Int J Peptide Protein Res **44**, 1994, 239–244
Printed in Belgium – all rights reserved

INTERNATIONAL JOURNAL OF
PEPTIDE & PROTEIN RESEARCH
ISSN 0367-8377

Secondary structure prediction from multiple sequence data: blood clotting factor XIII and *Yersinia* protein-tyrosine phosphatase

CRAIG D. LIVINGSTONE and GEOFFREY J. BARTON

Laboratory of Molecular Biophysics, University of Oxford, Oxford, UK

Received 16 November 1993, accepted for publication 16 February 1994

Predictions of protein structure are best tested without prior knowledge of the protein three-dimensional structure. Three-dimensional atomic models will soon be determined by X-ray crystallography for the α-subunit of human blood clotting factor XIII and members of the family of protein tyrosine specific phosphatases. Accordingly, we here present secondary structure predictions for each of these proteins. The secondary structure predictions were generated from aligned sets of protein sequences. This technique has previously provided reliable predictions for the Annexins and the SH2 domains. The factor XIIIα prediction contains 39 regions predicted in strand conformation (34% of the protein) with only 3 helices (4%). The protein tyrosine phosphatases have 12 predicted strands and 5 helices (30 and 17%, respectively). We expect greater reliability from regions of alignments that show clear patterns of residue conservation (61% of factor XIIIα and 57% of the protein tyrosine phosphatases). The aligned protein tyrosine phosphatases show two regions (L39–L80 and I138–E253) with clear patterns of residue conservation separated by a region of variable amino acid composition. We suggest this indicates that the tyrosine phosphatase fold comprises two domains separated by an exposed linker. Potential phosphate binding sites are identified in the protein tyrosine phosphatases.

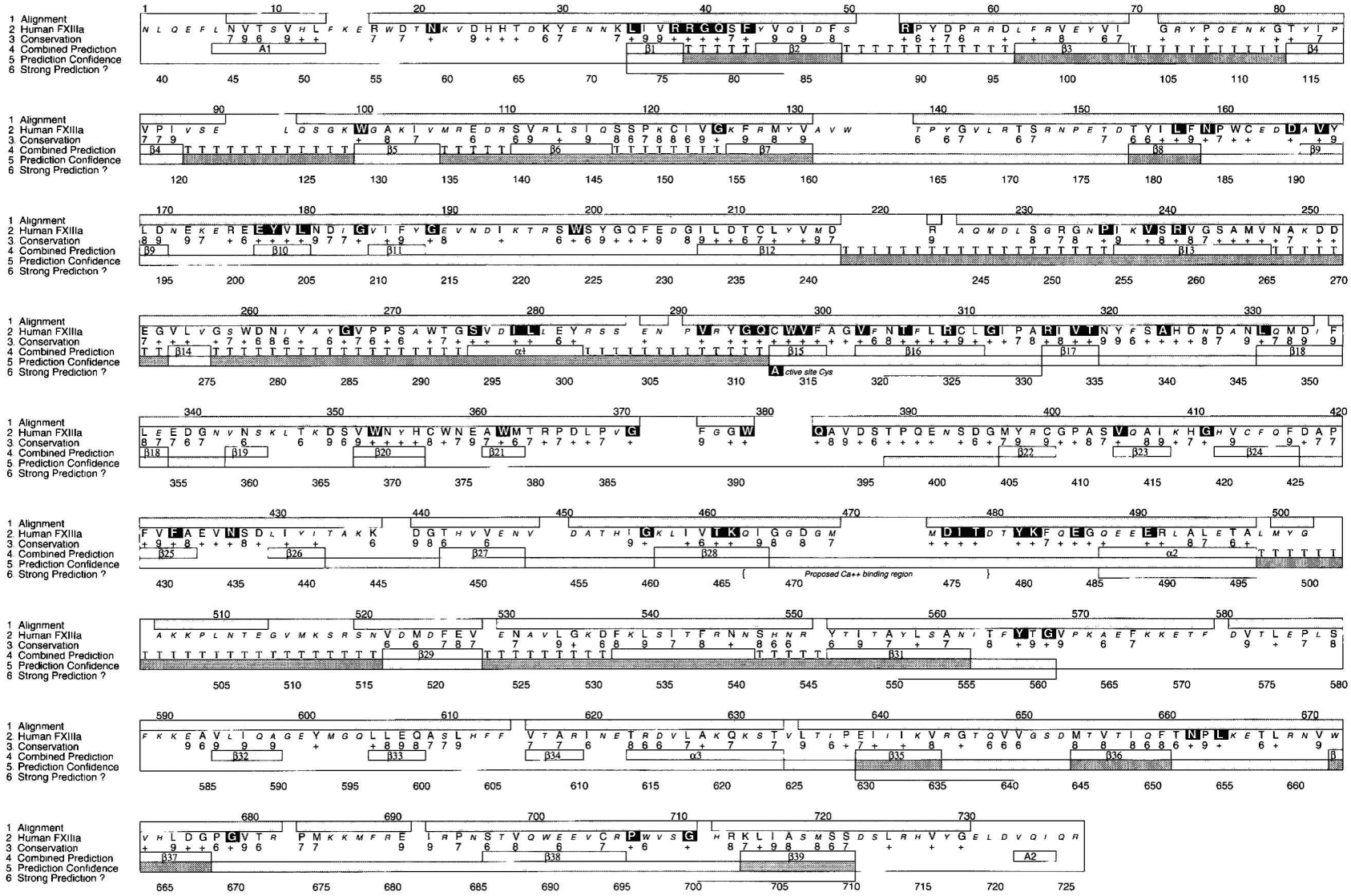
Key words: factor XIII; phosphatase; protein sequence alignment; structure prediction; transglutaminase

The average accuracy for a protein secondary structure prediction has long been ≃63% for 3-states; α-helix, β-strand, coil (1). This value has recently increased by up to 9% through the use of information from accurately aligned protein families (2–5). Whereas protein structures that are already known may be used for the development of prediction techniques, the most unbiased test of any prediction method is when applied 'blind' to a protein of unknown three-dimensional structure. Such blind tests of prediction from alignments have now been made for a number of proteins, including tryptophan synthase (6), protein kinases (7), annexins (8, 9), SH2 domains (2, 10) and SH3 domains (11), often with a high degree of accuracy (4).

As a consequence of our accurate prediction of the structure of the SH2 domain (2), we have been asked to apply our techniques to two families of proteins where a tertiary structure for a family member is close to being determined by X-ray crystallography. Accordingly, in this paper we present a summary of the secondary structure prediction for the α subunit of factor XIII (FXIIIα) (12) from an alignment of 11 sequences, and a prediction for the *Yersinia* protein-tyrosine-phosphatase (PTPase) (13) from 73 sequences. Both predictions present difficult challenges for our method, factor XIIIα because of the limited number of related sequences and their length (730 residues) and high conservation, and PTPases due to a large unconserved region (A81–T137). However, several regions of both proteins give clear predictions (61% of FXIIIα and 57% of the PTPases) and we draw particular attention to these segments. At the time this paper was submitted, the secondary and tertiary structures of the two proteins were unknown to us. When crystal structures become available, the comparison of prediction with reality may be used to guide the development of improved prediction methods.

METHODS

The prediction method applied to both protein families is essentially the same as that used for accurate structure predictions of the annexins and SH2 domains (2, 8, 10). The protocol has the following steps (1) multiple alignment of all related sequences that can be aligned reliably (14); (2) analysis of residue conservation and the location of gaps; (3) prediction on each sequence by three secondary structure (15–17), and two turn


1 Alignment
2 Human FXIIIa
3 Conservation
4 Combined Prediction
5 Prediction Confidence
6 Strong Prediction ?
Active site Cys
Proposed Ca++ binding region

prediction (18, 19) algorithms; (4) combination of the information from 2 and 3 to yield a summary prediction.

The combination in step 4 is performed by inspection starting with the assignment of surface loops, then filling in the regions between loops with α-helix or β-strand. At this stage we recognise that although all secondary structure predictions are subject to error, it is possible to assign the secondary structure of some regions with greater confidence than others. We apply the following hierarchy when assigning confidence to each predicted region: predicted loop (where insertions/deletions occur) > loop (conserved Gly/Pro/hydrophilic) > surface helix (with clear hydrophobic patterns) ≥ surface strand (with clear hydrophobic patterns) > buried strand (short run of conserved hydrophobic residues) > helix (no clear conserved hydrophobic pattern) > strand (no conserved pattern). Conservation patterns were identified with the aid of the AMAS program (20), while secondary structure predictions by the algorithm of Zvelebil *et al.* (5) were used to help resolve ambiguous predictions.

Figures 1 and 2 provide prediction summaries for factor XIIIα and *Yersinia* PTPase. A full multiple sequence alignment is available from the authors.

RESULTS AND DISCUSSION

FXIIIα is catalytic domain of a plasma transglutaminase and is the last enzyme utilised in the classical blood clotting cascade (21). The activity of Factor XIII is centred on a cysteine thiol (C313) which mediates a calcium-dependent reaction between the γ-carboxamide of glutamine and any primary amine, forming crosslinks between the fibrin monomers and other proteins which form a blood clot (12, 22, 23). FXIIIα shows sequence similarity to the tissue and keratinocyte transglutaminases, forming a family of enzymes whose function is to stabilise structures, making them resistant to chemical proteolysis and physical damage (22).

The striking feature of our FXIIIα secondary structure prediction (Fig. 1) is the predominance of β-strand over helix (strand 34%, helix 4%, loop 62%). There are three regions (61% of the FXIIIα sequence) where the patterns of residue conservation suggest that the prediction will be reliable (73–331, 486–561, 630–710). In the remaining sections (N–72, 332–485, 562–629, 711–C), assignment to a particular secondary structure type is difficult because the sequences show no clear patterns of conservation. The lack of clear conservation patterns can be due either to extremely low (few residues conserve physico-chemical properties, *e.g.* V562–P629), or high conservation (all residues conserve properties, *e.g.* A332–D437) between the sequences.

The regions L73–A331 and E630–S710 show clear loop predictions around sites of insertion and deletion (*e.g.* loop β2–β3). Other loops are predicted for regions showing low conservation and containing polar or turn-promoting residues. Strand conformation is predicted for sections with alternately conserved residues (*e.g.* β7) and for conserved hydrophobic regions (*e.g.* β1). α1 shows a clear pattern of conserved residues consistent with helical conformation. The active site cysteine (C313) is located at the *N*-terminal end of a series of three predicted strands (β15, β16 and β17) broken by loops at A317/G318 and following G328.

The most reliably predicted region (Q486–G561) includes a clear loop prediction between α2 and β29, a region of low conservation containing gaps and likely to be a surface loop. In FXIIIα, this contains the thrombin inactivation site (K512). Loops are reliably predicted between the remaining units of regular structure in this region. α2 has a pattern of conserved polar residues consistent with the conservation of one face of an α-helix. β29 predicts weakly as helix but is assigned to strand owing to the presence of conserved valines.

The calcium binding site proposed between Q467 and D478 (12), which shows some sequence similarity to the calmodulin high-affinity calcium binding site, occurs in one of the regions for which prediction confidence is low. The predicted conformation is loop, although there is some indication of helical propensity between M474 and T477 by both the combined prediction method and that of Zvelebil *et al.*

A secondary structure prediction of FXIIIα has previously been performed, by Takahashi *et al.*, although

FIGURE 1

Summary of the multiple alignment and structure prediction of 10 transglutaminases and the α subunit of human factor 13 (FXIIIα). Numbering at the top of each strip of the diagram indicates the position in the alignment. Numbering at the bottom shows the position relative to human FXIIIα. (1) Alignment: Shaded regions show sites where no insertions or deletions occur in the alignment. (2) Human FXIIIα: The sequence is shown with positions sharing identical amino acids with all other sequences in the alignment in white on a black background. Positions where at least six physico-chemical properties are shared with the remainder of the alignment are shown in plain type, while those sharing less than six properties are shown in small italics. (3) Conservation: Numbers indicate the number of properties shared by the amino acids at each sequence position in the remainder of the alignment (excluding FXIIIα) out of a maximum of 10 (sequence identity, '+'). Values less than six are omitted. (4) Combined prediction: The secondary structure prediction consists of a number of blocks identified as regular secondary structure (β-β-strand/extended structure, α-helix) separated by gaps indicating predicted loops. Positions where loops can be assigned with confidence are shown by 'T'. (5) Confidence: Dark shading indicates high confidence in the prediction, light shading indicates intermediate confidence and absence of shading indicates low confidence. (6) Reliable prediction: The regions where the patterns of residue conservation suggest that the prediction will be reliable are indicated, together with the location of the active site (C313) and the proposed calcium binding region.

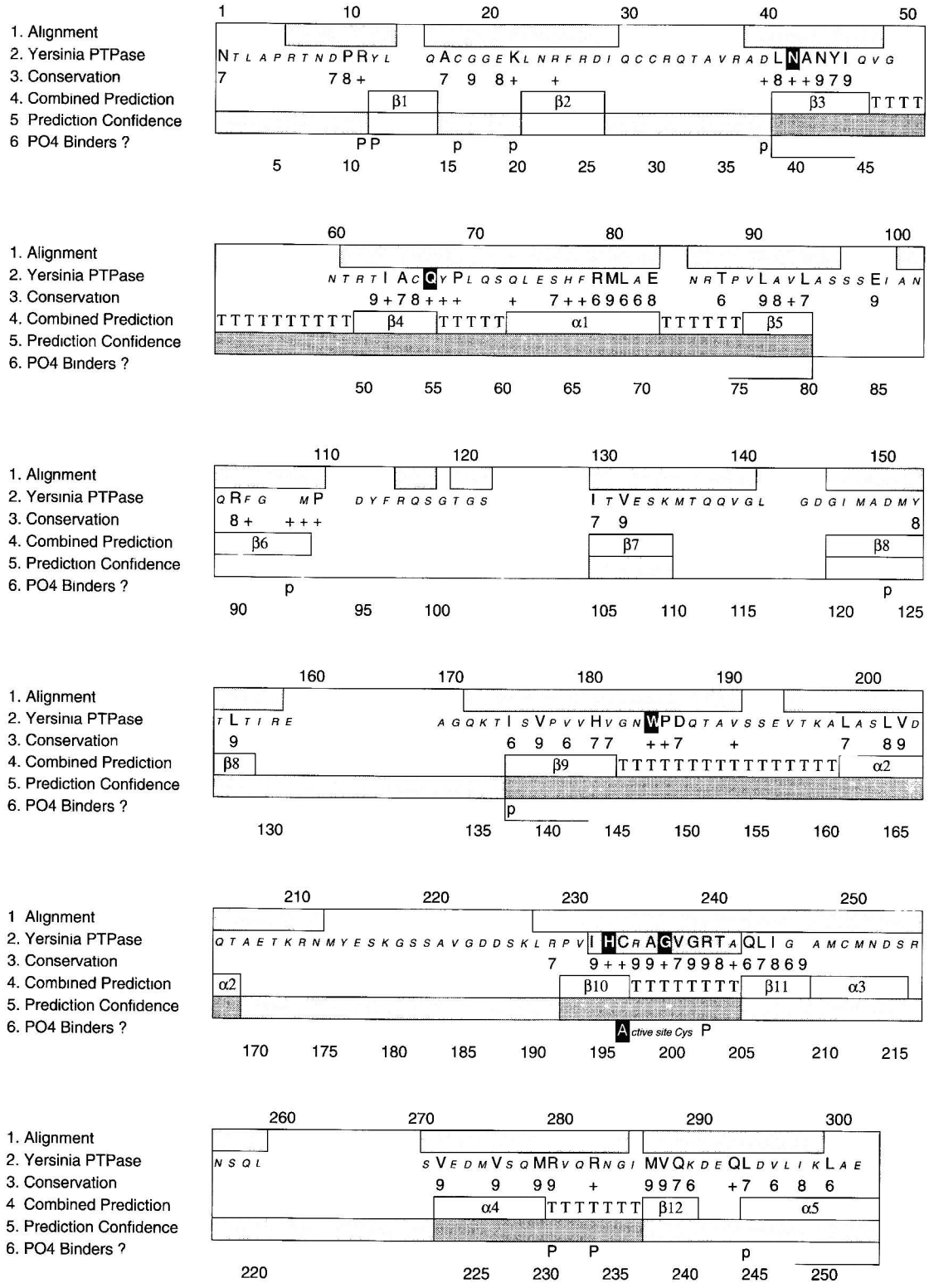


FIGURE 2

Summary of the multiple alignment and structure prediction for 73 protein tyrosine phosphatases (PTPases). The representation is similar to that for Fig. 1. Line 2 shows the sequence of *Yersinia* PTPase. Line 6 (reliable prediction) includes additional markers for potential phosphatase binding sites (*i.e.* R or Y) conserved across all sequences (P) and across all sequences excluding *Yersinia* PTPase (p). The location of the active site (C197) is indicated.

based only on a single sequence (12). This shows a similar proportion of the sequence in strand conformation (36%), but predicts significantly more helix (29%).

Changes in the phosphorylation of tyrosine residues, regulated by the opposing actions of the protein tyrosine kinases and the PTPases, have been identified as a key method for control of many cellular processes (24, 25). PTPase activity resides in a conserved domain of between 250 and 300 residues which may appear as a single copy or as a tandem repeat pair (26). At the centre of this domain is a sequence of 11 residues containing the motif xHCxAGxGRxG. The use of thiol-directed reagents has identified the cysteine (C197) at the centre of this motif as the probable catalytic residue (26).

Figure 2 summarises the PTPase prediction with reference to the *Yersinia* sequence. The prediction consists of two regions in which residues can be more reliably assigned to a secondary structure class (L39–L80, I138–E253), and two that are less clear (N1–D38, P94–T137). The regular secondary structure consists mainly of β-strand (29%) with 17% α-helix.

Several of the predicted loops (*e.g.* β3–β4) show variable composition across the set of sequences and are sites of a number of insertions or deletions; a pattern is usually indicative of an exposed surface loop.

The predicted loop between β4 and α1 commences with a conserved GP pair, likely to disrupt any preceeding regular secondary structure. The loop also contains charged or polar residues, although not in any ordered conservation pattern, typical of an exposed structure. Loop β9–α2, while relatively unconserved, is rich in Gly and Pro, is the site of a two-residue insertion in three sequences including *Yersinia* PTPase and of deletions in four sequences. It contains conserved prolines at 149 and 154 (V154 in *Yersinia* PTPase) and a conserved charged group, D150. The active site (C197) at the *C*-terminus of β10 is immediately followed by a flexible loop containing conserved Ala and Gly residues.

The pattern of conserved residues making up predicted helix 1 shows conserved charged residues at positions 64, 67, 68 and 71 in all but the *Yersinia* sequence. If this region is a helix these residues would form a charged surface down one face. α2 shows a weak pattern of conserved hydrophobic residues originating at L162 consistent with a partially exposed helix packing against the core of the protein. α4 shows a similar hydrophobic pattern starting at V223.

The region between I138 and V145 shows an alternating pattern of conserved and unconserved residues. This is consistent with a stretch of β-structure, where sequential sidechains point in opposite directions, and where every second residue is exposed to a conserved environment. With the exception of strands 3 and 4, which contain polar residues absolutely conserved across the alignment (N40 and Q55), the remaining strands in the well predicted regions are defined by short, highly conserved, stretches of residues consistent with buried β-structure. The conservation pattern of α5 is equivocal, being consistent with both strand and helix. A pattern of *unconserved* positions discernable at 256, 260 and 264 could form an exposed helical face. α3 shows no obvious pattern of conservation. β6, β7 and β8 are located by weak predictions of strand and some evidence of alternating patterns of unconserved positions.

The predicted structure for the PTPases is divided into two structurally distinct units separated by a region which shows little conservation in the family. This suggests that the tertiary fold comprises two subunits joined by a variable linker. In our predicted PTPase structure the *N*-terminal half of the protein (N-E132) is mostly β-strand (40% strand, 8.3% helix), while the *C*-terminal half (A133–C) contains roughly equal proportions of β-strand and α-helix (18% strand, 26% helix). The catalytic C197 of PTPase is located in the *C*-terminal reliably predicted region. The cAMP-dependent protein kinase structure also has two domains, with the catalytic D166 in the *C*-terminal (27). However, despite some superficial similarities the secondary structure predicted for the PTPases appears quite different to that of the protein kinases.

A number of candidate phosphate binding sites are seen, but those conserved within both the *Yersinia* sequence and the remaining 71 sequences are the most credible. These are R11, R231 and R234. R231 appears at the *C*-terminus of predicted helix 4. The dipole of helices favours phosphate binding at the *N*-terminus, so R231 is a less favoured candidate. R234, exposed in the following loop, is a good candidate as a phosphotyrosine binding site. The chain topology would allow this site to be placed close to the active site cysteine in the three-dimensional structure. The more distant R11, exposed at the *N*-terminus of a strand, may fold close this site.

Note added in proof

Shortly after this paper was accepted for publication the three-dimensional structure of human protein tyrosine phosphatase 1B (PTP1B) was reported by Barford *et al.* (28). Here, we briefly compare the results of crystallography with our prediction.

Fourteen of the 17 regular secondary structures in the conserved PTPase domain are correctly located. All five α-helices were correctly situated, although the *N*-terminus of helix 4 was predicted as strand. Eight of the 12 β-strands are predicted, with small differences between the lengths of the predicted regions and those observed in the X-ray structure.

The overall three-state accuracy of the prediction (Q3) is 68%, which compares favourably with the predicted Q3ave for the alignment of 83% given by the ASSP algorithm of Russel & Barton (29). The accuracy in the regions to which a high confidence in the pre-

diction was assigned is 83%. Regions of intermediate confidence show 72% accuracy, while regions of low confidence show 24% accuracy.

The proportion of strand in the X-ray structure is 24%, while 29% of the fold is helix. These differ from the proportions predicted (29 and 17%, respectively) owing to a general under-prediction of helix length, the incorrect prediction of the *N*-terminal half of helix 4 ($\alpha - 3$ of our prediction) as strand, the prediction of strands at the *N*-terminus of the conserved region and between helices 5 and 6 ($\alpha - 4$ and $\alpha - 5$ in our prediction), and the general over-prediction of strand in the region of low prediction confidence between A81 and T136. The region A81–T136 comprises a number of short strands isolated from the main sheet of the PTP1B fold, and is not highly conserved in the family. Accordingly, it is a probable site of structural heterogeneity between the 73 proteins in the PTPase alignment.

The phosphate binding site in PTP1B is composed of three residues whose equivalents in *Yersinia* PTPase are R203, Q240 and the active site cysteine (C197). The sites we predicted as phosphate binders (R11, R231 and R234) are all too distant from this site to be involved. The region containing Q240 was incorrectly predicted as a buried strand and was therefore not expected to be available for binding phosphate.

Our suggestion that the PTPase structure would consist of two domains separated by the variable-length loop between $\beta - 8$ and $\beta - 9$ ($\beta - 10$ and $\beta - 11$) is incorrect, since the PTP1B structure has a single domain. However, the predicted $\beta - 8/\beta - 9$ loop does lie in an exposed position on the surface of the structure consistent with our prediction.

Additional data regarding the predictions made in this paper may be obtained by anonymous ftp from geoff.biop.ox.ac.uk.

ACKNOWLEDGEMENTS

We thank Professor L.N. Johnson for encouragement and support, and R.B. Russell for his critical reading of the manuscript. C D L is supported by an MRC studentship award and is a member of Green College, Oxford. G.J.B. thanks the Royal Society for support. We thank Vivien Yee and David Barford for suggesting the FXIIIα and *Yersinia* PTPase problems.

REFERENCES

1. Holley, H.L. & Karplus, M. (1989) *Proc. Nat. Acad. Sci. USA* **86**, 152–156
2. Barton, G.J. & Russell, R.B. (1993) *Nature (London)* **361**, 505–506
3. Rost, B & Sander, C. (1993) *J. Mol Biol.* **232** (in press)
4. Russell, R.B & Barton, G.J. (1993) *J. Mol. Biol* (in press)
5. Zvelebil, M.J.J.M., Barton, G.J., Taylor, W.R. & Sternberg, M.J.E. (1987) *J. Mol. Biol.* **195**, 957–961
6. Crawford, I.P., Niermann, T. & Kirchner, K. (1987) *Proteins: Structure, Function and Genetics* **2**, 118–129
7. Benner, S. & Gerloff, D. (1990) *Adv. Enz Reg* **31**, 121–181
8. Barton, G.J., Newman, R.H., Freemont, P.F. & Crumpton, M.J. (1991) *Eur J Biochem* **198**, 749–760
9. Taylor, W.R. & Geisow (1987) *Protein Eng.* **1**, 183–187
10. Russell, R.B., Breed, J. & Barton, G.J. (1992) *FEBS Lett.* **304**, 15–20
11. Rost, B , Schneider, R. & Sander, C. (1993) *Trends Biochem. Sci.* 120–123
12. Takahashi, N., Takahashi, Y. & Putnam, F.W. (1986) *Proc. Natl Acad. Sci USA* **83**, 8019–8023
13. Guan, K. & Dixon, J.E. (1990) *Science* **249**, 553–556
14. Barton, G.J. (1990) *Methods Enzymol.* **183**, 403–428
15. Lim, V. (1974) *J. Mol. Biol.* **88**, 873–894
16. Chou, P.Y. & Fasman, G.D. (1978) *Adv. Enzymol* **47**, 45–148
17. Garnier, J., Osguthorpe, D.J. & Robson, B. (1978) *J Mol. Biol.* **120**, 97–120
18. Wilmot, A.C.M. & Thornton, J.M. (1988) *J Mol Biol.* **203**, 221–232
19. Rose, G D (1978) *Nature (London)* **272**, 586–591
20. Livingstone, C.D. & Barton, G.J. (1993) *Computer Applications in the Biosciences* (in press)
21. Davie, E.W (1986) *Protein Chem.* **5**, 247–253
22. Greenberg, C.S., Birckbichler, P.J. & Rice, R.H. (1991) *FASEB J.* **5**, 3071–3077
23 Bishop, P.D., Teller, D.C., Smith, R.A., Lasser, G W , Gilbert, T. & Seale, R.L. (1990) *Biochemistry* **29**, 1861–1869
24. Fischer, E.H., Charbonneau, H. & Tonks, N K. (1991) *Science* **253**, 401–406
25. Kaplan, R., Morse, B., Huebner, K., Croce, C., Howk, R., Ravera, M., Ricca, G., Jaye, M. & Schlessinger, J. (1990) *Proc Natl Acad. Sci USA* **87**, 7000–7004
26. Charbonneau, H., Tonks, N.K., Kumar, S., Diltz, C.D., Harrylock, M., Cool, D.E., Krebs, E.G., Fischer, E.H & Walsh, K.A. (1989) *Proc Natl Acad. Sci. USA* **86**, 5252–5256
27. Knighton, D , Zheng, J., Ten Eyck, L.F., Ashford, V A., Xuong, N.-h., Taylor, S.S. & Sowadski, J.M. (1991) *Science* **407–414**, 635–650
28. Barford, D., Flint, A., & Tonks, N (1994) *Science* **263**, 1397–1404
29. Russell, R. & Barton, G (1993) *J Mol. Biol* **234**, 951–957

Address

Dr Geoffrey J. Barton
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU
UK
Fax: 44-865-510454
e-mail: geoff@biop.ox.ac.uk