PROTEINS: Structure, Function, and Genetics 21:1–10 (1995)

PREDICTION REPORT

Predicted Secondary and Supersecondary Structure for the Serine–Threonine-Specific Protein Phosphatase Family

Thomas F. Jenny, Dietlind L. Gerloff, Mark A. Cohen, and Steven A. Benner
Department of Chemistry, Swiss Federal Institute of Technology (E.T.H.), CH-8092 Zurich, Switzerland

ABSTRACT A bona fide consensus prediction for the secondary and supersecondary structure of the serine–threonine specific protein phosphatases is presented. The prediction includes assignments of active site segments, an internal helix, and a region of possible 3_{10} helical structure. An experimental structure for a member of this family of proteins should appear shortly, allowing this prediction to be evaluated.

Key words: protein structure prediction, protein phosphatase, evolution

INTRODUCTION

Four developments have characterized recent work to develop useful approaches for converting sequence data into models of the conformation of a polypeptide chain. First, an explosive growth of sequence databases has permitted conformational models to be built from alignments of homologous sequences rather than from single sequences.[1–4] Second, the explosive growth of sequence databases has enabled the development of new, often highly detailed, models of how protein sequences diverge under functional constraints.[5–8] These have allowed the development of prediction tools that transcend simple averaging of classical predictions over a set of aligned homologous protein sequences, in particular, those that extract conformational information from patterns of conservation and variation within a multiple alignment.[9–11] The combination of abundant sequence data and detailed evolutionary models has underlain a third development, where tertiary structural information is explicitly introduced early in a structure prediction, providing a way free of the vicious cycle that arises because secondary structure is examined before tertiary structure is predicted, but tertiary interactions are the strongest determinants of secondary structure. Fourth, bona fide predictions, those that are announced before an experimental structure is available, have again had their value recognized as a key part of an effort to develop useful prediction tools, especially as a way to identify and avoid biases that inevitably compromise methods tested using a database known at the time the tools were developed.[12]

The primary disadvantage of bona fide predictions as research tools is that such predictions are made one at a time, and must await confirmation by subsequently determined experimental structures. Thus, it is difficult to generate rapidly a large number of predictions that can permit a statistically significant test of a prediction method. This implies that efforts to use bona fide predictions as a research tool must be sustained over a substantial period of time to allow accumulation of many predictions. Especially important in these efforts are venues (such as this) where such predictions can be published, as these permit a record to be established of the scope and success of prediction methods, creating an accountability that itself serves the research effort.

To date, over a dozen bona fide predictions have been made in four laboratories using tools that include an analysis of conservation and variation in homologous protein sequences. With the recent prediction of the structure of protein tyrosine phosphatase[13] by Livingston and Barton,[14] eight of these can now be evaluated using one or more subsequently determined experimental structures.[15] While no finite number of successful predictions can "prove" a method, these predictions have shown that these methods are transferrable from laboratory to laboratory, can provide some remarkably accurate secondary structure predictions, and can be the first step in the generation of tertiary structural models from secondary structural models. Particularly noteworthy is the first prediction contest that placed these tools in competition with the most advanced neural networks and classical methods averaged

Received July 12, 1994; revision accepted September 19, 1994.
Address reprint requests to Dr. Steven A. Benner, Department of Chemistry, E.T.H., CH-8092 Zurich, Switzerland.

over a set of aligned protein sequences.[16] In this contest, predictions based on an explicit analysis of patterns of conservation and variation assigned more residues correctly than either the neural networks or classical tools. More important, the number of misassignments that confused α-helices and β-strands in the prediction was vanishingly small, suggesting that the new tools yield secondary structural models especially suited as the starting point for tertiary structural modeling.[17] Again, however, this was a single case, and many more prediction contests must be arranged before the relative merits of different methods can be evaluated in any general way.

Recently, prediction efforts have focused on proteins and domains involved in signal transduction, including the Src homology 1 (SH1) domain, a protein kinase,[18] the Src homology 2 (SH2) domain,[19] a protein unit that binds peptides containing phosphotyrosine, the Src homology 3 (SH3) domain,[20,21] a protein unit that may bind proline-rich peptide sequences, the protein tyrosine phosphatases,[14] and the pleckstrin homology (PH) domain,[22,23] a domain with unknown function.[24,25] We turn now to the protein serine–threonine phosphatases,[26] a family where no experimental structure is known, but where an experimental structure may be imminent.[27] We report here a predicted consensus secondary structure for this protein family.

METHODS

Sequences of protein serine phosphatases were extracted from entries in SwissProt 27 and a multiple alignment built by DARWIN.[28] The multiple alignment was then adjusted by hand (Fig. 1). In particular, the nonaligning segments at the beginning and end of the multiple alignment, presumably representing noncore regions, were removed, and the insertion in human calcium-dependent calcineurin A (accession number P16298; sequence p, between alignment positions 073 and 074), presumably representing an intron-derived sequence, was removed. Surface and interior residues were then assigned by an automated computer program, following a design described in detail elsewhere.[8] The multiple alignment was then parsed (separated into segments that form independent secondary structure) by analysis of gaps in the multiple alignment, identifying conserved Pro and Gly residues adjacent to surface using procedures, and identifying parsing strings, using heuristics described elsewhere.[8,9,16,18] Positions that have conserved functionality in the context of strings of consecutive or nearly consecutive active site assignments were assigned to the active site, as discussed elsewhere.[9,16,18] Many of the automated computer programs used in this work are available on a server via electronic mail at the address cbrg@inf.ethz.ch (send a one word message "help" for instructions).

Secondary structures were then assigned to individual parsed segments of the multiple alignment from patterns of conservation and variation in the various assignments using an automated computer program that detects periodicity in these patterns. A 3.6 residue periodicity in interior and surface assignments was designated a surface helix, while alternating periodicity in these assignments, was designated a surface beta strand, as discussed elsewhere.[9,16,18] Short ($3 \leq 6$ residues) segments of interior assignments were designated as β-strands, while heuristics designed to distinguish between internal helices and internal strands (see below) were applied to longer segments. Segments of consecutive or nearly consecutive residues assigned near the active site assignments were assigned as "active site regions." In these regions, patterns of amino acid substitution are dominated by constraints relating to catalytic function, which often obscure patterns that reveal secondary structure.

The secondary structures predicted here are presented in a way that facilitates their use as the starting point for tertiary structure modeling. Important in this presentation is recent work that defines the limits to which a consensus secondary structural model built from a family of proteins can represent the conformation of any individual member of the family.[12,29,30] In building tertiary structural models from predicted secondary structure, the ϕ and ψ angles in segments assigned as helices or strands are held rigidly, while those in the coil segments are permitted to vary. Therefore, the most useful secondary structure model is presumably the one that assigns helical or strand conformations to the cores of these secondary structural units, those likely to be found in all members of a protein family, and leave flexible those regions that are not likely to be found in all family members.

Supersecondary structure, defined here to describe the relative orientation of consecutive secondary structure units, was predicted based on orientation of predicted secondary structure units with respect to active site assignments, and searching for compensatory covariation, as described elsewhere.[18]

RESULTS AND DISCUSSION

The secondary structure prediction is summarized in Table I, based on the multiple alignment, surface, interior, parsing and active site assignments recorded in Figure 1. Most of the predicted secondary structural elements were obtained by automated procedures that implement the prediction heuristics outlined in 1989 and 1991.[9,18] Certain of the secondary structure segments are canonical,[18] and should be highly reliable. In particular, helices (003–019), (119–130), and (236–243) are reliably assigned (Fig. 2), as are strands (059–062), (090–093), (135–139), (143–146), (175–180), and (228–232). A helix can also be assigned to segment (101–116), provided that

a weak surface assignment at position 107 is accepted.

The remaining assignments of secondary structure were not routine. For example, the interior and surface assignments made for segment (154–166) display a perfect 3.0 residue periodicity. This is expected either for an α-helix whose contacts with the remainder of the protein precess around the helix axis, or for a 3_{10} helix. 3_{10} helices are rarer than α-helices, but are not unknown in protein structures.[31,32] This is, however, the first time such patterns have been observed in the context of a bona fide prediction. We have listed the helix as being of the 3_{10} type to encourage more effort to investigate patterns of divergent evolution in such "nonstandard" secondary structures.

Internal helices are also difficult to find using methods that assign secondary structure based on periodicity in interior and surface assignments. An internal helix was missed in the protein kinase prediction,[18] while another was correctly identified in the hemorrhagic metalloprotease family.[16] In the protein phosphatases, an internal helix is assigned to positions 075–086. This assignment is based on two criteria: (1) the length of consecutive interior assignments compared with the overall size of the protein, and (2) the pairing of the helix with a β-strand following it (090–093), where the short parsing element (087–089) and the active site assignments within positions 063–074 and 094–098 require the two structural units to be antiparallel and roughly equivalent in length. This is noteworthy as an example where the assignment of secondary structure strongly relies on a tertiary structural hypothesis, and it will be interesting to learn whether it is correct.

The most problematic assignments for any method for predicting secondary structure are those made for segments near the active site. Parameters derived from a statistical treatment of protein sequences as a whole are the least representative in these regions. Further, near the active site, patterns of variation and conservation are dominated by the demands of catalytic function, which often obscure patterns that might indicate secondary structure type. Previous prediction efforts have shown how problematic these regions are in other protein families,[29] and similar regions are encountered in the protein phosphatase family. For example, the segment 029–050 includes a β-strand assigned to segment (030–034) and a short α-helix assigned to segment (043–049) separated by an active site region (035–042) with poorly defined secondary structure. The α-helix might be extended into the active site region. However, the conservation evidently arising from catalytic function precludes reliable tertiary structural assignments in this region that might secure this extension.

Further, the helix assigned to segment (246–262) contains a conserved tripeptide RxH that is plausibly (but not definitely, see below) placed at the active site. The segment displays convincing 3.6 residue periodicity (Fig. 2) if residue 254 is assigned to the surface. To observe this periodicity requires, however, assignment of a conserved R (251) to the surface and a conserved H (253) and a conserved G (259) to the inside. Further, the DG element at positions (258–259) is a weak parsing element. Thus, a second, weaker, assignment separates this segment into two shorter elements separated by an active site coil. In tertiary structure modeling, this alternative assignment must be considered.

Finally, the final segment (274–299) contains a conserved SAPNY that is also plausibly placed near the active site. The segment (274–278) might canonically be assigned as a β-strand, as is segment 290–295. This leaves, however, segment 279–289 problematic. Dipeptide parses at positions (281–282), (284–285), (288–289), and (291–292) suggest that this segment is a coil. However, the first is involved in a putative active site assignment, making the assignment lower in reliability.

While active site assignments are sources of ambiguity when assigning secondary structure, they make key contributions when building a tertiary structural model (Fig. 3). Several long, conserved, functionalized peptide segments are convincingly assigned to the active site. First, the segments (035–042), (063–074), (094–098), and (174–152) are clearly present in the active site. The isolated conserved functionalized residues at positions 83, 111, 177, 182, and 226 are not. The segment (279–283) is also assigned to the active site, even though it contains an unusual pattern of functionalization that suggests a noncatalytic role. Finally, the segment 251–253 is also assigned at the active site, primarily based on the conservation of His-253.

Based on these active site assignments and patterns of compensatory covariation, the protein kinase structure can be divided into several supersecondary structural elements (Fig. 3). For example, both strand (030–034) and strand (059–062) end at the active site. They are joined by helix (043–049), with the loop connecting helix (043–049) and the following strand not lying at the active site. This forces these three secondary structural elements into a β–α–β segment where the two β-strands are part of a parallel β-sheet. With only a few exceptions, this supersecondary structural element is right handed. Further, refinement of the secondary structural model in light of this tertiary structural modeling suggests that the helix (043–049) probably includes additional amino acids at the end, most probably residues (040–042) and (050). Together, these considerations support a good model of the conformation of these 35 amino acids.

Similarly, internal helix (075–086) starts near the active site, while the following strand (090–093)

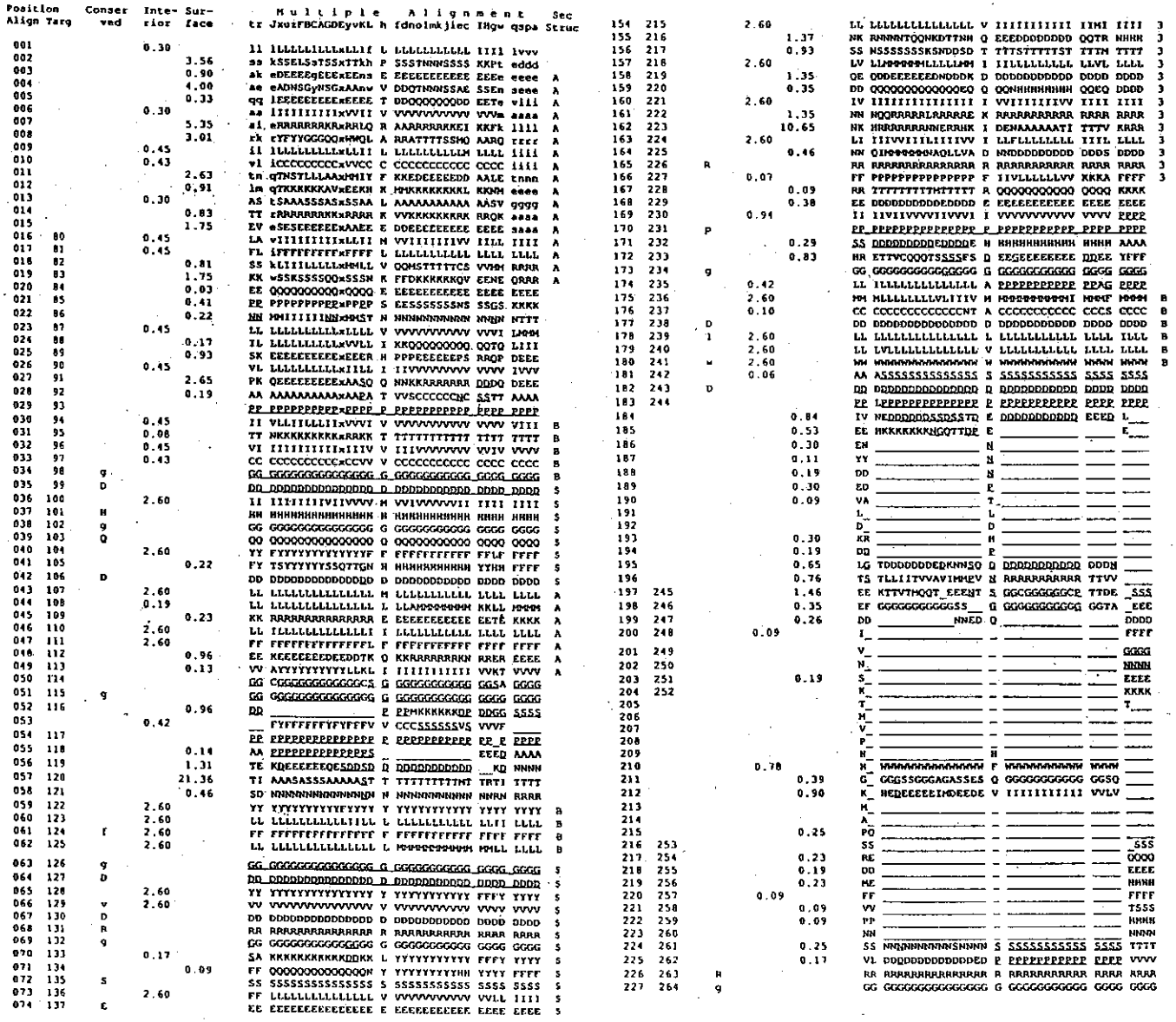


Fig. 1. Consensus secondary structure prediction for the protein serine–threonine phosphatases. Positions in the alignment are numbered consecutively. Target position numbers refer to the human calcineurin A2 sequence. Conserved residues bearing functional groups are given in upper case letters; others in lower case letters. Numbers in surface and interior assignments increase qualitatively with increasing strength of the assignment. Index numbers in the multiple alignment refer to sequences from SwissProt 27 as listed below. Parsing strings are underlined, with dipeptide parses chosen to be Gly or Pro adjacent to Gly, Pro, Ser, Asp, or Asn. Tripeptide and higher parsing strings are consecutive (Gly Pro Ser Asp Asn) residues. Lower case letters in the sequences indicate residues not aligned by DARWIN. Lower case x in the alignment indicate sequences that are fragments. Secondary structure: A, α-helix; B, β-strand; $, active site segment; 3, 3_{10} helix.

a—(P16299) protein phosphatase 2B (calmodulin-dependent calcineurin A2), *Homo sapiens*.

b—(P29349) protein–tyrosine–phosphatase (EC 3.1.3.48), *Drosophila melanogaster*.

c—(P23635) protein phosphatase PP2A-1, *Schizosaccharomyces pombe*.

d—(P23594) protein phosphatase PP2A-1, *Saccharomyces cerevisiae*.

e—(P23636) protein phosphatase PP2A-2, *Schizosaccharomyces pombe*.

f—(P23595) protein phosphatase PP2A-2, *Saccharomyces cerevisiae*.

g—(P32345) yeast protein phosphatase, *Saccharomyces cerevisiae*.

h—(P32838) protein phosphatase PP2A, *Saccharomyces cerevisiae*.

i—(P05323; P13197) protein phosphatase PP2A-α, *Homo sapiens*, rabbit, pig, and bovine.

j—(P13353) rat protein phosphatase PP2A-α, rat

k—(P11082) protein phosphatase PP2A-β, *Homo sapiens*, rat.

l—(P11493) protein phosphatase PP2A-β, (fragment), pig.

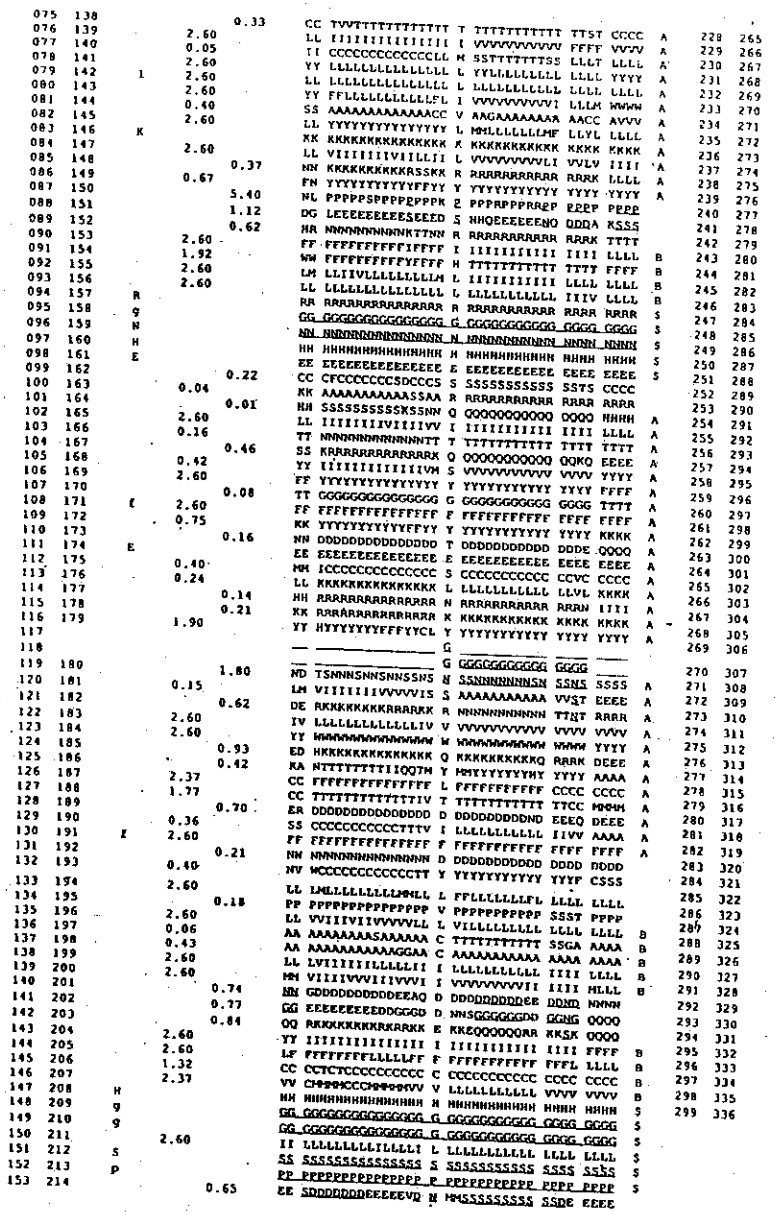
m—(P11611) protein phosphatase PP2A-β, rabbit.

n—(P23778) protein phosphatase PP2A (fragment), rape.

o—(P23696) protein phosphatase PP2A, *Drosophila melanogaster*.

p—(P16298) protein phosphatase 2B (calmodulin-dependent calcineurin A1), *Homo sapiens*.

q—(P20652; P12816) protein phosphatase 2B (calmodulin-dependent calcineurin A), rat and mouse.



f—(P23287) protein phosphatase 2B (calcineurin A1), *Saccharomyces cerevisiae.*

s—(P20651) protein phosphatase 2B (calmodulin-dependent calcineurin A), rat.

l—(P14747) protein phosphatase 2B (calcineurin A2), *Saccharomyces cerevisiae.*

u—(P13681) protein phosphatase PP1, *Schizosaccharomyces pombe.*

v—(P23733) protein phosphatase, *Trypanosoma brucei.*

w—(P20604) protein phosphatase PP1-1, *Saccharomyces cerevisiae.*

x—(P23880) protein phosphatase PP1, *Schizosaccharomyces pombe.*

y—(P23734) protein phosphatase, *Trypanosoma brucei.*

z—(P32598) protein phosphatase PP1-2, *Saccharomyces cerevisiae.*

A—(P08129; P22802; P20653) protein phosphatase PP1-α, *Homo sapiens,* rabbit, and rat.

B—(P12982) protein phosphatase PP1-87B, *Drosophila melanogaster.*

C—(P08128) protein phosphatase PP1-β, rabbit.

D—(P30366) protein phosphatase PP1, *Arabidopsis thaliana.*

E—(P23777) protein phosphatase PP1 (fragment), rape.

F—(P20654) protein phosphatase PP1, *Aspergillus nidulans.*

G—(P22198) protein phosphatase PP1, maize.

H—(P33172) protein phosphatase PP-X, *Homo sapiens*

I—(P11084) protein phosphatase PP-X, rabbit.

J—(P11612) protein phosphatase PP-Y, *Drosophila melanogaster.*

K—(P26570; Q00979) protein phosphatase PP-Z1, *Saccharomyces cerevisiae.*

L—(P32945) protein phosphatase SAL6, *Saccharomyces cerevisiae.*

TABLE I. Secondary Structure Prediction for the Serine/Threonine Specific Protein Phosphatases

Segment	Structure	Comments
003–019	Helix	Reliable, perfect amphiphilicity
021–022	Parse	PN, PS, string of 5 surface assignments
023–026	Coil	In some protein family members, possibly an edge strand
027–029	Parse	SPP, DSP, PXP, NP, SP, conserved P
030–034	Beta	Near active site region 035–042
035–042	Active site	Conserved D, HGQ, D
043–049	Alpha	Near active site region 035–042
050–058	Parse	Gap, PPSSN, GGDP, GGSP
059–062	Strand	Reliable
063–074	Active site	
075–086	Helix	Internal helix
087–088	Parse	PS, PN, PD
090–093	Strand	Reliable, central in a sheet
094–098	Active site	Conserved RGNHE
101–116	Helix	Reliable, residue 107 assignment must lie on the surface
117–120	Parse	Gap, GGNS, GNS
119–130	Helix	Reliable
134	Parse	Weak, partly conserved P, break in amphiphilicity of preceding helix
135–139	Strand	Reliable
140–142	Parse	NNS, DG
143–146	Strand	Reliable
147–152	Active site	Conserved HGGXSP
151–153	Parse	SPS, SPD
154–166	Helix	3_{10} geometry suggested by patterns of conservation and variation
169–174	Parse	PDSG, GGP, PP, GP
175–180	Strand	Reliable
181–225	Parse	SDPSGD, NNNP, PG, SP; surface loop distant from active site
228–232	Strand	Reliable, core of β-sheet
233–235	Parse	GPD, GSD
236–243	Helix	Reliable, surface
243–245	Parse	NNG, NNN
246–262	Helix	Middle passes near active site
263–273	Parse	Gap, DGG, PS
274–278	Strand	Unusual patterns of conserved functional groups
279–283	Active site	Conserved SAPNY
284–287	Coil	
288–289	Parse	GN
290–295	Strand	
297–299	Parse	DDS, end of core alignment

ends at the active site. These are therefore assigned as a supersecondary structural unit formed by two antiparallel structural units, a helix and a strand. The most plausible structural model assembles this supersecondary structural unit with the preceding supersecondary structural unit to yield a parallel β-sheet. These assignments suggest that this part of protein phosphatase forms an α–β parallel sheet fold, well known in structural biology. Assignments of the remaining supersecondary structural elements are less reliable, and await confirmation by a detailed covariation analysis.

One application of secondary structural models is to detect distant homology between protein families that is not established by statistically significant sequence similarities. In this application, the predicted secondary structure from one family of proteins is aligned either with a predicted secondary structure or an experimental structure from another protein family that shares poor sequence similarity. For example, in its search of the database, DARWIN identified and aligned six positions (169–174) of the serine–threonine protein phosphatases with a tyrosine-specific protein phosphatase (a PPSHAP sequence), suggesting that the tyrosine and the serine–threonine protein phosphatase families might be homologous despite their apparent mechanistic differences. Superimposition of the secondary structural elements of the two protein families, one predicted and the other experimental, clearly shows that these two families are not homologous, however. This suggests that this particular sequence similarity in this one tyrosine phosphatase arose by convergent, not divergent, evolution.

Finally, several distinct mechanistic classes of enzymes are known that catalyze the transfer of a

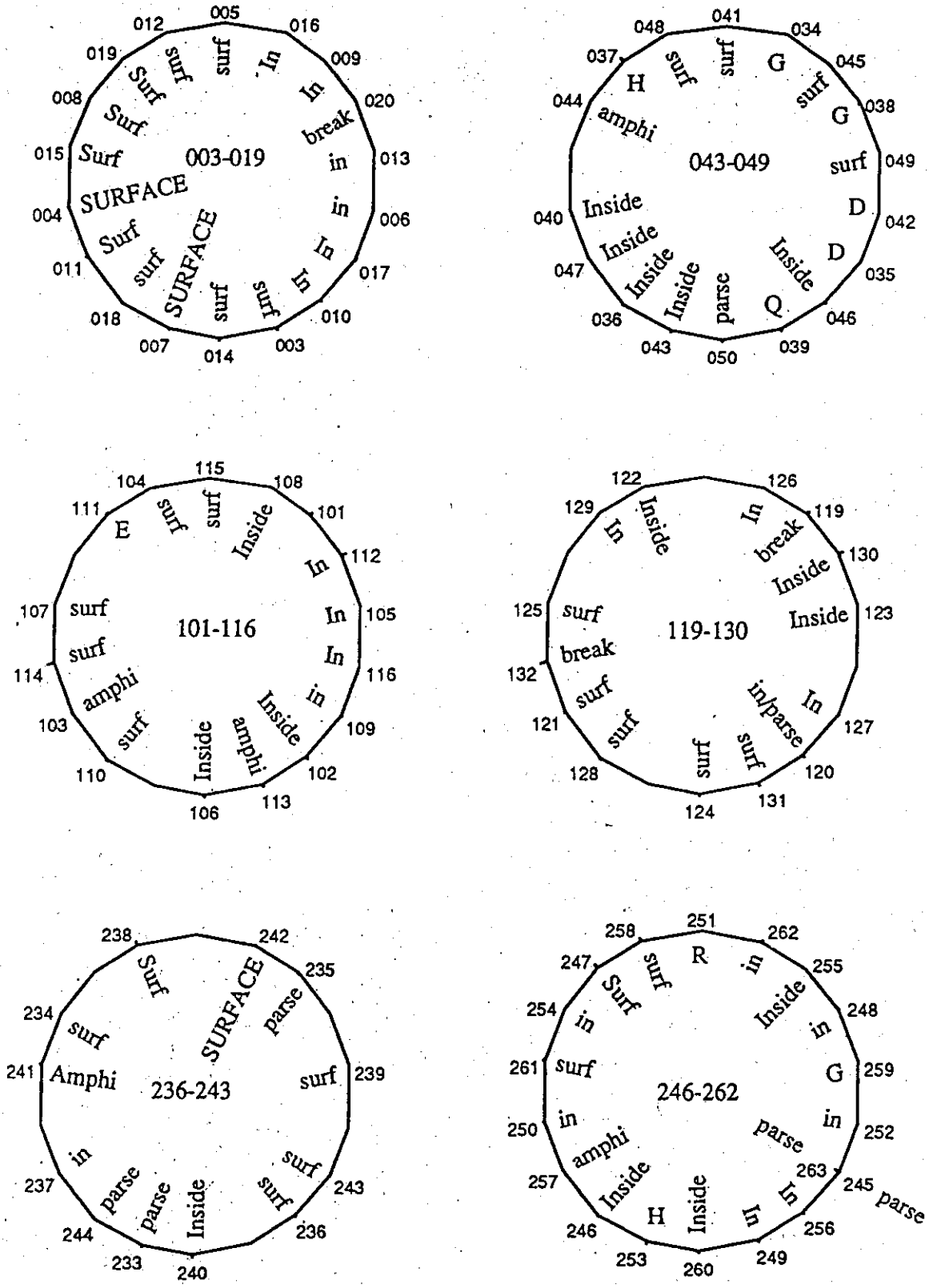


Fig. 2. Helix wheels for the serine–threonine protein phosphatases.

26. Cohen, P. The structure and regulation of protein phosphatases. Annu. Rev. Biochem. 58:435–508, 1989.
27. Barford, D., Keller, J. C. Cocrystallization of the catalytic subunit of the serine/threonine specific protein phosphatase 1 from human in complex with microcystin LR. J. Mol. Biol. 235:763–766, 1994.
28. Gonnet, G. H., Benner, S. A. Computational biochemistry research at ETH. Technical Report 154, Departement Informatik, March, 1991.
29. Gerloff, D. L., Jenny, T. F., Knecht, L. J., Gonnet, G. H., Benner, S. A. The Nitrogenase MoFe Protein: A secondary structure prediction. FEBS Lett. 318:118–124, 1993.
30. Jenny, T. F., Benner, S. A. Evaluating predictions of secondary structure in proteins. Biochem. Biophys. Res. Commun. 200:149–155, 1994.
31. Pavone, V., Di Blasio, B., Santini, A., Benedetti, E., Pedone, C., Toniolo, C., Crisma, M. The longest, regular polypeptide 310 helix at atomic resolution. J. Mol. Biol. 214: 633–635, 1990.
32. Barlow, D. J., Thornton, J. M. Helix geometry in proteins. J. Mol. Biol. 201:601–619, 1988.
33. Kim, E., Wyckoff, H. W. Reaction mechanism of alkaline phosphatase based on crystal structures. J. Mol. Biol. 218: 449–464, 1991.
34. Benner, S. A., Glasfeld, A., Piccirilli, J. A. Stereospecificity in enzymology: Its place in evolution. Topics Stereochem. 19:127–207, 1989.
35. Russell, R. B., Barton, G. J. The limits of protein secondary structure prediction accuracy from multiple sequence alignment. J. Mol. Biol. 234:951–957, 1993.