

# 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, United Kingdom  
j.nebel@kingston.ac.uk

**Abstract:** With the increasing availability of protein structures, the generation of biologically meaningful 3D patterns from their simultaneous alignment is an exciting prospect: active sites could be better understood, protein functions and structures could be predicted more accurately. Although patterns can be generated at the fold and topological levels, no system produces high resolution 3D patterns including atom and cavity positions. Here, we present a new approach allowing the generation of 3D patterns from alignment of proteins with rigid prosthetic groups. Using 237 proteins representing these proteins, our method was validated by comparing 3D templates generated from homologues with structures of the proteins they model. Atom positions were predicted reliably: 93% of them had an accuracy below 1.00 Å. Similar results were obtained regarding chemical group and cavity positions. Finally, a 3D template was generated for the active site of human cytochrome P450 CYP17. Its analysis showed it is biologically meaningful: our method detected the main patterns and motifs of the P450 superfamily. The 3D template also suggested the locations of a cavity and of a hydrogen bond between CYP17 and its substrates. Comparisons with independently generated 3D models comforted these hypotheses.

## 1 Introduction

The simultaneous alignment of several protein sequences using tools such as ClustalW [Th94] 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 pair wise protein structure comparisons based on atom coordinates [Ho94], [Gi96] and [Sh98], 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 [Or97] and [Sh04] and topological levels [Gi01]. Recently, the multiple alignment of protein 3D structures based on their residue positions (C-alpha) has been offered by the server CE-MC [Gu04]. 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 [Ol02], [Sc02] [Ca03], [Ja02] and [Jo04]. While they have been very efficient in detecting active site similarities between non homologue 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 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 (Figure 1). These proteins are of particular interest, because their rigid prosthetic group is a key element of their active sites. Therefore, generated patterns are expected to be biologically meaningful.

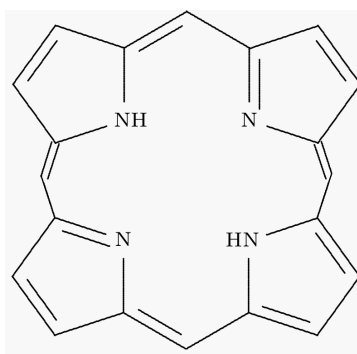


Figure 1: Porphyrin structure

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 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 Method

### 2.1 Principle

Our new method is based on the comparison of sets of homologue 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 three different types of structural information: atom, chemical group and cavity 3D positions. 3D patterns are generated by comparing each protein to all the others of the set in one to one comparisons. During these comparisons, elements which cannot be paired within a given threshold are discarded.

Seven non exclusive chemical groups were defined according to the properties of residue side chains: acidic, basic, amide, hydroxyl, aromatic, sulphur and non polar. In the group representation of a protein, each residue is replaced by one or two virtual atoms (e.g. tyrosine is represented by hydroxyl and aromatic groups) located at the centre of mass of the group they represent. Each protein is also pre-processed 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

Set of proteins are aligned by performing rigid transformations between them according to the atom positions of their rigid prosthetic group. When proteins are composed of several chains, only the first chain containing a prosthetic group is utilised. We implemented an algorithm developed by Horn [Ho87] 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 groups such as porphyrin rings is such the Root of Mean Square Deviations (RMSD) between atoms of aligned prosthetic groups is below 0.2 Å.

Once the proteins are aligned, the process of 3D pattern generation begins. Each protein is compared to all the others in one to one comparisons. It is important to ensure that all combinations of protein comparison are performed so that the order of the chosen proteins does not impact on the generation of the 3D pattern. Moreover, in order to provide consistency and fast computation only elements selected at a given stage of the comparison process are used during the next stage.

The pattern generation method is applied on three types of structural elements: protein atoms, chemical groups and solvent atoms produced by the cavity generation method. For each type of element, comparisons between two proteins involve the calculation of distances between similar elements: same atom type (carbon, oxygen, nitrogen or sulphur), equivalent chemical groups or solvent atoms. If an element cannot be paired within a given threshold, it is discarded so that it will not belong to the consensus pattern. At the end of the procedure, each protein contains an identical number of selected elements the positions and types of which are conserved among the set of

proteins being processed. The first output of the method is a sequential alignment of the proteins of the set based on their structural alignment. Secondly, a consensus 3D pattern is generated: 3D positions of the three types of consensus elements are calculated by averaging positions of elements conserved for the whole set of proteins, each consensus atom or group is also provided with a type (including consensus residue, if relevant).

### **2.3 Cavity generation**

Although efficient tools are available for the recognition of ligand binding sites and protein surface cavities [Li98] and [Br00], 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 visualisation 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 and they involve planar ring structures, they can be used 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 the 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. 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 empty 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 planar ring structure of reference (e.g. the directions can be given by the four nitrogen atoms of porphyrin rings, Figure 1).

First, empty grid elements are detected: each grid element whose neighbourhood does not contain any atom 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, cavity elements which are outside the protein need to be discarded. Rays are cast from each empty element at 45 degrees from each other; if at least three continuous lines of empty elements can be found the empty element is marked as outside the cavity. Finally, the cavity shape is rebuilt by continuity using the remaining empty elements. Since these empty 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) [Be00]. 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 1<sup>st</sup> February 2005). From that data set, 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 data set, each chain sequence was aligned with all the others using FASTA [Pe88]. Then for each chain, a set of homologue 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 homologue 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 (Figure 1): RMSD between the different sets of 24 atoms was found to be under 0.15 Å. Patterns were calculated only if at least 3 homologue 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 positives. Otherwise, they were marked as false positives.

### 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 et al. carried out a statistical study on a large number of proteins, the structures of which were determined at least twice by X-ray crystallography [Ey04]. 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.

Max E value	Models	Homologues	Atom distance in Å	Atoms	True positives
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%

Table 1: Properties of atom templates depending on set parameters

In Table 1 results are reported regarding atom templates generated from our data set. 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 positives.

Whatever the values of the parameters, the average number of true positives 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 positives is around 83%.

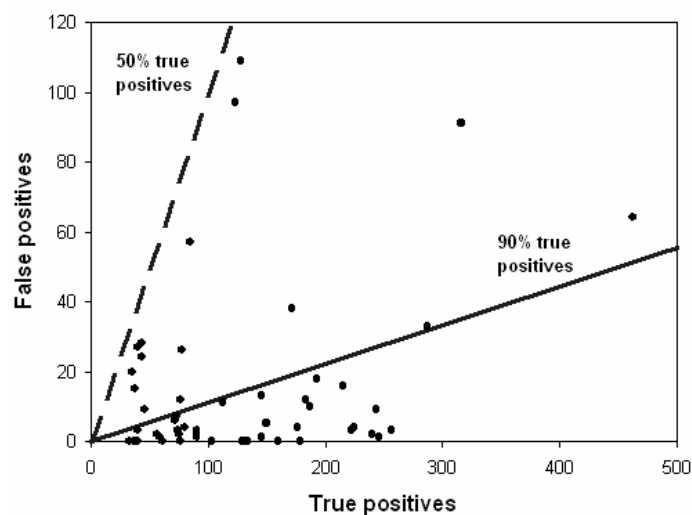


Figure 2: Number of true and false positive atoms

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  $1.25 \text{ \AA}$ , the number of atoms which are either true or false positives. The diagram shows that templates tend to cluster under the 90% true positive 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 positive rate and a decrease of the number of atoms.

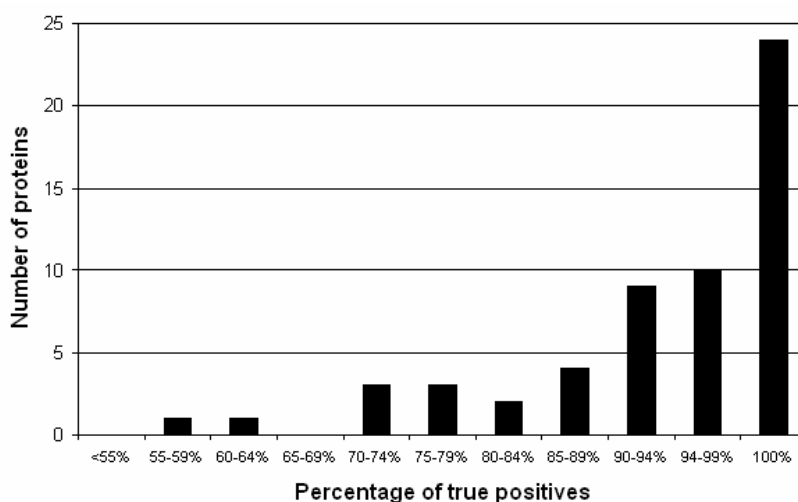


Figure 3: Distribution of true positive atoms

In Figure 3, we present the distribution of true positives in the data set shown in Figure 2. More than 75% of the templates have more than 90% of true positives; 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 positives which is under 90%. Moreover, just two proteins - 1U5U [OI05] and 1S05 [Be04] - have templates with a true positive rate below 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 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 data set. 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 [Le01]. Indeed, their haem-based active sites are involved in the metabolism of numerous substrates such as drugs, carcinogens and sex hormones.

		Average	Minimum
Atoms	Number	101	62
	True positives	97%	82%
Groups	Number	21	10
	True positives	93%	76%
Solvent atoms	Number	82	52
	True positives	92%	71%

Table 2: Templates generated for the P450 data set

This superfamily is represented by 18 proteins in our data set. Experimentally, we found that cavity shapes were best described using a solvent size of 4.8 Å. Table 2 shows results for the P450 data set, 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. 16 templates were generated based on 8.3 homologues in average. Compared with the results obtained with the whole data set, the results are significantly better. Whatever the type of structural element - cavity included -, the true positive rate is above 92% and minima are high. 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 with accuracy above 90% of true positives for thresholds within the 1.0 to 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, the detection of structural anomalies in the structure of the 1S05 protein 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 above 30% and accuracy rarely under a RMSD of 2.0 Å [Tr03]. 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. Firstly, it is a protein of great interest for the pharmaceutical community: it is involved in key steps leading to the biosynthesis of sex hormones and it is associated to several forms of cancer [He04] and [Ci04]. Secondly, although its 3D structure is still unknown, CYP17 has distant sequential homologues - their sequence similarity is below 30% - the structures of which are known. Thirdly, 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 [Ah04]. Once the template is generated, its relevance is assessed: first, it is verified that its main features can be explained by biological data; secondly, it is compared with 3D models of the active site of CYP17 which were generated independently.

### 4.1 3D template generation

Homologue 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 high overlap. These P450 proteins are quite distant homologues: they all belong to the twilight zone with pair wise sequence similarities found in the range 20 to 30%. They also come from very different sources, since the first two are from humans and the other three are from bacteria.

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
<i>1IZO_A</i>	<i>417</i>	<i>24.0%</i>	<i>146</i>	<i>0.00011</i>
<i>1AKD</i>	<i>417</i>	<i>29.5%</i>	<i>61</i>	<i>1.5</i>

Table 3: CYP17 homologues using FASTA on PDB50%

In this study, the consensus 3D template was generated using atom distances of 1.0 and 1.25 Å and a group distance of 2.0 Å. Atom and group templates were respectively named T1.00, T1.25 and Tg2.00. As previously, a solvent atom size of 4.8 Å was chosen. 3D templates are analysed either as sequences aligned with their homologues where residues taking part in the template are shown (Figure 4) or as a consensus 3D structure (Figure 5).



Figure 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

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 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 (Figure 5).

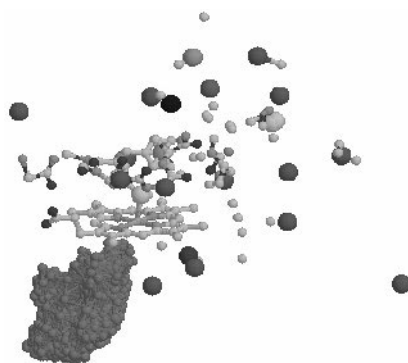


Figure 5: 3D template of human cytochrome P450 CYP17. Atoms are represented by small spheres, chemical groups are modelled by larger spheres and cavity is dark grey

## 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 [Fa02], 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, see Figure 4 (4). Moreover, this signature is located at the core of the longest sequential pattern of the template.

The Catalytic Site Atlas (CAS) [Po04] 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 [Hi00]. 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 [Po95]. That pattern, Figure 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 [Le01], are detected by our technique as shown in Figures 4 (2) and 4 (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 [Pe95].

### 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 3 groups (1 hydroxyl - threonine residue - and 2 non polar - alanine and glycine - groups) are very well conserved, i.e. RMSD < 1.2 Å (Figure 6). Moreover, alignment of the sequence of CYP17 with our set of proteins confirmed that these three residues are conserved in CYP17 too. 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, 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 [Ah04]. These models were then aligned to our 3D template using haem group positions as reference (Figure 6).

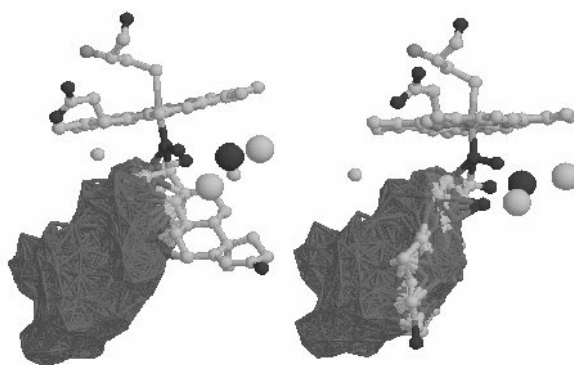


Figure 6: Alignment of CYP17 template with models of 1) 17alpha-hydroxylase and 2) lyase. Hydroxyl and non polar groups are respectively modelled by black and grey spheres

In both cases, the potential presence of a H-bond is confirmed since distances between the oxygen atoms of the threonine 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, Figure 6 (2). In the other case, Figure 6 (1), the mismatch between the substrate and the cavity 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 [Ah04].

#### **4.4 Discussion**

The analysis of the CYP17 3D template has shown that it 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 recognised 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 the consensus cavity was confirmed by the lyase complex model.

Results obtained with the CYP17 models are significant because they indicate our 3D template could be exploited by completing current active site models used for de novo design of novel inhibitors of enzymes such as cytochrome P450 proteins.

#### **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.

## References

- [Ah04] Ahmed,S.: The use of the novel substrate-heme complex approach in the derivation of a representation of the active site of the enzyme complex 17 $\alpha$ -hydroxylase and 17,20-lyase, *Biochem. Biophys. Res. Commun.*, 316(3), 595-598, 2004.
- [Be00] Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E.: The Protein Data Bank, *Nucleic Acids Research*, 28, 235-242, 2000.
- [Be04] Bertini,I., Faraone-Mennella,J., Gray,H.B., Luchinat,C, Parigi,G. and Winkler,J.R.: NMR-validated structural model for oxidized *Rhodospseudomonas palustris* cytochrome c(556), *J Biol Inorg Chem.*, 9(2), 224-30, 2004.
- [Br00] Brady,G.P. and Stouten,P.F.W.: Fast prediction and visualization of protein binding pockets with PASS, *J Computer-Aided Molecular Design*, 14, 383-401, 2000.
- [Ca03] Campbell SJ, Gold ND, Jackson RM, Westhead DR: Ligand binding: Functional site location, similarity and docking, *Curr Opin Struct Biol.*, 13(3):389-95, 2003.
- [Ci04] Cicek,M.S., Conti,D.V. ,Curran,A., Neville,P.J., Paris,P.L., Casey,G. and Witte,J.S.: Association of prostate cancer risk and aggressiveness to androgen pathway genes: SRD5A2, CYP17, and the AR, *Prostate*, 59(1), 69-76, 2004.
- [Ey04] Eyal,E., Edelman,M. and Sobolev,V.: The influence of crystal packing on protein structure, 3Dsig, Glasgow, UK, 2004.
- [Fa02] Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J., Hofmann,K. and Bairoch,A.: The PROSITE database, its status in 2002, *Nucleic Acids Res.*, 30, 235-238, 2002.
- [Gi96] Gibrat,J.F., Madej,T. and Bryant,S.H.: Surprising similarities in structure comparison, *Current Opinion in Structural Biology*, 6(3), 377-385, 1996.
- [Gi01] Gilbert,D., Westhead,D., Viksna,J. and Thornton,J.: A computer system to perform structure comparison using TOPS representations of protein structure, *Comput. Chem.*, 26, 23-30, 2001.
- [Gu04] Guda,C., Lu,S., Scheeff,E.D., Bourne.P.E. and Shindyalov,I.N.: CE-MC: a multiple protein structure alignment server, *Nucleic Acids Res.*, 32 (Web Server issue):W100–W103, 2004.
- [He04] Hefler,L.A., Tempfer,C.B., Grimm,C., Lebrecht,A., Ulbrich,E., Heinze,G., Leodolter,S., Schneeberger,C., Mueller,M.W., Muendlein, A. and Koelbl,H.: Estrogen-metabolizing gene polymorphisms in the assessment of breast carcinoma risk and fibroadenoma risk in Caucasian women, *Cancer*, 101(2), 264-269, 2004.
- [Hi00] Hishiki,T., Shimada,H., Nagano,S., Egawa,T., Kanamori,Y., Makino,R., Park,S.Y., Adachi,S.I., Shiro,Y. and Ishimura,Y.: 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, 2000.
- [Ho94] Holm,L. and Sander,C.: Searching protein structure databases has come of age, *Proteins*, 19, 165-173, 1994.
- [Ho87] Horn,B.K.P.: Closed-form solution of absolute orientation using unit quaternions, *J. Optical Soc. Am.*, 4, 629-642, 1987.
- [Ja02] Jambon M, Imberty A, Deléage G, Geourjon C: SuMo: a software that detects 3D sites shared by protein structures, *JOBIM*, 2002.

- [Jo04] Jones S, Thornton JM: Searching for functional sites in protein structures, *Curr Opin Chem Biol.*, 8(1):3-7, 2004.
- [Le01] Lewis,D.: *Guide to Cytochromes P450: Structure & Function* (London: Taylor & Francis), 2001.
- [Li98] Liang,J., Edelsbrunner,H. and Woodward,C.: Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design, *Protein Science*, 7, 1884-1897, 1998.
- [OI02] Oldfield TJ: Data mining the protein data bank: residue interactions, *Proteins*, 49(4):510-28, 2002.
- [OI05] Oldham,M.L., Brash,A.R., Newcomer,M.E.: The structure of coral allene oxide synthase reveals a catalase adapted for metabolism of a fatty acid hydroperoxide, *Proc Natl Acad Sci USA*, 102(2), 297-302, 2005.
- [Or97] Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M.: CATH- A hierarchic classification of protein domain structures, *Structure*, 5(8), 1093-1108, 1997.
- [Pe88] Pearson,W.R. and Lipman,D.J.: Improved tools for biological sequence comparison, *PNAS*, 85, 2444-2448, 1988.
- [Pe95] Peterson,J.A. and Graham-Lorence,S.E.: Bacterial P450s: structural similarities and functional differences, *Cytochrome P450* (Ortiz de Montellano Editor: Plenum), 151-180, 1995.
- [Po04] Porter,C.T., Bartlett,G.J. and Thornton,J.M.: The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data, *Nucleic Acids Res.*, 32, D129-D133, 2004.
- [Po95] Poulos,T.L., Cupp-Vickery,J. and Li,H.: Structural studies on prokaryotic cytochromes P450, *Cytochrome P450* (Ortiz de Montellano Editor: Plenum), 125-150, 1995.
- [Sc02] Schmitt,S., Kuhn,D. and Klebe,G.: A New Method to Detect Related Function among Proteins Independent of Sequence and Fold Homology, *J. Mol. Biol.*, 323:387-406, 2002.
- [Sh04] Shapiro,J. and Brutlag,D.: FoldMiner and LOCK 2: protein structure comparison and motif discovery on the web, *Nucleic Acids Res.*, 32(2), W536-W541, 2004.
- [Sh98] Shindyalov,I.N. and Bourne,P.E.: Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein Engineering*, 11(9), 739-747, 1998.
- [Th94] Thompson,J.D., Higgins,D.G. and Gibson,T.J.: 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, 1994.
- [Tr03] Tramontano,A. and Morea,V.: Assessment of homology-based predictions in CASP5, *Proteins*, 53 Suppl 6:352-68, 2003.