doi:10.1016/S0022-2836(02)00698-8 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) not known, 1–14

JMB

De Novo Prediction of Three-dimensional Structures for Major Protein Families

Richard Bonneau[1], Charlie E.M. Strauss[2], Carol A. Rohl[1] Dylan Chivian[1], Phillip Bradley[1], Lars Malmström[1], Tim Robertson[1] and David Baker[1]*

[1]*Department of Biochemistry University of Washington, Box 357350, J-567 Health Sciences Seattle, WA 98195-7350, USA*

[2]*Biosciences Division, Los Alamos National Laboratory Los Alamos, NM 87544, USA*

**Corresponding author*

We use the Rosetta *de novo* structure prediction method to produce three-dimensional structure models for all Pfam-A sequence families with average length under 150 residues and no link to any protein of known structure. To estimate the reliability of the predictions, the method was calibrated on 131 proteins of known structure. For approximately 60% of the proteins one of the top five models was correctly predicted for 50 or more residues, and for approximately 35%, the correct SCOP superfamily was identified in a structure-based search of the Protein Data Bank using one of the models. This performance is consistent with results from the fourth critical assessment of structure prediction (CASP4). Correct and incorrect predictions could be partially distinguished using a confidence function based on a combination of simulation convergence, protein length and the similarity of a given structure prediction to known protein structures. While the limited accuracy and reliability of the method precludes definitive conclusions, the Pfam models provide the only tertiary structure information available for the 12% of publicly available sequences represented by these large protein families.

Keywords: Rosetta; structure prediction; gene annotation; structural genomics; Pfam

Introduction

As the number of gene sequences in databases, public and private, increase dramatically, so do the number of genes of unknown function. Of the protein sequences currently available approximately one-quarter have known function and approximately one-quarter have a link, *via* sequence homology, to a known structure.[1,2] Additionally, many proteins of known function contain domains of putative or unknown function.[3–5] The number of sequences with unknown structure is unlikely to plateau soon, since 40–66% of genes in newly sequenced genomes do not have significant sequence homology to proteins of known structure.[6] Because structure–structure relationships are conserved across greater evolutionary distances than are sequence–sequence relationships,[7,8] protein three-dimensional structures can in some cases reveal distant relationships not apparent from sequence information alone.[9,10]

De novo structure prediction can potentially provide a large number of structure models with considerably less investment of time, money and human effort than experimental approaches, albeit producing models of far lower reliability and accuracy. The recent CASP3 and CASP4 structure prediction experiments[11,12] show that Rosetta is probably the best current method for *de novo* protein structure prediction.[13–15] Rosetta is based on a picture of protein folding in which local sequence segments rapidly alternate between different possible local structures, and folding occurs when the conformations and relative orientations of these local segments combine to form low energy global structures.[16–19] The distribution of conformations sampled by an isolated chain segment is approximated by the distribution of conformations adopted by that sequence segment and related sequence segments in the protein structure database. Non-local interactions are optimized by

Abbreviations used: CASP, critical assessment of structure prediction; ORF, open reading frame; PDB, Protein Data Bank; RMSD, root-mean-square deviation.

E-mail address of the corresponding author: dabaker@u.washington.edu

0022-2836/02/$ - see front matter

Table 1. Rosetta performance on test set

1	2	3	4	5	6	7	8	9	10	11
			Best model match to PDB		Cluster center match to native			Cluster center match to PDB		
PDB id	Length	cThresh	*Z* (best)	SCOP (best)	Best rms in 5	Len-maxsub	rms-maxsub	Name	*Z*	SCOP
1a1z	82	2.96	11.96	5	9.02	63	3.69	1a0p0	8.74	1
1a32	65	2.08	9.83	4	6.39	63	4.39	1aep0	9.98	1
1a3k	110	8.39	7.80	4	13.84	45	3.20	1ygs0	5.23	1
1a68	87	5.45	8.74	5	8.92	53	5.07	1dqeA	5.47	2
1a6m	139	6.68	18.70	4	11.41	107	5.52	1cqxA	9.46	4
1a6s	83	3.91	12.01	5	10.75	84	5.05	1csh0	11.43	1
1aa2	97	5.88	9.37	4	9.58	79	5.20	1b0b0	8.39	1
1aba	87	7.38	7.40	5	11.93	49	5.15	1abv0	7.36	0
1aca	81	3.21	10.61	5	11.91	53	4.95	1bmtA	10.20	1
1acf	123	7.26	10.95	5	6.37	110	5.28	2prf0	11.89	5
1acp	73	2.80	7.53	5	5.08	69	4.77	1af80	10.81	4
1adr	76	2.47	10.49	5	7.11	67	2.83	1lmb3	9.87	4
1ag2	97	7.39	9.76	3.5	12.13	52	5.29	1b5eA	8.19	1
1agi	125	8.09	3.87	0	12.79	53	5.38	1ftpA	6.14	0
1ah9	70	3.47	9.99	5	11.59	42	4.91	1bj70	7.98	1
1ail	67	4.26	9.07	5	6.39	58	4.93	1qkmA	9.87	2
1aj3	95	2.86	9.83	3.5	12.73	45	2.90	1aep0	10.45	1
1am3	57	2.29	8.52	3.5	2.57	52	5.01	1d1dA	8.69	3
1aoy	78	2.72	11.41	5	6.27	63	3.71	1hstA	8.42	3
1ap0	45	2.29	5.46	0	9.74	39	5.04	1eciA	5.12	0
1apf	45	3.35	4.17	0	9.56	37	5.87	4ull0	5.42	0
1bb8	71	4.54	8.03	5	9.10	48	4.63	1g6aA	8.90	3
1bdo	75	5.72	8.56	5	9.58	60	3.23	1mdc0	8.22	1
1beg	96	7.40	8.67	5	10.88	58	5.39	1azsA	9.30	1
1bfg	126	8.28	5.77	1	12.78	42	3.90	1ospO	6.85	1
1bgk	27	2.30	2.84	0	2.31	27	2.65	–	–	–
1bor	41	2.57	4.67	0	8.07	39	5.64	2ifeA	5.88	0
1buo	121	7.09	9.32	5	11.86	6	5.23	1ehkA	7.01	2
1bw6	56	2.47	8.36	5	4.24	52	3.31	1bw6A	8.30	5
1c5a	62	2.43	8.47	4	5.55	10	4.54	1d8bA	9.51	1
1cc5	72	4.58	10.29	5	6.96	67	5.43	1blxA	7.61	1
1cm4	127	6.25	11.64	3.5	13.74	5	5.20	1ocp0	7.95	1
1cmr	26	2.42	2.61	0	4.29	26	3.58	–	–	–
1coo	81	2.35	11.72	5	8.03	78	3.78	1coo0	9.87	5
1cpq	118	5.68	12.41	4	7.40	91	5.12	1jafA	12.42	4
1csp	64	3.13	9.67	4	10.75	36	2.70	1df3A	7.01	1
1ctf	63	1.57	10.14	5	7.41	59	3.01	1lxa0	6.84	0
1cxc	124	8.31	5.76	3.5	11.89	54	5.63	1bqv0	6.17	1
1ddf	72	2.44	7.20	0	14.20	57	5.27	1bl0A	7.94	1
1dec	35	2.92	4.53	4	7.91	30	5.13	3sdhA	4.47	0
1dhn	121	6.43	8.16	4	10.71	75	5.52	1dx0A	7.14	2
1dvc	98	5.72	8.38	5	8.55	77	4.31	1eq6A	8.26	1
1eca	136	6.55	18.39	5	9.53	121	3.99	1mba0	13.70	4
1erd	29	3.07	3.28	1	6.29	29	3.94	–	–	–
1fbr	93	8.02	9.51	5	12.77	43	4.99	1dp0A-1	6.01	2
1fwp	64	4.70	7.78	0	10.59	38	5.21	1f5qB	8.31	2
1gab	47	1.80	6.80	1	2.61	47	2.43	1cuk0	6.74	1
1gb1	54	1.89	7.79	5	4.18	54	4.02	2igg0	7.80	5
1gpt	44	3.72	6.08	4	8.99	35	3.59	1b4bA	6.19	0
1gvp	82	7.44	4.99	0	10.00	44	5.83	1dixA	6.26	2
1hfc	137	7.69	4.63	0	14.37	30	3.90	1dt9A	6.22	1
1hnr	47	2.45	6.56	5	5.45	44	3.56	2igg0	6.98	0
1hp8	68	3.22	8.74	5	4.61	66	4.04	1b91A	9.18	1
1hqi	85	4.87	6.51	4	9.72	60	4.72	1ckv0	6.84	4
1hsn	62	2.39	8.03	5	4.96	55	2.91	1cf7B	8.63	1
1hyp	67	3.05	9.09	5	7.19	56	5.32	1elrA	7.81	1
1iyv	73	7.61	10.34	4	11.92	73	4.64	1ac6A	8.24	1
1jvr	74	4.27	9.09	5	6.43	59	4.51	1a0p0	11.37	1
1kde	65	4.16	7.01	3.5	7.47	56	4.57	1fchA	7.78	2
1kjs	74	2.91	9.76	5	3.26	73	4.02	1kjs0	10.96	5
1ksr	92	4.26	9.93	5	12.24	50	4.95	1bwyA	9.13	1
1kte	96	4.52	14.76	4	3.68	99	3.75	1jhb0	14.07	4
1lea	72	2.40	10.92	5	2.53	72	2.88	1lea0	10.86	5
1leb	63	2.11	9.51	5	2.45	63	3.35	1lea0	9.67	5
1lfb	69	4.86	6.94	2	3.87	65	3.64	1cfr0	8.07	0
1lis	111	5.51	11.84	5	13.33	87	3.89	2lisA	11.06	5
1lz1	109	7.84	5.31	0	12.07	56	5.62	1a28A	5.25	0
1mai	119	8.02	7.98	5	10.01	43	3.90	2orc0	6.72	0

(*continued*)

Table 1 Continued

1	2	3	4	5	6	7	8	9	10	11
			Best model match to PDB		Cluster center match to native			Cluster center match to PDB		
PDB id	Length	cThresh	*Z* (best)	SCOP (best)	Best rms in 5	Len-maxsub	rms-maxsub	Name	*Z*	SCOP
1msi	60	4.61	4.04	3.5	6.19	48	4.86	1gh8A	6.86	2
1mzm	71	2.63	10.76	5	3.55	71	3.53	1afh0	10.92	5
1ner	74	2.38	10.45	4	8.13	47	4.62	1adr0	9.18	4
1ngr	80	3.02	11.45	5	10.76	77	4.18	1guxB	10.39	1
1nkl	70	2.22	10.61	5	7.43	50	3.30	1ffkO	7.69	2
1nre	66	2.32	9.98	5	8.27	47	5.01	2occE	9.14	1
1nxb	53	4.11	5.78	4	7.35	41	4.73	1rgeA	6.19	0
1orc	56	2.19	7.68	5	8.60	41	3.50	2orc0	7.62	5
1otg	125	7.51	10.72	5	12.59	5	3.97	1et0A	8.53	2
1pdo	129	6.35	11.80	5	5.93	117	4.26	1pdo0	12.21	5
1pft	34	2.38	4.47	0	6.56	34	4.84	1d0qA	4.93	3
1pgx	57	2.13	8.52	5	3.78	56	3.92	1i50K	8.69	4
1poc	125	8.15	5.67	1	15.18	55	5.00	1ghc0	6.78	1
1pou	68	2.22	10.61	5	11.71	56	4.21	1alo0-0	8.75	1
1ppa	113	8.22	5.92	0	17.37	17	3.78	1kvoA	7.20	4
1pse	61	4.31	5.14	0	11.19	21	3.80	1fy7A	6.05	2
1ptq	43	2.36	4.60	0	9.38	29	5.96	1f6vA	6.01	2
1qyp	42	2.53	5.88	5	4.89	42	4.64	3proC	5.82	0
1res	35	1.03	4.53	1	1.58	35	1.97	1a5j0	4.73	3
1rip	74	5.05	5.59	1	12.86	63	5.78	1cd1A	6.50	1
1ris	92	4.02	12.09	3.5	5.01	84	4.47	2hhmA	10.73	0
1sap	60	2.45	6.75	5	9.45	50	5.56	1rblM	6.71	0
1stu	68	2.33	9.57	4	5.05	67	3.70	1qu6A	10.45	4
1svq	90	5.38	11.74	5	5.24	75	4.72	1svq0	9.58	5
1tif	59	2.51	8.10	5	4.96	58	4.18	1tif0	8.13	5
1tih	33	2.44	4.93	3.5	7.06	36	5.81	1f53A	4.06	2
1tit	85	4.76	12.85	5	5.51	72	2.97	1fhgA	10.41	4
1tnt	65	3.64	9.83	3.5	6.09	65	3.58	1g4dA	8.48	4
1tpm	41	2.28	4.70	0	7.34	31	5.11	1fjgC	5.88	2
1tsg	94	6.79	4.76	0	9.81	55	5.22	2hddA	6.26	0
1tul	97	7.22	5.74	5	9.77	55	5.13	1f0yA	7.65	2
1uba	44	2.39	6.19	3.5	5.66	41	3.13	1bed0	6.44	1
1utg	62	2.29	7.59	4	9.99	48	4.94	1afh0	8.85	1
1uxd	43	2.20	6.07	3.5	3.32	39	4.78	1uxc0	6.25	5
1vcc	65	3.73	10.24	5	7.66	72	3.91	1vcc0	6.63	5
1vif	48	3.71	5.40	0	7.96	35	4.31	2ait0	6.35	1
1vls	126	5.16	15.24	5	11.00	108	5.50	1c17M	12.09	0
1vqh	86	7.47	5.32	0	11.13	45	5.29	1qhlA	5.97	0
1vtx	36	2.45	3.95	4	6.39	27	5.22	1neq0	4.93	0
1who	88	5.42	7.95	1	9.97	54	4.77	1dzkA	8.99	3
2acy	92	7.03	10.04	1	6.51	65	3.25	1qm9A	9.42	4
2af8	86	3.34	12.65	5	5.67	79	4.77	1af80	10.17	5
2bby	69	3.11	9.88	3.5	5.64	65	3.71	1fy7A	9.13	3
2end	137	8.33	6.04	0	15.35	43	3.90	1xvaA	9.80	0
2erl	35	2.41	4.53	1	8.67	27	4.74	1a6jA	4.73	0
2ezh	65	2.22	9.83	5	4.13	57	3.71	1cfr0	9.12	0
2ezk	93	3.85	10.38	5	12.26	77	3.64	1quuA	11.61	1
2ezl	99	4.68	12.84	5	6.85	79	4.49	2ezk0	10.65	5
2fow	66	2.72	7.78	5	5.41	57	3.25	1aci0	8.43	5
2gdm	137	6.13	15.30	3.5	12.09	109	4.44	1mba0	15.70	4
2hp8	56	2.48	7.45	5	3.82	55	4.13	1hp80	7.17	5
2ife	91	3.63	13.42	4	4.53	83	4.56	1tig0	13.42	4
2ktx	34	2.56	4.33	0	7.38	28	4.34	1dv0A	4.53	0
2lfb	100	6.32	9.82	0	10.96	70	3.91	2ng10	8.13	1
2orc	64	2.72	5.59	5	4.82	57	5.47	1et0A	6.58	2
2pdd	43	2.55	6.07	1	3.66	43	3.36	1alvA	6.25	1
2ptl	60	2.07	9.02	5	2.70	60	3.07	2ptl0	9.18	5
2sn3	46	2.42	5.69	4	8.78	34	3.59	1ge9A	6.39	2
2u1a	76	3.04	11.17	5	6.30	60	2.91	1cvjA	8.68	4
2vgh	34	2.39	3.27	0	7.75	22	5.45	3pviA	4.53	0
3ait	60	3.98	4.60	0	11.78	51	5.67	1gcuA	7.64	2
3lzt	129	8.32	5.72	0	13.12	68	5.45	1bqbA	7.51	0
4fgf	110	8.25	4.80	5	12.28	43	3.60	1at3A	5.19	1

Column 2 is the median length of the three sequences folded. Column 3 is the average distance of all members of the largest cluster to the center of the largest cluster. Smaller clustering thresholds indicate that the program converged while thresholds greater than approximately 7.5 Å indicate lower expected model quality. Columns 4 and 5 show the results of using the Rosetta model with the lowest RMSD value to the experimental structure (out of the entire decoy set) to compare to the PDB with Mammoth. Column 4 is the *Z*-score of the highest *Z*-score match between the PDB and this best model, while column 5 is the SCOP assessment of the similarity of this matched protein and the correct structure: 0 indicates no match, 1 indicates a class match, 2 indicates a fold match, 3 indi-

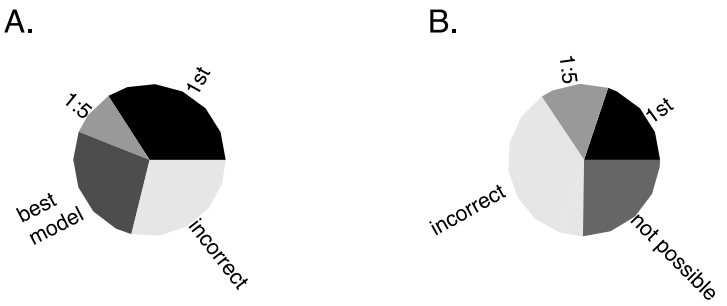

Figure 1. Overall performance of Rosetta on test set. (a) Ability of Rosetta models to identify the correct fold from the entire PDB for the test set. 1st indicates that the highest Z-score match of the top five models to the PDB identified the correct structure or a protein in the same superfamily as the correct structure (the first ranked fold identification was correct). 1:5 indicates that one of the top five models had a highest Z-score match that was correct. best model indicates that the best model in the entire decoy set (prior to model selection) for a given protein identified a protein in the correct SCOP superfamily as first rank. (b) Same as (a) except with pairs of proteins with significant sequence similarity removed in order to simulate the performance expected for Pfam-A families with no sequence-detectable links to known structures. not possible indicates that the removal of all sequence similar proteins from our non-redundant set of experimental structures removed all possible correct folds matches from the set according to SCOP.

a Monte Carlo search through the set of conformations that can be built from the ensemble of local structure fragments for each sequence segment. The procedure thus results in structures that have low free energy local and non-local interactions.

While high-resolution experimental structures are required for detailed functional and mechanistic insight into protein action, lower resolution *de novo* structure predictions can in some cases provide functional insights.[20] Methods for obtaining functional information from protein structures or predicted protein structures fall into two categories. The first class of methods uses libraries of motifs consisting of a small number of residues with specified spatial arrangements to search for specific types of functional sites[21–24] and can be readily combined with weak sequence pattern matches.[25,26] The second class of methods searches for larger, sequence-independent matches of the structure of the protein to previously determined protein structures;. These methods exploit the observation that two structures having a common fold often share at least some aspect of their function.[9,10,27] A number of previously described algorithms are available for carrying out the required structure–structure comparsions (Mammoth,[28] Dali,[29,30] CE[31]).

Large protein families for which no member has a known three-dimensional structure are particularly attractive candidates for *de novo* structure prediction because a single model can provide insights into the structure and function of a large number of sequences. In particular, the Pfam-A database[3] contains 2800 sequence families that represent 65–70% of the proteins in SWISSPROT and TrEMBL. The Pfam database has the additional advantage that sequences have been parsed, when possible, based on sequence homology patterns into single domains families.

Here we use Rosetta to generate models for Pfam-A domains of less than 150 amino acid residues in length without links to known structures. We then use a sequence-independent structure–structure comparison to the PDB[32] to identify proteins with similar structures that may have related functions. In order to assess the accuracy of the predictions and the reliability of the fold links obtained by structure matches to the PDB, structure–structure comparisons between Rosetta predictions and the PDB were made for a large training set of proteins of known structure. The models generated for Pfam families provide many interesting preliminary insights into the vast expanse of molecular evolution yet to be uncovered by structural genomics efforts.

Results and Discussion

Performance on test set

To provide a benchmark for deriving confidence measures for Pfam predictions made with Rosetta, we generated Rosetta models for a test set consisting of 131 proteins. Of these 131 proteins 101 were the test set used by Simons[33] and an additional 30 proteins in the size range of 110–150 residues were added to bring the size distribution of the test set into accordance with the size distribution of the 510 Pfam families for which predictions were generated. The fraction of β,α and α/β

cates a superfamily match, 3.5 indicates that one of the two proteins were not in SCOP at the time this study was carried out and that a Mammoth comparison of the two proteins gives a Z-score of 5.0 or greater, 4 indicates a family match, and 5 indicates that the match was to the same protein. Column 6 is the best RMSD value to the experimental structure over the entire protein length for the top five ranked models. Columns 7 and 8 show the number of superimposable residues and RMSD value of the MaxSub alignment of the best model in the top five ranked models to the correct structure. Columns 9–11 are the name, Z-score and SCOP match designation of the highest Z-score match to the PDB without removing the correct structure or sequence-similar proteins from the PDB prior to searching.

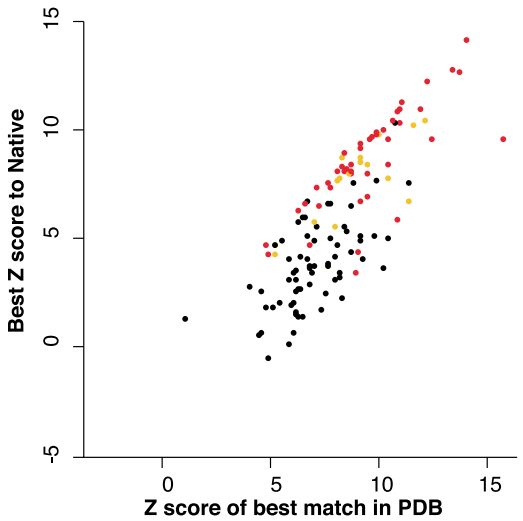

Figure 2. Relationship between best match *Z*-score and model quality. The *x*-axis is the best *Z*-score found when the top ten ranked models for each protein in the test set were searched against the PDB using Mammoth. The vertical axis shows the best *Z*-score of these top ten models to the correct structure, and is thus one way of expressing overall model quality for each protein. Proteins for which no model had a significant *Z*-score to any member of the PDB (including the correct structure) are not included in this Figure (models with a minus (−) designation in column 10 of Table 1). Points are colored according to the degree to which the correct SCOP superfamily was identified for that protein: red indicates that the highest *Z*-score match between any of the top ten models for a given protein and the PDB was in the correct SCOP superfamily; orange indicates that one of the top ten models had its highest ranking match to the correct SCOP superfamily and blue indicates that the correct SCOP superfamily was not identified by any of the top ten models.

proteins in the test set was similar to the proportions seen in the SCOP and CATH protein structure databases.[34,35]

Throughout this study we focus our attention on the top five ranked models for each protein (see Materials and Methods for the ranking procedure).[13,19] There are several useful metrics for describing the quality of these five models. One is to describe in absolute terms how structurally similar each model is to the experimentally determined structure. Columns 7 and 8 of Table 1 show the length and RMSD value of the best MaxSub[36] alignment between the correct structure and any of the top five ranked models for the test set (MaxSub[36] is designed to find the longest superposition of a model onto a correct structure). For 80 of the 131 proteins, 50 or more residues were superimposable on the experimental structure to within 6.0 Å RMSD (or, for proteins less than 50 residues in length, had a global RMSD value of less than 3.0 Å). This level of success is consistent with the performance of Rosetta in CASP4 and with previously published tests of the method.[13,14,16]

A second measure of success is the degree to which the structure-based searches of the PDB using the top ranked models can identify the correct fold family. To evaluate performance in this way we used a measure of success based on the SCOP fold classification database.[34] A prediction was considered successful if its closest structural match to the PDB using the Mammoth[28] program (see Materials and Methods) was in the correct SCOP superfamily (Figure 1(a) and Table 1). For 44 of the 131 proteins in the test set, the closest structural match to any of the top five ranked models belonged to the same SCOP superfamily as the experimental structure. For an additional 13 proteins in the test set, at least one of the top five models best matched a structure in the correct SCOP superfamily. Thus, for 57 of the 131 proteins the correct superfamily could be narrowed to five or fewer possibilities (Figure 1(a)).

For an estimate of performance on sequences for which no link to known structure is detectable by sequence homology, structure–structure matches that can be recognized solely by sequence similarity using methods such as PSI-BLAST[37] must be removed. Figure 1(b) shows the performance of the method when predictions were considered correct if the search of the PDB identified the correct SCOP superfamily excluding structure pairs with PSI-BLAST *E*-values of 0.001 or lower. For 26 of the 131 proteins in the test set, the closest structure match to the PDB was in the correct SCOP superfamily. For an additional 19 of the 131 proteins the correct superfamily was identified by one or more of the top five ranked models. For 33 of the proteins in the test set, removing all sequence similar proteins removes every member of the correct SCOP superfamily, so that for these 33 cases, even a perfect structure prediction would fail to identify any correct fold matches. Thus, the correct fold was identified to within five possibilities in 45 of the 98 cases for which it was possible to identify the correct superfamily after removing all sequence-recognizable links.

While absolute model accuracy is generally higher for smaller proteins (<50 residues), success in using the models to identify SCOP classifications is generally higher for larger proteins. Of the proteins in the test set under 50 residues in length, few succeeded in identifying the correct SCOP superfamily or a protein with a significant Mammoth *Z*-score to the experimental structure, even though many of the predictions for these small proteins were correct to RMSD values of less than 3.5 Å. This failure with small proteins may reflect either the reduced information content in a shorter structural match (which may be more likely the result of convergent evolution or random chance) or a breakdown of the statistics used in structure–structure comparison for very small proteins.

For most proteins in the test set, better models existed than were chosen as the top ranked models. Columns 4 and 5 of Table 1 show the

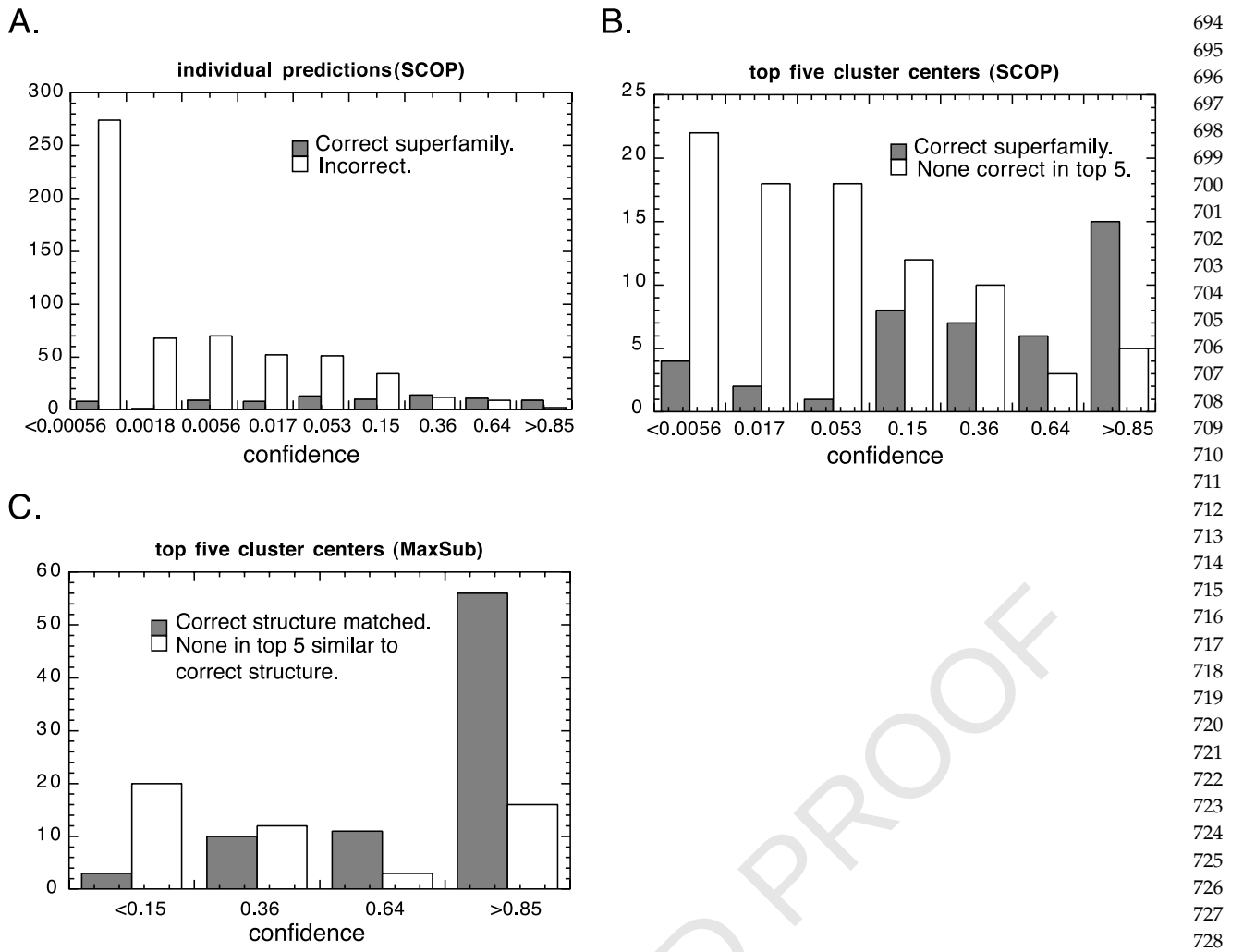

Figure 3. Prediction of model and annotation confidence The histograms reveal how well each of the three confidence functions discriminate false positives from true positives (and negatives). To varying degrees, the confidence functions concentrate the poor predictions on the left and the correct predictions on the right. Counts for successful predictions are shown in gray, while counts for incorrect predictions are shown in white. The three definitions of success are as follows: (a) The best Mammoth match identified for a given cluster center was in the correct SCOP superfamily. (b) At least one of the top five cluster centers was matched by Mammoth to a protein in the correct SCOP superfamily. (c) One of the top five models was superimposable on the correct structure for 50 residues or more (according to MaxSub) or has a global RMSD value of 3.0 Å or less.

results of using the model closest to the experimentally observed structure (as judged by global RMSD) to search against the PDB using the same procedureas used for the top ranked models. Using the model closest to the experimental structure,the correct superfamily according to SCOP is identified for 95 of the 131 proteins ("best model" in Figure 1), roughly double the success rate obtained when cluster centers are used to search the PDB. This result highlights the continued need for improved methods to rank models: better discrimination between correct and incorrect conformations could double the number of correct structural relationships identified.

The errors associated with Rosetta necessitate a method for estimating the likelihood that a given model is correct. Towards this end we have created confidence statistics for judging Rosetta predictions. The confidence functions used in this work are based on simulation convergence (as measured by clustering threshold; see Materials and Methods), protein length, and (when applicable) the structural similarity to the most similar structure in the PDB (Mammoth *Z*-score),[28] as described in Materials and Methods. The clustering threshold was previously shown to correlate inversely with model accuracy. Rosetta simulations that fail to converge tend to result in incorrect structure predictions, while tightly converged Rosetta simulations result more often in correct top ranked models.[13] A new result in this study is the observation that the degree to which top ranked models for a given protein match structures in the PDB is a strong indicator of the likelihood that one of the top ranked models is correct (Figure 2). Confidence functions were computed both for individual top

Table 2. Results for Pfam families with structures solved after predictions were made

Pfam	PDB	Model/fold	Confidence	Best Z in 5	Best pdb match (top 5)	Match to native Z
PF00015	1qu7-A	(+/+)	0.85/0.77	5.65	1ffk-S(7.74)	7.39
PF00164	1fjf-L	(−/+)	0.16/0.15	3.59	1ffk-A(5.52)	5.22
PF00253	1fjf-N	(−/−)	0.15/0.251	1.52	na	na
PF00533	1cdz-A	(+/−)	0.66/0.69	5.19	1bjx(5.56)	4.08
PF00570	1d8b-A	(+/+)	0.85/0.85	10.09	1ith-A(7.21)	12.4
PF00672	1joy-A	(−/−)	0.85/0.85	2.67	1e2a-A(7.60)	4.53
PF01393	1dz1-A	(−/−)	0.75/0.75	3.91	1hyw-A(5.96)	2.39
PF01909	1fa0-A	(+/−)	0.15/0.15	5.44	na	na
PF01984	1eij-A	(+/+)	0.85/0.84	6.26	1eo0-A(7.71)	5.63
PF02013	1e8r-A	(−/−)	0.15/0.15	1.84	na	na
PF02151	1qoj-A	(−/−)	0.15/0.15	3.69	na	na
PF02186	1d8j-A	(+/+)	0.85/0.69	5.43	1eeo-A(6.75)	5.89

Column 1 is the Pfam identification number. Column 2 is the PDB id for the family member recently solved. Column 3 provides a binary indication of prediction quality:the first + (−) indicates good (bad) model quality, and the second + (−) indicates a correct (incorrect) superfamily identification, according to the criteria given in Materials and Methods. The confidence that one of the top five ranked models is a correct structure prediction and the confidence that one of the top five ranked models provides a correct SCOP superfamily identification are reported, respectively, in column 4. Note that the confidence function erred in three cases (shown in bold), producing two false positives and one false negative. Column 5 is the best *Z*-score to the correct structure among the top five ranked models. For columns 6 and 7, the closest Mammoth matches in the PDB for each of the top five ranked models were each compared using Mammoth to the correct structure, and the match with the highest *Z*-score was selected. Column 6 gives the match name and *Z*-score to the closest cluster center. Column 7 shows the *Z*-score between the match shown in column 6 and the correct structure (column 2); a high value in this column denotes that Mammoth search with one of the top five ranked models identified a protein with significant structural similarity to the correct structure.

ranked models (Figure 3(a)) and for the top five ranked models as a group (Figure 3(b) and (c)). The group confidence reports on the reliability of predictions for a protein or protein family, while the individual confidence metrics help to rank predictions within a high confidence group. Although individual confidences are in principle more useful than the group confidences assigned to the top five ranked models, the latter are better determined by our training set and are likely to be more accurate.

Pfam results

Predicted models were generated for 510 Pfam families with average sequence lengths ranging between 35 and 150 residues using the same method as for the test set calculations described above (see Materials and Methods).[16] In the time elapsed since these predictions were made, a structure has been determined for one or more sequences in 12 of the families. Table 2 shows the quality of our predictions for these 12 protein domains. For six of these proteins, at least one of the top five ranked models matched the experimentally observed structure with a Mammoth *Z*-score of 4.0 or greater, roughly correlating to a correctly predicted region of greater than 50 residues. For five of the 12 solved families, one of the matches of the top five models to the PDB was in the same SCOP superfamily as the correct structure (Table 2, column 6). Our rate of success on these 12 families (50% good model quality, 41% correct fold links) is comparable to the success rates obtained for the test set (61% good model quality, ~34% correct fold links).

From analysis of these 12 protein families, as well as the results seen with the training set, we estimate that for 50–60% of the Pfam families for which predictions were made, one of the top five ranked models has significant structural similarity to the correct structure of the Pfam domain. These families represent 12% of publicly available protein sequences, and for most of these families, the Rosetta models generated here are the only three-dimensional structural information available.

When the top ranked models for each family are compared to the structures in a non-redundant subset of the PDB, significant structural matches are obtained for a large fraction of the families predicted, as expected from the results of the training set. While the majority of the fold linkages predicted here have not been previously reported using fold recognition methods, it is important to note that the primary goal of this study is not to improve upon or compete with established fold recognition methods, but to provide structural information for protein domain families for which no other structural information is currently available by any other method. Fold recognition methods cannot provide a prediction if a similar known structure does not exist, whereas *de novo* prediction methods can. Consequently, the most interesting models in the Rosetta-generated database are likely those for which structure matches to known structures cannot be detected.

Similarly, the 510 Pfam families for which models were generated include both extensively studied proteins and proteins about which little is known. The models are likely to be particularly useful for uncharacterized protein families: they may provide a framework for interpreting existing data or yield clues about function. In the present discussion, however, we focus on example Pfam families of known function because functional predictions based on fold matches can be directly compared to published annotations. There is

Table 3. Pfam prediction examples

pfam_id	Pdb	Chain	zscore	Single conf.	Length	1 in 5 conf.	GenThreader	SCOP superfamily
PF00601	1AJ3	0	12.23	0.31	105	>0.85	na	Spectrin repeat
PF00677	1EP3	B	9.23	0.39	85	>0.85	1i8dA1(0.63)	Ferredoxin reductase-like, FAD-binding (N-terminal) domain
PF00855	1B08	A	7.1	0.36	74	0.49	na	C-type lectin-like
PF00936	1F0Y	A	7.17	0.23	87	>0.85	na	NAD(P)-binding Rossmann-fold domains
PF00938	1ALL	A	10.5	0.71	92	>0.85	na	Globin-like
PF01047	1DPU	A	7.37	0.7	108	0.65	1b9mA1(0.61)	"Winged helix" DNA-binding domain
PF01059	1DI1	A	10.64	0.22	104	>0.85	na	Terpenoid synthases
PF01104	1SLT	A	8.34	0.57	82	>0.85	na	Concanavalin A-like lectins/glucanases
PF01124	3MDD	A	11.17	0.33	89	>0.85	na	Acyl-CoA dehydrogenase (flavoprotein), C-terminal domain
PF01155	1C1D	A	8.55	0.14	117	0.81	na	Aminoacid dehydrogenase-like, N-terminal domain
PF01277	1VRE	A	9.83	0.18	99	>0.85	na	Globin-like
PF01307	1HMC	A	7.91	0.11	99	0.77	na	4-helical cytokines
PF01392	1HP8	0	8.12	0.13	116	0.38	na	p8-MTCP1
PF01399	1BM9	A	9.68	0.67	75	>0.85	na	"Winged helix" DNA-binding domain
PF01445	2BID	A	7.79	0.21	56	0.82	na	Bcl-2 inhibitors of programmed cell death
PF01519	1HW1	A	9.34	0.17	116	0.64	na	"Winged helix" DNA-binding domain
PF01542	2NG1	0	7.82	0.17	75	0.85	na	Domain of the SRP/SRP receptor G-proteins
PF01675	1TAF	B	8.06	0.67	80	0.62	na	Histone-fold
PF01713	1EKE	A	8.8	0.38	75	>0.85	na	Ribonuclease H-like
PF01809	1NKL	0	7.32	0.56	68	0.69	na	Saposin
PF01883	2TSR	A	10.39	0.48	96	>0.85	na	Thymidylate synthase/dCMP hydroxy-methylase
PF01903	1C2Y	A	11.9	0.3	113	>0.85	1hrkA0(0.611)	Lumazine synthase
PF01918	1DCT	A	9.66	0.29	82	>0.85	na	*S*-adenosyl-L-methionine-dependent methyl-transferases
PF01938	1SRO	0	8.8	0.62	58	0.84	na	Nucleic acid-binding proteins
PF01938	1SRO	0	6.8	0.48	58	0.84	na	Nucleic acid-binding proteins
PF01985	1TIG	0	10.71	0.76	70	>0.85	na	Translation initiation factor IF3
PF02020	1DVK	A	8.8	0.57	79	>0.85	na	Functional domain of the splicing factor Prp18
PF02109	1DI1	A	10.3	0.23	112	>0.85	na	Terpenoid synthases
PF02346	1DF4	A	8.69	0.76	89	>0.85	na	Virus ectodomain
PF02379	1BYK	A	9.54	0.35	104	>0.85	na	Periplasmic binding protein
PF02379	2DHQ	A	12	0.63	104	>0.85	na	3-Dehydroquinate dehydratase
PF02440	1B56	0	9.67	0.15	75	>0.85	na	Lipocalins
PF02517	2CB5	A	9.06	0.14	89	>0.85	na	Cysteine proteinases
PF02519	1AQB	0	8.06	0.44	79	0.72	na	Lipocalins
PF02619	1BIA	0	8.26	0.34	95	0.57	na	Class II aaRS and biotin synthetases

Fold matches are shown for several of the Pfam families predicted. Columns 2 and 3 show the PDB id and chain of the protein match. Columns 4 and 5 show the *Z*-score of the match and the resultant confidence for the fold link shown. Column 7 shows the confidence that one of the top five models generated for this Pfam family identifies the correct SCOP superfamily. The SCOP superfamily for the PDB matched is given in column 9. Column 8 shows the GenThreader results (http://bioinf.cs.ucl.ac.uk/psiform.html: profile and secondary structure input were used) on the Pfam families shown. "na" indicates that GenThreader's top match was low confidence (low or guess as designated in the server output). For the remaining three hits of "medium" confidence, the PDB identifier of GenThreader's top hit and the confidence value for that hit are shown.

clearly much more information in the database of models than can be extracted here. In particular, for the Pfam families that cannot be linked to known structures by any method, one of the Rosetta predictions is expected to have significant structural similarity to the true structure for about half of the families. While potentially quite valuable, the models must be used with caution, as there is a significant probability that any given model is incorrect.

Individual predictions

Several representative predictions yielding interesting fold links are shown in Table 3. None of these fold links were detected by PSI-BLAST at the time this paper was prepared. In many of the cases described here, putative or previously existing annotations bolster the confidence assigned to a prediction by our automatic function. For each prediction we report the individual and group confidences provided by our automatic functions but do not attempt to quantify the degree to which functional similarities improve each prediction's likelihood of being correct (see Table 3).

For comparison to a previously described fold recognition method, we have used GenThreader†[1] with profile and secondary structure for each of the Pfam families represented in Table 3. For two of the examples in Table 3, GenThreader produced

† http://bioinf.cs.ucl.ac.uk/psiform.html

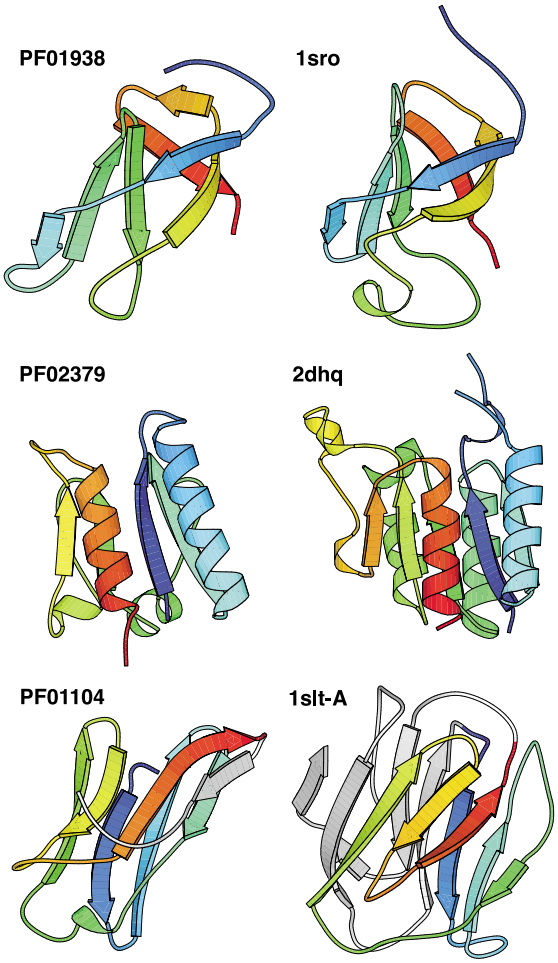

Figure 4. Pfam prediction examples. (a) The predicted model for PF01938 is shown beside the closest match to this model in the PDB, 1sro. (b) The model with the strongest match to the PDB for PF02379 is shown next to 2dhq. (c) The highest confidence model for PF01104 and 1slt-A (animal S-lectin), described in Table 3 and the text.

a top match to the same fold (PF01047 and PF00677). In one case, GenThreader produced a top match that was in a different SCOP superfamily than the fold predicted by Rosetta (PF01903). In all other cases GenThreader found no significant matches to the PDB, designating them low confidence or "guess". These results show that Rosetta derived fold linkages are at least partially orthogonal to template- based fold recognition techniques in most instances. How to best quantify the statistical significance of conflicts and/or agreements between Rosetta predictions and fold recognition methods is an important area of future research.

PF01938 to 1sro

The TRAM domain, PF01938, is a small domain of unknown function, suspected to be a nucleic acid binding protein. We find a strong match to 1sro,[38] a nucleic acid binding domain belonging to a SCOP superfamily containing many diverse RNA, DNA and ssDNA binding proteins, thus supporting the previously existing putative annotation (Figure 4(a)).

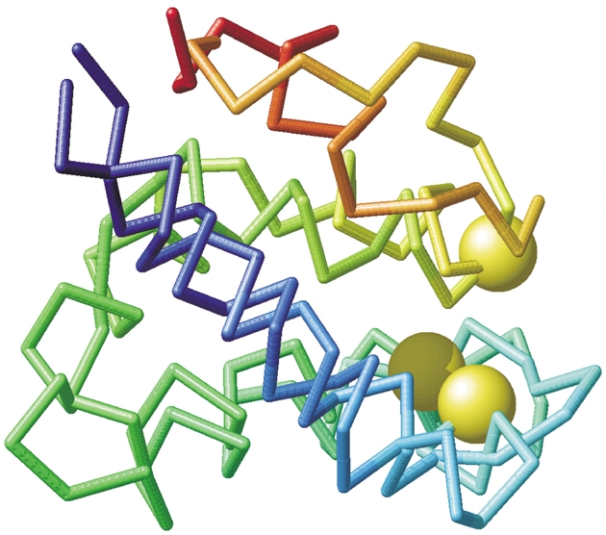

Figure 5. PF01809. The prediction for PF01809 is shown superimposed on 1nkl, the strongest match to the PDB for this model. The three conserved cystine residues found in PF01809 are indicated as yellow spheres on the predicted model for PF01809.

PF01809 to 1nkl

This domain of approximately 70 residues is found in short hypothetical proteins in many different bacteria and has no known function. One member of this family is SWISSPROT:Q44066, putatively annotated to have hemolytic activity (unpublished results). Mammoth search with the models generated for PF01809 identifies NK-lysin (1nkl)[39] (69% group confidence, 56% individual confidence), a hemolytic protein expressed in natural killer T-cells, supporting this putative annotation (Figure 5). In addition to the fold match, the proteins in family PF01809, have a large net positive charge; the protein predicted in Figure 7 has a net charge of +7 while NK-lysin and Bactereocin-AS-48 have net charges of +5 and +8, respectively. This net charge is critical to the mechanism proposed for these lytic proteins,[40] termed molecular electroporation, and is consistent with the fold prediction and the putative function annotation.

PF02379 to 1e2b

This Pfam family consists of the fructose-specific IIB subunit of the bacterial phosphoenolpyruvate: sugar phosphotransferase system (PTS). Structure similarity searches using the top ranked models for this Pfam family identify several different superfamilies that share a three-layer structure of two helices, a four-stranded beta sheet, and two

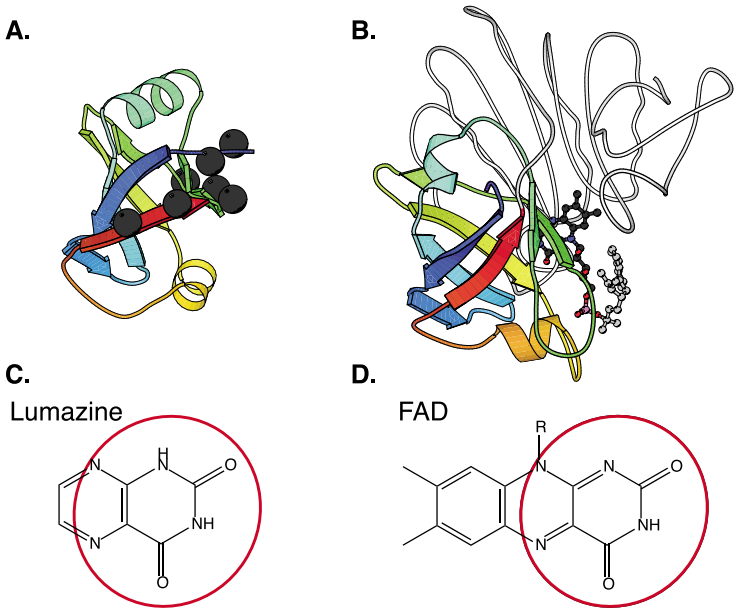

Figure 6. PF00677. The predicted model for PF00677 is shown next to 1ep3 chain B, the closest Mammoth match in the PDB. Proteins in PF00677 bind lumazine while 1ep3-B binds FAD, which is chemically similar to lumazine; this similarity in function supports our fold identification. Additionally, most of the strongly conserved residues for this Pfam family (indicated by grey spheres), cluster in a region of the fold where residues responsible for FAD binding are located in 1ep3-B.

helices. The strongest structure match to the PDB, 2dhq, is shown in Figure 4(b).[41] One of the proteins matched is 1e2b,[42] the IIB component of the phosphoenolpyruvate-dependent phosphotransferase system of *Escherichia coli* cellobiose transporter. Despite the obvious similarity in the functional descriptions of PF02379 and 1e2b, the relationship was not detected by PSI-BLAST or the current GenThreader server[1] at the time this paper was prepared.

PF00677 to 1ep3-B

This family consists of several proteins involved in Lumazine binding. The top ranked model for this family structurally matches 1ep3-B,[43] dihydroorotate dehydrogenase B, with a *Z*-score of 9.23, a group confidence of >85%, and an individual confidence of 39%. Dihydroorotate dehydrogenase binds FAD as a cofactor. Chemical similarities between FAD and lumazine and the sequence conservation pattern for the family support the prediction (Figure 6). GenThreader identifies a match to a protein in the same superfamily that was solved after our Rosetta and Mammoth calculations were carried out.

PF01713 to 1ekeA

PF01713 contains the small mutS related protein (Smr) and mutS2 (sometimes referred to as mutSB). These proteins, PF01713, have no detectable global sequence similarity to mutS as detected by PSI-BLAST, GenThreader and 3D-PSSM,[2] but they are functionally similar. We find a structure match between a top ranked model and 1ekeA,[44] a member of the ribonuclease-H SCOP superfamily. This SCOP superfamily contains several proteins that have functions related to mismatch repair and has strong structural similarity to the second domain of mutS (1ewq).[45]

Most of the above links are between protein domains of roughly equal size. Many of the predicted models for Pfam domains matched parts of known structures/domains, but because the proteins that contain Pfam domains usually have additional N or C-terminal domains, or regions that could complete the structural unit, partial structural matches can still be functionally significant. Two such examples are described below.

PF01059 to 1di1-A

This family is the NADH-ubiquinone oxidoreductase chain 4 Nterminus. Models generated for this family match 1di1-A, squalene synthase.[46] The chemical similarity between the isoprenoid tail of ubiquinone and squalene, as well as the fact that both proteins use either NADH or NADPH as a co-enzyme, suggest a distant evolutionary link (Figure 7). Another Pfam family, PF02326, containing plant proteins of unknown function, also showed a strong link (*Z*-score = 12.43, >85% group confidence, 22% individual confidence) to 1di1-A, suggesting a possible function involving terpenoid /isoprenoid binding for PF02326.

PF01104 to 1slt-A

This family contains Bunyavirus non-structural protein NS-s sequences from several members of the family *Bunyaviridae*. Our second model for this family matches animal S-lectin (1slt-A),[47] with a *Z*-score of 8.3, suggesting that the domain may interact with extracellular glycoproteins and be important for host-viroid interactions (Figure 4).

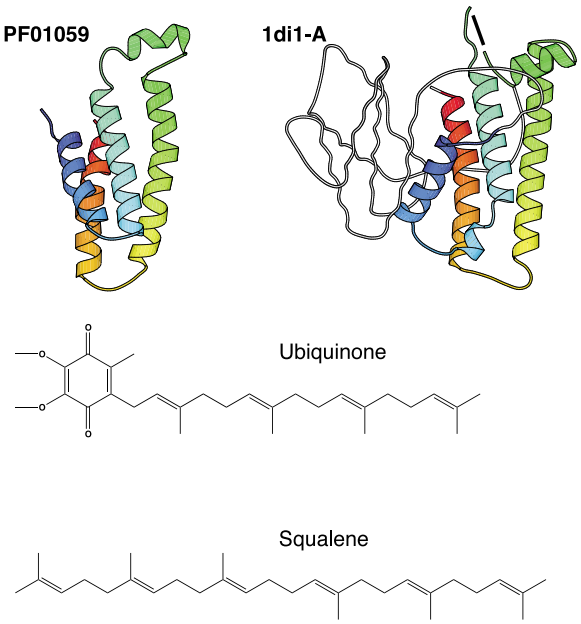

Figure 7. PF01059. The first ranked fold prediction (best *Z*-score) for PF01059 is shown on the left; the closest match in the PDB (1di1 chain A) is shown to the right. Proteins in the sequence family are known experimentally to bind ubiquinone while 1di1-A binds squalene.

The length of the PF01104 domain is considerably shorter than that of 1slt (80 *versus* 133 residues), and only part of the beta sandwich is included in the structural alignment. Additionally, there are high *Z*-score matches to other families, so this connection must be viewed with some caution.

Conclusions

The models generated in this paper provide low-resolution structural information for some of the largest known protein families. The two most pressing areas for current work are to improve the quality and reliability of the models and to develop better methods for extracting useful functional information from the models. The latter problem is obviously shared with all large-scale structural genomics projects, but is somewhat complicated in the case of structure predictions by the fact that the models may be inaccurate or even entirely incorrect. It is likely that use of weak sequence–sequence similarities in conjunction with the structure–structure similarities examined here will help to reliably identify distant evolutionary relationships. Active site recognition methods, such as those of Wallace & Fetrow,[21,22,24] may also be useful in interpreting structure predictions. We anticipate considerable complimentarity between the Rosetta-based approach used here and fold recognition methods[1,2] as they utilize sequence information in quite different ways. A significant advantage to the method described here over fold recognition methods is that *de novo* prediction produces models even when no recognizably similar folds exist in the PDB. Low-resolution structural models can be used to interpret experimental results, such as identifying clusters of functionally important residues and even multiple alternative models can be useful for developing alternate hypotheses that can be tested experimentally

It is difficult for us to judge the exact value of the Pfam models generated in this study. On one hand, they provide the only three-dimensional structural information available for a very large number of proteins, on the other, they contain inaccuracies and are often entirely incorrect, and the confidence functions only partially mitigate these failings. Ultimately the value of the models will be measured by the extent to which they help frame hypotheses about function that are tested experimentally, and we are eager to assist with this effort wherever possible. The entire database of models, along with confidence values and links to experimental structures where available can be accessed for academic use at the Pfam website.

Materials and Methods

Models were generated for three representative and diverse sequences for each alignment (Pfam and test set) as described.[16] Sequences shorter than 60% of the query sequence length were discarded. Sequences with greater than 60% or less than 20% sequence identity to the query were also discarded. The sequence of each protein in the test set was extracted directly from the corresponding PDB file and a multiple sequence alignment (MSA) was generated using PSI-BLAST[37] (three iterations, *E*-value cutoff of 0.01). For Pfam, alignments were taken directly from the Pfam database.

Model generation

The protocol for generating predictions was nearly identical to the protocol used during CASP4.[13] Large numbers of structures are generated and then filtered to remove overly local topologies and improbable strand arrangements. The remaining structures are then clustered (using pairwise C^{α} RMSD values as a distance metric) to produce a small set of models ranked by the size of the cluster they represent. Top ranked models are then compared to the PDB for matches to proteins of known structure.[32] Several subtle differences between the current protocol and the CASP4 protocol are, however, key to the success of the method in the current context.

During CASP4, predictions were manually selected from the top 20 cluster centers produced for each target. This manual reordering of the results was done in order to remove commonly seen systematic errors from the results and to rescue good predictions that were produced less frequently than incorrect models. This manual step was abandoned in this study primarily due to the difficulty of applying human judgment on a genomic scale. Additionally, we were unsure that manual intervention improved our CASP4 predictions: the majority of our correct predictions were top ranked prior to manual intervention and most of the failures were not recoverable by any means. We do not believe

the omission of this step significantly detracts from the performance of the method within the sequence size range of 35–150 residues.

The other important change involves the method used to detect structural similarities to our models. During CASP4 and previous studies we used DALI to detect structural similarities,[29,30] but have since found Mammoth[28] to be superior for our purposes. We found Mammoth *Z*-scores to be more highly correlated with model quality and thus more useful in the context of *de novo* structure prediction, where model quality is not a given. Mammoth is less likely to find non-contiguous alignments than DALI, and consequently, DALI finds many false positives that Mammoth avoids. While DALIs ability to find such disparate structural alignments is quite useful for matching experimental structures against the PDB, it is detrimental when non-protein-like conformations, such as those found in *de novo* structure predictions, are included in the analysis.

For each of the three representative sequences for each family, fragment libraries for each three and nine-residue segment of the chain are extracted from the protein structure database using a sequence profile–profile comparison method as described.[18] At no point is knowledge of the native structure used to select fragments or fix segments of the structure in the test set. The conformational space defined by these fragments is then searched using a Monte Carlo procedure with an energy function that favors compact structures with paired beta strands and buried hydrophobic residues. A total of 2000 independent simulations are carried out for each query representative sequence less than 110 residues and 4000 simulations for each sequence between 110 and 150 residues in length. The resulting structures are filtered and then clustered as described below and previously.[13,33] Prior to clustering, many of the structures produced by Rosetta are incorrect; for this reason, we refer to raw conformations generated by Rosetta as decoy conformations or decoys.

Decoy population filtering

Prior to clustering two filters were applied to remove incorrect conformations: a contact order-based filter and a strand arrangement filter; both filters were also used during the CASP4 experiment.[13,48] Estimation of the allowable contact order range for different length and secondary structure classes was carried out using a non-redundant set of proteins from 50 to 160 residues in length†.[49] Decoys having absolute contact orders in the lowest fifth percentile were discarded prior to clustering to rid populations of overly local conformations. Decoys with unpaired beta strands and other non-protein like strand arrangements were also explicitly removed from the populations prior to clustering.[48]

Clustering procedure

For each prediction, the combined decoy sets for the three homologous sequences were clustered based on C^{α} RMSD over the sets of residues common to all three.[16,50] The decoy with the 100 closest neighbors was located, and the distance to the 100th closest neighbor (or 3 Å, whichever was greater) was used as a cluster threshold. In each iteration, the decoy with the most neighbors within the threshold distance is identified as the top cluster center. All members of this cluster are then removed from the population, and the cycle repeated. The top five cluster centers are the top ranked models.

Structure matching

The Mammoth[28] program was used to search a non-redundant set (less than 50% sequence identity) of 3390 protein chains‡[49] for structures similar to the top five ranked models. Mammoth estimates the maximal, sequence-independent, structural superposition and reports a MaxSub *Z*-score representing the likelihood of finding a similar length match between similar sized proteins by chance.[49]

Definitions of model "correctness"

Three different definitions of correctness were used. First (Definition I), an individual model was considered correct if the strongest structure–structure match in the PDB was to a protein in the same SCOP superfamily as the correct structure (when one or both members of a structure–structure pair were not in the current SCOP database, the fold linkage was considered correct if the Mammoth *Z*-score between the two structures was >5.0).[34,51] Second (Definition II), the group of predictions made for a protein or family were considered successful if any one of the top five models matched a structure in the correct SCOP superfamily. Third (Definition III), the group of predictions made for a protein were considered successful if one of the top five models was aligned by MaxSub[36] to the correct structure for 50 residues or more with an RMSD value under 6 Å;for proteins smaller than 50 residues predictions were considered a success if lower than 3.0 Å in RMSD to the correct structure.

Confidence estimation

We developed separate confidence functions for assessing the likelihood that a given prediction is correct according to each of the three definitions of success described above.

Definition I. For the probability (p) that an individual model matches a PDB structure in the correct SCOP superfamily, we combined[52] the Mammoth *Z*-score of the match (Z), the degree of simulation convergence (C), the length of the protein (L), and the ratio of the lengths of the protein and its PDB match (L_H/L_Q) as follows (see Figure 3(a)):

$$\log\left(\frac{p}{1-p}\right) = 0.416(Z) + 0.00982(L) - 0.326(C) - 1.01(L_H/L_Q) - 2.17,$$

where C is the average RMSD value amongst the five top ranked cluster centers.[13]

Definition II. We fit the probability that one of the top five models has a strongest match in the PDB to a protein in the correct SCOP

† The non-redundant set of proteins structures is available from Roland Dunbrack at http://www.fccc.edu/research/labs/dunbrack/cullpdb.html

superfamily to a function of the best Mammoth *Z*-score (Z), the simulation convergence (C), and the protein length (L) as follows (see Figure 3(b)):

$$\log\left(\frac{p}{1-p}\right) = 0.527(Z) + 0.012(L) - 0.239(C) - 4.97$$

Definition III. For the probability that one of the top five models matched the correct structure, according to the MaxSub success criteria, we fit our results to a function of simulation convergence and length (see Figure 3(c)):

$$\log\left(\frac{p}{1-p}\right) = 0.0619(L) - 0.661(C) - 1.39$$

The discrimination of good and bad predictions provided by these logistic functions is shown in Figure 3(a)–(c). Because the small size of the test set precludes robust cross-validation, confidence estimates may be over-fit near the extremes of the confidence distribution. We have therefore truncated all confidences to a range of 0.15–0.85.

Acknowledgments

R.B. acknowledges the support of a Howard Hughes Medical Institute (HHMI) predoctoral fellowship. D.C. is a fellow of the Program in Mathematics and Molecular Biology at the Florida State University, with funding from the Burroughs Wellcome Fund Interfaces Program. This work was supported by HHMI. We thank Angel Ortiz, Mount Sinai School of Medicine department of Physiology and Biophysics, for the use of Mammoth. We thank Keith E. Laidig, Formix™, for effective and innovative administration of the computer resources necessary to complete this study. We thank Mhairi Marshal and Alex Bateman, Wellcome Trust Sanger Institute (Wellcome Trust Genome Campus, Hinxton, Cambridge, UK), for incorporating the data generated in this study into the Pfam website.

References

1. Jones, D. T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**, 797–815.
2. Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299**, 499–520.
3. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D. & Sonnhammer, E. L. (1999). Pfam 3.1:1313 multiple alignments and profile HMMs match the majority of proteins. *Nucl. Acids Res.* **27**, 260–262.
4. Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A. & Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucl. Acids Res.* **26**, 320–322.
5. Sonnhammer, E. L., Eddy, S. R. & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins: Struct. Funct. Genet.* **28**, 405–420.
6. Pieper, U., Eswar, N., Stuart, A. C., Ilyin, V. A. & Sali, A. (2002). MODBASE, a database of annotated comparative protein structure models. *Nucl. Acids Res.* **30**, 255–259.
7. Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A. *et al.* (1998). Protein folds and functions. *Structure*, **6**, 875–884.
8. Russell, R. B. & Ponting, C. P. (1998). Protein fold irregularities that hinder sequence analysis. *Curr. Opin. Struct. Biol.* **8**, 364–371.
9. Todd, A. E., Orengo, C. A. & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143.
10. Orengo, C. A., Todd, A. E. & Thornton, J. M. (1999). From protein structure to function. *Curr. Opin. Struct. Biol.* **9**, 374–382.
11. Moult, J., Hubbard, T., Fidelis, K. & Pedersen, J. T. (1999). Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins: Struct. Funct. Genet.* **37**, 2–6.
12. Moult, J., Fidelis, K., Zemla, A. & Hubbard, T. (2001). Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins: Struct. Funct. Genet.* **45**, 2–7.
13. Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C. E. M. & Baker, D. (2001). Rosetta in CASP4: progress in *ab initio* protein structure prediction. *Proteins: Struct. Funct. Genet.* **45**, 119–126.
14. Lesk, L. M., Conte, L. L. & Hubbard, T. J. P. (2001). Assessment of novel fold targets in CASP4. *Proteins: Struct. Funct. Genet.* **45**, S5.98–S5.118.
15. Orengo, C. A., Bray, J. E., Hubbard, T., LoConte, L. & Sillitoe, I. (1999). Analysis and assessment of *ab initio* three-dimensional prediction, secondary structure, and contacts prediction [In Process Citation]. *Proteins: Struct. Funct. Genet.* **37**, 149–170.
16. Bonneau, R., Strauss, C. & Baker, D. (2001). Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. *Proteins: Struct. Funct. Genet.* **43**, 1–11.
17. Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C. & Baker, D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins: Struct. Funct. Genet.* **34**, 82–95.
18. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225.
19. Simons, K. T., Bonneau, R., Ruczinski, I. & Baker, D. (1999). *Ab initio* protein structure prediction of CASP III targets using ROSETTA. *Proteins: Struct. Funct. Genet.* **37**, 171–176.
20. Bonneau, R., Tsai, J., Ruczinski, I. & Baker, D. (2001). Functional inferences from blind *ab initio* protein structure predictions. *J. Struct. Biol.* **134**, 186–190.
21. Wallace, A. C., Borkakoti, N. & Thornton, J. M. (1997). TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* **6**, 2308–2323.
22. Wallace, A. C., Laskowski, R. A. & Thornton, J. M. (1996). Derivation of 3D coordinate templates for searching structural databases: application to Ser-

His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.* **5**, 1001–1013.
23. Moodie, S. L., Mitchell, J. B. & Thornton, J. M. (1996). Protein recognition of adenylate: an example of a fuzzy recognition template. *J. Mol. Biol.* **263**, 486–500.
24. Fetrow, J. S., Godzik, A. & Skolnick, J. (1998). Functional analysis of the *Escherichia coli* genome using the sequence- to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* **282**, 703–711.
25. Jonassen, I., Eidhammer, I., Grindhaug, S. H. & Taylor, W. R. (2000). Searching the Protein Structure Databank with weak sequence patterns and structural constraints. *J. Mol. Biol.* **304**, 599–619.
26. Kasuya, A. & Thornton, J. M. (1999). Three-dimensional structure analysis of PROSITE patterns. *J. Mol. Biol.* **286**, 1673–1691.
27. Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147–164.
28. Ortiz, A. R., Strauss, C. E. M. & Olmea, O. (2002). MAMMOTH: matching molecular models obtained from theory. An automated method for protein model evaluation. *Protein Sci.* In the press..
29. Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138.
30. Holm, L. & Sander, C. (1995). Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.* **20**, 478–480.
31. Shindyalov, I. N. & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**, 739–747.
32. Westbrook, J., Feng, Z., Jain, S., Bhat, T. N., Thanki, N., Ravichandran, V. *et al.* (2002). The Protein Data Bank: unifying the archive. *Nucl. Acids Res.* **30**, 245–248.
33. Simons, K. T., Strauss, C. & Baker, D. (2001). Prospects for *ab initio* protein structural genomics. *J. Mol. Biol.* **306**, 1191–1199.
34. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
35. 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.
36. Siew, N., Elofsson, A., Ryclewski, L. & Fischer, D. (2000). MaxSub: an automated measure for the assessment of protein prediction quality. *Bioinformatics*, **16**, 776–789.
37. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., 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.
38. Bycroft, M., Hubbard, T. J., Proctor, M., Freund, S. M. & Murzin, A. G. (1997). The solution structure of the S1 RNA binding domain: a member of an ancient nucleic acid-binding fold. *Cell*, **88**, 235–242.
39. Liepinsh, E., Andersson, M., Ruysschaert, J. M. & Otting, G. (1997). Saposin fold revealed by the NMR structure of NK-lysin. *Nature Struct. Biol.* **4**, 793–795.
40. Gonzalez, C., Langdon, G. M., Bruix, M., Galvez, A., Valdivia, E., Maqueda, M. & Rico, M. (2000). Bacteriocin AS-48, a microbial cyclic polypeptide structurally and functionally related to mammalian NK-lysin. *Proc. Natl Acad. Sci. USA*, **97**, 11221–11226. In Process Citation..
41. Gourley, D. G., Shrive, A. K., Polikarpov, I., Krell, T., Coggins, J. R., Hawkins, A. R. *et al.* (1999). The two types of 3-dehydroquinase have distinct structures but catalyze the same overall reaction. *Nature Struct. Biol.* **6**, 521–525.
42. Ab, E., Schuurman-Wolters, G., Reizer, J., Saier, M. H., Dijkstra, K., Scheek, R. M. & Robillard, G. T. (1997). The NMR side-chain assignments and solution structure of enzyme IIB cellobiose of the phosphoenolpyruvate-dependent phosphotransferase system of *Escherichia coli*. *Protein Sci.* **6**, 304–314.
43. Rowland, P., Norager, S., Jensen, K. F. & Larsen, S. (2000). Structure of dihydroorotate dehydrogenase B: electron transfer between two flavin groups bridged by an iron–sulphur cluster. *Structure*, **8**, 1227–1238.
44. Lai, L. H., Yokota, H., Hung, L. W., Kim, R. & Kim, S. H. (2000). Crystal structure of archaeal Rnase Hii: a homologue of human major Rnase H. *Structure (Lond.)*, **8**, 897–904.
45. Obmolova, G., Ban, C., Hsieh, P. & Yang, W. (2000). Crystal structures of mismatch repair protein muts and its complex with a substrate DNA. *Nature*, **407**, 703–710.
46. Caruthers, J. M., Kang, I., Cane, D. E., Christianson, D. W. & Rynkiewicz, M. J. (2000). Crystal structure determination of aristolochene synthase from the blue cheese mold, *Penicillium roqueforti*. *J. Biol. Chem.* **275**, 25533–25539.
47. Liao, D. I., Kapadia, G., Ahmed, H., Vasta, G. R. & Herzberg, O. (1994). Structure of S-lectin, a developmentally regulated vertebrate beta-galactoside-binding protein. *Proc. Natl Acad. Sci. USA*, **91**, 1428–1432.
48. Ruczinski, I., Kooperberg, C., Bonneau, R. & Baker, D. (2002). Distributions of beta sheets in proteins with application to structure prediction. *Proteins: Struct. Funct. Genet.* **48**, 85–97.
49. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409–417.
50. Shortle, D., Simons, K. T. & Baker, D. (1998). Clustering of low-energy conformations near the native structures of small proteins. *Proc. Natl Acad. Sci. USA*, **95**, 11158–11162.
51. Murzin, A. G. (1999). Structure classification-based assessment of CASP3 predictions for the fold recognition targets. *Proteins: Struct. Funct. Genet.* **37**, 88–103.
52. Ripley, W. N. V. B. D. (1999). *Modern Applied Statistics With S-Plus. Statistics and Computing* (Chambers, J., ed.), 3rd edit., Springer, New York.

Edited by J. Thornton

(Received 22 March 2002; received in revised form 3 July 2002; accepted 8 July 2002)