PROTEINS: Structure, Function, and Genetics 23:356–369 (1995)

Threading a Database of Protein Cores

Thomas Madej, Jean-François Gibrat, and Stephen H. Bryant
Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894

ABSTRACT **We present an analysis of 10 blind predictions prepared for a recent conference, "Critical Assessment of Techniques for Protein Structure Prediction."[1] The sequences of these proteins are not detectably similar to those of any protein in the structure database then available, but we attempted, by a threading method, to recognize similarity to known domain folds. Four of the 10 proteins, as we subsequently learned, do indeed show significant similarity to then-known structures. For 2 of these proteins the predictions were accurate, in the sense that a similar structure was at or near the top of the list of threading scores, and the threading alignment agreed well with the corresponding structural alignment. For the best predicted model mean alignment error relative to the optimal structural alignment was 2.7 residues, arising entirely from small "register shifts" of strands or helices. In the analysis we attempt to identify factors responsible for these successes and failures. Since our threading method does not use gap penalties, we may readily distinguish between errors arising from our prior definition of the "cores" of known structures and errors arising from inherent limitations in the threading potential. It would appear from the results that successful substructure recognition depends most critically on accurate definition of the "fold" of a database protein. This definition must correctly delineate substructures that are, and are not, likely to be conserved during protein evolution.**

Key words: structure prediction, fold recognition, protein threading

INTRODUCTION

The challenge to threading methods is to recognize from sequence that a protein adopts a known fold, even though sequence similarity as measured traditionally is low or nonexistent. By "fold" one means a fragmentary substructure observed previously in another protein, with sufficient similarity to the true conformation that one might rightly infer some of a protein's biological properties. To recognize a fold, threading methods employ empirical free energy functions, searching, in the computer, for sequence–structure alignments that appear to represent a stable model structure. Current potentials measure rather general chemical properties of amino acid side chains such as their hydrophobicity or tendency to occur in close proximity.[2–4] As a consequence they are broadly sensitive to sequence–structure compatibility, when conformations are similar but not identical, but they are also relatively nonspecific. It is easy to show, for example, that the fraction of all possible sequence assignments that will appear to have lower energy than the native sequence is much higher than, say, the fraction that will show 35% residue identity.[5–7] In threading one considers very many alternative sequence–structure alignments, so many, perhaps, that the odds of encountering a low-energy model within any sequence may be high. From this perspective one is tempted to state the challenge to threading methods more precisely: to identify the signal of a truly similar fold amid the noise of favorable threading alignments one may expect to find by chance.

To obtain the greatest specificity they can from current potentials the groups working on threading methods have attempted to refine their definitions of the "folds" of known structures, hoping to exclude implausible alignments, and thus improve the signal-to-noise ratio of a threading search. Much is known about the nature of structural similarities that are likely to persist during protein evolution, namely that a core consisting of large secondary structure elements will be conserved, with considerable variation in the lengths and conformation of intervening loop regions.[8,9] To consider only alignments that are consistent with conservation of a core, all threading methods consider only sequence–structure alignments which encompass most or all of the secondary structure elements of a protein or protein domain. These alignments are furthermore constrained to have relatively few "gaps," or breaks in chain continuity in the alignment of residues from the sequence with sites in the structure. The precise definition of the "fold" of a known structure is spec-

Received April 3, 1995; revision accepted June 9, 1995.
Address reprint requests to Stephen H. Bryant, Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894.

ified, however, by the details of the alignment models employed.

Some threading methods adopt the alignment model used traditionally in sequence comparison. Any colinear alignment of residues from the sequence with sites in the structure is permitted in principle, but gap penalties effectively exclude alignments which are either too short, contain too many breaks in chain continuity, or perhaps require chain breaks at inappropriate locations within the structure.[10–18] In as much as they define the number and identity of residue sites which must be included in any high-scoring threading model, it is these gap penalties which effectively define the "fold" of a known structure. Other threading methods, including our own, adopt an alignment model based on explicitly defined "core elements."[19,20] Each corresponds to a β-strand or α-helix assumed to be present in any shared "core" substructure, and their residue sites must be aligned with a chain-continuous segment from the sequence under investigation. The remaining "loop" residues are assumed to occupy nonequivalent sites, and are neither aligned nor included in energy calculations. No gap penalties are employed, and it is instead the prior definition of core elements that explicitly constrains recruitment of residue sites from the known structure into the threading model.

One cannot be sure, of course, that any of these alignment models yet offers correct definition of the "folds" of database structures, or indeed that this factor was uniformly critical to the successes of the threading methods tested in the "Asilomar Challenge."[1] We found, however, as did other groups at Asilomar, that some threading alignments were quite inaccurate, and it is interesting to ask why. To better understand our own results we would like to know, in particular, whether alignment accuracy was limited by incorrect definitions of the "cores" of known structures, or instead by imprecision in the threading potential or alignment optimization algorithm. The answer to this question will not only help to tell us "what went right" and "what went wrong," but will also help to determine the direction of further research in fold recognition.

It is from this perspective that we undertake in this paper a critical analysis of the threading results we submitted for 10 Asilomar proteins. We compare these protein's true structures, made available afterward, to structures then in the Protein Data Bank,[21] asking whether any statistically significant substructure similarity indeed exists. We then compare the observed structural similarities to the predefined "cores" we used in the threading search, to determine whether our explicit definition of domain "folds" was consistent with the true structural similarities. With the results of this analysis in hand, we then interpret the successes and failures of the threading predictions, and find some interesting surprises. If the fold is defined incorrectly, it would appear that there is still some chance of identifying a protein with a true substructure similarity. But there is no chance, obviously, of finding a threading alignment that is sufficiently accurate for molecular modeling. If the fold is correctly defined, however, it would appear that even our current threading potential is capable of producing a reasonably accurate model, which might indeed serve as a starting point for more detailed modeling and prediction of functional properties.

METHODS

Protein Core Database

For the Asilomar contest we prepared a preliminary database of protein "cores," based on a nonhomologous subset of the April 1994 release of the Protein Data Bank,[21] the latest then available. This subset contained 435 chains exhibiting less than 35% pairwise residue identity, with representatives from homology groups selected automatically according to structure quality and completeness. "Core" motifs within these structures were identified using geometric criteria described below. The "fold" of each domain, to summarize, was defined as the set of alternative substructures encompassing core secondary structure elements, with explicit limitations as to the number of sequence residues assignable to nonequivalent "loop" sites. This core database was recorded as a set of annotations on each known structure and submitted to the contest organizers as part of the requested information on methods. It is available from us in machine-readable form.

Elements of secondary structure in each chain were identified automatically by an algorithm similar to that of Kundrot and Richards.[22] Helical and beta "ladder" substructures similar to 6- or 8-residue paradigms were identified on the basis of inter-α-carbon distances and merged to form complete helices and β-strands. "Bent" strands and helices were detected from the residuals of a fitted axis and split to form 2 or more approximately linear substructures. Secondary structure classifications by this algorithm agree well with author assignments from the Protein Data Bank[21] and appear insensitive to resolution and/or local distortions in crystallographic or NMR-derived structures. Domain secondary structure elements identified in this way were also used in the structure–structure comparison algorithm described below.

For each chain in the nonhomologous set compact domains were identified using an algorithm similar to that of Holm and Sander.[23] Domain boundaries were placed at either one or two of the midpoints of loops joining secondary structure elements, where breaks would result in the highest ratio of intra- to interdomain contacts, and wherever that ratio was 2 or greater. Both resulting domains were also re-

quired to have no isolated secondary structure elements, i.e., those forming no contacts,[19] and the smaller of the resulting domains was required to contain 25 or more residues within secondary structure elements. The domain boundaries identified are very similar in test cases to those reported.[23] Division of the nonhomologous subset into domains generated a total of 786 alternative domain structures, including the 435 derived from complete chains. Domains are identified below by nomenclature following the pattern *1RAA-B-2,* based on the PDB identification code *1RAA,* PDB chain code *B,* and an arbitrary domain number, respectively.

Core elements were defined as a subset of the residue sites within strands or helices that is likely to be conserved in a distantly related protein. As the most central residue sites of strands or helices are often conserved, we "trimmed" residue sites from the N- and C-terminal ends of each secondary structure element until a total of 60% of its contacts with other elements remained. Since small secondary structure elements within loops are often not conserved, we omitted from our definition of core elements any that contained less than 2 strand or 4 helical residues, or formed less than 60% of the average number of contacts per core element. These core elements form a nucleus of residue sites necessarily present in all threading models, but we note that the alignment model described below allows recruitment of additional sites by addition in chain-continuous fashion to the amino- or carboxy-terminal ends of these core elements.

Since they were intended to represent autonomously folding domains our core definitions also placed constraints on the numbers of residues that could be assigned to nonequivalent sites, representing loops in the predicted structure. The number of loop sites allowed was specified as a function of the spanning distance between the endpoints of adjacent secondary structure elements and of the observed loop length. Maxima were either the 75th percentile value in a table of observed loop lengths in the nonhomologous set, grouped by spanning distance in 3.4 Å bins, or 1.25 times the observed loop length, whichever was greater. Maxima were further increased by 0.75 times the number of residue sites "trimmed" from secondary structure elements. These constraints were intended to allow loops in threading models to be roughly twice as long as an "average" loop, but to exclude inserted domains. Loop length minima were adjusted automatically so as to exclude nonphysical models, where the number of loop residues is less than the number needed to span the distance between core element endpoints.

Threading

Sequences of the Asilomar contest proteins were threaded through core folding motifs using the alignment model and contact potential described previously.[19] In this procedure all core elements must be aligned with subsequence fragments, and we thus search for domains within each sequence that adopt one of the predefined "folds." Sequence length determines the number of alternative alignments possible for each core motif, and may exclude domains requiring more residues, but no gap penalties are employed and there is no a priori preference for core motifs of a given size.[5] For the 50 best-candidate core motifs we submitted raw threading energies $\Delta G_{R|M}$ and composition-corrected scores $Z_{R|M}$ as previously defined.[19] To facilitate comparison with other methods $Z_{R|M}$ was also expressed in normalized form relative to scores for the best 200 core motifs, Z_r, and we similarly submitted a score based on the correlation coefficients of pairwise contact energies μ_{rsd}[19] in the native versus model structure, normalized in an equivalent fashion, Z_x. Threading scores were submitted to the contest in order of decreasing "combined score," Z_c, which was defined as the sum Z_r and Z_x, again normalized across the best 200 "hits." We were unable to compute in time for the contest alignment-number corrected scores and chance-occurrence probabilities, but we show below values for the submitted model structures, calculated as previously described.[5]

Optimal threading alignments of sequence and core motif were identified not by enumeration,[19] but via a Gibbs Sampling algorithm.[7] This procedure results in an ensemble of alternative alignments ranked by threading energy $\Delta G_{R|M}$, and we submitted for the Asilomar contest the 10 best alignments for each sequence and core motif pair. In the sampling procedure subsequence fragments are initially aligned at random with core elements. Alternative alignments of an individual core element are then enumerated and $\Delta G_{R|M}$ calculated in the "field" defined by the current alignment of the others. A new alignment for that core element is then drawn at random based on Boltzmann probabilities derived from $\Delta G_{R|M}$ with an assumed "temperature" of 10 kT units. The procedure is then iterated for all core elements, for a fixed number of cycles. Recruitment of residue sites to the ends of core elements is performed in a similar stochastic fashion, based on Boltzmann probabilities in the "field" defined by the current alignment and endpoints of other core elements. The nucleus defined by the core elements tends to grow, in effect, when addition of new residue sites from the known domain structure results in favorable interactions as judged by the contact potential.[19] The Gibbs Sampling algorithm has been used in some previous work[5,24] and validated in control experiments to be reported elsewhere.[7]

Structure Comparison

Structures of contest proteins were compared to domain structures in our database using an algorithm that searches for similar spatial orientation

TABLE I. Proteins for Which Threading Results Were Submitted

Code	Full protein name	Species of origin
114	Ribosomal protein 114	*Bacillus stearothermopholis*
mystery	Fictional structure	*Hyper sapiens*
pbdg	6-Phospho-β-D-galactosidase	*Lactococcus lactis*
ppdk-1	Domain 1 of pyruvate phosphate dikinase	*Bacteroides symbiosus*
ppdk-2	Domain 2 of pyruvate phosphate dikinase	*Bacteroides symbiosus*
ppdk-3	Domain 3 of pyruvate phosphate dikinase	*Bacteroides symbiosus*
ppdk-4	Domain 3 of pyruvate phosphate dikinase	*Bacteroides symbiosus*
prosub	Propeptide of subtilisin BPN'	*Bacillus subtilis*
rtp	Replication terminator protein	*Bacillus subtilis*
staufen	Staufen 3 protein	*Drosophila melanogaster*

and connectivity of secondary structure elements.[25] Grossly similar substructures are identified using an exhaustive search procedure aided by a fast algorithm from graph theory.[26] The most surprising similarity is then identified by comparing root mean square (rms) superposition residuals for pairs of secondary structure elements to the distribution observed when such pairs are selected at random from the known-structure database. This calculation gives the odds, p_c, that a substructure alignment with this superposition residual would be observed by chance in a random draw of two substructures of that size. Statistical significance p_d is calculated as the product of p_c and the number of alternative substructure alignments that are possible in comparison of the two proteins, given the numbers of strands and helices they contain. A value of 0.05 indicates, for example, that a similarity as surprising as that observed should be expected to occur by chance in only 5% of pairwise protein comparisons. For a database search involving many pairwise comparisons a significance threshold reflecting the size of that database should be used.[27] Optimal C_α coordinate structural alignments are identified by a Monte Carlo sampling procedure that refines the precise alignment and boundaries of conserved secondary structure elements, making reference to empirical distributions of rms residual versus the number of aligned residue sites.[25] We report below the number of equivalent C_α sites in the common substructures identified, N, the fraction of the contest protein sites this represents, F, and the rms residual in Ångstroms, R.

To quantify the agreement between an observed common substructure and our prior definition of the "fold" of a database domain we compute two measures of error, "overprediction" or type 1 error, and "underprediction" or type 2 error. Type 1 error derives from residue sites in core elements which are not present in the true common substructure. An extra helix in the core definition leads to type 1 error, for example, which we report below as E_1, the ratio of extra sites to the number of sites in the common substructure, expressed as a percentage. Type 2 error arises from sites in the common substructure which could not be aligned correctly by threading, given the constraints imposed by our core definition. Helices or β-strands missing from our core definition give a simple measure of type 2 error, E'_2, the fraction of sites in the common substructure for which no corresponding core element was defined. We also report a more stringent measure of type 2 error which takes into account the interdependence of the alignment of different core elements. It is possible, for example, that a helix could be present in both the true common substructure and our core definition and yet could not possibly be aligned correctly by threading, due to loop-length constraints and/or the presence of type 1 error in the core definition. We report below this stringent measure, E_2, as the fraction of sites in the true common substructure that could, in the most favorable case, be placed in their correct structural alignment.

RESULTS

True Structural Similarities

To identify known structures that are similar to the Asilomar contest proteins we conducted an automatic structure–structure comparison search. In Table I we list the 10 proteins for which we submitted threading results, the complete set available just before the contest deadline of October 1, 1994. In Table II we list all significant structural similarities between these proteins and the 786 domain structures then available in our database. We also compare in Table II the substructures identified by structural alignment to our prior definition of the "core" of each domain. This comparison does not describe the accuracy of threading models, which we present below, but simply asks whether our representation of the domain "fold" enabled us to accurately reproduce the observed structural similarity.

Five of the 10 proteins show no significant structural similarity to any domain structure at $p_d < 10^{-5}$, a liberal significance threshold for search of a database so large. These proteins are *l14*,* *ppdk-1*,† *ppdk-2*,† *ppdk-3*,† and *staufen3*.‡ The *mystery* protein is similar to many database proteins but was suspected and later confirmed to be a fictional structure, and we omit it from Table II and from further

TABLE II. Structure–Structure Comparison of Contest Proteins and Database Domains

Contest protein	Database domain	R	N	F	$-\log p_c$	$-\log p_d$	E_1	E'_2	E_2
114	1RAA-B-1	2.0	21	17.2	6.8	3.4	—	—	—
pbdg	2TMD-A-3	3.5	178	39.1	29.8	16.8	16.9	10.1	56.7
pbdg	1BTC	2.8	205	45.1	31.4	15.2	9.3	11.7	56.1
pbdg	2MNR-1	2.3	147	32.3	26.1	14.5	9.5	16.3	75.5
pbdg	1PII-1	2.5	136	29.9	23.5	14.3	6.6	39.0	84.6
pbdg	1CHR-A-1	3.0	144	31.6	26.0	13.0	10.4	8.3	77.1
pbdg	1GOX	3.1	149	32.7	27.1	12.6	36.9	34.2	81.2
pbdg	1FCB-A-1	2.8	146	32.1	26.9	11.9	41.1	35.6	78.8
pbdg	1RAL	2.8	140	30.8	23.9	11.8	20.7	29.3	64.3
pbdg	1XIS	3.7	185	40.7	26.1	11.7	9.7	8.6	57.8
pbdg	1YPI-A	3.2	146	32.1	22.6	11.6	8.2	31.5	84.2
pbdg	1PII-2	2.9	131	28.8	24.0	11.5	9.9	21.4	74.8
pbdg	2MNR	2.3	147	32.3	26.1	11.4	46.9	23.1	87.8
pbdg	1CHR-A	3.0	144	31.6	26.0	10.6	47.9	23.6	86.1
pbdg	1FBA-A	3.6	161	35.4	23.9	10.5	29.2	6.2	65.2
pbdg	1FCB-A	2.9	148	32.5	26.9	9.9	56.1	23.0	79.1
pbdg	1TML-1	2.8	119	26.2	19.1	9.8	5.9	18.5	66.4
pbdg	2TMD-A	3.4	179	39.3	29.8	9.7	94.4	10.6	75.4
pbdg	1WSY-A	3.4	139	30.5	22.2	9.7	19.4	30.9	79.1
pbdg	9RUB-A	2.5	138	30.3	25.0	8.9	54.3	17.4	86.2
pbdg	1NAR	3.1	171	37.6	21.0	8.8	12.9	15.8	73.7
pbdg	4ENL-1	3.2	134	29.5	21.4	8.7	26.9	21.6	78.4
pbdg	9RUB-A-2	2.5	112	24.6	23.3	8.4	58.9	18.8	85.7
pbdg	1RLD-A	2.9	129	28.4	23.9	8.1	55.8	0.0	73.6
pbdg	6TAA	3.1	147	32.3	23.2	7.7	57.1	15.6	51.7
pbdg	1TML	2.8	124	27.3	19.1	7.6	13.7	12.1	52.4
pbdg	1PII	3.1	130	28.6	23.5	7.4	63.8	20.0	74.6
pbdg	1PII	2.5	133	29.2	24.0	7.2	80.5	39.8	91.0
pbdg	4ENL	3.2	135	29.7	21.4	5.8	63.0	33.3	85.9
pbdg	1CDG	3.8	151	33.2	24.9	5.5	127.8	13.9	94.7
pbdg	1ADD	3.7	103	22.6	18.8	5.4	69.9	27.2	66.0
pbdg	6TAA-2	2.6	81	17.8	13.9	4.3	—	—	—
ppdk-1	2TBV-A-1	3.3	31	11.3	8.8	4.8	—	—	—
ppdk-2	2PKA-A	4.5	34	19.8	6.4	4.9	—	—	—
ppdk-3	1BNH-1	3.0	40	19.5	9.9	4.7	—	—	—
ppdk-4	2MNR-1	2.7	153	41.5	27.9	18.4	3.3	7.8	100.0
ppdk-4	1PII-1	2.5	132	35.8	24.8	18.0	6.8	40.9	100.0
ppdk-4	1GOX	2.7	157	42.5	29.4	16.5	19.1	16.6	69.4
ppdk-4	1CHR-A-1	2.9	145	39.3	27.5	16.5	3.4	0.0	69.0
ppdk-4	1FCB-A-1	2.8	148	40.1	30.0	16.3	23.0	18.2	78.4
ppdk-4	1PII-2	2.8	139	37.7	27.1	16.1	4.3	13.7	79.9
ppdk-4	2MNR	2.7	153	41.5	27.9	14.8	39.2	13.1	100.0
ppdk-4	1FCB-A	2.8	150	40.7	30.0	14.0	38.0	7.3	67.3
ppdk-4	1CHR-A	3.0	145	39.3	27.5	13.5	40.7	12.4	76.6
ppdk-4	1RAL	3.3	121	32.8	24.3	13.1	16.5	16.5	52.1
ppdk-4	1XIS	4.8	157	42.5	25.3	11.9	26.1	19.1	100.0
ppdk-4	4ENL-1	3.3	117	31.7	22.8	11.5	30.8	19.7	70.1
ppdk-4	1PII	2.5	131	35.5	27.1	11.5	81.7	41.2	100.0
ppdk-4	1YPI-A	2.7	105	28.5	21.9	11.5	9.5	31.4	100.0
ppdk-4	1WSY-A	2.6	146	39.6	23.0	11.4	21.2	36.3	100.0
ppdk-4	1RLD-A	3.1	169	45.8	25.7	11.3	36.7	7.1	100.0
ppdk-4	1BTC	3.0	166	45.0	25.8	11.1	21.1	19.3	39.2
ppdk-4	9RUB-A-2	3.2	142	38.5	24.3	10.5	40.1	14.8	100.0
ppdk-4	1TML-1	2.9	118	32.0	18.4	10.5	0.0	12.7	61.0
ppdk-4	1PII	2.8	139	37.7	24.8	9.3	54.7	12.2	100.0
ppdk-4	9RUB-A	3.2	142	38.5	24.3	9.2	54.2	14.8	100.0
ppdk-4	1FBA-A	3.4	138	37.4	21.5	9.1	36.2	9.4	49.3
ppdk-4	2TMD-A-3	2.6	150	40.7	20.9	9.0	34.7	27.3	40.0
ppdk-4	1NAR	2.7	111	30.1	19.5	8.6	27.0	12.6	68.5
ppdk-4	4ENL	3.3	115	31.2	22.8	8.5	73.9	33.9	82.6
ppdk-4	1TML	3.0	120	32.5	18.4	7.6	8.3	6.7	60.8
ppdk-4	6TAA	4.3	122	33.1	21.3	7.4	73.8	15.6	100.0
ppdk-4	1ADD	3.4	127	34.4	20.6	7.3	55.9	29.1	100.0
ppdk-4	1GCA-2	2.8	54	14.6	14.1	6.2	66.7	25.9	100.0
ppdk-4	6TAA-2	2.4	64	17.3	14.6	5.6	25.0	0.0	51.6
ppdk-4	1CDG	4.9	133	36.0	21.8	4.7	NA	NA	NA
prosub	2BOP-A	2.8	39	54.9	8.6	8.6	0.0	23.1	35.9
prosub	2NCK-R	2.4	47	66.2	9.8	8.6	17.0	0.0	80.9
prosub	1APS	2.5	43	60.6	9.6	8.5	9.3	14.0	41.9
prosub	1RAA-B-2	2.4	37	52.1	8.4	7.7	16.2	16.2	64.9
prosub	1VNA	2.7	30	42.3	7.4	6.5	0.0	0.0	23.3
prosub	1FXD	3.5	34	47.9	7.5	6.2	17.6	17.6	47.1
prosub	1RAA-B	2.4	37	52.1	8.4	6.1	56.8	0.0	100.0
prosub	1TPL-A-1	2.2	36	50.7	9.7	5.9	191.7	0.0	100.0
prosub	1GPT	3.3	30	42.3	6.6	5.7	0.0	0.0	23.3
prosub	2BUS	2.9	30	42.3	6.5	5.6	0.0	0.0	20.0
prosub	1MUP	3.0	33	46.5	8.4	5.2	106.1	0.0	100.0
prosub	1MEE-I	3.1	30	42.3	6.7	5.2	20.0	20.0	56.7
prosub	1POH	3.4	42	59.2	6.8	5.1	11.9	0.0	100.0
prosub	1CBN	3.5	30	42.3	5.7	5.0	0.0	0.0	40.0
prosub	1TPL-A	2.3	36	50.7	9.7	4.6	—	—	—
rtp	1HST-A	1.9	52	44.8	8.1	7.5	0.0	0.0	17.3
rtp	1BIB-3	1.9	36	31.0	6.4	4.5	—	—	—
rtp	3GAP-A-1	3.8	25	21.6	6.0	3.9	—	—	—
staufen	2RSP-A	3.2	34	50	6.9	4.1	—	—	—

consideration. The 4 remaining proteins, *pbdg*,[§] *ppdk-4*,[†] *prosub*,[**] and *rtp*[††] all show significant structural similarities to one or more previously known domain structures. None, however, shows any significant sequence similarity to the corresponding domain as measured by Blast chance-occurrence probabilities.[27]

Proteins *pbdg* and *ppdk-4* are structurally similar to a large number of proteins containing an 8-stranded parallel beta motif commonly known as a "*TIM*-barrel." Triose phosphate isomerase (*TIM*) from yeast, *1YPI-A*,[28] is one of these structures. None of the similarities is very extensive, as they encompass only between 14 and 46% of the residue sites of *pbdg* or *ppdk-4*. Rms residuals range from 2.3 to 4.9 Å. Type 1 error E_1 of the predefined cores range from 0 to 150%, with only 1 of 62 comparisons showing no error. Most *TIM*-barrels contain additional strands and helices in loops or inserted domains extraneous to the barrel itself, and these were usually included in our geometrically defined cores. The type 2 error measure E'_2 was under 20% for more than half of the predefined cores, indicating that they often contained most or all of the strands and helices of the barrel, but the more stringent measure of type 2 error E_2 was under 50% in only 4 out of 64 comparisons. Our definition of the "core" of the database structures deliberately did not allow for deletion or insertion of large secondary structure elements. Because most database cores have such insertions, and because *pbdg* and *ppdk-4* themselves have additional strands and/or helices not equivalent to any in the domain database, we could not in any of the corresponding sequence–structure comparisons have achieved an even approximately correct threading alignment. Our definition of the "folds" of the database domains was too stringent, requiring more extensive structural similarity than observed.

prosub is structurally similar to a number of domains having a structure composed of a 4-stranded mixed parallel/antiparallel β-sheet with 2 α-helices, including ferredoxin, *1FXD*.[29] Structural similarities vary in extent from 42 to 66% of the residue sites in prosub, with RMS values ranging from 2.2 to 3.5 Å. Type 1 error E_1 relative to our predefined "cores" ranges from 0 to 344%, but is less than 20% in 11 of 15 comparisons, indicating that many cores contain few additional strands or helices beyond the observed common substructure. Type 2 error E'_2 is 20% or less in 14 of 15 comparisons, indicating that most predefined cores omit few of the common secondary structure elements. E_2 is never less than 36% for comparisons where the common substructure encompasses more than 50% of the residues in *prosub*, however, indicating that the combination of type 1 error and loop-length constraints allows only partially correct threading alignments. Together these results indicate that our representation of the domain "folds" was capable of reproducing an approximately correct sequence-structure alignment for *prosub*, but not a fully correct one.

rtp is structurally very similar only to the histone H5 globular domain from chickens, *1HST-A*,[30] a structure with 3 helices with a single β-hairpin ladder. Fifty-two residues of *rtp* superimpose to 1.9 Å rms residual, with the remaining nonequivalent sites falling largely within an additional carboxy-terminal helix and an amino-terminal extension, which together form the *rtp* dimerization domain. This represents 45% of the total number of residues in *rtp*, but a large fraction, 67%, of it's chain-continuous globular domain. Interestingly, *rtp* shows less extensive similarity to two other DNA-binding domains, *1BIB-3*,[31] the biotin repressor, and *3GAP-A-1*,[32] the catabolite gene activator protein. Errors E_1 and E'_2 relative to the *1HST-A* "core" are both zero, indicating that the predefined core elements corresponded exactly to the strands and helices conserved in *rtp*. E_2 is 17%, as the loop-length constraint based on the β-hairpin present in *1HST-A* was too short to accommodate the extension of the ladder observed in *rtp*, and forces a register-shift error in the carboxy-terminal beta strand, or that preceding it. On the whole, however, our representation of the "fold" of *1HST-A* was capable of reproducing a nearly correct sequence–structure alignment for *rtp*.

Accuracy of Threading Models

The similar substructures identified by three-dimensional coordinate comparison are a standard by which our threading results may be judged. For those contest proteins which are similar to a database domain we examine the distribution of threading scores to see if these true positives were sufficiently near the top of the list to focus attention on the correct "fold." For these models we also examine the accuracy of the sequence–structure alignment produced by threading, comparing it to the corresponding structural alignment. Alignment accuracy is a direct measure of agreement between the true common substructure and that predicted by threading, and is, perhaps, the most critical measure of the usefulness of a predicted model structure.

Proteins *pbdg* and *ppdk-4* had no possibility of a globally correct threading alignment, as discussed above, but *TIM*-barrel domains are nonetheless concentrated in the right tail of the score distributions shown in Figure 1. *2TMD-A*[33] and *1BTC*,[34] which ranked at positions 2 and 8 in the submitted score lists for *pbdg* and *ppdk-4*, respectively, are *TIM*-bar-

*C. Davies and S. White, Asilomar data submission.
†O. Herzberg, Asilomar data submission.
‡M. Bycroft, Asilomar data submission.
§C. Weismann, Asilomar data submission.
**T. Gallagher, P. Bryan, and G. Gilliland, Asilomar data submission.
††D. Bussiere and S. White, Asilomar data submission.

rel domains. The top-scoring domains for *pbdg* and *ppdk-4* are both large proteins with extensive β/α structure, however, as are most others with high scores. An analysis of threading alignments suggests that the method has primarily detected in these cases the presence of sequence patterns compatible with β–α–β local structure, in that we find that strands are often aligned with strands, and helices with helices, but in only some cases those which are superimposable in a global structural alignment. In the threading alignment of *pbdg* against the domain *2TMD-A*-3, for example, 4 of 14 superimposable core elements have alignment errors less than 4 residues, but in the submitted alignment with *2TMD-A* no core elements are aligned correctly, and the same is true for the submitted alignment of *ppdk-4* with *1BTC*. Mean errors for these alignments are 70.3 and 50.1 residues, respectively. Threading scores found for neither *pbdg* nor *ppdk-4* are statistically significant as judged by chance occurrence probabilities.[5]

Some core motifs in our database were capable of reproducing at least approximately the correct conformation for *prosub*. One of these, *1RAA-B-2*,[35] ranked second in the submitted threading scores, as shown in Figure 2. This domain structure also had the lowest chance occurrence probability, with a *P*-value of approximately 0.001. Other structurally similar domains scored less favorably than *1RAA-B-2*, most probably because the threading alignment with *1RAA-B-2* allowed recruitment of a larger number of additional loop sites into the model, including much of the structurally equivalent helix following strand 4, which was omitted from the core definition due to it's short length in *1RAA-B-2*. As may be seen in the "energy scaffold" in Figure 3, the threading model includes a large number of sites in the loop following strand 4, which are somewhat similar in length and conformation in *prosub* and *1RAA-B-2*, though not enough so to be included in the structural alignment.

The threading alignment of *prosub* and *1RAA-B-2* is shown in Figure 4. It does not agree perfectly with the structural alignment, but it is nonetheless reasonably accurate, certainly more so than would be predicted solely by the lengths of the core elements and *prosub* sequence. Strand 1 is aligned correctly, and helix 2 differs from its structural alignment by a "register shift" of one turn. Some displacement of helix 2 and strands 3 and 4 is forced by the *1RAA-B-2* core definition, which does not quite allow as long a loop as seen between strand 1 and helix 2 in *prosub*, or the observed truncation in *prosub* of strands 3 and 4. Strand 5 is also forced out of register by inclusion in the core definition of an extra "edge" β-strand that is not present in *prosub*. The mean alignment error of 4.0 residues seems perhaps as good as one might expect, given the inaccuracies of the prior core definition, and the fact that the common substructure accounts for only 52% of the residues in prosub.

The core motif based on *1HST-A* was able to quite accurately reproduce the common substructure see in *rtp*. Perhaps as a consequence, the threading model and alignment shown in Figures 3 and 4 were the most accurate of any of the threading results. Mean error for the submitted alignment is 2.7 residues, and no helices are displaced by more than a "register shift" of 1 turn, an accuracy that is completely nontrivial given that the additional carboxy-terminal sequence of *rtp* allowed displacements of 40 residues or more. The strictly correct alignments for each core element also had significant probability in the threading ensemble, as shown in Figure 4, even though the core definition disallowed simultaneous correct alignment of strands 4 and 5, due to an overshort loop-length constraint. This accuracy of alignment is perhaps the best one can expect for a substructure with 1.9 Å rms superposition residual, where true contact patterns differ, and loop conformations are quite different.[7]

Given the good alignment accuracy for *rtp*, it seemed at first surprising that *1HST-A* ranked only 10th in the list of submitted threading scores, and that we were able to identify it as one of three likely models only on the basis of its known function. By the purely statistical criterion of chance occurrence probability *1HST-A* is also not a strong candidate. We note, however, that *1HST-A* is similar to *rtp* only in its globular domain, encompassing 45% of the *rtp* structure. In statistical computations based on random shuffling of the *rtp* sequence, the extra residues "dilute" the signal of a domain alignment that is itself quite extensive, at 67% of *rtp* residues. A similar phenomenon accounts for false positives in the composition-corrected threading scores. The core folding motifs which score better are somewhat larger, mostly helical domains, which place the additional, long carboxy-terminal helix in helical regions, forming extensive contacts with the remainder of the structure. From a purely energetic point of view these may indeed be better models than the structure based on the *1HST-A* monomer, where the 30-plus carboxy-terminal residues are implicitly assumed to be disordered, and make no contacts with the remainder of the structure.

On the whole it seems reasonable to interpret the threading results for *pbdg* and *ppdk-4* as largely incorrect. Some "*TIM*-barrel" proteins did indeed score well, particularly for *pbdg*, but threading alignments were grossly inaccurate. By the same token it seems reasonable to interpret the threading results for *prosub* and *rtp* as largely correct, on the basis of good if not perfect alignments, and scores that were, prospectively, sufficient to focus attention on the correct "folds." We must stress, however, that the fold-recognition models for *prosub* and *rtp* still contain errors and do not approach the accuracy of ex-

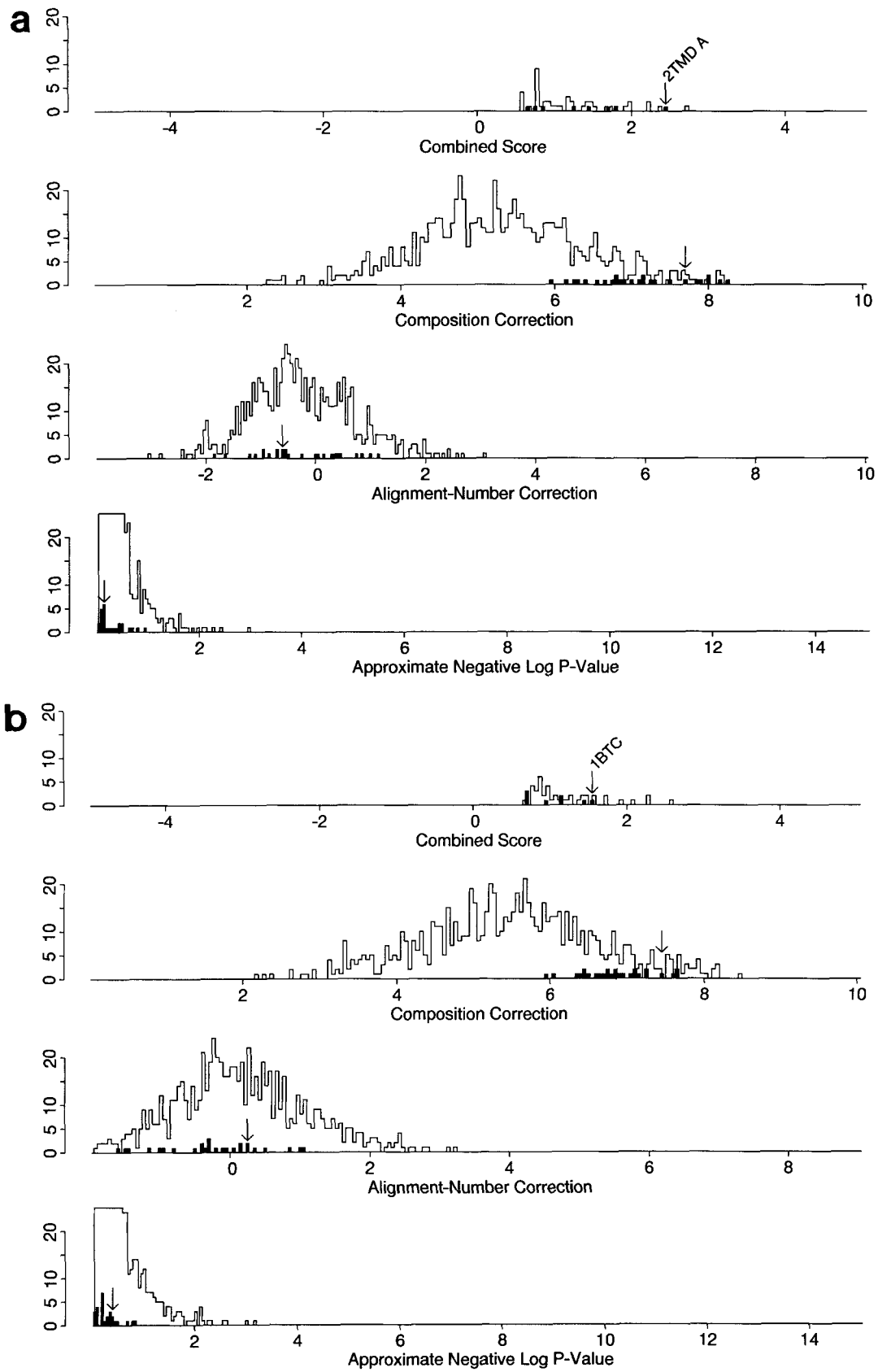

Fig. 1. Threading score distributions for *pbdg* (**a**) and *ppdk-4* (**b**), proteins where "core" definitions were incompatible with the observed common substructures. True negatives are shown as open, unshaded bars and true positives identified by structure comparison $p_d < 10^{-5}$ as solid bars. The uppermost panel shows "combined scores" Z_c for the top 50 hits, the values used to rank scores submitted for the contest. The next shows "composition correction" scores Z_{RIM} as defined previously,[19] including values for all database core motifs. "Alignment number correction" scores Z_N scale each composition-corrected score relative to the empirical distribution obtained by randomly shuffling the *pbdg* or *ppdk-4* sequence 50 times, each time optimally threading it through the corresponding core motif. Assuming this distribution to be normal, we obtain a chance occurrence probability by referring Z_N to the quantiles of the standard normal distribution.[5] The resulting *p*-values are shown as their negative logs to the base 10.

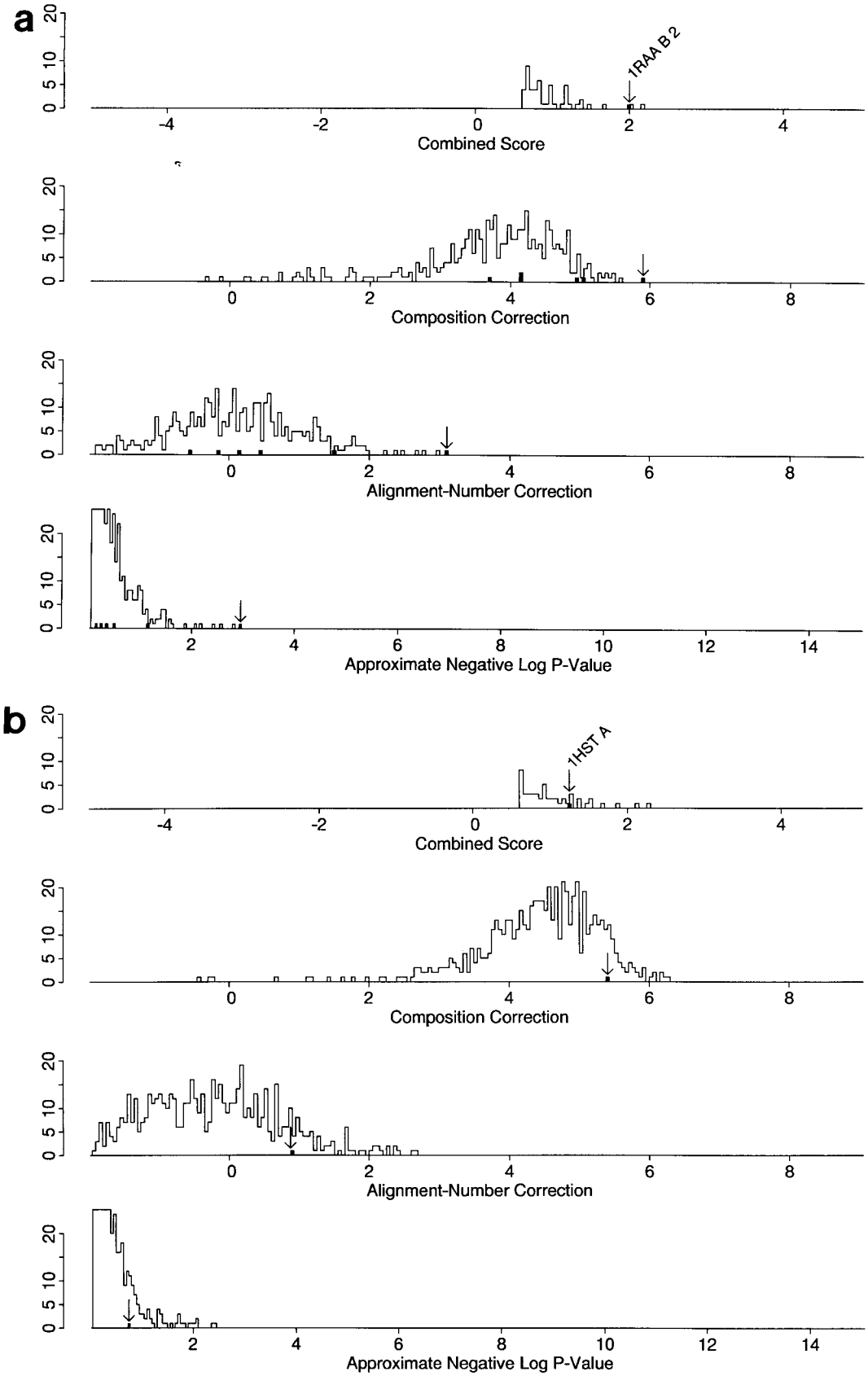

Fig. 2. Threading score distributions for *prosub* (**a**) and *rtp* (**b**), proteins where "core" definitions were compatible with the observed common substructures. Quantities are calculated and displayed as described in the caption to Figure 1.

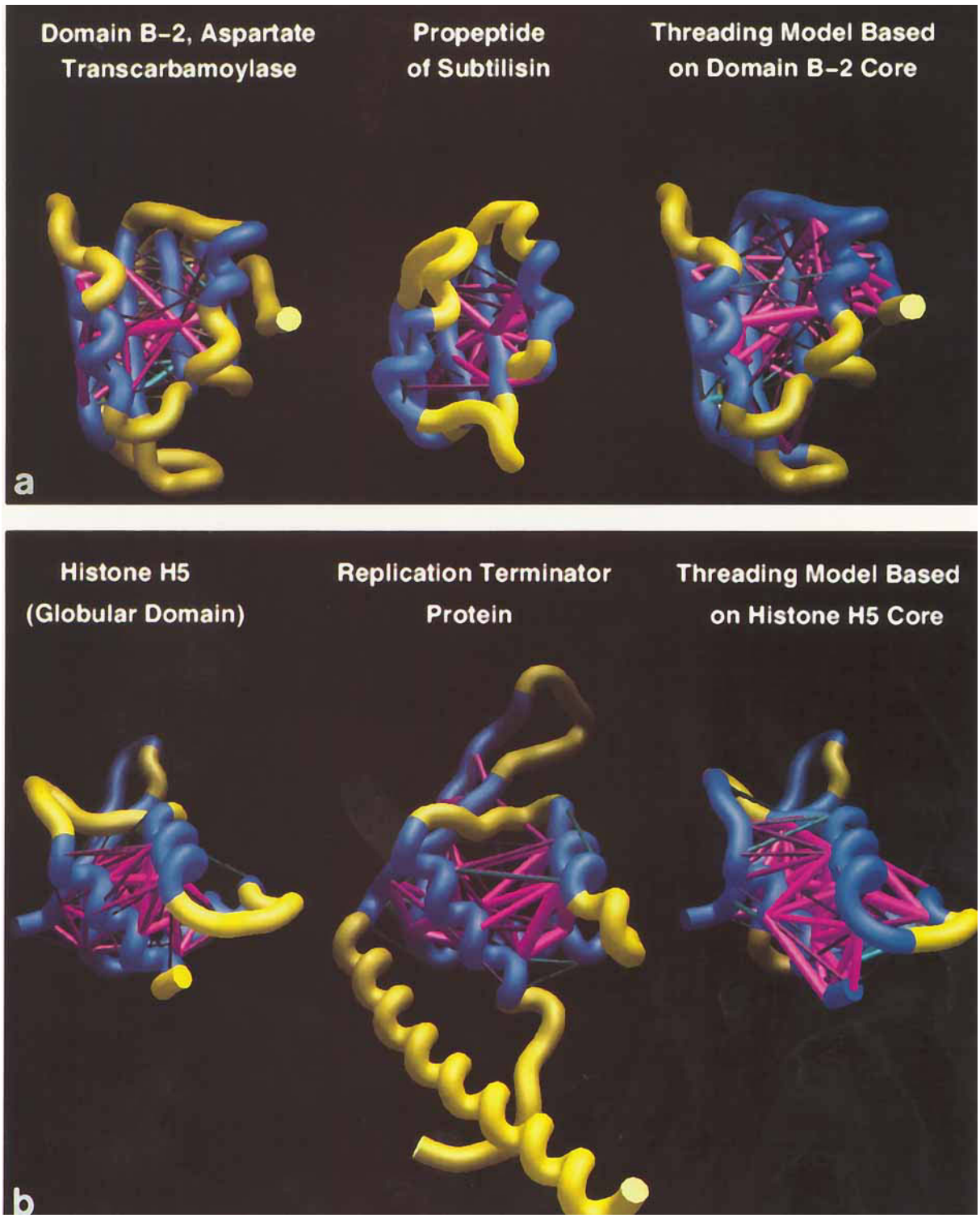

Fig. 3. Energy scaffold comparisons for *prosub* (**a**) and *rtp* (**b**). The left-most panels show the native structures of the database domains *1RAA-B-2* and *1HST-A*, with blue coloring of the C_α "worm" indicating the boundary of the common substructure shared with *prosub* or *rtp*, respectively. The large magenta-colored cylinders represent favorable pairwise interactions, and the cyan-colored ones unfavorable interactions, in the manner described previously.[19] The central panel shows in like manner the corresponding substructure in *prosub* or *rtp*. The right-most panel shows the submitted threading models for *prosub* or *rtp*. Blue coloring of the backbone worm indicates the regions of the known structure included in the model, the core elements plus additional recruited sites. The energy scaffold values derive from the sequences *prosub* or *rtp*, as aligned with the *1RAA-B-2* or *1HST-A* core structures. Figures were prepared with the program GRASP by Anthony Nicholls.[40]

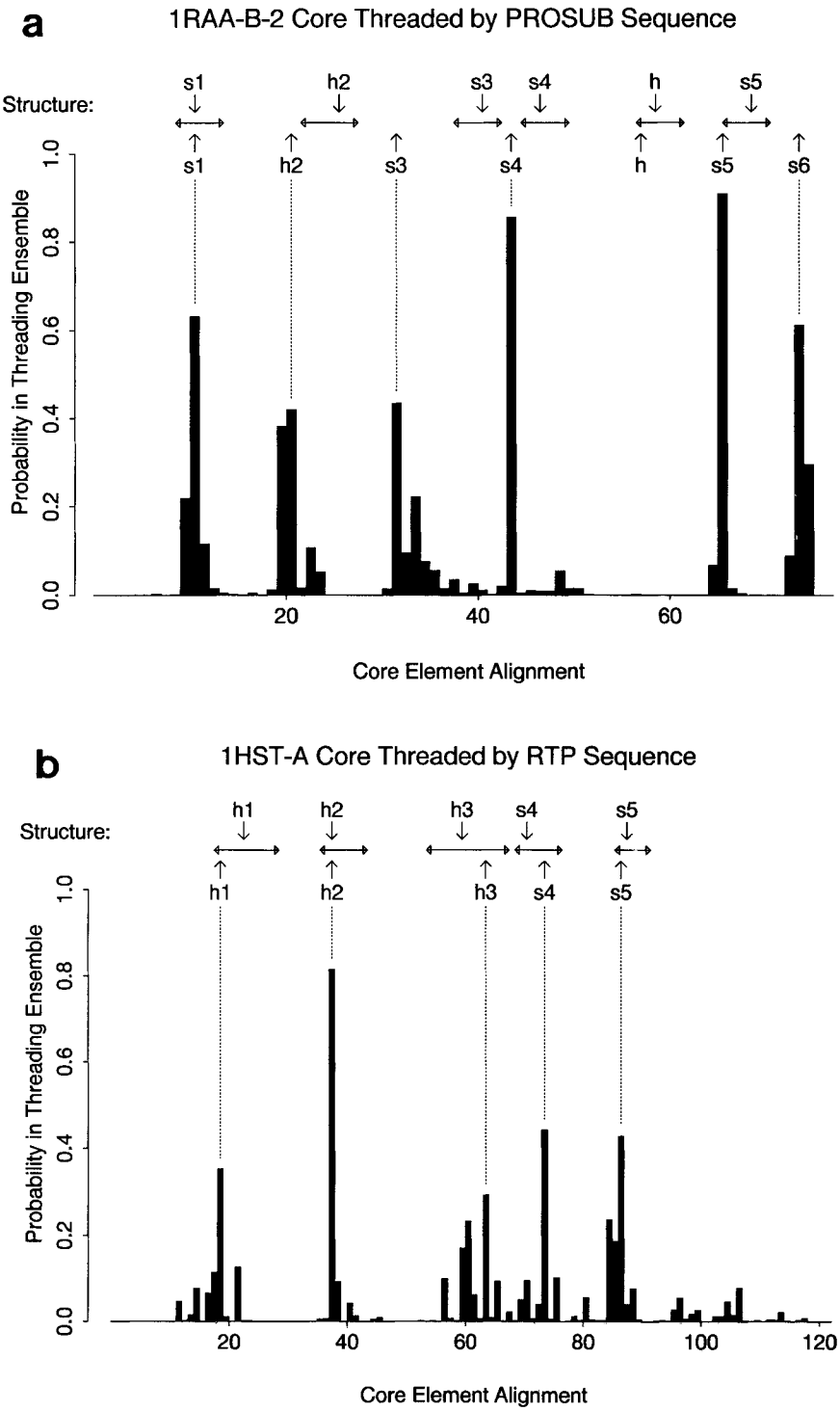

Fig. 4. Comparison of threading alignment ensembles for *prosub* and *1RAA-B-2* (**a**) and *rtp* and *1HST-A* (**b**) with their respective structural alignments. The histograms show the marginal probabilities with which *1RAA-B-2* or *1HST-A* core elements are aligned at alternative positions within the *prosub* or *rtp* sequences. The most probable alignments of individual core elements are indicated by vertical arrows. The regions of the *prosub* or *rtp* proteins which are structurally equivalent to the database domains are shown above the histogram, as horizontal lines labeled by the predominant secondary structure type of that segment, "h" for helix and "s" for strand, followed by an integer indicating chain order in the core definition. The arrows above these lines indicate the correct, structural alignment of the corresponding core element with *prosub* or *rtp.* They correspond to the "centers" of each core element, which serve as fiduciary marks for alignment, and which were defined automatically at residues 17, 29, 45, 59, 84, and 93 of *1RAA-B,* and residues 32, 51, 71, 82, and 93 of *1HST-A,* using PDB residue numbering. A threading alignment in perfect agreement with the structural superposition would align arrows above and below the line, and the probabilities of alternative "register shifts" may be read by the height of the bars in the histogram below and to either side of the correct alignment as indicated. Note that the core definition for *1RAA-B-2* omitted one helix structurally conserved between *prosub* and *1RAA-B-2.* Residues from the *prosub* sequence were aligned with this helix by recruitment of residues from the carboxy-terminal end of strand s4, however, and the resulting alignment is indicated. The core definition for *1RAA-B-2* also included a β-strand at its carboxy-terminus which is not present in *prosub,* and for which no correct alignment is possible.

perimental structures, and we present complete data on scores and alignments so that readers may judge for themselves "what went right" and "what went wrong."

DISCUSSION

Several results from these experiments support the conclusion that careful definition of the "fold" of a known structure or domain is essential to the success of blind threading predictions. The most obvious, perhaps, is that threading alignments for the *TIM*-barrels *pbdg* and *ppdk-4* were completely inaccurate, and necessarily so, since our prior definition of the "cores" of the database proteins did not agree well with the common substructures we subsequently identified. For the smaller proteins *prosub* and *rtp,* the cores we defined agreed reasonably well with the true common substructure, and the corresponding sequence–structure alignments were in turn reasonably accurate. The threading alignment produced for *rtp* was strikingly accurate for a protein with no detectable sequence similarity, with a mean alignment error over the common substructure of 2.7 residues corresponding entirely to "register shifts" of some secondary structure elements.

Given the inaccuracy of the *pbdg* and *ppdk-4* alignments it is tempting to conclude that we employed definitions of the database "folds" that were simply too specific, in that they supposed a too extensive substructure similarity. Another obvious conclusion from these experiments, however, is that threading scores identified the correct folds with a signal that is already at or below the level of the noise of false positives. The threading score of the correct fold for *prosub* was statistically significant, but not greatly so considering the size of the search database. There were other core motifs that scored nearly as well, for the most part other small domains with mixed β/α architecture, and it would appear that with only 52% of residues superimposable between *prosub* and *1RAA-B-2,* we are at the limit of distinguishing true positives. *rtp* yielded the most accurate threading alignment, but its threading scores were not statistically significant, and it was identified as a strong candidate only by using information concerning its biological function. The relatively low score may be explained by a structural similarity that encompasses only 45% of the sites, but the accuracy of the threading alignment is doubtless due to the high fraction of superimposable sites, 67%, in the globular domain that is structurally similar to *1HST-A.* It would appear from these considerations that the most important factor contributing to a favorable threading score for true positives is a large fraction of residues aligned with core sites in the known domain structure, as opposed to assigned to nonequivalent loop sites. Accuracy in alignment may not require global similarity in this way, but nonetheless requires a high percentage of equivalent core sites in a compact domain. From these observations one may argue that the definition of "folds" should in fact require quite extensive similarity to known "core" substructures, since this is what we can reliably detect using current potentials.

This is not to say that the database of protein "cores" prepared for the Asilomar contest was a satisfactory representation of known "folds," and in retrospect we can easily suggest improvements. Sometimes, as in the case of *prosub,* the minimum sizes we assumed for core elements forced inclusion of nonequivalent sites into the threading model. Minimum size constraints on individual core elements are not critical to the alignment model, since it is the small number of core elements, not their size, that leads to the critical reduction of the space of alternative alignments. These constraints may in future be easily relaxed or eliminated. As seen for both *prosub* and *rtp,* the automatically defined loop length constraints also can force partial misalignment. This aspect of core definition was intended originally for reverse folding searches with protein families where patterns of loop length variations are known from multiple sequences or structures,[19] and it is probably inappropriate when cores are defined automatically by geometric criteria. It may be preferable to require only that a certain fraction of residues from the domain subsequence be assigned to core sites, rather than nonequivalent "loop" sites. This change is easily accommodated in the algorithm used for alignment optimization, and it is also possible, given the formalism of Gibbs sampling, to include prior probabilities reflecting one's expectation that particular sites are likely to be conserved in any common "core" substructure.

Results of the structure–structure similarity searches suggest that one should also consider a definition of "folds" based on substructure recurrence.[36,37] A very large number of substructure similarities are found for *pbdg, ppdk-4,* and *prosub,* and *rtp* shows similarity, albeit remote, to other DNA binding proteins in addition to *1HST-A.* By having compared the similar domains in the known-structure database one might have identified structural changes accommodated by protein evolution and/or folding, and better incorporated this information into a quantitative description of their folds. One might distinguish secondary structure elements which are always present, for example, from those which are deleted in certain structures. Such information can also be treated in the Gibbs sampling algorithm as a prior probability,[38] and doing so might have allowed us to detect, for example, that *prosub* lacks an edge β-strand present in *1RAA-B-2.*

Threading techniques are new and they represent a first generation of recognition methods for structure prediction. Their scoring potentials, alignment algorithms, and representations of the "folds" of

known structures may certainly be improved, and better understanding of the statistics of score distributions may allow more reliable predictions with quantitative confidence limits. It is nonetheless striking to the present authors that we find ourselves discussing the limits to success in 2 of 4 cases, rather than failure in them all. It would seem that the various workers in this area have indeed learned how to use the known-structure database for nontrivial predictions, and that a recent prediction of progress,[39] at least, has proven correct.

ACKNOWLEDGMENTS

This work is has been supported by the NIH Intramural Research Program. We thank the many structural biologists who kindly provided the "targets" for blind prediction, and the organizers of the Asilomar contest.[1] We also thank Anthony Nicholls for assistance in data transfer between GRASP[40] and PKB[41] and Chris Hogue for his implementation of an algorithm[42] to derive backbone-atom coordinates from a C_α model.

REFERENCES

1. Moult, J., Pedersen, J., Fidelis, K., Balhor, R., Judson, R., Stevens, W. (organizers) Meeting on the Critical Assessment of Techniques for Protein Structure Prediction, Asilomar Conference Center, California, December 4–8, 1994.
2. Rost B., Sander, C. Structure prediction of proteins—where are we now? Curr. Opinion Struct. Biol. 5:372–380, 1994.
3. Fetrow, J.S., Bryant, S.H. New programs for protein tertiary structure prediction. Bio/Technology 11:479–484, 1993.
4. Wodak, S.J., Rooman, M.J. Generating and testing protein folds. Curr. Opinion Struct. Biol. 3:247–259, 1993.
5. Bryant, S.H., Altschul, S.F. Statistics of sequence-structure threading. Curr. Opinion Struct. Biol. 5:236–244, 1995.
6. Jones, D.T. De novo protein design using pairwise potentials and a genetic algorithm. Protein Sci. 3:567–574, 1994.
7. Bryant, S.H. In preparation.
8. Chothia, C., Lesk, A.M. The relation between the divergence of sequence and structure in proteins. EMBO J. 5:823–826, 1986.
9. May, A.C.W., Blundell, T.L. Automated comparative modelling of protein structures. Curr. Opinion Biotechnol. 5:355–360, 1994.
10. Zhang, K.Y., Eisenberg, D. The three-dimensional profile method using residue preference as a continuous function of residue environment. Protein Sci. 3:687–695, 1994.
11. Jones, D.T., Taylor, W.R., Thornton, J.M. A new approach to protein fold recognition. Nature (London) 358:86–89, 1992.
12. Godzik, A., Kolinski, A., Skolnick, J. Topology fingerprint approach to the inverse folding problem. J. Mol. Biol. 227: 227–238, 1992.
13. Sippl, M.J., Weitckus, S., Flockner, H. In search of protein folds. In: "The Protein Folding Problem and Tertiary Structure Prediction." Merz, K.M., Le Grand, S.M. (eds.). Boston, MA: Birkhauser, 1994: 353–408.
14. Abagyan, R., Frishman, D., Argos, P. Recognition of distantly related proteins through energy calculations. Proteins 19:132–140, 1994.
15. Goldstein, R., Luthey-Shculten, Z.A., Wolynes, P.G. Protein tertiary structure recognition using optimized Hamiltonians with local interactions. Proc. Natl. Acad. Sci. U.S.A. 89:9029–9033, 1993.
16. Johnson, M.S., Overington, J.P., Blundell, T.L. Alignment and searching for common protein folds using a data bank of structural templates. J. Mol. Biol. 231:735–752, 1993.
17. Nishikawa, K., Matsuo, Y. Development of pseudoenergy potentials for assessing protein 3-D-1-D compatibility and detecting weak homologies. Protein Eng. 6:811–820, 1993.
18. Ouzounis, C., Sander, C., Scharf, M., Schneider, R. Prediction of protein structure by evaluation of sequence-structure fitness. Aligning sequences to contact profiles derived from three-dimensional structure. J. Mol. Biol. 232:805–825, 1993.
19. Bryant, S.H., Lawrence, C.E. An empirical energy function for threading protein sequence through folding motif. Proteins 16:92–112, 1993.
20. Lathrop, R.H., Smith, T.F. A branch and bound algorithm for optimal protein threading with pairwise (contact potential) amino acid interactions. In: "Proceedings of the 27th Hawaii International Conference on System Sciences," Vol. 5. Hunter, L. (ed.). Los Alamitos, CA: IEEE Computer Society Press, 1994:365–376.
21. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., Weng, J.C. Protein data bank. In: "Crystallographic Databases: Information Content, Software Systems, Scientific Applications." Allen, F.H., Bergerhoff, G., Sievers, R. (eds.). Bonn: Int. Union of Crystallography, 1987:107–132.
22. Richards, F.M., Kundrot, C.E. Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure. Proteins 3:71–84, 1988.
23. Holm, L., Sander, C. Parser for protein folding units. Proteins 19:256–268, 1994.
24. Koonin, E.V., Mushegian, A.R., Tatusov, R.L., Altschul, S.F., Bryant, S.H., Bork, P., Valencia, A. Eukaryotic translation elongation factor 1γ contains a glutathione transferase domain—Study of a diverse, ancient protein superfamily using motif search and structural modeling. Protein Sci. 3:2045–2054, 1994.
25. Gibrat, J.-F., Madej, T., Bryant, S.H. Finding the most surprising structural similarities. In preparation.
26. Grindley, H.M., Artymuik, P.J., Rice, D.W., Willett, P. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. J. Mol. Biol. 229:707–721, 1993.
27. Altschul, S.F., Boguski, M.S., Gish, W., Wootton, J.C. Issues in searching molecular sequence databases. Nature Genet. 6:119–129, 1994.
28. Lolis, E., Alber, T., Davenport, R.C., Rose, D., Hartman, F.C., Petsko, G.A. Structure of yeast triosephosphate isomerase at 1.9 Ångstroms resolution. Biochemistry 29: 6609–6618, 1990.
29. Kissinger, C.R., Sieker, L.C., Adman, E.T., Jensen, L.H. Refined crystal structure of ferredoxin II from Desulfovibrio gigas at 1.7 Ångstroms resolution. J. Mol. Biol. 219:693–715, 1991.
30. Ramakrishnan, V., Finch, J.T., Graziano, V., Lee, P.L., Sweet, R.M. Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. Nature (London) 362:219–223, 1993.
31. Wilson, K.P., Shewchuk, L.M., Brennan, R.G., Otsuka, A.J., Matthews, B.W. The E. coli biotin holoenzyme synthetase/bio repressor crystal structure delineates the biotin- and DNA-binding domains. Proc. Natl. Acad.Sci. U.S.A. 89:9257–9261, 1992.
32. Weber, I.T., Steitz, T.A. Structure of a complex of catabolite gene activator protein and cyclic AMP refined at 2.5 Ångstroms resolution. J. Mol. Biol. 198:311–326, 1987.
33. Barber, M.J., Neame, P.J., Lim, L.W., White, S., Matthews, F.S. Correlation of x-ray deduced and experimental amino acid sequences of trimethylamine dehydrogenase. J. Biol. Chem. 267:6611–6619, 1992.
34. Mikami, B., Sato, M., Shibata, T., Hirose, M., Aibara, S., Katsube, Y., Morita, Y. Three-dimensional structure of soybean beta-amylase determined at 3.0 Ångstrom resolution: Preliminary chain tracing of the complex with alpha-cyclodextrin. J. Biochem. 112:541–546, 1992.
35. Kosman, R.P, Gouauz, J.E., Lipscomb, W.N. Crystal structure of CTP-ligated T state aspartate transcarbamoylase at 2.5 Ångstroms resolution: Implications for ATCase mutants and the mechanism of negative cooperativity. Proteins 15:147–176, 1993.

36. Holm, L., Sander, C. Searching protein structure databases has come of age. Proteins 19:165–173, 1994.
37. Orengo, C.A., Jones, D.T., Thornton, J.M. Protein superfamilies and domain superfolds. Nature (London) 372:631–634, 1994.
38. Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., Wootton, J.C. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. Science 262:208–214, 1993.
39. Thornton, J.M., Flores, T.P., Jones, D.T., Swindells, M.B. Protein structure. Prediction of progress at last. Nature (London) 354:105–106, 1991.
40. Nicholls, A., Sharp, K.A., Honig, B. Protein folding and association: Insights from the thermodynamic properties of hydrocarbons. Proteins 11:281–296, 1991.
41. Bryant, S.H. PKB: A program system and data base for analysis of protein structure. Proteins 5:233–247, 1989.
42. Rey, A., Skolnick, J. Efficient algorithm for the reconstruction of a protein backbone from the alpha-carbon coordinates. J. Comp. Chem. 13:443–456, 1992.