Update on protein structure prediction: results of the 1995 IRBM workshop

Tim Hubbard, Anna Tramontano and the 1995 IRBM workshop team*

Computational tools for protein structure prediction are of great interest to molecular, structural and theoretical biologists due to a rapidly increasing number of protein sequences with no known structure. In October 1995, a workshop was held at IRBM to predict as much as possible about a number of proteins of biological interest using *ab initio* prediction of fold recognition methods. 112 protein sequences were collected via an open invitation for target submissions. 17 were selected for prediction during the workshop and for 11 of these a prediction of some reliability could be made. We believe that this was a worthwhile experiment showing that the use of a range of independent prediction methods and thorough use of existing databases can lead to credible and useful *ab initio* structure predictions.

Address: Instituto di Ricerche di Biologia Molecolare (IRBM), P. Angeletti, Via Pontina Km 30.600, 00040 Pomezia, Roma, Italy.

Correspondence: Tim Hubbard, Centre for Protein Engineering, Medical Research Council (MRC) Centre, Cambridge, CB2 2QH, UK. e-mail: th@mrc-cpe.cam.ac.uk

*The 1995 IRBM workshop team: Geoff Barton, David Jones, Manfred Sippl, Alfonso Valencia, Arthur Lesk, John Moult, Burkhard Rost, Chris Sander, Reinhard Schneider, Armin Lahm, Raphael Leplae, Christiane Buta, Miriam Eisenstein, Ola Fjellström, Hannes Floeckner, J Guenter Grossmann, Jan Hansen, Manuela Helmer Citterich, Flemming Steen Joergensen, Aron Marchler-Bauer, Joel Osuna, Jong Park, Astrid Reinhardt, Lluis Ribas de Pouplana, Arturo Rojo-Dominguez, Vladimir Saudek, John Sinclair, Shane Sturrock, Ceslovas Venclovas and Carla Vinals.

Electronic identifier: **1359-0278-001-R0055**

Folding & Design 01 Jun 1996, 1:R55–R63

© Current Biology Ltd ISSN 1359-0278

Introduction

In December 1994, there was a meeting to evaluate the first ever large-scale protein structure prediction competition, which ran for most of 1994 [1,2]. The results were instructive in that fold recognition methods [3] were shown to frequently identify folds compatible with a target sequence in the absence of detectable sequence homology and useful *ab initio* predictions were made for targets with many homologous sequences [4]. We felt that this progress had to be exploited by bringing together the authors of the most successful methods to produce models of proteins of biological interest.

The scientific community was invited, via announcements on the internet, to propose suitable target proteins for this experiment. The criteria were set such that the prediction of the proposed proteins would be helpful to the biological community and that no homologous protein of known structure should be present in the database. All 112 submitted protein sequences were automatically analyzed in order to collect as much information as possible before the workshop and screen out targets with obvious homology to known structures.

At the beginning of the IRBM workshop, the authors of this report selected a subset of 17 proteins, judged to be suitable for prediction by a number of published and unpublished methods (Table 1), and during the next 10 days attempts were made to predict as much as possible about them. A flow chart of the steps typically used for predictions made during this workshop is shown in Figure 1. Detailed information and references for most methods are publicly available via the World Wide Web (WWW) together with the relevant bibliography on the selected target proteins and the full workshop reports [5]. A summary of the results of the different methods used for each of the 17 proteins is shown in Table 2. For 11 of these proteins, a reliable prediction at a useful level of detail could indeed be obtained and is critically reviewed here.

Predictions

For one of the target proteins (T0092) a cluster of secondary structural units could clearly be identified, but little concrete information could be obtained about the way they interact in three dimensions. In two cases (T0098 and T0218) some specific long-range interactions could be identified with some confidence, but there were insufficient data to determine the entire or exact fold. In the remaining cases, either the relative position in space of most secondary structural segments could be accounted for (T0167), or a possible match to a known fold could be identified (T0112, T0119, T0127, T0129, T0149, T0174 and T0217).

Target T0092 is the nitrogenase δ-chain of *Rhodobacter capsulatus*, an enzyme that catalyzes the reduction of molecular nitrogen to ammonia in nitrogen-fixing microorganisms. Nitrogenase consists of two metallo-proteins, the Fe-protein and the MoFe-protein. Their structures have been determined recently and show that both are α/β proteins. The Fe-protein is composed of two identical subunits connected by a 4Fe–4S cluster, while the MoFe-protein is an $(\alpha\beta)_2$ tetramer with structurally similar α and β subunits. Each αβ dimer coordinates two types of metal centres: the FeMo-cofactor and the P-cluster pair. At low levels of Mo, an apparently iron-only protein

Table 1

Programs used at the workshop.

Program [ref]	Type
BLAST [6], FASTA [7] and SSEARCH [8]	Pairwise sequence database searching.
BLOCKS [9]	Search against BLOCKS database of conserved regions using BLOCKSEARCH program [10].
MOTIFS [11]	Search against PROSITE motif database.
MPSEARCH	Server implementation of the Smith–Waterman alignment on a massively parallel machine.
SCANPS	Database scanning using a derivative of the Smith–Waterman algorithm (G Barton, unpublished data).
MaxHom [12]	Multiple sequence alignment.
CLUSTALW [13]	Multiple sequence alignment.
AMPS [14]	Multiple sequence alignment. The AMPS package also has many other functions.
GCG [15]	Sequence analysis package.
DSSP [16]	Pre-calculated dictionary of secondary structure.
SCOP [17]	Structural classification database.
THREADER [18]	Fold recognition: uses double dynamic programming to align a target sequence to a structure while evaluating the match using continuous statistically derived potentials.
ProFIT [19]	Fold recognition: aligns a target sequence to a structure while evaluating the match using a continuous statistically derived potential.
MAP	Fold recognition: reduces a secondary structure prediction to a string of secondary structural units and then searches the structure database for compatible domains (G Barton, unpublished data).
Topits [20]	Fold recognition: takes secondary structure prediction and accessibility prediction of PHD as input.
HMM [21]	Hidden Markov models are derived from multiple sequence alignments and can be used to search sequence databases for distant relationships.
PHD [22]	Predicts secondary structure, accessibility and transmembrane helices.
RUNPRED	A collection of existing secondary structure prediction methods (G Barton, unpublished data).
CORRELATION [23]	Prediction of long-range contacts between residues from correlated mutations.
SequenceSpace [24]	Prediction of functionally important residues.
PREDBB [25]	Prediction of long-range interactions between β-strand residues in β-sheets.
GLASS	'Graphical Language for Assembly of Secondary Structures' used to combine predicted secondary structures, predicted long-range interactions and information from multiple sequence alignments to enable all this information to be displayed while the predicted secondary structures can be manipulated as objects in 3D with the graphics program Insight II (R Leplae, unpublished data).

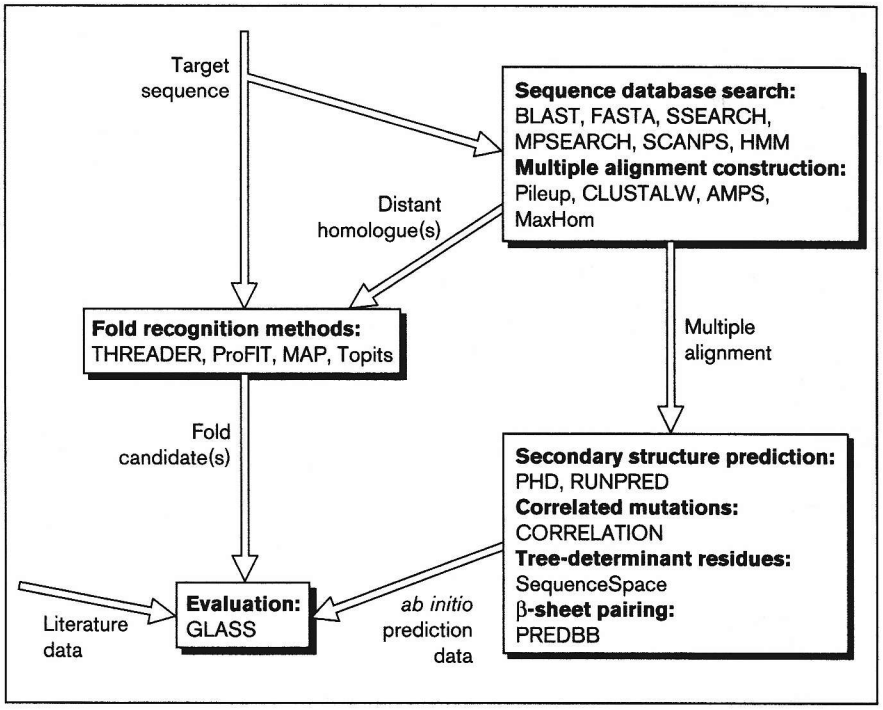
(FeFe-protein) is expressed with a shorter β subunit (lacking the N-terminal domain which wraps around the α subunit) and, intriguingly, the complex contains two additional δ subunits, whose structure and function are unknown. Our results indicated that the nitrogenase δ-chain is mainly helical. The single β-strand predicted by PHD is incompatible with an isolated folding unit, so it is either incorrectly predicted or must be part of a protein–protein interaction, perhaps forming an interface with another subunit in the $(\alpha\beta\delta)_2$ hexamer. Consistent with the above, fold recognition programs did not produce a plausible model for this β-strand, but their alignments with the two four α-helix structures (256B and 2HMQ) place hydrophobic residues in the core of the structure, as would be expected if the model was roughly correct. It has been proposed that the δ-chain plays a role in the stabilization of the quarternary structure of the hexamer, and that it is located near the N-terminal region of the β subunit, taking on the role of the missing short N-terminal domain of the β subunit in the MoFe-proteins. This latter fragment comprises four α-helices and a β-strand, as would the δ subunit according to our prediction. This similarity may be coincidental, however, as the sequence of the δ subunit is about twice as long as the N-terminal domain of the β subunit.

Target T0098 is the preprotein of the tick-borne encephalitis virus envelope glycoprotein M (prM), which interacts with the envelope glycoprotein E (of known structure) and blocks its pH-dependent fusion activity. A model of prM could shed light on the mechanisms of virus replication, activation and receptor binding. The programs consistently predicted all-β proteins of three main families: immunoglobulin (IG), plastocyanin (PLC) and retinol-binding-protein (RBP) like folds. The RBP fold was discarded because it has many more β-strands than predicted for prM and its fold is incompatible with the predicted presence of three disulfide bridges. An initial analysis of the model based on several IG-like folds and a PLC fold revealed that only in the latter could the six conserved cysteines form three disulfide bridges. As prM is believed to interact with E, we looked for correlated mutations between the two multiple sequence alignments: the correct prM model should cluster any predicted correlated residues onto the external surface. When the strongly predicted contacts are mapped onto the two possible models for prM, only in the IG-based fold do they cluster together on the surface. Neither an IG-based or a PLC-based model is therefore consistent with all our results. PrM is likely to be an all-β 'sandwich' or a 'barrel' fold but we cannot exclude a different topological arrangement of strands consistent with both cysteine distribution and correlated mutation localization.

Human αA-crystallin (target T0112) is an eye lens protein, usually found in large aggregates with αB-crystallin. αA-

Figure 1

Schematic guide to steps used in structure prediction at the workshop. References for the methods are given in Table 1.



and B-crystallins share 50–60% sequence identity. Previously published secondary structure predictions suggest the presence of two similar hydrophobic β-sheet-rich motifs connected by an hydrophilic α-helical region. The fold recognition result with the most convincing alignment was from ProFIT to β- and γ-crystallins, which both have the same fold but share only 37% sequence identity. ProFIT also identifies β- and γ-crystallin for the αB-crystallin with plausible alignments. There are four pairs of conserved residues in the alignment of α- and γ-crystallin which all map at the surface of a region of the second domain of γ-crystallin. While γ-crystallin has seven conserved cysteines, α-crystallins have only two, but in the γ-crystallin-based model they are in a plausible conformation to form a disulphide bridge. This evidence reinforced our view that the α-crystallin family is compatible with the β/γ-crystallin fold.

Target T0119 is the human arylamine *N*-acetyltransferase 1 (NAT1), a cytosolic enzyme that catalyzes the acetylation of arylamines from acetyl coenzyme A. It is widely expressed in human tissue and, together with its polymorphic homologue NAT2, is responsible for metabolism of a number of xenobiotic compounds. The possibility of a domain structure with separate binding sites for coenzyme and substrate (acetyl CoA and arylamine) had been proposed by the group that submitted the protein. Alignments produced by ProFIT for the two potential domains were compared with the PHD secondary structure prediction for the target protein and the DSSP information for the proposed fold. The alignment to 1CB1 is convincing and although the alignment to 1KNB requires numerous insertions and deletions, none interrupts a secondary structural element. The CORRELATION results were visualized using GLASS and used to map the NAT1 secondary structure elements predicted by PHD onto the corresponding secondary structure elements in 1CB1 and 1KNB. The results very convincingly suggest that NAT1 consists of two domains, the first an α-helical region similar to calbindin and the second a β-sandwich with a fold similar to that of 1KNB.

Target T0127 is human phosphatidylcholine-sterol acyltransferase (LCAT), a central enzyme in the extracellular metabolism of plasma lipoproteins. Although there is no overall sequence similarity between LCAT and other lipases, the sequence contains the PROSITE lipase pattern. Fold recognition using ProFIT not only identified a lipase fold, but produced alignments such that the active site residues are perfectly aligned between LCAT and both 1TCA and 1THG. A model of the protein based on the identified fold has two potential problems. It is known that LCAT contains two disulphide bonds (Cys74–Cys98 and Cys337–Cys380), but in the ProFIT alignment to 1THG only one pair of cysteines map to residues sufficiently close in space, and for the alignment to 1TCA neither do. Furthermore, the PHD secondary structure prediction, obtained using a single sequence, and the secondary struc-

Table 2

Prediction results.

Target T0092	**Nitrogenase δ-chain of *Rhodobacter capsulatus*.**
Sequence length	132
Family size	5
% identity	35–53%
PHD	α α α α β
RUNPRED	Similar to PHD.
THREADER	1DSB A chain, insertion domain (two slanted α-hairpins) and 2HMZ (four α-helix bundle).
ProFIT	1PVA (EF hand: two α-hairpins).
MAP	256B and 2HMQ (up-down-up-down four α-helix bundles).
Prediction	Up-down four α-helix bundle.
Submitted by	Eugen Krahn, Faculty of Chemistry (ACI), University of Bielefeld, Germany.
Target T0097	**Dichloromethane dehalogenase repressor DcmR.**
Prediction	None.
Submitted by	Stephane Vuilleumier, Mikrobiologisches Institut ETH-Zurich, Switzerland.
Target T0098	**Propeptide, envelope glycoprotein M (prM) from tick-borne encephalitis virus.**
Sequence length	91
Family size	19; six completely conserved cysteines; few large insertions and deletions.
% identity	23–92%
PHD	6 or 7 β: high reliability, except for first β-segment, which because of its length may contain two β-strands.
PREDBB	8 β: extra β-strand predicted in PHD β6–7 loop; β6 and β7 internal and antiparallel; β8 edge β-strand.
HMM	7/10 hits are IG-like folds. Second hit (2TBV) is also a β-sandwich.
TOPITS	IG-like in top five and other all β-folds.
ProFIT	IG-like (second hit 1CD8) and 1PLC.
THREADER	IG-like and RBP.
CORRELATION	Four strong contact positions between the prM and E molecules.
Prediction	β-sandwich or barrel, possibly IG or plastocyanin fold.
Submitted by	Aron Marchler-Bauer, Research Institute of Molecular Pathology (IMP), Vienna, Austria.
Target T0111	**Macrophage migration inhibitory factor.**
Prediction	None.
Submitted by	Graeme Wistow, 6/222, NIH, Bethesda, Maryland 20892-2730, USA.
Target T0112	**Human αA-crystallin.**
Sequence length	173
Family size	67: alignment optimized using MPSEARCH. No significant sequence similarity was found to N and C termini of small heat-shock proteins and these were excised, as were large inserts in some αA-crystallin sequences.
% identity	16–96%
PHD	Predominantly β.
RUNPRED	Predominantly β.
THREADER	2MSB-A (α/β protein), 1AAJ (α/β protein) amicyanin (nine β-strands) but poor Z scores.
HMM	Interleukin 1 (all-β protein).
ProFIT	Flavodoxin (α/β) and γ- and β-crystallin.
TOPITS	1GOF galactose oxidase (all-β). Poor Z scores.
PREDBB	Signal for one pair of parallel β-strands.
MAP	Many β-sandwiches.
Prediction	β/γ-crystallin fold.
Submitted by	Graeme Wistow, 6/222, NIH, Bethesda, Maryland 20892-2730, USA.

Table 2 (continued)

Prediction results.

Target T0119	**Human arylamine *N*-acetyltransferase type 1.**
Sequence length	290
Family size	14
% identity	28–95%
PHD	α α α α β β β β β α α β α β β β β α β β
ProFIT	In fragmentation mode: residues 1–78 Calbindin D9K (1CB1); 71–256 fibre protein from human adenovirus type 5 (1KNB).
CORRELATION	Specific interactions predicted between the N-terminal domain and residues on one face of the β-sandwich domain created on the 1KNB template.
Prediction	Identified as multi-domain: N-terminal α-helix bundle; C-terminal α/β fold.
Submitted by	John Sinclair, University Department of Pharmacology, Oxford, UK.
Target T0127	**Human phosphatidylcholine-sterol acyltransferase precursor.**
Sequence length	440: residues 1–24: signal peptide.
Family size	5 + 2 proteins of unknown function.
% identity	20–93%
PROSITE	Lipase family (residues 175–184).
PHD (single sequence)	19 β-strands and six α-helices.
ProFIT	1TCA and 1THG (lipases).
TOPITS	3/6 flavocytochromes; 3/25 lipases.
Prediction	Lipase fold.
Submitted by	Carla Vinals, URC Molecular Biology - FUNDP 5000 Namur, Belgium.
Target T0129	**Growth arrest and DNA damage inducible protein (Gadd45).**
Sequence length	165
Family size	5
% identity	55–96%
PHD	α β α β α β α β
THREADER	1PNE, 1ACF (profilins: α+β) and 3CHY (flavodoxin-like fold: doubly wound α/β).
ProFIT	3CHY (flavodoxin) and 1PFL (profilin).
MAP	Domain II of the A-chain of 1PFK; domain I of the B-chain of 1WSY and 3CHY (all α/β with mainly β-sheets).
PREDBB	Same β-strands as PHD and consistent with parallel topology.
Prediction	Flavodoxin-like fold.
Submitted by	Jong Park, MRC Centre for Protein Engineering, Cambridge, UK.
Target T0149	**NifA.**
Sequence length	240
Family size	47
% identity	29–75%
PHD	α β α β α β β α β α α β α α α
CORRELATION	Many correlations between predicted secondary structure elements except involving the first α-helix, the third β-strand and the last α-helix.
MAP	String of secondary structural elements used for searching excluded three listed above: mononucleotide-binding folds (1ETU, 5P21, 3ADK etc.); L-arabinose binding protein like (2LIV).
THREADER, ProFIT, TOPITS	Parallel β-sheet surrounded by some α-helices. No nucleotide-binding folds.
HMM	2LIV.
Prediction	Classic mononucleotide-binding fold.
Submitted by	Joel Osuna, Cuernavaca, Morelos, Mexico.

Table 2 (continued)

Prediction results.

Target T0167	**E7 protein (VE7_HPV16) from human papillomavirus type 16.**
Sequence length	122
Family size	48 (partial sequences discarded).
% identity	20–76%
CLUSTALW	Final alignment obtained by manual adjustment; two conserved Cys-X-X-Cys motifs.
PHD	α β β β α β (low reliability: α-helix$_1$, β-strand$_1$ and β-strand$_4$).
RUNPRED	Consistent with high reliability prediction of PHD.
TOPITS	1PRT (pertutoxin α+β).
ProFIT	Complete sequence: 1PRT; residues 45–98: 5PTI, 1DXT, 1KNT (5/20 BPTI-like folding class, ββα unit).
MAP	3GRS domain III (residues 365–478: α+β).
THREADER	Residues 45–98: 5PTI.
CORRELATION	Mostly between α1 and β3; β2 and β3; β2 and β4. Weaker contacts are predicted between β2, β3 and α2.
PREDBB	PHD predicted strands confirmed; possible additional antiparallel strand at C terminus; β2 and β3, β1 and β4 antiparallel.
Prediction	Zinc-binding domain with BPTI-like motif.
Submitted by	Peter Hjelmstrom, Department of Molecular Medicine, Karolinska Institute, Stockholm, Sweden.
Target T0174	**Small subunit of acetohydroxyacid synthase III from *E. coli* (ILVH).**
Sequence length	160
Family size	11
% identity	32–97%
PHD	β β β β α β β α β β α β
RUNPRED	Generally agrees with PHD, but α-helix$_1$ could also be a β-strand.
ProFIT	1NDC (nucleotide diphosphate kinase). Same hit with the yeast homologue (30% identity).
MAP	Mainly β proteins.
CORRELATION	Mostly between β1 and β3, β7 and β9; β9 and α2; β6 and α1 and β5 and β6.
Prediction	NDP kinase.
Submitted by	Tsiona Elkayam, Ben-Gurion University of the Negev, Beer-Sheva, Israel.
Target T0176	**Synaptobrevin homologue 2.**
Prediction	None.
Submitted by	Miriam Eisenstein, Department of Chemical Services Weizmann Institute of Science, Rehovot, Israel.
Target T0205	**ParR or StbB.**
Prediction	None.
Submitted by	Kenn Gerdes, Odense University Department of Molecular Biology, Campusvej Odense, Denmark.
Target T0217	**FixJC.**
Sequence length	76: C-terminal domain
Family size	50
% identity	23–55%
PHD	α α α α
PROSITE	Helix-turn-helix motif.
PREDBB	Complete protein: identified two-domain structure: FixJC: no β-strands; FixJN: results in agreement with the homologous known structure.
CORRELATION	Mostly between α1 and α2; α1 and α3; α2 and α3; α3 and α4.
MAP	1AVR, 1LMB, 1UTG and 2HMQ (all α-helical).
THREADER	1HYP and 1LEA (both α-helical).
ProFIT	In fragmentation mode: FixJN: 1NTR FixJC: 1FIA and 1HCR (all α-helical DNA binding).
Prediction	Helix-turn-helix, DNA binding.
Submitted by	Daniel Kahn, INRA/CNRS, Castanet-Tolosan, Cedex, France.

Table 2 (continued)

Prediction results.

Target T0218	**60S acidic ribosomal protein P1α (*S. cerevisiae*).**
Sequence length	61 (N terminus).
Family size	33
% identity	40–60%: 21% between P1 and P2
PHD	α α α α
TOPITS	Four α-helix bundles, repressors (three α-helices) and globins (all α-helical).
ProFIT	All α-helical proteins.
THREADER	All α-helical, especially small repressors and DNA-binding proteins.
CORRELATION	Mostly between: α1 and α4; α2 and α3.
Prediction	Four α-helices.
Submitted by	Alfonso Valencia, CNB-CSIC, Madrid, Spain.
Target T0220	**Heat-shock/chaperone protein Grpe (Hsp24) from *E. coli*.**
Prediction	None.
Submitted by	Alfonso Valencia, CNB-CSIC, Madrid, Spain.
Target T0221	**C-terminal domain of α-tubulin from *Sus scrofa*.**
Prediction	None.
Submitted by	Alfonso Valencia, CNB-CSIC, Madrid, Spain.

Databank codes [26]: 1AAJ, amicyanin; 1ACF, profilin I; 1AVR, annexin V; 1CB1, calbindin; 1CD8, cd8; 1DSB, dsba (disulfide bond formation protein); 1DXT, haemoglobin; 1ETU, elongation factor Tu (domain I); 1FIA, Fis protein; 1GOF, galactose oxidase; 1HCR, Hin recombinase; 1HYP, hydrophobic protein from soybean; 1KNB, adenovirus type 5 fibre protein; 1KNT, collagen type vi; 1LEA, LexA repressor DNA-binding domain; 1LMB, lambda repressor/operator complex; 1NDC, nucleoside diphosphate kinase; 1NTR, NTRC receiver domain; 1PFK, phosphofructokinase; 1PFL, profilin I; 1PLC, plastocyanin; 1PNE, profilin; 1PRT, pertussis toxin; 1PVA, parvalbumin; 1TCA, lipase; 1THG, lipase; 1UTG, uteroglobin; 1WSY, tryptophan synthase; 256B, cytochrome *b*562; 2HMQ, hemerythrin; 2HMZ, hemerythrin; 2LIV, leucine/isoleucine/valine-binding protein; 2MSB, mannose-binding protein a (lectin domain); 3ADK, adenylate kinase; 3CHY, CheY protein; 3GRS, glutathione reductase; 5P21, Ras-p21 protein; 5PTI, trypsin inhibitor.

ture of these folds overlap well only around the active site region in the ProFIT alignments. This could just reflect the high variability of the lipase fold, however, and so we still believe that LCAT adopts a lipase-like fold.

Gadd45 (target T0129) is involved in growth arrest in the case of severe DNA damage upon ionizing radiation or contact with mutagenic substances, which is a crucial event in preventing cell death and propagation of heritable genetic errors. Gadd45 seems to bind to two domains of the proliferating cell nuclear antigen (PCNA) with its N-terminal 95 residues. Gadd45 is predicted here to be an αβ structure with either a flavodoxin or a profilin fold. The effect of profilin on the action of epidermal growth factors hinted at a possible biological relationship to Gadd45, but when we searched SWISSPROT with an HMM built from the alignment of 24 profilin sequences the sequence of Gadd45 was not found. THREADER and ProFIT were run for the sequence least homologous to the target in the alignment (mouse MyD118 protein). The highest scoring structure with both programs was 3CHY and this fold is also consistent with the parallel β-sheet interactions predicted by PREDBB. These results suggest that the flavodoxin-like fold is more plausible than the profilin fold. The threading programs did not align the predicted N-terminal helix of Gadd45, but it is worth noting that this region contains several conserved negatively charged residues that may interact with a positively charged groove on PCNA.

NifA (target T0149) belongs to a class of bacterial enhancer-binding proteins that stimulate the expression of genes required for nitrogen fixation. NifA is composed of three functionally different domains. Experimental evidence indicates that the isolated central domain (240 residues) retains its biological function to stimulate DNA transcription. From the data obtained, we propose that this is a classic mononucleotide-binding fold. The 3ADK (adenylate kinase) template best fits with the predicted clusters of correlated mutations.

Target T0167 (human papillomavirus 16 E7) belongs to a family of transforming proteins involved in the pathogenesis of human cervical cancer. E7 is homologous to the adenovirus E1A oncoprotein and might have a similar transforming mechanism. E7 binds zinc in a 1:1 molar ratio and contains two Cys-X-X-Cys motifs in the C-terminal part, important for zinc binding and dimerization, but not for pRB binding. Either both motifs chelate the

same zinc ion or each zinc is coordinated by two Cys-X-X-Cys motifs, one from each monomer. The PHD prediction correlates very well with the CD results in the absence of zinc (PHD: α-helix: 16%; β-strand: 28%; CDapo: α-helix 16%; β-strand 27%; CDzinc: α-helix 29%; β-strand 11%). According to this prediction, the first Cys-X-X-Cys motif is located between strands β2 and β3. Fold recognition predicts 1PRT-like and BPTI-like folds. The structural alignment to these templates do not correlate well with the PHD secondary structure prediction, but all these folds are consistent with a tetrahedrally chelated zinc by the two Cys-X-X-Cys motifs, one within the β2–β3-hairpin and the other at the end of $\alpha\text{-helix}_2$. Correlated mutations and prediction of β-strand pairing are consistent with such a zinc-binding motif. Fold recognition is unlikely to find a model for the entire sequence due to the presence of the zinc, as none of the potentials used by these programs takes into account the effects of metal ions.

ILVH (target T0174) is part of a multimeric complex and interacts with a dimeric large subunit that belongs to the acetolactate synthases family. Very little is known about the small subunit. The fold recognition results, the pattern of conserved residues in the multiple sequence alignment and the correlation between β-strands in a putative central β-sheet support the existence of a conserved central core composed of three β-strands and one α-helix which is compatible with the 1NDC hit found by ProFIT. Although the 1NDC secondary structure and the PHD prediction do not perfectly overlap (the second predicted helix corresponds to a strand in 1NDC), the threading alignment is very convincing. It maps some of the ILVH conserved residues to positions conserved in the 1NDC family as well. These conserved residues cluster in space when mapped onto the 1NDC structure, thus defining two regions, one corresponding to the binding site, the other to interface regions of 1NDC. We are confident that the latter represents a reasonable model for ILVH.

FixJ (target T0217) is involved in the transcriptional activation of nitrogen fixation genes. It is formed by an N-terminal phosphorylated regulatory domain (128 residues) and a C-terminal transcriptional activator domain, FixJC. The structure of the FixJN domain can be modelled by homology (30% homology to 1NTR). The structure of FixJC domain shows no obvious homology with known structures and it is presently being studied by NMR spectroscopy. All our results are consistent with the hypothesis that FixJC is an all-helical protein with a helix-turn-helix motif similar to that of 1FIA and 1HCR. For our rough modelling using GLASS we used 1HCR, the DNA-binding domain of Hin recombinase complexed with DNA. This allowed us to tentatively position a DNA molecule interacting with the FixJC model and verify that the residue distribution in the interacting region is consistent with our model. The last predicted helix of FixJC was not modelled as there is no corresponding segment in 1HCR.

Target T0218 is a protein called P1α which belongs to a family of very acidic ribosome-binding proteins (P proteins) that are phosphorylated when bound to ribosomes. There are two subfamilies (P1 and P2) sharing 21% sequence identity. P proteins form heterodimers (P1–P2) and two such dimers form a pentameric complex with the P0 protein. The N-terminal domain of P proteins is needed for P1–P2 complex formation while the C-terminal part of P proteins is highly charged and likely to be exposed to the solvent. P1 is predicted to contain four α-helices, with contacts predicted between α1 and α4 (strong) and α2 and α3 (weak), but these interactions could be either intermolecular or intramolecular. Conserved, perhaps functionally important, residues are found in the region between α2 and α3 and SEQUENCE-SPACE identified tree determinant residues (i.e. residues able to discriminate between sub-families in a multiple sequence alignment) in α1 and α3 of P1 and α2 of P2. Fold recognition algorithms failed to identify a clear candidate fold. Despite the sequence similarity between P1 and P2 and the similarity of the secondary structure predictions, there are obvious differences in the distribution of correlated mutations and tree determinant residues. There are also clear asymmetries in the predicted contacts between P1 and P2. These predicted structural differences are likely to correlate with the functional differences between the two proteins.

Conclusions

One important conclusion of this experiment is that most of the target proteins selected could be predicted with some reliability by taking advantage of the availability of a number of different methods. Interpretation of the results was helped by critical evaluation from the authors of each method and, in a number of cases, from an expert in the biological and experimental background of the target taking part in the prediction.

One diagnostic of a reliable prediction that emerged during the workshop was the agreement between the results of different independent methods. Whether or not this will turn out to be a reasonable criterion will be verified only if and when an experimental structure is determined. It is encouraging to note that in the few cases where there was already suggestive (but not significant) information about the structure, the prediction results were able to independently support this. For example, αA-crystallin was predicted to have a fold similar to those of other crystallins, and a lipase structure was predicted for a sequence containing a diagnostic lipase motif. Similarly, the presence of a helix-turn-helix sequence motif in the FixJC sequence was noticed only after this fold had already been correctly identified.

As mentioned above, predictions were made to different levels of detail. In some cases a clear model structure could be identified and this allowed most of the important features of the protein to be mapped into three dimensions. In other cases only the rough arrangement of secondary structural elements could be predicted, but experiments could be designed to both test and improve such predictions.

It should be noted that the length of the workshop imposed limits on both the number of targets that could be selected and the number of methods that could be used on each of these. We expect that a number of the non-selected targets could also be predicted with similar levels of confidence and hope that the public availability of the raw analysis data, via the WWW-based database [5], will facilitate and encourage the prediction of these too.

Acknowledgements

We would like to express our gratitude to IRBM staff and in particular to Professor Riccardo Cortese for encouragement and advice; to the Information System and Technology Department for invaluable technical help; to Silicon Graphics, Q-Associates and Biosym for providing part of the hardware and software used during the workshop. We are also grateful to all the people who submitted target sequences. We are grateful to IRBM for financial support. T Hubbard is grateful to the MRC and Zeneca for financial support.

References

1. Moult, J., Pedersen, J.T., Fidelis, K. & Judson, R. (1995). Results of the 1994 Structure Prediction Competition and meeting 'Critical assessment of techniques for protein structure prediction'. *Proteins* **23,** ii–iv; http://iris4.carb.nist.gov/.
2. Shortle, D. (1995). Protein fold recognition. *Struct. Biol.* **2,** 91–93.
3. Lemer, C.M.-R., Rooman, M.J. & Wodak, S.J. (1995). Protein structure prediction by threading methods: evaluation of current techniques. *Proteins* **23,** 337–355.
4. Defay, T. & Cohen, F.E. (1995). Evaluation of current methods for *ab initio* protein structure prediction. *Proteins* **23,** 431–445.
5. Hubbard, T. & Tramontano, A. (1996). IRBM Practical Course Frontiers of Protein Structure Prediction. World Wide Web <URL: http://www.mrc-cpe.cam.ac.uk/irbm-course95/>\.
6. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.
7. Pearson, W.R. & Lipman, D.J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448.
8. Pearson, W.R. (1995). Comparison of methods for searching protein sequence databases. *Protein Sci.* **4,** 1145–1160.
9. Henikoff, S. & Henikoff, J.G. (1991). Automated assembly of protein blocks for database searching. *Nucleic Acids Res.* **19,** 6565–6572.
10. Fuchs, R. (1993). Block searches on VAX and Alpha computer systems. *Comput. Appl. Biosci.* **9,** 587–591.
11. Bairoch, A. & Bucher, P. (1994). PROSITE, recent developments. *Nucleic Acids Res.* **22,** 3583–3589.
12. Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9,** 56–68.
13. Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680.
14. Barton, G.J. & Sternberg, M.J. (1990). Flexible protein sequence patterns. A sensitive method to detect weak structural similarities. *J. Mol. Biol.* **212,** 389–402.
15. Devereux, J., Haeberli, P. & Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic. Acids Res.* **12,** 387–395.
16. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22,** 2577–2637.
17. Murzin, A., Brenner, S.E., Hubbard, T.J.P. & Chothia, C. (1995). scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536–540.
18. Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992). A new approach to protein fold recognition. *Nature* **358,** 86–89.
19. Sippl, M.J. & Floeckner, H. (1996). Threading thrills and threats. *Structure* **4,** 15–19.
20. Rost, B. (1995). *TOPITS: Threading One-Dimensional Predictions into Three-Dimensional Structures, Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology.* (Rawlings, C.J., Clark, D., Altman, R., Lengauer, T., Wodak, S., eds.), pp. 314–321, AAAI Press, Menlo Park, CA, Cambridge, UK.
21. Eddy, S.R., Mitchison, G. & Durbin, R. (1995). Maximum discrimination Hidden Markov models of sequence consensus. *J. Comp. Biol.* in press.
22. Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232,** 584–599.
23. Gobel, U., Sander, C., Schneider, R. & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins* **18,** 309–317.
24. Casari, G., Sander, C. & Valencia, A. (1995). Sequencespace: a tool for family analysis. *Nat. Struct. Biol.* **2,** 171–178.
25. Hubbard, T.J.P. & Park, J. (1995). Fold recognition and *ab initio* structure predictions using Hidden Markov models and β-strand pair potentials. *Proteins* **23,** 398–402.
26. Bernstein, F.C., *et al.*, & Tasumi, M. (1977). The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* **112,** 532–542.