	Epsilon subunit of F1F0 ATP synthase
MG399	Cert	1.000	-452.2	-11.5	1293	376	467	382	1BMF-A	Bovine mitochondrial F1-ATPase (F1-ATPase)
MG400	Cert	1.000	-78.9	2.7	173	111	122	279	1BMF-A	Bovine mitochondrial F1-ATPase (F1-ATPase)
MG401	Cert	1.000	-668.8	-22.9	1226	478	487	518	1BMF-A	Bovine mitochondrial F1-ATPase (F1-ATPase)
MG402	Cert	1.000	-162.6	-6.6	238	102	105	176	1ABV	Delta subunit of the F1F0-ATP synthase fragment: N-terminal domain
MG403	High	0.968	-166.1	-0.3	65	144	144	208	1LPE	Apolipoprotein-E3 (LDL receptor binding domain)
MG407	Cert	1.000	-707.0	-13.4	1531	432	436	458	1ONE-A	Enolase (2-phospho-D-glycerate hydrolase)
MG412	Cert	1.000	-265.5	-18.2	541	299	321	377	1QUK	Phosphate-binding protein mutant
MG419	High	0.825	-105.5	-9.0	40	168	184	287	1IDO	Integrin fragment: i-domain; (a-domain)
MG420	High	0.802	-142.9	10.1	82	211	220	260	1AK2	Adenylate kinase isoenzyme-2 (ATP:AMP phosphotransferase)
MG423	High	0.805	-180.0	-2.6	138	182	221	561	1BME	Metallo-beta-lactamase (class b beta-lactamase)
MG424	High	0.998	-10.2	-3.3	298	86	88	86	1AB3	Ribosomal RNA binding protein S15
MG425	High	0.811	-283.9	16.3	57	404	651	449	1PJR	PCRA DNA helicase
MG429	Cert	1.000	-613.9	-10.0	1199	521	873	572	1DIK	Pyruvate phosphate dikinase (ppdk) biological unit: dimer
MG430	High	0.978	-299.3	12.0	166	381	449	507	1ALK-A	Alkaline phosphatase
MG431	Cert	1.000	-377.1	-18.9	696	241	250	244	1TPF-A	Triosephosphate isomerase
MG433	Cert	1.000	-437.8	-14.6	865	281	282	298	1EFU-A	Elongation factor Tu (elongation factor for transfer: heat unstable: EF-Tu)
MG442	High	0.856	-104.9	5.9	113	139	177	270	1ETU	Elongation factor Tu (domain i) - guanosine diphosphate complex
MG444	High	0.801	-32.6	-2.6	42	94	97	119	1RIS	Ribosomal protein S6
MG449	Cert	1.000	-107.9	-2.1	247	142	785	144	1PYS-B	Phenylalanyl-tRNA synthetase
MG451	Cert	1.000	-537.5	-27.3	1198	393	405	394	1EFT	Elongation factor Tu (EF-Tu)
MG455	Cert	1.000	-401.4	-19.0	847	310	319	407	1TYA-E	Tyrosyl-transfer RNA synthetase
MG458	Cert	1.000	-275.4	-11.1	336	171	173	175	1HGX-A	Hypoxanthine-guanine-xanthine phosphoribosyl transferase (hgxpptase)
MG460	Cert	1.000	-493.9	-18.5	851	304	309	312	1HYH-A	L-2-Hydroxyisocaproate dehydrogenase (L-hicdh)
MG462	Cert	1.000	-445.3	-30.2	958	465	468	484	1GLN	Glutamyl-tRNA synthetase
MG463	Cert	1.000	-188.5	2.1	172	217	393	259	2ADM-A	Adenine-N6-DNA-methyltransferase <i>TaqI</i>
MG468	Cert	1.000	-278.4	-1.2	645	270	294	291	1TFR	T4 RNase h (T4 5' to 3' exonuclease)
MG469	High	0.843	-209.3	0.8	50	198	220	437	1AK2	Adenylate kinase isoenzyme-2 (ATP:AMP phosphotransferase)
MG470	Cert	1.000	-272.8	1.5	269	252	283	269	1NIP-A	Nitrogenase iron protein

ORF, open reading frame identifier; Category, confidence category for prediction; Net, strength of prediction (network output); E_{pair} , pairwise energy sum for predicted fold; E_{solv} , solvation energy sum for predicted fold; AlnSc, sequence alignment score; Alen, number of aligned residues; Len1, length of template protein sequence; Len2, length of target protein sequence; PDB code, matched chain from fold library; Description, description of matched protein structure.

Table 4. ORFs with ABC (ATP-binding) domains

MG014	Transport ATP-binding protein (msbA)
MG015	Transport ATP-binding protein (msbA)
MG042	Spermidine/putrescine transport ATP-binding protein
MG065	Heterocyst maturation protein (devA)
MG079	Oligopeptide transport ATP-binding protein (amiE)
MG080	Oligopeptide transport ATP-binding protein (amiF)
MG119	Methylgalactoside permease ATP-binding protein (mglA)
MG179	Haemolysin secretion ATP-binding protein (hlyB)
MG180	Membrane transport protein (glnQ)
MG187	ATP-binding protein (msmK)
MG239	ATP-dependent protease (lon)
MG290	ATP-binding protein P29
MG303	Membrane transport protein (glnQ)
MG304	Membrane associated ATPase (cbiO)
MG390	Lactococin transport ATP-binding protein (lcnDR3)
MG410	Peripheral membrane protein B (pstB)
MG421	Excinuclease ABC subunit A (uvrA)
MG467	Heterocyst maturation protein (devA)

These domains match several GTP and ATP binding folds in the fold library at high and certain confidence. Another six ORFs also match these domains, but with borderline alignment scores.

Finally, as further corroborating evidence it can be seen in Figure 5(c) that the secondary structure prediction for ORF MG276 is in very good agreement with the observed secondary structure of 1HGX.

ORF MG353: an ORF with no known function

ORF MG353 represents a more interesting class of problem. In this case, no function has been assigned to this ORF by searching sequence data banks for homology to proteins of known function. Despite this, GenTHREADER produces a match to a PDB entry (1HUE) with very high confidence. The alignment is shown in Figure 6, and clearly again the predicted secondary structure for MG353 is in good agreement with that observed in 1HUE. The fold assigned in this case (shown in Figure 6(a)) is that of a superfamily of prokaryotic DNA-bending proteins, often described as “histone-like” proteins. Although the precise function

Table 5. A list of “novel” structural assignments which cannot be recognised when the pairwise and solvation energy components are not considered

ORF	Category	TIGR annotation	PDB Code	Description
MG006	High	Thymidylate kinase	1GKY	Guanylate kinase
MG007	High	DNA polymerase III subunit	1AK2	Adenylate kinase isoenzyme-2
MG008	Cert	Thiophene and furan oxidizer	1HUR-A	Human ADP-ribosylation factor 1 (harf1: arf1: marf1)
MG013	High	5,10-Methylene-tetrahydrofolate dehydrogenase	1GDH-A	D-Glycerate dehydrogenase (apo form)
MG023	High	Fructose-bisphosphate aldolase	1WSY-A	Tryptophan synthase
MG050	High	Deoxyribose-phosphate aldolase	1RVV-A	Riboflavin synthase
MG101	High	-	1OCT-C	Oct-1 (pou domain)
MG112	High	D-Ribulose-5-phosphate 3 epimerase	1IGS	Indole-3-glycerolphosphate synthase (igps)
MG125	High	Hypothetical protein	1JUD	L-2-haloacid dehalogenase
MG127	High	Hypothetical protein	1EGO	Glutaredoxin (oxidized) (NMR: 20 structures)
MG140	High	-	1PJR	PCRA DNA helicase
MG146	High	Hemolysin	1AK5	Inosine-5'-monophosphate dehydrogenase (impdh)
MG156	High	Ribosomal protein L22	1CTF	L7/L12 50 S ribosomal protein (C-terminal domain)
MG159	High	Ribosomal protein L29	1DKZ-A	Substrate binding domain of dnaK
MG186	High	Hypothetical protein	2SNS	Staphylococcal nuclease
MG213	Cert	-	4DFR-A	Dihydrofolate reductase complex with methotrexate
MG248	High	Major sigma factor	1VID	Catechol O-methyltransferase (comt)
MG259	High	Protoporphyrinogen oxidase	2ADM-A	Adenine-N6-DNA-methyltransferase <i>TaqI</i>
MG265	High	Hypothetical protein	1JUD	L-2-Haloacid dehalogenase
MG290	High	ATP-binding protein P29	1HUR-A	Human ADP-ribosylation factor 1 (harf1: arf1: marf1)
MG298	High	115 kDa protein	1SIG	RNA polymerase primary sigma factor fragment
MG308	High	ATP-dependent RNA helicase	1PJR	PCRA DNA helicase
MG335	Cert	Hypothetical protein	5P21	c-H-ras p21 protein (amino acids 1-166)
MG342	Cert	-	5NLL	Flavodoxin biological unit: monomer
MG347	High	Hypothetical protein	1XVA-A	Glycine N-methyltransferase
MG353	Cert	-	1HUE-A	Hu protein
MG359	High	Holliday junction DNA helicase	1GKY	Guanylate kinase complex with guanosine monophosphate
MG372	High	-	1GPM-A	GMP synthetase (XMP aminase)
MG380	Cert	Glucose inhibited division protein	1VID	Catechol O-methyltransferase (comt)
MG384	High	GTP-binding protein	5P21	c-H-ras p21 protein (amino acids 1-166)
MG387	High	GTP-binding protein era homolog	1ETU	Elongation factor Tu (domain i)
MG403	High	ATP synthase B chain	1LPE	Apolipoprotein-E3 (LDL receptor binding domain)
MG419	High	-	1IDO	Integrin fragment: i-domain (a-domain)
MG420	High	DNA polymerase III subunit	1AK2	Adenylate kinase isoenzyme-2 (ATP:AMP phosphotransferase)
MG425	High	ATP-dependent RNA helicase	1PJR	PCRA DNA helicase
MG442	High	Hypothetical protein	1ETU	Elongation factor Tu (domain i)
MG444	High	Ribosomal protein L19	1RIS	Ribosomal protein S6
MG469	High	Chromosomal replication initiator protein	1AK2	Adenylate kinase isoenzyme-2 (ATP:AMP phosphotransferase)

The author annotations for each ORF are given alongside the annotation that is implied from the matched protein fold.

of these proteins remains unknown, they are able to wrap DNA and protect it from denaturation.

Again further corroboration for assigning MG353 to the histone-like protein superfamily by considering the conserved residues in the alignment shown in Figure 7(b). The most striking region of conservation in this alignment (shown in bold) is over the region starting at Arg66 and ending at Pro83 in MG353, which is aligned to the region of 1HUE involved in DNA binding. Figure 7(a) shows the dimeric structure of 1HUE, with the two DNA binding arms (one per subunit) indicated.

Discussion

It has been demonstrated here that a relatively simple approach to the fold recognition problem can achieve extremely good results when applied to genome analysis. One thing that must be emphasised is that the main goal of this method is not just to predict protein structure, but also to infer possible evolutionary relationships. For genome annotation this is a very important distinction to make. By concentrating on proteins with likely common ancestry it is also much more straightforward to produce a reliable estimator of confidence

in a given prediction. This is an important distinction between this approach to fold recognition and more general threading methods. For almost all threading methods it would be possible to generate a prediction for every gene product from a given genome. The problem would be to determine which of the predictions are correct and which are just random noise. By using the three-stage filtering procedure outlined here, however, the decision as to which predictions should be taken seriously and which should be ignored is easy to make.

Other groups have also performed analyses similar to the present study, though with different methods. Fischer & Eisenberg (1997) have recently reported the results of applying their automated fold recognition method to the *M. genitalium* genome. They found that 22% of the ORFs could be assigned to a known fold at high confidence, with 16% showing enough sequence similarity to be picked up by sequence comparison alone. For the ORFs predicted both by Fischer & Eisenberg (1997) and those listed in Tables 3 and 4, there seems to almost complete agreement in the fold assignments. Interestingly, Fischer & Eisenberg (1997) report some fold assignments which are not detected by GenTHREADER at high confidence. This suggests that there is definite scope for

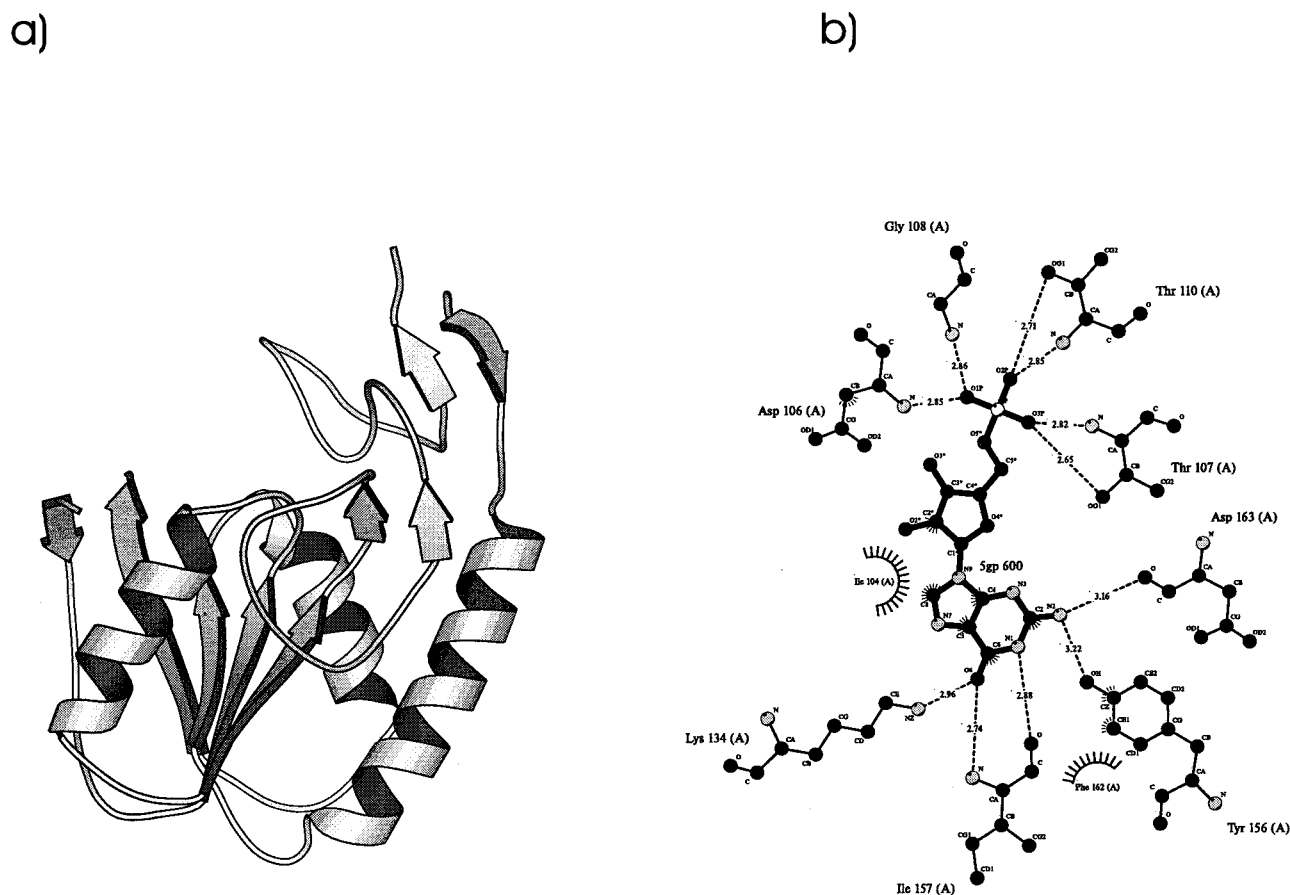


Figure 5 (a) and (b) (legend opposite)

match seems to make some sense from the biology, PDB entry 1EAG is an aspartic proteinase, and ORF MG067 has been tentatively annotated as a glutamic acid specific protease. However, indicators point to this being an incorrect match. Firstly, a full threading search using THREADER 2 finds a trypsin-like serine protease as the most compatible fold (with a plausible alignment which conserved the catalytic His, Asp and Ser). Sec-

ondly, the alignment with 1EAG is not consistent with the functionally important residues in aspartic proteinases. Thirdly, although some evidence now exists that aspartic proteinases are found in bacteria (Hill & Philip 1997), they appear to be relatively rare. The weight of evidence thus seems to point more towards common ancestry with trypsin-like serine proteases than with aspartic proteinases, but is by no means a clear cut case. This is a

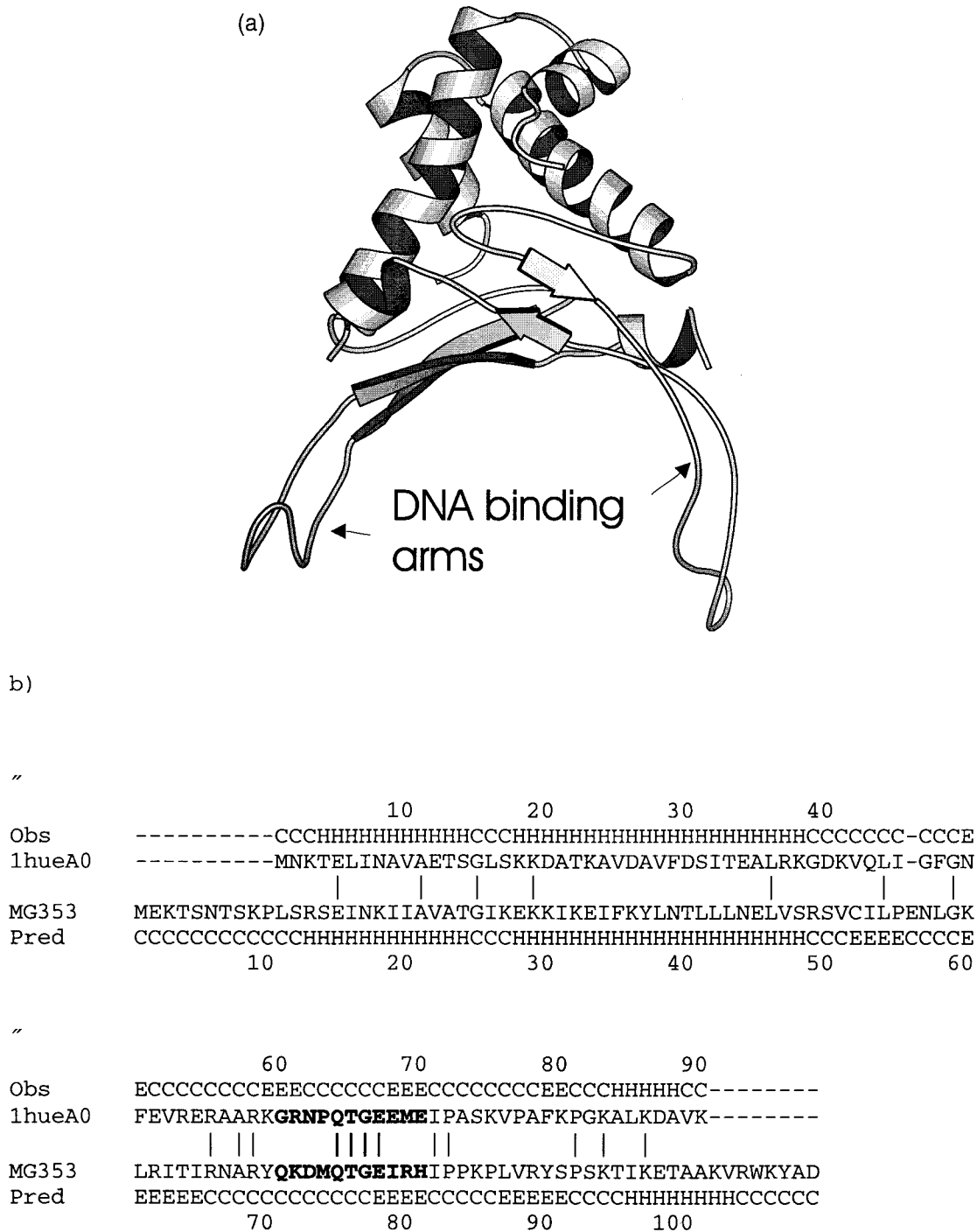


Figure 6. An example of a GenTHREADER prediction for an ORF of unknown function (MG353). (a) A MOLSCRIPT diagram showing the best matching structure as a dimer. The DNA binding "arms" of the protein are marked. (b) The GenTHREADER alignment with the region corresponding to the DNA-binding arm shown in bold.

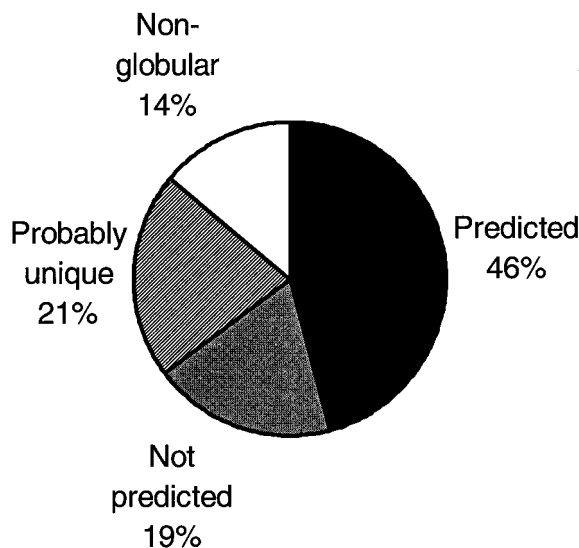


Figure 7. Overall summary of results for *M. genitalium* predictions. A total of 46% of the ORFs can apparently be at least partially modelled on a known protein fold. Of the remainder, 14% of the ORFs are predicted to encode non-globular proteins (transmembrane proteins and some repetitive sequences), 21% are expected to have expected to have folds which are not yet represented in the fold library, and 19% probably do contain a domain which resembles a known fold, but these have not been predicted at high confidence by GenTHREADER.

very good example of the value added by the second and third-level filters in GenTHREADER; the initial sequence alignment, whilst producing a significant score, was rejected on the basis of poor compatibility with the matched protein fold. Although it is still impossible to rule out the assignment of MG067 as an aspartic proteinase, the fact that the implied model was of low quality indicates that despite the significant local sequence similarity the match should be assigned a very low confidence.

Perhaps a more interesting question is to ask how much further we might expect to go with the assumption that a method might be devised for finding almost all the related folds in the proteome from sequence. It might be possible, for example, to pool the results from several methods. Figure 7 shows a summary of the fraction of the *M. genitalium* ORFs which can be confidently predicted (46%) by the method described here. From transmembrane segment prediction and detection of repeats it is estimated that outside the 46% of ORFs with at least one "predictable" domain, 14% of the ORFs represent entirely non-globular proteins (note that some of the predictable proteins are transmembrane also, though generally with just a single membrane spanning segment). From comparison of protein structures (Orengo *et al.*, 1994, 1997; C. A. Orengo, personal communication) the fraction of domains of known structure which

have a similar fold to another protein yet do not have a significant degree of sequence similarity is 75%. This implies that 75% of the globular protein domains in *M. genitalium* should in principle be amenable to some kind of protein fold recognition method. Ignoring the implications of domain structure, we would therefore expect that 64.5% (75% of 86%) of the ORFs should have a known fold. From this we would expect that an ideal fold recognition method should extend the predictable region in Figure 4 from 46% to 64.5%, but on a more positive note this implies that GenTHREADER is already detecting 71% of the structural similarities in the *M. genitalium* genome. The complication of domain organization casts some doubts on this estimate of course. Although 35% of the ORFs match a protein of known structure over more than 50% of their length, the remaining 11% only match within a local region. For these 11% of the ORFs, additional unique domains may be present, which would reduce the estimate of effectiveness from 71% to a lower value. However, in some cases these other domains have also been matched to a known structure (Table 3 shows only the single top scoring match). In light of this, therefore, it is estimated that GenTHREADER is correctly recognizing from 65-71% of the *M. genitalium* domains which are homologous to proteins of known structure.

Apart from further developments of GenTHREADER itself, an approach to tackle the missing 29% of the structural similarities might be to switch to a full threading method, and not one which is tuned to detecting structures with common ancestry. Indeed, this might well be expected from the distribution of superfolds (Orengo *et al.*, 1994) in non-homologous proteins, where in our most recent analysis (Orengo *et al.*, 1997) we found that 31% of the protein domain folds match one of the ten known superfolds (fold families which include pairs of proteins with no common ancestry). This might imply that most of the similarities which GenTHREADER is not able to recognise are in fact between superfolds where no evolutionary relationship is implied. This is rather speculative and optimistic of course, but nonetheless it is likely that superfolds will be extremely prevalent in the "unpredictable" set of ORFs, and these will require a more sophisticated fold recognition method (i.e. threading) to be used before they can be identified.

At least 20% of the domains in *M. genitalium* are expected to have unique folds based on the above estimates. One of the possible applications for a fold recognition method might be to prioritize these gene products for X-ray crystallographic experiments. By focusing structure determination on the gene products which cannot be predicted by fold recognition methods, the generality of the fold library can be greatly increased. It might well be hoped that when the folds of all the globular domains in the *M. genitalium* proteome are presented in the fold library then a very large fraction

of larger proteomes might well be amenable to large scale structure prediction.

References

- Abagyan, R. A. & Batalov, S. (1997). Do aligned sequences share the same fold? *J. Mol. Biol.* **273**, 355-368.
- Abagyan, R., Frishmani, D. & Argos, P. (1994). Recognition of distantly related proteins through energy calculations. *Proteins: Struct. Funct. Genet.* **19**, 132-140.
- Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). Protein Data Bank. In *Crystallographic Databases*, pp. 107-132, Data Commission of the International Union of Crystallography, Bonn, Cambridge, Chester.
- Altschul, S. E., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.
- Annereau, J. P., Wulbrand, U., Vankeerberghen, A., Cuppens, H., Bontems, F., Tummeler, B., Cassiman, J. J. & Stoven, V. (1997). A novel model for the first nucleotide binding domain of the cystic fibrosis transmembrane conductance regulator. *FEBS Letters*, **407**, 303-308.
- Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein databank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.
- Bleasby, A. J., Akrigg, D. & Attwood, T. K. (1994). OWL - a non-redundant composite protein sequence database. *Nucl. Acids Res.* **22**, 3574-3577.
- Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164-170.
- Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073-6078.
- Bryant, S. H. & Lawrence, C. E. (1993). An empirical energy function for threading protein-sequence through the folding motif. *Proteins: Struct. Funct. Genet.* **16**, 92-112.
- Defay, T. R. & Cohen, F. E. (1996). Multiple sequence information for threading algorithms. *J. Mol. Biol.* **262**, 314-323.
- Fischer, D. & Eisenberg, D. (1996). Protein fold recognition using sequence-derived predictions. *Protein Sci.* **5**, 947-955.
- 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.
- Fischer, D., Elofsson, A., Rice, D. W., LeGrand, S. & Eisenberg, D. (1996). Assessing the performance of fold recognition methods by means of a comprehensive benchmark. In *Proceedings of the Pacific Symposium on Biocomputing*, pp. 300-318, World Scientific Press, Hawaii.
- Flöckner, H., Braxenthaler, M., Lackner, P., Jarirz, M., Ortner, M. & Sippl, M. J. (1995). Progress in fold recognition. *Proteins: Struct. Funct. Genet.* **23**, 376-386.
- Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M. & Fuhrmann, J., et al. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397-408.
- Godzik, A. & Skolnick, J. (1992). Sequence-structure matching in globular proteins: application to super-secondary and tertiary structure determination. *Proc. Natl Acad. Sci. USA*, **89**, 12098-12102.
- Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355-4358.
- Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M. J. (1990). Identification of native protein folds amongst a large number of incorrect models: the calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **216**, 167-180.
- Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915-10919.
- Hill, J. & Phylip, L. H. (1997). Bacterial aspartic proteinases. *FEBS Letters*, **409**, 357-360.
- Huynen, M., Doerks, T., Eisenhaber, R., Orengo, C., Sunyaev, S., Yuan, Y. P. & Bork, P. (1998). Homology-based fold predictions for *Mycoplasma genitalium* proteins. *J. Mol. Biol.* **280**, 323-326.
- Jones, D. T. (1997). Progress in protein structure prediction. *Curr. Opin. Struct. Biol.* **7**, 377-387.
- Jones, D. T. & Thornton, J. M. (1996). Potential-energy functions for threading. *Curr. Opin. Struct. Biol.* **6**, 210-216.
- Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, **358**, 86-89.
- Jones, D. T., Miller, R. T. & Thornton, J. M. (1995). Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. *Proteins Struct. Funct. Genet.* **23**, 387-397.
- Kocher, J. P. A., Rooman, M. J. & Wodak, S. J. (1994). Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J. Mol. Biol.* **235**, 1598-1613.
- Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946-950.
- Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994). Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.* **235**, 1501-1531.
- Lathrop, R. H. & Smith, T. F. (1996). Global optimum protein threading with gapped alignment and empirical pair score functions. *J. Mol. Biol.* **255**, 641-665.
- Lemer, C. M. R., Rooman, M. J. & Wodak, S. J. (1995). Protein structure prediction by threading methods - evaluation of current techniques. *Proteins: Struct. Funct. Genet.* **23**, 337-355.
- Lüthy, R., Xenarios, I. & Bucher, P. (1994). Improving the sensitivity of the sequence profile method. *Protein Sci.* **3**, 139-146.

- Madej, T., Gilbrat, J.-F. & Bryant, S. H. (1995). Threading a database of protein cores. *Proteins: Struct. Funct. Genet.* **23**, 356-369.
- Maierov, V. N. & Crippen, G. M. (1992). Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* **227**, 876-888.
- Miyazawa, S. & Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623-644.
- Nishikawa, K. & Matsuo, Y. (1994). Development of pseudoenergy potentials for assessing protein 3-D-1-D compatibility and detecting weak homologies. *Protein Eng.* **6**, 811-820.
- Orengo, C. A., Brown, N. P. & Taylor, W. R. (1992). Fast structure alignment for protein databank searching. *Proteins: Struct. Funct. Genet.* **14**, 139-167.
- Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature*, **372**, 631-634.
- Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH - a hierarchic classification of protein domain structures. *Structure*, **5**, 1093-1108.
- Ouzounis, C., Sander, C., Scharf, M. & Schneider, R. (1993). Prediction of protein structure by evaluation of sequence-structure fitness. Aligning sequences to contact profiles derived from three-dimensional structures. *J. Mol. Biol.* **232**, 805-825.
- Park, B. H., Huang, E. S. & Levitt, M. (1997). Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* **266**, 831-846.
- Russell, R. B., Copley, R. R. & Barton, G. J. (1996). Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol.* **259**, 349-365.
- Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56-68.
- Shortle, D. (1997). Folding proteins by pattern recognition. *Curr. Biol.* **7**, R151-R154.
- Taylor, W. R. (1986). Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.* **188**, 233-258.
- Taylor, W. R. (1988). A flexible method to align large numbers of biological sequences. *J. Mol. Evol.* **28**, 161-169.
- Wallace, A. C., Laskowski, R. A. & Thornton, J. M. (1995). Ligplot - a program to generate schematic diagrams of protein ligand interactions. *Protein Eng.* **8**, 127-134.
- Yi, T. M. & Lander, E. S. (1994). Recognition of related proteins by iterative template refinement (ITR). *Protein Sci.* **3**, 1315-1328.
- Yoshida, M. & Amano, T. (1995). Common topology of proteins catalyzing ATP-triggered reactions. *FEBS Letters*, **359**, 1-5.

Edited by B. Honig

(Received 31 March 1998; accepted 19 January 1999)