PROTEINS: Structure, Function, and Genetics 27:450–458 (1997)

A Predicted Consensus Structure for the N-Terminal Fragment of the Heat Shock Protein HSP90 Family

Dietlind L. Gerloff,[1] Fred E. Cohen,[1,2] Chantal Korostensky,[3] Marcel Turcotte,[5] Gaston H. Gonnet,[5] and Steven A. Benner[4,5]*

Departments of [1]Cellular and Molecular Pharmacology, and [2]Pharmaceutical Chemistry, Biochemistry and Biophysics, and Medicine, University of California San Francisco, San Francisco, California
[3]Institute for Scientific Computation, and [4]Department of Chemistry, ETH Zurich, Switzerland
[5]Department of Chemistry, University of Florida, Gainesville, Florida

ABSTRACT **A secondary structure has been predicted for the heat shock protein HSP90 family from an aligned set of homologous protein sequences by using a transparent method in both manual and automated implementation that extracts conformational information from patterns of variation and conservation within the family. No statistically significant sequence similarity relates this family to any protein with known crystal structure. However, the secondary structure prediction, together with the assignment of active site positions and possible biochemical properties, suggest that the fold is similar to that seen in N-terminal domain of DNA gyrase B (the ATPase fragment). Proteins 27:450–458, 1997.** © **1997 Wiley-Liss, Inc.**

Key words: protein structure prediction; prediction contest; protein sequence alignment

INTRODUCTION

An important problem in modern protein chemistry asks the biological chemist to deduce the secondary structure of a protein from sequence information alone (primary structure). Both at the ETH in Zurich[1] and elsewhere,[2–6] much progress toward solution of this problem has come through an analysis of patterns of conservation and variation in the sequences of homologous proteins that is based on rules transparent to the scientist.[7,8] Such an analysis is especially powerful when it is aided by detailed models of divergent evolution.[9,10] Predictions made using this approach are "consensus" models for conformation of a protein family, and assume that proteins related by common ancestry have similar conformations.[11] To date, some two dozen bona fide predictions, those made and announced before an experimental structure is known, have been made using these methods (reviewed in ref. 8). Many of these have been rather accurate.[8]

In most cases where successful bona fide secondary structure predictions have been made, expert biochemists or molecular modelers have manually contributed to the sequence analysis. This follows the tradition of conformational analysis in organic chemistry generally, where problems have been solved by individual chemists aided both by training and intuition long before computational tools became available that automated chemical expertise.

Manual sequence analysis is tedious, however, difficult to transfer from laboratory to laboratory, and prone to idiosyncrasies. Now that the understanding of protein structure prediction has advanced to the point where high-quality secondary structure predictions by manual analysis are almost routine, it is appropriate to attempt to develop computer tools that reproduce automatically the expertise of the biochemist successful at predicting secondary structures manually. Recently, we have been working to prepare an automated computer tool that generates secondary structure predictions by using the procedure that we have described in manual form in earlier papers.[8] These tools will be useful to make predictions, and they will also serve as tools for learning how to make predictions, since the rules underlying the program are "transparent," unlike those underlying neural networks,[12] for example, which have had success in bona fide secondary structure predictions.[13]

As noted earlier, the testing of automated tools is best when both predictions (against protein families with unknown secondary structure) and retrodictions (against structures already known in the database) are combined. The submission of yeast heat shock protein HSP82, a member of the HSP90 family, as a contest entry for Phase 2 of the Critical Assessment of Techniques for Protein Structure Prediction (CASP) project[14] offers an opportunity to present the first comparison of a fully automated secondary structure prediction tool based on a transparent design (as opposed to, for example, a neural

*Correspondence to: Dr. Steven A. Benner, Department of Chemistry, University of Florida, Gainesville, FL 32611-7200. E-mail: Benner@chem.ufl.edu *or* Department of Chemistry, ETH Zurich, CH-8092 Switzerland. E-mail: Benner@chem.ethz.ch
Received 3 September 1996; accepted 9 September 1996.

network) against a bona fide secondary structure prediction generated by hand.

Further, the setting allows us to use the ability to predict the relative three-dimensional orientation of secondary structure elements toward a putative active site of the protein in the tertiary structure prediction of a medium-sized protein fragment (220 residues).[20]

METHODS

A multiple alignment (Fig. 1) for the heat shock protein HSP90 family was built from sequences extracted from SwissProt[15] (Version 33) and GenBank (National Center for Biotechnology Information, URL www.ncbi.nlm.nih.gov) using the DARWIN system.[16,17] Gaps in the alignment were shifted by using a procedure that identifies misaligned gaps of identical length in nearby regions of the multiple alignment, and shift residues to align the gaps (Korostensky, unpublished). This improves the placement of gaps, but does not guarantee that the globally optimum multiple alignment is found. The improvement in the multiple alignment was followed using the multiple alignment scoring tool of Korostensky and Gonnet.[18]

Surface and interior residues were assigned by automated procedures similar to those described elsewhere,[19] the multiple alignment was parsed into units forming independent secondary structures automatically, and elements of secondary structure were predicted within the parsed segments from patterns of interior and surface assignments, as described elsewhere.[1,8,10,19,20] Many of the automated routines used in this prediction are available to the public on a server accessible via electronic mail at the address cbrg@inf.ethz.ch, or using the World Wide Web with URL http://cbrg.inf.ethz.ch/.

"Parsing strings," consecutive positions that contain Pro, Gly, Ser, Asn, or Asp, were also used to assign breaks in secondary structure. Recent work in these laboratories (T. F. Jenny and M. Turcotte, unpublished) has suggested that these are significantly more reliable than gaps in assigning breaks in secondary structure.

Separately, secondary structure predictions were assigned manually by two of our group (D.L.G., S.A.B.) following rules outlined previously for manual prediction purposes.[8]

SECONDARY STRUCTURE PREDICTION

Figure 1 reports the multiple alignment, surface and interior assignments, parsing assignments, active site assignments, and a secondary structure assignment, all made fully automatically (Auto). The final column are the assignments made by the experts manually, before and after refinement in light of "low resolution" tertiary structure model building.

TERTIARY STRUCTURE ANALYSIS

One use for predicted secondary structural models is to detect long-distance homology between protein families where divergence has been so great that no statistically significant sequence similarities remain, even though the overall fold is similar. Preliminary reports that HSP90 interacts with ATP[21] focused our attention on other ATP binding enzymes, ATPases in particular.[22] The nature and sequence of secondary structural elements and the location of biochemically expected active site functionalities in the HSP90 prediction were compellingly similar to those found in large parts of the experimentally determined N-terminal fragment of DNA gyrase B (ATPase fragment).[23] Table 1 proposes a correlation between the predicted secondary structural elements of the HSP90 family and the experimental elements in gyrase. We are indebted to Dale B. Wigley (University of Oxford) for forwarding us the gyrase coordinates, thereby allowing us to examine the structures more closely.

The gyrase domain adopts a unique fold with a central eight-stranded β sheet, which can be subdivided into two antiparallel sheets with six and two strands joined by a parallel strand-pairing. The ATPase active site is located in the middle of the sheet surface near a long helical segment, which provides residues that bind to the nucleotide, and is covered by a "lid" segment approximately 34 residues long, containing both short α-helical and coil segments. The lid is connected to the core at two short glycine-rich hinge sites. Movement of the lid is likely to account for conformational changes observed upon the binding of ATP to the protein.

In fitting the proposed secondary structure prediction for HS90 to the known structure of DNA gyrase B, several suggestions arose as to how the multiple alignment might be adjusted from this "knowledge-based" perspective. For example, the two structures (predicted for HSP90 and experimental for gyrase) fit somewhat better if the gap placed at positions 126–127 were moved further down in the alignment (see below). Further application of the optimization heuristic found multiple alignments with improved scores if the gap was shifted in this direction.

Likewise, the four residue insertion at positions 178–181, interpreted in the prediction as reflecting introduction of a single turn of a helix, might be shifted down as well. As placed in the automated tool, this gap prevents the tool from identifying a helix found by the "expert." Further application of the optimization heuristic (not shown in Fig. 1) shifted this gap and improved the score of the resulting multiple alignment. These results illustrate that the gap-shifting heuristic is, of course, not an algorithm. It is not guaranteed to find the optimal alignment. However, the combination of the scoring algorithm and the gap-shifting heuristic apparently

Cross reference (Tue Aug 20 05:16:01 1996):
a — (P02829) HS82_YEAST HEAT SHOCK PROTEIN HSP90.
Saccharomyces cerevisiae (baker's yeast).
b — (P15108) HS83_YEAST HEAT SHOCK COGNATE PROTEIN HSC82.
Saccharomyces cerevisiae (baker's yeast).
c — (P46598) HS90_CANAL HEAT SHOCK PROTEIN 90 HOMOLOG.
Candida albicans (yeast).
d — (P41887) HS90_SCHPO HEAT SHOCK PROTEIN 90 HOMOLOG.
Schizosaccharomyces pombe (fission yeast).
e — (P33125) HS82_AJECA HEAT SHOCK PROTEIN 82.
Ajellomyces capsulata (histoplasma capsulatum).
f — (Q04619) HS9B_CHICK HEAT SHOCK COGNATE PROTEIN HSP 90-BETA.
Gallus gallus (chicken).
g — (P33126) HS82_ORYSA HEAT SHOCK PROTEIN 82.
Oryza sativa (rice).
h — (Q03930) HS81_ARATH HEAT SHOCK PROTEIN 81 (HSP81-1).
Arabidopsis thaliana (mouse-ear cress).
i — (P36181) HS80_LYCES HEAT SHOCK COGNATE PROTEIN 80.
Lycopersicon esculentum (tomato).
j — (Q08277) HS82_MAIZE HEAT SHOCK PROTEIN 82.
Zea mays (maize).
k — (P04809) HS83_DROPS HEAT SHOCK PROTEIN 83 (HSP 82) (FRAGMENT).
Drosophila pseudoobscura (fruit fly).
l — (P46633) HS9A_CRIGR HEAT SHOCK PROTEIN HSP 90-ALPHA (HSP 86).
Cricetulus griseus (chinese hamster).
m — (P07900) HS9A_HUMAN HEAT SHOCK PROTEIN HSP 90-ALPHA (HSP 86).
Homo sapiens (human).
n — (P02828) HS83_DROME HEAT SHOCK PROTEIN 83 (HSP 82).
Drosophila melanogaster (fruit fly).
o — (P08238) HS9B_HUMAN HEAT SHOCK PROTEIN HSP 90-BETA (HSP 84).
(HSP 90). Homo sapiens (human).
p — (P11501) HS9A_CHICK HEAT SHOCK PROTEIN HSP 90-ALPHA.
Gallus gallus (chicken).
q — (P06660) HS85_TRYCR HEAT SHOCK LIKE 85 KD PROTEIN.
Trypanosoma cruzi.
r — (P24724) HS90_THEPA HEAT SHOCK PROTEIN 90 (HSP90).
Theileria parva.
s — (P27741) HS83_LEIAM HEAT SHOCK PROTEIN 83 (HSP 83).
Leishmania amazonensis.
t — (P12861) HS83_TRYBB HEAT SHOCK PROTEIN 83.
Trypanosoma brucei brucei.
u — (P36183) ENPL_HORVU ENDOPLASMIN HOMOLOG PRECURSOR.
(GRP94 HOMOLOG). Hordeum vulgare (barley).
v — (P35016) ENPL_CATRO ENDOPLASMIN HOMOLOG PRECURSOR.
(GRP94 HOMOLOG). Catharanthus roseus (rosy periwin).
w — (P08110) ENPL_CHICK ENDOPLASMIN PRECURSOR (TRANSFERRIN-BINDING PROTEIN).
Gallus gallus (chicken).
x — (P41148) ENPL_CANFA ENDOPLASMIN PRECURSOR (94 KD GLUCOSE-REGULATED PROTEIN) (GRP94).
Canis familiaris (dog).
y — (P14625) ENPL_HUMAN ENDOPLASMIN PRECURSOR (94 KD GLUCOSE-REGULATED PROTEIN) (GRP94).
Homo sapiens (human).
z — (P08113; P11427) ENPL_MOUSE ENDOPLASMIN PRECURSOR (94 KD GLUCOSE-REGULATED PROTEIN) (GRP94).
Mus musculus (mouse).
A — (P44516) HTPG_HAEIN HEAT SHOCK PROTEIN HTPG.
Haemophilus influenzae.
B — (P10413) HTPG_ECOLI HEAT SHOCK PROTEIN HTPG.
Escherichia coli.
C — (P46208) HTPG_BACSU HEAT SHOCK PROTEIN HTPG HOMOLOG.
Bacillus subtilis.
D — (Gb_ro:S45392/PID:g256089) HEAT SHOCK PROTEIN 90. Rattus sp. brain (rat).
E — (Gb_pl:Phnhsp83a/PID:g169296) HEAT SHOCK PROTEIN 83 (HSP83) GENE.
Pharbitis nil (strain violet).

Pos	C	AB	decba	r	tqs	jEhig	nkpmlDof	wzyx	uv	SIA	Auto	Manual		3D ref.
71	–	––	––A–M	–	–––	–––––	EEEEEEED	EEEE	NS	S				
72	–	––	––K–A	–	–––	EED-E	EEEEEEEE	KKKK	SD	s				
73	–	––	––V–S	–	TTT	TAA–T	AAVVVVVV	SSSS	AA	s				
74	–	EE	EEEEE	E	EEE	EEEEE	EEEEEEEE	EEEE	EE	s		e		
75	–	TT	TTTTT	V	TTT	TTTTT	TTTTTTTT	KKKK	KK	·		E		
76	–	RR	FFHFF	Y	FFF	FFFFF	FFFFFFFF	FFFF	FF	i	e	E	e	
77	–	GG	KEEEE	A	AAA	AAAAA	AAAAAAAA	AAAA	EE	s	e	E	e	
78	F	FF	FFFFF	F	FFF	FFFFF	FFFFFFFF	FFFF	FF	i	e	E	e	
79	K	QQ	DQTQQ	N	QQQ	QQQQQ	QQQQQQQQ	QQQQ	QQ	s	e	E	e	
80	A	SS	WAAAA	A	AAA	AAAAA	AAAAAAAA	AAAA	AA	i	e	E	e	
81	E	EE	EEEEE	D	EEE	EEEEE	EEEEEEEE	EEEE	EE	s	e	E	e	
82	S	VV	IIIII	I	III	IIIII	IIIIIIII	VVVV	VV	i	e	e	e	
83	K	KK	SSSTT	S	NNN	NNNNN	AAAAAAAA	NNNN	SS		S	e		
84	R	QQ	QQQQQ	Q	QQQ	QQQQQ	QQQQQQQQ	RRRR	RR		i	e	h	h

Pos	C	AB	decba	r	tqs	jEhig	nkpmlDof	wzyx	uv	SIA	Auto	Manual		3D ref.
85	L	LL	LLLLL	L	LLL	LLLLL	LLLLLLLL	MMMM	LL		I		H	h
86	L	LL	MLMMM	L	MMM	LLLLL	MMMMMMMM	MMMM	MM		I		H	h
87	D	QH	SSSSS	S	SSS	SSSSS	SSSSSSSS	KKKK	DD		s	H	H	H
88	M	LL	LLLLL	L	LLL	LLLLL	LLLLLLLL	LLLL	II		I	H	H	H
89	M	MM	IIIII	I	III	IIIII	IIIIIIII	IIII	II		I	H	H	H
90	I	II	IIIII	I	III	IIIII	IIIIIIII	IIII	II		i	H	H	H
91	N	HH	NNNNN	N	NNN	NNNNN	NNNNNNNN	NNNN	NN		s	H	H	H
92	S	SS	TTTTT	A	TTT	TTTTT	TTTTTTTT	SSSS	SS		i	H	H	H
93	I	LL	VVVVV	F	FFF	FFFFF	FFFFFFFF	LLLL	LL		I	H	H	H
94	Y	YY	YYYYY	Y	YYY	YYYYY	YYYYYYYY	YYYY	YY		i	H	H	H
95	T	SS	SSSSS	S	SSS	SSSSS	SSSSSSSS	KKKK	SS		s	H	H	H
96	Q	NN	NNNNN	N	NNN	NNNNN	NNNNNNNN	NNNN	NN		s		H	
97	K	KK	KKKKK	K	KKK	KKKKK	KKKKKKKK	KKKK	KK		s	A	a	h
98	E	EE	EEEEE	E	EEE	EEEEE	EEEEEEEE	EEEE	DD		s		a	H
99	I	II	IIIII	I	III	IIIII	IIIIIIII	IIII	II		i	A	a	H
100	F	FF	FFFFF	F	FFF	FFFFF	FFFFFFFF	FFFF	FF		i	A	a	H
101	L	LL	LLLLL	L	LLL	LLLLL	LLLLLLLL	LLLL	LL		i	A	a	H
102	R	RR	RRRRR	R	RRR	RRRRR	RRRRRRRR	RRRR	RR		s	A	a	H
103	E	EE	EEEEE	E	EED	EEEEE	EEEEEEEE	EEEE	EE		s		a	H
104	L	LL	LLLLL	L	LLV	LLLLL	LLLLLLLL	LLLL	LL		I	A	a	H
105	I	II	IIIII	I	III	IIIII	IIIIIIII	IIII	II		i	A	a	H
106	S	SS	SSSSS	S	SSS	SSSSS	SSSSSSSS	SSSS	SS	P	.	A	a	H
107	N	NN	NNNNN	N	NNN	NNNNN	NNNNNNNN	NNNN	NN	P	s		a	H
108	S	AA	AFAAA	A	SSA	AASSS	AASSSAAA	AAAA	AA	P	i	A	a	H
109	S	SS	SSSSS	S	SSS	SSSSS	SSSSSSSS	SSSS	SS	P	.	A	a	H
110	D	DD	DDDDD	D	DDD	DDDDD	DDDDDDDD	DDDD	DD	P	s	A	a	H
111	A	AA	AAAAA	A	AAA	AAAAA	AAAAAAAA	AAAA	AA		i		a	h
112	I	AA	LLLLL	L	CCC	LLLLL	LLLLLLLL	LLLL	LL		I		a	h
113	D	DD	DDDDD	E	DDD	DDDDD	DDDDDDDD	DDDD	DD		s	A	a	
114	K	KK	KKKKK	K	KKK	KKKKK	KKKKKKKK	KKKK	KK		s		e	
115	I	LL	IIIII	I	III	IIIII	IIIIIIIII	IIII	II		I	E	E	e
116	Y	RR	RRRRR	R	RRR	RRRRR	RRRRRRRR	RRRR	RR		s	E	E	E
117	Y	FF	YYYYY	Y	YYY	FFFFF	YYYYYYYY	LLLL	FF		I	E	E	E
118	K	KR	QKQQK	E	QQQ	EEEEE	EEEEEEEE	IIII	LL		S	E	E	E
119	A	AA	SAAAS	A	SSS	SSSSS	SSSTSSSS	SSSS	AA		s	E	E	E
120	L	LL	LLLLL	I	LLL	LLLLL	LLLLLLLL	LLLL	LL		I	E	E	E
121	T	SS	SSSSS	K	TTT	TTTTT	TTTTTTTT	TTTT	TT	P	.			e
122	D	NN	DDDDD	D	NND	DDDDD	DDDDDDDD	DDDD	DD	P	s			
123	D	PP	PPPPP	P	QQP	KKKKK	PPPPPPPP	EEEE	KK	P	S			
124	A	AD	HSSKK	K	SAS	SSSSS	SSSSSSSS	NNNN	EE	P	S			
125	L	LL	AKQQQ	Q	VVV	NKKKK	KKKKKKKK	AAAA	VI		S			
126	_	__	_____	_	___	_____	_____	____	ML	P	.			e*
127	_	__	_____	_	___	_____	_____	____	GG	P	i			e*
128	T	YY	LLLLL	I	LLL	VLLLL	LLLLLLLL	LLLL	EE		i			e*
129	F	EE	DEEEE	E	GGG	NDDDD	DDDDDDDD	AASA	GG	P	S			e*
130	D	GG	ASSTT	D	DDD	AAGGA	SSSSSSST	GGGG	DD	P	S			e*
131	K	DD	EDEEE	Q	EEA	QQQQQ	GGGGGGGG	NNNN	TT	P	S			
132	D	GG	KKPPP	P	PST	PPPPP	KKKKKKKK	EEEE	AA	P	S			
133	S	DE	DDEDD	D	HHR	EEEEE	EEDEEEED	EEEE	KK	P	S		E	
134	Y	LL	LLLLL	Y	LLL	LLLLL	LLLLLLLL	LLLL	LL		I	E	E	E
135	Y	RR	FRFFF	Y	RRC	FFFFF	YYKHHKKK	TTTT	EE		S	E	E	E
136	I	VV	IIIII	I	IIV	IIIII	IIIIIIII	VVVV	II		I	E	E	E
137	K	RR	RDRRR	R	RRR	RRRHH	KKNNNDDD	KKKK	QQ		S	E	E	E
138	V	VV	IIIII	L	VVV	LLLII	LLLLIIII	IIII	II		I	E	E	E
139	A	SS	TTITT	Y	IVV	VVVIV	IIIIIIIV	KKKK	KK		i	E	E	E
140	A	FF	PPPPP	A	PPP	PPPPP	PPPPPPPP	CCCC	LL	P	i		e	
141	D	DD	DDQKK	D	DDD	DDDDD	NNNNNNNN	DDDD	DD	P	s			
142	K	AK	KKKPP	K	RKK	KKKKK	KKKKKPPP	KKKK	KK	P	S			
143	D	DD	EEDEE	N	VAE	ATSAA	TTHQQQQR	EEEE	EE	P	S			
144	A	KK	NNQEQ	N	NNN	SNNNS	AADDDEED	KKKK	NK	P	S			
145	R	GR	KKKKK	N	KKK	KKKNN	GGRRRARP	NNNN	KK	P	S			
146	T	TT	ITVVV	T	TTT	TTTTT	TTTTTTTT	MLLL	II		I	E	E	E
147	L	IL	LLLLL	L	LLL	LLLLL	LLLLLLLL	LLLL	LL		I	E	E	E
148	T	TT	TTEEE	T	TTT	SSSTS	TTTTTTTT	HHHH	SS		s	E	E	E
149	I	II	IIIII	I	VVV	IIIII	IIIIILLL	VVVV	II		I	E	E	E
150	S	SS	RRRRR	E	EEE	IIIII	IIVVVVVL	TTTT	RR		S	E	E	E
151	D	DD	DDDDD	D	DDD	DDDDD	DDDDDDDD	DDDD	DD	P	S	A	E	E
152	T	NN	TTSSS	S	STN	SSSSS	TTTTTTTT	TTTT	RR	P	s		E	e
153	G	GG	GGGGG	G	GGG	GGGGG	GGGGGGGG	GGGG	GG	P	i			
154	I	IV	IIIII	I	III	VVIII	IIIIIIII	IVVV	VI	P	I		e	
155	G	GG	GGGGG	G	GGG	GGGGG	GGGGGGGG	GGGG	GG	P	i		e	
156	M	MM	MMMMM	M	MMM	MMMMM	MMMMMMMM	MMMM	MM		I		e	
157	T	TT	TTTTT	T	TTT	TATTT	TTTTTTTT	TTTT	TT		I		e	
158	K	RR	KKKKK	K	KKK	KKKKK	KKKKKKKK	KRRR	KK		s			
159	D	ED	NAAAA	A	AAA	SAAAS	SSAAAAAA	EEEE	EE		s	H	h	H
160	E	QE	DDDEE	D	DED	DDDDD	DDDDDDDD	EEEE	DD		S	H	h	H

Pos	C	AB	decba	r	tqs	jEhig	nkpmlDof	wzyx	uv	SIA	Auto	Manual		3D ref.
161	L	VV	LLLLL	L	LLL	LLLLL	LLLLLLLL	LLLL	LL		I	H	h	H
162	E	II	IVVII	V	VVV	VVVVV	VVVIIIIV	IVVV	II		i	H	h	H
163	Q	DD	NNNNN	N	NNN	NNNNN	NNNNNNNN	KKKK	KK		s	H	h	H
164	H	HH	NNNNN	N	NNN	NNNNN	NNNNNNNN	NNNN	NN		s	H	h	H
165	L	LL	LLLLL	L	LLL	LLLLL	LLLLLLLL	LLLL	LL		i	H	h	h
166	G	GG	GGGGG	G	GGG	GGGGG	GGGGGGGG	GGGG	GG		i	H	h	
167	T	TT	VTTTT	T	TTT	TTTTT	TTTTTTTT	TTTT	TT		I			
168	I	II	IIIII	I	III	IIIII	IIIIIIII	IIII	II		i			
169	A	AA	AAAAA	A	AAA	AAAAA	AAAAAAAA	AAAA	AA		i			
170	K	KK	KRKKK	K	RRR	RRRRR	KKKKKKKK	KKKK	KK		S			
171	S	SS	SSSSS	S	SSS	SSSSS	SSSSSSSS	SSSS	SS		.	A		
172	G	GG	GGGGG	G	GGG	GGGGG	GGGGGGGG	GGGG	GG		i	A		
173	S	TT	TTTTT	T	TTT	TTTTT	TTTTTTTT	TTTT	TT		i		H	h
174	L	KK	KKKKK	R	KKK	KKKKK	KKKKKKKK	SSSS	SS		s		H	H
175	A	ES	QQSAA	A	SAA	EEEEE	AAAAAAAA	EEEE	AA		S		H	H
176	F	FF	FFFFF	F	FFF	FFFFF	FFFFFFFF	FFFF	FF		i		H	H
177	K	LL	MMMMM	M	MMM	MMMMM	MMMMMMMM	LLLL	VV		i		H	H
178	K	TE	_____	_	___	_____	_____	NNNN	__	P	s		h	
179	_	__	_____	_	___	_____	_____	KKKK	__	P	s		h	
180	_	__	_____	_	___	_____	_____	MMMM	__	P	i		h	
181	_	__	_____	_	___	_____	_____	TTTT	__	P			h	
182	E	AS	EEEEE	E	EEE	EEEEE	EEEEEEEE	EEEE	EE		s		H	H
183	N	LL	AAAAA	A	AAA	AAAAA	AAAAAAAA	MAAA	KK		.		H	H
184	E	GG	ALLLL	L	LLL	LLLLL	LLLLLLLL	QQQQ	MM		i		H	H
185	L	QS	ATSSS	Q	EEE	AQQAA	QQQQQQQQ	DEEE	QQ		s		h	h
186	K	DD	SAAAA	A	AAA	AAAAA	AAAAAAAA	DDDD	TT		s			
187	D	QQ	GGGGG	G	GGG	GGGGG	GGGGGGGG	SGGG	GS		s			
188	G	AA	AAAAA	S	GGA	AAAAA	AAAAAAAA	QQQQ	GG		s			
189	_	KK	_____	_	___	T____	_____	SSSS	__	P	s			
190	_	ND	DDDDD	D	DDD	DDDDD	DDDDDDDD	TTTT	DD		s			
191	H	SS	IIVVV	M	MMM	VVVVV	IIIIIIII	SSSS	LL		i			
192	D	QQ	SSSSS	S	SSS	SSSSS	SSSSSSSS	EEEE	NN		s			
193	I	LL	MMMMM	M	MMM	MMMMM	MMMMMMMM	LLLL	LL		I			
194	I	II	IIIII	I	III	IIIII	IIIIIIII	IIII	II		i		a	
195	G	GG	GGGGG	G	GGG	GGGGG	GGGGGGGG	GGGG	GG		i		a	
196	Q	QQ	QQQQQ	Q	QQQ	QQQQQ	QQQQQQQQ	QQQQ	QQ		.		a	
197	F	FF	FFFFF	F	FFF	FFFFF	FFFFFFFF	FFFF	FF		i		a	
198	G	GG	GGGGG	G	GGG	GGGGG	GGGGGGGG	GGGG	GG	P	i		a	
199	V	VV	VVVVV	V	VVV	VVVVV	VVVVVVVV	VVVV	VV	P	i		a	
200	G	GG	GGGGG	G	GGG	GGGGG	GGGGGGGG	GGGG	GG	P	i		a	H
201	F	FF	FFFFF	F	FFF	FFFFF	FFSFFFFF	FFFF	FF		i		a	H
202	Y	YY	YYYYY	Y	YYY	YYYYY	YYYYYYYY	YYYY	YY		i		a	H
203	A	SS	SSSSS	S	SSS	SSSSS	SSSSTSSS	SSSS	SS		s		e	H
204	A	AA	AALLL	A	AAA	AAAAA	AAAAAAAA	AAAA	VV		I	E	e	H
205	F	FF	YYFFF	Y	YYY	YYYYY	YYYYYYYY	FFFF	YY		I	E	e	H
206	M	II	LLLLL	L	LLL	LLLLL	LLLLLLLL	LLLL	LL		I	E	e	h
207	V	VV	VVVVV	V	VVV	VVVVV	VIVVVVVV	VVVV	VV		I	E	e	h
208	A	AA	AAAAA	A	AAA	AAAAA	AAAAAAAA	AAAA	AP	P	i		e	
209	D	DD	DDDDD	D	DDD	DEEEE	DDEEEEEE	DDDD	DD	P	s			
210	V	KK	KKHRR	K	RRR	RKKKR	KRKKKKKK	RKKK	YY		S	E	E	E
211	V	VV	VVVVV	V	VVV	VVVVV	VVVVVVVV	VVVV	VV		i	E	E	E
212	T	TT	QTQQQ	T	TTT	MIVVV	TTTTTVVV	IIII	EE		s	E	E	E
213	V	VV	VVVVV	V	VVV	VVVVV	VVVVVVVV	VVVV	VV		i	E	E	E
214	I	KR	VIIII	V	VVT	TTTTT	TTIIIIII	TTTT	VI		s	E	E	E
215	S	TT	SSSSS	S	SSS	TTTTT	SSTTTTRT	SSSS	SS		s		e	E
216	K	RR	KKKKK	K	KKK	KKKKK	KKKKKKKK	KKKK	KK		s			
217	A	AA	_____	_	___	_____	_____	_____	__	P	i			
218	L	AA	_____	_	___	_____	_____	_____	__	P	i			
219	G	GG	_____	_	___	_____	_____	_____	__	P	i			
220	_	EE	HSHNS	N	NNN	HHHHH	NNHHHHHH	HHHH	HH		s			
221	_	EK	NNNNN	N	NNN	NNNNN	NNNNNNNN	NNNN	NN	P				
222	S	AP	DDDED	A	EDS	DDDDD	DDDDDDDD	NNNN	DD	P	S			
223	E	DE	DDDDD	D	DDD	DDDDD	DDDDDDDD	DDDD	DD	P	s			
224	E	KN	EEEEE	D	DEE	EEEEE	EEEEEEEE	TTTT	KK		S			e
225	A	AG	QQQQQ	Q	AAV	QQQQQ	QQQQQQQQ	QQQQ	QQ		.		e	E
226	Y	VV	YYYYY	Y	YYY	YYYYY	YYYYYYYY	HHHH	YY		I		e	E
227	K	LF	IIVII	V	TTV	VIVVV	VVAAAAAA	IIII	VI		i		e	E
228	W	WW	WWWWW	W	WWW	WWWWW	WWWWWWWW	WWWW	WW		i		e	E
229	E	EE	EEEEE	E	EEE	EEEEE	EEEEEEEE	EEEE	EE		s	A	A	E
230	S	SS	SSSSS	S	SSS	SSSSS	SSSSSSSS	SSSS	SS		.	A	A	
231	A	AA	SNNNN	T	SSS	QQQQQ	SSSSSSSS	DDDD	KK		S			
232	G	GG	AAAAA	A	AAA	AAAAA	AAAAAAAA	SSSS	AA		i			
233	A	EE	GGGGG	S	GGG	GGGGG	GGGGGGGG	NNNN	DD	P	s			
234	D	GG	GGGGG	G	GGG	GGGGG	GGGGGGGG	_____	GG	P	s			
235	G	EE	STKSS	H	TTT	SSSSS	SSSSSSSS	EEEE	SA	P	S		e	
236	Y	YY	FFFFF	F	FFF	FFFFF	FFFFFFFF	FFFF	FF		I	E	E	E

Pos	C	AB	decba	r	tqs	jEhig	nkpmlDof	wzyx	uv	SIA	Auto	Manual		3D ref.
237	T	ST	TKTTT	T	TTT	TTTTT	TTTTTTTT	SSSS	AA		s	E	E	E
238	I	VV	VVVVV	V	VVI	VVVVV	VVVVVVVV	VVVV	II		I	E	E	E
239	E	AA	TTTTT	K	TTT	TTTTT	_____	IIII	SS	P	s		E	E
240	P	DD	LQLLL	K	SPS	HRRRR	RKRRRRRR	DAAA	EE	P	S		E	E
241	C	II	DDDDD	D	TTA	DDDDD	AALTTAAT	DDDD	DD	P	s		e	E
242	E	ET	TDEEE	D	PPP	TVVTT	DDDDDDDD	PPPP	TV	P	S			
243	K	KK	DDTVV	S	DDE	TDDSS	NNNTTHHH	RRRR	WW	P	S			
244	D	KE	GGNNN	H	CCS	GGGGG	SSGGGGGG	GGGG	NN	P	S			
245	S	SD	PREEE	E	DDD	EEEEE	EEEEEEEE	NNNN	EE	P	S			
246	V	RR	RARRR	P	___	QQPNQ	PPPPPPPP	TTTT	PP	P	S			
247	_	__	LILII	L	LLM	LLLLL	LLLMMIII	LLLL	LL	P	i			
248	_	__	LGGGG	K	KKK	GGGGG	GGGGGGGG	GGGG	GG	P	s			
249	_	__	RRRRR	R	RRL	RRRRR	RRRRRRMR	RRRR	RR	P	s			
250	G	GG	GGGGG	G	GGP	GGGGG	GGGGGGGG	GGGG	GG	P	.		e	E
251	T	TT	TTTTT	T	TTA	TTTTT	TTTTTTTT	TTTT	TT		I	E	E	E
252	D	DE	EKMVI	R	RRR	KKKKK	KKKKKKKK	TTTT	EE		S	E	E	E
253	I	VI	IMLLL	L	III	IIIMI	IIVVVVVV	IIII	II		I	E	E	E
254	I	IT	RIRRR	I	VVT	TTTVT	VVIIIIII	TTTT	KR		s	E	E	E
255	L	LL	LLLLL	L	LLL	LLLLL	LLLLLLLL	LLLL	LL		i	E	E	E
256	K	HH	FHFFF	H	HHH	FFFYY	YYHHHHHY	VVVV	HH		s	E	E	E
257	I	LL	MLLLL	L	LLL	LLLLL	IILLLLLL	LLLL	LL		I	E	E	e
258	K	RR	KKKKK	K	KKK	KKKKK	KKKKKKKK	KKKK	RR		s	E	e	
259	E	EE	EDEDD	E	EEE	DEDED	EEEEEEEE	EEEE	DD		S	E		
260	N	DG	DEDDD	D	DDD	DDDDD	DDDDDDDD	EEEE	EE		s			
261	T	EE	QQQQQ	Q	QQQ	QQQQQ	QQQQQQQQ	AAAA	AA		i			
262	E	KD	LTLLL	T	QQL	LLLLL	TTTTTTTT	SSSS	KQ		S			
263	D	EE	QEEEE	E	EEE	EEEEE	DDEEEEEE	DDDD	EE		S			
264	D	__	_____	_	___	_____	_____	____	__	P	s			
265	S	__	_____	_	___	_____	_____	____	__	P	.			
266	Y	__	_____	_	___	_____	_____	____	__	P	i			
267	D	__	_____	_	___	_____	_____	____	__	P	s			
268	E	__	_____	_	___	_____	_____	____	__	P	s			
269	F	FF	YYYYY	Y	YYY	YYYYY	YYYYYYYY	YYYY	YY		I		e	h
270	L	LL	LLLLL	L	LLL	LLLLL	LLLLMLLL	LLLL	LL		I		e	h
271	E	ND	ENEEE	E	EEE	EEEEE	EEEEEEEE	EEEE	ED		S			h
272	E	ED	EEEEE	E	EEA	EEEEE	EEEEEEEE	LLLL	EE		S	H		h
273	Y	WW	KSKKK	R	RRR	RRRRR	SSRRRRRR	DDDD	GF		S	H	h	H
274	R	RR	TKRRR	R	RRR	RRRRR	KKRRRRRR	TTTT	KK		S	H	H	H
275	L	LV	IIIII	L	LLL	LILLL	IIIIIVVV	VIII	LL		I	H	H	H
276	K	RR	KKKKK	K	KKK	KKKKK	KKKKKKKK	KKKK	KK		s	H	H	H
277	A	ES	DEEEE	E	DDE	DDDDD	EEEEEEEE	NNNN	DE		S	H	H	H
278	I	II	TVVVV	L	LLL	LLLLL	IIIIIVVV	LLLL	LL		I	H	H	H
279	I	II	VVVII	V	III	VVVII	VVVVVVVV	VVVV	VV		I	H	H	H
280	K	GS	KKKKK	K	KKK	KKKKK	NNKKKKKK	KRKK	KK		s	H	H	H
281	K	KK	KKKRR	K	KKK	KKKKK	KKKKKKKK	KKKK	KR		s	H	H	H
282	Y	YY	HQHHH	H	HHH	HHHHH	HHHHHHHH	YYYY	YY		I	H	H	H
283	S	SS	SSSSS	S	SSS	SSSSS	SSSSSSSS	SSSS	SS		.	H	H	H
284	D	DD	EEEEE	E	EEE	EEEEE	QQQQQQQQ	QQQQ	EE		s	H	H	H
285	F	HH	FFFFF	F	FFF	FFFFF	FFFFFFFF	FFFF	FF		I	H	H	H
286	I	II	IIVVV	I	III	IIIII	IIIIIIII	IIII	II		I	H	H	H
287	R	GA	SFAAA	S	GGG	SSSSS	GGGGGGGG	NNNN	NN	P	s		h	
288	Y	LL	YYYYY	F	YYY	YYYYY	YYYYYYYY	FFFF	FF	P	I			
289	P	PP	PPPPP	P	DDD	PPPPP	PPPPPPPP	PPPP	PP		s		e	
290	I	VV	IIIII	I	III	IIIII	IIIIIIII	IIII	II		I	E	E	E
291	K	EE	QYQQQ	S	EEE	YYYSS	KKRTTTTT	YYYY	YY		S	E	E	E
292	M	MI	LLLLL	L	LLL	LLLLL	LLLLLLLL	VVVV	LL		I	E	E	E
293	D	LE	VHVLV	S	MMM	WWWWW	LLFFFYYY	WWWW	WW		s	E	E	E
294	T	TK	VVVVV	V	VVV	TTIVT	VVVVVLLV	SSSS	AA		.			E
295	T	KR	TLTTT	E	EEE	EEEEE	EEEEEEEE	SSSS	TS		s			e
296	I	EE	RKKKK	K	NKK	KKKKK	KKKKKKKK	KKKK	KK		S			
297	N	YE	EEEEE	T	TAT	TTTTT	EEEEEEEE	TTTT	EE		S			
298	K	DK	VNVVV	Q	TTT	TTTIT	RRRRRRRR	EEEE	VV		S			
299	P	D_	EEEEE	E	EEE	EEEEE	EEDDDEEE	TTTT	DE		S			e
300	K	E_	KKKKK	T	KKK	KKKKK	KKKKKKKK	VVVV	VV		.			e
301	E	__	EEEEE	E	EEE	EEEEE	EEEEEEEE	EEEE	EE		s			e
302	G	__	VVVVV	V	VVV	IIIII	VVVVVIIV	EEEE	VV		.			e
303	S	__	PPPPP	T	TTT	SSSSS	SSSSSSSS	PPPP	PP	P	s			
304	E	__	EDEII	D	DDD	DDDDD	DDDDDDDD	VLMM	AA	P	s			

Fig. 1. Residue-by-residue consensus secondary structure prediction for the heat shock protein HSP90 family. The SIA records assignments of positions to the surface (S, s), interior (I, i), or near the "active site" (A, a). Automated assignments are given, with the output generated by DARWIN. Services of DARWIN are available by server to the user on the Web (URL http://cbrg.inf.ethz.ch/). Secondary structure is indicated by E (strong strand assignment), e (weak strand assignment), H (strong helix assignment), and h (weak helix assignment). Sequences, designated using single letters, are from the SwissProt database and Genbank, as summarized below. Sequence "a" is the target sequence. The column marked "Auto" contains output from the fully automated secondary structure prediction tool. The column marked "Manual" contains assignments from semimanual analysis of the same data. The column marked "3D refined" contains secondary structure assignments made after comparison with the experimentally determined structure of the N-terminal domain of DNA gyrase B, where an asterisk (*) indicates where a shift in the alignment is required.

Table I. Refined secondary structure assignments for the heat-shock protein 90 family

Unit	Alignment Positions	Comments	Approximately Corresponding Region in E. coli DNA Gyrase B (ref. 23)
strand 0	76–82	prediction weakened by model; non-core, possibly a strand in dimeric form	coil/strand (9–14)
parse	83	weak parse	
helix A	84–95	relatively buried	helix (17–24)
parse	96–97	surface parse	
helix B	98–112	possibly 3/10 at C-end	helix (35–55)
parse	113–114	active site	
strand 1	115–121	amphiphilic	strand (59–65)
parse	122–125	DPS tripeptide parse, exposed	
strand I1	126–130	rearranged alignment, exposed weak prediction, edge strand?	—
parse	131–133	DGD tripeptide, PD dipeptide parses, exposed	
strand I2	134–139	amphiphilic	—
parse	140–145	PDP tripeptide parse, exposed	
strand 2	146–152	amphiphilic	strand (69–74)
parse	153–158	DxGxG (151–155) possible hinge, near active site	DxGxG (73–77)
helix C	159–165	short, oriented towards active site	[insufficient]
active site	166–172	conserved S at 171	[correspondence]
helix D	173–185	10 residues in target sequence; possible break in the middle	[to match region]
parse	186–190	GGD tripeptide and gap	
coil/parse	191–199	note strand possibility in sequences a-t, E, D (191–194); GxxGxG (195–200) possible hinge	GxxGxG (114–119)
helix E	200–207	highly conserved hydrophobic segment; prediction from model	helix (119–126)
parse	208–209	weak parse	
strand 3	210–215	amphiphilic, but weakly	strand (131–136)
parse	216–223	NNDD tetrapeptide and gaps	
strand 4	224–229	buried, oriented towards a separate functional site?	strand (140–146)
parse	230–235	SNAGGS hexapeptide and gap	
strand 5	236–241	amphiphilic/exposed	strand (154–160)
parse	242–249	strong polypeptide parses, gaps	
strand 6	250–257	amphiphilic	strand (164–170)
parse	258–268	surface parse and insertion in sequence C	
helix F	269–286	amphiphilic; N-terminus overrides weak strand prediction and possible surface parse (271–274)	helix (184–200)
parse	287–289	GxP parse	
strand 7	290–295	amphiphilic, but weakly	strand (202–207)
parse	296–298	surface parse	
strand 8/coil	299–302	possibly coil, predicted from model only	strand (215–219)

reevaluate the multiple alignment much as it is done by eye, given enough computation time.

The fitting also assisted in assigning secondary structure near the active site, where patterns of variation and conservation that normally might otherwise indicate particular types of secondary structure are obscured by patterns that reflect catalytic or binding function, and suggested that some of the predicted secondary structural elements should be reevaluated. For example, a strand is predicted in a region (positions 204–207) that aligns against a short internal helix in gyrase. Internal helices are well known for being difficult to predict using the transparent methods applied here.[24] The automated program notices that a helix might be assigned to positions 207–212, but rejects it in favor of two strand assignments at positions 204–207 and 210–214. Most "experts" would prefer the two β strands as well. Inspection of the gyrase multiple alignment (data not shown) suggests that both the manual and automated procedures would probably have misassigned this segment of conserved hydrophobic positions in gyrase as well. Thus, in a "knowledge-based" environment, one might find support in this analysis for distant homology even if this particular secondary structure unit were predicted incorrectly.

The first strand in the predicted HSP90 model forms an extended coil at the N terminus of the gyrase structure; the strand prediction is weakened by the comparison, as this segment is presumably noncore. A region at the putative active site between positions 98 and 110 is predicted to be a long helix contributing amino acid side chains that serve as ligands to Mg. To accommodate the predicted inser-

tion in the HS90 proteins over positions 123–145, an additional short strand segment is predicted to pair with the strong amphiphilic pattern at 134–139 (see below). The remainder of the secondary structure prediction (excluding positions 158–194, discussed below) fits well with the experimentally determined secondary structural elements in gyrase up to the final eight residues (positions 297–304). In the gyrase structure, this final segment forms an exposed edge strand leading into the following domain, and this may also be the case with HSP90. We list this as a possible assignment in Table 1, even though the assignment would not be made from the multiple sequence alignment alone.

The secondary structure prediction derived from an analysis that incorporates information from the gyrase structure is shown in Table 1. This output represents a combination of de novo (or ab initio) approaches and "knowledge-based" modeling akin to threading (fold recognition).[25]

If our proposed fitting were correct, there would be three regions where the folds of the heat shock protein 90 and the N-terminal domain of gyrase B might differ. Most important, we propose an additional antiparallel hairpin structure between strands 1 and 2 in the gyrase structure. The apparently strong exposure to solvent of the weakly predicted strand at 126–130 (in the rearranged alignment) suggests that this segment would form the edge of a β sheet. Hence, while the exact location of the inserted hairpin remains speculative, it is not likely to be part of the main sheet in the domain.

Next, the sequence of the "lid" segment of DNA gyrase B (not shown, residues 36–113 in the gyrase from *E. coli*)[23] is not sufficiently similar to any segment in the corresponding region of HSP90 to permit a speculative alignment in this region. While the segment is still predicted to contain helical and coil segments and to form a "lid" anchored at the glycine-rich sequence motifs DXGXG (alignment positions 151–155) and GXXGXG (195–200), the tertiary structure must be remodeled ab initio to obtain a more precise definition of conformation. As a biochemical clue for the modeling, the conserved serine at position 171 might be the site of the autophosphorylation events observed by Csermely and colleagues.[26] As an alternative explanation for the poor correspondence in the "lid" segment, ATP might not be bound in the exact same conformation by the two proteins. Finally, the N-terminal 25 residues (corresponding to alignment positions 71–95 for the heat shock proteins) are not part of the core in our template. Thus, the relative orientation of the predicted helix at positions 85–95 and the extended N-terminus could be slightly different.

In conclusion, this prediction report shows that the output of a fully automated secondary structure prediction tool can, at least in this case, produce essentially the same secondary structural model as an "expert" manually analyzing the same multiple sequence alignment. Further, it provides a test case for the use of such an output to identify very long distant homologs by comparison of experimentally predicted secondary structural elements with those generated by the automated tool. These approaches are now being used by several groups (e.g., ref. 26). Further, these results suggest that members of the HSP90 family form the same overall fold as the N-terminal domain of gyrase B. If this suggestion is correct, it indicates that the automated program and the "expert" both mispredict an internal helix.

REFERENCES

1. Benner, S.A. Patterns of divergence in homologos proteins as indicators of tertiary and quaternary structure. Adv. Enzymol. Reg. 28:219–236, 1989.
2. Pananoyotou, G., Bax, B., Gout, I., Federwisch, M., Wroblowski, B., Dhand, R., Fry, M.J., Blundell, T.L., Wollmer, A., Waterfield, M.D. Interaction of the p85 subunit of PI 3-kinase and its N-terminal SH2 domain with a PDGF receptor phosphorylation site. EMBO J. 11:4261–4272, 1992.
3. Russell, R.B., Breed, J., Barton, G.J. Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains. FEBS Lett. 304:15–20, 1992.
4. Musacchio, A., Gibson, T., Lehto, V.-P., Saraste, M. SH3. An abundant protein domain in search of a function. FEBS Lett. 304:15–20, 1992.
5. Bazan, J.F. Structural design and molecular evolution of a cytokine receptor superfamily. Proc. Natl. Acad. Sci. USA 87:6934–6937, 1990.
6. Moe, G.R., Koshland Jr., D.E. Transmembrane signalling through the aspartate receptor. In: "Microbial Energy Transduction, Genetics, Structure and Function of Membrane Proteins." Youvan, D.C., Daldal, F. (eds.). Cold Spring Harbor, NY: Cold Spring Harbor Press, 1986.
7. Benner, S.A., Gerloff, D.L., Jenny, T.F. Predicting protein crystal structures. Science 265:1642–1644, 1994.
8. Benner, S.A. Chelvanyagam, G., Turcotte, M. Bona fide predictions of protein secondary structure using transparent analyses of multiple sequence alignments. Chem. Rev. 1996. (submitted).
9. Benner, S.A., Ellington, A.D. Evolution and structural theory: The frontier between chemistry and biochemistry. Bioorg. Chem. Front. 1:1–70, 1990.
10. Benner, S.A. Predicting the conformation of proteins from sequence data. Prot. Eng. 71:99, 1996.
11. Chothia, C., Lesk, A.M. The relation between the divergence of sequence and structure in proteins. EMBO J. 5:823–826, 1986.
12. Rost, B., Sander, C. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 231:584–599, 1983.
13. DeFay, T., Cohen, F.E. Evaluation of current techniques for ab initio predictions. Proteins 23:431, 1995.
14. Moult, J., Pedersen, J.T., Judson, R., Fidelis, K. A large-scale experiment to assess protein structure prediction methods. Proteins 23:R2, 1995.
15. Bairoch, A., Boeckmann, B. The SWISS-PROT protein sequence data bank. Nucleic Acids Res. 20:2019–2022, 1992.
16. Gonnet, G.H., Benner, S.A. Computational biochemistry research at ETH. Tech. Rep. 154, 1991.
17. Gonnet, G.H., Cohen, M.A., Benner, S.A. Exhaustive matching of the entire protein sequence database. Science 256: 1443–1445, 1992.
18. Korostensky, C., Gonnet, G.H. Evaluation measures of multiple sequence alignments. Symp. Discrete Algorithms, 1997. (in preparation).

19. Benner, S.A., Badcoe, I., Cohen, M.A., Gerloff, D.L. Bona fide predicton of aspects of protein conformation: Assigning interior and surface residues from patterns of variation and conservation in homologous protein sequences. J. Mol. Biol. 235:926–958, 1994.
20. Benner, S.A., Gerloff, D. Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: The catalytic domain of protein kinases. Adv. Enzymol. Reg. 31:121–181, 1991.
21. Csermely, P., Kahn, C.R. Yhe 90-kDa heat shock protein (hsp-90) possesses an ATP binding site and autophosphorylating activity. J. Biol. Chem. 266:4943–4950, 1991.
22. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247:536–540, 1995.
23. Wigley, D.B., Davies, G.J., Dodson, E.J., Maxwell, A., Dodson, G. Crystal structure of an N-terminal fragment of the DNA gyrase B protein. Nature 351:624–629, 1991.
24. Jenny, T.F., Benner, S.A. Evaluating predictions of secondary structure in proteins. Biochem. Biophys. Res. Commun. 200:149–155, 1994.
25. Bowie, J.U., Luthy, R., Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170, 1991.
26. Russell, R.B., Copley, R.R., Barton, G.J. Protein fold recognition by mapping predicted secondary structures. J. Mol. Biol. 259:349–365, 1996.