Biochemical and Biophysical Research Communications **254,** 70–76 (1999)
Article ID bbrc.1998.9884, available online at http://www.idealibrary.com on IDEAL®

Evolutionary, Mechanistic, and Predictive Analyses of the Hydroxymethyldihydropterin Pyrophosphokinase Family of Proteins

Dietlind L. Gerloff,* Gina M. Cannarozzi,† Marcin Joachimiak,* Fred E. Cohen,* David Schreiber,† and Steven A. Benner†,‡

**Department of Cellular and Molecular Pharmacology, University of California, San Francisco, California 94143; †Department of Chemistry, University of Florida, Gainesville, Florida 32611; and ‡Department of Anatomy and Cell Biology, University of Florida, Gainesville, Florida 32611*

Received November 16, 1998

A prediction has been prepared *ab initio* for the secondary structure of the hydroxymethyldihydropterin pyrophosphokinase (HPPK) family of proteins starting from a set of aligned homologous protein sequences. Attempts to identify a fold by threading failed, judging by the inability to find a threading "hit" that had a secondary structure that was plausibly congruent to the predicted secondary structure for the HPPK family. Therefore, a set of tertiary structure models was assembled *ab initio,* where alternative models were built and used to select between alternative secondary structure models. This prediction report illustrates the importance of non-computational approaches to structure prediction at its present frontier, which is to obtain medium resolution models of tertiary structure. © 1999 Academic Press

Three general conclusions can be drawn from recent experience testing tools for modeling protein folding using *bona fide* predictions (1), those made and announced before an experimental result is known (as distinct from tests that retroactively apply a tool to a database of known structures, a process that is also frequently termed "prediction") (2). These have recently been reviewed (1).

First, methods based on an analysis of multiple sequences of homologous proteins frequently yield accurate models for the core elements of secondary structure (those that are shared by all proteins in a family). The use of *bona fide* predictions to test these methods has helped dispel much of the skepticism concerning secondary structure prediction that has been expressed by experimentalists in the past (3).

Second, the accuracy of these models has been sufficient to make them useful for solving biological problems, including detecting distant homology (4), selecting targets for pharmaceutical development programs, and inferring physiological roles of proteins from genomic sequences (1).

Third, the prediction of the conformation of a specific sequence at the level of atomic coordinates generally remains far beyond reach. Too little is known about fundamental chemistry (the structure of water, its interaction with solutes, and the interaction between solutes in water, for example) to make the task tractable even in cases where the computational complexity of the problem can be handled. Indeed, our understanding of underlying chemistry appears to be insufficient to solve formally "simple" problems, such as the prediction of the conformation of a protein sequence that is a close homolog of a protein whose atomic coordinates are already known (5). Further work developing a fundamental understanding of the interactions between solvents and solutes in smaller molecules will need a high priority.

These conclusions suggest that for now, measurements of improvements in structure prediction methods must focus on how well those methods convert a secondary structure prediction (including one that contains ambiguities) into "medium resolution" models of the fold "topology", as this is the process that is frequently, but not universally, successful. As in other areas of conformational analysis in organic chemistry, we expect that the development of an understanding of this process will involve human interaction with data, not fully automated computer analysis (6).

An especially useful tool to represent protein structures at this level of resolution is the "segment contact" representation introduced by Lesk (7). The Lesk representations present a protein fold at the position between where prediction success and failure abut. A recent characterization of hydroxymethyldihydro-

0006-291X/99 $30.00

pterin pyrophosphokinase (HPPK, E.C. 2.7.6.3), an enzyme in the folate biosynthesis pathway, suggested that a crystal structure of this protein would be of interest as HPPK is a potential target for antimicrobial and antifungal therapeutic agents (8). Confirming this suggestion is the fact that this protein was submitted by Xiao, Yan, and Ji to the project known as "Critical Assessment of Structure Prediction" (CASP3, http://PredictionCenter.llnl.gov). With 13 identifiably homologous sequences, the target seemed ideal to test a broad range of tools based on multiple sequence alignments. In particular, we document a specific case where tertiary structural modeling is used to explore ambiguities in a secondary structure prediction.

MATERIAL AND METHODS

The multiple sequence alignment shown in Figure 1 was prepared using DARWIN 2.0 (9), improved to incorporate new gap placement heuristics. Positions in the multiple sequence alignment containing amino acids whose side chains lie on the surface of the fold, in the interior, in the active site and in parses in the protein fold were assigned as reviewed recently, using the DARWIN tool available via a server (www.cbrg.inf.ethz.ch). Consensus secondary structure predictions were made from these using the program Structure Assignment with Informative Transparency (SAINT) using procedures recently reviewed (1), supplemented by expert analysis.

Maximum likelihood trees were prepared using the DARWIN server, and compared with maximum parsimony trees prepared using the MacClade program (10). Reconstructed ancestral sequences were used to calculate the ratios of expressed/silent substitution using a program that implemented the method of Li et al. (11, 12). Tertiary structure modeling followed analyses used for the prediction of the tertiary fold of protein kinase (13), synaptotagmin (14), and phospho-beta-galactosidase (15), and is discussed below.

RESULTS

A secondary structure model (Table I, Figure 1) was first generated for HPPK. The "transparent" prediction method (1) suggested that two helices (marked as "H"), three strands (marked as "E"), and two active site regions (the conserved RXXDXD and PH elements) are "reliable". These form the center of any attempt to model the tertiary structure of the protein. One additional helix and one additional strand are assigned tentatively, with the possibility reserved during tertiary structural modeling that the assignments could be reversed (but in no case be modeled as coil regions). Two additional segments are modeled tentatively as strands, with the option during tertiary structural modeling of regarding these as coils (but not helices).

The PepPep search tool within Darwin was used to search for long distance homologs for HPPK. This recovered the sequence for pyruvate phosphate dikinase (PPDK, also known as pyruvate orthophosphate phosphotransferase, E.C. 2.7.9.1) from the database. PPDK converts ATP, inorganic phosphate and pyruvate into AMP, pyrophosphate, and phosphoenolpyruvate. The enzyme has been proposed to proceed via an intermediate where the enzyme is phosphorylated on His (16). The chemical similarities between the reactions catalyzed by PPDK and HPPK, the presence of a conserved His in HPPK assigned to the active site, and an RXXD sequence in both protein families putatively involved in catalysis provided suggestive evidence supporting distant homology between PPDK and HPPK. The similarities in the sequences of the two proteins is clearly sub-significant, however, extends over only part of the HPPK sequence, and does not include the putative catalytic histidine in PPDK.

In other cases, similarities of this type have proven not to be conclusive statements either supporting or denying long distance homology. In the ribonucleotide reductase superfamily, for example, (4) similarly poor sequence similarity is found between the B12-dependent and Fe-dependent enzymes; the mechanistic analogies joining the two protein families turned out to provide a correct prediction of long distance homology and analogous fold (17). In contrast, the protein kinase/adenylate kinase pair of proteins displayed more sequence similarities as well as analogous reaction types. Nevertheless, the inference that these proteins were homologous, which was used in three dimensional modeling by several laboratories (18, 19), proved to be incorrect (20).

Structure prediction can be used to confirm or deny conjectured homology based on sub-significant similarities and mechanistic analogy (13). Predictions are first made for the secondary (and, if possible, the tertiary) structure of the two protein families. These are then compared to ascertain whether their core elements are congruent. If they are, the conjecture is supported; if they are not, the conjecture is denied. This approach is especially easy to follow for PPDK, as an experimental structure is known for the protein (16). In PPDK, the active site His (His 454) is embedded in a phosphohistidine domain that folds as an eight fold alpha-beta barrel. Even considering ambiguities in the prediction of HPPK, the protein family cannot adopt the same conformation as the PPDK family. Thus, it is predicted that the conjectural similarities in the sequences of the two proteins is not indicative of either distant homology or analogous folds.

A second approach for detecting possible long distance homologs is based on a proposal from Ornston, who suggested that enzymes catalyzing consecutive steps in a metabolic pathway may have evolved from a common ancestor (21, 22). In particular, the dihydrofolate reductase (Brookhaven 1ra2) in the pathway had a similar size and similar overall secondary structure composition as HPPK. Inspection of alternative tertiary folds (see below) failed to find congruency between the HPPK predicted and the 1ra2 experimental structures, and this was ruled out as a possible homolog.

Threading was then applied using the UCLA DOE server (http://www.doe-mbi.ucla.edu/people/frsvr/frsvr.

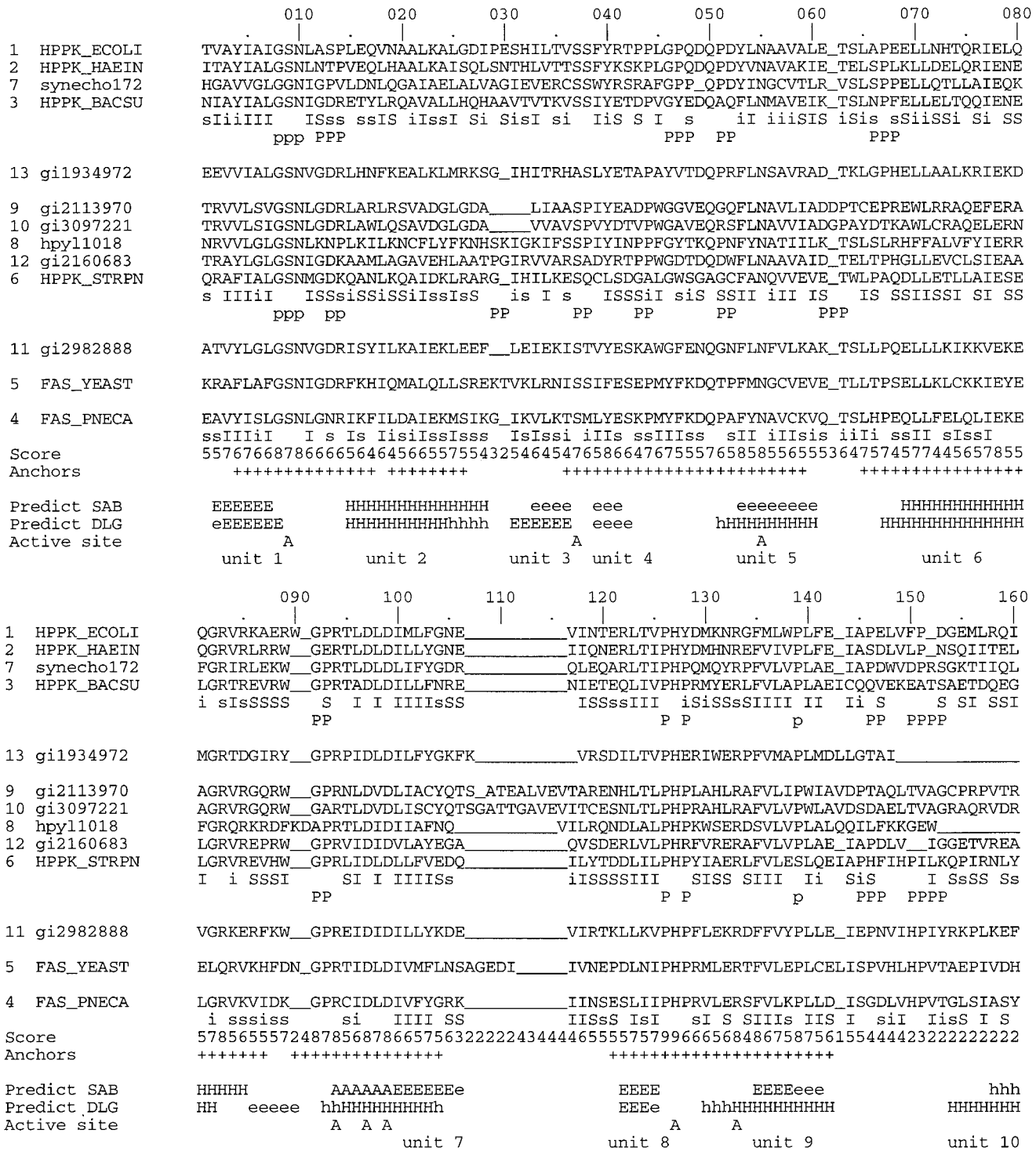

FIG. 1. Multiple sequence alignment, surface (S,s), interior (I,i), active site (A) and parse (P) assignments, and secondary structure prediction for the hydroxymethyldihydropterin pyrophosphokinase (HPPK) family of proteins. Helix (reliable and suggested, H and h) and strand (reliable and suggested, E and e) residues are designated below the sequences, given in the one letter code for amino acids. Underscore (_) designates a deletion/insertion.

html) (23) and the ProFit tool (http://www.horus.com/sippl/). The UCLA server did not generate any "significant" hits, and the one borderline hit did not match any of the features of the *ab initio* predicted structure for HPPK. The ProFit tool generated 15 hits, including some all-helix proteins, some all strand proteins, and

TABLE I

Assignment of Elements of Secondary Structure to the Hydroxymethyldihydropterin Pyrophosphokinase (HPPK) Family of Proteins

Segment number	Alignment positions	Reliable assignments	Support	Tentative/proposed
1	002–007	**strand**	buried (6)	
	008–014	parse		
2	015–028	**helix**	amphiphilic	
	029–032	parse	PG/gap	
3	033–036			**strand,** not buried
	037–038	parse	SP dipeptide	
4	039–041			**strand,** not buried
	042–051	parse	PPxGPx_xP	
5	052–060			**helix,** if conserved N on surface
	061–063	parse	DGP/deletion	
	064–065			coil
	066–068	parse	SPP	
6	069–085	**helix**	amphiphilic	
	086–093	parse	xPxx_GP	
	094–099	active site	RxxDxD	
7	100–106	**strand**	buried	must bury 4 residues
	107–116	parse	gap placement	
8	117–125		LTV, LKV	**strand 122–125**
	126–128	active site	PHP	parse also
9	129–138	unknown	135–138	**strand 135–138**
	139–142			possible extension of strand
	143–146	parse	_xxPD	
	147–149			coil
	150–157	parse	P_N	
	158–172	non-core		**helix** possible

some alpha-beta proteins, all with rather similar rankings. Interestingly, ProFit identified dihydropteridine reductase (Brookhaven 1dhr) as a possible hit. Again, however, the proposed threading alignment was not congruent with the secondary structure prediction. The most likely known structure to plausibly fit the HPPK prediction is that for phosphofructokinase (PFK), although specific issues arising when attempting to superimpose the two structures (for example, the orientation of secondary structural element 6 and following elements) forced the conclusion that HPPK is not a distant homolog of PFK.

Failing to find a recognizable homolog of HPPK in the database, we turned to building by direct assembly a tertiary structure model, following combinatorial procedures (24) similar to those used to build models for protein kinase (13), synaptotagmin (14), and phospho-beta-galactosidase (15). The HPPK protein family proved to be especially difficult because its members are joined by a problematic evolutionary tree where deep branchings are connected by short edges; this means that the connectivity of the tree is unreliable.

This sub-optimal tree is, in part, responsible for ambiguities in the secondary structural model (Table I). It also influences the reliability of several tools for assembling predicted secondary structural elements into a tertiary structure. For example, compensatory covariation, rather successful in the protein kinase prediction, (13) proved inapplicable for HPPK. It is known that the success of such an approach depends on the evolutionary distance separating sequences (25), and few pairs of members with the appropriate distances were found in the HPPK family.

Therefore, assembling a tertiary structural model began by identifying the least buried strands in the collection of strands and placing them at the edge of the sheet. In the initial set of predicted secondary structural units, segments 3 and 4 would provide the least buried strands. Because they are joined by only a dipeptide, however, it was not possible to plausibly place them at opposite ends of a sheet in the model building. Therefore, the first round of model building chose one of the two as the edge, and sought another strand in the collection to form the other edge. This was assigned as segment 8. In a protein as small as HPPK, only one or two sheets are conceivable. The absence of clear choices for four edge strands (in fact, only segments 3 and 4 are truly good candidates for edge strands) and the absence of clearly amphiphilic strands restricted the models to those containing a single sheet.

For a six stranded sheet, nearly 300 connectivities are possible. Adding the predicted helices to these sheets increased the number of possible medium resolution models for the protein to over 2000. Once edge strands are chosen, 72 distinct sheets are possible. Each of these was explicitly built, and the collection

TABLE II

Lesk Segment Contact Tableaux* for the Top Six Tertiary Structure Models for the Hydroxymethyldihydropterin Pyrophosphokinase (HPPK) Family of Proteins

Model 1	1	2	3	4	5	6	7	8	9
	β	α	β	β	α	α	β	β	α
1 β		OS	KK	-	OS?	PE	HH	-	-
2 α			OS	-	-	-	-	-	PE
3 β				HH	OS	-	-	-	-
4 β					PD	-	-	-	-
5 α						RT	-	-	-
6 α							OT	-	-
7 β								HH	PE
8 β									OT

Model 2	1	2	3	4	5	6	7	8	9
	β	α	β	β	α	α	β	β	β
1 β		OS	KK	-	OS?	PE	-	-	HH
2 α			OS	-	-	-	-	-	PD
3 β				HH	OS	-	-	-	-
4 β					PD	-	-	-	-
5 α						RT	-	-	-
6 α							OT	-	OT
7 β								HH	KK
8 β									-

Model 3	1	2	3	4	5	6	7	8	9
	β	α	β	β	β	α	β	β	α
1 β		OS	-	-	HH	PE	HH	-	-
2 α			OS	-	PD	-	-	-	PE
3 β				HH	HH	-	-	-	-
4 β					-	-	-	-	-
5 β						OT	-	-	-
6 α							OT	-	-
7 β								HH	PE
8 β									OT

Model 4	1	2	3	4	5	6	7	8	9
	β	α	β	β	α	α	β	β	β
1 β		OS	KK	HH	OT	-	-	-	-
2 α			OS	-	-	-	-	-	PD
3 β				-	OT	PD	-	-	HH
4 β					PE	-	-	-	-
5 α						LS	-	-	-
6 α							OS	-	OS
7 β								HH	KK
8 β									-

Model 5	1	2	3	4	5	6	7	8	9
	β	α	β	β	α	α	β	β	α
1 β		OS	KK	HH	OT	-	-	-	-
2 α			OS	-	-	-	-	-	PD
3 β				-	OT	PD	HH	-	-
4 β					PE	-	-	-	-
5 α						LS	-	-	-
6 α							OS	-	-
7 β								HH	PD
8 β									OS

Model 6	1	2	3	4	5	6	7	8	9
	β	α	β	β	β	α	β	β	α
1 β		OS	KK	HH	-	OS	-	-	-
2 α			OS	-	PE?	-	-	-	PD
3 β				-	HH	PD	-	-	-
4 β					-	-	-	-	-
5 β						PD	KK	-	OS
6 α							OS	OS?	-
7 β								HH	PD
8 β									OS

* The matrix is symmetric around the diagonal. Indices (top row and left column) designate secondary structural elements (numbered consecutively; underlined elements are the same in all *tertiary* structural models. A "-" indicates that the index elements are not in contact. HH and KK denote antiparallel and parallel relations between strands adjacent in the same sheet. Letters designate orientation of other secondary structural elements (D, 0–90°; R, 45–135°; T 90–180°; O, 135–225°; S, 180–270°; L, 225–315°; E, 270–360°; P, 315–45°).

culled by excluding those that severely violated tertiary folding rules (26), or that did not assemble the predicted active site residues in a way that chemical reaction mechanism theory suggested would allow HPPK to catalyze its reaction. These included the placement of a putative metal binding site (the Asp residues in the RxxDxD sequence and the conserved N, for example), and the H in the PHP sequence as a general base, and the conserved R to stabilize the transition state.

The process of tertiary structural modeling caused us to rethink the assignments made for elements 5 and 9. In particular, segment 9 was considered to be a possible helix because it was viewed as possibly burying parts of the sheet that the surface-interior assignments suggested were not exposed. Assigning segment 5 as a strand was considered to make possible a more optimal orientation of active site residues. This led to alternative tertiary structural models. From several thousand alternative models, these considerations generated six medium resolution models for the tertiary structure of HPPK, which were then ranked for their overall ability to accommodate active site residues, conform to empirical rule for tertiary folds, and conform to the secondary structure prediction. These are represented in Table II using the Lesk "segment contact" formalism (6), and graphically in Figure 2.

DISCUSSION

The weaknesses of contemporary secondary structure prediction tools based on an analysis of a set of aligned homologous protein sequences have recently been reviewed (1). Such tools work best when presented with a well balanced set of sequences, have difficulties assigning secondary structure near active sites, and occasionally have difficulties distinguishing between exposed strands and coils.

The HPPK family of proteins presented a challenge because of the unbalanced nature of the tree describing the family of proteins. This contributed to several ambiguities in the secondary structural assignment and the absence of useful compensatory covariation hits, both important in building tertiary structural models. Accordingly, tertiary structure modeling relied more than usual on combinatorial sheet construction and a concept of the mechanism by which the enzyme catalyzed a reaction. The value of this Prediction Report in recording this prediction ultimately resides in its ability to be re-examined once an experimental structure is known to see whether these subjective tools were useful in a *bona fide* prediction setting. If so, and if their use is confirmed in other cases, it will be worthwhile to develop automated tools that incorporate them when modeling tertiary structure from predicted secondary structural elements.

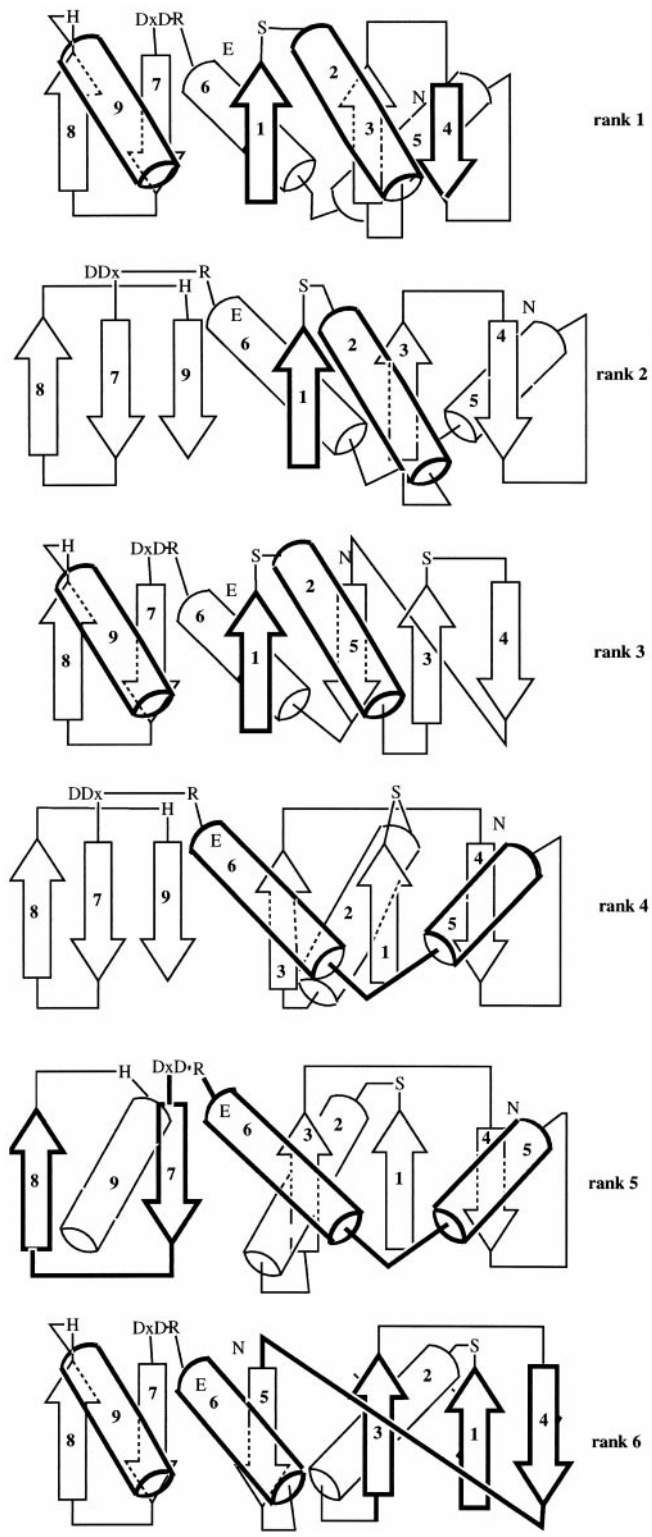

FIG. 2. The top six tertiary structure models for hydroxy-methyldihydropterin pyrophosphokinase (HPPK) ranked in order of reliability. The model of rank 1 was built in SYBYL and submitted as a coordinate model to the CASP3 structure prediction contest. Secondary structural elements are numbered consecutively (see Figure 1 and Tables I and II). Strands are indicated by "arrows"; helices by cylinders; coils by lines.

REFERENCES

1. Benner, S. A., Cannarozzi, G., Chelvanayagam, G., Turcotte, M. (1997) *Chem. Rev.* **97,** 2725–2843.
2. Robson, B., Garnier, J. (1993) *Nature* **361,** 506.
3. Hunt, T., Purton, M. (1992) *Trends Biochem. Sci.* **17,** 273.
4. Tauer, A., Benner, S. A. (1997) *Proc. Natl. Acad. Sci.* **94,** 53–58.
5. Moult, J., Pedersen, J. T., Judson, R., and Fidelis, K. (1995) *Proteins* **23,** ii–v.
6. Pauling, L., and Corey, R. B. (1950) *J. Am. Chem. Soc.* **72,** 5349.
7. Lesk, A. M. (1995) *J. Mol. Graphics.* **13,** 159–164.
8. Talarico, T. L., Ray, P. H., Dev, I. K., Merrill, B. M., and Dallas, W. S. (1992) *Journal of Bacteriology* **174**(18), 5971–5977.
9. Gonnet, G. H., Cohen, M. A., and Benner, S. A. (1992) *Science* **256,** 1443–1445.
10. W. P. Maddison, D. R. Maddison, and MacClade. (1992) Analysis of Phylogeny and Character Evolution. Sinauer Associates, Sunderland MA.
11. Li, W.-H., Wu, C.-I., and Luo, C.-C. (1985) *Mol. Biol. Evol.* **2**(2), 150–174.
12. Li, W. H. (1993) *J. Molec. Evol.* **36,** 96–99.
13. Benner, S. A., and Gerloff, D. L. (1991) *Adv. Enz. Reg.* **31,** 121–181.
14. Gerloff, D. L., Chelvanayagam, G., and Benner, S. A. (1995) *Proteins. Struct. Funct. Genet.* **21,** 299–310.
15. Gerloff, D. L., and Benner, S. A. (1995) *Proteins. Struct. Funct. Genet.* **21,** 273–281.
16. Herzberg, O., Chen, C. C. H., Kapadia, G., McGuire, M., Carroll, L. J., Noh, S. J., and Dunaway-Mariano, D. (1996) *Proc. Nat. Acad. Sci.* **93,** 2652–2657.
17. S. S., Holler, T. P., Yu, G. X., Bollinger, J. M., Booker, S., Johnston, M. I., and Stubbe J. (1992) *Biochemistry* **31,** 9733–9743.
18. Sternberg , M. J. E., and Taylor, W. R. (1984) *FEBS Lett.* **175,** 387–392.
19. Fry, D. C., Kuby, S. A., and Mildvan, A. S. (1986) *Proc. Natl. Acad. Sci. U. S. A.* **83,** 907–911.
20. Knighton, D. R., Zheng, J., Ten Eyck, L., Ashford, F. V. A., Xuong, N. H., Taylor, S. S. , and Sowadski, J. M. (1991) *Science* **253,** 407–414.
21. Yeh, W. K., and Ornston, L. N. (1980) *Proc. Nat. Acad. Sci.* **77,** 5365–5369.
22. Yeh, W. K., Fletcher, P., and Ornston, L. N. (1980) *J. Biol. Chem.* **255,** 6342–6346.
23. Rice, D. W., and Eisenberg, D. (1997) *J. Mol. Biol.* **267,** 1026–1038.
24. Cohen, F. E., and Kuntz, I. D. *in* Prediction of Protein Structure and the Principles of Protein Conformation (Fasman, G. D., Ed.): New York, NY, pp. 647–705, 1989.
25. Chelvanayagam, G., Eggenschwiler, A., Knecht, L., Gonnet, G. H., and Benner, S. A. (1997) *Protein Engineering* **10,** 307–316.
26. Richardson, J. S., and Richardson, D. C. (1989) *in* Prediction of Protein Structure and the Principles of Protein Conformation (Fasman, G. D., Ed.) New York, NY, 1989.