118

PROTEINS: Structure, Function, and Genetics 21:273–281 (1995)

PREDICTION REPORT

A Consensus Prediction of the Secondary Structure for the 6-Phospho-β-D-Galactosidase Superfamily

Dietlind L. Gerloff and Steven A. Benner
Department of Chemistry, Swiss Federal Institute of Technology, CH-8092 Zürich, Switzerland

ABSTRACT **Two separate unrefined models for the secondary structure of two subfamilies of the 6-phospho-β-D-galactosidase superfamily were independently constructed by examining patterns of variation and conservation within homologous protein sequences, assigning surface, interior, parsing, and active site residues to positions in the alignment, and identifying periodicities in these. A consensus model for the secondary structure of the entire superfamily was then built. The prediction tests the limits of an unrefined prediction made using this approach in a large protein with substantial functional and sequence divergence within the family. The protein belongs to the (α–β class), with the core β strands aligned parallel. The supersecondary structural elements that are readily identified in this model is a parallel β sheet built by strands C, D, and E, with helices 2 and 3 connecting strands (C + D) and (D + E), respectively, and an analogous β–α unit (strand G and helix 7) toward the end of the sequence. The resemblance of the supersecondary model to the tertiary structure formed by 8-fold α–β barrel proteins is almost certainly not coincidental.** © 1995 Wiley-Liss, Inc.

Key words: prediction contests, α–β barrel, protein sequence alignment

A central problem in protein chemistry challenges the chemist to deduce the conformation (secondary and tertiary structure) of a protein from sequence information (primary structure). Both at the ETH in Zurich[1] and elsewhere,[2–6] progress toward solution of this problem has come through an analysis of patterns of conservation and variation in the sequences of homologous proteins.[7] Such an analysis is especially powerful when it is aided by detailed models of divergent evolution.[8] 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.[9]

The value of these methods can be explored by using them to make bona fide predictions, those published before an experimental structure becomes available. To date, over a dozen bona fide predictions have been made using these methods [reviewed in refs. 7 and 10]. For about half of these, a subsequently determined crystal structure has allowed these predictions to be evaluated. In many cases, the predictions have proven to be remarkably accurate.[10] It is now clear that predictions are possible that miss no core secondary structural elements, misassign no α helices as β strands (or vice versa), and do not overpredict any significant secondary structural element.[11] Predictions meeting these criterion are satisfactory as starting points to assemble a tertiary structural model of a protein family. Predicted secondary structures for pleckstrin homology domain,[12,13] hemorrhagic metalloproteinases,[14] and Src homology 2 domains[2,3] come close to meeting this standard.

Ongoing bona fide prediction efforts are necessary to define the scope of prediction methods. Over time, a large set of examples will emerge that will become statistically representative of proteins as a whole. As this set has accumulated to date, it has become clear that misassignments made by evolutionary analyses come in five principal types:

1. where multiple alignment is incorrect;
2. where the secondary structure of homologous proteins has diverged;
3. when attempting to distinguish between surface β strands and surface loops;
4. when attempting to distinguish between long internal β strands and internal helices;
5. when attempting to assign secondary structure to active site regions.

The first two (and often the first three) problems are interrelated. When the secondary structure has diverged, this often creates bad multiple alignments. Further, the distinction between a surface strand and a surface loop is often difficult even when the experimental data are in hand, and these elements

Received November 11, 1994; revision accepted December 8, 1994.

Address reprint requests to Steven A. Benner, Department of Chemistry, ETH, CH-8092 Zürich, Switzerland.

often undergo substantial divergence in conformation during divergent evolution.

A challenge was issued on October 10, 1994, to predict the secondary and elements of the supersecondary structure of the 6-phospho-β-D-galactosidase superfamily. The prediction was due before November 1. This protein family appeared to be an excellent target for placing the method to an extreme test. The protein is large; the target sequence has 468 amino acids. The family appears to adopt quaternary structure, at least in some cases. Both the thioglucosidase from *Brassica napus* (rape) and the thioglucosidase from *Sinapis alba* (white mustard) are reported to be homodimers, while the β-galactosidase from *Sulfolobus solfataricus* (not shown in the alignment) is a homotetramer. This implies that quaternary contacts might bury some residues that are on the surface of subunits, complicating the secondary structure prediction. Further, biological function has diverged substantially within the protein family, as measured by a wide divergence in substrate specificity in the member proteins. Finally, with only a few days to make a prediction, the example tests the ability of a prediction method to produce an accurate model without the benefit of extensive refinement.

A multiple alignment for the protein family was built from sequences extracted from SwissProt 29[15] using the DARWIN system.[16,17] Surface and interior residues were assigned by automated procedures similar to those described elsewhere,[18] the multiple alignment was parsed into units forming independent secondary structures, and elements of secondary structure were predicted within the parsed segments from patterns of conservation and variation, as described elsewhere.[10,12,14,19] 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 (WWW) with URL http://cbrg.inf.ethz.ch/.

The secondary structure prediction is presented residue-by-residue in Figure 1, and summarized in Table 1. A summary of the secondary structure prediction follows:

Strand A (a009–a011; b049–051) is a short internal segment confirmed in both subfamilies.

Strand B (a014–020; b053–060) is separated from strand A by a GG dipeptide in a well anchored region of the alignment. It is largely internal. This region is interesting from a methodological point of view, as a strong assignment would not have been possible if only one of the two subfamilies were available. In both subfamilies considered alone, an internal helix would be possible. Together, however, a GG dipeptide parse at (a012–103) and a GG dipeptide parse at positions (b062–063), together with strong alignment anchoring excludes an internal helix in this region.

Parse region (a048–059; b071–081) is problematic. Subfamily b could contain a β strand in this region (b073–078). However, it is matched with a parsing string (PGDSG; a050–054) in sequence e of subfamily a, and a strand was not assigned.

Helix 1 (a072–084; b095–107) is reliably assigned in both subfamilies, is well anchored, and displays good amphiphilicity.

Strand C (a089–093; b111–115) is problematic in subfamily a, in part because of the small number of sequences available in this subfamily. In subfamily b, the surface and interior assignments display alternating periodicity, which confirms the strand assignment.

Active site a (a095–102; b117–125) contains conserved Arg, Ser, Trp, and Arg. It is strongly assigned.

Helix 2 (a116–130; b138–153) is reliably assigned in both subfamilies, is well anchored, and displays good amphiphilicity.

Strand D (a136–140; b159–163) is well parsed in subfamily 2, is confirmed in both subfamilies, and is largely internal.

Active site b (a141; b164–166) contains conserved Thr (part of the preceding strand) and His.

Helix 3 (a158–177; b181–198) is reliably assigned in both subfamilies, is well anchored, and displays good amphiphilicity.

Strand E (a182–185; b205–208) is assigned to a region that is near the active site, where conservation associated with active site function often obscures patterns of variation and conservation that might be used to assign secondary structure.

Active site c (a184–187; b205–209) is relatively weak, and is based ultimately on a single conserved Asn. A conserved Thr two residues before supports this conclusion. Interesting, a Trp two residues earlier is almost completely conserved in the superfamily as well, as is a Glu immediately following the conserved Asn.

A region (b212–215) following this active site segment might be assigned as a β strand in subfamily b. It is not paired with a reliable assignment in subfamily a, which contains repeated parsing elements that almost certainly exclude a standard secondary structural element in this area. A similar β strand might be assigned in subfamily b (b219–221); this again has no corresponding element in subfamily a, and might form a β hairpin with the preceding strand. The alignment is poorly anchored in this region, and considerable sequence divergence between the two subfamilies is evident.

Helix 4 (a212–226; b248–268) is cleanly amphiphilic up to position a227, when an interior assignment appears on the surface arc of the amphiphilic helix. The following segment also forms a short (8 residues) amphiphilic helical pattern. In subfamily b, the helix is largely internal. Nevertheless, to the extent that amphiphilicity is detected, it extends past the position where the amphiphilic pattern is broken. This indicates that the contacts made in subfamily a are different from those made in subfamily b. Interestingly, this helix contains a conserved His (a218; b255) and a nearly conserved His (a224; b261).

Strand x (a242–245) is cleanly parsed in subfamily a, and is canonically assigned as a short β strand. The segment is disrupted by parsing elements in subfamily b, which appears to be well anchored. It is possible to identify a plausible β segment in this subfamily. Our experience, however, has been that the experimental assignments made for such regions depend strongly on the experimental secondary structure assignment tool.

Helix x (a259–273) is not cleanly amphiphilic (position a269), but is assigned nevertheless when considering subfamily a alone. A gap is placed in its middle in subfamily b (positions b311–312). If the multiple alignment of subfamily b2 is rearranged, a helix can be detected from positions (b303–317; total length 13 positions). If the multiple alignment of subfamily b1 is adjusted, and the sequence with the deletion discarded, a weak helix can also be found. The ambiguous alignment makes all of these assignments insecure, however, and there is significant possibility that the conformations of different members of the superfamily are quite different.

Strand y (a275–280) is assigned in subfamily a only. It corresponds to a parsed region in subfamily b. Two interior residues (b323–324) might form a corresponding structure, however, in subfamily b.

The amphiphilicity of helix 5 (a286–293; b332–342) is difficult to detect when examining the alignment overall. Examining subalignments, especially of subfamily b1 and subfamily b2, makes the amphiphilicity clearer.

The region (a314) might be assigned as a short helix (7–10 residues) if the left side of subfamily a is examined alone. There is no confirmation of this helix elsewhere, however, as this region of the alignment has undergone massive sequence divergence.

Strand F (a323–327; b381–388) is badly parsed in subfamily a. The segment is conceivably a continuation of a putative helix that may follow. In subfamily b, the strand is more reliably assigned. An excellent set of anchors aligns the subalignments, and we have chosen on these grounds to make the assignment definitive in the consensus secondary structure model.

Helix y (a329–339) is short, and contains a problematic residue at position (a336). There is no confirmation for a helix assignment in subalignment b. The ambiguous alignment makes this assignment further insecure, and there is a significant possibility that the conformations of different members of the superfamily are quite different.

Strand z (a375–382) is assigned in a region of the multiple alignment that has undergone massive sequence divergence, and where DARWIN had extreme difficulties achieving a plausible matching. It has plausible amphiphilicity in subfamily a. Therefore, the multiple alignment in subfamily b was collapsed in an effort to obtain regions that might also form β strands. For subfamily b1, segment (b446–452) displayed an alternating pattern. For subfamily b2, this was not possible, although it cannot be excluded that further rearrangement of the multiple alignment upon refinement could find an analogous region. As time was inadequate to do a complete search of different possible multiple alignments, no strand was assigned in this region in the consensus model.

Helix 6 (a385–398; b456–469) is well parsed, well anchored, amphiphilic, and confirmed in both subfamilies. It might, however, be missing one turn in some proteins in subfamily b.

Strand G (a404–407; b476–479) is well parsed, internal, and confirmed in both subfamilies.

Active site d (a408–410; b480–482), containing conserved Glu, Asn, and Gly, is not strongly assigned by analysis of the sequences themselves. It is, however, supported by biochemical work.[20]

Helix 7 (a431–448; b497–517) is well parsed, well anchored, amphiphilic, and confirmed in both subfamilies.

Residues (a451–a482; b522–554) form a remarkable segment. In subfamily b, the segment is not parsed for 35 residues, has a large number of interior residues, and apparently contains more than one secondary structural element. The first task is to parse this section. To this end, four additional columns were added to the multiple alignment by recognizing that lactase phlorizin hydrolase has multiple internal repeats. Interestingly, in two of these repeats, a parsing string PG appears. However, the repeats that contain this parsing string are cleaved proteolytically during the posttranslational modification.[1] These repeats are also missing Glu (b480), presumed to be part of an active site. Thus, there is no guarantee that these repeats have divergently evolved under functional constraints. This example makes an important point regarding the analysis of homologous sequences in the prediction of a protein structure.

In this region, an internal helix must be considered. Assignment of internal helices (as opposed to internal strands) relies on accurate parsing. The two subalignments were first carefully anchored. A reliable parse at (a471) was matched with a weak parse at (b541). A dipeptide GP parse in subfamily a (a460–462) was used to divide the first part of this segment. The conserved Asp was assumed to also indicate a break in secondary structure (as opposed to being an indicator of an active site position). This led to the assignment of four secondary structural elements in this region as follows:

Subfamily a

Pos	cba	def	SIAPred	Parse
001	M M	M	.	
002	S T	S	s	
003	KKK	N-	s	
004	QTT	P-	.	
005	LLL	FFF	I	
006	PPP	PPP	.	
007	QEK	EAE	S	
008	DDD	SHT	s	
009	FFF	FFF	I	strand internal
010	VII	LLL	I	strand
011	MFF	WWW	I	strand
012	GGG	GGG	.	parse
013	GGG	GGG	.	parse
014	AAA	AAA	I	strand internal
015	TTT	LIT	i	strand
016	AAA	AAA	I	strand
017	AAA	AAA	I	strand
018	YYY	NNN	i	strand
019	QQQ	QQQ	A	strand
020	VAA	SVV	i	strand
021	EEE	EEE	A	active site?
022	GGG	GGG	.	
023	AAA	AAA	I	
024	TTT	FYW	i	
025	KNH	RLQ	S	
026	ETT	ETE	s	
027	DDD	GDD	s	parse
028	GGG	DGG	s	parse
029	KKK	KKK	A	
030	GGG	GGG	.	
031	___	LLI	i	
032	___	TSS	s	
033	___	TTT	a	
034	___	VSS	.	
035	___	DDD	a	
036	___	MLL	i	
037	___	IQQ	s	
038	___	PPP	.	
039	___	HQH	s	
040	___	GGG	.	
041	___	_IV	i	
042	___	_FM	i	
043	___	_GG	.	
044	___	_EK	S	
045	___	_IM	i	
046	___	EVE	S	
047	___	HTP	s	
048	RRP	RRR	S	parse
049	VVV	MQI	i	
050	LAA	APL	i	parse
051	WWW	VGG	i	parse
052	DDD	KDK	S	parse
053	DTK	LSE	S	parse
054	FYY	GGN	s	parse
055	LLL	LII	I	
056	DEE	EKK	S	
057	KED	K__	S	
058	QNN	R__	s	
059	GYY	F__	i	
060	___	Q__	a	
061	___	L__	.	
062	___	R__	a	
063	___	D__	a	
064	___	D__	a	
065	___	E__	a	
066	RWW	F__	s	
067	FYY	Y__	i	
068	KTT	P__	s	
069	PAA	S__	i	parse
070	DEE	HDD	S	parse
071	PPP	EVV	s	parse
072	AAA	AAA	I	helix amphiphilic
073	ASS	TII	i	helix
074	DDD	DDD	A	helix
075	FFF	FFF	I	helix
076	YYY	YYY	I	helix
077	HNH	HHH	S	helix
078	RRK	RRR	S	helix
079	YYY	YYY	I	helix
080	DPP	KPP	S	helix, parse?
081	EVV	EQE	s	helix
082	DDD	DDD	A	active site?
083	LLL	III	I	helix
084	AEE	AAA	.	helix
085	LLL	LLL	I	break to the inside
086	ASA	MFF	i	
087	EEE	AAA	.	
088	KKE	EEE	S	
089	YFY	MMM	I	strand?
090	GGG	GGG	.	strand?
091	HVV	FFF	i	strand?
092	QNN	KTT	s	strand?
093	VGG	VCC	i	strand?
094	III	FLL	I	
095	RRR	RRR	A	active site
096	VII	TII	i	
097	SSS	SSS	A	active site
098	III	III	I	
099	AAA	AAA	I	
100	WWW	WWW	I	active site
101	SSS	STA		
102	RRR	RRR	A	active site
103	III	LII	I	
104	FFF	FFF	I	
105	PPP	PPP	.	parse
106	DNT	QQQ	S	parse
107	GGG	GGG	.	parse
108	AYY	___	i	
109	GGG	DDD	s	
110	EEE	EEE	A	
111	VVV	IAV	I	
112	ENN	TEE	S	parse
113	PPE	PPP	S	parse
114	___	NNN	a	parse
115	___	QEE	s	
116	RKK	QAA	s	helix amphiphilic
117	GGG	GGG	.	helix
118	VVV	ILL	I	helix
119	AEE	AAA	.	helix
120	FYF	FFF	I	helix
121	YYY	YYY	I	helix
122	HHH	RDD	s	helix
123	KKK	SRR	s	helix
124	LLL	VLL	I	helix
125	FFF	FFF	I	helix
126	AAA	EDD	s	helix
127	DEE	EEE	s	helix
128	CCC	CLM	I	helix
129	AHH	KAA	s	helix
130	AKK	KKQ	S	helix
131	HRR	YYA	.	break to surface
132	HHH	GGG	.	
133	IVV	III	I	
134	EEE	EQK	S	
135	PPP	PPP	.	parse?
136	FFF	LLL	I	strand
137	VVV	VVV	I	strand
138	TTT	TTT	A	strand
139	LLL	LLL	I	strand
140	HHH	CSS	.	strand
141	HHH	HHH	A	active site
142	FFF	FYY	I	
143	DDD	DEE	s	
144	TTT	VMM	i	
145	PPP	PPP	.	
146	EEE	MYY	.	
147	RVA	HGG	s	
148	LLL	LLL	I	
149	HHH	VVV	i	
150	EKS	TEK	S	parse

Subfamily b

Pos	e	ba	dc	f	lj	ig	kh	SIAPred	Parse
040	-	--	--	-	L	LL	--	i	
041	-	--	--	-	S	SS	--	a	
042	-	--	--	-	R	SS	--	s	
043	-	--	--	-	S	KK	--	s	
044	-	--	--	-	S	NN	--	s	
-045	F	FF	FF	F	FF	FF	FF	I	
=046	P	PP	PP	P	AP	GG	PP	.	parse
-047	A	KK	SQ	G	PR	KK	EE	s	parse
-048	T	GD	DD	D	GG	DD	GG	s	parse
=049	F	FF	FF	F	FF	FF	FF	I	strand internal
-050	M	LI	KM	L	VI	II	IV	I	strand
-051	W	WW	WW	F	FF	FF	WW	I	strand
=052	G	GG	GG	G	GG	GG	SS	.	
053	T	AS	VT	V	TA	VV	AT		strand internal
-054	S	AA	AA	A	AG	AA	AS	i	strand
055	T	TT	TT	T	SS	SS	ST		strand
-056	S	AA	AA	A	SS	SS	AA	i	strand
-057	S	SA	AA	S	AA	AA	AA	i	strand
058	Y	YY	YY	F	FY	YY	YF	I	strand
=059	Q	QQ	QQ	Q	QQ	QQ	QQ	A	strand
060	I	II	II	I	YF	II	II	I	strand
=061	E	EE	EE	E	EE	EE	EE	A	active site?
=062	G	GG	GG	G	GG	GG	GG	.	parse
-063	G	AA	AA	S	AA	GG	AA	i	parse
-064	T	WY	YY	T	AV	___	WW	i	parse
065	D	NN	NQ	K	FN	___	RR	S	parse
066	E	EE	EE	A	EE	___	AA	s	parse
=067	G	DD	DD	D	DG	RR	DD	s	parse
=068	G	GG	GG	G	GG	GG	GG	.	parse
-069	R	KK	RR	R	KR	RR	KK	s	parse
-070	T	GG	GG	K	GG	GG	GG	s	parse
071	P	EE	ML	P	PP	VV	LL	.	parse
072	S	SS	SS	S	SS	NN	SG	s	parse
073	I	II	II	I	II	VV	II	I	
=074	W	WW	WW	W	WW	WW	WW	I	
=075	D	DD	DD	D	DD	DD	DD	A	
076	T	RR	TT	A	TT	GG	TT	s	
-077	F	FF	FF	F	FF	FF	FF	I	
078	C	TS	AA	C	TT	SS	ST		
079	Q	HH	HH	N	HH	HH	HH	s	
080	I	QT	TT	M	KK	RR	TT	S	
081	_	___	___	_	YY	YY	___	i	
082	P	KP	PP	P	PP	PP	PR	S	parse
083	G	RG	GG	G	EE	EE	LL	s	parse
084	K	NN	KK	H	KK	KK	RK	S	
085	V	II	VV	V	II	SA	VI	i	
086	I	LA	KF	F	___	GG	EE	s	parse
087	G	YD	NN	G	KR	SS	NN	s	parse
088	_	___	___	_	DD	DD	DD	a	parse
089	G	GG	GG	R	RG	LL	AD	S	parse
090	D	HH	DD	H	TS	KK	II	s	parse
091	C	NT	NN	N	NN	NN	___	s	parse
092	G	GG	GG	G	GA	GG	GA	.	parse
093	D	DD	NN	D	DD	DD	DD	s	parse
094	V	VV	VV	I	VI	TT	VV	I	
-095	A	AA	AA	A	AT	ST	AA	S	helix amphiphilic
096	C	CC	CC	C	IV	CC	CC	I	helix
-097	D	DD	DD	D	DD	EE	DD	s	helix
098	H	HH	SS	H	EQ	SS	SS	s	helix
-099	F	YY	YY	Y	YY	YY	YY	I	helix
-100	H	HH	HH	N	HH	TT	HH		helix
-101	H	RR	RR	R	RR	RR	KK	s	helix
-102	F	FY	VY	W	YY	WW	II	I	helix
103	K	EE	EE	E	KK	KQ	AS	S	helix
104	E	EE	EE	E	EE	KK	EE	s	helix
=105	D	DD	DD	D	DD	DD	DD	A	helix active site
-106	V	VI	VI	L	IV	VV	LV	I	helix
107	Q	SK	QR	D	GG	ED	VV	S	helix
-108	L	LI	LL	L	II	IV	TA	I	break to inside
109	M	MM	LM	I	MM	MM	LL	I	
110	K	KK	KK	K	KK	GG	QQ	s	
111	Q	EE	DE	E	DD	EE	NN	s	strand
112	L	LI	LL	M	MQ	LL	LL	I	strand
-113	G	GG	GG	G	NN	NN	GA	s	strand
114	F	LI	VI	V	LM	AA	VV	I	strand
115	L	KK	KR	E	DD	TT	ST	s	strand
116	H	AS	VT	A	AS	GG	HH	s	
117	Y	YY	YY	Y	YY	YY	YY	I	active site
=118	R	RR	RR	R	RR	RR	RR	A	active site
119	F	FF	FF	F	FF	FF	FF	I	active site
=120	S	SS	SS	S	SS	SS	SS	A	active site
-121	V	II	IV	L	II	FF	II	I	
-122	A	AS	SS	A	SS	AA	SS	i	
=123	W	WW	WW	W	WW	WW	WW	I	active site
-124	P	TP	PP	P	PP	SS	SS	s	
=125	R	RR	RR	R	RR	RR	RR	A	active site
126	I	II	VI	I	VI	II	II	I	
-127	M	FF	LF	I	LL	VI	LL	I	
=128	P	PP	PP	P	PP	PP	PP	.	parse
-129	A	DE	QN	D	KK	KK	DD	s	parse
=130	_	GG	GG	G	GG	GG	GG	.	parse
131	_	___	___	_	KK	KK	TT	s	parse
132	_	___	___	_	LL	VV	___	i	parse
133	A	FT	TD	F	SS	SS	TT	s	parse
134	G	GG	GG	G	GG	RR	RN	S	
-135	I	TK	EE	P	GG	GG	YY	s	
136	I	VL	VV	I	VI	VV	II	I	
137	N	NN	NN	N	NN	DN	NN	S	break
138	E	QQ	RQ	E	RH	QQ	EE	S	helix amphiphilic
139	E	KK	AE	K	EE	AG	AA	s	helix
=140	G	GG	GG	G	GG	GG	GG	.	helix
-141	L	LL	LL	L	II	LL	LL	I	helix
142	L	ED	DD	D	NK	DD	NN	s	helix
-143	F	FF	YY	F	YY	YY	YY	I	helix
=144	Y	YY	YY	Y	YY	YY	YY	I	helix
-145	E	DK	HH	D	NN	HH	VV	s	helix
-146	H	RR	RR	R	NN	NK	RR	S	helix
-147	L	LL	LV	L	LL	LL	LL	I	helix
148	L	IT	VV	V	II	II	II	I	helix
149	D	NN	DD	D	NN	DD	DD	s	helix
150	E	KL	EL	G	EE	AA	TA	s	helix
-151	I	LL	LL	C	VL	LL	LL	I	helix
152	E	VL	LN	K	LL	LL	LL	i	helix
153	L	EE	AD	A	AA	EE	AA	s	helix
154	A	NN	NN	R	NN	KK	AA	s	break
155	G	GG	GG	G	GG	NN	SN	S	parse
156	L	II	II	I	MI	II	II	I	
-157	I	EM	EE	K	QQ	TT	QK	S	
=158	P	PP	PP	T	PP	PP	PP	s	parse?
159	M	VA	FF	Y	YF	FF	QQ	I	strand
-160	L	VI	CC	A	VV	VV	VV	I	strand
=161	T	TT	TT	T	TT	TT	TT	A	strand
-162	L	LL	LL	L	LL	LL	IM	I	strand
163	Y	YY	YY	Y	FF	FF	YY	I	strand
=164	H	HH	HH	H	HH	HH	HH	A	active site
165	W	WW	WW	W	WW	WW	WF	I	
=166	D	DD	DD	D	DD	DD	DD	A	active site
167	L	LL	LL	L	VL	LL	LL	I	
=168	P	PP	PP	P	PP	PP	PP	.	parse
169	Q	QQ	QQ	L	QQ	QQ	QQ	I	
170	W	KK	AA	T	AV	TT	TA	s	surface strand?
-171	I	LL	LL	L	LL	LL	LL	I	
172	E	QQ	QQ	M	EE	QQ	QQ	.	
173	D	DD	DD	G	DD	DD	DD	s	parse

```
151 ADN EKN        S parse
152 ___ YHY        s parse
153 GGG GGG        . parse
154 DDD SGG        s parse
155 WFF WWW        I
156 LLL RGA        s
157 SNN NNN        s
158 QRR RRR        s helix
159 EKE KLA        S helix
160 MTN LTV          helix
161 LII VII        I helix
162 DDE EDD        S helix
163 DYH FCH        S helix
164 FFF FFF        I helix
165 VVI SEE        s helix
166 ADD RRH        S helix
167 YYY YYY        I helix
168 AAA AAA        I helix
169 KEA RRR        S helix
170 FYF TTT        i helix
171 CCC CVV        I helix
172 FFF FFF        I helix
173 EKE EAT        S helix
174 EEE ARR        s helix
175 FFF FYY        I helix
176 SPP DRQ        S helix
177 EEE GHH        s helix
178 ___ LKK        s
179 VVV VVV        I
180 KKN KKA        S
181 YYY YRL        s
182 WWW WWW        I strand
183 ITT LLL        i strand
184 TTT TTT        A strand
185 IFF FFF        I strand
186 NNN NNN        A active site
187 EEE EEE        A
188 PII III        i parse

189 TGG NNN        s parse
190 SPP IMM        i parse
191 MII MSS        i
192 AGG LLL        i
193 VDD HHH        .
194 QGG SAA        .
195 QQQ PPP        .
196 YYY FFF        I
197 TLL STT        s
198 ___ GGG        .
199 TVV AVV        i
200 GGG GGG        .
201 TKK LLL        .
202 FFF VPA        i parse
203 PPP FPE        S parse
204 PPP EDE        S parse
205 AGG ESS        s parse
206 EII GDG        s parse
207 SKK EKE        S
208 GYY NAA        s



209 RDD Q__        S
210 FFL D__        .
211 DEA Q__        S
212 KKK VAE        s helix mostly internal
213 TVV KIV        s helix
```

```
-174 E IK QA D   EE EE __     s   parse
 175 _ __ __ _   YY YY VV     i   parse
-176 G GG GG G   RG EE GG     s   parse
-177 G GG GG G   GG GG GG     .   parse
-178 W WW WW W   FF FF WW     I
-179 T AK GG A   LL LL EE     s
-180 Q NN SN S   GN DD NN     S   break
 181 R PR RR R   RS RR EE     s   helix amphiphilic
 182 E ED IR S   NG QQ TT     S   helix
 183 T IT TT T   IV II II     I   helix
-184 I VT II A   VI II VV     I   helix
 185 Q ND DQ H   DN QQ QQ     S   helix
 186 H YY AA A   DD DD RR     i   helix
=187 F YF FF F   FF FF FF     I   helix
 188 K FT AV Q   RR KK KK     s   helix
 189 T DE EQ R   DD DD EE     s   helix
=190 Y YY YF Y   YY YY YY     I   helix
=191 A AS AA A   AT AA AA     i   helix
 192 S ME EE K   ED DD DD     S   helix
 193 V LV LT T   LL LL VV     I   helix (break to inside)
-194 I VI MM V   CC CC LL     I   helix
-195 M IF FF M   FF FF FF     I   helix
 196 D NK KR A   KK KK QQ     s   helix
-197 R RN EE R   EE EE RR     s   helix
-198 F YL LF L   FF FF LL     I   helix
 199 G KG GH G   GG GG GG     s   parse, see gap
 200 E DD GG D   DD GG DD     s   parse, see gap
 201 R KI KK R   RR KK KK     S
-202 I VV II L   VV VV VV     I
-203 N KP KQ D   KR KK KK     S
 204 W KI QH A   HY NH FF     S
-205 W WW WW V   WW WW WW     I   strand
 206 N IF IL A   IS II II     i   strand
=207 T TT TT T   TT TT TT     A   strand
-208 I FH FF F   LL II LL     I   strand
=209 N NN NN N   NN NN NN     A   active site
-210 E EE EE E   EE QQ EE     s
 211 P PP PP P   PP LL PP     i   parse

 212 Y YG WW W   WW YY FF     i   strand, internal
 213 C CV CC C   GV TT VV     i   strand
 214 A IV MI A   VF VV IV     I   strand
 215 S AS AA V   SS PP AA     i   strand
 216 I FL FF W   MN TT YY     i
 217 L LL LL L   NS RR QH     s
 218 G GG SS S   AG GG GG     .   parse
 219 Y YH NN H   YY YY YY     i
 220 G FF YM L   AA AA GG     i
 221 T HL LL Y   YL LI YT     I
 222 G GG GG G   GG GG GG     .
 223 E II VV V   TT TT TL
 224 H HH HH H   FN DD AY     s
-225 A AA AA A   AA AA AA     I
-226 P PP PP P   PP PP PP     .   parse
 227 G GG GG G   GG GG GG     .   parse
 228 _ __ __ _   RR RR VI     s
 229 _ __ __ _   CC CC SY     s
 230 _ __ __ _   SS SS NF     s   parse
 231 _ __ __ _   DA PP RR     i   parse
 232 _ __ __ _   WS KM __     s   parse
 233 _ __ __ _   LN VV __     .   parse
 234 _ __ __ _   KV DD __     S
 235 _ __ __ _   LA TT __     i
 236 _ __ __ _   NK KK __     S
 237
 238
 239 _ __ __ _   __ QH __     S
 240 _ __ __ _   __ RR __     s
 241 _ __ __ _   C_ CC __     i
 242 _ __ __ _   T_ YY __     s
 243 _ __ __ _   GP GG __     .   parse
 244 H II NL E   GG GG __     .   parse
 245 E KK KT R   DD NN __     s   parse
 246 N DD DN N   SS SS PP     .   parse
-247 W FL LL M   GG SS GG     .   parse
 248 R KR QQ E   RT TT TT     s   helix mostly internal
 249 E VT LT A   EG EE AA_     s  helix (b1)
 250 A AS AA A   PP PP PP     i   helix (b1)
```

```
214 FFF YYY        I helix
215 QQQ QQQ        A helix
216 ASS AAA        i helix
217 EHH AII        s helix
218 HHH HHH        A helix
219 NNN HHH        . helix
220 QMM QQQ          helix
221 MMM LLL        I helix
222 VVV VVV        I helix
223 AAS AAA        i helix
224 HHH SSS        . helix
225 AAA AAA        I helix
226 RRR LRR        s helix
227 IAA AAA        I break to inside
228 VVV TVV        i
229 NKK KKK        s
230 LLL IAA        I
231 YFY ACC        I
232 KKK HHH        s
233 SDD EDS        S parse
234 MGK VML        S parse


235 QGG NIL        s parse
236 LYY PPP        i parse
237 GKK QDE        S parse
238 GGG NAA        s parse
239 QEE QQK        S
240 III VII        I
241 GGG GGG        .
242 IVV CNN        . strand
243 VVV MMM        I strand
244 HHH LLL        i strand
245 AAA ALL        I strand
246 LLL GGG        i parse
247 QPP GAG        s parse
248 TTT NML        s parse
249 VKK FLV        . s
250 YYY YYY        I
251 PPP PPP        . parse?
252 YFY YLL        I
253 SDD STT          parse
254 DPP CSC        s parse
255 SSE KKQ        S parse
256 ANN PPP        S parse
257 VPP ___        .
258 _EA            S


259 DDD EEQ        S helix
260 HVV DDD        s helix
261 HRR VVM        . si
262 AAA WML        I helix
263 AAA AEQ        s helix
264 EEE ASA        S helix
265 LLL LLM        I helix
266 QEE EHE        S helix
267 DDD KQE        S helix
268 AII DNN        s helix
269 LII RRR          helix
270 EHH EER        S helix
271 NNN NWW        s helix
272 RKK LLM        s helix
273 LFF FFF        I helix
274 YII FFF        I break to inside


275 LLL IGG        i strand
276 DDD DDD          strand
277 GAA VVV        i strand
278 TTT QQQ        i strand
279 LYY AVA        I strand
280 ALL RRR        . strand

281 GGG GGG        .
282 EKH TAQ        S
```

```
-251 F ML II L   YY YY YY     i    helix
 252 T DE DD A   LI II II     si   helix
 253 A VV VV A   AV VV VV     I    helix
 254 A VS SG M   AT AA GG     .    helix
=255 H HH HH H   HH HH HH     A    helix active site
-256 H SN HH H   YN NN NN     s    helix
 257 I LL LL I   QQ QQ LL     I    helix
-258 L ML LL N   LI LL II     i    helix
 259 M LL VV L   LL LL KK     i    helix
-260 C SS AA A   AA AA AA     i    helix
-261 H HH HH H   HH HH HH          helix
-262 G FG GG G   AA AA AA     i    helix
-263 I KK RL F   AE AT EE     S    helix
-264 A VA AS G   AA IV AA     i    helix
 265 S VV VV V   AV VV WW     i    helix
-266 N KK TR E   RH DD HH     S    helix
-267 L AL LR A   LV LL LL     .    helix
 268 H VF FF S   YY YY YY     i    helix
 269 K KR RR R   KK RR NN     s
-270 E EE EE H   TT TT DD     s
 271 K NM LL V   KK NK VV     S
 272 _ __ __ _   YY YY YY
 273 _ __ __ _   QQ __ RR     s
 274 _ __ __ _   AA AK AA     S
-275 G NN GG A   SY FF SS     .
-276 L II IT P   QQ QQ QQ     i    parse
-277 T DD SS K   NK NK GG     S    parse
-278 G VA GG V   GG GG GG     i    parse
-279 K EQ EQ P   IK KK VV     S
-280 I VI II V   II II II     I
-281 G GG GG G   GG GG SS     .    parse
 282 I II II L   II PP II     i    parse
 283 T TA AA V   TT VV TT     I
 284 L LL PP L   LL MM II     i    parse
 285 N NN NN N   VV II SS     i    parse
 286 M LL TV A   SS TT SS     i
 287 E TS SS H   HN RR DD     s
 288 H PY WW S   WW WW WW     i
-289 V VH AA A   FL FF AA     I

-290 D YY VV I   EM LL EE     s
-291 A LP PP P   PP PP PP     .    parse?
-292 A QA YY A   AL YF RR     i
-293 S TS RS S   SD DD DD     S    parse
 294 E EE RT D   KD EE PP     S    parse
 295 R R_ TS G   EN SS SS     s    parse
 296 _ L_ __ _   __ __ __     .
 297 _ G_ __ _   __ __ __     .
 298 _ Y_ __ _   __ __ __     .
 299 _ KK __ _   KS DD NN     s    parse
 300 P VA KE E   _I PP QQ     s
 301 E SE EE A   AP AA EE     s
-302 D ED DD D   DD CS DD     s
 303 V II MK L   VI II VV     .
 304 A EE EA K   DK EE EE     S
-305 A RA AA A   AA AA AA     i
-306 A EA CC A   AA AA AA     i
-307 I ME LA E   KE EE RK     S
 308 R VL RR R   RR RR RR     s
 309 R SS VT A   GS MM YY     s
 310 D LF NI F   LL NN VV     s
 311 _ S_ __ _   __ __ __     s
 312 _ S_ __ _   __ __ __     s
 313 G QS GS Q   DD QQ QQ     s
 314 F LL WL F   FF FF FF     I
 315 I DA SH H   MQ FF MM     s
 316 N NG GS N   LF HH GG     s
 317 R QR DD G   GG GG GG     s
 318 W LW WW A   WL WW WW     I    strand
-319 F FY YF F   FF YY FF     I    strand
 320 A LL LL F   MM MM AA     I    strand
-321 E DD DQ D   HE EE HH     S
 322 P PP PP P   PQ PP PP     s
 323 L VV II V   LL LL II     I
 324 F LL YY F   TT TT FF     I
 325 N KK FQ K   KT KK KK     s
 326 _ __ __ _   __ __ NN     a    parse
=327 G GG GG G   GG GG GG     .    parse
 328 K SR ES E   RD RR DD     S    parse
```

```
283 YYY YYY        I
284 HSS PPP        s parse
285 QRD AGG        S parse
286 EEK YYY        S helix
287 TTT SMM        i helix
288 LMM AHQ        i helix
289 AEE RRR        s helix
290 LGG VYF        i helix
291 VVV FFF        I helix
292 KQN RRR        S helix
293 EHH EED        s helix
294 III ___        i
295 LLL ___        i
296 DSA ___        s
297 AVE ___        S
298 NNN ___        a parse
299 HGG KQH        s parse
300 QGG GGN        S parse
301 PKE VII        S parse
302 MLL TTT        i
303 F__ ILI        i
304 QND NNE        S
305 SIL KIM        s
306 TTR ATT        S
307 PDD PAE        S parse
308 QEE GQS        s parse
309 EDD DDD        s parse
310 MYF DKA        s parse
311 KAQ ___        S
312 AIA ___        s
313 ILL ___        I

314 DDD EQE        S helix
315 EAA IDD        s helix
316 AAA LLL        i helix
317 AKK KKK        s helix
318 HDD NAH        s helix
319 QLL TTT        i helix
320 LNN VVV        is

321 DDD DDD        A
322 FFF FFF        I strand?
323 VLL VII        I strand?
324 GGG SSS        . strand?
325 VII FFF        I strand?
326 NNN SSS          strand?
327 NYY YYY        . strand?
328 YYY YYY        I break to inside
329 FMM AMM        I helix
330 SSS STT        s helix
331 KDD RGG        S helix
332 WWW CCC        I helix
333 LMM AVV        I helix
334 RRQ STS        S helix
335 AGA ATH        s helix
336 YYF EDD        s helix
337 HDD MEE        S helix
338 GGG NAS        s helix
339 KEE AQI        s helix
340 SST NLN        S break to surface
341 EEE NEK        S
342 TII ___        .
343 ITI ___        S
344 HHH ___        a
345 NNN ___        a parse
346 GAG ___        . parse
347 DTK ___        S parse

=329 Y YY YY Yi YY YY YY     I
-330 P PP PP P  PS PP NN     s   parse
-331 E QE KQ As EK DD EE     S   parse
-332 D KN FF Es SS II VV     s   helix
 333 M LA ML Mi MM MM MM     I   helix
 334 V LL LV Mi RR RR KK     s   helix
-335 E DK DD Es YR QQ TT     s   helix
 336 W YL WW Ai LI II RQ         helix
-337 Y LY YF Li VV VV II     i   helix
 338 G VK EA Gs RK GG RR     s   helix
-339 T QK NE Ds KN SS DE     S   helix
 340 Y KK LQ Rs RR RR RR         helix
 341 L DG GG Ms LL LL SS         helix
 342 N LI YA P  PP PP LL     i   helix
 343 _ __ __ _  __ __ __
 344 _ __ __ _  __ __ __
 345 _ __ __ _  __ __ __     .
 346 _ __ __ _  __ __ __
 347 _ __ __ _  __ __ __
 348 _ __ __ _  __ __ AA
 349 _ __ __ _  __ __ AA
 350 _ __ __ _  __ __ GG
 351 _ __ __ _  __ __ LL
 352 _ __ __ _  __ __ NN
 353 _ __ __ _  __ __ KE
 354 _ __ __ _  __ __ SS
 355 G LE KT _  __ __ RR
 356 L DL PV _  __ __ LL
 357 D SS PP _  __ __ PP
 358 F QF II V  __ __ __
 359 V KP VQ V  __ __ __
 360 _ __ __ E  __ __ __
 361 _ __ __ _  __ __ __     .
 362 _ __ __ _  __ __ __
 363 _ __ __ _  __ __ __
 364 _ A_ __ _  __ __ __
 365 _ L_ __ _  __ __ __
 366 _ S_ __ _  KK NN EE     s
 367 _ M_ __ _  FF FF FF     i

 368 Q Q_ __ _  SS TT TT     s
 369 P QE DD A  TK EE EE     s   parse
 370 G ED GG E  EF AE SS     S   parse
 371 D VD DD D  EE EE EE     s   parse
 372 M KL MM L  SS AA KK     s
 373 E EK ED G  KS EE RR     S   coil
 374 L NL LI I  EL LL RR     s
 375 I FI II I  LV VV II     I
 376 Q IS HG S  TN AA NN     s
 377 Q FQ QE Q  GG GG GG     s
 378 P PP PP K  SS SS TT     s   parse
 379 G _I II L  FF YY YY     i   parse
=380 D DD DD D  DD DD DD     A   parse
-381 F FF FM W  FF FF FF     I   strand
-382 L LI II W  LI LL FF     I   strand
-383 G GA GG G  GG GG GG     .   strand
-384 I IF II L  LI LL FF     I   strand
-385 N NN NN N  NN NN NN         strand
-386 Y YN YY Y  YY YY HH     i   strand
=387 Y YY YY Y  YY YY YY     I   strand
 388 T TS TS T  SS VV TT     i   strand
-389 R RS SM P  SS TT TT     s
 390 S AE SS M  YS QQ VV     s
 391 I VF MV R  YY YY LL     i
 392 I RI NN V  AI AA AA
-393 R LK RR A  AS KQ YY
 394 S YY YF D  KN PP NN     S
 395 T DD NN D  AA KK LF     i
 396 N EP PP A  PP PP NN     S   parse
 397 _ __ __ _  RS NN __     s   parse
 398 _ __ __ _  IH __ __     s
 399 _ __ __ _  PG __ __     .
 400 _ __ __ _  NN __ __     a
 401 _ __ __ _  AA PP __     i   parse
 402 _ __ __ _  RK YY YY     s
 403 _ __ __ _  PP PP AP     .   parse
 404 _ __ __ _  AS SS TS     S   parse
 405 _ __ __ _  IY EE AI     s
 406 _ __ __ _  QS TT IM     s

348 GGG ___        . parse
349 TDE ___        S
350 KKK ___        a
351 GGG ___        . parse
352 SGS SKN        S parse
353 SSS STA        s parse
354 VKK ARQ        S
355 AYY AGG        i
356 RQQ NNN        s
357 LLI VII        I
358 QKK ___        s
359 GGG ___        .
360 VVV ___        i
361 GGG ___        .
362 EQR ___        S
363 ERR ___        s
364 KE_ ___        s
365 LF_ VLL        i
366 _D_ KNN        s
367 _VV SMM        i
368 _DA LVI        S
369 P P RPP        s parse
370 D D NNN        s parse
371 ___ PPP        . parse
372 G Y YYH        s parse
373 IVV LLL        I
374 EPP QEK        S

375 TRR VSS        s strand, amphiphilic
376 TTT SSS        i strand
377 DDD DEE        s strand
378 WWW WWW        I strand
379 DDD GGG        s strand
380 WWW WWW        I strand
381 SMI GQQ        s strand
382 III III        I strand
383 YYY DDD        . parse
384 PPP PPP        . parse
385 RQE LLV        S helix amphiphilic
386 GGG GGG        . helix
387 MLL LLL        I helix
388 YYY RRR        . helix
389 DDD IYV        . helix
390 IQQ TLL        i helix
391 LII MLL        I helix
392 MMM NNN        is helix
393 RRR MFT        s helix
394 IVV MLL        I helix
395 HVK YYW        S helix
396 NKN DDD        S helix
397 DDD RRR        s helix
398 YYY YYY        I helix
399 PPP QQQ        . parse

400 LNN ___        s parse
401 VYY ___        i
402 PHK KKK        S distributed parse
403 VKK PPP        s distributed parse
404 TII LLL        i strand
405 YYY FFF        I strand
406 VII LII        I strand
407 TTT VVV        i strand
408 EEE EEE        A active site
409 NNN NNN        A active site
410 GGG GGG        . active site
411 ILL LLL        I
412 GGG GGG        .

 407 _ __ __ _  TT HH SS     .
 408 _ __ __ _  DN TT ST     s
 409 _ __ __ _  SP AA FV     i
 410 _ __ __ _  LM LM DD     s
 411 _ __ __ _  IT MM AA     i
 412 D __ __ T  NN DD DD     s   parse
 413 _ __ __ P  AI AA RR     s   parse
 414 _ NS G_ G  __ GG GG     s   parse
 415 _ SS __ V  __ VV VV     i
 416 _ SE EE E  __ DK AA     S
 417 _ WS __ _  __ LL SS     i
 418 _ IG __ _  TS TT II     s
 419 _ FF __ _  FF FY AV     i
 420 _ PS __ _  EE ND DD     S
 421 _ __ __ _  HK NN RR     s
 422 _ __ __ _  NH SS SS     s
 423 _ __ __ _  __ RR WW     .
 424 _ __ __ _  GG GG PP     s   parse
 425 _ __ __ _  KI EE DG     s   parse
 426 _ _P __ _  PP YF SS     i   parse
 427 _ _A AA F  LL PL GG     i   parse
 428 A _N G_ P  GG GG SS     s   parse
 429 S _S GG A  PP PP FY     .
 430 L _I MF T  MR VL WW
 431 L _L LL M  AA FF LL     i
 432 Q _E SQ P  AA AV KK     s
 433 V _K SS A  SS EE MM     s
 434 E _F EE P  __ __ __     s
 435 Q __ AE A  __ __ __     s
 436 V I_ II V  __ __ __     i
 437 H R_ SN S  __ __ __     s
 438 M W_ MM _  __ __ __     i
 439 E E_ GG _  __ __ __     s
 440 E H_ AL D  __ __ __     s
 441 P P_ PP V  SI _D __     .
 442 _ A_ __ _  WW _K __
 443 _ G_ __ _  LI _V __
 444 _ __ __ _  CY DN __
 445 _ EE __ _  IV AG __

 446 V YK KV K  __ __ __     .   strand (b1)
-447 T TT TT T  __ __ __         strand (b1)
-448 D ED DD D  __ __ __         strand (b1)
 449 M MM II I  __ __ __         strand (b1)
-450 G GG GG G  YY NN TT         strand (b1)
-451 W WW WW W  PP SS PP         strand (b1)
 452 E EI EP E  -Y YY FF     s/i strand (b1)
-453 I VI IV V  -M YY GG     i
 454 H FY YE Y  -F YY FF     i
-455 P PP AS A  -I PP RR     s
-456 E QE ER P  -Q KK RR     s   helix
-457 S GG GG A  -E GG II     i   helix
 458 F LL LL L  -D II LL     i   helix
 459 Y FY YY H  -F YY NN     i   helix
-460 K DD DE T  -E YY WW     s/i helix
-461 L LL LV L  -I VV LI     i   helix
-462 L LL LL V  -F MM KK     i   helix
 463 T IM RH E  -C DD EE     s   helix
-464 R WL YY T  -Y YY EE     s   helix
 465 I IL TL L  -I FF YY     i   helix
 466 E KD AQ Y  -L KK __     s   helix
-467 K ER D_ E  -K TT __     s   helix
-468 D SD KK R  -I KK __     s   helix
-469 F YY YY Y  -N YY __     S   helix
 470 _ __ __ _  -I __ __     i
 471 S __ __ _  -_ __ __     s
 472 K PG GG D  -T NG NN     S   parse
 473 G QK NN L  -I ND DN     S   parse
 474 L IP PI P  -L PP PP     .   parse
 475 P PN TD E  -Q LL PP     s   parse
-476 I II LI C  -F II II     i   strand
 477 L YV YY Y  -S YY YY     i   strand
-478 I II II I  -I IV VV     i   strand
-479 T TS TT T  -T TT TT     .   strand
=480 E EE EE E  -E EE EE     a   active site
=481 N NN NN N  -N NN NN     a   active site
=482 G GG GG G  -G GG GG     .   active site
 483 A AA AA A  -M IF VV     i
 484 A AA CC C  -N SS SS     i
```

```
413 LYY AAA        I
414 KKK KKK        A
415 EDD DDD        s
416 SEE EKS        S
417 LFF FIV        I
418 PIV AEE        s
419 EED AEA        S
420 NSN NND        S parse
421 A__ GGG        . parse
422 T__ EDS        S parse
423 P__ ___        . parse
424 DE_ ___        s parse
425 TK_ ___        s
426 VTT ___        .
427 IVV III        I
428 EHY NYQ        S
429 DDD DDD        A parse
430 PDD DDD        s parse
431 KAG YYY        . helix
432 RRR RRR        A helix, amphiphilic
433 III III        I helix
434 DDD SRA        S helix
435 YYY YYY        I helix
436 VVV LLL        I helix
437 KRK RNN        S helix
438 KQQ EDD        s helix
439 YHH HHH        i helix
440 LLL ILL        I helix
441 SNE RVV        S helix
442 AVV AQQ        i helix
443 MIL MVV        I helix
444 AAS GGN        S parse
445 DDD GEE        s parse
446 AAA TAA        i helix
447 III III        I helix
448 HIA ADA        s helix

449 DDD DDD        A
450 GGG GGG        . strand

451 AAA IVV        I strand
452 NNN PED        S strand
453 VVV LVI        I strand
454 KKK MLM        s strand
455 GGG GGG        i
456 YYY YYY        I strand
457 FFF TTT        i strand
458 III TCS        . strand
459 WWW WWW        I strand
460 SSS GGG        . parse
461 LLL CPP        i parse
462 QMM III        i
463 DDD DDD        A active site
464 QVV LLL        i
465 FFF VVV        I
466 SSS SSS        A
467 WWW AAA        I
468 TSS CSS        s
469 NNN TKH        S
470 GGG GAS        .
471 ___ EEQ        S
472 YYY MMM        I
473 SEE SSS        s
474 KKK KKK        A
475 RRR RRR        A
476 YYY YYY        I
477 GGG GGG        .
478 LLL FFF        I strand
479 FFF VII        I strand
480 FYY FYY        I strand
481 VVV VVV        I strand
482 DDD DDD        A strand
483 FFF RRR        .
484 PED DDD        S parse
485 ___ DDD        a parse
486 ___ AAN        S parse
487 ___ GGG        . parse
488 ___ NHE        S parse
489 ___ GGG        . parse
```

```
 485 M YF YI Y  -E TT QH      s
-486 R NK NN N  -F PP RR      s   parse
-487 D DD DD M  -N GS EG      S   parse
-488 E IE GE G  -D SS ED      S   parse
-489 L VI LV V  -A EE TS      s
 490 V TG SV E  -T SN DY      S   parse
 491 N ES LN N  -L RR LL      s   parse
 492 G DN DG G  -P CE NN      S   parse
 493 Q GG GK Q  -V __ __      s   parse
 494 _ KK R_ _  -E __ __      S
 495 _ __ __ _  -E EQ __      S
 496 _ __ __ _  -A AA __      i

-497 I VI IV V  -L II __      i   helix, amphiphilic
-498 E HE HQ N  -L AA __      s   helix
-499 D DD DD D  -N DD DD      s   helix
 500 T ST QD Q  -T YY TT      s   helix
-501 G KK RR P  -Y KK AT      s   helix
=502 R RR RR R  -R RR RR      a   helix
-503 H II II L  -I II II      i   helix
-504 G EQ DS D  -D ND YY      S   helix
=505 Y YY YY Y  -Y YY YY      i   helix
 506 I LL LM Y  -Y LL LL      i   helix
-507 E KK AQ A  -Y CC RR      s   helix
-508 E QD MQ E  -R SS TS      s   helix
-509 H HY HH H  -H HH YY      i   helix
-510 L FL LL L  -L LL II      i   helix
-511 K ET IV G  -Y CC NN      s   helix
-512 A AQ QQ I  -Y FF EE      i   helix
-513 C AA AV V  -I LL AA      i   helix
 514 H RH SH A  -R RR LL      s   helix
 515 R KR RR D  -S KK KK      S   helix
-516 F AA AT L  -A VV AA      i   helix
-517 I II II I  -I II VV      i   helix
 518 _ __ __ _  -R RK _Q      S
 519 E EQ EH R  -A EE QQ      S
-520 E ND DD D  -_ KK DD      s
-521 G GG GG G  -G GG KK      s   strand

 522 G VV IL Y  -S VV VV      i   strand
-523 Q DN NH P  -N NN DD      s   strand
-524 L LL LV M  -V IV LL      i   strand
 525 K RK KK R  -K RR RR      S   strand
-526 G GA GG Gi -G GG GG      .
-527 Y YY YY Yi -F YY YYi     i   strand
-528 F FY MM Fi -Y FF TTi     i   strand
 529 V VL EA As -A AA VVi     s   strand
=530 W WW WW Wi -W WW WWi     i   strand
 531 S SS SS S  -S AA STi         strand
 532 F LL LL Li -F LL ALi     i   strand
-533 L ML ML Mi -L GG MMi     i   strand
=534 D DD DD D  -D DD DD      a   active site
 535 N NN NN N  -C NN NN      s
 536 F FF FF F  -N YY FF      i
 537 E EE EE E  -E EE EE      a
 538 W WW WW W  -W FF WW      i
 539 A AA AA A  -F CC AY      i
 540 W MY EE E  -A KK TT      s
-541 G GG GG G  -G GG GG      .

 542 Y YY YY Y  -F FF FF      i
 543 S TN GN R  -T TT SS      s
-544 K KK MM M  -V VV ED      S
=545 R RR RR R  -R RR RK      S
 546 F FF FF F  -F FF FF      i
=547 G GG GG G  -G GG GG      .
 548 I II LM L  -L LL LL      i   strand
-549 V IV VI V  -N SS HH      i   strand
-550 H YH HH H  -F YY FF      i   strand
-551 I VV VV V  -V VV VV      i   strand
-552 N DN DD D  -D NN NN      s   strand
 553 Y YF YF Y  -- WW YY      i   strand
-554 E ED DR Q  -- DE SS      S   strand
 555 _ __ __ _  -- DD DD      a
 556 _ __ __ _  -- LL PP      i
```

```
490 TTT TSS        .
491 QQQ LLL        i
492 NEE TET        S
493 RRR RRR        A
494 YYY TRT        s
495 IPP HRR        s break
496 KKK RKK        s helix, amphiphilic
497 QKK KKK        s helix
498 SSS SSS        A helix
499 AAA FFF        I helix
500 EYH WYG        S helix
501 WWW WWW        I helix
502 FYY YYY        I helix
503 KKK KQA        s helix
504 SEK KSE        S helix
505 VLL VVV        I helix
506 SAA III        i helix
507 EEE A-K        s helix
508 TTT S-T        . helix
509 H-Q N-R        S helix
510 I-V ---        .
511 I-I ---        .
```

```
-557 T TT TT T  -- DD SS      i
-558 Q QL LQ Q  -- __ LL      .
 559 E KE VV V  -- DD PP      s
=560 R RR RR R  -- RR RR      a
 561 T IK TT T  -- NN II      s
-562 P KI PP V  -- LL PP      s
-563 K KK KK K  -- KK KR      S   helix, amphiphilic
 564 Q DD DE N  -- EE AE      S   helix
=565 S SS SS S  -- SS SS      a   helix
-566 A FG FY G  -- GG AA      i   helix
 567 L YY YY K  -- KK KK      s   helix
-568 W FW WW W  -- WW FF      i   helix
-569 F YY YY Y  -- YY YY      i   helix
-570 K QK KR S  -- -- AA      s   helix
 571 Q QE GN A  -- -- SS      s   helix
-572 M YV VV L  -- -- VI      i   helix
-573 M II IV A  -- -- VV      i   helix
-574 A KK SS S  -- -- RR      s   helix
 575 K EN RN -  -- -- CC      s   helix
 576 N NN GN -  -- -- NN      s   helix
 577 - -- -- -  -- -- GG      .
 578 - -- -- -  -- -- FF      i
 579 - -- -- -  -- -- PP      .
 580            -- -- DD      a
 581            -- -- PP      .
 582            -- -- AA      i
 583            -- -- TE      S
 584            -- -- GG      .
```

Fig. 1. Multiple alignment of the two subfamilies of the 6-phospho-strand-D-galactosidase. Underlined residues correspond to parses. In regions where the alignment has been readjusted by hand, surface and interior assignments may not correspond to those produced by the automated computer output. These should be ignored.

Subfamily a (a b c j k o):
a—(p11546) lacg_lacla 6-phospho-strand-galactosidase (EC 3.2.1.85) (strand-D-phosphogalactoside galactohydrolase). *Lactococcus lactis* (subsp. *lactis*) (*Streptococcus lactis*).
b—(p11175) lacg_staau 6-phospho-strand-galactosidase (EC 3.2.1.85) (strand-D-phosphogalactoside galactohydrolase). *Staphylococcus aureus*.
c—(p14696) lacg_lacca 6-phospho-strand-galactosidase (EC 3.2.1.85) (strand-D-phosphogalactoside galactohydrolase) (p-strand-gal) (pbg). *Lactobacillus casei*.
d—(p24240) ascb_ecoli 6-phospho-strand-glucosidase (EC 3.2.1.86). *Escherichia coli*.
e—(p26206) arbb_erwch 6-phospho-strand-glucosidase (EC 3.2.1.86). *Erwinia chrysanthemi*.
f—(p11988) bglb_ecoli 6-phospho-strand-glucosidase (EC 3.2.1.86). *Escherichia coli*.

Subfamily b (d e f g h i l m n p q r):
a—(p26208) bgla_clotm strand-glucosidase a (EC 3.2.1.21) (gentiobiase) (cellobiase) (strand-D-glucoside glucohydrolase). *Clostridium thermocellum*.
b—(p10482) bgls_calsa strand-glucosidase a (EC 3.2.1.21) (gentiobiase) (cellobiase) (strand-D-glucoside glucohydrolase) (amygdalase). *Caldocellum saccharolyticum*.
c—(p22073) bgla_bacpo strand-glucosidase a (EC 3.2.1.21) (gentiobiase) (cellobiase) (strand-D-glucoside glucohydrolase) (amygdalase). *Bacillus polymyxa*.
d—(q03506) bgla_bacci strand-glucosidase (EC 3.2.1.21) (genitobiase) (cellobiase) (strand-D-glucoside glucohydrolase) (amygdalase). *Bacillus circulans*.
e—(p22505) bglb_bacpo strand-glucosidase b (EC 3.2.1.21) (gentiobiase) (cellobiase) (strand-D-glucoside glucohydrolase) (amygdalase). *Bacillus polymyxa*.
f—(p12614) bgls_agrsp strand-glucosidase (EC 3.2.1.21) (gentiobiase) (cellobiase) (strand-D-glucoside glucohydrolase) (amygdalase). *Agrobacterium* sp. (strain atcc 21400).
g—(q00326) myro_brana myrosinase precursor (EC 3.2.3.1) (sinigrinase) (thioglucosidase). *Brassica napus* (rape).
h—(p09849) lph_rabit pos 1361 to 1926 of lactase-phlorizin hydrolase precursor (EC 3.2.1.108) (EC 3.2.1.62) (lactase-glycosylceramidase) (lph). *Oryctolagus cuniculus* (rabbit).
i—(p29092) myr3_sinal myrosinase mb3 precursor (EC 3.2.3.1) (sinigrinase) (thioglucosidase). *Sinapis alba* (white mustard).
j—(p26204) bgls_trirp noncyanogenic strand-glucosidase precursor (EC 3.2.1.21). *Trifolium repens* (creeping white clover).
k—(p09849) lph_human pos 1361 to 1927 of lactase-phlorizin hydrolase precursor (EC 3.2.1.108) (EC 3.2.1.62) (lactase-glycosylceramidase). *Homo sapiens* (human).
l—(p26205) bglt_trirp cyanogenic strand-glucosidase precursor (EC 3.2.1.21) (linamarase) (fragment). *Trifolium repens* (creeping white clover).

TABLE 1. Consensus Secondary Structure Prediction for the 6-Phospho-β-D-galactosidase Superfamily*

Strand A	009–011	Strand A	049–051	
Strand B†	014–020	Strand B†	053–060	Internal
Helix 1†	072–084	Helix 1†	095–107	Amphiphilic
Strand C?	089–093	Strand C	111–115	Amphiphilic
Act site a	095–102	Act site a	117–125	
Helix 2†	116–130	Helix 2†	138–153	Amphiphilic
Strand D†	136–140	Strand D†	159–163	Internal
Act sit b†	141	Act sit b†	164–166	
Helix 3†	158–177	Helix 3†	181–198	Amphiphilic
Strand E	182–185	Strand E	205–208	
Act sit c	184–187	Act sit c	207–209	
Helix 4†	212–226	Helix 4†	248–268	Largely internal
Strand x	242–245			
Helix x	259–273			Ambiguous alignment
Strand y	275–280		318–320	Shifted alignment
Helix 5†	286–293	Helix 5†	332–342	Amphiphilic
Strand F	323–327	Strand F	381–388	Interior
Helix y	329–339	Gap		Ambiguous alignment
Strand z	375–382	Strand z	446–452‡	Amphiphilic
Helix 6†	385–398	Helix 6†	456–469	Amphiphilic
Strand G	404–407	Strand G†	476–479	Internal
Act site d†	408–410	Act site d†	480–482	
Helix 7†	431–448	Helix 7†	497–517	Amphiphilic
Strand H†	450–454	Strand H†	521–525	Amphiphilic
Strand I†	456–459	Strand I†	527–530	Interior
Strand J†	464–467	Strand J†	535–539	Interior
Strand K†	478–482	Strand K†	548–554	Interior
Helix 8†	496–509	Helix 8†	563–576	Amphiphilic

*Assignments in the consensus model (which applies to the entire superfamily) are designated with upper case letters A–K (for β strands) and Arabic numerals 1–8 (for α helices). Strands and helices designated by "x," "y," and "z" are not part of the consensus model, and may be present in only some members of the superfamily. Assignments marked with "?" are weak within one subfamily, but confirm a stronger assignment in the other subfamily.

†Reliable assignments.

‡The multiple alignment is ambiguous; see text.

Strand H (a450–454; b521–525) is amphiphilic and confirmed in both subfamilies.

Strand I (a456–459; b527–530) is interior and confirmed in both subfamilies. It may be longer by two residues in subfamily b.

Strand J (a464–467; b535–539) is interior and confirmed in both subfamilies.

Strand K (a478–482; b548–554) is interior, well anchored, and confirmed in both subfamilies.

Finally, helix 8 (a496–509; b563–576) is amphiphilic, well anchored, and confirmed in both subfamilies.

In examining the consensus secondary structural model reported in Table 1, it is difficult not to notice the secondary structural pattern characteristic of an 8-fold α–β barrel protein. This tertiary structural hypothesis does not rest solely on pattern recognition. The model is, in fact, enforced by the active site assignments designated in Table 1. Here, β strands C, D, E, and G all must terminate near the active site of the protein, as in an 8-fold α–β barrel. While other topologies could also bring these residues together, this is our preferred tertiary structural model.

REFERENCES

1. Benner, S.A. Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. Adv. Enzyme Reg. 31:121–181, 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. and Saraste, M. SH3—an abundant protein domain in search of a function. FEBS Lett. 307–55–61, 1992.
5. Bazan, J.F. Structural design and molecular evolution of a cytokine receptor superfamily. Proc. Natl. Acad. Sci. U.S.A. 87:6934–6937, 1990.
6. Moe, G.R., Koshland, D.E., Jr. 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: 163–168.
7. Benner, S.A., Gerloff, D.L., Jenny, T.F. Predicting protein crystal structures. Science 265:1642–1644, 1994.
8. Benner, S.A., Ellington, A.D. Evolution and structural theory: The frontier between chemistry and biochemistry. Bioorg. Chem. Frontiers 1:1–70, 1990.
9. Chothia, C., Lesk, A.M. The relation between the divergence of sequence and structure in proteins. EMBO J. 5:823–826, 1986.
10. Benner, S.A., Gerloff, D.L., Jenny, T.F., Chelvanayagam,

G., Knecht, L.J., Cohen, M.A., Gonnet, G.H. Predicting conformation in proteins from homologous protein sequences. Nature Struct. Biol., submitted.

11. Jenny, T.F., Benner, S.A. Evaluating predictions of secondary structure in proteins. Biochem. Biophys. Res. Commun. 200:149–155, 1994.
12. Jenny, T.F., Benner, S.A. A prediction of the secondary structure of the pleckstrin homology domain. Proteins: Struct. Funct. Genet. 20:1–4, 1994.
13. Musacchio, A., Gibson, T., Rice, P., Thompson, J., Saraste, M. The pH domain—a common piece in the structural patchwork of signaling proteins. Trends Biochem. Sci. 18: 343–348, 1993.
14. Gerloff, D.L., Jenny, T.F., Knecht, L.J., Benner, S.A. A secondary structure prediction of the hemorrhagic metalloprotease family. Biochem. Biophys. Res. Commun. 194: 560–565, 1993.
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. Technical Report 154, Departement Informatik, March, 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. Benner, S.A., Badcoe, I., Cohen, M.A., Gerloff, D.L. Bona fide prediction 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.
19. 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. Enzyme Reg. 31:121–181, 1991.
20. Wacker, H., Keller, P., Falchetto, R., Legler, G., Semenza, G. Location of the two catalytic sites in intestinal lactase-phlorizin hydrolase. J. Biol. Chem. 267:18744–18752, 1992.