Structure Conservation in Cytochromes P450

Jordi Mestres
Chemogenomics Laboratory, Research Unit on Biomedical Informatics, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, Barcelona, Spain

ABSTRACT **The recent availability of crystal structures for several diverse cytochromes P450 (CYPs) offers the possibility to perform an up-to-date comparative analysis to identify the degree of structure conservation among this superfamily of enzymes specially relevant for their involvement in drug metabolism and toxicity. A set of 9 CYPs sharing between 10% and 27% sequence identity was selected, including 7 class I (CYP 101, 107, 108, 119, 121, 51, and 55) and two class II (CYP 102, and 2C5) structures. After obtaining a multiprotein structure superimposition, a structure-based sequence alignment was derived. Mapping the level of three-dimensional structural conservation onto the sequence alignment revealed that over 28% of the alignment positions have the C_α carbons of their residues within a root-mean-square deviation (RMSD) of 2 Å. This degree of structure conservation is found to be generally preserved, even when the structure undergoes dramatic conformational changes. Performing the analysis on 4 members of the CYP2 family (CYP 2B4, 2C5, 2C8, and 2C9), the percentage of alignment positions within 2 Å RMSD amounted to 73%, increasing to over 85% when only structures in a closed conformation are considered. The present findings suggest that it should be plausible to derive models of overall good quality for the major CYP2 metabolizing forms (CYP 2A6, 2C19, 2D6, and 2E1), whereas high levels of uncertainty are still likely to be expected in models for the remaining 2 major P450 metabolizing forms (CYP 1A2 and 3A4), with the corresponding implications for their potential applicability in drug design activities. Proteins 2005;58:596–609.** © 2004 Wiley-Liss, Inc.

Key words: comparative analysis; structure superimposition; positional conservation indices; homology modeling; drug metabolism

INTRODUCTION

Cytochromes P450 (CYPs) constitute a superfamily of heme–thiolate enzymes that catalyze the monooxygenation of a variety of hydrophobic substrates.[1–3] The catalytic center in CYPs is located around a porphyrin–heme complex confined within different amino acid sequences, which alter the topography of the active site and allow the binding of considerably diverse substrates.[4,5] The monooxygenation reactions require a redox partner to transfer electrons from reduced pyridine nucleotides to the P450 for the binding and activation of the iron-bound molecular oxygen.[6] Depending on the nature of the redox partner, CYPs can be divided into 4 different classes. Class I enzymes are found in most bacteria and in the mitochondrial membranes of eukaryotes, and require a flavin adenine dinucleotide (FAD)-containing reductase and an iron–sulfur protein. Class II enzymes comprise the membrane-bound eukaryotic microsomal enzymes and interact directly with a reductase-containing FAD and flavin mononucleotide (FMN). Class III enzymes do not require an exogenous source of electrons, as they already employ oxygen-containing substrates. Class IV enzymes receive electrons directly from reduced pyridine nucleotides, without the intervention of an electron carrier. The need to accommodate a large variety of substrates of various shapes and sizes, and to interact with different redox partners confers to the superfamily of P450 enzymes a significant degree of sequence and structural variability.[7–9]

In the last few years, an increasing number of P450 crystal structures has been determined and deposited in the Protein Data Bank (PDB).[10] This wealth of structural data has resulted in comparative analyses between P450 structures, which have revealed that despite having sequence identities in the range of 10–30%,[11] these proteins possess the same tertiary structure and have a well-conserved heme-binding structural core.[12,13] Cytochrome P450 enzymes are pivotal in drug metabolism and toxicity,[14] and thus represent a major concern to the pharmaceutical industry for successfully developing a drug.[15] The availability of P450 structures has opened an avenue to address these important issues through P450 modeling,[16–20] especially since the first crystal structure of a mammalian P450 became accessible.[21]

Despite the recent significant advances in P450 structure determination, and given the low sequence identities across families, deriving overall quality models of CYPs for which no member of the family has yet been crystallized remains a challenging task.[20] In this respect, comparative

Grant sponsor: Instituto de Salud Carlos III (Ministerio de Sanidad y Consumo); Grant number: 02/3051.

Correspondence to: Jordi Mestres, Chemogenomics Laboratory, Research Unit on Biomedical Informatics, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, Passeig Marítim de la Barceloneta 37-49, 08003, Barcelona, Spain. E-mail: jmestres@imim.es

Received 19 May 2004; Accepted 30 August 2004

Published online 22 December 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20354

analyses of P450 structures are very valuable, because they can help mapping the expected local accuracy of cytochrome P450 models. In particular, previous comparative analyses[8] identified a well-conserved structural core around the heme consisting of a 4-helix bundle of helices D, E, I, and L, and helices J and K. Because of the presence of several sequence markers, the sequence alignment in the regions comprising the structural core can be established unambiguously; thus, good local quality should be expected for any P450 model. In contrast, high variability is found in the surface helices A, B′, F, G, and H, thought to be involved in substrate recruitment and binding.[8] These regions are significantly different in amino acid sequence and length, and have an inherent structural flexibility. Consequently, modeling these regions with reasonable accuracy is still difficult.[20]

Key to a comparative analysis of protein structures is the method used to superimpose structures and the metric applied to quantify the degree of global and local similarity between structures. A large variety of methods exist for superimposing protein structures and evaluating their global structural similarity.[22–24] Some of them have been designed more toward identifying local structural similarities[25–29] and performing postsuperposition analyses that allow the visual detection of common structural motifs between related proteins.[30,31] These analyses are particularly relevant for the comparative modeling of proteins, because they can provide important information to project the local quality of the models.

With the current availability of representative structures for several P450 domains, the present work aims at performing a comparative analysis of a set of 9 diverse P450 structures to derive a structure-based sequence alignment from which both the global and local levels of sequence and structure conservation among them can be assessed. The results are then contrasted with those obtained when comparing some recently reported structures representative of 4 mammalian P450 members of the CYP2 family. The implications of these findings for the modeling of novel P450 domains are discussed. A description of the methods used for evaluating protein structure similarities and the degree of sequence and structure conservation at each position of a sequence alignment are given next.

MATERIALS AND METHODS

Structural Similarity and Superposition

In this work, protein structure similarity is evaluated using the Gaussian-based approach implemented in the program GAPS.[32] GAPS has been already validated against other protein superimposition methods[32] and shown to obtain sensible structural superimpositions between sets of proteins having low sequence identities.[30,32] Within a Gaussian-based approach, each atom i located at $\boldsymbol{R}_i$ is represented by a Gaussian function g_i as

$$g_i(\boldsymbol{r}) = \alpha_i \cdot \exp\left(-\beta_i|\boldsymbol{r} - \boldsymbol{R}_i|^2\right) \tag{1}$$

where the coefficient α_i and the exponent β_i determine the value of its maximum height at the origin and its decay, respectively. The coefficients α_i are generally defined by $\alpha_i = 0.4798{*}Z_i^{3.1027}$, with Z_i being the atomic number of atom i.[33] In contrast, the coefficients β_i depend on the user-defined extent of the Gaussian function.[32] Then, a Gaussian-based representation of the structure of a protein A, P_A, can be defined as

$$P_A(\boldsymbol{r}) = \sum_{i \in A} g_i(\boldsymbol{r}). \tag{2}$$

Since the regular features of the secondary structure of a protein are clearly defined by the trace of C_α carbons,[34] in order to speed up calculations, only C_α carbons will be considered to evaluate structural similarities and derive protein superimpositions. For a C_α atom, the α coefficient in Eq. (1) is 124.5748.

Once a Gaussian representation of the protein structure is defined, the structural similarity between two proteins A and B, S_{AB}, is assessed by evaluating the overlap integral, Z_{AB}, between their respective representations, P_A and P_B, as

$$Z_{AB}(\boldsymbol{t},\theta) = \int P_A(\boldsymbol{r})P_B(\boldsymbol{r})d\boldsymbol{r}, \tag{3}$$

which can then be normalized using a cosine-like index[35]:

$$S_{AB}(\boldsymbol{t},\theta) = Z_{AB}(\boldsymbol{t},\theta) \,/\, (Z_{AA} \cdot Z_{BB})^{1/2} \tag{4}$$

The values of S_{AB} range from 0 to 1. A value of 1 is achieved only in the limit case of identity, whereas any dissimilarity between the 2 proteins will be reflected in a value smaller than 1.

Exploration of the structural similarity between a pair of proteins is performed using a systematic spherical search.[32] Basically, one of the proteins is kept fixed (the reference protein) while the target protein is systematically placed in a number of unique starting orientations about the reference protein. Then, from each starting orientation, the structural similarity between the 2 proteins [S_{AB} in Eq. (4)] is optimized in all translational ($\boldsymbol{t}$) and rotational (θ) degrees of freedom using common gradient-seeking techniques. This procedure ensures a wide and uniformly distributed exploration of the similarity landscape defined by the structural characteristics of the 2 proteins. The sampling of the search depends on the rotational step of the sphere used to define the starting orientations. In the present case, a 45° search was performed for all pairwise superimpositions. Following previous studies,[32] a Gaussian representation vanishing at 5 Å from the atom centers was used to explore the similarity space between pairs of proteins. Under this constraint, the β coefficient in Eq. (1) for a C_α atom is 0.2851.

Finally, the structural similarity between a set of M proteins, S, can be defined as[30]

$$S(\boldsymbol{t},\theta) = 2\sum_{A>B} S_{AB}(\boldsymbol{t},\theta) \,/\, (M^2 - M), \tag{5}$$

where A and B are any 2 proteins belonging to the protein set. A multiple protein structural superimposition was

TABLE I. List of Cytochromes P450 Analyzed in This Work

CYP	Class	PDB ID	Resolution	Date	Source
101 (cam)	I	1phc	1.60	31/10/1993	*Pseudomonas putida*
102 (bm3)	II	2bmh	2.00	31/07/1994	*Bacillus megaterium*
107 (eryF)	I	1oxa	2.10	07/12/1995	*Saccharopolyspora erythraea*
108 (terp)	I	1cpt	2.30	31/01/1994	*Pseudomonas* sp.
119	I	1io7	1.50	28/02/2001	*Sulfolobus solfataricus*
121	I	1n40	1.06	04/02/2003	*Mycobacterium tuberculosis*
51	I	1e9x	2.10	01/11/2000	*Mycobacterium tuberculosis*
55 (nor)	I	1rom	2.00	15/10/1997	*Fusarium oxysporum*
2C5	II	1dt6	3.00	27/09/2000	*Oryctolagus cuniculus*
2B4	II	1po5	1.60	07/10/2003	*Oryctolagus cuniculus*
2C8	II	1pq2	2.70	13/01/2004	*Homo sapiens*
2C9	II	1og2	2.60	17/07/2003	*Homo sapiens*

Also included are class, PDB code of the representative structure used, resolution (in Å), release date, and enzyme source.

then obtained by optimizing S in all translational ($\boldsymbol{t}$) and rotational (θ) degrees of freedom using the optimized pairwise superimpositions as starting orientations. In this case, the fuzziness of the Gaussian representation was reduced and a Gaussian extent of 2 Å from the atom centers was used. Under this constraint, the β coefficient in Eq. (1) for a C_α atom is 1.7819. As for S_{AB}, values for S can range from 0 to 1.

Positional Conservation Indices

Once a multiprotein structure superimposition has been obtained, a structure-based sequence alignment can be derived. Structure and sequence positional conservation will be evaluated only at alignment positions containing residues from all proteins in the set. At each of these alignment positions, the level of structure conservation will be assessed by calculating the root-mean-square deviation (RMSD) between the C_α atoms of the residues at that position. In addition, the degree of sequence conservation will also be determined using an information–theoretical approach.[36,37] Within this approach, the entropy, E^I, at each position, I, of a sequence alignment between M proteins is defined as

$$E^I = -\sum_{i=1}^{20} \rho_i^I \cdot \ln\, \rho_i^I;\ \rho_i^I = p_i^I/M, \qquad (6)$$

where ρ_i^I and p_i^I are, respectively, the probability and the population of each amino acid, i, in position, I, of a sequence alignment. This entropy was originally proposed as a measure of the average information per symbol contained in a message,[38] but the concept can be applied as well to measure sequence variability by quantitatively describing the spreading of the amino acid distribution in each position of a sequence alignment. The values of E^I range between 0, reflecting full sequence identity, and a maximum number, E_{max}, reflecting complete sequence variability. For a set of $M \leq 20$ proteins, $E_{max} = \ln M$. Therefore, in order to provide a more intuitive upper-bound measure of sequence variability, positional entropy values are treated as if they were positional E_{max} values and then expressed in terms of their corresponding amino acid variability, $V^I = e^{E^I}$. For a set of $M \leq 20$ proteins, the values of V^I will range between 1, full positional identity, and M, complete positional variability.

RESULTS AND DISCUSSION

The set of CYPs analyzed in this work is listed in Table I, together with details of the actual representative structures used. The list includes 7 class I CYPs, namely, 101 (cam),[39] 107 (eryF),[40] 108 (terp),[41] 119,[42] 121,[43] 51,[44] 55 (nor),[45] and 5 class II CYPs, namely, 102 (bm3),[46] 2C5,[47] 2B4,[48] 2C8,[49] and 2C9.[50] The first 9 structures are representative of the sequence and structural diversity found in P450s and will be the primary focus of this study. A comparative analysis of these 9 structures should provide a good indication of the extent and conservation of the common structural core among the entire superfamily. The last 4 structures belong to the CYP2 family and offer the possibility to analyze the expected degree of conservation within a class II P450 family and discuss the consequences for the homology modeling of other closely related members of the family.

Comparative Analysis of CYP 101, 102, 107, 108, 119, 121, 51, 55, and 2C5

For historical reasons, the crystal structure of the bacterial CYP101 enzyme (1phc) was taken as the reference structure onto which the other structures were superimposed considering C_α carbons only. Once optimal pairwise superimpositions were obtained, all structures were then allowed to modify their relative orientation to maximize the overlap between all 9 proteins. That generated the final multiprotein structural superimposition from which a structure-based sequence alignment was then derived.

The pairwise structural similarities and sequence identities obtained from the multiprotein structural superimposition and structure-based sequence alignment, respectively, are given in Table II. As can be observed, sequence identities between protein pairs are all found within the so-called twilight zone, between 10% and 27%, although it

TABLE II. Percentage of Sequence Identity and Structural Similarity (Expressed as $100 \cdot S_{AB}$) Between 9 Diverse P450 Domains and Their Representative Structures

	1phc	2bmh	1oxa	1cpt	lio7	1n40	1e9x	1rom	1dt6
1phc	—	57.3	76.0	71.9	68.0	69.6	55.6	67.0	54.9
2bmh	13.7	—	60.5	62.8	59.4	59.4	58.4	61.5	61.8
1oxa	18.9	11.4	—	71.9	74.6	76.0	57.6	71.9	56.1
1cpt	22.0	16.6	20.6	—	72.6	69.2	54.8	73.3	57.8
1io7	18.6	15.6	25.8	20.8	—	71.8	53.6	72.1	57.1
1n40	18.4	13.1	22.1	18.6	21.1	—	55.6	73.2	59.6
1e9x	10.1	17.2	16.1	14.4	12.0	11.0	—	55.9	53.4
1rom	20.1	11.1	27.4	24.0	21.4	23.5	13.7	—	59.4
1dt6	14.5	17.2	14.0	13.6	13.4	11.9	14.1	13.0	—

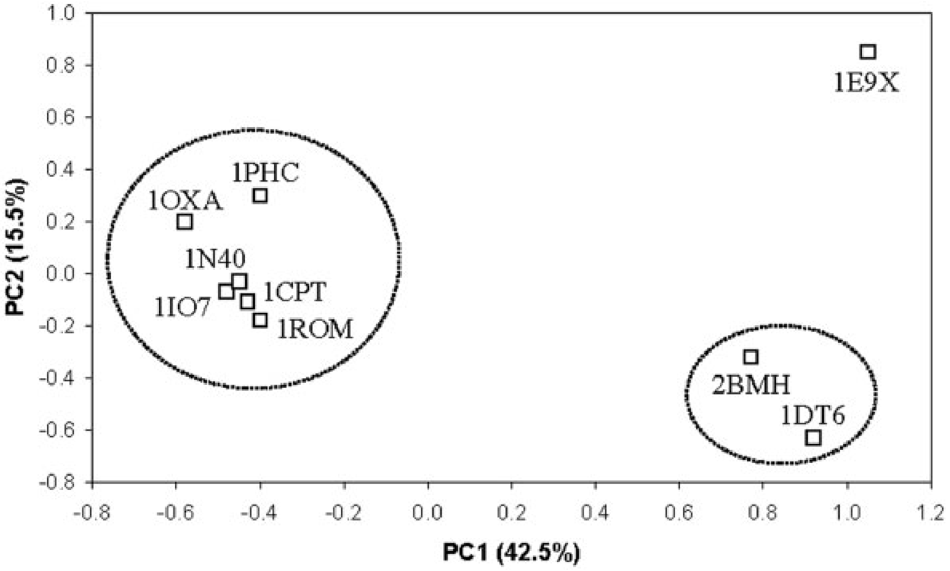


Fig. 1. Principal component analysis of the pairwise similarity matrix obtained from the final multiprotein structure superimposition. In parenthesis, the contribution of each principal component to the total variance.

is worth pointing out that sequence identities derived from structure-based sequence alignments are usually slightly lower than those obtained from sequence alignments based solely on sequence conservation.[12] In spite of these low sequence identity values, significant structural overlap is obtained, with pairwise similarity values ranging between 53% and 76%. Interestingly, some of the structural similarity values collected in Table II agree well with previously reported similarities between pairs of P450 structures, thus confirming the validity of the Gaussian-based approach to protein structure similarity used in this work. For example, the structure of CYP121 (1n40) was found to be most similar to CYP107 (1oxa), in good agreement with a comparative analysis done by other means,[43] and the structure of CYP2C5 (1dt6) was found to be most similar to CYP102 (2bmh), reflecting the findings from previous analyses.[47]

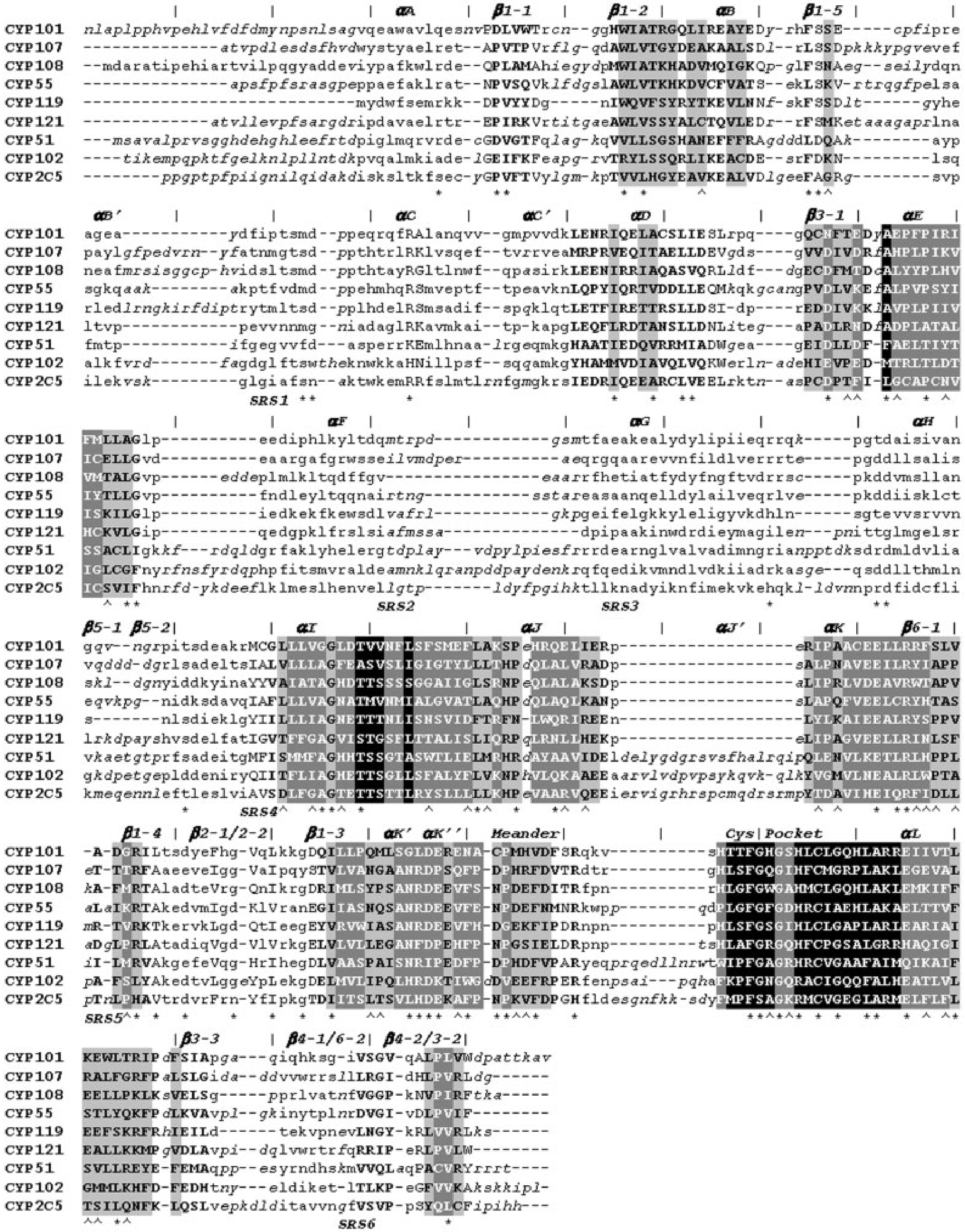
To provide a more visual impression of the structural resemblance between all 9 P450 structures as measured from the Gaussian-based approach used in this work, a principal component analysis of the pairwise structural similarity values given in Table II is presented in Figure 1. As can be observed, 3 clusters are clearly formed in the space described by the first 2 principal components representing 58% of the total variance. The most populated cluster contains all class I CYPs, with the exception of CYP51 (1e9x), which appears completely discriminated from the other structures. For the remainder of the article, this well-clustered set of 6 class I CYPs will be referred to as the class I* set, indicating that the structure of CYP51 (1e9x) is not included in the set. In fact, the CYP51 structure (1e9x) has been already reported to have striking structural differences with respect to other P450 structures determined, with a bent I helix in its N-terminal site and an open conformation of the BC loop.[44] Despite these obvious structural alterations, it was decided to keep CYP51 (1e9x) in the study in order to identify the minimum set of positions that remain structurally conserved even when the P450 structure undergoes dramatic conformational changes. The third cluster is formed by the 2 class II CYPs, 102 (2bmh) and 2C5 (1dt6). This result reflects the fact that, despite being a bacterial enzyme, CYP102 was originally proposed as the most adequate prototype for microsomal CYPs[51] and, in the absence of mammalian P450 structures, was generally used as the preferred template structure for homology modeling of mammalian sequences.[52]

The structure-based sequence alignment derived from the optimized multiprotein structural superimposition is shown in Figure 2. The ordering of the CYPs in the alignment reflects the clustering obtained in Figure 1, with the 6 P450 enzymes of the class I* set first, then the singleton formed by CYP51, and finally the 2 class II CYPs. Since the aim is to incorporate structural information in the sequence alignment to establish a direct connection between amino acid sequences and their relative position in three-dimensional (3D) space, all sequences were directly extracted from the structure files in the PDB. Therefore, some of the sequences may not be complete, as some highly variable structural features could not be resolved satisfactorily and are consequently missing in some of the PDB structure files. For example, this is the case for the FG loop in CYP108 (1cpt) and CYP2C5 (1dt6). In order to include 3D structural information in the sequence alignment, each alignment position in Figure 2 has been colored according to some ranges of RMSD values for the C_α carbons corresponding to the amino acids assigned to that position. This way of presenting sequence alignments allows for locating those regions where the assignment of amino acids to a given alignment position is tightly linked to having those amino acids in close proximity in 3D space, representing a simple illustrative means of directly identifying the most structurally conserved regions among all proteins. It maybe worth stressing at this point that the coloring of the alignment positions in Figure 2 reflects the local structural conservation of the amino acids assigned to each alignment position as they are found in 3D space in the optimized global multiprotein structural superimposition. Identification of common substructure blocks would result in a different picture in several regions, which would complement the structural information contained in Figure 2 from a more local perspective.[13]

As can be extracted from the structure-based sequence alignment presented in Figure 2, the most structurally conserved regions comprise helix E, the C-terminal half of helix I, helices J and K, the β6-1 and β1-3 sheets, helices K′ and K″, the Cys-pocket, helix L, and the β3-2 sheet. Other regions also found to be relatively well conserved structurally are located in the β1-2 sheet, helices B and D, the β3-1 and β1-4 sheets, and the Meander. Finally, the main secondary structure elements found in general to be structurally most poorly conserved are helices A, B′, C, C′, F, G, and H, as well as the N-terminal part of helix I. Overall, 21, 86, 143, 181, and 200 alignment positions, out of a total of 498, were found to have the C_α carbon atoms corresponding to the aligned residues within 1.0, 1.5, 2.0, 2.5, and 3.0 Å RMSD, respectively. Taking arbitrarily an RMSD value of 2 Å as an acceptable level of spatial conservation (all alignment positions with shaded background in Fig. 2), this means that around 30% of a P450 structure is found to be generally well preserved structurally within this superfamily. On this basis, this set of 143 residues will be referred to as the "structural core" in the remainder of the article.

These findings can be better analyzed visually in 3D space in Figure 3, where each C_α carbon in the CYP101 structure (1phc) has been colored according to the RMSD value of the sequence alignment position to which it has been assigned. Nonaligned positions were arbitrarily assigned an RMSD value of 10 Å. The picture reveals clearly that the most structurally conserved regions are mainly located around the heme group and the proximal site of the P450 enzyme forming two interaction clusters, one comprising the β1-2 sheet, helix B, the β6-1 and β1-3 sheets, helices K′ and K″, and the Cys-pocket, and the other containing helices D and E, the C-terminal half of helix I, and helices J, K and L. The spatially well conserved proximal surface (conserved even when the protein undergoes dramatic conformational changes) may constitute evidence of the existence of an "evolutionary trace" of the interaction between the P450 and its redox partner.[53–55] In contrast, the most variable regions are found in the distal site, consistent with those elements being involved in substrate recruitment and binding and thus inherently requiring some degree of structural flexibility.[8] In this respect, it is interesting to observe how this static picture provides a fairly good indication of the dynamic fluctuations in the backbone of a P450 structure and, in particular, the extent at which the spatial movements of the B′, F, and G helices are further transmitted to the adjacent secondary structure elements, namely, helices C, C′, H, and the N-terminal site of helix I.

At this stage, one may be tempted to suggest that the inclusion of the open conformation in CYP51 (1e9x) is perhaps strongly limiting the degree of conservation (variation with number and class of structures considered) and extent (number of residues involved) of the structural core identified above. In order to investigate this aspect, the structural cores defined by the subsets of P450 enzymes belonging to class I* (all class I except CYP51), class II, and class I*+II (all class I and II except CYP51) were compared to the one defined by the entire set (class I+II). The results are depicted in Figure 4, where each circle represents a position in the alignment with an RMSD value below 2 Å. A total of 204 and 166 most structurally conserved positions were identified for classes I* and II, respectively. The majority of these highly conserved positions are shared between these two sets but some clear differences are also noticeable. These differences are particularly localized in the high degree of 3D spatial conservation observed in helices A and B and throughout the regions between helices K and K′ in the class I* set, and in helices C and H, and the C-terminal site of helix G in the class II set. When these 2 sets are combined (class I*+II set), the number of most structurally conserved positions is reduced to 159. The distribution of these positions reflects mainly the existing commonalities between the two sets but also, to a much lesser extent, the relative number of structures contributing from the two sets to the combined set. Thus, as a result of the higher weight of the class I* structures in the combined set, the latter effect is



```
                  |        |        |  αA   |  β1-1  |   β1-2 |    αB  |  β1-5 |       |
CYP101   nlaplpphvpehlvfdfdmynpsnlsagvqeawavlqesnvPDLVWTrcn--ggHWIATRGQLIREAYEDy-rhFSSE-----cpfipre
CYP107   --------------atvpdlesdsfhvdwystyaelret--APVTPVrflg-qdAWLVTGYDEAKAALSDl--rLSSDpkkkypgvevef
CYP108   --mdaratipehiartvilpqgyaddeviypafkwlrde--QPLAMAhiegydpMWIATKHADVMQIGKQp-glFSNAeg--seilydqn
CYP55    ---------------apsfpfsrasgpeppaefaklrat--NPVSQVklfdgslAWLVTKHKDVCFVATS--ekLSKV-rtrqgfpelsa
CYP119   ----------------------------mydwfsemrkk--DPVYYDg-----nIWQVFSYRYTKEVLNNf-skFSSDlt------gyhe
CYP121   -------------atvllevpfsargdripdavaelrtr--EPIRKVrtitgaeAWLVSSYALCTQVLEDr--rFSMKetaaagaprlna
CYP51    ---msavalprvsgghdehghleefrtdpiglmqrvrde-cGDVGTFqlag-kqVVLLSGSHANEFFFRAgdddLDQAk--------ayp
CYP102   ----tikempqpktfgelknlpllntdkpvqalmkiade-lGEIFKFeapg-rvTRYLSSQRLIKEACDE--srFDKN---------lsq
CYP2C5   --------ppgptpfpiignilqidakdiskslt kfsec-yGPVFTVylgm-kpTVVLHGYEAVKEALVDlgeeFAGRg--------svp
                                               *     **            * *       ^         **^

         αB'        |        |        |  αC   |     αC' |   αD |        |   β3-1 |    αE  |
CYP101   agea-----------ydfiptsmd--ppeqrqfRAlanqvv--gmpvvdkLENRIQELACSLIESLrpq---gQCNFTEDyAEPFPIRI
CYP107   paylgfpedvrn--yfatnmgtsd--ppthtrlRKlvsqef--tvrrveaMRPRVEQITAELLDEVgds----gVVDIVDRfAHPLPIKV
CYP108   neafmrsisggcp-hvidsltsmd--ppthtayRGltlnwf--qpasirkLEENIRRIAQASVQRLldf---dgECDFMTDcALYYPLHV
CYP55    sgkqaak--------akptfvdmd--ppehmhqRSmveptf--tpeavknLQPYIQRTVDDLLEQMkqkgcangPVDLVKEfALPVPSYI
CYP119   rledlrngkirfdiptrytmltsd--pplhdelRSmsadif--spqklqtLETFIRETTRSLLDSI-dp----rEDDIVKKlAVPLPIIV
CYP121   ltvp------------pevvnnmg--niadaglRKavmkai--tp-kapgLEQFLRDTANSLLDNLiteg---aPADLRNDfADPLATAL
CYP51    fmtp-----------ifgegvvfd---asperrKEmlhnaa-lrgeqmkgHAATIEDQVRRMIADWgea----gEIDLLDF-FAELTIYT
CYP102   alkfvrd-------fagdglftswtheknwkkaHNillpsf--sqqamkgYHAMMVDIAVQLVQKWerln-adeHIEVPED-MTRLTLDT
CYP2C5   ilekvsk---------glgiafsn--aktwkemRRfslmtlrnfgmgkrsIEDRIQEEARCLVEELrktn--asPCDPTFI-LGCAPCNV
                       SRS1 **        *                     *   *   **             * ^^  *^  * ^

                  |        |   αF  |        |        |   αG  |        |        |   αH  |
CYP101   FMLLAGlp----------eediphlkyltdqmtrpd------------gsmtfaeakealydylipiieqrrqk-----pgtdaisivan
CYP107   ICELLGvd----------eaargafgrwsseilvmdper----------aeqrgqaarevvnfildlverrrte-----pgddllsalis
CYP108   VMTALGvp------eddeplmlkltqdffgv----------------eaarrfhetiatfydyfngftvdrrsc-----pkddvmsllan
CYP55    IYTLLGvp----------fndleyltqqnairtng-----------sstareasaanqelldylailveqrlve-----pkddiisklct
CYP119   ISKILGlp----------iedkekfkewsdlvafrl------------gkpgeifelgkkyleligyvkdhln------sgtevvsrvvn
CYP121   HCKVLGip----------qedgpklfrslsiafmssa--------------dpipaakinwdrdieymagilen---pnittglmgelsr
CYP51    SSACLIgkkf---rdqldgrfaklyhelergtdplay---vdpylpiesfrrrdearnglvalvadimngrianpptdksdrdmldvlia
CYP102   IGLCGFnyrfnsfyrdqphpfitsmvraldeamnklqranpddpaydenkrqfqedikvmndlvdkiiadrkasge---qsddllthmln
CYP2C5   ICSVIFhnrfd-ykdeeflklmeslhenvellgtp------ldyfpgihktllknadyiknfimekvkehqkl-ldvnnprdfidcfli
          ^ **                       SRS2                  SRS3               *             **

         β5-1 β5-2|        |  αI    |        |   αJ   |        |   αJ' |    αK  |   β6-1 |
CYP101   gqv--ngrpitsdeakrMCGLLLVGGLDTVVNFLSFSMEFLAKSPeHRQELIERp------------------eRIPAACEELLRRFSLV
CYP107   vqddd-dgrlsadeltsIALVLLLAGFEASVSLIGIGTYLLLTHPdQLALVRADp------------------sALPNAVEEILRYIAPP
CYP108   skl--dgnyiddkyinaYYVAIATAGHDTTSSSSGGAIIGLSRNPeQLALAKSDp------------------aLIPRLVDEAVRWTAPV
CYP55    eqvkpg--nidksdavqIAFLLLVAGNATMVNMIALGVATLAQHPdQLAQIKANp------------------sLAPQFVEELCRYHTAS
CYP119   s-------nlsdieklgYIILLLIAGNETTTNLISNSVIDFTRFN-LWQRIREEn-------------------LYLKAIEEALRYSPPV
CYP121   lrkdpayshvsdelfatIGVTFFGAGVISTGSFLTTALISLIQRPqLRNLIHEKp------------------eLIPAGVEELLRINLSF
CYP51    vkaetgtprfsadeitgMFISMMFAGHHTSSGTASWTLIELMRHRdAYAAVIDEldelygdgrsvsfhalrqipQLENVLKETLRLHPPL
CYP102   gkdpetgeplddeniryQIITFLIAGHETTSGLLSFALYFLVKNPhVLQKAAEEaarvlvdpvpsykqvk-qlkYVGMVLNEALRLWPTA
CYP2C5   kmeqennleftleslviAVSDLFGAGTETTSTTLRYSLLLLLKHPeVAARVQEEiervigrhrspcmqdrsrmpYTDAVIHEIQRFIDLL
                     *  SRS4^  ^**^ *        ^    ^*^  *   *^  ^                  ^ ^     *  **^^^ ^

            β1-4 | β2-1/2-2|   β1-3  | αK' αK''|  Meander|        |      Cys|Pocket |    αL  |
CYP101   -A-DGRILtsdyeFhg-VqLkkgDQILLPQMLSGLDERENA-CPMHVDFSRqkv----------sHTTFGHGSHLCLGQHLARREIIVTL
CYP107   eT-TTRFAaeeveIgg-VaIpqySTVLVANGAANRDPSQFP-DPHRFDVTRdtr----------gHLSFGQGIHFCMGRPLAKLEGEVAL
CYP108   kA-FMRTAladteVrg-QnIkrgDRIMLSYPSANRDEEVFS-NPDEFDITRfpn----------rHLGFGWGAHMCLGQHLAKLEMKIFF
CYP55    aLaIKRTAkedvmIgd-KlVranEGIIASNQSANRDEEVFE-NPDEFNMNRkwpp--------qdPLGFGFGDHRCIAEHLAKAELTTVF
CYP119   mR-TVRKTkervkLgd-QtIeegEYVRVWIASANRDEEVFH-DGEKFIPDRnpn----------pHLSFGSGIHLCLGAPLARLEARIAI
CYP121   aDgLPRLAtadiqVgd-VlVrkgELVLVLLEGANFDPEHFP-NPGSIELDRpnp---------tsHLAFGRGQHFCPGSALGRRHAQIGI
CYP51    iI-LMRVAkgefeVqg-HrIhegDLVAASPAISNRIPEDFP-DPHDFVPARyeqprqedllnrwtWIPFGAGRHRCVGAAFAIMQIKAIF
CYP102   pA-FSLYAkedtvLggeYpLekgDELMVLIPQLHRDKTIWGdDVEEFRPERfenpsai---pqhaFKPFGNGQRACIGQQFALHEATLVL
CYP2C5   pTnLPHAVtrdvrFrn-YfIpkgTDIITSLTSVLHDEKAFP-NPKVFDPGHfldesgnfkk-sdyFMPFSAGKRMCVGEGLARMELFLFL
         SRS5^* *  *     *   *  * *   ^^  **** ^*  **^^*   *                **^*^* * *  **  * ^ ^*

                  |β3-3     | β4-1/6-2| β4-2/3-2|
CYP101   KEWLTRIPdFSIApga---qiqhksg-iVSGV-qALPLVWdpattkav
CYP107   RALFGRFPaLSLGida--ddvvwrrsllLRGI-dHLPVRLdg------
CYP108   EELLPKLKsVELSg-----pprlvatnfVGGP-kNVPIRFtka-----
CYP55    STLYQKFPdLKVAvpl--gkinytplnrDVGI-vDLPVIF--------
CYP119   EEFSKRFRhIEILd------tekvpnevLNGY-kRLVVRLks------
CYP121   EALLKKMPgVDLAvpi--dqlvwrtrfqRRIP-eRLPVLW--------
CYP51    SVLLREYE-FEMAqpp--esyrndhskmVVQLaqPACVRYrrrt----
CYP102   GMMLKHFD-FEDHtny---eldiket-lTLKP-eGFVVKAkskkipl-
CYP2C5   TSILQNFK-LQSLvepkdlditavvngfVSVP-pSYQLCFipihh---
         ^^ *^                     SRS6      *
```

Fig. 2. Structure-based sequence alignment of 9 diverse cytochromes P450. Main secondary elements are marked above the sequences, whereas SRSs are indicated below the sequences. Nonshaded and shaded alignment positions reflect RMSD values above and below 2Å, respectively: *a*: nonaligned; a: >3.0; A: [2.5,3.0]; **A**: [2.0,2.5]; light-gray shaded **A**: [1.5,2.0]; dark-gray shaded white **A**: [1.0,1.5]; black shaded white **A**: ≤1.0. Positions with $V^l < 3$ are marked with * and those with $V^l > 5$ and RMSD < 2 Å are marked with ∧.

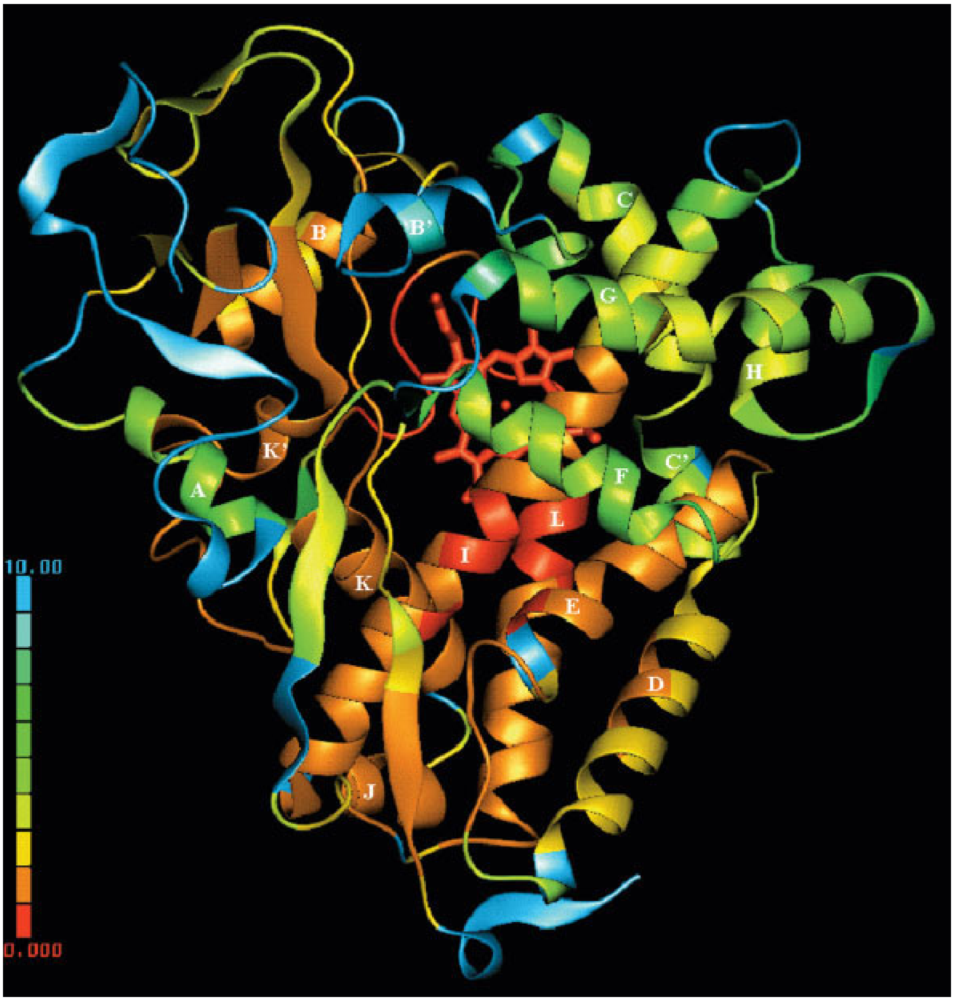


Fig. 3. Structure of CYP101 (1phc) with each C_α carbon colored according to the RMSD value of the sequence alignment position to which it has been assigned to (see Fig. 2). The heme group is depicted in red solid bonds, with a sphere locating the central Fe atom. Values in the color spectrum reflect the degree of structural conservation, with blue and red being the two limit cases of maximum and minimum spatial variability among the structures of 9 diverse cytochromes P450, respectively.

observed in the few positions retained in helix B and the slight reduction in structural conservation in helices C, D, E, G, and H. Finally, when CYP51 is included in the analysis (class I+II set), a total of 143 positions are found within 2 Å RMSD (see also shaded positions in Fig. 2). It is remarkable to observe that the distribution obtained for the class I*+II set is essentially retained in the complete set. Only 16 positions have been lost by the addition of a structure showing striking conformational changes with respect to the other structures. These positions are concentrated in helices C, H, and the N-terminal site of helix I, consistent with those elements involved in the dynamics of P450 structures identified above (see Fig. 3).

Having examined the extent and degree of structure conservation among the 9 diverse P450 structures, the analysis turns now to investigating the existence and extent of local levels of sequence conservation and the correlation with the observed degree of structural conservation. To this aim, entropy values, expressed in terms of sequence variability indices (V^I), were plotted against RMSD values obtained at each position along the sequence alignment. The results are presented in Figure 5. As can be seen, the clear increase in structural conservation achieved within the different secondary structure elements is not accompanied by a corresponding complementary decrease in sequence variability. This is not surprising, given the overall low sequence identities between the different P450 enzymes (see Table II). However, a tendency toward lower levels of sequence variability is observed in the most structurally conserved regions. In order to locate precisely where these positions are, positions with a sequence variability index lower than 3 have been marked (*) in Figure 2. These positions can be used as sequence markers to assist in obtaining the sequence alignment of other CYPs from different families, a step of key importance in the process of deriving quality structural models of cytochromes P450 by homology to existing structures.

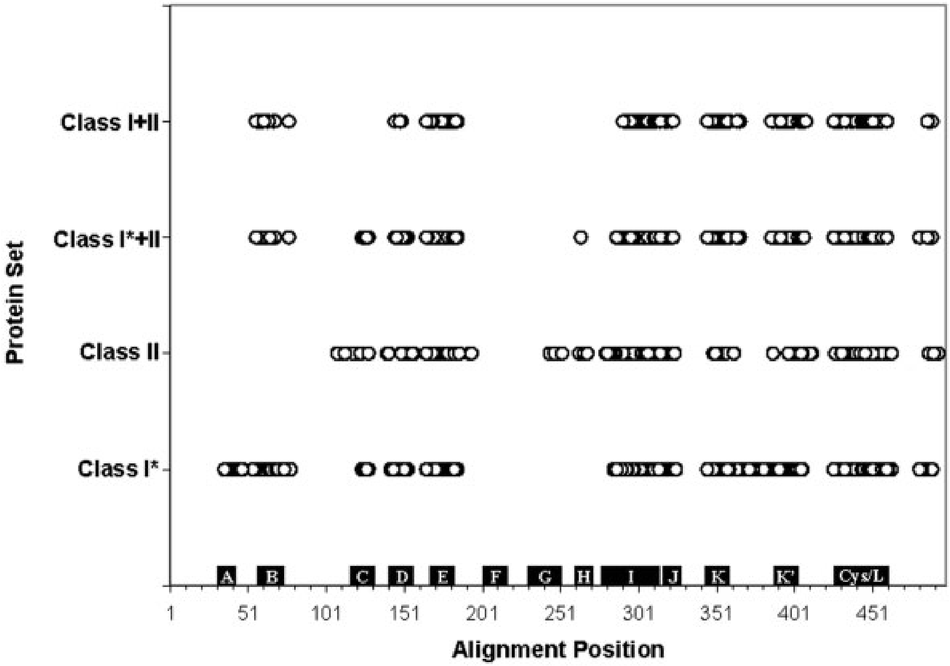


Fig. 4. Structure conservation profile within 2 Å RMSD along the sequence alignment for different subsets of 9 diverse cytochromes P450 (see text for subset details). The approximate location of the main secondary structure elements is also given.

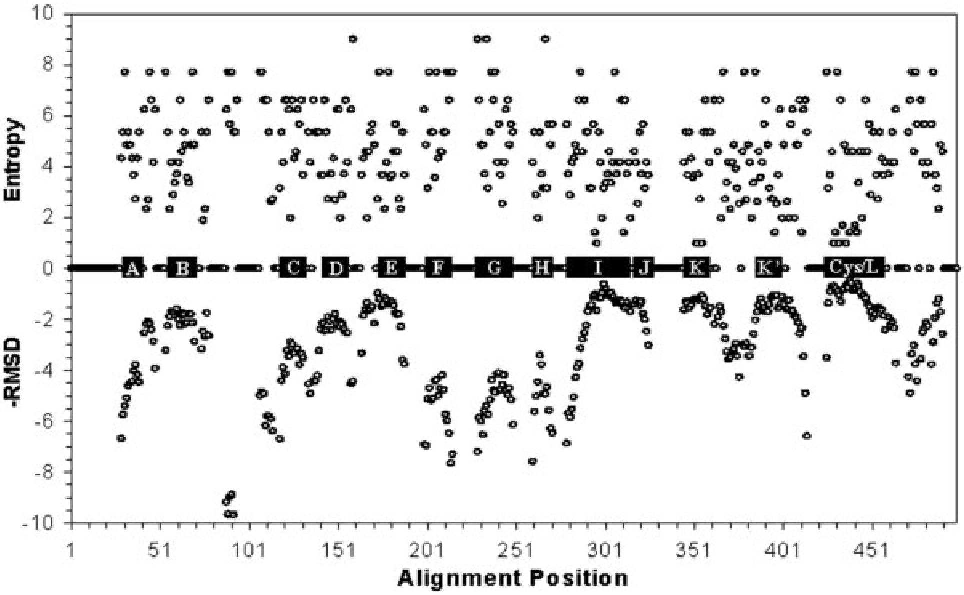


Fig. 5. Sequence variability (entropy, V^i) and structure conservation (RMSD) along the sequence alignment of 9 diverse cytochromes P450. The approximate location of the main secondary structure elements is also given.

Unfortunately, as noted in Figure 5, high sequence conservation is not always correlated with high structure conservation. This is better reflected in Figure 6, where sequence variability has been plotted against structural conservation at the 498 positions of the structure-based sequence alignment. As can be observed, no correlation exists between sequence and structure conservation, points covering almost uniformly the entire sequence–structure variability space. A dotted rectangle encloses the 60 positions having sequence variability index below 3. Among those, 39 positions have their residues placed in 3D space within 2 Å RMSD. The remaining structurally less localized 21 positions are located mainly at the edges of well conserved secondary structure elements, thus justifying their susceptibility to larger spatial variations (see Fig. 2). In contrast, a dashed rectangle encloses the set of 34

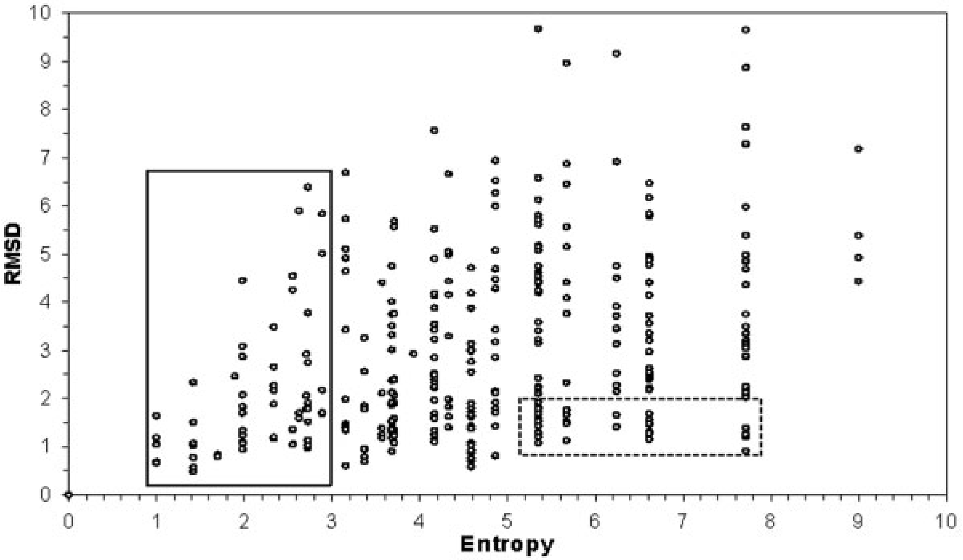


Fig. 6. Correlation between sequence variability (entropy, V') and structure conservation (RMSD) at the different positions of the sequence alignment between 9 diverse cytochromes P450.

positions having high sequence variability (variability index higher than 5) but high structural conservation as well (RMSD values lower than 2 Å). For the sake of clarity, these positions have also been marked (∧) in Figure 2, where they can be found within the most structurally conserved regions often next to some of the positions with lowest sequence variability. In this respect, it has been recommended the use of average conservation indices over a window of positions in order to smooth out the often abrupt variations in positional indices and be able to detect consensus sequence motifs, which are generally more informative than the identification of conserved isolated positions.[37]

To this aim, average conservation indices were calculated over different window sizes (w = 3, 6, 9, and 12) at those alignment positions fully occupied within the range defined by the window. In each case, the alignment position with the lowest average sequence variability index within each of the main secondary structure elements was stored, together with its corresponding average rmsd value. The results are plotted in Figure 7, where the lowest average entropy value found for a window of 3 alignment positions can be compared with the value obtained for the most conserved isolated position at each secondary structure element. In general, averaging of positional conservation indices results consistently in a significant increase of the lowest sequence variability found within each of the main structural elements. For example, within the Meander region the most conserved isolated position was position 411, with a sequence variability index of 1.4174 reflecting a consensus R residue. When averaging over a window of 3 positions, the lowest average sequence variability index (4.0311) is obtained at position 406, which contains within a 3-position window the consensus sequence motif PXXF (see Fig. 2). In addition, the radius of each circle in Figure 7 reflects the value of the (average) RMSD at the corresponding alignment position, providing an illustrative means for comparing the structural conservation among the different secondary structure elements. On this basis, one can clearly extract that the best consensus sequence and structure conservation over 7-residue segments is located in the Cys-pocket region and helices I, K, K′, and L. In contrast, the highest spatial variability can be found in the regions defined by helices A, B′, F, G, and H, in good agreement with the findings presented above. However, it should be stressed that some of these regions may be locally superimposable but are found displaced to some extent after the global multiple protein superposition due to concerted movements of loops and secondary structure elements. As highlighted above, performing the same analysis on individual substructure elements would allow for identifying local patterns of sequence and structure conservation irrespective of their relative global disposition.[13]

Ordering of all alignment positions according to the average sequence variability index over a window of w positions provides a ranking of all $(2w + 1)$-residue segments. A sequence variability index below 2.0000 in the original isolated position is then taken as a threshold for the assignment of a consensus residue to an alignment position within a segment, an X residue being assigned otherwise. A segment is considered a consensus sequence motif if more than one consensus residue is found within the segment. On the basis of these criteria, Table III collects the top 4 consensus sequence motifs identified when averaging the positional conservation indices over a window of 3 positions, and the top motifs when expanded windows of 6, 9, and 12 positions are considered. Also given in Table III are the location of the motif and the average entropy and rmsd values at the central position. Remarkably, consensus sequence motifs representative of each of the most conserved secondary structure elements are recovered when positional conservation indices are averaged over a window of 3 positions, in good agreement

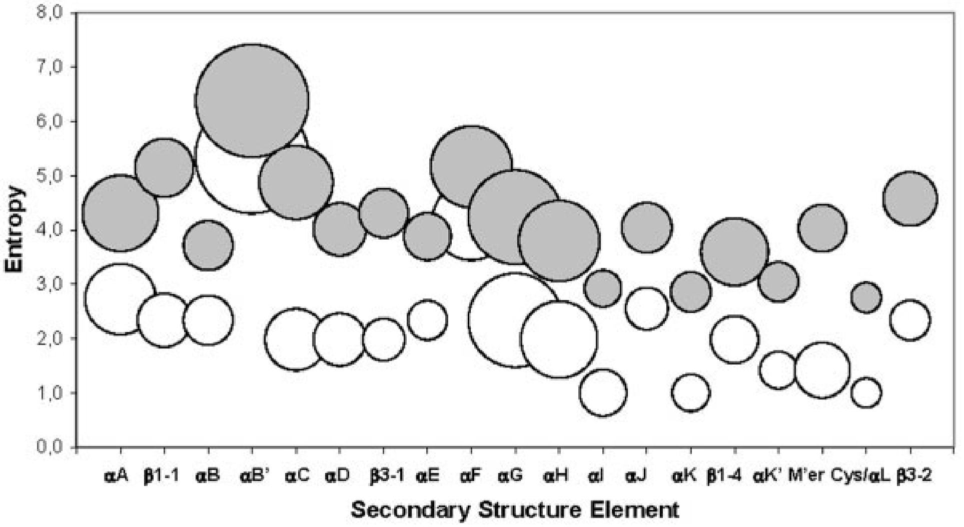


Fig. 7. Effect of averaging positional conservation indices over a window of 3 positions (circles in gray) compared to the most conserved isolated positions (circles in white) within each of the main secondary structure elements. The radius of the circle reflects the value of the average RMSD in the alignment position with the lowest average entropy at each secondary structure element. M'er stands for Meander.

TABLE III. Consensus Sequence Motifs Identified by Averaging Positional Conservation Indices at Each Alignment Position Over a Window of *w* Positions (*w* = 3,6,9,12).

i ± *w*	Motif	Location	|Entropy|	|RMSD|
439 ± 3	**C**XGXXLA	Cys-Pocket/αL	2.76	0.68
352 ± 3	XXX**E**XX**R**	αK	2.86	1.21
298 ± 3	A**GG**XXTXX	αI	2.92	1.04
397 ± 3	XXDXXXF	αK′	3.05	1.19
440 ± 6	HX**C**XGXXLAXXEX	Cys-Pocket/αL	3.11	0.78
436 ± 9	XX**F**GX**G**XHX**C**XGXXLAXXE	Cys-Pocket/αL	3.25	0.81
303 ± 12	XXXXA**G**XXTXXXXXXXXXXXXLXXXP	αI	3.91	1.25

Also given are the location of the motif and the average entropy (expressed as V^1) and RMSD values at the central position (*i*). The absolutely conserved residues are highlighted in bold.

with previous recommendations.[37] These include the well-recognized EXXR motif at the C-terminal end of helix K, the CXGXXLA motif in the Cys-pocket and the AGXXT motif in helix I, comprising some of the absolutely conserved residues among all CYPs, and the DXXXF motif in helix K′. Apart from their low average sequence variability, all these motifs share a high structure conservation, with average RMSD values between 0.68 and 1.21 Å. Enlarging the window size allows for identifying longer consensus sequence motifs within the region formed by the Cys-pocket and helix L and within helix I, which are extensions of the motifs identified previously with shorter window sizes. In particular, a 17-residue motif, FGXGXHX-CXGXXLAXXE, is identified in the Cys-pocket/αL region when the averaging is performed over a window of 9 positions, and a 21-residue motif, AGXXTXXXXXXXXXXXX-LXXXP, is located in helix I when the averaging is done over a window of 12 positions. Despite their length, both motifs are found to be structurally well conserved, with average RMSD values of 0.81 and 1.25 Å, respectively.

In view of the present findings, it is also worth stressing their potential implications for modeling by homology the active site of novel diverse CYPs with sequence identities within the twilight zone and for which no experimental structure is available yet in order to be able to assess the scope of applicability of the models for drug design purposes. To this aim, the approximate location of the well-recognized 6 substrate recognition sites (SRSs)[56] has been also added in the structure-based sequence alignment of Figure 2. With respect to getting the sequence alignment right for the residues covered within each SRS, only SRS2 and SRS3 located at the C-terminal end of helix F and N-terminal end of helix G, respectively, could be problematic in this respect, as no clear sequence markers are found in the vicinities of these regions. However, even though the correct sequence alignment could be derived for the remaining SRSs, the expected quality of the structural model in these regions is still questionable, with average rmsd values for the C_α carbons of the residues within those regions in the range of 2 to 4 Å.

Finally, it may be worth comparing the alignment derived in the present work (Fig. 2) with other multiple sequence alignments available for this set of CYPs. To this aim, it is remarkable to notice that, despite using a completely different superposition method and similarity metric, the structure-based sequence alignment available from the HOMSTRAD database[57] agrees well with the alignment presented here. This is particularly reflected by the fact that the alignment of the structural core, as defined in this work, shows a perfect match in 98% of its alignment positions (this is, 140 out of the 143 alignment positions shaded in Fig. 2). The differences between the 2 alignments are thus essentially concentrated in the structurally variable regions highlighted previously, including some of the substrate recognition sites. In this respect, exactly the same alignment is derived in the regions defining SRS4 and SRS6. However, some minor differences are identified in SRS1, SRS3, and SRS5, and significant variations are found in SRS2, which reinforces the statement made above about the expected quality of any structural model derived for these regions. In contrast, comparison with a purely sequence-based multiple sequence alignment[58] reveals that only 75% of the structural core (i.e., 107 out of the 143 alignment positions shaded in Fig. 2) is aligned in the same way. In addition, major differences are found in the structurally more variable regions, with only SRS4 having exactly the same alignment among all substrate recognition sites. This strongly emphasizes the need for structural information and the relevance of deriving structure-based sequence alignments in this particular family of enzymes.

Comparative Analysis of CYP 2B4, 2C5, 2C8, and 2C9

The recent determination of crystal structures for four mammalian P450s of the CYP2 family (namely, 2B4, 2C5, 2C8, and 2C9) offers an unique opportunity to perform a comparative analysis of representative structures of a class II P450 family (see Table I). In addition, because of its association with the metabolism of drugs in man, the availability of these structures represents a significant step forward toward applying structure-based approaches to drug metabolism through the modeling of the major forms of CYPs involved in these processes (namely, 1A2, 2A6, 2C9, 2C19, 2D6, 2E1, and 3A4). Accordingly, following the same protocol described above, the crystal structure of CYP2C5 (1dt6) was taken as the reference structure onto which the other structures were superimposed considering C_α carbons only. Once optimal pairwise superimpositions were obtained, all structures were then allowed to modify their relative orientation to maximize the overlap between all 4 proteins. That generated the final multiprotein structural superimposition from which a structure-based sequence alignment was then derived.

TABLE IV. Percentage of Sequence Identity and Structural Similarity (Expressed as $100 \cdot S_{AB}$) Between 4 Members of the CYP2 Family and Their Representative Structures

	1po5	1dt6	1pq2	1og2
1po5	—	75.5	75.4	75.2
1dt6	46.1	—	90.9	89.6
1pq2	50.1	68.2	—	92.9
1og2	47.1	71.5	78.4	—

The pairwise structural similarities and sequence identities obtained from the multi-protein structural superimposition and structure-based sequence alignment, respectively, are given in Table IV. It is worth emphasizing that among the set of CYP2 structures, the structure of CYP2B4 (1po5) adopts an open conformation and has thus significant structural differences with the other structures.[48] This notwithstanding, structure similarity values between CYP2B4 and any member of the CYP2C subfamily are over 75%, with sequence identities around 50%. In contrast, sequence identities between the 3 CYP2C subfamily members are all around 70% or higher, with structure similarities around 90%. A comparison of the sequence identity and structural similarity ranges for the set of four CYP2 structures (46–78% and 75–93%, respectively) with those reported previously for the set of 9 diverse P450s (10–27% and 53–76%, respectively) justifies convincingly the inclusion of only one CYP2 structure in the previous set (CYP2C5, 1dt6) and the need for a separate, more focused, analysis of the CYP2 family.

The structure-based sequence alignment for the four CYP2 family members is presented in Figure 8. Following the same criteria as above, all sequences were extracted directly from the PDB structure file, implying that for CYP2C5 (1dt6) the sequence for the FG loop is missing. Since all sequences have approximately the same length and given the relatively high sequence identity, comparison with other purely sequence-based multiple alignments available[21,59] exposes, not surprisingly, only minor differences. One of them involves the alignment of the highly conserved PPGP motif at the N-terminus of CYP2 sequences. Inspection of the multiple structure superimposition unambiguously reveals a 2-residue shift of the PPGP motif in CYP2C5 with respect to the other 3 sequences. A second difference involves the alignment of helix C, where instead of a full sequence alignment, the structure-based alignment incorporates a 4-residue shift in CYP2B4. Finally, a subtle difference can be observed in the alignment of the GNFK motif located in the loop connecting the Meander and the Cys-pocket. The structure superimposition discloses the presence of an additional residue in that loop in CYP2C5.

The same coloring criteria as above were also used for the mapping of structural information in the sequence alignment. Overall, 254, 321, 345, 358, and 376 alignment positions, out of a total of 471, were found to have the C_α carbon atoms corresponding to the aligned residues within 1.0, 1.5, 2.0, 2.5, and 3.0 Å RMSD, respectively. Taking arbitrarily an RMSD value of 2 Å as an acceptable level of structural conservation (all alignment positions shaded in Fig. 8), this means that at least 73% of a P450 structure is found to be generally well preserved structurally within the CYP2 family, even when members of this family undergo dramatic conformational changes. Compared to the percentage of structure generally conserved within the P450 superfamily proposed above (ca. 30%), a value of 73% represents a significant expansion in structure conservation within a given class II P450 family. In order to be able to assess the additional coverage of structure conservation

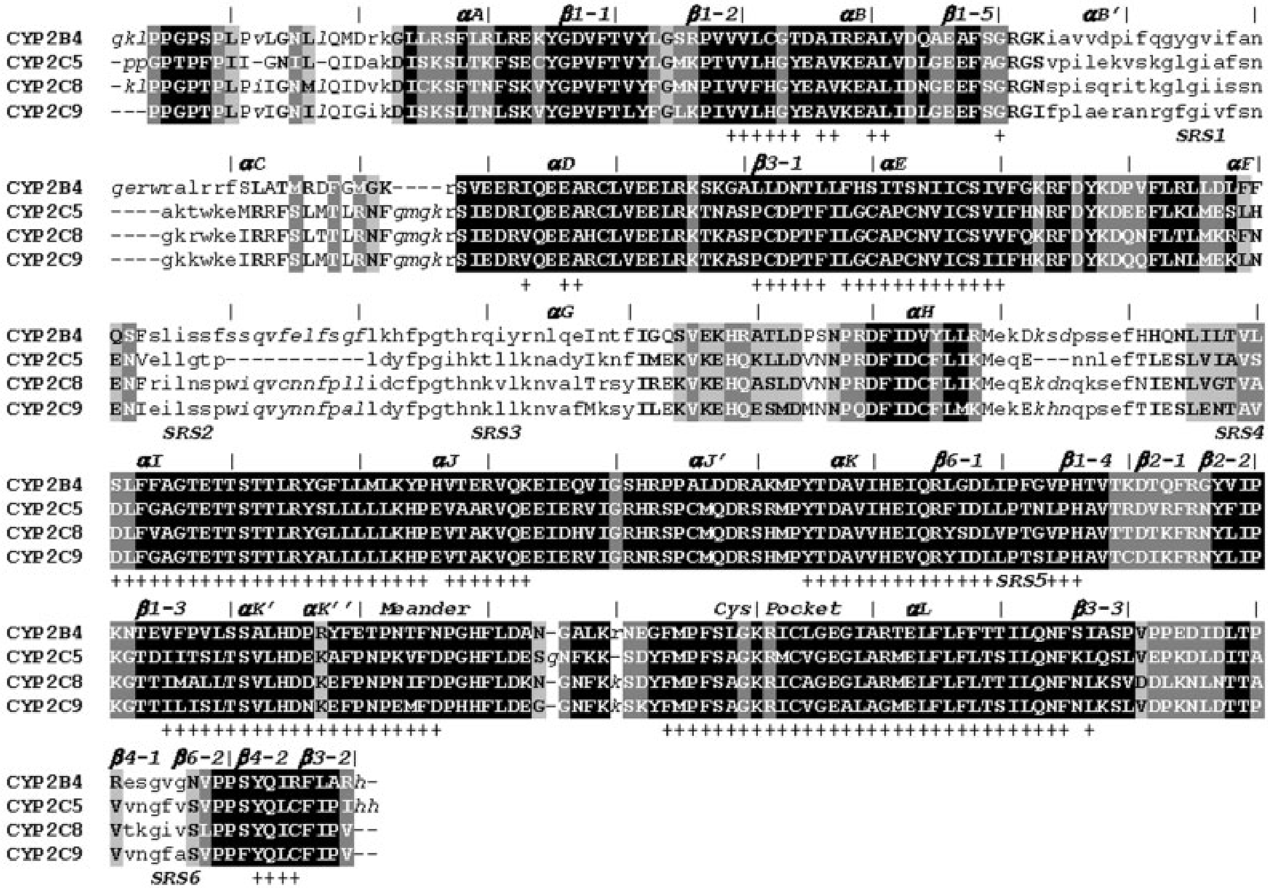


Fig. 8. Structure-based sequence alignment of 4 members of the CYP2 family. Main secondary elements are marked above the sequences, whereas SRSs are indicated below the sequences. Nonshaded and shaded alignment positions reflect RMSD values above and below 2Å, respectively: *a*: nonaligned; a: >3.0; A: [2.5,3.0]; **A**: [2.0,2.5]; light-gray shaded **A**: [1.5,2.0]; dark-gray shaded white **A**: [1.0,1.5]; black shaded white **A**: ≤1.0. Positions within 2 Å RMSD identified previously among a set of 9 diverse cytochromes P450 (see Fig. 2) are marked with +.

obtained for the CYP2 family, the alignment positions within 2 Å RMSD identified previously among the set of nine diverse CYPs (see Fig. 2) have been marked (+) in Figure 8. As can be observed, the only regions where high spatial variability is found are essentially located in helices B′ and C, the C-terminal end of helix F, helix G, and the N-terminal site of helix I, consistent with these secondary structure elements being the ones mainly involved in the conformational changes of P450 structures.

The extension of the highly variable regions identified above may have been somehow enlarged by the inclusion of the open conformation of CYP2B4 (1po5). To investigate further the extent at which inclusion of a structure with an open conformation limits the overall structure conservation found within the CYP2 family, the degree of structure conservation within the CYP2C subfamily was compared to that obtained when considering all CYP2 structures available. The results are depicted in Figure 9, where each circle represents a position in the alignment with an RMSD value below 2 Å. Indeed, when the structure of CYP2B4 (1po5) is excluded from the analysis, the only regions identified with appreciable spatial variability are located in the B′ helix and in the FG and HI loops. A total of 402 most structurally conserved positions were identified for the subset of 3 structures representative of the CYP2C subfamily. This represents 57 additional positions to those identified from all CYP2 structures, and extents potentially the level of overall structure conservation within the CYP2 family to over 85% for structures in a closed conformation.

The implications of these results for the homology modeling of other members of the CYP2 family are important. A degree of over 85% of structure conservation should guarantee that models of good overall quality can be derived for the CYP2 family members of the major P450 forms involved in drug metabolism, namely, CYP 2A6, 2C19, 2D6, and 2E1.[21] In addition, the local quality of the models in the 6 SRSs should also be acceptable. As derived from the conservation profile exhibited by the 3 CYP2C structures in a closed conformation (Fig. 9), only SRS1 would retain a certain degree of spatial uncertainty. However, given the low sequence identities with members of the CYP2 family, the same degree of structural conservation cannot be guaranteed for models derived for the remaining two P450 forms of key relevance in drug metabolism, namely, CYP1A2 and CYP3A4. In these cases, the availability of representative structures for the CYP1 and CYP3 families will be an important milestone

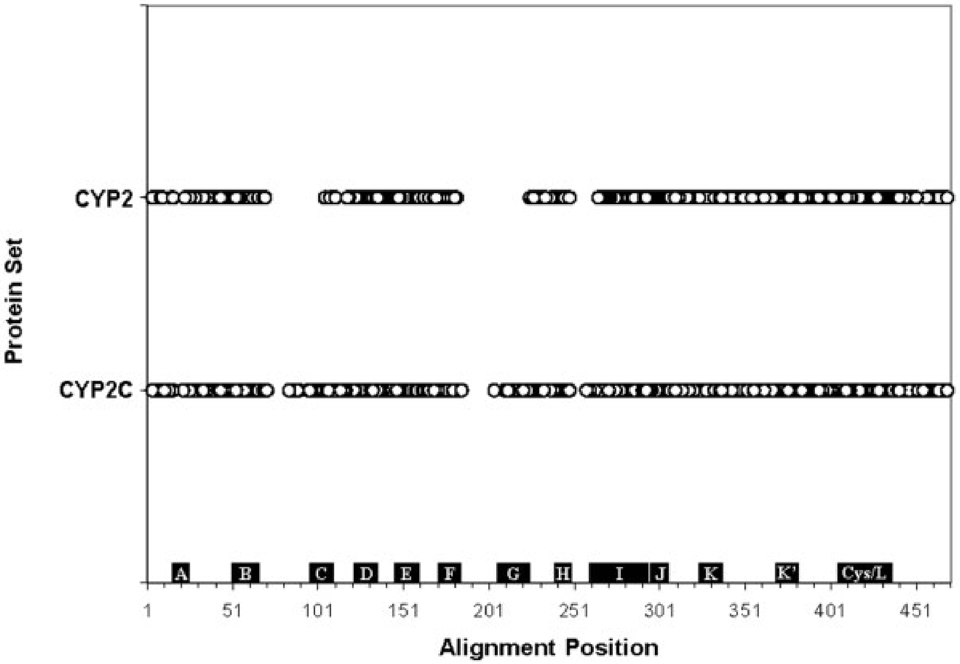


Fig. 9. Structure conservation profile within 2 Å RMSD along the sequence alignment for different subsets of 4 members of the CYP2 family (see text for subset details).

towards developing generic and more predictive structure-based approaches to drug metabolism.[60]

CONCLUSIONS

Due to their key role in metabolism and detoxification processes, the superfamily of cytochromes P450 continues to generate renovating interest with a constantly increasing number of X-ray crystal structures being determined and particularly the recent availability of several representative structures of mammalian enzymes. Gaining a better understanding of the levels of structure conservation, both generically along the entire superfamily and specifically within a given family, may help addressing drug metabolism issues from a structure-based design perspective. The present study provides a precise mapping of the location and degree of structure conservation among representative structures of 9 diverse CYPs and 4 members of the CYP2 family. The results suggest that models within 2 Å RMSD in over 85% of the actual experimental structure are attainable for CYP 2A6, 2C19, 2D6, and 2E1. Together with CYP 2C9, for which representative structures are already available,[50] they complete a list of P450 forms responsible for the metabolism of over 40% of the clinically available drugs. This structural information can then be used effectively in the design and optimization of ligands devoid of affinity for these xenobiotic metabolizing enzymes. Unfortunately, the results suggest also that at present models derived for CYP 1A2 and 3A4 can be guaranteed to be within 2 Å RMSD in only around 30% of the actual experimental structure, thus currently limiting their potential applicability in drug design activities.

REFERENCES

1. Ortiz de Montellano PR (Editor). Cytochrome P450: structure, mechanism and biochemistry. New York: Plenum Press; 1995.
2. Guengerich FP. Reactions and significance of cytochrome P450 enzymes. J Biol Chem 1991;266:10019–10022.
3. Porter TD, Coon MJ. Cytochrome P450: multiplicity of isoforms, substrates, and catalytic and regulatory mechanisms. J Biol Chem 1991;266:13469–13472.
4. Smith DA, Ackland MJ, Jones BC. Properties of cytochrome P450 isoenzymes and their substrates Part 1: Active site characteristics. Drug Discov Today 1997;2:406–414.
5. Smith DA, Ackland MJ, Jones BC. Properties of cytochrome P450 isoenzymes and their substrates: Part 2. Properties of cytochrome P450 substrates. Drug Discov Today 1997;2:479–486.
6. Sono M, Roach MP, Coulter ED, Dawson JH. Heme-containing oxygenases. Chem Rev 1996;96:2841–2887.
7. Degtyarenko KN. Structural domains of P450-containing monooxygenase systems. Protein Eng 1995;8:737–747.
8. Peterson JA, Graham SE. A close family resemblance: the importance of structure in understanding cytochromes P450. Structure 1998;6:1079–1085.
9. Werck-Reichhart D, Feyereisen R. Cytochromes P450: a success story. Gen Biol 2000;1:1–9.
10. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
11. Nelson DR, Kamataki T, Waxman DJ, Guengerich FP, Estabrook RW, Feyereisen R, Gonzalez FJ, Coon MJ, Gunsalus IC, Gotoh O, Okuda K, Nebert DW. The P450 superfamily: update on new sequences, gene mapping, accession numbers, early trivial names of enzymes and nomenclature. DNA Cell Biol 1993;12:1–51.
12. Hasemann CA, Kurumbail RG, Boddupalli SS, Peterson JA, Deisenhofer J. Structure and funtion of cytochromes P450: a comparative analysis of three crystal structures. Structure 1995;3: 41–62.
13. Jean P, Pothier J, Dansette PM, Mansuy D, Viari A. Automated multiple analysis of protein structures: application to homology modeling of cytochromes P450. Proteins 1997;28:388–404.
14. Smith DA, Jones BC, Walker DK. Design of drugs involving the concepts and theories of drug metabolism and pharmacokinetics. Med Res Rev 1996;16:243–266.

15. Raucy JL, Allen SW. Recent advances in P450 research. Pharmacogenomics J 2001;1:178–186.
16. Szklarz GD, Halpert JR. Molecular modeling of cytochrome P450 3A4. J Comput Aided Mol Des 1997;11:265–272.
17. Lozano JJ, López-de-Briñas E, Centeno NB, Guigó R, Sanz F. Three-dimensional modelling of human cytochrome P450 1A2 and its interaction with caffeine and MeIQ. J Comput Aided Mol Des 1997;11:395–408.
18. de Groot MJ, Ackland MJ, Horne VA, Alex AA, Jones BC. Novel approach to predicting P450-mediated drug metabolism: development of a combined protein and pharmacophore model for CYP2D6. J Med Chem 1999;42:1515–1524.
19. de Groot MJ, Alex AA, Jones BC. Development of a combined protein and pharmacophore model for cytochrome P450 2C9. J Med Chem 2002;45:1983–1993.
20. Kirton SB, Baxter CA, Sutcliffe MJ. Comparative modelling of cytochromes P450. Adv Drug Deliv Rev 2002;4:385–406.
21. Lewis DFV. Modelling human cytochromes P450 involved in drug metabolism from the CYP2C5 crystallographic template. J Inorg Biochem 2002;91:502–514.
22. Koehl P. Protein structure similarities. Curr Opin Struct Biol 2001;11:348–353.
23. Godzik A. The structural alignment between two proteins: is there a unique answer? Protein Sci 1996;5:1325–1338.
24. Feng ZK, Sippl MJ. Optimum superimposition of protein structures: ambiguities and implications. Fold Des 1996;1:123–132.
25. Orengo CA, Taylor WR. A local alignment method for protein structure motifs. J Mol Biol 1993;233:488–497.
26. Koch I, Lengauer T, Wanke E. An algorithm for finding maximal common subtopologies in a set of protein structures. J Comput Biol 1996;3:289–306.
27. Escalier V, Pothier J, Soldano H, Viari A. Pairwise and multiple identification of three-dimensional common substructures in proteins. J Comput Biol 1998;5:41–56.
28. Lehtonen JB, Denessiouk K, May ACW, Johnson MS. Finding local structural similarities among among families of unrelated protein structures: a generic non-linear alignment algorithm. Proteins 34;1999:341–355.
29. Ochagavía ME, Wodak S. Progressive combinatorial algorithm for multiple structural alignments: application to distantly related proteins. Proteins 2004;55:436–454.
30. Maggiora GM, Rohrer DC, Mestres J. Comparing protein structures: a Gaussian-based approach to the three-dimensional structural similarity of proteins. J Mol Graph Mod 2001;19:168–178.
31. Sheridan RP, Holloway MK, McGaughey G, Mosley RT, Singh SB. A simple method for visualizing the differences between related receptor sites. J Mol Graph Model 2002;21:71–79.
32. Mestres J. Gaussian-based alignment of protein structures: deriving a consensus superposition when alternative solutions exist. J Mol Model 2000;6:539–549.
33. Bader RFW. Atoms in molecules: a quantum theory. Oxford, UK: Oxford University Press; 1990.
34. Oldfield TJ, Hubbard RE. Analysis of Ca geometry in protein structures. Proteins 1994;18:324–337.
35. Carbó R, Leyda L, Arnau M. How similar is a molecule to another?: an electron density measure of similarity between two molecular structures. Int J Quantum Chem 1980;17:1185–1189.
36. Shenkin PS, Erman B, Mastrandrea LD. Information-theoretical entropy as a measure of sequence variability. Proteins 1991;11: 297–313.
37. Pei J, Grishin NV. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 2001;17:700–712.
38. Shannon CE, Weaver W. The mathematical theory of communication. Urbana: University of Illinois Press; 1949.
39. Poulos TL, Finzel BC, Howard AJ. Crystal structure of substrate-free *Pseudomonas putida* cytochrome P450. Biochemistry 1986;25: 5314–5322.
40. Cupp-Vickery JR, Poulos TL. Structure of cytochrome P450eryF involved in erythromycin biosynthesis. Nat Struct Biol 1995;2:144–153.
41. Hasemann CA, Ravichandran KG, Peterson JA, Deisenhofer J. Crystal structure and refinement of cytochrome P450terp at 2.3 Å resolution. J Mol Biol 1994;236:1169–1185.
42. Park SY, Yamane K, Adachi S, Shiro Y, Sligar SG. Thermophilic cytochrome P450 (Cyp119) from *Sulfolobus sulfataricus*: high resolution structural origin of its thermostability and functional properties. Acta Crystallogr D Biol Crystallogr 2000;56:1173–1175.
43. Leys D, Mowat CG, McLean KJ, Richmond A, Chapman SK, Walkinshaw MD, Munro AW. Atomic structure of *Mycobacterium tuberculosis* CYP121 to 1.06 Å reveals novel features of cytochrome P450. J Biol Chem 2003;278:5141–5147.
44. Podust LM, Poulos TL, Waterman MR. Crystal structure of cytochrome P450 14α-sterol demethylase (CYP51) from *Mycobacterium tuberculosis* in complex with azole inhibitors. Proc Natl Acad Sci USA 2001;98:3068–3073.
45. Park SY, Shimizu H, Adachi SI, Nakagawa A, Tanaka I, Nakahara K, Shoun H, Obayashi E, Nakamura H, Iizuka T, Shiro Y. Crystal structure of nitric oxide reductase from denitrifying fungus *Fusarium oxysporum*. Nat Struct Biol 1997;4:827–832.
46. Li H, Poulos TL. Modeling protein–substrate interactions in the haem domain of cytochrome P450BM3. Acta Crystallogr D Biol Crystallogr 1995;51:21–32.
47. Williams PA, Cosme J, Sridhar V, Johnson EF, McRee DE. Mammalian microsomal cytochrome P450 monooxygenase: structural adaptations for membrane binding and functional diversity. Mol Cell 2000;5:121–131.
48. Scott EE, He YA, Wester MR, White MA, Chin CC, Halpert JR, Johnson EF, Stout CD. An open conformation of mammalian cytochrome P450 2B4 at 1.6 Å resolution. Proc Natl Acad Sci USA 2003;100:13196–13201.
49. Schoch GA, Yano JK, Wester MR, Griffin KJ, Stout CD, Johnson EF. Structure of human microsomal cytochrome P450 2C8. J Biol Chem 2004;279:9497–9503.
50. Williams PA, Cosme J, Ward A, Angove HC, Vinkovic DM, Jhoti H. Crystal structure of human cytochrome P450 2C9 with bound warfarin. Nature 2003;424:464–468.
51. Ravichandran KG, Boddupalli SS, Hasemann CA, Peterson JA, Deisenhofer J. Crystal structure of hemoprotein domain of P450BM3, a prototype for microsomal P450s. Science 1993;261: 731–736.
52. Szklarz GD, Halpert JR. Use of homology modeling in conjunction with site-directed mutagenesis for analysis of structure–function relationships of mammalian cytochromes P450. Life Sci 1997;61: 2507–2520.
53. Stayton PS, Poulos TL, Sligar SG. Putidaredoxin competitively inhibits cytochrome *b*5–cytochrome P450cam association: a proposed molecular model for a cytochrome P450cam electron-transfer complex. Biochemistry 1989;28:8201–8205.
54. Sevrioukova IF, Li H, Zhang H, Peterson JA, Poulos TL. Structure of a cytochrome P450–redox partner electron-transfer complex. Proc Natl Acad Sci USA 1999;96:1863–1868.
55. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
56. Gotoh O. Substrate recognition sites in cytochrome P450 Family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences. J Biol Chem 1992;267:83–90.
57. Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: A database of protein structure alignments for homologous families. Protein Sci 1998;7:2469–2471. http://www-cryst.bioc.cam.ac.uk/homstrad
58. http://drnelson.utmem.edu/p450bpub237.html
59. http://drnelson.utmem.edu/p450apub192.html
60. Astex Technology, Ltd. Crystal structure of cytochrome P450 3A4 and its use. Priority patent US60241063, 25 October 2002.