Threading Structural Model of the Manganese-Stabilizing Protein PsbO Reveals Presence of Two Possible β-Sandwich Domains

Florencio Pazos,[1†] Pedro Heredia,[2†] Alfonso Valencia,[1] and Javier de las Rivas[2*]
[1]*Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas, Madrid, Spain*
[2]*Instituto de Recursos Naturales y Agrobiología, Consejo Superior de Investigaciones Científicas, Salamanca, Spain*

ABSTRACT **The manganese-stabilizing protein (PsbO) is an essential component of photosystem II (PSII) and is present in all oxyphotosynthetic organisms. PsbO allows correct water splitting and oxygen evolution by stabilizing the reactions driven by the manganese cluster. Despite its important role, its structure and detailed functional mechanism are still unknown. In this article we propose a structural model based on fold recognition and molecular modeling. This model has additional support from a study of the distribution of characteristics of the PsbO sequence family, such as the distribution of conserved, apolar, tree-determinants, and correlated positions. Our threading results consistently showed PsbO as an all-beta (β) protein, with two homologous β domains of approximately 120 amino acids linked by a flexible Proline-Glycine-Glycine (PGG) motif. These features are compatible with a general elongated and flexible architecture, in which the two domains form a sandwich-type structure with Greek key topology. The first domain is predicted to include 8 to 9 β-strands, the second domain 6 to 7 β-strands. An Ig-like β-sandwich structure was selected as a template to build the 3-D model. The second domain has, between the strands, long-loops rich in Pro and Gly that are difficult to model. One of these long loops includes a highly conserved region (between P148 and P174) and a short α-helix (between E181 and N188)). These regions are characteristic parts of PsbO and show that the second domain is not so similar to the template. Overall, the model was able to account for much of the experimental data reported by several authors, and it would allow the detection of key residues and regions that are proposed in this article as essential for the structure and function of PsbO. Proteins 2001;45:372–381.** © 2001 Wiley-Liss, Inc.

Key words: fold recognition; manganese-stabilizing protein; protein structure; PsbO; threading

INTRODUCTION

The higher plant photosystem II (PSII) oxygen-evolving complex has three extrinsic polypeptides associated with the luminal side of the thylakoid membranes. These polypeptides, with apparent molecular weights of 33, 23, and 17 kDa, are important components of the oxygen-evolving apparatus because they stabilize the correct function of the manganese cluster, the site of water splitting and oxygen evolution. These polypeptides are encoded by three nuclear genes known as *psbO*, *psbP,* and *psbQ*. They are all present in the photosystem II of algae and higher plants, but PsbO is the only one also present in cyanobacteria, the most primitive oxyphotosynthetic organisms.

Experimental data on the specific role of the polypeptide PsbO associated with the manganese cluster are controversial because this protein is not directly involved in the binding of the manganese or calcium required for oxygen evolution. This lack of clear, functional data comes together with a limited knowledge of the PsbO structure. Several groups have carried out analyses of the protein by Fourier transform infrared (FTIR), circular dichroism (CD), far-UV CD, and other biophysical techniques and have reported diverse data concerning its secondary structure.[1–5] The FTIR and CD data seem to indicate that PsbO presents a major β–structure component, but they differ substantially about the explicit composition, with values for the α-helix content ranging from 8% to 27%. Hydrodynamic studies have indicated that the shape of the manganese-stabilizing protein is elongated in solution, with approximate dimensions of 12.6 nm × 3.0 nm, yielding an axial ratio of 4.2.[6] Folding studies on isolated PsbO have suggested that the kinetics of unfolding and refolding of this protein are similar to those of the immunoglobulin light chain.[7] Recent studies have presented an open discussion about the structure or conformation assigned to PsbO in solution, with some authors suggesting that it has a "natively unfolded" structure[8,9] and others proposing that it attains a "molten globule" structure.[10] In a recent publication the first X-ray structure of photosystem II from a thermophilic cyanobacteria (*Synechococcus elongatus*) at 3.8 Å resolution presented.[11] This work provided

Grant sponsor: CICYT; Grant number: PB98-0480.

†Florencio Pazos and Pedros Heredia contributed equally to this work.

*Correspondence to: Javier de las Rivas, Instituto de Recursos Naturales y Agrobiología, Consejo Superior de Investigaciones Científicas, c/ o Cordel de Merinas 40–52 (P.O. Box 257), 37071 Salamanca, Spain. E-mail: jrivas@gugu.usal.es

Received 4 January 2001; Accepted 6 August 2001

some new information about PsbO, such as the assignment of an electron density region of the extrinsic luminal side of photosystem II crystals to half the molecular mass of PsbO. This region showed a cylindrical structure 35 Å in length and corresponded to a group of β-strands of uncertain connectivity.[11] Because these X-ray data were low resolution, they did not provide clear knowledge about the PsbO fold. The model published (PDB = 1FE1) only gave coordinates for alpha-carbons without assignment of residue types.[11] For this reason it was not possible to know which part of the PsbO sequence corresponded to the electron density observed. This experimental work,[11] published after the first submission of our threading model, showed good agreement with our model, which in turn may help to further the assignment of residues in the PSII X-ray density map.

Protein structure prediction using computational methods is an expanding field that can provide interesting insight into protein function and biochemistry. In the current work, in which we used fold recognition methods, we made significant progress on previous structural predictions about PsbO.[12] "Threading," "fold recognition," and "remote homology modeling" are names given to a set of techniques to predict a protein's three-dimensional structure in the absence of obvious sequence similarity with proteins of known structure. Threading methods attempt to accommodate the problem sequence into the 3-D coordinates of several proteins of known structure and, with the aid of an appropriate scoring system, to select the best sequence–structure fit. Many scoring schemas have been proposed, ranging from simple concordance of secondary structure elements[13] to potentials of mean force,[14] statistical potentials,[15] fitness of sequence profiles,[16] and many others. More sophisticated methods, combining different potentials using neural networks technology, followed the initial approaches.[17] According to the results reported at several CASP (Critical Assessment of Techniques for Protein Structure Prediction) meetings, threading methods are increasingly reliable and a genuine alternative method for predicting protein structures.[18] According to reports at recent CASP and CAFASP events it has been also shown that using a combination of several threading methods for fold recognition considerably increases the correctness of predictions (http://www.cs.bgu.ac.il/~dfischer/CAFASP2).

There were three steps in our process of making a prediction about the PsbO structure: (1) obtaining possible models by using a series of complementary threading methods—TOPITS, GONPM, H3P2, 3-PSSM, HFR, THREADER2, and HMM; (2) comparing the results to select the most consistent predictions; and (3) assessing the possible models using information from the study of the PsbO protein structure family, including the distribution of conserved, apolar, correlated, and subfamily-specific residues. This approach is largely based on our previous systematic assessment of threading models with sequence-derived properties.[19] Finally, we compared the best model with the available experimental information and made suggestions to further experimental work. Doing structural analyses on PsbO increased our understanding of the relations between the structure and function of this protein.

METHODS

Threading Methods

Following are descriptions of the basic principles of the seven threading methods used in this work, including explanations of their scoring systems and associated reliability, that is, their chances of being correct in predicting, or percentage of certainty, along with the addresses of their Web sites. Four methods use *Z*-score schemas: $z\text{-score} = (raw - mean) / s$, where *raw* is the query-template alignment score, *mean* is the mean score of the query with all the possible folds in the fold library, and *s* is the standard deviation of the score distribution:

- TOPITS[13] (http://dodo.cpmc.columbia.edu/predictprotein/) is based on the matching between the predicted secondary structure and solvent accessibility of the problem sequence and the known secondary structure and accessibility of the templates. The results are ranked by *Z*-score. Hits with *Z*-scores of more than 3.0 are expected to correspond to correct predictions in more than 60% of the cases.
- GONPM[20] (http://fold.doe-mbi.ucla.edu/) uses sequence–sequence replacement tables and sequence-derived properties of the query protein, including the predicted secondary structure. The results are also arranged by a *Z*-score value, and when it is greater than 4.0, it indicates reliable predictions. The authors do not provide a quantitative value of certainty for the scores.
- H3P2[21] (http://fold.doe-mbi.ucla.edu/) is similar to the previous ones and uses a five-dimensional sequence-structure substitution matrix derived from known structures and predicted secondary structure. *Z*-scores are also used to classify the results. The authors presented an evaluation of the reliability associated with the *Z*-scores performed by counting the number of false positives in 243 examples using a fold library of 811 proteins. The probability of a false positive for *Z*-scores of 2.5 was 1.2×10^{-3}, decreasing to 1.9×10^{-4} for Z-scores of more than 3.0.
- 3D-PSSM[22] (http://www.bmm.icnet.uk/~3dpssm/) combines multiple-sequence profiles with structure-based profiles (which include solvation potentials derived from known structures and predicted secondary structure). The results are sorted by e-values. The authors indicate that an *e*-value of 0.97 corresponds to a certainty of correct prediction of 71.7% and an *e*-value of 1.11 to a certainty of 67.2%. Therefore, *e*-values < 1.0 are taken as significant.
- HFR[23] (http://www.cs.bgu.ac.il/~bioinbgu/) is a hybrid method that collects results from five threading programs, combining them in a search for the most consistent fold prediction among them. It also takes into consideration evolutionary information. The ranking of the hits is based on a combined score, which for values of more than 9.0 corresponds to correct predictions. A

quantitative value of certainty associated with the score was not provided.

- THREADER2[15,24] (http://insulin.brunel.ac.uk) is quite different from previous methods; it combines sequence information with pseudo-energies obtained from solvation and contact potentials previously derived from known protein structures. This method provides a *Z*-score to calibrate the significance of a prediction. The authors indicate that for a *Z*-score of greater than 3.5 the result is very significant, with a more than 60% chance of being correct.
- HMM[16] (http://www.cse.ucsc.edu/research/compbio/HMM-apps/) is perhaps the more different approach because it is based on a direct sequence comparison potential provided by a Hidden Markov Model engine. In this approach the query sequences are compared with collections of profiles derived for each one of the known protein structures (fold library). The score provided for each fold measures the probability that a query sequence belongs to such class or group of sequences with the same fold. A high score (such as ±9.0) indicates that the sequence of interest is probably a member of the selected class. Explicit certainty values associated with the scores are not provided.

Evaluating Threading Models by Sequence Features

Threading models were evaluated according to the fitting of sequence-extracted features derived from the study of the PsbO protein family. The clustering of apolar residues, of conserved residues, of correlated residues,[19] and of tree-determinant amino acids (Pazos & Valencia, unpublished) were the four properties studied. The idea was to discard those models in which, for example, apolar residues do not form a cluster, as statistically expected from the analysis of known globular proteins.

The tendency of sets of residues to cluster together was calculated as the difference between the distribution of distances among all pairs of residues and the distance distribution of the residues on which we are focusing: apolar, conserved, correlated, and tree-determinants. The explicit calculation was carried out with the previously derived formula[25]:

$$Xd = \sum_{i=1}^{n} \frac{P_{ic} - P_{ia}}{d_i \cdot n}$$

where n is the number of distance bins (there are 15 equally distributed bins from 4 to 60 Å); d_i is the upper limit for each bin (e.g., 8 for the 4–8 bin, normalized to 60); P_{ic} is the percentage of apolar, conserved, correlated, or tree-determinant pairs with distance between d_i and d_{i-1}; and P_{ia} is the same percentage for all pairs in the model. Defined in this way, $Xd = 0$ indicates no separation between the two distance populations, $Xd > 0$ indicates positive cases where the population of apolar, conserved, correlated, or tree-determinant pairs is shifted to smaller distances with respect to the population of all pairs (clustering).

RESULTS AND DISCUSSION

Threading Consistently Assigns an All-β Structure to PsbO

We explored the possible threading models for PsbO by applying seven programs to three divergent PsbO sequences: PsbO-spiol, a sequence from the plant *Spinacea oleracea;* PsbO-chlre, from the algae *Chlamydomonas reinhardtii;* and PsbO-synec, of the cyanobacteria *Synechocystis* sp. PCC6803. These three sequences cover the main range of known PsbO sequences down to 43% sequence identity.

The results obtained (Table I) are given in terms of the possible framework protein in PDB,[26] followed by the score proposed for each query-template alignment obtained with each threading program (e.g., 1fnf 2.32) and the structural characteristics that define each template according to SCOP,[27] that is, structural class (e.g., all-β), type of fold (e.g., sandwich), number of β-strands of a domain (e.g., seven strands), topology of the domain (e.g., Greek key), number of domains, and SCOP code (e.g., 1.002.001.002.001.003 for 1fnf, included for comparison purposes and not as an absolute reference because it cannot be expected to be stable over time).

All the PDB proteins selected by the threading methods as candidates for modeling PsbO are all-β proteins. Seven of the 11 proteins with significant scores (bold in Table I) are all-β sandwich proteins. Despite the consistent detection of an all-β-type structure, two methods (TOPITS and H3P2) did not predict any protein with a significant score and two other methods (3D-PSSM and THREADER2) produced only one best prediction above their confidence threshold, as shown in Table I. This could indicate that PsbO is a difficult protein to predict and that the more simple automatic programs cannot find any reasonable model for it. The best relative scores were produced by HFR and THREADER2 for models based on 1prr and 1hnf. The relative score 4.39/3.5, given by THREADER2, corresponds to a certainty greater than 80%.

According to the meaning of the scores (see the Methods section) and considering the agreement observed in Table I, we can estimate about a 90% level of confidence in the prediction of PsbO as an all-β protein, and about an 80% level of confidence in the prediction as an all-β sandwich protein. This estimation is prudent but has to be taken with caution because the accuracy of different threading methods has not been compared quantitatively in any published study, making it difficult to make a precise estimation of certainty.

PsbO Seems to Match Ig-Like β Sandwich Fold Type

Even if all the proteins presented in Table I belong to the very general all-β structural class, it would be necessary to do additional investigation into the possibility of predicting a more specific fold. For this reason the relative scores for different templates of the most significant methods (3D-PSSM, HFR, and THREADER2) are presented in Figure 1. The results consistently show the abundance of all-β sandwich and Ig-like β sandwich among the top five

TABLE I

Method	TOPITS Rost (1995)	GONPM Fischer & Eisenberg (1996)	H3P2 Rice & Eisenberg (1997)	3D-PSSM Sternberg (1999)	HFR Fischer (2000)	THREADER2 Jones (1999)	HMM Karplus (1998)
METHOD BASED ON	binary structure and accessibility or SOLVATION					OTHER METHODS	
Scores (confidence threshold)	> 3.0	> 4.0	> 2.5	< 1.0	> 9.0	> 3.5	< −9.0
PsbO-spiol 248 aa plant sequence identity 100%	1fnf (2.32) All-β sandwich 7 strands greek-key 4 β domains (7 × 4) 1.002.001.002.001.003 fibronectin cell adhesion protein	1mfn 4.2 All-β sandwich 7 strands greek-key 2 β domains (7 + 8) 1.002.001.002.001.004 fibronectin cell adhesion protein	2ayh (2.2) All-β sandwich 12–14 strands complex 1.002.026.001.002.002 beta glucanasa	> 1.0	1prr 9.90 All-β sandwich 8 strands greek-key 2 β-domains (8 + 8) 1.002.010.001.002.001 spore coat proteins binding protein	8fab A (3.03) All-β sandwich 7 strands greek-key 2 β-domains (9 + 7) 1.002.001.001.001.012 1.002.001.001.002.022 immunoglobulin Fab fragment	1cne −9.01 All-β barrel 6 strands greek-key 1.002.038.001.001.007 nitrate reductase core domain
PsbO-chlre 239 aa algae sequence identity 63%	< 3.0	1mfn 4.3 All-β sandwich 7 strands greek-key 2 β domains (7 + 8) 1.002.001.002.001.004 fibronectin cell adhesion protein	< 2.5	1avg I 0.70 All-β barrel 8 strands meander 1.002.053.001.003.001 thrombin inhibitor	1prr 14.2 All-β sandwich 8 strands greek-key 2 β-domains (8 + 8) 1.002.010.001.002.001 spore coat proteins binding protein	1hnf 4.39 All-β sandwich 7 strands greek-key 2 β domains (9 + 7) 1.002.001.001.001.007 1.002.001.001.003.006 T-lymphocyte adhesion glycoprotein	1cne −9.92 All-β barrel 6 strands greek-key 1.002.038.001.001.007 nitrate reductase core domain
PsbO-synec 246 aa cyanobacteria sequence identity 46%	< 3.0	1mfn 4.2 All-β sandwich 7 strands greek-key 2 β domains (7 + 8) 1.002.001.002.001.004 fibronectin cell adhesion protein	< 2.5	1fnh A (2.70) All-β sandwich 7 strands greek-key 3 β dom. (*7 + 7 + 6) 1.002.001.002.001.003 fibronectin integrin binding	1fnh A 9.10 All-β sandwich 7 strands greek-key 3 β dom. (7 + 7 + 6) 1.002.001.002.001.003 fibronectin integrin binding	< 3.5	2bb2 −9.03 All-β sandwich 8 strands greek-key 2 β domains (8 + 8) 1.002.010.001.001.005 eye lens protein

Fig. 1. Graphical representation of the threading results**.** The first five proteins selected by three threading methods (3D-PSSM, HFR, and THREADER2) for PsbO sequence of spinach (PsbO-spiol), *Chlamydomonas* (PsbO-chlre), and *Synechocystis* (PsbO-synec). The bar size corresponds to the values of the score given for each protein over the confidence threshold of the corresponding threading method. The gray bars correspond to all-β proteins, the bars in black correspond to Ig-like β sandwich protein, and the bars in white correspond to other folds.

proteins obtained with these methods. In some cases the prediction was all-β barrel: for example 1avgI, predicted by 3D-PSSM for *Chlamydomonas* PsbO. This kind of structure is less represented than the all-β sandwich. Indeed, the models of PsbO based on all-β barrel structures are usually partial, leaving important parts of PsbO out of the model (data not shown). The comparison of three threading methods in Figure 1 reflects that 3D-PSSM gives less confident predictions than do HFR and THREADER2, indicated by the lower values of the relative scores. As can be inferred from the discussion of the scoring schemas presented in Methods, a score/threshold parameter of 1 corresponds to a certainty level of around 60% to 70%.

PsbO Seems to Include Two Similar β Domains

A common and interesting feature of the threading models based on all-β sandwich proteins is that most of them correspond to two-domain proteins, with a few others

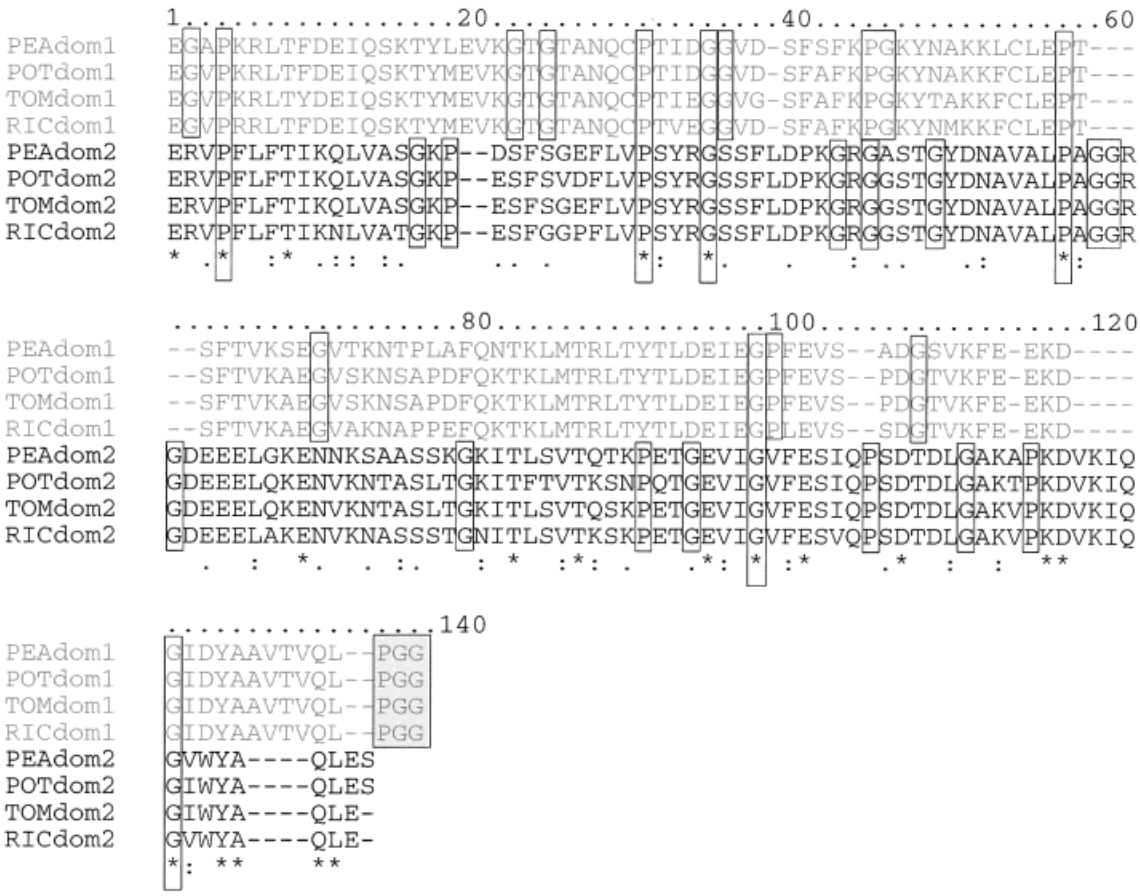

Fig. 2. Domain distribution shown of the multiple-sequence alignment of selected PsbO sequences: two regions from pea, potato, tomato, and rice. The first region includes 118 residues and the PGG final motif, marked within a gray box. The second region includes 127 residues. In each region the conserved Pro and Gly are marked with boxes. There are 20 positions fully conserved in the alignment, six of which are Pro.

(1fnhA, 1fnf) corresponding to proteins with three or four domains. Moreover, the best threading scores for the PsbO sequence were obtained with two-domain proteins. We interpret these results as a first indication of the presence of two structural domains in PsbO.

To use some independent or different information to test further the idea of a two-domain protein structure, we searched for possible domain boundaries in a multiple sequence alignment containing 15 PsbO sequences.[12] A clue came from the presence of a conserved PGG motif, from position 118 to position 120 in the pea sequence (117–119 in the spinach sequence). The flexible structure of this motif and its likely role as a domain linker or hinge can indicate the presence of two parts or regions in a protein with 118 and 127 residues, respectively.

A multiple alignment of four PsbO plant sequences (pea, potato, tomato, and rice) was generated to check the internal identity of the proposed domains. The result, shown in Figure 2, yielded a remarkable homology (20 totally conserved residues, with 16.9% identity and 47.4% similarity). It is surprising that such a symmetry had not been detected previously. Perhaps this is because the similarity between the two domains in the algae and cyanobacteria sequences is much lower, in turn making the similarity less obvious when all the known PsbO sequences are compared. The discovery of a symmetry in the protein, with two regions of clear sequence similarity, strongly supports the two-domain structure of PsbO initially suspected from analyzing the threading models. This observed two-domain structure cannot be detected using the standard domain databases ProDom[28] and Pfam.[29]

Separated Threading of Two PsbO Regions Also Selects Ig-Like β Sandwich Proteins

Because different methods—structure- and sequence-based—had indicated a two-domain structure for PsbO, independent threading runs with each of the two PsbO regions were carried out: first, with the N-terminal region, from residue 1 to residue 120, and second, with the C-terminal region, from residue 117 to residue 248. The threading results with each separated region pointed again to all-β sandwich proteins as the most favorable models, with a similar distribution to the one presented in Table I, in which the threading experiments had been carried out with the full sequence. Some templates proposed were: 1ksr = actin binding protein 120, for the N-terminal region, and 1srdA = Cu,Zn-superoxide dismutase, for the C-terminal second region, both of which

TABLE II

PDB Id	Threading score/threshold	Fitting values of PsbO properties in each template: Conserved residues	Apolar residues	Tree-determinants	Correlated mutations	Σ properties
8fab A	0.866	0.685	1.008	0.467	0.470	2.630
1fnf	0.773	0.385	1.093	0.063	−2.189	−0.648
1fnh A	1.011	0.376	0.305	−1.110	−0.650	−1.079
1hnf*	1.254*	0.387	1.067	0.553	2.442	4.449*
1mfn	1.050	0.258	1.193	0.832	3.068	5.351
2bb2	1.003	−1.129	0.377	−0.246	−0.414	−1.412
1prr	1.100	1.306	−0.053	−0.667	−2.521	−1.935

corresponded to the Ig-like β sandwich fold. These results can be viewed as confirming the proposed domain structure and fold. For the second domain the 1srdA template provided a better alignment than did the templates obtained for the whole protein. This protein (1srdA, superoxide dismutase) has longer loops between the β-strands that fit better with the PsbO loops in this region.[12] As a whole, the threading runs with the individual domains did not lead to higher scores, and they produced models that only covered half the protein.

Selection of the Best Frameworks Among the all-β Sandwich Candidates

A search of the FSSP database to compare protein structures[30] (http://www2.embl-ebi.ac.uk/dali/fssp/fssp.html), using the seven all-β sandwich protein candidates (Table I), led us to distinguish two groups of all-β sandwiches. In the first group were five proteins—1hnf, 8fabA, 1fnf, 1fnhA, and 1mfn; in the second group two, 1prr and 2bb2. These groups correspond to two different folds in the SCOP classification, the first to Ig-like β sandwiches and the second to crystallines. To select the optimum possible template to model PsbO, each of these β proteins was assessed using two criteria: threading alignment coverage and conservation of PsbO sequence properties.

The threading alignments of PsbO with each template protein showed that the coverage of the aligned sequence (248 aa) was quite complete for the first group of proteins (Ig-like β sandwich fold) but only partial for the second group (crystalline fold). Such partial alignment is reflected in the number of residues not aligned in the N-terminus or C-terminus of the models:

	PDB Code	Not Aligned Residues
1st—	1fnf	3 Nt or Ct
	1fnhA	3 Nt or Ct
	1hnf	19 Ct
	8fabA	9 Nt and 15 Ct
	1mfn	39 Nt
2nd—	2bb2	55 Ct
	1prr	85 Nt

These differences in model completeness may indicate that PsbO fits better to the number of secondary structure elements and topology characteristic of an Ig-like β sandwich fold.

Additional information was required to make the selection of a specific template from among different possible Ig-like β sandwich frameworks for modeling PsbO. Our strategy was to evaluate the distribution of a number of relevant PsbO sequence characteristics in each model. We evaluated in detail the distribution of the conserved, apolar, tree-determinant, and correlated residues that were calculated from the sequences of the PsbO protein family. The visualization of the sequence characteristics in the various models and the numerical evaluation of the significance of their distribution (see Methods section) was carried out with the program Threadlize.[31]

The distribution of the various residues (Table II) shows 1mfn and 1hnf as the best candidate models. The table also shows that 1prr and 2bb2 have negative scores, indicating they are not appropriate models for PsbO, which confirmed our previous deduction. In general, the lowest values were obtained for non-Ig-like folds. A combined score adding the four sequence features (Table II, last column) also selects some of the folds as better frameworks for holding the PsbO sequence attributes in a consistent spatial proximity. Such an addition of the fitting values produced the best numbers for 1mfn (5.351) and 1hnf (4.449). Several reasons formed the basis for the final selection of the best template from between these two proteins: (1) as indicated above, the coverage of the alignment of PsbO with 1mfn leaves out 39 N-terminal residues but only 19 C-terminal residues with 1hnf; (2) the 1mfn PDB corresponds to 20 structures determined by NMR that are more difficult to use for modeling than is the 1hnf X-ray structure; and (3) the relative threading score (Table II, second column) was best for 1hnf.

FSSP Structural Family Assigned to PsbO and Some Functional Clues

Because we had inferred there was a two-domain structure for PsbO, the structural families that best fit each domain had to be considered in the model building. Both are homologous Ig-like β sandwich domains, but they have some topological differences in their β-strands. The first domain of 1hnf belongs to the FSSP structural family led by 1qa9A (see the FSSP database[30] at http://jura.ebi.ac.uk:8765/holm/qz?filename=/data/research/fssp//1qa9A.fssp), which includes 8fabA, 1fnf, 1fnhA, 1cdy, and 1dr9A. These proteins are mentioned because several were detected as possible templates for PsbO by the threading methods (see 8fabA, 1fnf, and 1fnhA in Table I) or because they are structurally and functionally very closely homologous to 1hnf (1cdy and 1dr9A). The second domain fits better in

the FSSP structural family led by 1mfmA (see the FSSP database[30] at http://jura.ebi.ac.uk:8765/holm/qz?filename=/data/research/fssp//1mfmA.fssp), which also includes proteins 8fabA, 1fnhA, 1fnf, 1cdy and 1dr9A.

From a study of these two FSSP families we noted some important features about them. First, both families include proteins with very different functions. For instance, T-lymphocyte glycoprotein fragment, T-cell receptor fragment, fibronectin, superoxide reductase, and cytochrome f (cited in order of decreasing *Z*-value in the FSSP) are all part of the 1qa9A family; and superoxide dismutase, fibronectin, Fab fragment from human Ig, beta-amylase, and T-cell surface glycoprotein fragment are members of the 1mfmA family. Second, many of these functions are related to cell adhesion or macromolecule interaction. And, third, we observed that the folds of the redox-related enzymes superoxide dismutase and superoxide reductase are assigned to PsbO. It has been shown that proteins with similar folds can perform completely different functions,[32] and so the use of a T-lymphocyte adhesion glycoprotein (1hnf) as a template to model an oxygen-evolving protein (PsbO) should not be very surprising. However, the structural correlation found in this work may also include some unknown functional relation. In this respect, it was very interesting to find that PsbO (involved in the production of oxygen from water) has the same fold as that of superoxide reductase (an enzyme that reduces superoxide to hydrogen peroxide) and superoxide dismutase (an enzyme that catalyses a similar reaction with superoxide, producing hydrogen peroxide plus oxygen). If this predicted structural correlation includes any functional meaning, we would suggest that PsbO might be involved in the quenching of oxygen radicals produced during oxygen evolution. This is a totally new but sensible hypothesis that should be tested experimentally.

Proposal of a 3-D Remote Model for PsbO

The construction of a good alignment between template and query proteins is essential for obtaining sensible models. The global alignment of spinach PsbO against the best template selected, 1hnf, was obtained by the THREADER2 method, with some small regions taken from an alignment obtained with the HFR method (Fig. 3). Some small modifications were carried out because of the FSSP structural alignments of the two groups of proteins that correspond to each domain.

Most of the 1hnf region aligned with PsbO corresponds to the structural core of its FSSP structural families. Therefore, the main secondary structure features (mostly the β-strands) of 1hnf and those predicted for PsbO are in good agreement. This is particularly clear for the first domain because the second domain of PsbO has some specific regions (mainly loops) that could not be aligned (Fig. 3).

The alignment was prepared with SWISS-PDBviewer,[33] Version 3.6 (http://www.expasy.ch/spdbv/mainpage.htm) and submitted to SWISS-MODEL[33] (http://www.expasy.ch/swissmod/SWISS-MODEL.html) for the automatic building of a protein model. The quality of the model obtained was fairly good according to the scores of the evaluation programs (WhatCheck[34]), taking into account that it corresponds to a remote homology modeling case. For example, the first-generation packing quality was −3.733, Ramachandran plot appearance was −3.883, inside/outside distribution was 1.142. The structural model generated by SWISS-MODEL can be visualized with any representation software. As expected, given the low level of sequence identity between the query and the framework, the core of each domain was fairly well formed, but other regions, such as surface loops, presented serious modeling difficulties. The regions corresponding to insertions and deletions (see alignment in Fig. 3) corresponded mainly to loops in the structure, an indication that the structural core of the template matches the PsbO predicted structure. Long loops in the PsbO second domain were impossible to model, including the regions from residues Gly152 to Gly163 and from Gly177 to Asn198 (boxes in Fig. 3). The first loop corresponds to a very conserved PsbO region[12] and the second to a RGD cell-attachment motif (residues 178–180), followed by a short predicted α-helix, which does not have any counterpart in 1hnf. The size of the insertions confirms that PsbO second domain is less similar to the template than the first domain.

The structure of the PsbO model is represented in Figure 4; the corresponding coordinates are available from http://gredos.cnb.uam.es/pazos/PsbO. Given that the model corresponds to a two-domain protein linked with a flexible region, it is possible that the relative positions of both domains in PsbO would be different from those of 1hnf.

PsbO Model Based on 1hnf Is Congruent With Known Structural Data

Biophysical and biochemical data on the structure of PsbO are available from several sources. A few residues of significant importance for the PsbO structure have been identified, such as two Cys (Cys28 and Cys51) close to the N-terminus that form a disulfide bridge essential to maintaining native PsbO tertiary structure.[35,36] Trp241 is the only Trp in the protein; it is closed to the C-terminus, and its fluorescence emission spectrum indicates it is buried in a nonpolar niche, probably inside the PsbO core.[7,10]

The proposed model, based on 1hnf (Fig. 4), locates the cysteines in the first β-domain in a way that could interact to form the disulfide bridge and stabilize the two β-sheets that compose such a domain. In the model Trp241 is in the last β-strand in the interior of the second domain, where it forms part of the hydrophobic core. This is in good agreement with the previous experimental data. In addition, reports on refolding kinetics experiments based on the intrinsic fluorescence of Trp have indicated that PsbO presents an unfolding pattern similar to the characteristic pattern of inmunoglobulins,[7] providing additional indirect support for the proposed structural model.

Two other structural characteristics experimentally identified for PsbO are its elongated form, with a length-to-depth ratio of 4:1,[6] and its flexible architecture,[8,9] which corresponds to its composition being rich in Gly and Pro pairs (Fig. 1). The proposed model, based on 1hnf, is in

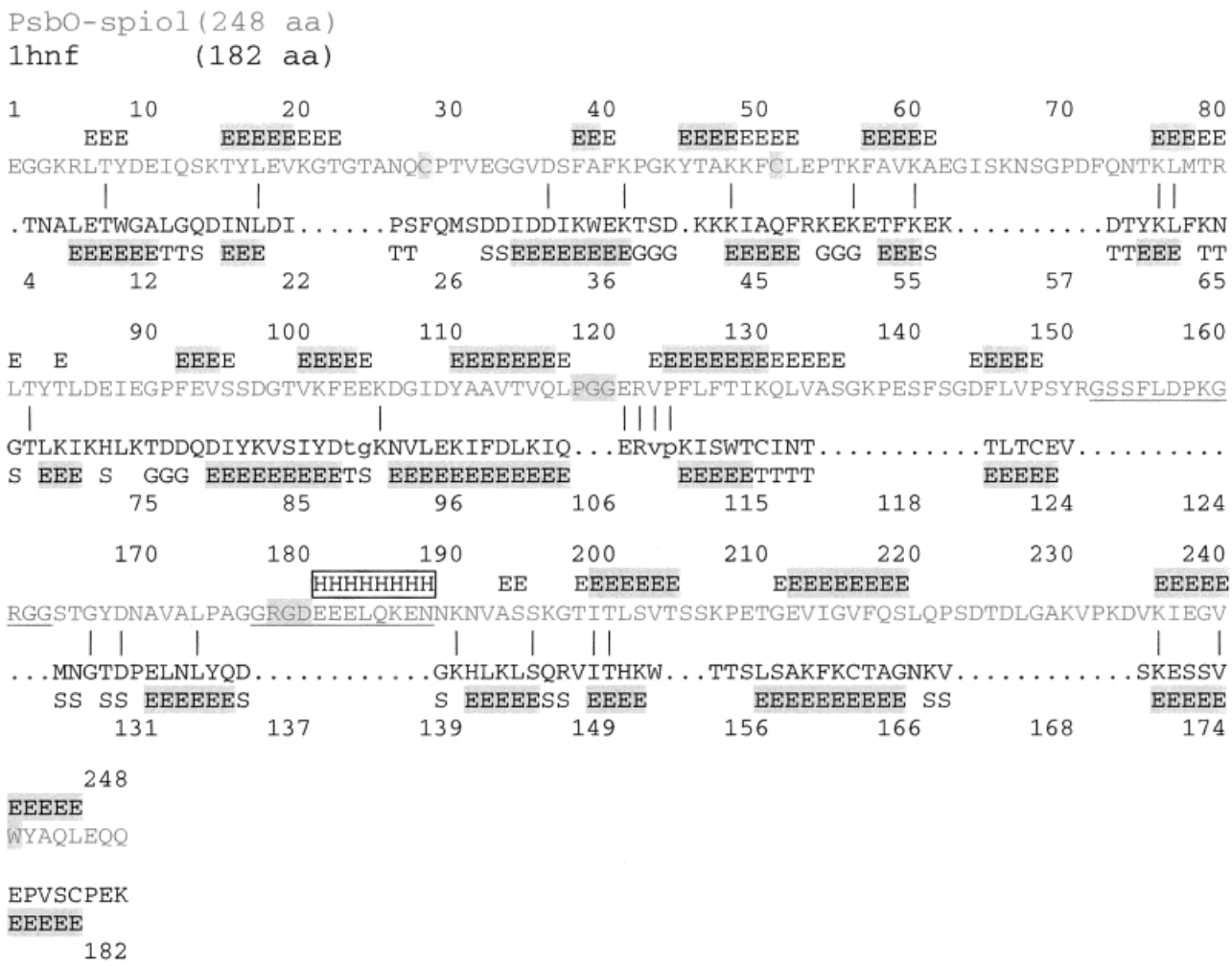

Fig. 3. Sequence to structure alignment of PsbO and 1hnf. Vertical bars mark the conserved residues. The predicted PsbO secondary structure is shown above its sequence, and the actual 1hnf secondary structure is indicated bellow its sequence (E corresponds to β-structure and H to α-helix). The residues in β-strands are shown in shadow if the reliability of the secondary structure prediction is above 90%. Some significant PsbO residues are marked in gray boxes on the sequence, including the two Cys (C28 and C51), the PGG motif (from residue 118 to residue 120), the RGD cell attachment domain (from residue 178 to residue 180), and the only Trp present in PsbO (W241). The numbering corresponds to the spinach sequence.

good agreement with both features because it corresponds to an elongated protein structure and contains a considerable proportion of loops rich in GG, GP, and PP pairs. The location of the conserved PGG triplet in the middle of two β-domains also provides additional flexibility to the protein.

Most recent structural information on PsbO has been provided by X-ray data at 3.8 Å resolution about cyanobacterial PSII,[11] which showed that a 35-Å section of PsbO has a β-sheet with at least five strands. Our model agrees with these data because both the domains in the PsbO model are 32–40 Å long and include two β-sheets with more than five strands. From an examination of the matching of the strands in size and position, we also suggest that the region presented in the X-ray low-resolution model corresponds to the N-terminal region of our PsbO model.

CONCLUSION

The use of computational methods to analyze, compare, and predict protein structure is a great aid in guiding experiments in protein biochemistry. Fold recognition is part of the protein structure prediction methodology, and it was used in this work to gain new insight into a protein with an unknown detailed structure. The structure of PsbO has been a controversial subject in recent years, with some researchers claiming that it is a "natively unfolded" protein,[8,9] and others suggesting that it has a "molten globule"–type structure.[10] Our results indicate that PsbO has a well-defined β core, surrounded by long flexible loops. This proposition is well supported by the new X-ray data about PSII.[11] Our model also agrees well with the idea that the protein has a elongated flexible structure, including a possible hinge point between two domains. This proposed PsbO architecture, with two domains con-

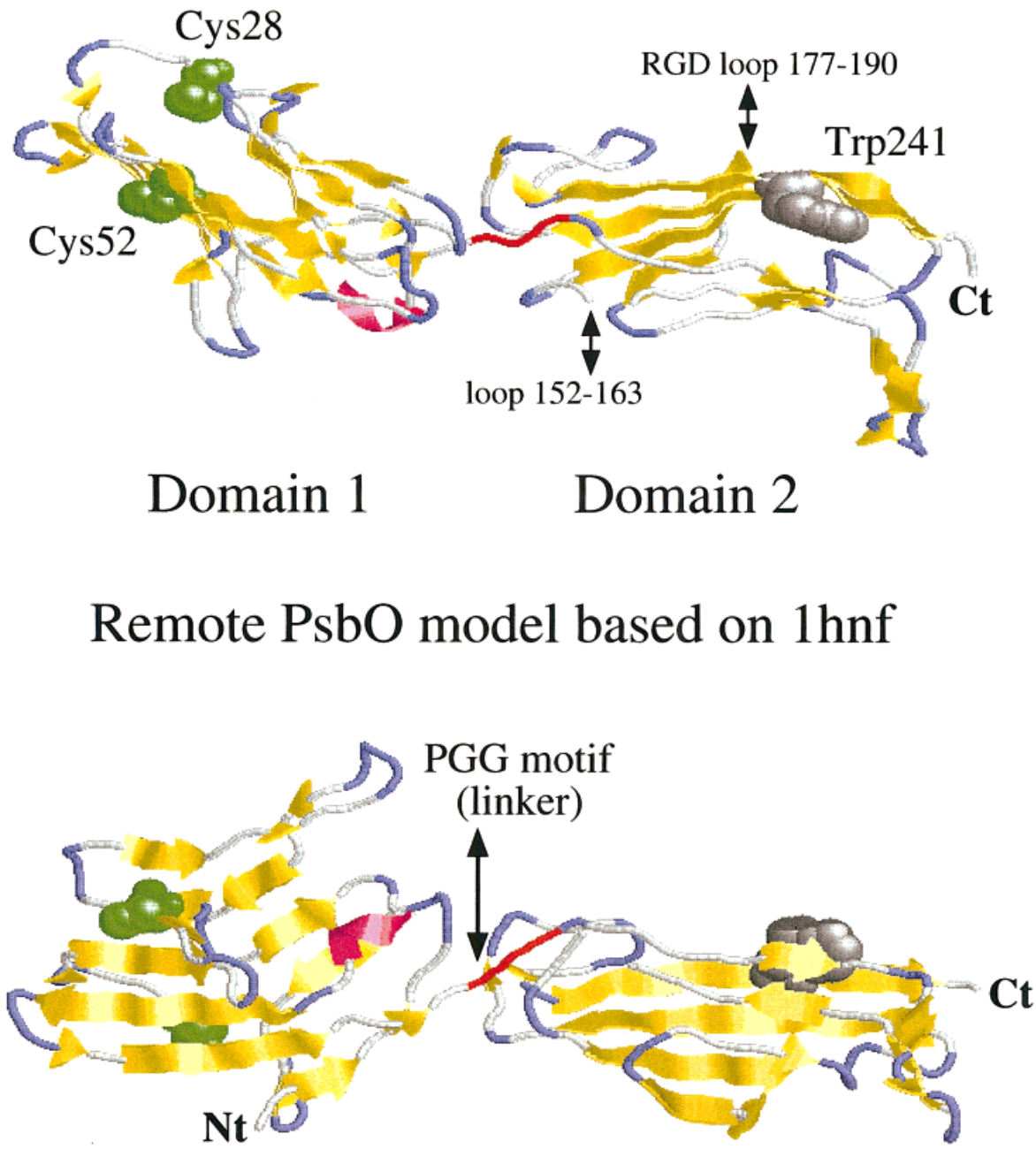

Fig. 4. Graphical representation of the PsbO model based on the 1hnf framework. The PsbO backbone of the model is represented. Note that it is based on a remote similarity with the 1hnf template. Consequently, large deviations are expected from the real structure of PsbO.[37] The representation was made with RASMOL.[38] The critical residues C28, C51, and W241 are presented as balls. Different regions are indicated corresponding to the two possible structural domains, the linker region (PGG motif) and the position of the two long insertions that were not included in the structural model.

nected by a linker region, would be ideal for providing conformational flexibility required to close and give stability to the manganese cluster in the thylakoid membrane or to be released free in the lumen when PSII is inactivated.[5] Some of the ideas derived from the model can now serve as a guide for additional experimental approaches that may validate it and may provide new insights into the detailed mechanism of PsbO function. Some proposals for future experiments are: (1) the cloning and expression of each separated PsbO domain to check their independent fold and folding kinetics and to test how it is their particular docking or interaction works with the oxygen-evolving complex; (2) construction of mutants in the highly conserved P148–P174 loop to see how they affect the oxygen-evolving activity; and (3) testing the ability of the PsbO protein to react with hydrogen peroxide or with other oxygen radicals. Some of these lines of research are currently being undertaken in our laboratory.

The modeling steps followed in this work can be also seen as a general strategy for the selection of templates for

proteins of unknown structure based on the combination of results from different threading methods, sequence information, and experimental data.

ACKNOWLEDGEMENTS

We thank members of the Protein Design Group for continual support and stimulating discussions. We also thank the Vasc Government for the predoctoral grant awarded to P. Heredia.

REFERENCES

1. Xu Q, Nelson J, Bricker TM. Secondary structure of the 33 kDa, extrinsic protein of photosystem II: a far-UV circular dichroism study. Biochim Biophys Acta 1994;1188(3):427–431.
2. Ahmed A, Tajmir-Riahi HA, Carpentier R. A quantitative secondary structure analysis of the 33 kDa extrinsic polypeptide of photosystem II by FTIR spectroscopy. FEBS Lett 1995;363(1–2): 65–68.
3. Sonoyama M, Motoki A, Okamoto G, Hirano M, Ishida H, Katoh S. Secondary structure and thermostability of the photosystem II manganese-stabilizing protein of the thermophilic cyanobacterium Synechococcus elongatus. Biochim Biophys Acta 1996;1297(2): 167–170.
4. Shutova T, Irrgang KD, Shubin V, Klimov VV, Renger G. Analysis of pH-induced structural changes of the isolated extrinsic 33 kilodalton protein of photosystem II. Biochemistry 1997;36(21): 6350–6358.
5. Hutchison RS, Betts SD, Yocum CF, Barry BA. Conformational changes in the extrinsic manganese stabilizing protein can occur upon binding to the photosystem II reaction center: an isotope editing and FT-IR study. Biochemistry 1998;37(16):5643–5653.
6. Zubrzycki IZ, Frankel LK, Russo PS, Bricker TM. Hydrodynamic studies on the manganese-stabilizing protein of photosystem II. Biochemistry 1998;37(39):13553–13558.
7. Tanaka S, Kawata Y, Wada K, Hamaguchi K. Extrinsic 33-kilodalton protein of spinach oxygen-evolving complexes: kinetic studies of folding and disulfide reduction. Biochemistry 1989; 28(18):7188–93.
8. Lydakis-Simantiris N, Betts SD, Yocum CF. Leucine 245 is a critical residue for folding and function of the manganese stabilizing protein of photosystem II. Biochemistry 1999;38(47):15528–35.
9. Lydakis-Simantiris N, Hutchison RS, Betts SD, Barry BA, Yocum CF. Manganese stabilizing protein of photosystem II is a thermostable, natively unfolded polypeptide. Biochemistry 1999;38(1): 404–14.
10. Shutova T, Irrgang K, Klimov VV, Renger G. Is the manganese stabilizing 33 kDa protein of photosystem II attaining a "natively unfolded" or "molten globule" structure in solution? FEBS Lett 2000;467(2–3):137–140.
11. Zouni A, Witt H-T, Kern J, Fromme P, Krauβ N, Saenger W, Orth P. Crystal structure of photosystem II from *Synechococcus elongatus* at 3.8 Å resolution. Nature 2001;409:739–743.
12. De Las Rivas J, Heredia P. Structural predictions on the 33 kDa extrinsic protein associated to the oxygen evolving complex of photosynthetic organisms. Photosynthesis Research 1999;61(1):11–21.
13. Rost B. TOPITS: threading one-dimensional predictions into three-dimensional structures. In: C Rawlings, D Clark, R Altman, L Hunter, T Lengauer, & S Wodak, editors. The third international conference on Intelligent Systems for Molecular Biology (ISMB). Cambridge, U.K.: AAAI Press; 1995. p 314–321.
14. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures of globular proteins. J Mol Biol 1990;213: 859–883.
15. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
16. Karplus K, Barrett C, Hughey R. Hidden Markov Models for detecting remote protein homologies. Bioinformatics 1998;14(10): 846–856.
17. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
18. Sippl MJ, Lackner P, Domingues FS, Koppensteiner WA. An attempt to analyze progress in fold recognition from CASP1 to CASP3. Proteins 1999;3:226–230.
19. Olmea O, Rost B, Valencia A. Effective use of sequence correlation and conservation in fold recognition. J Mol Biol 1999;295:1221–1239.
20. Fischer D, Eisenberg D. Fold recognition using sequence-derived predictions. Protein Science 1996;5:947–955.
21. Rice D., Eisenberg D. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J Mol Biol 1997;267:1026–1038.
22. Kelley LA, MacCallum RM, Sternberg MJE. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299(2):501–522.
23. Fischer D. Hybrid fold recognition: combining sequence derived properties with evolutionary information. Maun Lani, HI: Pacific Symp Biocomputing, 2000. p 119–130.
24. Jones DT, Miller RT, Thornton JM. Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. Proteins 1995;23:387–397.
25. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated Mutations contain information about protein–protein interaction. J Mol Biol 1997;271(4):511–523.
26. Bernstein FC, Koetzle TF, Williams GJ, Meyer, Jr., EE, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
27. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540
28. Sonnhammer E, Kahn D. Modular arrangement of proteins as inferred from analysis of homology. Protein Sci 1994;3:482–492.
29. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. The Pfam Protein Families Database. Nucl Acids Res 2000;28:263–266.
30. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–602.
31. Pazos F, Rost B, Valencia A. A platform for integrating threading results with protein family analyses. Bioinformatics 1999;15:1062–1063.
32. Russell RB, Sasieni PD, Sternberg MJ. Supersites within superfolds. Binding site similarity in the absence of homology. J Mol Biol 1998;282:903–918.
33. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–2723
34. Hooft RWW, Vriend G, Sander C, Abola EE. Errors in protein structures. Nature 1996;381:272.
35. Burnap RL, Qian M, Shen JR, Inoue Y, Sherman LA. Role of disulfide linkage and putative intermolecular binding residues in the stability and binding of the extrinsic manganese-stabilizing protein to the photosystem II reaction center. Biochemistry 1994; 33(46):13712–13718.
36. Betts SD, Ross JR, Hall KU, Pichersky E, Yocum CF. Functional reconstitution of photosystem II with recombinant manganese-stabilizing proteins containing mutations that remove the disulfide bridge. Biochim Biophys Acta 1996;1274(3):135–142.
37. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5:823–826.
38. Sayle R, Milner-White E. RASMOL: biomolecular graphics for all. Trends Biochem Sci 1995;20(9):374–376.