BIOINFORMATICS ORIGINAL PAPER
Vol. 22 no. 10 2006, pages 1183–1189
doi:10.1093/bioinformatics/btl040

Structural bioinformatics

Generation of 3D templates of active sites of proteins with rigid prosthetic groups

Jean-Christophe Nebel

Faculty of Computing, Information Systems & Mathematics, Kingston University, Kingston-upon-Thames, Surrey KT1 2EE, UK

Received on November 10, 2005; revised on January 15, 2006; accepted on February 3, 2006
Advance Access publication February 10, 2006
Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: With the increasing availability of protein structures, the generation of biologically meaningful 3D patterns from the simultaneous alignment of several protein structures is an exciting prospect: active sites could be better understood, protein functions and protein 3D structures could be predicted more accurately. Although patterns can already be generated at the fold and topological levels, no system produces high-resolution 3D patterns including atom and cavity positions. To address this challenge, our research focuses on generating patterns from proteins with rigid prosthetic groups. Since these groups are key elements of protein active sites, the generated 3D patterns are expected to be biologically meaningful.

Results: In this paper, we present a new approach which allows the generation of 3D patterns from proteins with rigid prosthetic groups. Using 237 protein chains representing proteins containing porphyrin rings, our method was validated by comparing 3D templates generated from homologues with the 3D structure of the proteins they model. Atom positions were predicted reliably: 93% of them had an accuracy of 1.00 Å or less. Moreover, similar results were obtained regarding chemical group and cavity positions. Results also suggested our system could contribute to the validation of 3D protein models. Finally, a 3D template was generated for the active site of human cytochrome P450 CYP17, the 3D structure of which is unknown. Its analysis showed that it is biologically meaningful: our method detected the main patterns of the cytochrome P450 superfamily and the motifs linked to catalytic reactions. The 3D template also suggested the position of a residue, which could be involved in a hydrogen bond with CYP17 substrates and the shape and location of a cavity. Comparisons with independently generated 3D models comforted these hypotheses.

Availability: Alignment software (Nestor3D) is available at http://www.kingston.ac.uk/~ku33185/Nestor3D.html

Contact: j.nebel@kingston.ac.uk

1 INTRODUCTION

The simultaneous alignment of several protein sequences using tools such as ClustalW (Thompson *et al*., 1994) is now an essential step in protein analysis. Multiple comparisons allow the alignment of distant homologues and the detection of patterns which can be further investigated by biochemists: protein functions can be suggested and residues potentially involved in protein activities can be detected. With the increasing availability of protein 3D structures, the simultaneous alignment of 3D structures is an exciting prospect which should allow the generation of biologically meaningful 3D patterns. These patterns could then be used either for predicting functions of proteins the 3D structures of which are known or, as templates, for modelling 3D structures.

While many powerful tools have been developed to allow pairwise protein structure comparisons based on atom coordinates (Holm and Sanders, 1994; Gibrat *et al*., 1996; Shindyalov and Bourne, 1998), the simultaneous alignment of protein structures and the generation of 3D patterns are still very challenging problems. Common patterns can be generated automatically from protein structures at the fold (Orengo *et al*., 1997; Shapiro and Brutlag, 2004) and topological levels (Gilbert *et al*., 2001). Recently, the multiple alignment of protein 3D structures based on their residue positions (C-alpha) has been offered by the server CE-MC (Guda *et al*., 2004). However, to date no method explores local atomic-level similarity based on multiple comparisons.

An alternative line of research has been the description of active sites in terms of geometry, charge, and hydrophobic/hydrophilic character. Techniques are generally based on the detection of surface cavities and on their abstract description (Oldfield, 2002; Schmitt *et al*., 2002; Campbell *et al*., 2003; Jambon *et al*., 2005; Jones and Thornton, 2004). While they have been very efficient in detecting active site similarities between non-homologous proteins, they do not offer multiple comparisons and do not provide atomic descriptions.

In this piece of research, we investigated the simultaneous structural alignments of proteins to generate high-resolution 3D patterns of the regions of their active sites. These patterns not only provide 3D positions of atoms, but also positions of chemical groups and cavity locations. We focused our efforts on proteins with rigid prosthetic groups such as porphyrin rings (Fig. 1). These proteins are of particular interest, because they include haem and chlorophyll proteins which are critical components in biological processes ranging from photosynthesis and aerobic respiration to drug detoxification. In particular, P450 enzymes are haem proteins which are responsible for the metabolism of drugs and therefore influence drug clearance, toxicity and activation. Consequently, they have been drug targets for decades. Moreover, since the rigid prosthetic group of those proteins is a key element of their active sites, generated patterns based on these groups are expected to be biologically meaningful.

In this paper, we present the technique we developed to generate 3D patterns from the alignment of multiple protein 3D structures. Then, our method is validated by processing all representatives of proteins with porphyrin rings. Finally, a 3D template is produced for

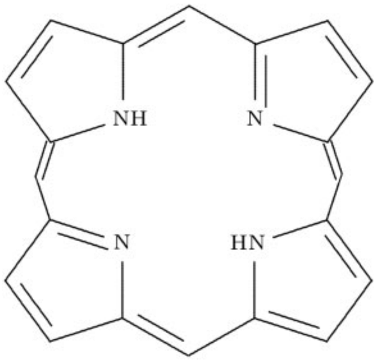

Fig. 1. Porphyrin structure.

the region of the active site of a P450 protein the 3D structure of which is unknown. Its validity is accessed by confrontation with biological data and 3D models which were created independently.

2 METHODS

2.1 Principle

Our new method is based on the comparison of sets of homologous proteins the 3D structures of which are known. Protein structures are aligned according to the position of their rigid prosthetic group. Then, from that multiple alignment, a consensus 3D pattern is produced. It contains 3D positions of three different types of structural information: protein atoms, chemical groups and solvent atoms which define the cavity space associated to the active site. By comparing each protein to all the others of the set in one-to-one comparisons 3D patterns are generated. During these comparisons, atoms, chemical groups and solvent atoms which cannot be paired within a given threshold are discarded.

Seven non-exclusive chemical groups were defined according to the characteristics of residue side chains relevant to potential biochemical interactions with ligands: acidic (D and E), basic (R, H and K), amidic (N and Q), hydroxyl (S, T and Y), aromatic (F, W and Y), non-polar (A, G, I, L, P, V and M) and potential for a sulphide bond (C). In the group representation of a protein, each residue is replaced by one or two virtual atoms (e.g. tyrosine is represented by both hydroxyl and aromatic groups) located at the centre of mass of the group they represent. Each protein is also preprocessed so that the cavity, which is involved in the active site defined by its prosthetic group, is filled in a regular manner with solvent atoms.

2.2 Protein alignment and pattern generation

Proteins are aligned by performing rigid transformations between them according to the atom positions of their rigid prosthetic group. Atom correspondences are based on the ‘Atom name’ field in the HETATM record of the PDB file format. When proteins are composed of several chains, only the first chain containing a prosthetic group is utilized. We implemented an algorithm developed by Horn (1987) to determine the translation and rotation that will align atoms in one coordinate system to corresponding atoms in another coordinate system, while minimizing the total distance between the two sets of atoms. The rigidity of porphyrin rings is such the root of mean square deviations (RMSDs) between the 24 atoms of the aligned porphyrin rings is <0.2 Å.

Once the proteins are aligned, the process of 3D pattern generation begins. For each of the three types of structural elements—atoms, chemical groups and solvent atoms produced by the cavity generation method (see Section 2.3)—the same process is followed. For a given structural element, e.g. atoms, each protein of a set is compared with all the others in one-to-one comparisons.

During a pairwise comparison, each atom of the first protein, P_i, is paired with the nearest atom of the same type (carbon, oxygen, nitrogen or sulphur) belonging to the second protein, P_j. If their distance is above a given threshold, that atom is discarded so that it will not belong to the consensus pattern of P_i, C_{ij}. At the end of the comparison, both proteins only contain atoms which could be paired and, therefore, constitute their common pattern. The same process is performed until each protein is compared with all the other, so that at the end of the process each protein comprises the atoms belonging to the consensus pattern of the set of proteins. The atomic consensus 3D pattern is then generated by averaging the 3D positions of the atoms conserved for the whole set of proteins. Each atom of the pattern has a type and may also be linked to a consensus residue, if relevant. The pattern generation method is similar for chemical groups and solvent atoms.

The process of generation of a consensus pattern, *C*, is described below in pseudocode, where *n* is the number of proteins in the set:

```
For i=1 to n-1
  For j=i+1 to n
    Structural alignment of $P_i$ & $P_j$
    Keep consensus structural elements of $P_i$ & $P_j$: $C_{ij}$ & $C_{ji}$
    $P_i=C_{ij}$
    $P_j=C_{ji}$
  endFor
endFor
For i=1 to n
  C += $P_i/n$
endFor
```

The consensus pattern can also be shown as a partial sequential alignment of the proteins of the set based on their structural alignment. Since atoms are structurally aligned, the residues they belong to can be considered as structurally aligned too. Therefore, a partial multiple sequential alignment can be generated where structurally aligned residues are shown (an example is given in Fig. 4).

2.3 Cavity generation

Although efficient tools are available for the recognition of ligand binding sites and protein surface cavities (Liang *et al*., 1998; Brady and Stouten, 2000), experiments showed they are not reliable when dealing with large cavities such as those of haem-based active sites. Furthermore, since they were made essentially for visualization purposes, the data they output would require further processing to be adapted to our application. Consequently, we developed a new cavity generation method taking into account the specificity of the data we deal with (proteins with rigid prosthetic groups) and the requirements of the 3D pattern generation algorithm.

Since rigid prosthetic groups are an active element of the protein active site, their atoms can be used to define a plane (e.g. using planar ring structures) to divide the protein space in two areas: one area where the group is attached to the protein and another one where the binding of the group with a ligand takes place. For example, the iron fifth ligand of the haem group of haem-thiolate proteins is a cysteine situated on one side of the haem group, while the cavity is on the other side. The direction, which is perpendicular to the division plane and which goes away from the covalent bond between the protein and the group, is defined as the cavity direction. Using the cavity direction and the centre of the prosthetic group as the origin of the cavity, the cavity is filled in a regular manner with cavity elements. A 3D regular grid is created in the cavity half space: the *z*-axis of the grid is aligned with the cavity direction, the *x*- and *y*-axes are defined by atoms from the division plane (e.g. the directions can be given by the four nitrogen atoms of porphyrin rings, Fig. 1).

First, empty grid elements are detected: each grid element whose neighbourhood does not contain any atom centre is set to empty. Then, we reject empty elements which cannot be reached from the origin of the cavity by a continuous chain of empty elements. Since part of the grid is outside the protein space, empty elements not belonging to the prosthetic group cavity may be connected to the cavity origin. Therefore, empty grid elements which are outside the protein space need to be discarded. We consider an empty grid element is outside the protein space if at least three straight lines at 45° from each other can be drawn from the grid element without intersecting any non-empty grid element. Finally, the cavity shape is built by continuity using only the empty elements which are inside the protein space. Since these empty grid elements can be seen as solvent atoms, cavity structures associated with rigid prosthetic groups can then be processed using the pattern generation algorithm.

3 VALIDATION

3.1 Methodology

In order to validate our method, 3D templates were generated using homologues for all representatives of proteins containing porphyrin rings present in the Protein Data Bank (PDB) (Berman *et al*., 2000). Each template is then compared with the PDB structure of the protein it models.

The family of prosthetic groups we are interested in is represented by 12 different PDB molecule codes including haem (e.g. HEM, HEC and HEA) and chlorophyll groups (e.g. BCL and CLA). The PDB holds 1551 proteins containing these groups, i.e. 5.3% of PDB entries (as of February 1, 2005). From that dataset, homologues with at least a 50% sequence identity were removed (PDB50%). Then, identical chains and chains that are not involved with a prosthetic group were removed. Finally, 237 chains were kept as representatives of the class of proteins containing porphyrin rings.

Within the dataset, each chain sequence was aligned with all the others using FASTA (Pearson and Lipman, 1988). Then for each chain, a set of homologous proteins, which were defined as having an *E*-value below a given threshold, was selected. A 3D pattern was calculated from their 3D structures and was used as a 3D template of the active site of the protein they are homologous to. The selected proteins were aligned by rigid registration according to the positions of the 24 atoms (20 carbon and 4 nitrogen atoms) lying on the rigid plane of their porphyrin group (Fig. 1): RMSD between the different sets of 24 atoms was found to be under 0.15 Å. Atom correspondences are unambiguous since the nomenclature of haem and chlorophyll groups in the PDB depends on the position of the covalent bond between the porphyrin ring and the peptide chain and the position of the side chains of the porphyrin ring.

Patterns were calculated only if at least three homologous proteins were available. Finally, each protein and its associated predicted template were compared using the same parameters as the ones used for generating the template. Structural elements of the template which could be paired with elements of the protein were marked as true predictions. Otherwise, they were marked as false predictions.

3.2 Results

In this section, we present the results of comparisons between generated templates and the protein structures they model. Templates were produced using a variety of *E*-value and distance thresholds. The different ranges of distance thresholds were set according to the following considerations. Eyal carried out a statistical study on a large number of proteins, the structures of which were determined at least twice by X-ray crystallography (Eyal *et al*., 2005). They observed significant differences between structures: RMSD based on alpha carbon atoms could reach 0.9 Å. Therefore, a minimum threshold of 1.0 Å was chosen when generating consensus atom positions. Moreover, since the length of a covalent bond between two atoms of a protein can reach 1.5 Å, this value was selected as the upper limit of 'atomic' resolution. Similarly, a value of 4.0 Å was set as the upper limit of 'group' resolution, since it is the maximal distance between two adjacent alpha carbon atoms.

Table 1. Properties of atom templates depending on set parameters

Max *E*-value	Templates	Homologues	Atom distance in Å	Atoms	True predictions (%)
1.0e−8	54	6.0	1.00	103	93
			1.25	153	91
			1.50	245	90
1.0e−6	66	6.2	1.00	86	93
			1.25	130	92
			1.50	208	92
1.0e−4	78	7.1	1.00	68	94
			1.25	92	92
			1.50	160	92
1.0e−2	91	8.3	1.00	63	95
			1.25	92	93
			1.50	141	92

In Table 1 results are reported regarding atom templates generated from our dataset. It provides for each *E*-value threshold, the number of produced templates and the average number of homologues used for generating them. Then, it gives for different atom distances the average number of atoms that are present in the produced templates and the average percentage of true predictions.

Whatever the values of the parameters, the average number of true predictions is between 90 and 95%. Parameter values mainly impact on the number of atoms present in the generated templates: the lower is the *E*-value threshold or the higher is the atom distance threshold, the more atoms constitute the templates. Similar results are found with group positions, where the number of true predictions is around 83%.

Figure 2 shows for each of the 66 templates, which were produced using an *E*-value threshold of 10e−6 and an atom distance of 1.25 Å, the number of atoms which are either true or false predictions. The diagram shows that templates tend to cluster under the 90% true prediction threshold. The variability in the number of elements generated by our method is linked to the number of homologue structures: their increase produces an increase of the true prediction rate and a decrease of the number of atoms.

In Figure 3, we present the distribution of true predictions in the dataset shown in Figure 2. It is observed that >75% of the templates have >90% of true predictions; 30% of them reach a rate of 100%. Similar results were found using the other parameter settings. By averaging the results obtained in the experiments reported in Table 1, it appears only 22.4% of templates have a number of true predictions which is under 90%. Moreover, just two proteins—1U5U (Oldham *et al*., 2005) and 1S05 (Bertini *et al*., 2004)—have templates with a true prediction rate <53%. The poor quality of our templates for these two models can be explained by studying their PDB files: 1U5U is a protein fragment and 1S05 is

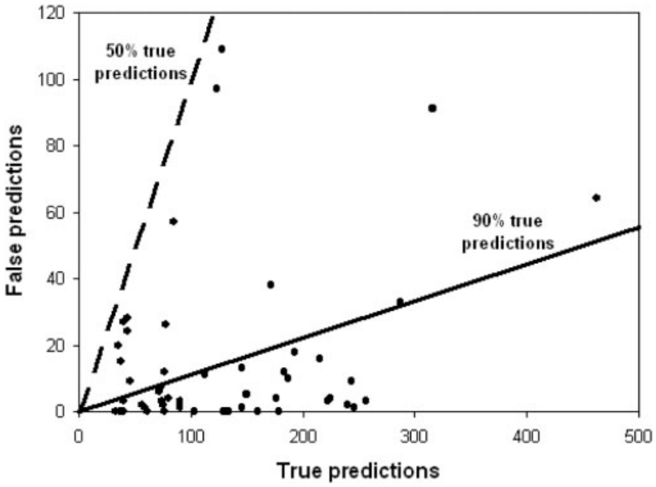

Fig. 2. Number of atom positions as true and false predictions.

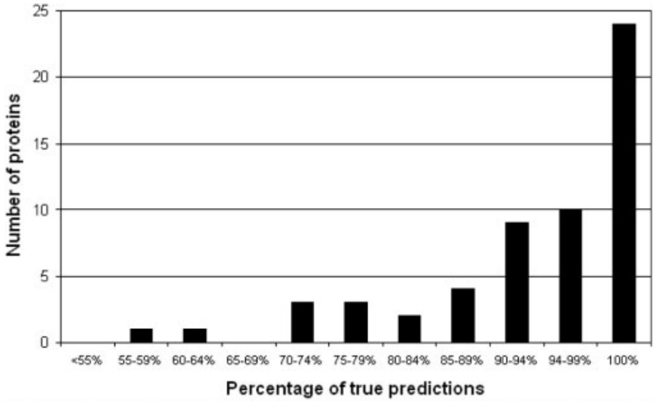

Fig. 3. Distribution of true predictions of atom positions.

actually a structural model which was validated using a restricted set of NMR experiments.

Since solvent size is an essential parameter in the cavity generation algorithm and it is related to the type of molecule which can potentially bind to the active site, a unique size cannot be used for processing the whole dataset. Therefore, the validation of the generation of cavity templates was done by considering proteins belonging to a well-defined family. We chose the cytochrome P450 superfamily (P450s), because these enzymes have been the most popular research topic in biochemistry and molecular biology over the past half century (Lewis, 2001). Indeed, their haem-based active sites are involved in the metabolism of numerous substrates such as drugs, carcinogens and sex hormones.

This superfamily is represented by 18 proteins in our dataset. The 3D grid used for detecting cavities had a sampling step of 1.25 Å. Experimentally, we found that cavity shapes were best described using cubic elements (or solvent) of side length 4.8 Å. Therefore, any empty cubic space of side length superior to 6.05 Å is detected. This value is consistent with CASTp (Liang *et al.*, 1998) where solvent space is detected if centres between atoms are distant of at least 5.6 Å for two oxygen and 6.3 Å for two carbon atoms (distance is solvent diameter, 2.8 Å, plus sum of van der Waals radii).

Table 2. Templates generated for the P450 dataset

		Average	Worst case
Atoms	Number	101	62
	True predictions	97%	82%
Groups	Number	21	10
	True predictions	93%	76%
Solvent atoms	Number	82	52
	True predictions	92%	71%

Table 2 shows results for the P450 dataset, where we used an *E*-value of 1.0e−6, atom and solvent distances of 1.25 Å and a group distance of 3.0 Å. A total of 16 templates were generated based on 8.3 homologues in average. Compared with the results presented previously, the results are significantly better. Whatever the type of structural element, cavity included, the average true prediction rate is >92%. Moreover the true prediction rate is high even in the worst cases. That is explained by the fact the active sites of the P450 superfamily are very well conserved.

3.3 Discussion

The processing of PDB representatives of proteins containing porphyrin rings showed the validity of our approach since the positions of the structural elements predicted agreed extremely well with their positions as recorded in the PDB. Atom positions were predicted accurately with >90% of true predictions for thresholds within the 1.0–1.5 Å range and similar results were obtained for chemical group and cavity positions. In addition, experiments with a well-defined protein family suggest that prediction rates can significantly be increased if homologues are carefully chosen. In a preliminary study, we aligned ATP binding proteins according to the position of the adenine group of the ATP molecule. Results were very encouraging and suggest our technique can be applied to a large range of proteins binding either rigid or semi-rigid molecules, i.e. 20% of PDB entries.

The analysis of consensus 3D patterns generated by our technique will no doubt permit a better understanding of the properties of active sites. Moreover, we anticipate our method will allow improving the quality of protein structure prediction by providing templates which could be used as additional constraints for current structure modelling techniques. Similarly, as the detection of structural anomalies with the 1S05 model suggests, our method could also play a part in the validation process of predicted structures. Finally, we believe our 3D templates will contribute to rational drug design by providing high-resolution data, up to 1.0 Å, for active sites the structures of which are unknown. Our technique would allow overcoming some of the limitations of traditional homology modelling such as reliance on sequence identities >30% and accuracy rarely under a RMSD of 2.0 Å (Tramontano and Morea, 2003). An example of the latter application is given in the next section.

4 APPLICATION: 3D TEMPLATE OF HUMAN CYTOCHROME P450 CYP17

The human cytochrome P450 CYP17 (CYP17) was chosen to illustrate an application of our method. First, it is a protein of great

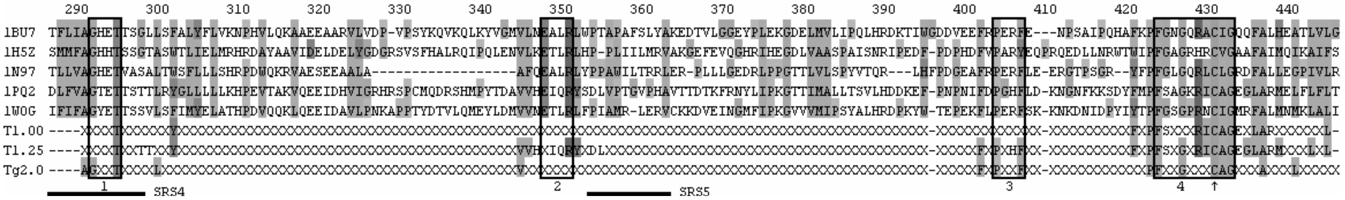

Fig. 4. Alignment of homologue and template sequences using ClustalW. Four regions representing significant P450 patterns are highlighted: (1) oxygen-binding site, (2) ion-pair, (3) PERF motif and (4) cytochrome P450 cysteine haem-iron ligand signature. The putative substrate recognition sites SRS4 and SRS5 are underlined.

Table 3. CYP17 homologues using FASTA on PDB50%

PDB ID	Length	Identity (%)	Overlap	*E*-value
1PQ2_A	476	28.7	478	9.0e−42
1W0G_A	485	28.3	481	2.5e−26
1BU7_A	455	29.4	269	1.2e−16
1H5Z_A	455	24.1	262	6.7e−09
1N97_A	389	26.4	231	4.0e−05
1IZO_A	*417*	*24.0*	*146*	*0.00011*
1AKD	*417*	*29.5*	*61*	*1.5*

interest for the pharmaceutical community: it is involved in key steps leading to the biosynthesis of sex hormones and its potential association to several forms of cancers—breast (Hefler *et al*., 2004; Shin *et al*., 2005) and prostate (Cicek *et al*., 2004; Douglas *et al*., 2005)—is an active field of research. Second, although its 3D structure is still unknown, CYP17 has distant sequential homologues—their sequence similarity is <30%—the structures of which are known. Third, previous results showed the P450 superfamily is a good candidate for our pattern generation method. Finally, the modelling of its active site has already been attempted (Ahmed, 2004). Once the template is generated, its relevance is assessed: first, it is verified that its main features can be explained by biological data; second, it is compared with 3D models of the active site of CYP17 which were generated independently.

4.1 3D template generation

Homologous proteins were selected by aligning, using FASTA, the sequence of human CYP17 to sequences from PDB50%. For each homologue class defined by the 50% threshold, a non-mutated representative was chosen (Table 3). Out of the main hits, the first five were selected because they combine low *E*-value and the size of their best overlapping region, 'Overlap', is quite large. These P450 proteins are quite distant homologues: they all belong to the twilight zone with pairwise sequence similarities found in the range of 20–30%, 'Identity', for their best overlapping region. They also come from very different sources, since the first two are from humans and the other three are from bacteria.

In this study, the consensus 3D template was generated using atom distances of 1.00 and 1.25 Å and a group distance of 2.00 Å. Atom and group templates were, respectively, named T1.00, T1.25 and Tg2.00. As previously, cubic elements of side length 4.80 Å were chosen to detect cavity space. Three-dimensional templates are analysed either as sequences aligned with their homologues where residues taking part in the template are shown (Fig. 4) or as a consensus 3D structure (Fig. 5). As expected, there is a high concentration of conserved atom and group positions in the area surrounding the haem group. For example, T1.00 is made of 43 atoms belonging to 16 different residues and Tg2.0 contains 17 groups. In particular, the cysteine binding to the haem group and its neighbours are very well conserved in the atom templates. Moreover, the consensus cavity space is large and well defined as a 13 Å long and 5 Å diameter pocket located at an angle of around 60° from the haem plane (Fig. 5).

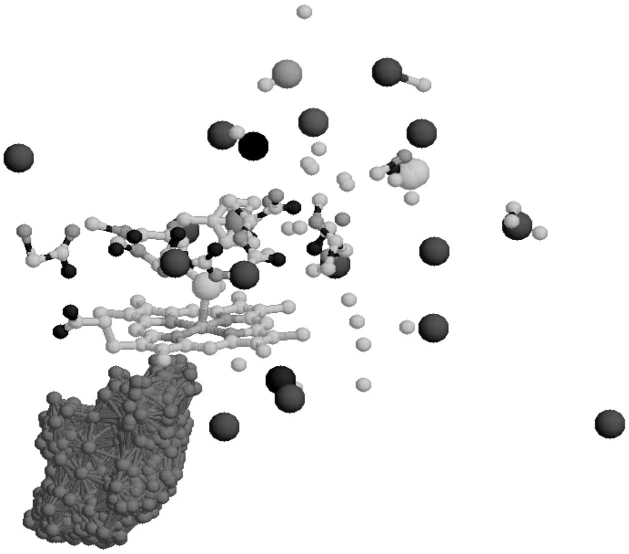

Fig. 5. 3D template of human cytochrome P450 CYP17. Atoms are represented by small spheres, chemical groups are modelled by larger spheres and cavity space is dark grey.

4.2 Comparison with known P450 sequence patterns and motifs

Figure 4 shows consensus elements cluster in four regions of the aligned sequences. Analysis of 3D structures shows these clusters are geometrically close to each other. In the PROSITE database (Falquet *et al*., 2002), which holds biologically significant patterns, there is a single entry for P450s. Pattern PS00086 is the P450 cysteine haem-iron ligand signature which is defined as

[FW]-[SGNH]-x-[GD]-x-[RKHPT]-x-C-[LIVMFAP]-[GAP].

All amino acids of the signature are represented in T1.25 [Fig. 4 (4)]. Moreover, this signature is located at the core of the longest sequential pattern of the template. The Catalytic Site Atlas (CAS, Porter *et al*., 2004) is a new database documenting enzyme active sites and catalytic residues in enzymes of 3D structure. CAS also contains one entry for P450 proteins: 1AKD (P450cam) contains a proton transfer network composed of Asp-251 and Thr-252 (Hishiki *et al*., 2000). These residues belong to the P450 conserved tetrapeptide G-x-[DEH]-T which is believed to represent an oxygen-binding site and point of access for an incoming dioxygen molecule (Poulos *et al*., 1995). That pattern, [Fig. 4 (1)], is also present in the template: the threonine is conserved among all the proteins of our set and is represented in T1.00. In addition, this threonine is completed by the glycine of the tetrapeptide in Tg2.0. Finally, the regions of the E-x-x-R and P-E-R-F motifs, which are among the most conserved motifs in all P450s (Lewis, 2001), are detected by our technique as shown in Figure 4 (2) and (3). The ion-pair, E-x-x-R, seems to be particularly important since it is thought to participate in both the redox partner interaction and haem binding (Peterson and Graham-Lorence, 1995).

Finally, locations of the consensus elements were compared with the six restricted regions of P450s termed substrate recognition sites (SRS) (Gotoh, 1992). Since they predetermine the substrate specificity of the enzymes, a method based on consensus positions should not be able to deliver much information regarding those regions. Only SRS4 and SRS5, (Fig. 4), contain some consensus elements, which is consistent with the fact that both sites are known as being highly conserved in their location relative to the haem group (Johnson, 2003).

4.3 Comparison with other CYP17 models

In the active site region, on the side of the haem group which is not linked to the cysteine, the 3D positions of three groups (1 hydroxyl group, threonine residue, and two non-polar groups, alanine and glycine) are very well conserved, i.e. RMSD < 1.2 Å (Fig. 6). Moreover, alignment of the sequence of CYP17 with our set of proteins confirmed that these three residues are conserved in CYP17 too (Thr-306, Ala-302 and Gly-303). Among these residues, the threonine is of particular interest because its hydroxyl group can potentially produce hydrogen bonds (H-bonds). Since it is located near the cavity space, it may be involved in H-bonds with substrates.

In order to test this hypothesis about putative H-bonds, our CYP17 template was compared with models of enzyme complexes involving human CYP17. Models of 17alpha-hydroxylase and 17,20-lyase (lyase) were provided by Ahmed: they had been produced using a novel molecular modelling technique, named the substrate-haem complex (Ahmed, 2004). These models were then aligned to our 3D template using haem group positions as reference (Fig. 6).

In both cases, the potential presence of a H-bond is confirmed since distances between the oxygen atoms of the Thr-306 and the substrate are between 2.5 and 2.8 Å. Furthermore, there is a very good match between the lyase and consensus cavity positions (Fig. 6b). In the other case (Fig. 6a), the mismatch between the substrate and the consensus cavity space can be explained by the fact that CYP17 is known to contain two cavity lobes whereas all proteins from our set of homologues contain only one (Ahmed, 2004).

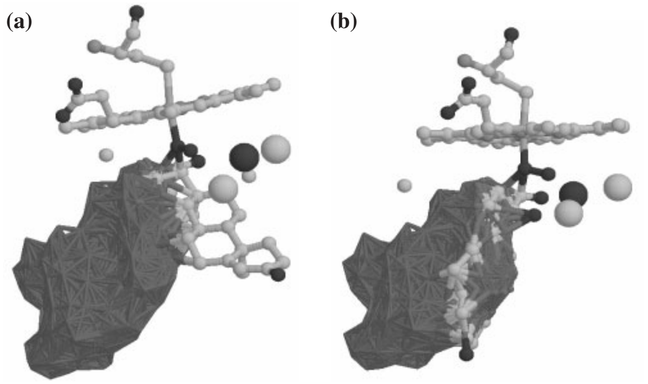

Fig. 6. Alignment of CYP17 template with models of (**a**) 17alpha-hydroxylase and (**b**) lyase. Hydroxyl and non-polar groups are, respectively, modelled by black and grey spheres.

4.4 Discussion

The analysis of the CYP17 3D template has shown that, although it does not provide a complete description of the active site region, it supplies high-resolution structural information which is biologically meaningful. Our method detected the main patterns of the cytochrome P450 superfamily and the motifs linked to catalytic reactions. In addition, it highlighted other residues which may have not yet been recognized as important elements of the protein activity. Comparisons between our 3D template and independently generated 3D models of the active site of CYP17 showed some evidence of the presence of a H-bond predicted by our template. Moreover, the shape and location of consensus cavity space was confirmed by the lyase complex model.

Results obtained with the CYP17 models are significant because they indicate our 3D template could be exploited either by completing current active site models used for *de novo* design of novel inhibitors or by providing constraints for 3D structure modelling.

5 CONCLUSION

In this paper we have introduced a novel method for the generation of high-resolution 3D patterns from the alignment of protein 3D structures. It can be applied to any type of proteins with rigid prosthetic group and produces accurate structural information which appears to be biologically meaningful. We are also confident our technique can be used with proteins binding semi-rigid molecules such as ATP. The analysis of the consensus 3D patterns generated by our technique will no doubt permit a better understanding of the properties of active sites. These patterns complete sequential patterns and motifs by providing structural information at the atomic, chemical group and cavity levels. Moreover, they may extend or detect patterns which are not conserved at the residue level. Results not presented in this paper also showed that although ClustalW alignments were generally consistent with our structural consensus, in some cases they could be refined using the generated 3D patterns.

Furthermore, we anticipate our method will allow improving the quality of protein structure prediction by providing templates which could be used as additional constraints for current structure prediction techniques. Similarly, our method could also play a part in the validation of predicted structures. Finally, we believe 3D templates will contribute to drug design by providing high-resolution structural information about active sites of proteins the structure of which is unknown.

ACKNOWLEDGEMENTS

The author would like to thank Dr Sabbir Ahmed for helpful discussions and for providing the CYP17 models. Funding to pay the Open Access publication charges for this article was provided by Kingston University.

Conflict of Interest: none declared.

REFERENCES

Ahmed,S. (2004) The use of the novel substrate-heme complex approach in the derivation of a representation of the active site of the enzyme complex 17alpha-hydroxylase and 17,20-lyase. *Biochem. Biophys. Res. Commun.*, **316**, 595–598.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bertini,I. *et al.* (2004) NMR-validated structural model for oxidized *Rhodopseudomonas palustris* cytochrome c(556). *J. Biol. Inorg. Chem.*, **9**, 224–230.

Brady,G.P. and Stouten,P.F.W. (2000) Fast prediction and visualization of protein binding pockets with PASS. *J. Computer-Aided Mol. Design*, **14**, 383–401.

Campbell,S.J. *et al.* (2003) Ligand binding: functional site location, similarity and docking. *Curr. Opin. Struct. Biol.*, **13**, 389–395.

Cicek,M.S. *et al.* (2004) Association of prostate cancer risk and aggressiveness to androgen pathway genes: SRD5A2, CYP17, and the AR. *Prostate*, **59**, 69–76.

Douglas,J.A. *et al.* (2005) Identifying susceptibility genes for prostate cancer—a family-based association study of polymorphisms in CYP17, CYP19, CYP11A1, and LH-beta. *Cancer Epidemiol. Biomarkers Prev.*, **14**, 2035–2039.

Eyal,E. *et al.* (2005) The limit of accuracy of protein modeling: influence of crystal packing on protein structure. *J. Mol. Biol.*, **351**, 431–442.

Falquet,L. *et al.* (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.

Gibrat,J.F. *et al.* (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.

Gilbert,D. *et al.* (2001) A computer system to perform structure comparison using TOPS representations of protein structure. *Comput. Chem.*, **26**, 23–30.

Gotoh,O. (1992) Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences. *J. Biol. Chem.*, **267**, 83–90.

Guda,C. *et al.* (2004) CE-MC: a multiple protein structure alignment server. *Nucleic Acids Res.*, **32** (Web Server issue), W100–W103.

Hefler,L.A. *et al.* (2004) Estrogen-metabolizing gene polymorphisms in the assessment of breast carcinoma risk and fibroadenoma risk in Caucasian women. *Cancer*, **101**, 264–269.

Hishiki,T. *et al.* (2000) X-ray crystal structure and catalytic properties of Thr252Ile mutant of cytochrome P450cam: roles of Thr252 and water in the active center. *J. Biochem.*, **128**, 965–974.

Holm,L. and Sander,C. (1994) Searching protein structure databases has come of age. *Proteins*, **19**, 165–173.

Horn,B.K.P. (1987) Closed-form solution of absolute orientation using unit quaternions. *J. Optical Soc. Am.*, **4**, 629–642.

Jambon,M. *et al.* (2005) The SuMo server: 3D search for protein functional sites. *Bioinformatics*, **21**, 3929–3930.

Johnson,E.F. (2003) Deciphering substrate recognition by drug-metabolizing cytochromes P450. *Drug Metab. Dispos.*, **31**, 1532–1540.

Jones,S. and Thornton,J.M. (2004) Searching for functional sites in protein structures. *Curr. Opin. Chem. Biol.*, **8**, 3–7.

Lewis,D. (2001) *Guide to Cytochromes P450: Structure & Function*. Taylor & Francis, London.

Liang,J. *et al.* (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.*, **7**, 1884–1897.

Oldfield,T.J. (2002) Data mining the protein data bank: residue interactions. *Proteins*, **49**, 510–528.

Oldham,M.L. *et al.* (2005) The structure of coral allene oxide synthase reveals a catalase adapted for metabolism of a fatty acid hydroperoxide. *Proc. Natl Acad. Sci. USA*, **102**, 297–302.

Orengo,C.A. *et al.* (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Peterson,J.A. and Graham-Lorence,S.E. (1995) Bacterial P450s: structural similarities and functional differences in Cytochrome P450. In Ortiz de Montellano (ed.), *Structure, Mechanism, and Biochemistry*, 2nd edn. Plenum Press, NY, pp. 151–180.

Porter,C.T. *et al.* (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.

Poulos,T.L., Cupp-Vickey,J. and Li,H. (1995) Structural studies on prokaryotic cytochromes P450, Cytochrome P450. In Ortiz de Montellano Editor (ed.), *Structure, Mechanism, and Biochemistry.* Plenum Press, NY, pp. 125–150.

Schmitt,S. *et al.* (2002) A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.*, **323**, 387–406.

Shapiro,J. and Brutlag,D. (2004) FoldMiner and LOCK 2: protein structure comparison and motif discovery on the web. *Nucleic Acids Res.*, **32**, W536–W541.

Shin,M.H. *et al.* (2005) Genetic polymorphism of CYP17 and breast cancer risk in Korean women. *Exp. Mol. Med.*, **37**, 11–17.

Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.

Thompson,J.D. *et al.* (1994) ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Tramontano,A. and Morea,V. (2003) Assessment of homology-based predictions in CASP5. *Proteins*, **53** (Suppl. 6), 352–368.