U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

COMPUTATIONAL STUDIES ON CYTOCHROMES P450

Rute A. Rodrigues da Fonseca

Departamento de Química

2006

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

COMPUTATIONAL STUDIES ON CYTOCHROMES P450

Rute A. Rodrigues da Fonseca

Doctoral Thesis in Chemistry

2006

Acknowledgments

I would first like to thank the person that offered me the opportunity to do this work, Maria João Ramos, who has warmly received me and provided an attentive guidance through the years I spent in the Theoretical Chemistry group. Next I would like to express my gratitude to André Melo for his unconditional support and his availability for any discussion. I would also like to thank Cristina Menziani, whose assistance was very important in the first part of my PhD. These first years were also made easier by Elsa, who was always ready to lend a hand on a rookie. Thanks also to Agostinho, who had a major role on the last part of my PhD and my upcoming future, for his helpfulness. Thanks to all the people (present and past) in lab 3.28 for the friendship, the interesting discussions and the readily available help. To Susana *arigato* for the *découvertes à deux*. Of course I must also thank Cristina and Nelson, my lunch and teatime buddies, for everything. Thanks also to the other members in the lab, for their companionship. Finally I would like to thank those that are always there for me, my family and João, and particularly my cousin Augusta, for the early lessons in Science.

I would like to acknowledge the founding sources that through FCT (Fundação para a Ciência e a Tecnologia) have provided my scholarship (SFRH/BD/7089/2001):

Ciência.Inovação 2010 Programa Operacional Ciência e Inovação 2010

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

UNIÃO EUROPEIA

Fundo Social Europeu

To João and Augusta

'It is a capital mistake to theorise before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.'

Sherlock Holmes

in *A Scandal in Bohemia* by Sir Arthur Conan Doyle

Abstract

The work presented in this thesis is the outcome of the application of a variety of computational methods to study different aspects of Cytochromes P450 (CYPs) biological profile. CYPs constitute a superfamily of enzymes involved in the oxidative metabolism of a wide range of compounds, with roles both in anabolism (e.g. biosynthesis of steroids) and catabolism (e.g. degradation of fatty acids). Their role in the disposal of xenobiotics interferes with human health in both positive and negative ways, with some of its members being accounted for the controlled biotransformation of pharmaceutical drugs and others taking part in the activation of carcinogenic compounds.

The initial approach focused on exploring protein structure and function. Homology models were built for human and rat CYP1A2, an enzyme involved in the activation of carcinogenic heterocyclic amines present in cooked red meat. Two of such amines were docked in the active site of the models and conclusions were drawn in relation to the different metabolites produced by the two enzymes.

The following step involved the study of the interaction of human CYP1A2 with known inhibitors, two groups of naturally occurring flavonoids. The physicochemical characteristics of these molecules were compared and related to the differential inhibitory character they exhibited towards human CYP1A2. This was accompanied by a thorough analysis of the specific interactions established by the flavonoids inside the active site, with the correspondent quantification by calculation of ligand/receptor binding energies.

Finally, various computational genomics methods were used to analyse the functional divergence of CYP2 enzymes, largely involved in the metabolism of different pharmaceutical agents. Phylogenetic studies were carried out together with statistical analyses of detection of functional divergence and positive selection using both amino acid and nucleotide sequences. All the results were critically discussed considering the available structural data of CYP2 enzymes.

Resumo

O trabalho apresentado nesta tese é o resultado da aplicação de vários métodos computacionais ao estudo de diferentes aspectos do papel biológico dos Citocromos P450 (CYPs). Os CYPs constituem uma superfamília de enzimas envolvidas no metabolismo oxidativo de uma grande variedade de compostos, intervindo tanto em processos anabólicos (biossíntese de esteróides) como catabólicos (degradação de ácidos gordos). O papel destas enzimas no metabolismo de compostos xenobióticos interfere ambos positiva e negativamente na saúde humana, tomando parte tanto na biotransformação controlada de medicamentos como na activação de compostos carcinogénicos.

O primeiro estudo realizado abordou a estrutura e a função destas enzimas. Foram construídos modelos por homologia para o CYP1A2 humanos e do rato, enzimas implicados na activação de aminas heterocíclicas carcinogénicas, compostos existentes em carne vermelha cozinhada. Foi estudada a maneira de ligação de dois destes compostos ao centro activo da enzima e retiradas ilações relativas à produção de diferentes metabolitos pelas duas enzimas. A etapa seguinte implicou o estudo da interacção do CYP1A2 humano com dois grupos de flavonóides naturais, conhecidos inibidores da enzima. As características físico-químicas destas moléculas foram comparadas e relacionadas com o seu carácter inibitório diferencial relativo ao CYP1A2 humano. Isto foi acompanhado por uma análise exaustiva das interacções específicas estabelecidas por flavonóides dentro do centro activo, com a respectiva quantificação por cálculo das energias de ligação ligando/receptor.

Por fim, vários métodos de genómica computational foram utilizados para analisar a divergência funcional das enzimas CYP2C que estão largamente envolvidos no metabolismo de diferentes agentes farmacêuticos. Foram feitos estudos filogenéticos conjuntamente com análises estatísticas de detecção de divergência funcional e de selecção positiva utilizando ambas sequências de aminoácidos e de nucléotidos. Os resultados foram analisados face às suas consequências nas estruturas tridimensionais disponíveis de enzimas CYP2.

Resumé

Le travail présenté dans cette thèse est le résultat de l'application de plusieurs méthodes informatiques à l'étude de différents aspects du profil biologique des Cytochromes P450 (CYPs). Les CYPs parfont une superfamille d'enzymes engagées dans le métabolisme oxidatif d'une grande variété de composés, intervenant dans des processus anaboliques (biosynthèse de stéroïdes) et cataboliques (dégradation d'acides gras). L'intervention de ces enzymes dans le métabolisme des xénobiotiques joue un rôle aussi positif que négatif pour la santé humaine, puis qu'elles biotransforment des médicaments mais peuvent aussi activer des composés carcinogéniques.

La première étude réalisée a abordée la structure et la fonction de ces enzymes. La modélisation par homologie a été utilizé pour construire des structures des CYP1A2 humaine et du rat, enzymes impliquées dans l'activation d'amines hétérocycliques carcinogéniques, composés qu'existent dans la viande rouge cuite. L'interaction de deux de ces composés avec le centre actif de l'enzyme a été étudiée et des illations concernant les différents métabolites produits par les deux enzymes ont été déduites.

L'étape suivante a impliqué l'étude de l'interaction du CYP1A2 humain avec deux groupes de flavonoïdes naturels, que sont des inhibiteurs de cet enzyme. Les caractéristiques physicochimiques de cettes molécules ont été comparées et rapportées avec le respective pouvoir inhibitoire du CYP1A2 humain. Ceci a été accompagné par une analyse exhaustive des interactions spécifiques établies par les flavonoïdes à l'intérieur du centre actif, avec la respective quantification par calcul des énergies de liaison.

À la fin, plusieurs méthodes de bioinformatique ont été utilisées pour analyser la divergence fonctionnelle des enzymes CYP2C qui sont engagées dans le métabolisme de plusieurs agents pharmaceutiques. Des études phylogénétiques ont été realizées conjointement avec des analyses statistiques de détection de divergence fonctionnelle et de sélection positive en utilisant les séquences de acides aminés et de nucléotides. Les résultats ont été analysés à la perspective de leurs conséquences dans les structures tridimensionnelles disponibles d'enzymes CYP2.

Contents

1. Introduction

1.1. Cytochromes P450

1.1.1. The Enzyme Superfamily

Cytochromes P450 (CYPs) comprise a superfamily of enzymes involved in various physiological functions [1;2]. Their name arises from their intense absorption band at 450 nm when complexed with carbon monoxide that first led to their identification in liver tissues [3;4]. CYPs first appeared in prokaryotes, when the atmosphere was poor in molecular oxygen, before the development of eukaryotes. This explains why these enzymes are almost ubiquitous in the biosphere, being present in all eukaryotes, most prokaryotes and Archea [5]. CYPs share a common fold and present a high structural conservation in the core of the protein, which reflects a conserved mechanism [5] (see Figure 1 for an overall view of the fold).

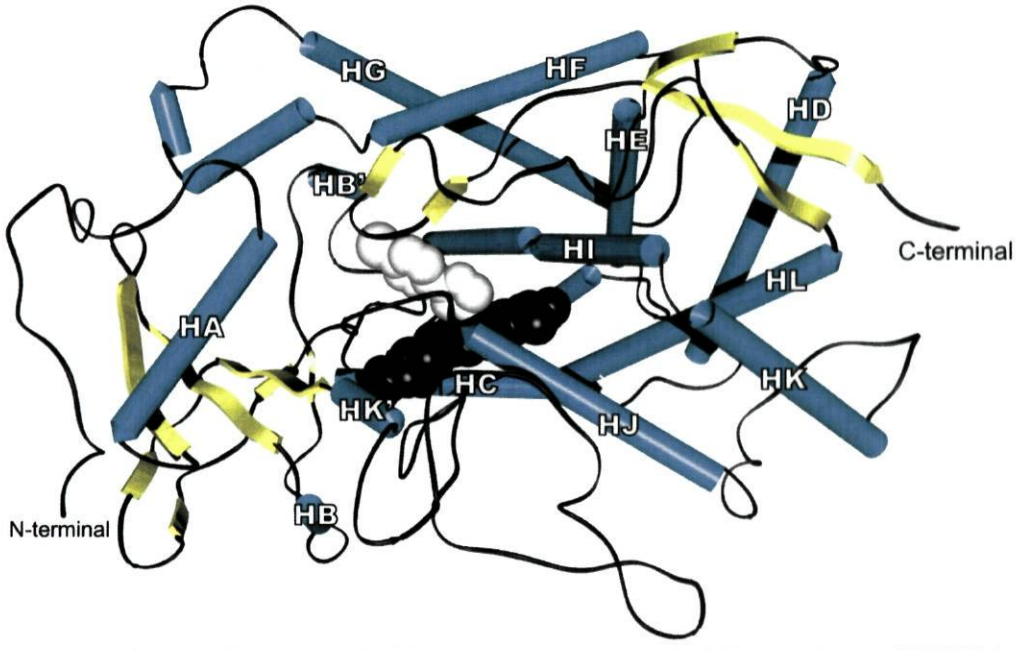


Figure 1. Overall view of a Cytochrome P450 structure: structure of rabbit cytochrome CYP2B4 bound to 4-(4-2 chlorophenyl)imidazole (pdb code 1SUO [6]). The heme is represented in black and the ligand in white.

Plant CYPs are mostly involved in the biosynthesis of natural products. However, in mammals, they have diverse anabolic and catabolic roles. These include the synthesis of steroid hormones thromboxane, cholesterol and bile acid and the degradation of endogenous compounds such as fatty acids, retinoic acids and steroids [7]. CYPs also play a major role in transforming xenobiotic substances into products easier to remove from the body. These include several drugs and environmentally available chemicals, such as carcinogenic

compounds present in food [1], which is why CYPs are so important in pharmaceutical research [2;8].

CYPs catalyse a variety of reactions: carbon hydroxylation (e.g., from a steroid, alkane, etc.), heteroatom oxygenation (P450s have been shown to add oxygen to N, S, P, and I) and epoxidation (products of which can be unstable and react with nucleophilic groups in macromolecules such as the P450 enzyme itself and , e.g, DNA) are the most common [7;9]. They have also been shown to play a role in the desaturation of fatty acids and ring expansion/formation [7].

CYPs have a heme prosthetic group – a protoporphyrin IX macrocycle bound to an iron atom (Figure 2) – which is coordinated to a cysteinate. This amino acid residue is part of the P450 consensus sequence (Phe-X-X-Gly-X-Arg-X-Cys-X-Gly) present in the heme pocket (Figure 4). The sixth ligand varies along the enzyme's reaction cycle.

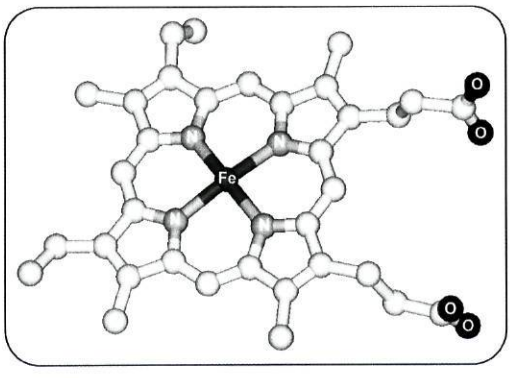


Figure 2. Protoporphyrin IX macrocycle bound to an iron atom (hydrogens atoms are omitted; carbon atoms are represented in white).

The active site of these enzymes is a buried cavity. The substrate enters the cavity through the movement of helices F and G (observed in a mammalian CYP2B4 X-ray structure [10]; Figure 3). Molecular dynamics studies have pointed out the region between B' helix and helix G as the most likely entrance channel in rabbit CYP2C5 [11] .

In the three-dimensional structure of CYPs six substrate recognition sites (SRSs) can be defined [12] (Figure 4). Variability in these areas affects both substrate and product shapes and chemical characteristics.

Microbial enzymes are soluble proteins while eukaryotic CYPs are intrinsic membrane proteins that are present in the endoplasmic reticulii of plant, fungal and animal cells. Animals also possess CYPs in the inner membrane of mitochondria [13].

One very important feature in this kind of enzymes, that divides CYPs in two major groups, is related to a key step in the P450 catalytic cycle – electron transfer from a redox partner.

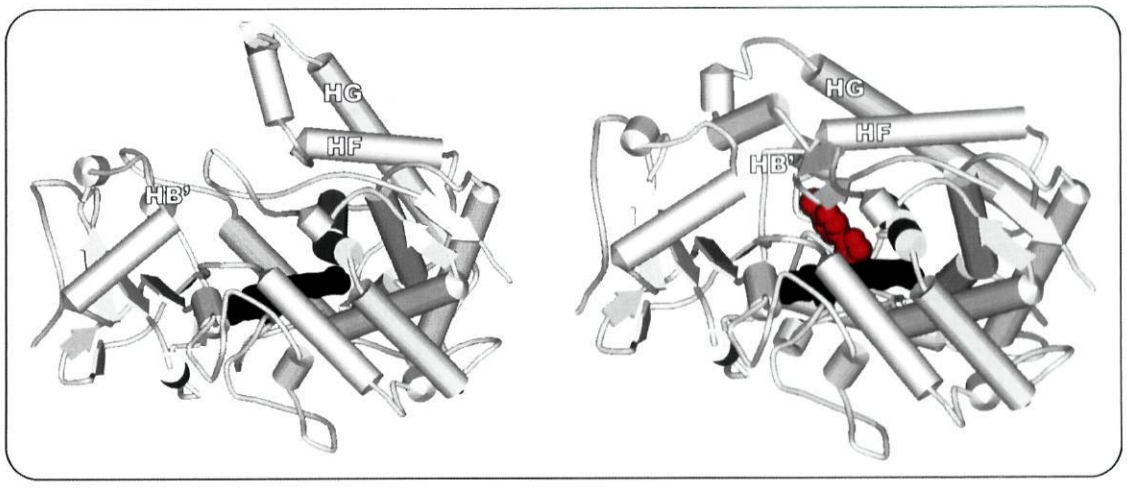


Figure 3. Structure of rabbit cytochrome CYP2B4 in an open conformation (pdb code 1PO5 [10]) and bound to 4-(4-2 chlorophenyl)imidazole (pdb code 1SUO [6]). The heme group is represented in black and the ligand in red.

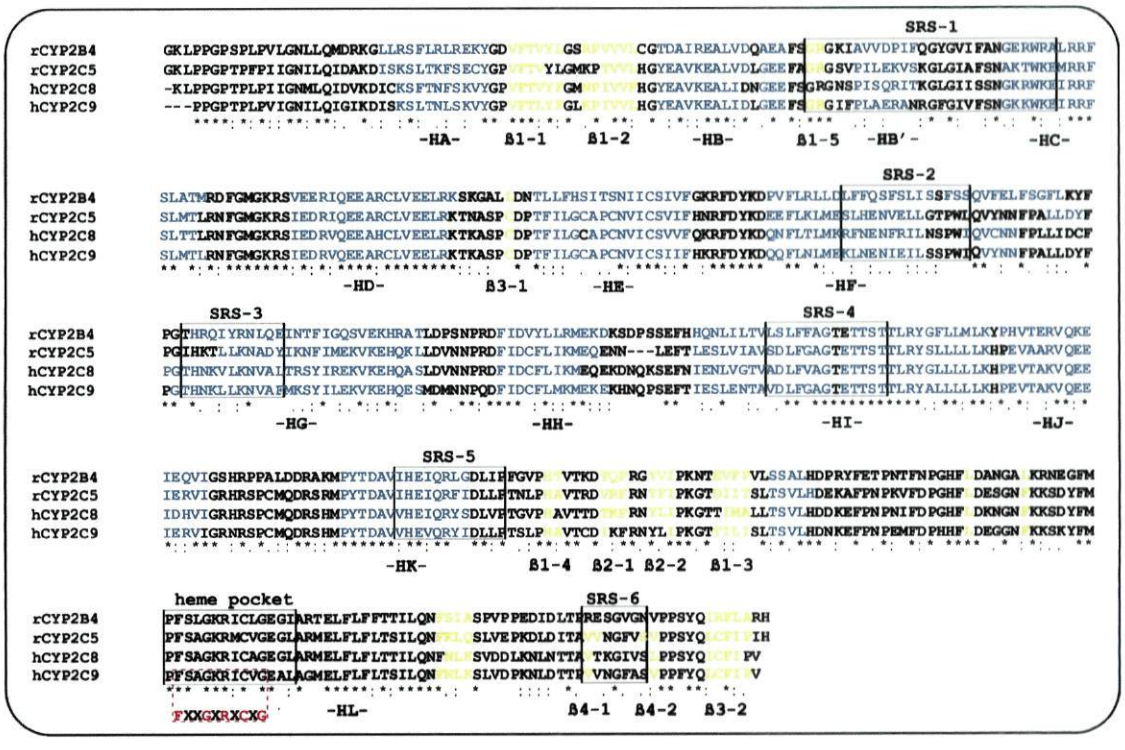


Figure 4. Amino acid sequences of CYP2 family enzymes with available X-ray structure: human CYP2C8 (pdb code 1PQ2); human CYP2C9 (pdb code 1OG5); rabbit CYP2B4 (pdb code 1SUO); rabbit CYP2C5 (pdb code 1NR6). The main structural features, alpha helices (blue) and beta-sheets (yellow) are annotated (the same color scheme used in Figure 1). Substrate binding areas (SRS) are show in boxes.

Class I enzymes use a flavin adenine dinucleotide (FAD) containing reductase and a soluble iron-sulphur protein. Class II enzymes use a FAD and a flavin mononucleotide (FMN) dependent flavoprotein reductase as a redox partner – cytochrome P450 reductase (CPR) [14;15].

The CYP superfamily is classified into families and subfamilies (> 40% or 55% amino acid sequence identity, respectively) [16]. In this thesis we will be looking more closely into CYP families 1 and 2.

CYP1 family includes elements responsible for the activation of carcinogenic compounds to reactive mutagens, the polycyclic aromatic hydrocarbon-inducible CYP1A1 and CYP1A2 enzymes. These enzymes are responsible for the activation of heterocyclic amines (HAs), procarcinogens that result from the pyrolization of creatine or creatinine and amino acids in meat juice when red meat (beef, pork or lamb) is cooked at high temperatures [17] (Figure 5).

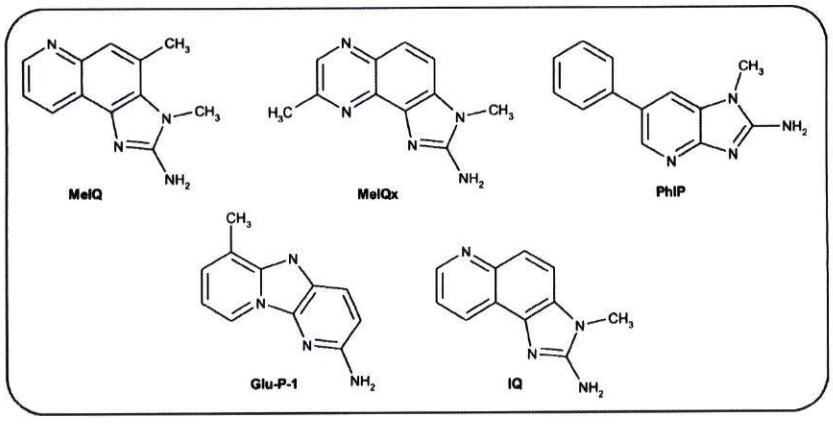


Figure 5. HAs metabolized by CYP1A enzymes.

After phase II metabolism transformations (glucoronidation, sulfation, *O*-methylation), these compounds can cause DNA damage, which results in the development of breast, colorectal and lung cancer [18-20] (Figure 6).

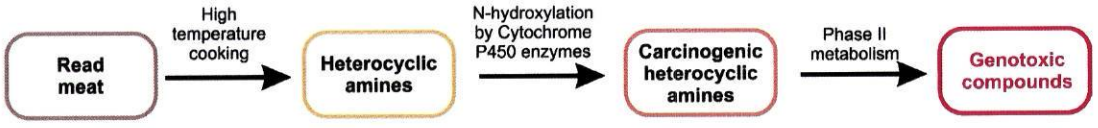


Figure 6. CYPs role in HAs-related carcinogenesis.

In humans and rodents both CYP1A1 and CYP1A2 are inducible by several chemicals, including tobacco smoke [21;22], and exhibit tissue-specific distribution, in which they differ greatly as CYP1A1 exists mainly in extrahepatic tissues and CYP1A2 is preferentially expressed in the liver [18;21;22]. Additionally, CYP1A2 exhibits polymorphic distribution in humans which means that N-hydroxylation of HAs and the associated risk factor for cancer development will be more significant in some populations than others [19;23]. In rodents both CYP1A1 and CYP1A2 carry out the reaction but, in humans, it is mainly CYP1A2 the responsible for it [18] (see Figure 7 and Figure 8 for details on HAs biotransformation).

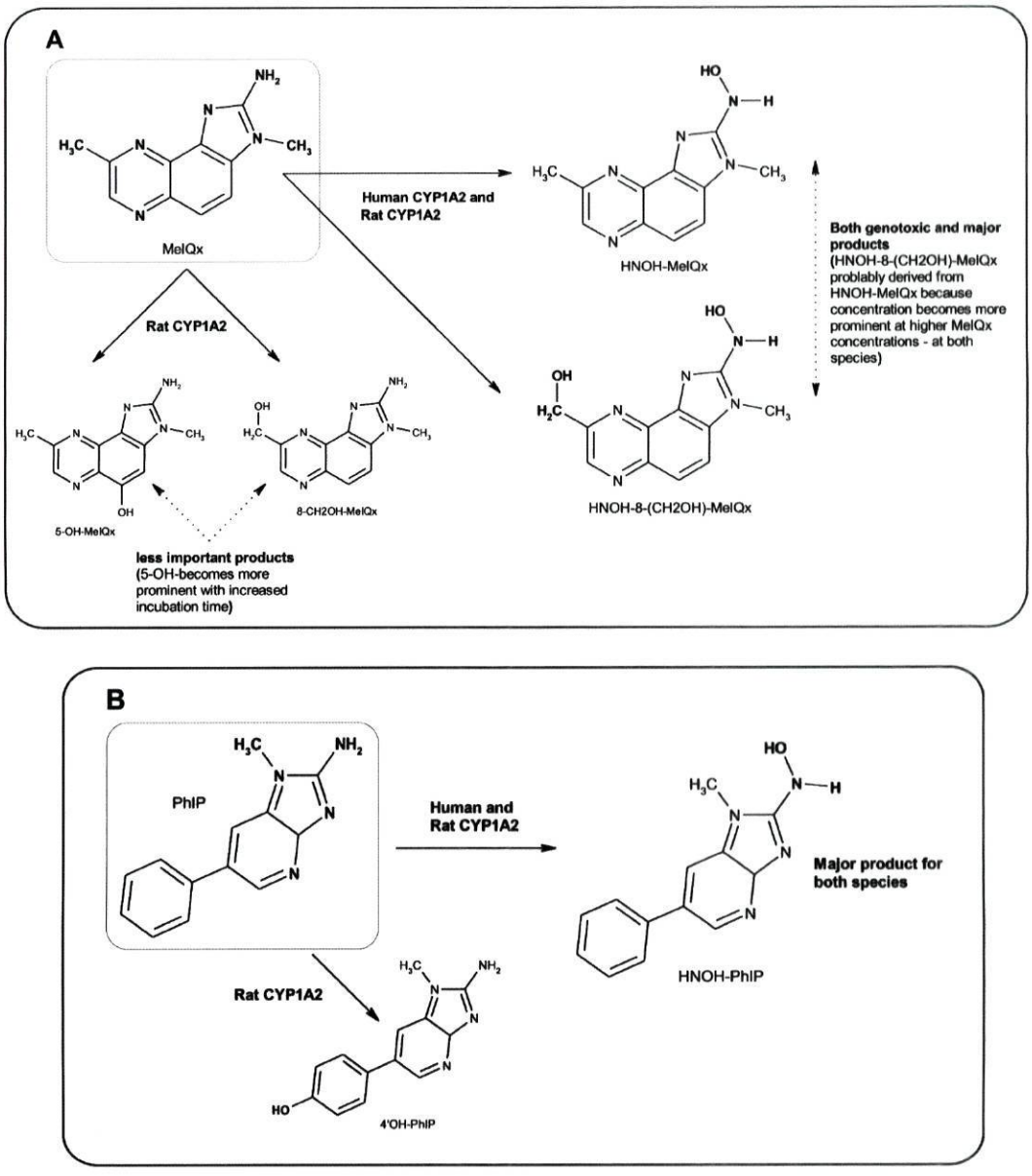


Figure 7. Schematic representation of the experimental results obtained for the activation of two commonly found heterocyclic amines by human and rat CYP1A2 [19]: a) MeIQx (2-amino-1-methyl-6-phenylimidazo[4,5-*f*]pyridine); b) PhIP (2-amino-1-methyl-6-phenylimidazo[4,5-*f*]pyridine).

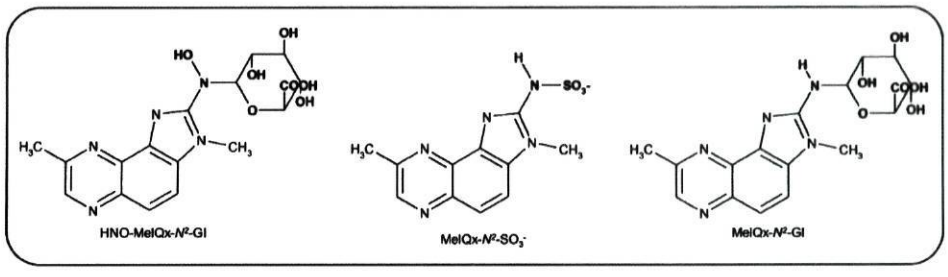


Figure 8. Carcinogenic phase II metabolism products of activated MeIQx (human and rat CYP1A2) [23].

Flavonoids reduce the risk of DNA damage by competing with the HAs for binding to the CYP1A active site [24-29] and therefore inhibiting its catalytic activity. Some of the natural occurring flavonoids that inhibit and react with CYP1A2 are shown in Figure 9.

Flavonoids are polyphenolic compounds constituting one of the best studied and structurally richest group of plant secondary metabolites. These natural compounds are responsible for odor, taste and coloration. They are ubiquitous in various constituents of the human diet such as vegetables, fruit, tea and red wine.

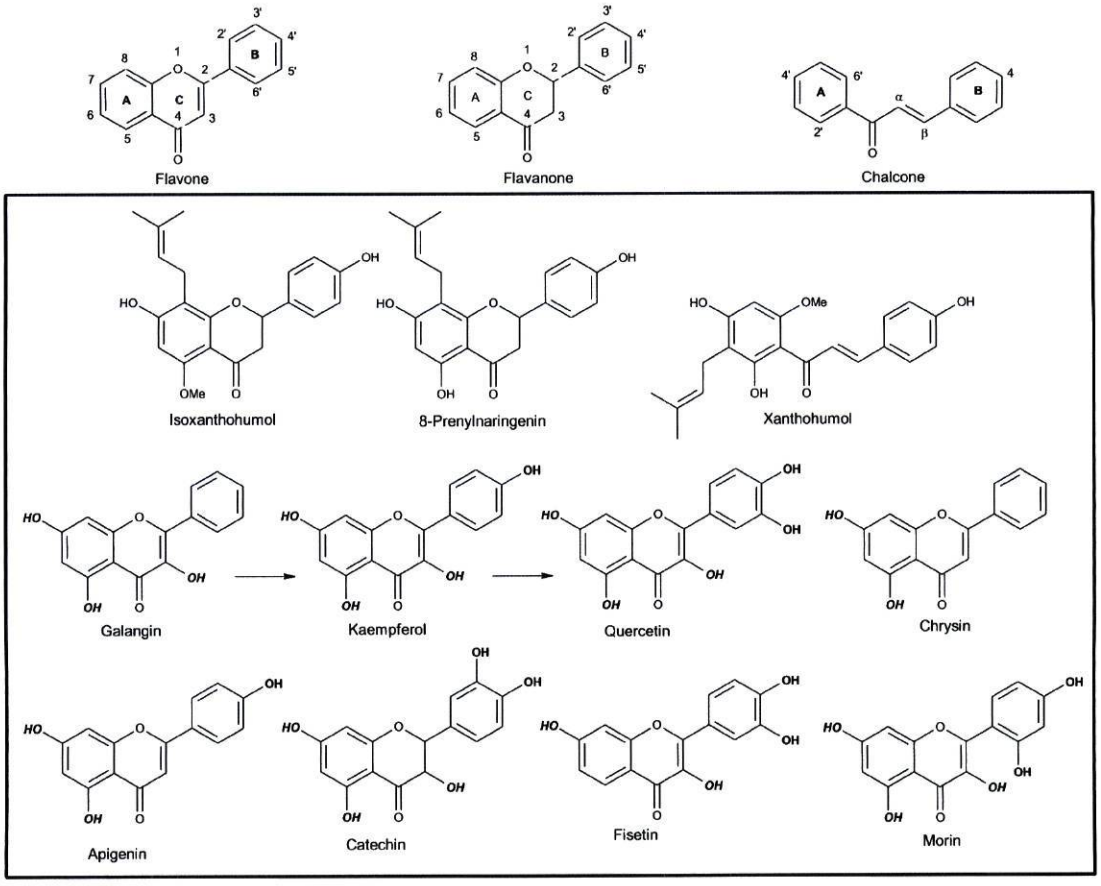


Figure 9. Naturally occurring flavonoids that interact with CYP1A2 [24-26;29-31]. The arrows represent CYP1A2 oxidative activity [32-34].

Their high antioxidant activity has been associated with prevention against diseases caused by oxidative damage and their pharmacological relevance includes also anti-inflammatory and antiviral action [26;27;31;35]. CYPs are involved in interactions with flavonoid compounds in at least three ways: (i) flavonoids induce the biosynthesis of several CYPs; (ii) enzymatic activities of CYPs are modulated (inhibited or stimulated) by these compounds; and (iii) flavonoids are metabolized by several CYPs [28]. As far as CYP1A enzymes are concerned, flavonoids constitute both substrates and reversible inhibitors.

The interaction of flavonoids with CYPs can be clinically significant [27]. This is the case for CYP3A4, the predominant human hepatic and intestinal CYP, which is responsible for the metabolism of around 50% of the current therapeutic agents. Flavonoids such as 7,8-benzoflavone and tangeretin have been described as enzyme stimulators, while flavonolignan and hyperforin from St. John's worth extracts ('hipericão' in Portuguese) act as inhibitors [27].

CYP2 family is widely involved in the metabolism of a variety of different pharmaceutical agents [1;2]. It is the largest and most diverse of CYP families [2], comprising several subfamilies, such as CYP2A (> 10 different enzymes), CYP2B (17 enzymes), CYP2C (40 different enzymes) and CYP2D (> 20 enzymes) [36]. In CYP2A subfamily, one of the members expressed in humans, CYP2A6, is active towards some carcinogenic compounds, and is induced by barbiturates [1;2]. CYP2B members are involved in the metabolism of amphetamines and benzodiazepines [1;2]. CYP2C subfamily metabolizes among others non-steroid anti-inflammatory agents, S-warfarin and phenytoin [1;2]. CYP2C9 alone accounts for approximately 17-20% of the human liver total CYPs content [2]. CYP2D6 is one of the most clinically relevant enzymes as its genetic polymorphisms alter the oxidative metabolism of a wide variety of compounds, including codeine, fluoxetine and fluvoxamine [1;2].

1.1.2. Reaction Mechanism

The crystallographic structures that correspond to the cycle's stable intermediates have been resolved for the widely studied CYPcam from *Pseudomonas putida* [37]. The generalized CYP reaction cycle is shown in Figure 10.

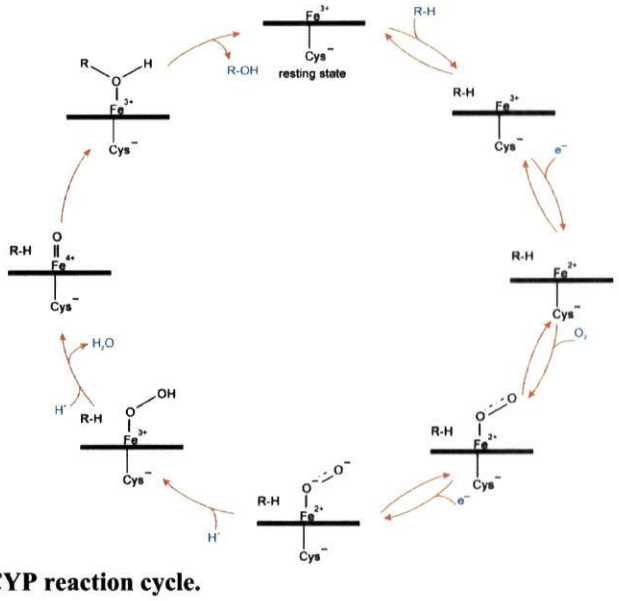


Figure 10. Generalized CYP reaction cycle.

The resting state of the enzyme presents a water molecule as a sixth ligand. The metal is a low-spin Fe^{III} characterized by a spectral of 420 nm wavelength. The substrate displaces the water molecule that is coordinated to the heme and the iron becomes high spin with a wavelength displacement to 390 nm. In the CYPcam X-ray structure it is possible to see a five-coordinated iron atom lying outside the porphyrin ring plane with no ordered water molecules around.

The enzyme receives its first electron from the redox partner and a dioxygen molecule occupies the sixth ligand position in the iron coordination sphere. In the CYPcam X-ray structure this molecule is in an end-on position [37], and two ordered water molecules can be observed between it and Thr252/Gly248/Asp251 (Figure 11). The hydrogen bond between Thr252 and Gly248 gives rise to a kink in helix I (this is visible in the rabbit cytochrome CYP2B4 structure represented Figure 1).

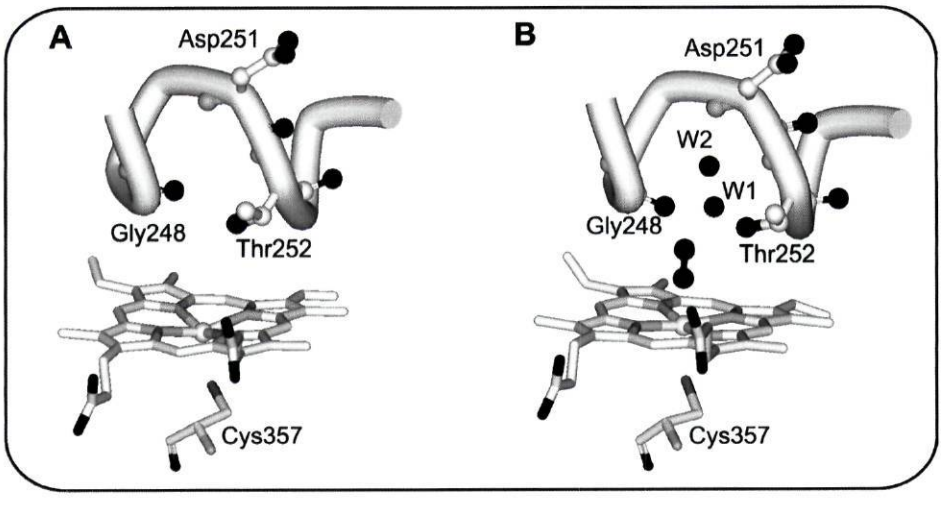


Figure 11. Ferrous pentacoordinated (a) and oxy (b) complexes of CYPcam (PDB codes 1DZ6 and 1DZ8, respectively).

The second electron transfer occurs, which is the rate limiting step in many CYPs [38], usually indistinguishable from following proton transfer step [39]. However, in a D251N mutant of CYPcam, the proton transfer becomes the limiting step, and reduced oxy complex intermediate in now detected [40]. Also Thr252 has been shown to be essential for effective proton transfer (a T252A mutant of CYPcam [41] shows an increased rate of hydrogen peroxide formation, the so-called uncoupling reaction). A proton transfer mechanism was proposed: solvent accessible residues Asp182/Lys178/Arg186 would provide protons to the close lying Asp251, which would act as a carboxylate switch delivering them to Thr252 [39;42]. Water molecules could serve as intermediates in this last step, as shown by kinetic

isotope effect studies [39], and could explain the ordered water molecules seen in the oxy crystallographic structure (W1 and W2 in Figure 11).

The first proton transfer originates an hydroperoxo complex, and the O-O bond is cleaved after a second proton transfer, with the formation of a water molecule and the incorporation of one oxygen atom in the substrate [43].

Irreversible inhibitors use the reaction cycle to form reactive intermediates that establish covalent bonds with the enzyme. One example is the hCYP1A2 specific inhibitor furafylline [44-46], used widely in pharmacological research to detect the activity of that enzyme against developing drugs. Figure 12 shows the proposed mechanism for this inhibitor.

1.1.3. Gating the cycle

The cysteinate ligand is important in controlling the redox chemistry of the powerful electron acceptor Fe(III) complex. It turns it into a poorer electron acceptor than a heme devoid of proximal ligand [47], preventing pointless and nonspecific electron transfer, but induces a quicker electron transfer than a histidine ligand would (as shown for a CYPcam C357H mutant [48]), avoiding uncoupling reactions. It also has a key role in the heterolytic cleavage of the O-O bond [47;48].

What seems to be another checkpoint in the cycle is the binding of the substrate. Initially it was thought that substrate binding was coupled with the first electron transfer by inducing a positive shift in the potential. This conclusion arose from the first photochemical measurements of redox potentials of CYPcam [49]. The resting state had an $E^{o'}$ of -340 mV and after substrate binding there was a positive shift to -173 mV allowing CYPcam to receive an electron from its redox partner, putidaredoxin, which exhibited an $E^{o'}$ of -196 mV when bound to the enzyme. Similar results have been obtained recently by cyclic voltametry (a positive shift of 136 mV between the free and camphor-bounds CYPcam) [50]. Various electrochemistry approaches have been used to measure CYPs redox processes and the E^{o} values have been shown to vary between enzymes and depend greatly on the conditions of the experiment, electrode type, mediator and oxidizing/reducing agent used [51;52]. As an example, when measuring the redox potential using CYPcam adsorbed on the surface of clay-modified electrode, the substrate binding did not interfere with the measured redox potential [53]. Curiously, quite recently, the heme redox potential of the bacterial CYPcin was shown to be unaffected by substrate binding [54]. Also for the bacterial CYPbm3, a recent study contradicted previous redox potential measurements and reported that no shift occurred in the present of different substrates [55]. It seems that what was initially thought to be the coupling

of substrate binding with the first electron transfer to avoid wasting electron cycling was a particular case in the vastness of CYPs present in Nature.

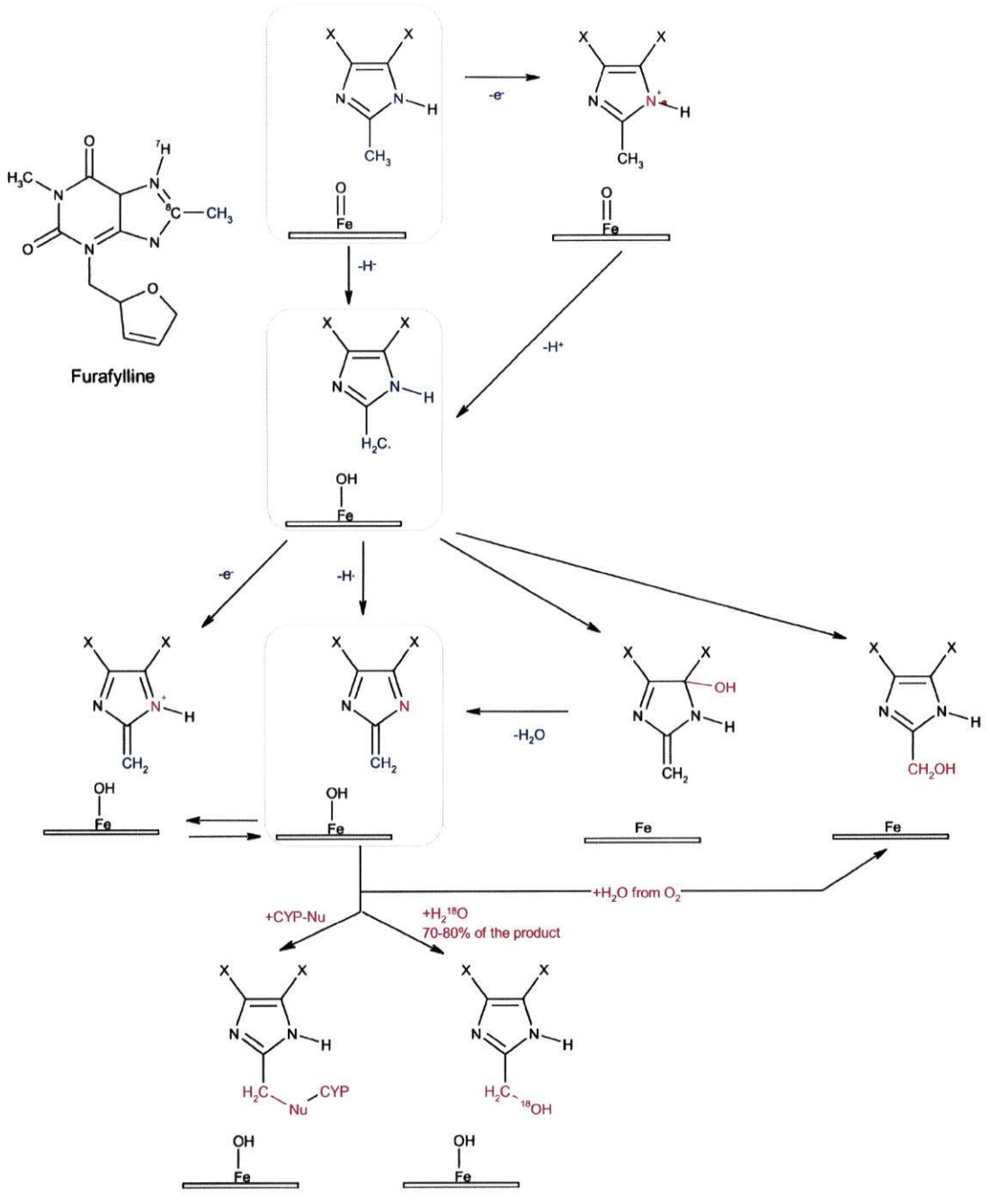


Figure 12. Mechanistic proposal for the irreversible inhibition of hCYP1A2 by furafylline [46].

Given the variety of CYPs present in each organism, and knowing that they all receive electrons from the same source, one can speculate that:

- the modulation of CYPs redox potentials may also be associated with the enzyme's importance in the cell;

- the protein-protein interactions between CYP and CPR may also be involved in the triggering of CYPs cycle (e.g. putidaredoxin's redox potential changes as it binds CYPcam).

1.2. Molecular Modelling

1.2.1. Molecular Mechanics

Potential energy is associated with the configuration (or arrangement) of a system. Molecular potential energy functions quantify the energy of a molecule or set of molecules in different conformations. Molecular Mechanics (MM) potential energy functions assume various approximations regarding the molecular systems, namely merging nuclei and electrons in atom-like particles. This means they are assumed to be spherical balls and the bonds between them are viewed as springs. These approximations allow the faster calculation of molecular properties when compared to electronic structure methods (described ahead), albeit with a limited molecular description, being unable to handle bond-breaking/making reactions. MM is very useful for exploring the geometry and relative energies of conformers of the same molecule, testing the effect of substituents on geometry and strain energy, to study the docking of substrates into active sites, to refine X-ray structures and to determine structures from NMR data.

Individual potential energy functions are used to describe bonded and nonbonded interactions between these particles and rely on both empirically derived and computationally calculated parameters. The molecular potential energy results from the sum of these functions, and can be generally represented as:

$$V = V_{bond} + V_{angle} + V_{dihe} + V_{imp} + V_{vdW} + V_{elec} \qquad \textbf{Eq. 1}$$

The bonding terms consist of bond stretching (V_{bond}) and angular distortions $(V_{angle/dihe/imp})$ (bond angle bending, dihedral torsional terms, and, sometimes, inversion terms). The nonbonding terms are the van der Waals long-range attraction and short-range repulsion between two electron densities (V_{vdW}), and the electrostatic term describing the partial ionic character of polar covalent atoms (V_{elec}). Additional terms, such as cross terms, can be found in some force fields to account for small variations in structure by describing the coupling between different internal degrees of freedom (e.g. when the H-O-H angle is squeezed in the water molecule the H-O bonds stretch slightly due to repulsion between the hydrogens). The potential energy functions and the parameters used for evaluating interactions are termed a force field. It should be pointed out that MM energies have no meaning as absolute quantities as they represent the sum of the interactions between atom-like particles and should only be used for comparison between molecular systems [56;57].

It is important to keep in mind that each force field is tested for accuracy within the range of molecules it intends to use, as the parameters are developed to fit these particular molecules. Within a force field, the same group of bonded atoms in different molecules is expected to be well described by the same parameters. This transferability of parameters between analogous groups of atoms within a force field is no longer valid between different force fields. Each force field should be regarded as a single entity and one should choose the force field that was most adequately developed to study the intended molecular system.

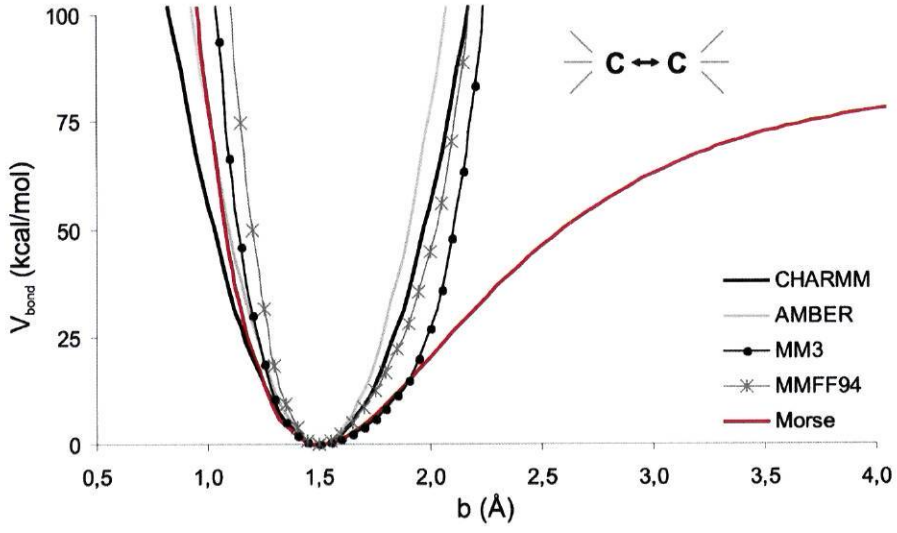


Figure 13. Potential energy associated with the stretching of the bond between two aliphatic carbons calculated by different force fields.

The available force fields can be generally divided into two major groups, depending on the complexity of their functional form and the number of parameters considered in the calculation. Some force fields, like CHARMM22 [58] and AMBER [59], are intended for simulations of bulk phases, macromolecules like proteins and nucleic acids. These generally have a simple form with harmonic terms for bond stretching and bending and usually a Lennard-Jones term for the Van der Waals interaction. On the other had, force fields like MM3 [60] and MM4 [61], widely used in organic chemistry, are designed to accurately determine structures and vibrational frequencies, and use high order polynomials to calculate bond stretches and bends and a Hill potential (or occasionally a Morse potential) to describe the Van der Waals interaction. Also, these include many cross terms for improved agreement with experiment (MM4 has been added more of these to improve accuracy relatively to MM3). Accuracy is thus obtained through a tedious parameterization procedure and higher computational time usage than the previously presented force fields. MMFF94 force field [62] shares the methodological approach with MM3 and MM4 but intends also to achieve good

results with condensed-phase systems in molecular dynamics simulations, having been developed to handle chemical systems of interest to both organic and medicinal chemists. One particular characteristic of this force field is the fact that the core portion of its parameters has been derived from high-quality computational data. Figure 13 shows the bond stretching potential energy between two aliphatic carbons described by various force fields. It is interesting to see that all of them represent fairly well the equilibrium geometry related to the minimum of energy of the system.

In this work, the CHARMM22 [58] force field within the CHARMM program [63] was chosen to study the Cytochrome P450 enzyme. This force field has been designed to handle small molecules besides macromolecules. The parameterization has been made by using both structural data and *ab initio* calculations. Like in the general equation presented above, the total potential energy is calculated as a sum of internal (or bonded) terms, which describe the bonds, angles and bond rotations in a molecule, and a sum of external (nonbonded) terms, which account for interactions between nonbonded atoms or atoms separated by three or more covalent bonds (Coulombic and van der Waals interactions). The energy is a function of the atomic positions of all the atoms in the system (usually expressed in terms of Cartesian coordinates).

$$E_{CHARMM22} = \sum_{bonds} k_b (b - b_0)^2 + \sum_{UB} k_{UB} (S - S_0)^2 + \sum_{angle} k_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} k_\chi (1 + \cos(n\chi - \delta)) + \sum_{impropers} k_{imp} (\varphi - \varphi_0)^2 + \sum_{i<j} e_{ij} \left[\left(\frac{R_{\min_{ij}}}{r_{ij}} \right)^{12} - \left(\frac{R_{\min_{ij}}}{r_{ij}} \right)^{6} \right] + \sum_{i<j} \frac{q_i q_j}{\varepsilon r_{ij}} \qquad \textbf{Eq. 2}$$

k_b, k_{UB}, k_θ, k_χ and k_{imp} are the bond, Urey-Bradley, angle, dihedral angle, and improper dihedral angle force constants, respectively. b, S, θ, χ and φ are the bond length, Urey-Bradley 1,3-distance, bond angle, dihedral angle, and improper torsion angle, respectively, with the subscript zero representing the reference values for the individual terms (Figure 14). n is the multiplicity (the number of minimum points in the function) and γ is the phase factor (determines where the torsion angle reaches its minimum value). The last two terms represent the Lennard-Jones 6-12 potential and the Coulomb interaction; ε_{ij} is the well depth and $R_{\min ij}$ is the distance at the Lennard-Jones minimum, q_i is the partial atomic charge on atom i, ε_1 is the effective dielectric constant, and r_{ij} is the distance between atoms i and j. The expression used for bond length is a harmonic approximation of the real system (better

described by a Morse potential) and only has physical meaning when we deal with values close to equilibrium (Figure 13).

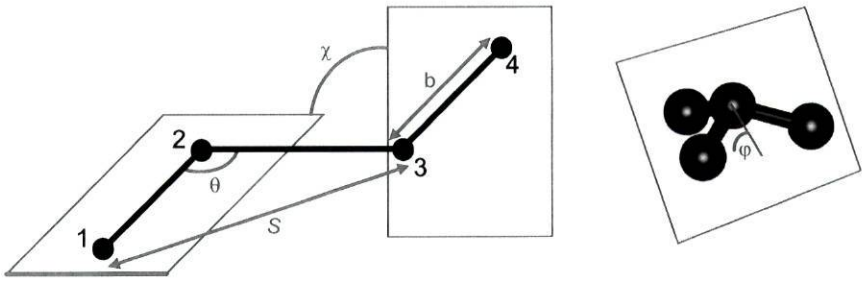


Figure 14. Molecular geometrical features generally considered by molecular mechanics force fields.

For large displacements from equilibrium the anharmonic effects become important. For shorter distances the repulsive interaction are dominant and the potential energy should rise sharply than in the harmonic model. For large distances, the potential energy should flatten out due to bond breaking. Also the bond angle energy is described by a harmonic potential, and is only realistic for small displacements from equilibrium. In order to improve the behaviour of the force field in relation to angle bending, a cross-term reflecting the interaction between 1,3 atoms is included (Urey-Bradley term). As far as the dihedral angle potential is concerned, one should notice that k_χ is not equal to the barrier height involved in the torsion, as there is a significant contribution from non-bonded interactions between the 1,4 atoms. CHARMM22 also considers improper torsion potential terms to guarantee, *e.g.*, the maximization of π-bonding energy, such as in phenyl rings.

As mentioned before, atoms can be regarded as spheres. Each possesses a particular radius called van der Waals radius. Two nonbonded atoms will not approach each other closer than the sum of their van der Waals radius (van der Waals repulsion) but they want to be close to each other (van der Waals attraction, a short-range effect). The way energy changes with the distance between two nonbonded atoms can be described by a Lennard-Jones 6-12 potential [57]:

- the energy climbs rapidly (repulsive part that varies as r_{ij}^{-12}) when the distance between the two atoms becomes shorter than the sum of their van der Waals radius;
- the energy increases slower (towards zero) when they become separated beyond the optimal distance (where the energy is $(-\varepsilon_{ij})$).

1.2.2. Quantum Mechanics

Our idea of an atomic system has changed a lot since the beginning of this century. Earlier models presented the atom as composed of moving particles, the electrons, which described fixed orbits around another particle, called the nucleus. However, the properties of atoms were not correctly described by this model. They have wave-like characteristics besides the particle-like ones – like standing waves, they are also a quantized phenomenon. Quantum mechanics developed the idea of electrons moving in random paths around the nucleus in a sort of a cloud. This region of space where the electron is most likely to be found has been called 'orbital'.

In this thesis, the computational chemistry methodology is used to describe stationary molecular states. The energy of such systems is calculated using several approximations when solving the non-relativistic time-independent Schrödinger's equation:

$$H\Psi = E\Psi \qquad \textbf{Eq. 3}$$

where Ψ is a state function dependent on the nuclear and electronic positions, H is the Hamiltonian operator and E represents the total energy of the system. The Hamiltonian for a system of k particles is:

$$H = T + V = \left[-\frac{\hbar^2}{2} \sum_k \frac{1}{m_k} \nabla_k^2 \right] + \left[\frac{1}{4\pi\varepsilon_0} \sum_j \sum_{k<j} \frac{q_j q_k}{r_{jk}} \right] \qquad \textbf{Eq. 4}$$

where T is the kinetic energy operator (derived from de Broglie's wave description), V is the potential energy operator (the same as in classic physics), $\hbar$ is Planck's constant divided by 2π, $4\pi\varepsilon_0$ is the vacuum permittivity, q_a is the charge on particle a, r_{ab} is the distance between particles a and b and ∇_k^2 is the Laplacian operator of particle k:

$$\nabla_k^2 = \frac{\partial}{\partial x_k^2} + \frac{\partial}{\partial y_k^2} + \frac{\partial}{\partial z_k^2} \qquad \textbf{Eq. 5}$$

In the particular case of a molecular system, the motion of nuclei can be considered as negligible when compared to that of the electrons, which have a much smaller mass than the former. This is Born-Oppenheimer's approximation, which results in a Schrödinger equation for many-electron molecules as the sum between individual electrons kinetic energies and electronic and nuclear potential terms [64;65]. What we in fact will be calculating for real molecular systems are the electronic energies for fixed nuclear positions and the correspondent Hamiltonian will be:

$$H_{el} = -\sum_i \frac{\hbar^2}{2m_e} \nabla_i^2 - \sum_i \sum_k \frac{e^2 Z_k}{4\pi\varepsilon_0 r_{ik}} + \sum_{i<j} \frac{e^2}{4\pi\varepsilon_0 r_{ij}} \qquad \textbf{Eq. 6}$$

where i and j run over electrons, k and l run over nuclei, m_e is the mass of the electron, m_k is the mass of the nucleus, e is the charge of the electron, Z_k is the atomic number of nuclei k and r_{ab} is the distance between particules a and b. The electronic Schrödinger equation will thus be:

$$H_{el}\Psi_{el}(q_i) = E_{el}\Psi_{el}(q_i) \qquad \textbf{Eq. 7}$$

where H_{el} includes the electronic terms from Equation 7 and E_{el} is the electronic energy and Ψ_{el} is the polielectronic state function. The electronic coordinates q_i are independent variables but the nuclear coordinates q_k are parameters. Ψ_{el} has no exact physical meaning but by integrating $|\Psi|^2$ over a certain region of space, one obtains the 'probability density', which corresponds to the probability of finding an electron in that region. The total energy of a molecular system with nuclear positions fixed is then calculated as the sum of the electronic energy and the nuclear repulsion:

$$E_{tot} = E_{el} + \sum_{k<l} \frac{e^2 Z_k Z_l}{4\pi\varepsilon_0 r_{kl}} \qquad \textbf{Eq. 8}$$

Computational methods intend not only to calculate the energy for a given configuration (in this case, E_{el}), but also to predict the geometry of that of the lowest energy. The energy values can be calculated by solving the Schrödinger equation, which implicates finding the appropriate wave function.

The electronic Schrödinger equation can only be analytically solved for monoelectronic systems. In order to deal with chemical and biological polielectronic systems additional simplifications have to be assumed. In this context, the Hartree-Fock (HF) method is one of the simplest approaches [57;65]. According to this formalism, the polielectronic state function $\psi(q_i)$ is expressed usually as a combination of an appropriate guess of molecular orbitals (ϕ_i). This combination should satisfy the antisymmetry principle, which states that the state function changes signal with the exchange of any pair of electrons preserving the same electronic density distribution. Consequently, a specific molecular orbital can describe no more than two electrons with opposite spins (Pauli exclusion principle). Also, HF accounts for electron exchange which results in the reduction of the Coulombic repulsion energy between two electrons bearing the same spin – which have a low probability of being close to each other.

As the exact form of the orbitals is unknown, a guess can be prepared. According to the Variational Principle, the expectation value of the energy based on the choice of an

appropriate Ψ will always be higher (or equal) than the exact energy of the system. This means that when we are looking at different possible wave functions in order to define the ground state of a system, the best will be that which has the lowest energy associated. The HF method uses this theory to search for the best approximation to the real wave function.

In HF, electron correlation, which makes the solving of Schrödinger's equation analytically impossible, is treated in a simplified manner. The HF method considers that all electrons except one are forming a cloud of electric charge through which that electron moves. The procedure starts with guess wave functions for all occupied molecular orbitals (MOs) which are then used to construct one-electron operators. HF equations are solved and the new calculated orbitals are used instead of the initial guess. This iterative procedure continues until the difference between the newly determined orbitals and the preceding ones is below a pre-determined convergence criterium.

To enhance the mathematical feasibility of HF computations, the Linear Combination of Atomic Orbitals (LCAO) formalism was introduced, in which molecular orbitals are treated as combinations of sets N atomic orbitals (AOs) (φ_k), the so-called 'basis sets' [66]:

$$\phi_i = \sum_{K=1}^{N} a_{K_i} \varphi_K \qquad \textbf{Eq. 9}$$

Consequently, the iterative process is simplified. In fact, only the coefficients (a_{k_i}) are now variationally optimized. The computation of these coefficients was permitted by the matrix algebraic equations developed by Roothaan. These atomic functions can be characterized by three quantum numbers n ($n = 1,2,3,...$) or principal quantum number, $l (l = n-1)$ or azimuthal quantum number, and m ($M = -l,-(l-1),...,0,...,(l-1),l$) or magnetic quantum number.

The mathematical functions more commonly used to mimic AOs are the Slater-type orbitals (STOs) and the gaussian type orbitals (GTOs), and are called 'basis functions'. STOs are functions that reproduce hydrogen-atom type orbitals and were initially used to describe molecular properties. However, they are computationally impractical when several electrons are being considered. GTOs, although less accurate in describing the chemical system, are mathematically simpler to use and computationally much faster than STOs. Most conveniently, linear combinations ('contractions') of GTOs were developed to reproduce STOs as accurately as possible, keeping a high computational performance. The GTOs used in the contraction are called 'primitives' and each linear combination of GTOs is a 'contracted basis function'. When using contracted GTOs to model STOs the functions are designated as STO-*x*G, where *x* represents the number of GTOs using in the contraction. In STO-3G, three

GTOs are combined to describe each atomic-like orbital. Such a basis set is called 'minimal' or 'single-ζ' basis set, as there is one and one only contracted basis function defining each type of orbital from core to valence. Although a contracted GTO might give a good approximation to an atomic orbital, it lacks any flexibility to expand or shrink in the presence of other atoms in a molecule. The solution to this problem is to add extra basis functions beyond the minimum number required to describe each atom.

This means that instead of having a basis function as a combination of three GTOs, we can use two basis functions for each AO – e.g. a GTO by itself and the result of a contraction of the other two. This would be called a 'double-ζ' basis set. Depending on the number of basis functions we intend to use, we can construct multiple-ζ basis sets with increasing quality in the description of the system. From a chemical point of view, this type of flexibility will be particularly significant when applied to valence orbitals which are directly involved in chemical bonding phenomena. To account for this, 'split-valence' or 'valence-multiple-ζ' basis sets were developed, where core orbitals are represented by a contracted basis function, while valence orbitals are split into many functions. 6-31G is a commonly used split-valence basis set – the core orbitals are described by a GTO that results from the contraction of 6 primitives, and the valence orbitals are described by 4 GTOs, three of which are contracted and one is used independently.

To improve the description on molecular orbitals, it is necessary to use functions other than those centered on the individual atoms, like *s* and *p*. This is achieved by adding basis functions that correspond to one quantum number of higher angular momentum than the valence orbitals, such as a *d* GTO for a first row element. The notation 6-31G* indicates that a set of *d* functions will be used to polarize *p* functions. This can also be written as 6-31G(d). The notations 6-31G** or 6-31G(d,p) indicate that polarization was also added to the hydrogen atoms.

Besides this, diffuse functions can be used to better describe weakly bound electrons that localize far from the remaining density, such as in the case of anions or highly excited electronic states. These functions are indicated by a '+' in the basis set name: in 6-31++G the first plus indicates the presence of diffuse functions in heavy atoms and the second indicates their presence in hydrogen.

The choice of the basis set used in a calculation must be pragmatic. The more orbitals we use, the better the description of the molecular orbital space, with accompanying growing computational time. Also, we should use functions that have large amplitude in regions of

space where the electron probability density is large, and small amplitudes where the electron probability density is small, using them wisely in a chemical sense.

When it comes to molecular systems that include very heavy elements, the number of basis functions needed to describe all the electrons is impractical. Mostly core electrons, they can be well characterized with analytical functions that represent the combined nuclear-electronic core to the remaining electrons – effective core potentials (ECPs) [66]. They include the Coulombic repulsion effects and the relativistic effects found in core electrons from very heavy elements (such effects could not be treated with the above considered non-relativistic Hamiltonian operator). An important decision when choosing an ECP is choosing how many electrons we want to include in the core. Large-core ECPs include everything but the outermost shell, while small-core ECPs scale back to the next lower shell. For metals, given the fact that polarization of the sub-valence shell can be chemically important, the small-core ECPs should be chosen.

An example of small-core ECP is the Stuttgart/Dresden's [67] which is implemented in Gaussian03 [68] (invoked by the SDD keyword). This has been used, e.g., for the quantum chemical treatment of the iron atom of the porphyrin ring using the B3LYP formalism [69]. In such case, it represents all electrons correspondent to a neon configuration as the nuclear-electronic core. The *s* valence orbitals are described by seven functions, one of them being a contraction of three, the *p* valence orbitals are described by four functions, one of them being a contraction of four, the *d* valence orbitals are described by four functions, one of them being a contraction of four, and there is one function describing *f* orbitals.

The applicability of HF theory is limited both from chemical and practical points of view. In a chemical sense, the oversimplified treatment of electron correlation is a rather serious approximation, and although HF yields very good bond lengths in molecules, the binding energies are in general not in good agreement with experimentally determined ones. Computationally speaking, HF scales to N^4 *(N = total number of basis functions)* – this will be the total number of rather complicated integrals that need to be solved – resulting in a daunting calculation.

The fastest computational methodology that accounts for electron correlation is the Density Functional Theory (DFT). This approach uses the physical observable electron density (ρ) which is univocally associated with the state function Ψ_{el}. This quantity, integrated over all space, gives the total number of electrons N:

$$N = \int \rho(r) dr \qquad \textbf{Eq. 10}$$

where $\rho(r)$ is the total electron density at a particular point in space r. DFT considers that there is a relationship between the total electronic energy and the overall electronic density. Hohenberg and Kohn demonstrated that the ground-state energy and other properties of a system are uniquely determined by the electron density [70]. In this way, instead of the $3N$ variables that are needed to describe the N electrons that exist in the molecular system, we are now dealing with only 3 variables that describe the electron density. We say that the electronic energy is a functional of the density. However, the Hohenberg-Kohn theorem does not provide us with the form of the functional dependence of energy on the density. The purpose of DFT methods is to propose appropriate functionals that describe this relationship.

Kohn-Sham orbitals (ϕ') are used analogously to those in HF methods, and can be also expressed in terms of a set of basis functions. They should not, however, be confused with HF ones, as Kohn-Sham orbitals don't share the same physical meaning and were developed specifically to calculate ρ:

$$\rho = \sum_{i}^{M} \left|\phi_i'\right|^2 \qquad \textbf{Eq. 11}$$

Generally, the energy of a system according to DFT can be represented as:

$$E_{DFT} = E_{N-N} + E_{N-e} + E_{e-e} + T_e + E_x + E_c \qquad \textbf{Eq. 12}$$

The terms for nuclear-nuclear repulsion (E_{N-N}), nuclear-electron attraction (E_{N-e}) and the classical electron-electron Coulomb repulsion (E_{e-e}) are the same as those used in HF theory, and those for the kinetic energy of the electrons (T_e) and the non-classical electron-electron exchange energies (E_x) are different. E_c describes the correlated movement of electrons of different spin and was not accounted for in Hartree-Fock theory. The electron correlation energy is sometimes included in a single exchange-correlation functional (E_{xc}).

The DFT calculation is a self-consistent, iterative process, where at each stage the calculated set of orbitals is used to calculate the density. The process stops when the density and the exchange-correlation energy have converged to within some tolerance.

Consequently, inaccuracy in DFT calculations arises specially from the calculation of the exchange and correlation functionals. There are various approaches within DFT to calculate these terms. The 'Local Spin Density Approximation' (LSDA) [71] assumes that the exchange correlation energy value depends solely on the local value of the electron density. B88 [72] and LYP [73] are based on a different approach, the 'Generalized Gradient Approximation' (GGA), which accounts for the variation of electron density in space, with the exchange and correlation functionals being dependent also on that gradient.

These are the so-called 'pure' DFT methods. Another group of functionals, called 'hybrid', combines energy terms calculated through an HF approach with those from the DFT methodology to obtain the E_{xc} energy with a higher accuracy than the previous ones [66]. The inclusion of the HF exchange term in the hybrid functional results in a cancellation of errors that improves the energy barrier determination in chemical reactions (usually underestimated by a pure GGA functional and overestimated with HF) and the determination of the chemical bond lengths (usually underestimated with HF and overestimated with the pure DFT functionals) [66].

The most popular hybrid functional to date, and the one used in the work presented in this thesis, is B3LYP (Becke-3-parameter-Lee-Yang-Parr) [74]. This calculates the combined exchange-correlation term from a sum of HF exchange energy with several DFT terms, scaling the different parcels with three empirically determined constants (a = 0.20, b = 0.72 and c = 0.81):

$$E_{xc}^{B3LYP} = (1-a)E_x^{LSDA} + aE_x^{HF} + b\Delta E_x^{B88} + (1-c)E_c^{LSDA} + cE_c^{LYP} \qquad \textbf{Eq. 13}$$

Comparisons made with experimental results and other functionals have shown its extremely good performance [66]. However, unlike in the HF theory, where the use of a complete basis set will lead us to an energy value that is always superior to the real energy of the system because the electron correlation is not accounted for (HF limit), DTF is not such a straightforward variational approach, because the exchange and correlation functionals have an unknown form. Thus, it is not known whether the energy value is above or below the real one, and the SCF procedure leads us to the smaller value of energy that is possible to obtain with the chosen functional. Besides, the inclusion of empirical parameters makes it impossible for us to be aware of the absolute energy values, as these are shifted into a different scale. DFT calculations formally scale as the third power of the number of basis functions.

1.2.3. Solvation Models

When handling a chemical system using computational methods, it is possible to evaluate many of its properties in vacuum. Although this is not at all a realistic approach to the problem, it can sometimes be considered as a reasonable approximation. However, for a chemical system that includes functional groups that are highly polarizable or charged, simulating an aqueous environment or the usually more hydrophobic interior of a protein can make a difference. This is the case of protein-ligand binding processes, where the polar/charged ligand is initially in aqueous solution and after binding can become completely isolated from the solvent in a protein hydrophobic cavity. By adding the solvation effects, one

should be able to evaluate if when binding to the protein the ligand compensates for the lost stabilizing interactions with the solvent. A solvent can interact with a solute either through 'short-range' effects (hydrogen bonding or a preferential orientation of the solvent molecules near an ion) or 'long-range' effects (which generate a dielectric constant different from 1). The first are mainly concentrated in the first solvation sphere and can be modelled using explicit solvent molecules. The long-range effects are either modelled with a large number of solvent molecules or by treating the solvent as a continuum medium.

In MM approaches, solvent molecules can be parametrically described, such as the TIP3P [75] water molecules within the CHARMM force field.

As pointed out before, the solvent can also be modelled as a continuum, an uniform polarizable medium with a dielectric constant of ε (reaction field) with a solute placed in a cavity inside the medium. In this case, the solvation free energy (ΔG_{solv}) associated with the solute/solvent system can be represented as:

$$\Delta G_{solv} = \Delta G_{electrostatic} + \Delta G_{cavity} + \Delta G_{vanderWaals} \qquad \textbf{Eq. 14}$$

where $\Delta G_{electrostatic}$ and $\Delta G_{vanderWaals}$ result from the electrostatic and van der Waals (dispersion) interactions between the solvent and the solute, and ΔG_{cavity} is the free-energy required to form the solute cavity within the solvent.

$\Delta G_{vanderWaals}$ and ΔG_{cavity} terms, often referred to as the nonpolar component of ΔG_{solv}, can be calculated as follows [76]:

$$\Delta G_{nonpolar} = \gamma A + b \qquad \textbf{Eq. 15}$$

where A is the surface area traced out by spherical particle of a given radius rolling on the van der Waals surface (solvent accessible surface area, SASA) or calculated using the van der Waals radius, and γ and b are constants derived from experimentally determined free energies.

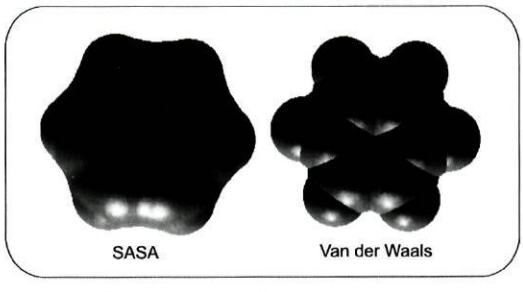


Figure 15. Solvent accessible and Van der Waals surfaces for benzene.

ΔG_{cavity} depends linearly on the cavity surface because the solvent molecules that have to be reorganized due to the presence of the solute will be roughly those in the first solvation shell. The number of such molecules will be approximately proportional to the cavity surface. The dependence is observed for van der Waals interactions, which quickly vanish with an increasing interaction distance.

The electrostatic component of ΔG_{solv} can be calculated using approaches based on *ab initio* methods or classical electrostatics.

In the computational approaches using quantum chemical methods, we used the Conductor-like Polarizable Continuum Model (C-PCM) [77] model of solvation implemented in Gaussian03. This method models the cavity using interlocking atomic spheres (van der Waals type cavity) and the cavity/dispersion contributions are derived using classical approaches based on the surface area. The electronic wave function of the solute will influence the computation of the reaction field. On the other hand, the solute's wave function will be influenced by the reaction field surrounding it. This is an iterative process, a Self-Consistent Reaction Field methodology. In the end, it can be used to calculate the properties of the solute in any solvent at the same level of theory as it is done *in vacuo*.

In the computational approaches using molecular mechanics methods, we used the program DelPhi [78] to calculate the electrostatic component of solvation. In this case, the solute is described classically, using the atomic charges from a standard force field, and as body of low dielectric constant. Delphi places the solute on a grid, and allocates the atomic charges to the eight surrounding grid points. Then it defines a boundary between the solute and the solvent in order to assign values of dielectric constant to each grid point. By using either the van der Waals or the SASA surfaces it will assign the solute's dielectric constant to all the points inside and at the surface of the cavity. It then solves the Poisson-Boltzmann equation using a finite difference formula. The grid size will influence the results, which will be more accurate as the lattice becomes finer. To have good results with less computational expenses, one can use the focusing technique, which implies performing consecutive calculations with the system occupying a greater fraction of the total box. This will improve the estimates of the potential value at the boundary. It should be pointed out that a different orientation of the solute within the grid can influence the results. This means that calculations done for comparative purposes (such as the solvation contribution to the formation of intermolecular complexes) should attempt to keep the solute in the same reference position within the cubic grid.

1.2.4. Potential Energy Surface *vs* Energy Minimization

The potential energy is a multidimensional function of the coordinates of the atoms. The energy is a function of $3N$ (N = number of atoms) Cartesian coordinates:

$$E = f(x_1, y_1, z_1, x_2, y_2, z_2, ..., x_N, y_N, z_N) \qquad \textbf{Eq. 16}$$

Atoms can also be described by reference to other atoms in the molecule: one atom is set to be the origin of all atoms. Then the distance between this and a second atom is defined. A third atom is defined as being at a certain distance from the second atom and at a certain angle from the second and first atoms. A fourth atom will be at a certain distance from the third atom, at a certain angle from the third and second atom and at a certain torsion angle from the third, second, and first atoms. All subsequent atoms must be defined such that they are at a certain distance, angle and torsion angle from previously defined atoms. These are called internal coordinates and correspond to the vibrational components of the movement. In this way the translational (three coordinates relative to the movement of the molecule in space) and rotational (three coordinates relative to the rotation of the molecule) components are ignored and the total number of coordinates to be considered is $3N-5$ for linear molecules and $3N-6$ for molecules with more than 3 atoms). The number of coordinates considered corresponds to the degrees of freedom in the system.

In molecular modelling we are interested in looking at the lowest energy conformations for the molecules, as these correspond to stable states of the system (we should bear in mind, though, that the biologically active conformation may not correspond to the global minimum in energy (Figure 16)). However, it is not feasible to calculate all the points in the potential energy surface to obtain this value. What is done instead is to make small adjustments to a molecule's geometry until these lead to an increase in the energy, in a process called geometry optimization. It is important to start with a promising structure, especially if we are dealing with complex systems.

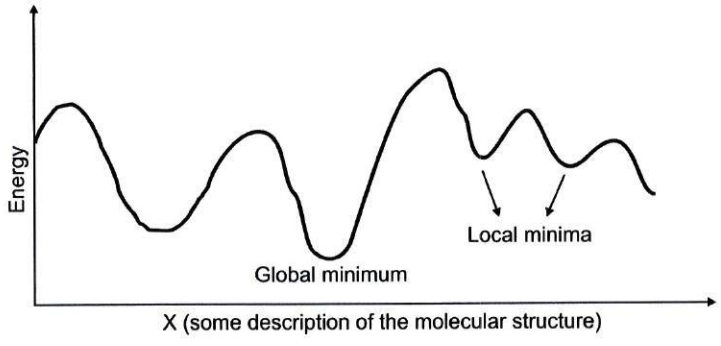


Figure 16. Example of a potencial energy surface.

From the many different minimization algorithms available, the derivative methods are generally very effective and can be found in all computational chemistry packages. Some examples are the Steepest Descent and the Conjugate Gradient.

These methods alter the atomic positions of the starting structure towards an optimal geometry by examining the effects of each modification on the total energy of the molecular system. They will accept the geometry if the first derivative of the energy function with respect to each of the variables (the gradient) is negative and will continue the process until the gradient is zero and the second derivatives are all positive, meaning that a minimum in the potential energy surface was reached (Figure 16; Figure 17). However, given the complexity of the potential energy surface for most molecular systems, one should be aware that this geometry may not correspond to the global minimum (Figure 17).

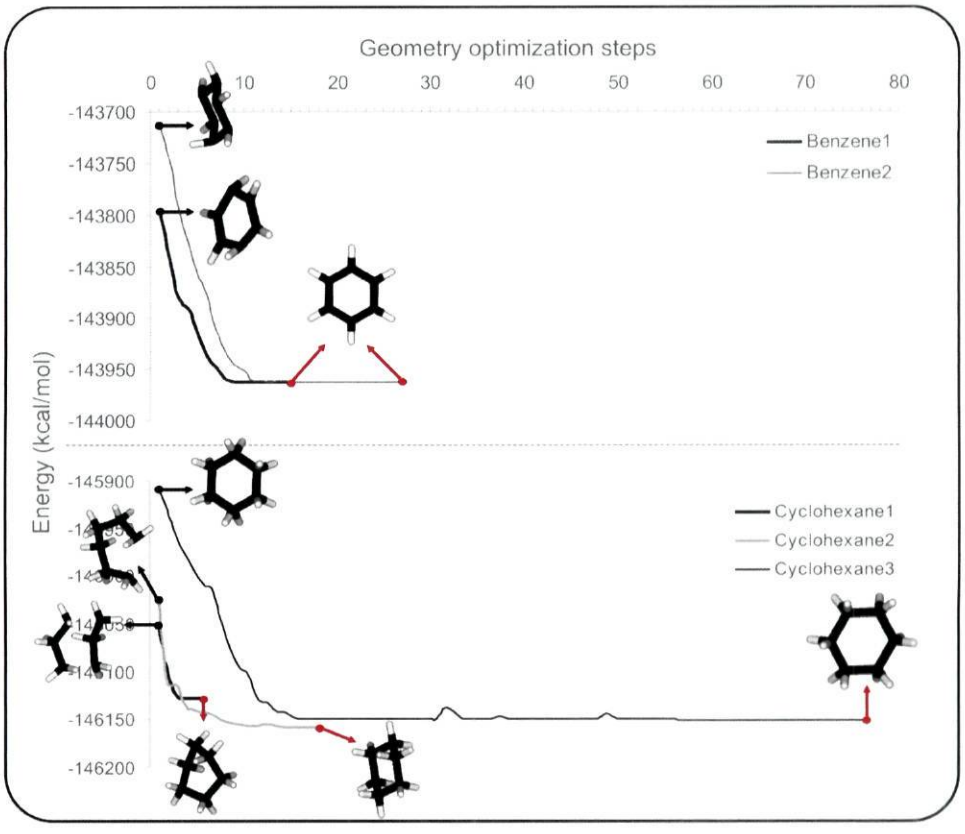


Figure 17. Geometry optimization of benzene and cyclohexane distorted structures using HF with a 3-21G basis function in Gaussian03. Cyclohexane has a higher number of geometrical degrees of freedom and that results in multiple minima in the potential energy surface. Starting geometries correspondent to different points in the surface lead to different minima.

1.2.5. Molecular Interactions

One of the most interesting challenges in molecular modelling is to understand how two molecules interact such as in enzyme/ligand and protein-protein recognition systems. These interactions can be described qualitatively and quantitatively. Qualitative descriptions explore

the geometry and the electrostatic pattern of the interacting molecules. Geometrical characteristics, such as shape, volume and surface, are very important for stereochemical complementarity reasons - molecules with similar geometrical features are likely to interact with the same receptors. Protein-ligand interactions are also known to be dependent on electrostatics, with the complementarity of the molecular electrostatic potential surfaces of both molecules being used as an indication of a potentially favourable interaction. Quantitative descriptions involve the calculation of a binding energy after determining the geometry of the complex formed by the two molecules.

Molecular recognition processes are usually involved in the approach of the ligands to the active site entrance and its subsequent binding. In this context, the electrostatic pattern recognition has been demonstrated to have a crucial role [79-81]. The molecular electrostatic potential can provide us with the electrostatic pattern that might be favored in molecular recognition processes and also with the potential sites for H-bonding and other noncovalent interactions formation, which could be very important for a correct orientation of the inhibitor inside the enzyme, with consequences in all mentioned steps [80;82;83]. The electrostatic potential V created by the nuclei and electrons on point r is:

$$V(r) = \sum_A \frac{Z_A}{|R_A - r|} - \int \frac{\rho(r')dr'}{|r'-r|} \qquad \textbf{Eq. 17}$$

where Z_A is the charge on nucleus A, located at R_A, and $\rho(r)$ is the electronic density function. The local minima in the potential surface correspond to areas which are susceptible of electrophilic attack (in blue in Figure 18) while the regions predisposed to nucleophilic interactions (in red in Figure 18) are only recognizable when displayed at a certain distance from the nucleus (the highest positive peak of electrostatic potential in the molecule), which is why the MEP is usually mapped onto a molecular surface.

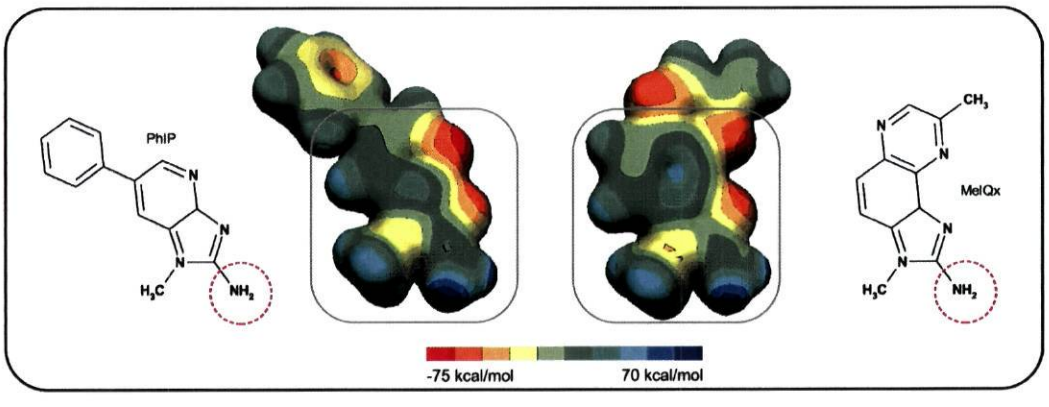


Figure 18. PhIP and MeIQx MEPs mapped on a 0.01 e^-/bohr^3 electron density surface. Red circles indicate the amine group oxidated by hCYP1A2 (see Figure 7).

It is possible to compare the affinities of different ligands towards the same enzyme by calculating appropriate thermodynamic quantities. One that is commonly used to describe such interaction is the binding free energy (ΔG_{bind}). It is possible to obtain a ΔG_{bind} using molecular mechanics and classical continuum solvation approximations. The ΔG_{bind}, corresponding to the association process of molecules A and B, can be defined as [84]:

$$\Delta G_{bind} = G_{aq}(A:B) - \left(G_{aq}(A) + G_{aq}(B)\right) \qquad \textbf{Eq. 18}$$

where, $G_{aq}(A)$ and $G_{aq}(B)$ correspond to the Gibbs free energies of molecules A and B and $G_{aq}(A:B)$ is the Gibbs free energy of the complex they form. Using a thermodynamic cycle such as that represented in Figure 19, we can calculate the Gibbs free energy of any generic species as [84]:

$$\Delta G_{aq} = \Delta G_{gas} + \Delta\Delta G_{solv} \qquad \textbf{Eq. 19}$$

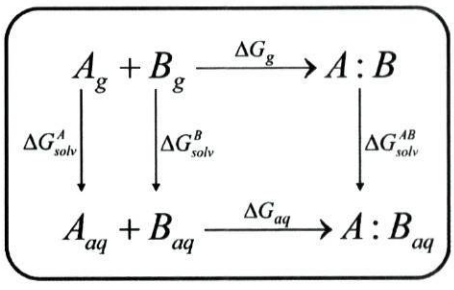


Figure 19. Thermodynamic cycle correspondent to the binding of species *A* and *B* in vaccum/solvent.

In equation [19] ΔG_{gas} is the binding free energy of the species in gas phase and $\Delta\Delta G_{solv}$ is its solvation free energy. ΔG_{gas} corresponds to the sum of the energy of the species in gas phase with the correspondent entropic contribution. The latter can be regarded as non-differential when comparing two complexes, as it refers to the process of association of similar ligands to the same protein [84].

In order to predict binding conformations one can use a docking approach. It is possible to carry out manual docking using all the known information regarding the products of the enzyme's metabolism or the binding modes seen in X-ray structures using computer graphics. Automated molecular docking explores the binding modes of two interacting molecules and calculates the energy of the resulting molecular complex. In the end, topographic features and energy-based considerations produce conformations where the interactions are the most favourable.

GOLD [64] is an example of an automated docking program. It uses a genetic algorithm to explore the conformational variability of a flexible ligand. The obtained binding modes are

scored with one of two available scoring functions: GoldScore and ChemScore. Both of these functions are based on empirically derived geometric parameters.

GoldScore fitness function evaluates the hydrogen bond and van der Waals energies resulting from the protein-ligand interaction, and the ligand's internal and torsional strain energies. The values for the terms included in the scoring function are empirical parameters. ChemScore calculates ΔG_{bind} as a sum of terms that are the product of a scale factor determined by regression and the magnitude of a particular physical contribution (hydrogen bonding, lipophilic atoms interaction, rotatable bonds freezing terms, clash penalties, ligand internal torsional strains, covalent bond formation and user-introduced constrains terms).

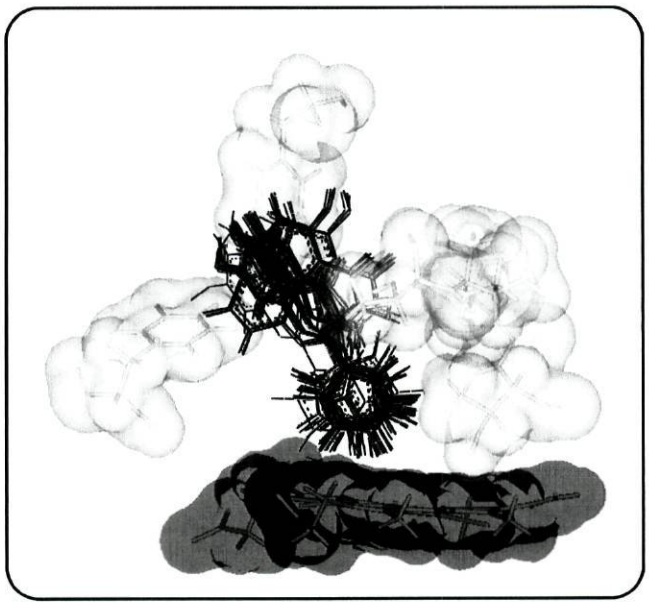


Figure 20. Example of GOLD output: the five best scored docked conformations on the active site of the human CYP1A2 model of flavone derivatives (heme and active site representative residues involved in a black/white van der Waals surface, respectively).

1.2.6. Protein Homology Modelling

In order to explore a protein's function using computational methods, it is necessary to have access to a three-dimensional structure of the molecule. In case there is none, two possibilities arise:

- Predicting the structure with energy-based calculations using the amino acid sequence;
- Modifying a closely related (homologous sequence), functionally analogous molecule whose three-dimensional structure is known.

The latter is the fastest and the most pragmatic way of obtaining a protein three-dimensional structure. It requires the existence of X-ray or NMR structures of homologous proteins. In order to obtain these, one can check databases that have protein structure data, using either

keywords or the amino acid sequence of the protein, in e.g. one of the following search engines:

- Protein Data Bank, "the single worldwide repository for the processing and distribution of three-dimensional biological macromolecular structure data" (http://www.rcsb.org/pdb/)
- Blastp from the BLAST engine (http://www.ncbi.nlm.nih.gov/blast/)
- 3D-PSSM, a "protein fold recognition engine using 1D and 3D sequence profiles coupled with secondary structure and solvation potential information" (http://www.sbg.bio.ic.ac.uk/~3dpssm/index2.html).

After retrieving sequences/structures of our protein's homologues, it is necessary to perform a sequence alignment. This can be done using automated approaches (e.g. ClustalX [85]) or manually. As proteins diverge, they become different with the occurrence of substitutions, insertions and deletions. When performing an alignment, it is necessary to introduce gaps to maximize the number of identical matches. However, one must consider that it might be biologically meaningful to align nonidentical residues. Scoring matrices that weigh matches between nonidentical residues based on evolutionary substitutions rates have been devised to address this need (e.g. Dayhoff mutation data matrix). Such tools increase the sensitivity of the alignment, especially in a situation where sequence identity is low.

When multiple structural templates are available, it is also important to identify similar substructures, regions that fit well. For that, we calculate the root-mean-square deviation/distance (RMSD) between corresponding atoms in two superimposed structures:

$$RMSD = \left[\sum_{i}^{N} (| u_i - v_i |)^2 \right]^{\frac{1}{2}} \qquad \textbf{Eq. 20}$$

(square root of the sum of the squares of the distances between corresponding atoms). u_i and v_i are the corresponding vector distances of the *i*th atom in the two structures containing N atoms. This is a useful measure of how similar the structures are. The result is a measure of how each atom in the structure deviates from each other, and a RMSD value of 0-3 Å signifies strong structural similarity.

Related proteins tend to retain similar folding patterns, a common core. However, as the amino acid sequences diverge, the distortions increase in magnitude. Therefore, it is useful to use the core structural elements alone to make a good fit of the structures. The superposition should be made using the main-chain atoms (N, Cα, C, O) of the correspondent amino acids in each structure. After finding the common fold substructures, one should go back to

improving the sequence alignment in order to maximize the superposition of these elements, since, in general, the three-dimensional structures are more likely to be conserved than are the corresponding amino acid sequences between distantly related proteins.

The prediction of protein structures using the information retrieved from the sequence and three-dimensional structures of homologues can be performed using two approaches:

- transferring the backbone conformation of the protein from a single template to the unknown protein – piecing together rigid bodies taken from the template protein;
- constructing a framework by averaging the structures from a number of protein templates;
- automated homology modelling using spatial restraints

The last method is implemented in the program Modeller [86]. Its output is a three-dimensional model containing all mainchain and sidechain non-hydrogen atoms. The side-chains conformations chosen come from libraries containing the most common rotamers present in high-resolution X-ray structures. Modeller calculates spatial restraints from:

- the alignment (distance and dihedral angles);
- statistical analysis of the relationships between various features of the protein structure (distances between alpha-carbons, residue solvent accessibilities or side-chain torsion angles); the form of these restraints was obtained from a statistical analysis of the relationships between many pairs of homologous structures (105 family alignments that included 416 proteins with known three-dimensional structure);
- CHARMM energy terms.

The restraints are expressed as conditional probability density functions, each of which is a smooth function which gives the distribution of the features as a function of the related variables. For example, probabilities for different values of the mainchain dihedral angles are calculated from:

- sequence similarity between the two proteins;
- mainchain conformation of an equivalent residue in a related protein;
- residue type.

Subsequently, the individual probability density functions are combined to give a function which is optimised using a combination of conjugate gradients and molecular dynamics with simulated annealing.

Modeller can provide multiple conformations for any loops that must be built in the structure. Loop conformations may also be obtained by searching the protein databank for stretches of

polypeptide chain that contain the appropriate number of amino acids and also have the correct spatial relationship between the two ends. SwissPDBViewer [87] (http://www.expasy.org/spdbv/) provides an interface with a loop database that can be used in standalone modelling when the results of the automated procedure are not satisfactory. It is also possible to use an *ab initio* approach to predict fold, using an energy function to judge the quality of the loop.

After determining the coordinates for the model this must go through a quality check. Problems may arise from low identity with the templates or from errors that the later may carry. There might also be deviations in geometrical features such as bond lengths, bond and torsion angles.

For proteins in particular, a Ramachandran plot is a useful stereochemical quality check. It plots the ψ main-chain torsion angles versus the ϕ main-chain torsion angles for every amino acid residue in the protein with the exception of the N-terminal residue (which has no ϕ) and the C-terminal residue (which has no ψ) (Figure 21). It allows an easy detection of which amino acids present deviations to the geometric parameters most commonly observed. PROCHECK [88] includes all stereochemical analyses already mentioned, plus the indication of bad contacts between nonbonded atoms and deviations to planar geometries.

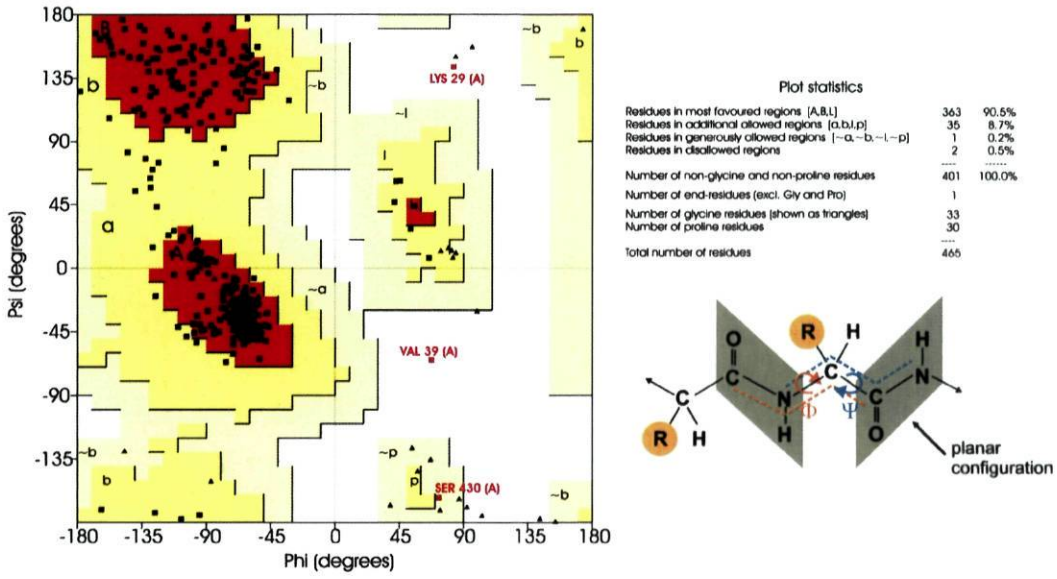


Figure 21. Ramachandran plot of the X-ray structure of rabbit cytochrome CYP2B4 (pdb code 1SUO) obtained with PROCHECK.

Another feature that should be examined in the model is the distribution of polar and apolar residues in order to avoid misfolded models. A PROFILE-3D analysis, such as that implemented in InsightII (http://www.accelrys.com), measures the compatibility of an amino

acid sequence with a three-dimensional protein structure. Each amino acid will be classified according to its exposure to solvent, polar and apolar protein environment, and the resulting score can be used to assess the fold quality or to compare evolutionary related proteins.

Model refinement following an automated modelling procedure demands a careful intervention of the human eye. When the protein alignment has ambiguously assigned indels, different models can be generated for different alignments. Either case should be followed by visual comparison of the target and the templates.

One should also identify and evaluate any specific stereochemical problems and eventually try to solve them using local energy minimization procedures. These should be short runs, as long minimizations tend to introduce errors in torsion angles, without any major benefits to the structure. Another way that proved efficient in improving stereochemistry was using the automated procedure in a self-consistent manner:

- building one or more models using one or more multiple alignments;
- choose a representative structure (which can result from the fusion of different models) and use it as a single template on the automated modelling program to build another model of the same structure.

Side-chains conformations provided by the automated procedure should also be examined, substituting rotamers in order to:

- avoid bad contacts;
- reproduce any conformations that should resemble those observed in the templates – e.g. such as those of residues involved in catalysis;
- adequately dispose polar side-chains present at the protein surface that might be curled into the bulk structure.

In the end, protein modelling can be seen as an iterative process that profits from a pragmatic approach of the human modeller to the automated procedures.

1.3. Molecular Evolution and Phylogenetics

Deoxyribonucleic acid (DNA) contains the information needed for a living organism to develop on sequence encoded in genes (some viruses depend on ribonucleic acid (RNA)). DNA and RNA are polymers of nucleotides (see Figure 22), each monomer containing:

- a sugar pentose (deoxyribose (DNA) and ribose (RNA));
- a nitrogen-containing base, either a purine (adenine or guanine) or a pyrimidine [thymine (DNA) / uracyl (RNA) or cytosine];
- one to three phosphate groups.

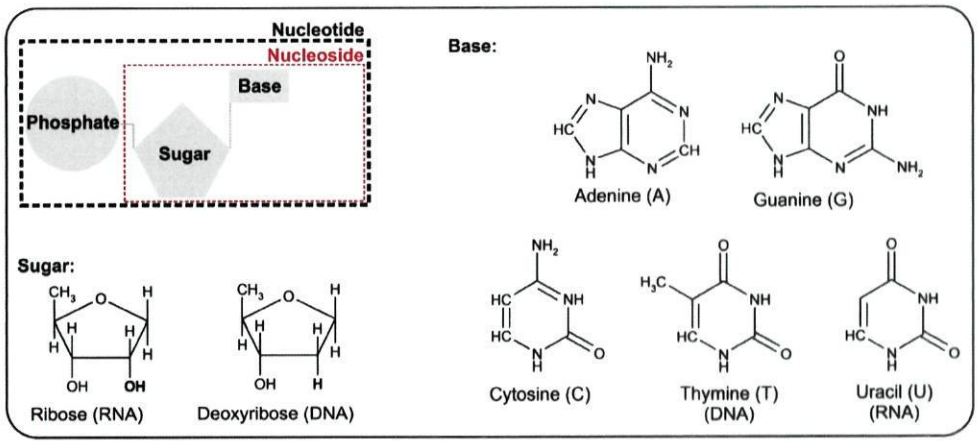


Figure 22. DNA and RNA nucleotide composition.

DNA is a double-stranded helix and the two chains are held together by hydrogen bonding between specifically paired bases: adenine pairs with thymine and guanine pairs with cytosine (Figure 23).

DNA is read by RNA polymerase from 3' to 5' and produces an mRNA (messenger RNA) transcript by adding nucleotides to the 3' end (Figure 24). The RNA transcript is thus antiparallel to the DNA template strand. The mRNA molecule in translated into a protein amino acid sequence according to the genetic code (see Table1).

A particular amino acid is coded by a sequence of three nucleotides called a codon. There are more codons than there are different amino acids in proteins, and some amino acids correspond to more than one codon. The genetic code is said to be redundant.

Evolution is based in the modification (nucleotide mutations) or the increase/decrease (indels) of nucleic acid sequences. By comparing DNA sequences it is possible to study the evolutionary relationships among organisms. In this thesis, we focused on how molecular evolution is interfering with the function of a particular protein (molecular adaptation; [89;90]).

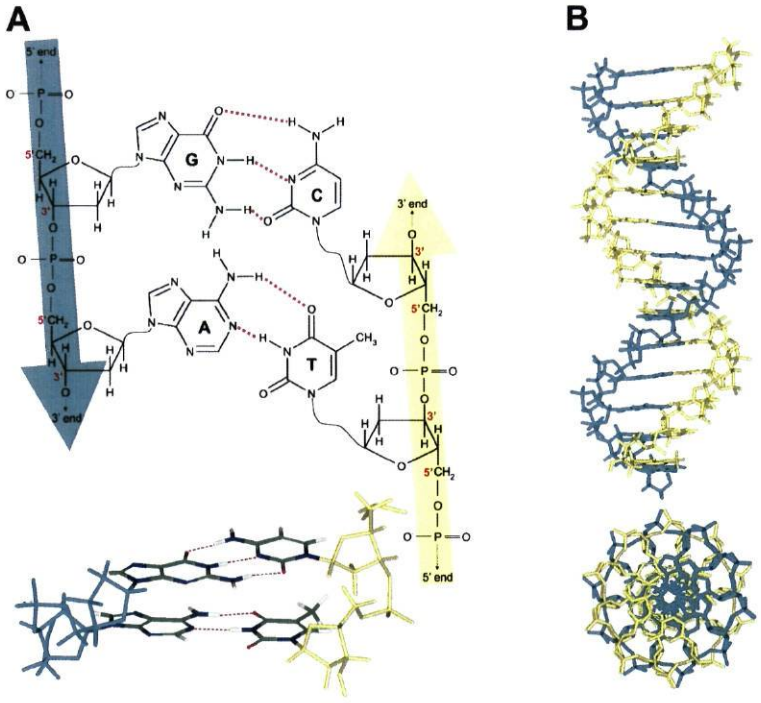


Figure 23. a) Base pairing in DNA – the hydrogen bonds between paired bases are represented as dotted red lines; b) DNA helix (side and top views).

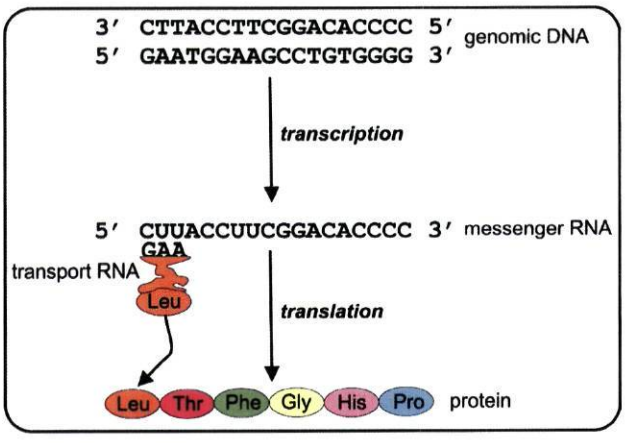


Figure 24. Schematic representation of DNA transcription and and mRNA translation.

The standard or 'universal' genetic code is used for both prokaryote and eukaryote genes with a few exceptions (mitochondrial genes, nuclear genes of ciliated protozoans and genes of the prokaryotic *Mycoplasma capricolum*). With the exception of tryptophan and metionine, all other amino acids are encoded by more than one codon. This means that mutations that affect only one of the nucleotides in the codon may or not change the amino acid it encodes, resulting in nonsynonymous or synonymous mutations, respectively (Figure 25).

Table 1. Standard genetic code (U, C, A and G stand for uracyl, cytosine, adenine and guanine nucleotides, respectively).

Codon	Amino acid	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	**UAA**	**Ter**	**UGA**	**Ter**
UUG	Leu	UCG	Ser	**UAG**	**Ter**	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGG	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGG	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	**Met**	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

These mutational changes can be substitutions of a nucleotide by another, insertion or deletion of nucleotides and inversion of nucleotides (e.g. in the sequence T*A*TGCG the adenine nucleotide changes place, and the sequence becomes TTG*A*CG). Insertions or deletions may shift the reading frame of a nucleotide sequence, thus are called frameshift mutations. Nucleotide substitutions can be either transitions [substitution of a purine (adenine or guanine) for another purine] or transversions [substitution of a pyrimidine (thymine or cytosine) for another pyrimidine]. Mutations that result in stop codons are named nonsense.

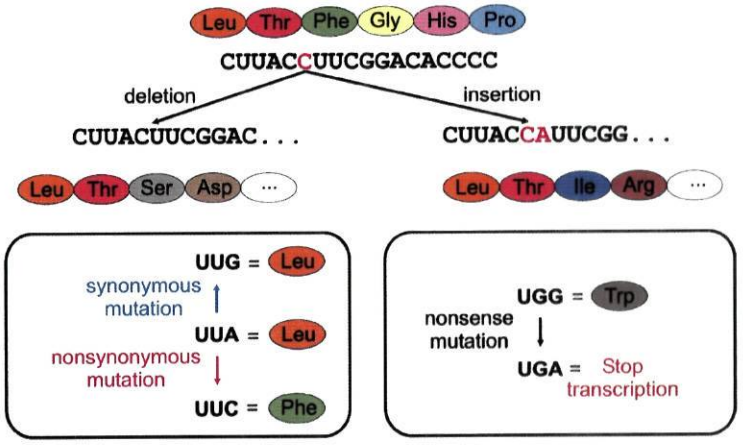


Figure 25. Types of mutations.

In order to obtain a mutation rate we calculate the number of synonymous substitutions per synonymous site, $d_S = 2r_S t$, and the number of nonsynonymous substitutions per nonsynonymous site, $d_N = 2r_N t$, where t is the divergence time [91].

Under neutral evolution, the rates of synonymous (d_S) and nonsynonymous substitutions (d_N) should be equal to each other. However, the rate of synonymous substitutions is usually higher than that of nonsynonymous substitutions ($d_S > d_N$), in order to maintain the function of the genes – negative or purifying selection ($\omega = d_N / d_S < 1$). When positive Darwinian selection influences the mutations rate in a gene, the rate of nonsynonymous substitutions is higher than that of synonymous substitutions ($\omega > 1$) [91].

If mutational events were random, the probability of occurrence of one of the four nucleotides in each nucleotide site should be the same. Also the relative frequency of codons encoding the same amino acid should be equal on average. However, changes in nucleotide sequences are dependent on:

- the codon usage bias: some codons are used more often than others because the correspondent tRNAs are more abundant (this abundance is correlated with the number of copies of the gene that encodes the tRNAs); codons which are less expressed will be eliminated by purifying selection, particularly for highly expressed genes, to improve the efficiency in protein synthesis;
- the biased mutation pressure: the relative frequency of nucleotides G and C (GC content) is known to vary from about 25 to 75%;
- the fact that mutations in the first position are mostly nonsynonymous, and consequently will be influenced by functional constraints and mutation pressure;
- the functional constraints that control mutations at the second position – as these always result in the change of the coded amino acid;
- the fact that mutations in the third position are predominantly synonymous, and will be mostly under mutation pressure; functional constraints will also interfere with the mutation rate at a lower extent.

This means that when measuring the rate of nucleotide substitution such constraints should be taken into account.

Phylogenetic trees are used to depict the divergence over time in a graphic form, using either the amino acid or nucleotide sequences. The former are useful for studying long-term evolution of genes or species because they are more conserved than the latter.

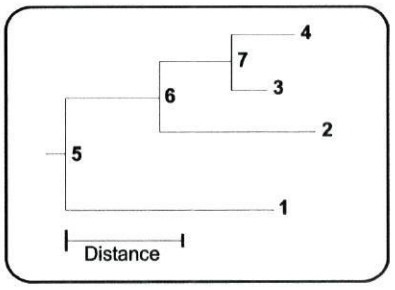


Figure 26. Phylogenetic tree.

In these representations, one can easily distinguish those genes that show the least number of nucleotide sequence differences (e.g. nos. 3 and 4 in Figure 26). Also, the relationship between such a pair of genes and a third one, towards which they both present a large number of differences (e.g. no. 2 when compared with nos. 3 and 4 in Figure 26) is readily seen.

There are several approaches to measure amino acid evolutionary distances. The simplest is the *p* distance which corresponds to the proportion of different amino acids between two sequences. More complex approaches include the Poisson-correction distance (which accounts for multiple amino acid substitutions at the same site), the Gamma distance (considers that amino acid sites of different functional importance will present different substitution rates from the others) and the Grishin distance (which includes a factor that relates the substitution rate with the chemical characteristics of the amino acid pair being considered) [91].

Measuring nucleotide differences involves the use of complex mathematical models, although, as mentioned, it is also possible to use a *p* distance (that measures the number of different nucleotides between two sequences). A frequently used model is the Tamura and Nei [92], which accounts for transition/transversion and GC content bias.

There are two main approaches to evaluate phylogenetic relationships. The computationally faster are the distance-based methods, which calculate distances between each pair of sequences, forming a distance matrix which will be used for the rest of the analysis. This approach is used in the Neighbor-Joining method [91].

Character-based methods, such as maximum parsimony (MP) [91] and maximum likelihood (ML) [91], compare all sequences in the alignment simultaneously. Maximum parsimony methods choose the nucleotide sequences for the extinct ancestors on the tree using the minimum number of changes. The goal of MP is to find the tree that requires the fewest base or amino acid substitutions, when mutational distances from each sequence to each ancestral node, and between ancestral nodes, are added up. The most parsimonious tree will be that with the smallest tree score/length.

ML methods will try to maximize the likelihood of observing a given set of sequence data using a specific codon substitution model for each topology. It builds the tree using an ancestral reconstruction based approach, which makes it very time consuming as it considers all possible nucleotides at each interior node (e.g. nos. 5, 6 and 7 in Figure). The parameters to be considered are the branch lengths for each topology, and the likelihood is maximized to estimate branch lengths. The topology that gives the highest maximum likelihood is chosen as the final tree.

The quality of the inferred trees can be evaluated by doing a bootstrap resampling of the sequence data. In this method, *n* nucleotide sites are randomly chosen with replacement from the original set of sequences. One can choose the number of replicate datasets to be created (default is usually 100) each containing positions sampled at random from the sequence alignment. In each set, some positions will be overrepresented, and others underrepresented. A large enough set of replicates should ensure that all parts of the sequence are equally biased among the replicates as a whole. The topology of the trees generated with these new sequence sets are compared to the original one, and the similarities are scored. If the data are robust, meaning that a given branch appears regardless of which sites are omitted from the sample, then that branch is strongly supported by the data, and will be attributed a high bootstrap value (>95%).

1.4. References

1. Anzenbacher P, Anzenbacherova E: **Cytochromes P450 and metabolism of xenobiotics**. *Cellular and Molecular Life Sciences* 2001, **58:**737-747.

2. Omiecinski CJ, Remmel RP, Hosagrahara VP: **Concise review of the cytochrome P450s and their roles in toxicology**. *Toxicol.Sci.* 1999, **48:**151-156.

3. Klingenberg M: **Pigments of rat liver microsomes**. *Arch.Biochem.Biophys.* 1958, **75:**376-386.

4. Garfinkel D: **Studies on pig liver microsomes. I. Enzymic and pigment composition of different microsomal fractions**. *Arch.Biochem.Biophys.* 1958, **77:**493-509.

5. Werck-Reichhart D, Feyereisen R: **Cytochromes P450: a success story**. *Genome Biol.* 2000, **1:**reviews3003.1-3003.9.

6. Scott EE, White MA, He YA, Johnson EF, Stout CD, Halpert JR: **Structure of mammalian cytochrome P4502B4 complexed with 4-(4-chlorophenyl) imidazole at 1.9-angstrom resolution - Insight into the range of P450 conformations and the coordination of redox partner binding**. *Journal of Biological Chemistry* 2004, **279:**27294-27301.

7. Guengerich FP: **Common and uncommon cytochrome P450 reactions related to metabolism and chemical toxicity**. *Chem. Res. Toxicol.* 2001, **14:**611-650.

8. de Graaf C, Vermeulen NP, Feenstra KA: **Cytochrome p450 in silico: an integrative modeling approach**. *J.Med.Chem.* 2005, **48:**2725-2755.

9. Guengerich FP: **Cytochrome P450 oxidations in the generation of reactive electrophiles: epoxidation and related reactions**. *Arch. Biochem. Biophys.* 2003, **409:**59-71.

10. Scott EE, He YA, Wester MR, White MA, Chin CC, Halpert JR, Johnson EF, Stout CD: **From The Cover: An open conformation of mammalian cytochrome P450 2B4 at 1.6-A resolution**. *PNAS* 2003, **100:**13196-13201.

11. Schleinkofer K, Sudarko, Winn PJ, Lüdemann SK, Wade RC: **Do mammalian cytochrome P450s show multiple ligand access pathways and ligand channelling?** *EMBO reports* 2005, **6:**584-589.

12. 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.

13. 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.

14. Sevrioukova IF, Li H, Zhang H, Peterson JA, Poulos TL: **Structure of a cytochrome P450-redox partner electron-transfer complex**. *PNAS* 1999, **96:**1863-1868.

15. Boddupalli SS, Hasemann CA, Ravichandran KG, Lu J, Goldsmith EJ, Deisenhofer J, Peterson JA: **Crystallization and Preliminary X-Ray Diffraction Analysis of P450terp and the Hemoprotein Domain of P450BM-3, Enzymes Belonging to Two Distinct Classes of the Cytochrome P450 Superfamily**. *PNAS* 1992, **89:**5567-5571.

16. Nelson DR, Zeldin DC, Hoffman SM, Maltais LJ, Wain HM, Nebert DW: **Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants**. *Pharmacogenetics* 2004, **14:**1-18.

17. Sinha R, Rothman N: **Role of well-done, grilled red meat, heterocyclic amines (HCAs) in the etiology of human cancer**. *Cancer Lett.* 1999, **143:**189-194.

18. Boobis AR, Lynch AM, Murray S, de la Torre R, Solans A, Farre M, Segura J, Gooderham NJ, Davies DS: **CYP1A2-catalyzed conversion of dietary heterocyclic**

amines to their proximate carcinogens is their major route of metabolism in humans. *Cancer Res.* 1994, **54:**89-94.

19. Turesky RJ, Constable A, Richoz J, Varga N, Markovic J, Martin MV, Guengerich FP: **Activation of heterocyclic aromatic amines by rat and human liver microsomes and by purified rat and human cytochrome P450 1A2**. *Chem. Res. Toxicol.* 1998, **11:**925-936.

20. Garner RC, Lightfoot TJ, Cupid BC, Russell D, Coxhead JM, Kutschera W, Priller A, Rom W, Steier P, Alexander DJ, Leveson SH, Dingley KH, Mauthe RJ, Turteltaub KW: **Comparative biotransformation studies of MeIQx and PhIP in animal models and humans**. *Cancer Lett.* 1999, **143:**161-165.

21. Iba MM, Fung J: **Pulmonary Cyp1A1 and CYP1A2 levels and activities in adult male and female offspring of rats exposed during gestation and lactation to 2,3,7, 8-tetrachlorodibenzo-p-dioxin**. *Biochem. Pharmacol.* 2001, **62:**617-626.

22. Wei C, Caccavale RJ, Kehoe JJ, Thomas PE, Iba MM: **CYP1A2 is expressed along with CYP1A1 in the human lung**. *Cancer Lett.* 2001, **171:**113-120.

23. Langouet S, Welti DH, Kerriguy N, Fay LB, Huynh-Ba T, Markovic J, Guengerich FP, Guillouzo A, Turesky RJ: **Metabolism of 2-amino-3,8-dimethylimidazo[4,5-f]quinoxaline in human hepatocytes: 2-amino-3-methylimidazo[4,5-f]quinoxaline-8-carboxylic acid is a major detoxification pathway catalyzed by cytochrome P450 1A2**. *Chem. Res. Toxicol.* 2001, **14:**211-221.

24. Zhai S, Dai RK, Friedman FK, Vestal RE: **Comparative inhibition of human cytochromes P450 1A1 and 1A2 by flavonoids**. *Drug Metab. Dispos.* 1998, **26:**989-992.

25. Zhai S, Dai R, Wei X, Friedman FK, Vestal RE: **Inhibition of methoxyresorufin demethylase activity by flavonoids in human liver microsomes**. *Life Sci.* 1998, **63:**L119-L123.

26. Bear WL, Teel RW: **Effects of citrus flavonoids on the mutagenicity of heterocyclic amines and on cytochrome p450 1A2 activity**. *Anticancer Research* 2000, **20:**3609-3614.

27. Hodek P, Trefil P, Stiborová M: **Flavonoids - potent and versatile biologically active compounds interacting with cytochromes P450**. *Chemico-Biological Interactions* 2002, **139:**1-21.

28. Lee H, Yeom H, Kim YG, Yoon CN, Jin C, Choi JS, Kim BR, Kim DH: **Structure-related inhibition of human hepatic caffeine N3-demethylation by naturally occurring flavonoids**. *Biochem. Pharmacol.* 1998, **55:**1369-1375.

29. Edenharder R, Rauscher R, Platt KL: **The inhibition by flavonoids of 2-amino-3-methylimidazo[4,5-f]quinoline metabolic activation to a mutagen: a structure-activity relationship study**. *Mutat. Res.* 1997, **379:**21-32.

30. Miranda CL, Yang YH, Henderson MC, Stevens JF, Santana-Rios G, Deinzer ML, Buhler DR: **Prenylflavonoids from Hops Inhibit the Metabolic Activation of the Carcinogenic Heterocyclic Amine 2-Amino-3-methylimidazo[4,5-f]quinoline, Mediated by cDNA-Expressed Human CYP1A2**. *Drug Metab. Dispos.* 2000, **28:**1297-1302.

31. Tsyrlov IB, Mikhailenko VM, Gelboin HV: **Isozyme-Specific and Species-Specific Susceptibility of Cdna-Expressed Cyp1A P-450S to Different Flavonoids**. *Biochimica et Biophysica Acta-Protein Structure and Molecular Enzymology* 1994, **1205:**325-335.

32. Silva ID, Gaspar J, da Costa GG, Rodrigues AS, Laires A, Rueff J: **Chemical features of flavonols affecting their genotoxicity. Potential implications in their use as therapeutical agents**. *Chemico-Biological Interactions* 2000, **124:**29-51.

33. Otake Y, Hsieh F, Walle T: **Glucuronidation versus oxidation of the flavonoid galangin by human liver microsomes and hepatocytes**. *Drug Metab. Dispos.* 2002, **30:**576-581.

34. Otake Y, Walle T: **Oxidation of the flavonoids galangin and kaempferide by human liver microsomes and CYP1A1, CYP1A2, and CYP2C9**. *Drug Metab. Dispos.* 2002, **30:**103-105.

35. Heller W: **Topics in the biosynthesis of plants**. *Acta Horticulturae* 1994, **381:**46-73.

36. Nelson DR, Koymans L, Kamataki T, Stegeman JJ, Feyereisen R, Waxman DJ, Waterman MR, Gotoh O, Coon MJ, Estabrook RW, Gunsalus IC, Nebert DW: **P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature**. *Pharmacogenetics* 1996, **6:**1-42.

37. Schlichting I, Berendzen J, Chu K, Stock AM, Maves SA, Benson DE, Sweet RM, Ringe D, Petsko GA, Sligar SG: **The Catalytic Pathway of Cytochrome P450cam at Atomic Resolution**. *Science* 2000, **287:**1615-1622.

38. Meunier B, de Visser SP, Shaik S: **Mechanism of oxidation reactions catalyzed by cytochrome p450 enzymes**. *Chem. Rev.* 2004, **104:**3947-3980.

39. Aikens J, Sligar SG: **Kinetic Solvent Isotope Effects during Oxygen Activation by Cytochrome P-450cam**. *J.Am.Chem.Soc.* 1994, **116:**1143-1144.

40. Benson DE, Suslick KS, Sligar SG: **Reduced oxy intermediate observed in D251N cytochrome P450cam**. *Biochemistry* 1997, **36:**5104-5107.

41. Martinis SA, Atkins WM, Stayton PS, Sligar SG: **A conserved residue of cytochrome P-450 is involved in heme-oxygen stability and activation**. *J.Am.Chem.Soc.* 1989, **111:**9252-9253.

42. Gerber NC, Sligar SG: **Catalytic mechanism of cytochrome P-450: evidence for a distal charge relay**. *J.Am.Chem.Soc.* 1992, **114:**8742-8743.

43. Shaik S, Kumar D, de Visser SP, Altun A, Thiel W: **Theoretical perspective on the structure and mechanism of cytochrome P450 enzymes**. *Chem.Rev.* 2005, **105:**2279-2328.

44. Clarke SE, Ayrton AD, Chenery RJ: **Characterization of the inhibition of P4501A2 by furafylline**. *Xenobiotica* 1994, **24:**517-526.

45. Kunze KL, Trager WF: **Isoform-selective mechanism-based inhibition of human cytochrome P450 1A2 by furafylline**. *Chem. Res. Toxicol.* 1993, **6:**649-656.

46. Racha JK, Rettie AE, Kunze KL: **Mechanism-based inactivation of human cytochrome P450 1A2 by furafylline: detection of a 1:1 adduct to protein and evidence for the formation of a novel imidazomethide intermediate**. *Biochemistry* 1998, **37:**7407-7419.

47. Ogliaro F, de Visser SP, Shaik S: **The 'push' effect of the thiolate ligand in cytochrome P450: a theoretical gauging**. *J. Inorg. Biochem.* 2002, **91:**554-567.

48. Auclair K, Moenne-Loccoz P, Ortiz dM: **Roles of the proximal heme thiolate ligand in cytochrome p450(cam)**. *J. Am. Chem. Soc.* 2001, **123:**4877-4885.

49. Sligar SG, Gunsalus IC: **A Thermodynamic Model of Regulation: Modulation of Redox Equilibria in Camphor Monoxygenase**. *PNAS* 1976, **73:**1078-1082.

50. Kazlauskaite J, Westlake ACG, Wong L-L, Hill HAO: **Direct electrochemistry of cytochrome P450cam**. *Chem.Commun.* 1996,2189.

51. Shumyantseva VV, Bulko TV, Archakov AI: **Electrochemical reduction of cytochrome P450 as an approach to the construction of biosensors and bioreactors**. *J. Inorg. Biochem.* 2005, **99:**1051-1063.

52. Bistolas N, Wollenberger U, Jung C, Scheller FW: **Cytochrome P450 biosensors-a review**. *Biosens. Bioelectron.* 2005, **20:**2408-2423.

53. Lei C, Wollenberger U, Jung C, Scheller FW: **Clay-bridged electron transfer between cytochrome p450(cam) and electrode**. *Biochem. Biophys. Res. Commun.* 2000, **268:**740-744.

54. Aguey-Zinsou K-F, Bernhardt PV, De Voss JJ, Slessor KE: **Electrochemistry of P450cin: new insights into P450 electron transfer**. *Chem.Commun.* 2003,418-419.

55. Fleming BD, Tian Y, Bell SG, Wong LL, Urlacher V, Hill HA: **Redox properties of cytochrome P450BM3 measured by direct methods**. *Eur.J.Biochem.* 2003, **270:**4082-4088.

56. Goodman JM: *Chemical Applications of Molecular Modelling*. Cambridge: The Royal Society of Chemistry; 1998.

57. Leach AR: *Molecular Modelling*. Harlow: Pearson Education Limited; 2001.

58. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M: **All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins**. *J.Phys.Chem.B* 1998, **102:**3586-3616.

59. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA: **Development and testing of a general amber force field**. *J.Comput.Chem.* 2004, **25:**1157-1174.

60. Allinger NL, Yuh YH, Lii J-H: **Molecular Mechanics. The MM3 Force Field for Hydrocarbons. 1.** *J.Am.Chem.Soc.* 1989,8551-8565.

61. Allinger NL, Chen K, Lii J-H: **An Improved Force Field (MM4) for Saturated Hydrocarbons**. *J.Comput.Chem.* 1996, **17:**642-668.

62. Halgren TA: **Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization and Performance of MMFF94**. *J.Comput.Chem.* 1996, **17:**490-519.

63. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M: **CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations**. *J. Comp. Chem.* 1983, **4:**187-217.

64. Jones G, Willet P, Glen RC, Leach AR, Taylor R: **Development and validation of a genetic algorithm for flexible docking**. *Journal of Molecular Biology* 1997, **267:**727-748.

65. Jensen F: *Introduction to Computational Chemistry*. Chichester: John Wiley & Sons Ltd.; 1999.

66. Cramer CJ: *Essentials of Computational Chemistry*, edn 2nd. Chichester: John Wiley & Sons; 2004.

67. Dolg M, Wedig U, Stoll H, Preuss H: **Energy-adjusted *ab initio* pseudopotentials for the first row transition elements**. *J.Chem.Phys.* 1997, **86:**866-872.

68. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery Jr JA, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski J, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA. Gaussian 03, Revision B.04. 2003.

69. Scherlis DA, Estrin DA: **Structure and spin-state energetics of an iron porphyrin model: an assessment of theoretical methods**. *Int.J.Quantum Chem.* 2002, **87:**158-166.

70. Hohenberg P, Kohn W: **Inhomogeneous Electron Gas**. *Phys.Rev.* 1964, **136:**B864.

71. Vosko SH, Wilk L, Nusair M: **Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis**. *Can.J.Phys.* 1980, **58:**1200-1211.

72. Becke AD: **Density-functional exchange-energy approximation with correct asymptotic behavior**. *PHYSICAL REVIEW.A* 1988, **38:**3098-3100.

73. Lee C, Yang W, Parr RG: **Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density**. *PHYSICAL REVIEW B.CONDENSED.MATTER* 1988, **37:**785-789.

74. Becke AD: **Density-functional thermochemistry. III. The role of exact exchange.** *J.Chem.Phys.* 1993, **98:**5648-5652.

75. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML: **Comparison of simple potential functions for simulating liquid water**. *J.Chem.Phys.* 1983, **79:**926-935.

76. Sitkoff D, Sharp KA, Honig B: **Accurate calculation of hydration free-energies using microscopic solvent models**. *J. Phys. Chem.* 1994, **98:**1978-1988.

77. Cossi M, Rega N, Scalmani G, Barone V: **Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model**. *J. Comput. Chem.* 2003, **24:**669-681.

78. Rocchia W, Alexov E, Honig B: **Extending the Applicability of the Nonlinear Poisson-Boltzmann Equation: Multiple Dielectric Constants and Multivalent Ions.** *J. Phys. Chem. B* 2001, **105:**6507-6514.

79. Murray JS, Politzer P: **The Molecular Electrostatic Potential**. In *Molecular Orbital Calculations for Biological Systems*. Edited by Sapse A. Oxford University Press; 1998:49-84.

80. Narayszabo G, Ferenczy GG: **Molecular Electrostatics**. *Chemical Reviews* 1995, **95:**829-847.

81. Narayszabo G: **Protein-ligand interactions**. In *Molecular Interactions*. Edited by Sheiner S. John Wiley & Sons Ltd; 1997:335-350.

82. Politzer P, Murray JS: *Reviews in Computational Chemistry* 1991, **2:**273-312.

83. Portela C, Afonso CM, Pinto MM, Ramos MJ: **Computational studies of new potential antimalarial compounds--stereoelectronic complementarity with the receptor**. *J. Comput. Aided Mol. Des.* 2003, **17:**583-595.

84. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE: **Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models**. *Acc. Chem. Res.* 2000, **33:**889-897.

85. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools**. *Nucleic Acids Res.* 1997, **25:**4876-4882.

86. Sali A, Blundell TL: **Comparative protein modelling by satisfaction of spatial restraints**. *J. Mol. Biol.* 1993, **234:**779-815.

87. Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling.** *Electrophoresis* 1997, **18:**2714-2723.

88. Laskowski RA, MacArthur MW, Moss DS, Thornton JM: **PROCHECK: a program to check the stereochemical quality of protein structures**. *J.Appl.Cryst.* 1993, **26:**283-291.

89. Liberles DA, Wayne ML: **Tracking adaptive evolutionary events in genomic sequences**. *Genome Biol.* 2002, **3:**REVIEWS1018.

90. Vallender EJ, Lahn BT: **Positive selection on the human genome**. *Hum.Mol.Genet.* 2004, **13:**R245-R254.

91. Nei M, Kumar S: *Molecular Evolution and Phylogenetics*. New York: Oxford University Press; 2000.

92. Hasegawa M, Kishino H, Yano T: **Dating of the human-ape splitting by a molecular clock of mitochondrial DNA**. *J.Mol.Evol.* 1985, **22:**160-174.

2. Results and Discussion

The PhD work presented in this thesis has been published/submitted as research articles in peer-reviewed international journals, which make up this *Results and Discussion* chapter.

This research work involved the use of multiple computational methodologies to understand some of the biochemical features of different Cytochromes P450. Initially, human and rat Cytochromes P450 1A2 (CYP1A2) were studied. This enzyme has been related to the development of tumours in several tissues which have been associated to its role in the metabolism of heterocyclic amines (HAs). These compounds are ubiquitous in red meat (beef, pork or lamb) cooked at high temperatures (e.g. grilled in charcoal). The human and rat enzymes present a high sequence similarity (75% conserved residues) but differences in the way they metabolize the two HAs studied (both in the catalytic activity and the regioselectivity): 2-amino-1-methyl-6-phenylimidazo[4,5-*f*]pyridine (MeIQx) and 2-amino-3,4-dimethylimidazo[4,5-*f*]quinoline (MeIQ). In the first article, entitled "*Modeling the metabolic action of human and rat CYP1A2 and its relationship with the carcinogenicity of heterocyclic amines*", a meticulous methodology for building homology models for the CYP1A2 enzymes using different multiple sequence alignments and quality checks were used thoroughly to obtain good quality structures, both stereochemically and according to literature information. Manual docking was used to place the ligands in the active site and molecular mechanics was used to optimize the geometry of the complexes using explicit solvation. The interaction of the enzymes with the two heterocyclic amines was analyzed shedding some light into the consequences that small variations on the active site have on substrates binding modes, resulting in the production of different metabolites by the two enzymes studied.

The next step involved the study of the inhibition of the human CYP1A2 enzyme. We looked specifically at naturally occurring compounds that were shown to inhibit this enzyme. Two groups of flavonoids were chosen. The first was composed by 8-Prenylnaringenin, Isoxanthohumol and Xanthohumol, that occur in hops and beer. The second group contained six flavone hydroxylated derivatives that exist in vegetables, fruit, tea and red wine. The second and third articles entitled "*Computational insight into anti-mutagenic properties of CYP1A flavonoid ligands*" and "*Molecular interactions between human CYP1A2 and flavones derivatives*", respectively, present a thorough analysis of the interactions between these compounds and the model structure of the human enzyme modeled in the previous work. The goal of these projects was to find which were the chemical properties and geometrical characteristics (such as shape, volume and surface areas) that were involved in enzyme-ligand complementarity. Initially,

molecular electrostatic potential maps for the ligands were analyzed to find common patterns of molecular recognition. By observing the molecular electrostatic pattern it is possible to detect the potential sites for H-bonding and other noncovalent interactions formation, which can be very important for a correct orientation of the ligand inside the enzyme. The ligands were then docked into the active site of the enzyme and the main points for electrostatic and other stabilizing interactions between the ligands and the enzyme were thoroughly analyzed. Total stabilization energy resultant from the binding of the ligands was calculated using molecular mechanics and either a classical continuum solvation approach or explicit solvent. Also, some of the components of the total stabilization energy concerning the interaction between the ligands and specific groups of amino acids were shown in detail. The results were successfully correlated with experimentally determined inhibitory power of the flavonoids towards human CYP1A2.

Finally, the selective mutational pressures influencing Cytochromes P450 genetic variability and its structural consequences on protein structure and function were studied using both gene and protein level approaches. The fact that cytochromes P450 are involved in the metabolism of xenobiotics, and that these environmentally available substances vary throughout times, suggested that an accelerated evolution could be interfering with the metabolic activity of these enzymes. In the fourth article, entitled "*Functional divergence and diversifying selection on mammalian cytochromes P450 2C*", the effect of natural selection on CYP2 mammalian enzymes was examined. This is the largest and most diverse of CYP families. CYP2 enzymes metabolize a variety of different pharmaceutical agents. Besides, there are four available three-dimensional structures for CYP2 mammalian enzymes, which allow correlating the genetic variability with structural/physicochemical variations at specific sites on the proteins. Statistical tests that detect variation in selective pressures in nucleotide and amino acid sequences were used. The former measure if the rate of nonsynonymous substitutions is higher than that of synonymous substitutions, indicating that positive selection is acting on particular sites or areas of the enzyme, accelerating the fixation of mutations that change the amino acid sequence (otherwise, if only random mutations were responsible for changing the DNA sequence, both types of mutations would have a similar probability of occurrence). Otherwise, if the rate of synonymous substitutions is higher than that of nonsynonymous substitutions, negative selection is said to be occurring, which indicates that there is a high conservation of the amino acid sequence and thus, of the functional role of that particular domain. The other approach evaluated statistically significant physicochemical amino acid changes. All methods confirmed that the broadening and changing of CYP2s substrate specificity is a result of an accelerated rate of mutations on the

active site areas that are related to substrate binding while the areas related to maintaining the highly conserved catalytic mechanism (close to the heme prosthetic group) show signatures of negative selection.

1.1. Modeling the metabolic action of human and rat CYP1A2 and its relationship with the carcinogenicity of heterocyclic amines

MOLECULAR PHYSICS, 10 SEPTEMBER 2003, VOL. 101, NO. 17, 2731–2741

Taylor & Francis
Taylor & Francis Group

Modelling the metabolic action of human and rat CYP1A2 and its relationship with the carcinogenicity of heterocyclic amines†

RUTE DA FONSECA[1], MARIA CRISTINA MENZIANI[2], ANDRÉ MELO[1] and MARIA JOÃO RAMOS[1]*

[1]REQUIMTE/Departamento de Química, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal

[2]Dipartimento di Chimica, Università di Modena, Via Campi 183, 41100 Modena, Italy

(*Received 6 January 2003; accepted 6 July 2003*)

Cytochrome P450 (CYP) is a family of enzymes responsible for organism detoxification. However, some of the members of the CYP1A subfamily also catalyse the activation of heterocyclic amines (HAs), present in cooked meat, to carcinogenic compounds which have been shown to increase the risk of breast, colorectal and lung cancer. In humans, CYP1A2 is the enzyme with the most significant action in HA metabolism but in rodents CYP1A1 is also important in this biotransformation. Understanding the metabolic action of these enzymes is essential to predict the factors that enable the formation of the carcinogenic products. We have built two models of CYP1A2, one for the human enzyme and one for the rat homologue. The templates chosen include the only X-ray structure published to date for a mammal CYP, a quimeric C2A5 from rabbit, as well as CYPs belonging to *Bacillus megaterium* (CYPBm-3), *Pseudomonas putida* (CYPcam), *Pseudomonas sp.* (CYPterp), and *Saccharopolyspora erythraea* (CYPeryf). Two HAs, MeIQ (2-amino-3,4-dimethylimidazo[4,5-*f*]quinoline) and MeIQx (2-amino-3,8-dimethylimidazo[4,5-*f*]quinoxaline), known substrates of human and rat CYPIA2, were docked in the active site of the models, providing information regarding the different catalytic rates associated with the metabolisms in both enzymes. This is important for analysing the behaviour of animal models concerning the testing of anticancer drugs.

1. Introduction

Cytochrome P450 (CYP) is a family of enzymes which has a major role in allowing an organism to dispose of xenobiotic substances (substances alien to the living organism, such as pesticides, anaesthetics and food additives, among others) by transforming them into products that are easier to remove from the organism [1, 2]. Their activity occurs principally in the liver, the main organ responsible for drug and toxin removal, and they are polymorphically expressed [3, 4]. This detoxification mechanism exists in many life forms and the diversity of cytochromes P450 illustrates their adaptative ability to survive new natural toxins and noxious compounds to which organisms are constantly exposed [5].

Besides this detoxification action, the human cytochromes have other important roles such as the synthesis of steroid hormones, prostacyclin, thromboxane, cholesterol, and bile acids, and also the degradation of endogenous compounds, such as fatty acids, retinoic acid and steroids.

CYPs exist in all eukaryotes, most prokaryotes and Archea. Microbial enzymes are soluble proteins while eukaryotic CYPs are intrinsic membrane proteins that are present in the endoplasmic reticula of plant, fungal and animal cells. Animals also possess CYPs in the inner membrane of mitochondria [2].

Despite the benefits of its overall action, cytochrome P450 catalysis of some xenobiotics can transform them into reactive toxins and mutagens [2]. An example is the biotransformation of heterocyclic amines (HAs), which can be found in significant amounts in red meat (beef, pork or lamb) cooked at high temperatures. HAs are a result of the pyrolysation of creatine or creatinine and amino acids in meat juice. *In vitro* and animal studies show that HAs are genotoxic mutagens causing DNA damage, which results in the formation of tumours in a variety of tissues in several species [6, 7, 8]. The risk of cancer development in humans owing to HAs is yet to be established quantitatively, but consumption of well-done/very well-done red meat has already been

†Supporting information (SUP 16148) held with the British Library.

*Author for correspondence. e-mail: mjramos@fc.up.pt

Molecular Physics ISSN 0026-8976 print/ISSN 1362-3028 online © 2003 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/00268970310001603112

associated to an increase in the risk of developing breast, colorectal and lung cancer [8].

The carcinogenic HAs are formed by metabolic activation of those HAs present in cooked meat. The first step in this activation consists of an *N*-hydroxylation of the HAs catalysed by P450 cytochromes of the polycyclic aromatic hydrocarbon-inducible CYP1A subfamily, CYP1A1 and CYP1A2. In humans and rodents both CYP1A1 and CYP1A2 are inducible by several chemicals, including tobacco smoke [9, 10], and exhibit tissue-specific distribution, in which they differ greatly as CYP1A1 exists mainly in extrahepatic tissues and CYP1A2 is preferentially expressed in the liver [9–11]. Additionally, CYP1A2 exhibits polymorphic distribution in humans which means that N-hydroxylation of HAs and the associated risk factor for cancer development will be more significant in some populations than others [3, 12]. In rodents both CYP1A1 and CYP1A2 carry out the reaction but, in humans, it is mainly CYP1A2 that is responsible for it [12].

Two of the amines that are substrates of both human and rat CYP1A2 are MeIQx (2-amino-3,8-dimethylimidazo[4,5-f]quinoxaline) [2, 8, 12] and MeIQ (2-amino-3,4-dimethylimidazo[4,5-*f*]quinoline) [6, 13, 14]; they can be seen in figure 1.

According to the best sequence alignment between the human and rat isoforms of CYP1A2, it is possible to observe that 75% of the residues are conserved. With such high sequence identity, we expect that there will be a significant structural similarity between these structures [15]. However, small amino acid differences result in a significant difference both in the catalytic activity and the regioselectivity of the two CYP1A2s [4]. Therefore, a comparison of their active sites and the relative position of known docked substrates should provide some information on the differences in metabolic rates and products resulting from the activity of both enzymes. The ability to establish comparisons between the metabolic activity of CYP1A2 enzymes of both human and laboratory animals would be a step forward to the evaluation of the actual human health risk towards HAs. As there is no existing three-dimensional (3D) structure for these enzymes, building homology models is a necessary first step for establishing such a comparison.

The human and rodent genomes contain up to 50 CYP genes, which can be distributed through ten families. The mammalian genome CYPs show partial overlap but distinct substrate functions. Using comparative sequence analysis methods it is possible to focus on the stretches of sequence which are involved in recognition or binding of substrates, and in that way draw some conclusions about the different substrate affinity. Gotoh [5] has defined six putative substrate recognition sites (SRSs) in the CYP2 family. Gotoh related this variability to an increase in the number of metabolizable substrates—an adaptative evolution that in animals constitutes a response to the diversification of toxic materials in food.

One can establish comparisons amongst the several groups of CYPs using these SRSs, as they correspond to the active site region, which is more conserved through evolution than any of the loop regions, as well as other particular regions in the structure, such as the membrane binding domains in eukaryotes. One very important feature in this kind of enzyme that divides CYPs in two major groups is related to a key step in the P450 catalytic cycle—electron transfer from a redox partner. Those CYPs that use a flavin adenine dinucleotide (FAD) containing reductase and a soluble iron–sulphur protein for getting the electrons needed for their reactions are class I whereas class II enzymes use an FAD- and a flavin mononucleotide (FMN)-dependent flavoprotein reductase as a redox partner, namely cytochrome P450 reductase (CPR) [16, 17].

Structural comparisons among bacterial P450s show that they share a common fold [1], even though they have a very low sequence similarity (around 20%) and present quite significant sequence modifications in details that are involved in specificity of substrate interaction [18]. This remains true when the comparison is made with the eukaryotic P450s, as it is now possible to do after the determination of the first structure of a mammalian P450 [2, 19]. The most remarkable difference between the two groups of structures relies on the fact that the eukaryotic P450s are membrane bound

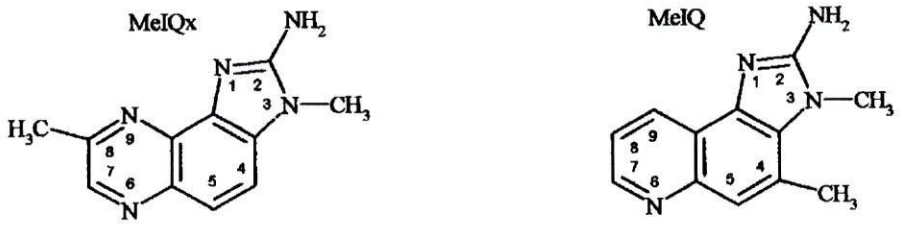


Figure 1. Two HAs, MeIQx (2-amino-3,8-dimethylimidazo[4,5-*f*]quinoxaline) and MeIQ (2-amino-3,4-dimethylimidazo [4,5-*f*]quinoline, substrates to both human and rat CYP1 A2.

proteins, while bacterial ones are fully soluble proteins [I]. Besides the *N*-terminal anchor, in eukaryotic P450 there is evidence for a strong interaction of the F-G loop with the membrane [16], meaning that it is expected that, structurally, this region differs significantly between both groups. Also, between class I and class II cytochromes, there should be some particular differences concerning the binding place for the redox partner. The regions responsible for these two types of interactions may therefore originate ambiguities in the multiple sequence alignment of enzymes, which are evolutionarily so far apart [18, 19]. Despite these differences, it has been shown that the overall fold of the active site region of the microbial P450s is conserved, and that the same kind of fold can be observed for the core of mammalian enzymes; the recent X-ray structure of a chimeric rabbit CYP2C5-2C3 [2, 19] has contributed to confirming this knowledge.

Up to now, because of the lack of structural information on mammalian cytochromes, homology models for human CYPs have been built [20] based on the available structures, determined by X-ray diffraction for both class I CYPs and class II CYPs. The newly available structure for a chimeric rabbit CYP2C5-2C3 [2] widens the possibilities for building more reliable models for human CYPs [18]. The impact of incorporating the 2C5 crystal structure into comparative models of several human CYPs has already been mentioned in literature [21, 22].

2. Methodology

2.1. *Multiple alignment*

After using the BLAST [23, 24] engine for finding sequence homologues of both human and rat CYP1A2, the coordinates for five CYPs, with X-ray structures available, were extracted from the Protein Data Bank (PDB) [25, 26] (table 1). The sequences for CYP1A2 from rat and human were extracted from the SWISS-PROT protein sequence database (references P04799 and P05177, respectively) [27]. The information regarding the sequence and structure of all the other structures was extracted from the corresponding PDB files.

In order to build the models, we first attempted an automatic alignment of all sequences using CLUSTALW [28, 29], and then performed a comparative analysis of the secondary structure of the X-ray template structures, together with the prediction, using the amino acid sequence of the rat and human enzymes. This was done using the homology module from the InsightII molecular modelling software from Accelrys Inc. [30]. This analysis proved the incapacity of the automatic method to correctly align the conserved secondary structure elements [20] of all the CYPs, which is undoubtedly a consequence of the relatively low amino acid sequence similarity between mammalian and bacterial CYPs (about 20%). Any other options of automated alignment, such as the profile approach from CLUSTALX [31] or an automatic alignment using both sequence homology and secondary structure information [32] were unable to achieve a good alignment, as the only region that was acceptably matched was a ten residue stretch corresponding to the Cys-pocket, the consensus pattern of the CYP superfamily, that includes the cysteine residue binding the iron atom from the heme. In the end, the best way of obtaining a good alignment was found to be the use of a combined profile/structural manual alignment approach. We began by aligning sequences of CYPs that were closer in evolutionary terms. Initially, the class I CYP sequences were aligned, one by one, with careful monitoring of how changes in the alignment affected the overall fit of the corresponding X-ray structures, with a view to optimizing the fit in the most conserved regions. This was done using the viewer and homology modules from InsightII. Then the CYPBm-3 sequence was added, followed by the mammalian CYP sequences already aligned with

Table 1. Templates chosen for building the CYP1A2 models. The Ramachandran plot was part of the PROCHECK analysis and the sequence similarity was calculated according to the alignment we have used with the five templates (%similarity = number of similar residues/total number of residues of the CYP1A2).

Class	Name	Species	PDB code	Resolution (Å)	Ramach. plot (%)[a]	3D profile score[b]	Sequence similarity with CYP1A2 (%)	
							Rat	Human
I	P450cam	*Pseudomonas putida*	1PHB	1.6	89.1	180.3	11.2	12.0
	P450eryf	*Saccharopolyspora erythraea*	1EUP	2.1	90.8	178.8	12.5	11.6
	P450terp	*Pseudomonas sp.*	1CPT	2.3	89.0	185.5	11.4	12.4
II	P450bm-3	*Bacillus megaterium*	1BU7	1.6	93.1	217.2	16.9	17.5
	P450 2C5	Rabbit	1DT6	3.0	71.3	169.8	28.0	30.2

[a]Percentage of residues in most favoured regions of the ϕ–ψ Ramachandran plot.
[b]This analysis was done using the unsolvated structures.

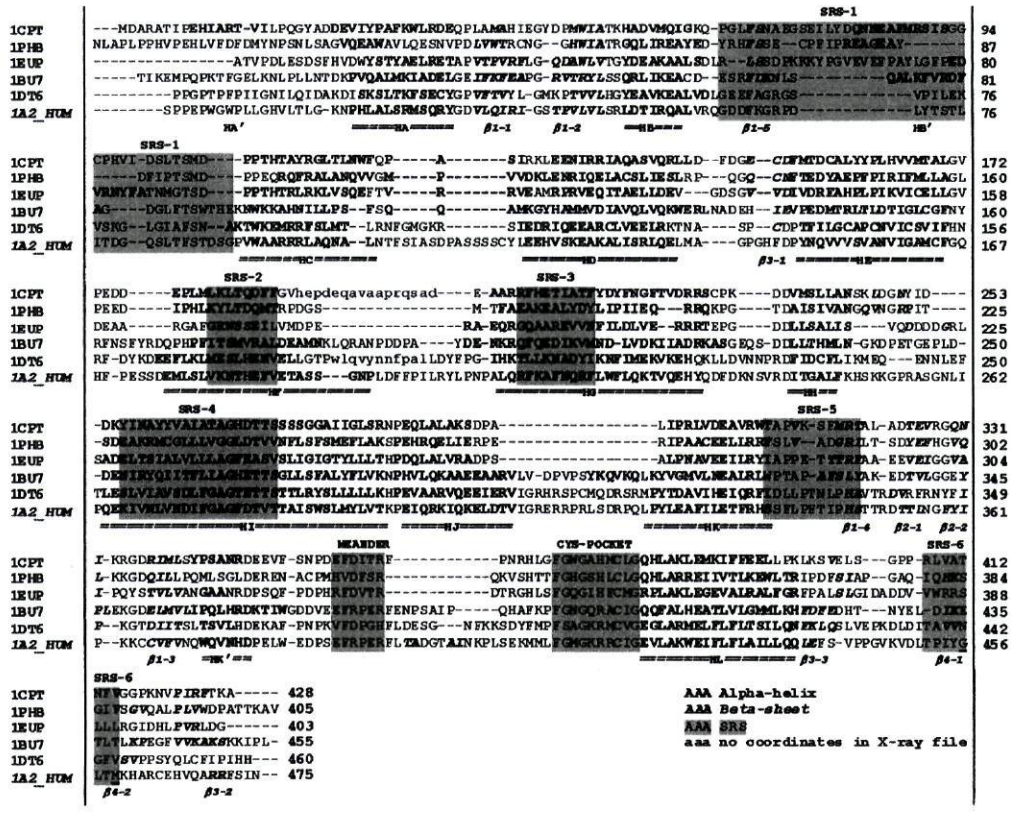


Figure 2. Alignment of the human CYP1A2 amino acid sequence with class I CYPs, P450cam, P450eryf and P450terp, and class II CYPs, CYPBm-3 and the chimeric CYP2C5-2C3.

each other. Misalignments due to comparison of low similarity sequences were therefore reduced, as the secondary structure elements were steadily aligned amongst each pair of the most homologous sequences, with CYPBm-3 serving as a link between eukaryotes and prokaryotes. This procedure also took into account some extra information such as experimental data on residues that are important for catalysis [13, 14, 33], the available information on the possible binding site for CPR [2, 15] and the relative position of CYP1A2 towards the membrane [2].

This alignment included both class I and class II CYPs (CYPcam, CYPeryf, and CYPterp, all from bacterial organisms) (figure 2). We have also made another alignment including class II templates only—the CYPBm-3 from *Bacillus megaterium* and the quimeric CYP2C5-2C3 from rabbit, i.e. the two templates more related to the human CYP1A2 (figure 3)—using the same monitoring procedure described above. By using multiple templates and two different alignments we hope to ensure an utter sampling of the conformational space of the protein, and therefore expect to build better quality models [19].

For generating the rat model, a single alignment with the human CYP sequence was used (figure 4). We ought to mention the fact that initially we followed the same procedure for rat CYP1A2 as for human CYP1A2, that is we carried out two alignments with five and two homologous sequences. Simultaneously, we have also attempted the alignment that we report here, basically because the sequence similarity for the two CYP1A2 sequences is very high (75%). The results and quality checks obtained with this last alignment were much better than with the other two and therefore we have adopted it to perform the rest of the work.

2.2. *Homology modelling*

The first residues, corresponding to the membrane anchor N-terminal helix, were ignored while building the model, as there is no template structure available for it (i.e. the first 40 residues in the human and the first 39 residues in the rat are missing from the model; the numbering of the residues on each model is made considering the first residue in the alignment as number one).

The coordinates from residues 191 to 206 were missing from the 1CPT PDB file as well as the coordinates for residues 211 to 222 from the IDT6 PDB file. The PDB files were then truncated one residue

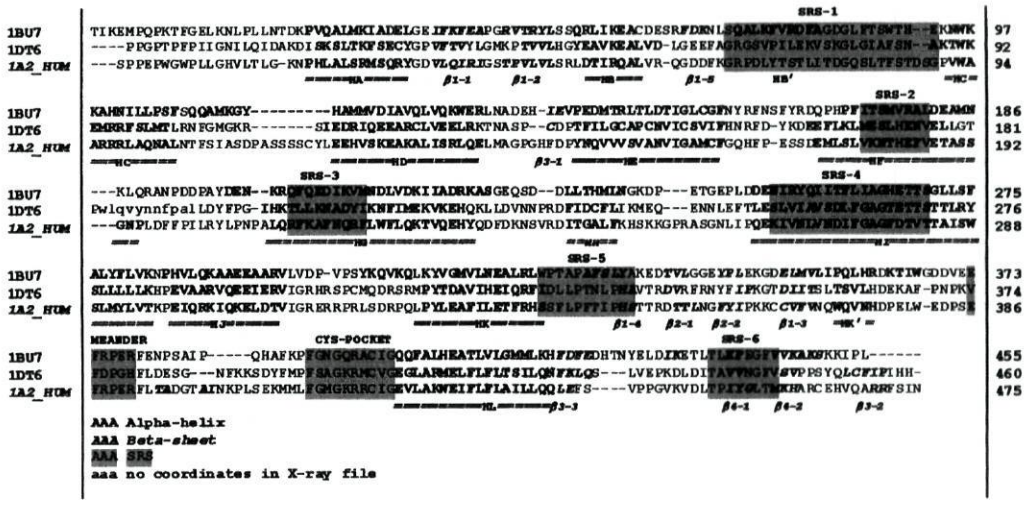


Figure 3. Alignment of the human CYP1A2 amino acid sequence with two class II templates, CYPBm-3 and the chimeric CYP2C5-2C3.

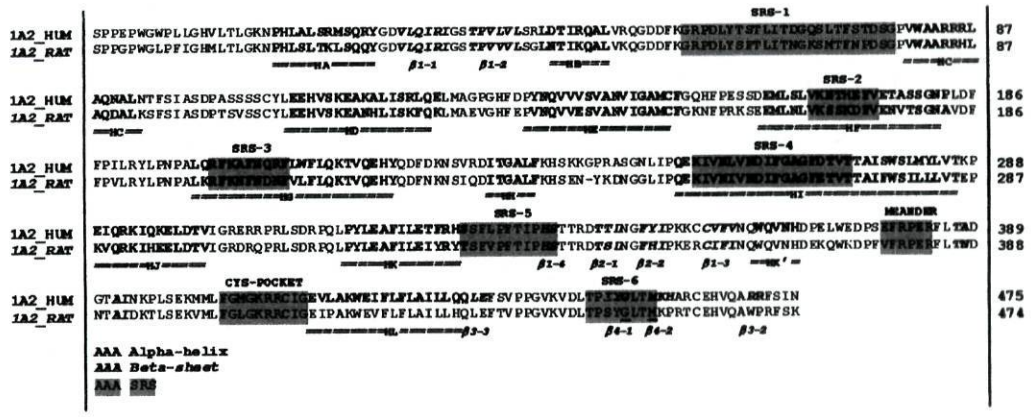


Figure 4. Alignment used to build the model for the rat CYP1A2 model.

more at the extremities formed by these deletions before the modelling procedure.

The models were built with the software Modeller [34]. Briefly, the program Modeller derives many distance and dihedral angle restraints in the target sequence from its alignment with the template 3D structures. Then, the spatial restraints and the energy terms enforcing proper stereochemistry are combined into an objective function, and the final models are obtained by optimizing the objective function in Cartesian space, employing methods of conjugate gradients and molecular dynamics with simulated annealing.

The first model to be built was the human CYP1A2. Several models were generated for each alignment, using the randomization of the Cartesian coordinates, allowing a deviation of ± 4 Å and the loop building option. The best of each set was chosen, considering the quality checks from Procheck [35], profile 3D analysis from the Quanta software package from Accelrys Inc. [30], and the main-chain root mean square deviation (RMSD) for conserved regions. The quality geometry factor indicator of Procheck measures how atypical each residue is concerning its covalent geometry and dihedral angles, with respect to the 'typical' distribution defined by the PDB structures. The 3D profile analysis [36] provides an evaluation of the overall fold and side-chain packing of the models. The method expresses the 3D structure of a protein as a table, the profile, which represents the local environment of each residue, characterized in terms of the statistical preferences of the profiled residue for neighbours of specific residue types, main-chain conformations, or secondary structure. The score of an amino acid sequence, aligned with the 3D profile, reflects its compatibility with the profiled structure. Therefore, the overall score obtained as the sum over all

the amino acids of the protein is a measure of the accuracy of the model and can be compared with the accuracy of the templates used in the modelling.

The final structure for the human CYP1A2, HUM1, combined features from both types of models generated by the different alignments, considering both the quality of the fragments of each model and the available information on the human CYP1A2. In general, the model obtained with the alignment with five templates was better than the one obtained with two templates. It has a better stereochemistry quality and presents a lower RMSD concerning the most homologous class II templates. However, regions that have structural importance exclusively for class II CYPs, such as the region close to the probable place of interaction with the redox partner, residues 100–123, as well as the region that includes the membrane insertion loop, residues 175–216, were taken from the model made with class II templates. One of the loops (residues 379–407), for which automated modelling was unable to present a reasonable choice, was added using the best fitting option given by the PDB loop database search from SwissPDB viewer [37, 38].

Both the stereochemical quality and the packing of the final model were much improved by generating a new human CYP1A2 model (HUMFINAL) using the patchwork structure (HUM1) as a single template in Modeller [33]. Again several models were generated using the randomization of the Cartesian coordinates, allowing a deviation of ± 4 Å and the loop building option. The best model was chosen following the criteria already mentioned. As expected, the most significant changes occurred in the loops, and we observed that the RMSD values between HUMFINAL and the templates concerning the conserved regions are very similar, and slightly better for all templates with the exception of 1DT6 (see table 2), which might mean that a slight bias of HUM1 structure towards 1DT6 structure has been lowered.

As we mentioned previously, using the human model as a single template generated the best rat model.

2.3. *Model refinement*

The coordinates for the haem molecule were extracted from the X-ray structure of 1DT6 and fitted into the active site of the models. The H-bonding interactions were predicted according to the existing interactions in the chimeric CYP2C5-2C3. The structure of the models was then refined with the program CHARMM and the corresponding force field [39]. The explicit hydrogen bond term was used in the CHARM energy calculations [39]. The united atom force-field parameters, and a 12 Å non-bonded cut-off distance were used. Solvent was treated explicitly by adding an 8 Å layer of TIP3P water to each model before the minimization procedure. Initially, the water molecules were minimized together with the hydrogens from the protein structure with the non-hydrogen atoms kept fixed. Then, an energy minimization was performed for each model maintaining the protein backbone fixed. Minimizations used 500 steps of steepest descent followed by a conjugate gradient.

2.4. *Substrate docking*

We have attempted to dock the known substrates MeIQx and MeIQ on the active site of both models to find out which were the residues directly involved in ligand binding. We have tried to mimic the first stage of the P450 cycle, when the iron from the heme is coordinated to a water molecule which will be displaced by the entrance of the substrate. As we are particularly interested in the carcinogenic product originated by the oxidation of the 2-amino group, we have orientated this amine group towards the iron of the heme, where the activation of molecular oxygen takes place. The

Table 2. Quality checks for the human CYP1A2 model done at several stages of refinement (HUM 2templates is the model obtained using the alignment with only class II CYP templates, HUM 5templates is the model obtained using all templates available, HUM1 is the result of the fusion of the models from these two types of alignments and HUMFINAL resulted from the use of an alignment between the human CYP1A2 sequence with itself, using the HUM1 coordinates as the single template).

Model	Ramachandran plot (%)[a]	3D profile score[b]	Main chain RMSD with templates[c]				
			1CPT	1PHB	1EUP	1BU7	1DT6
HUM 2templates	80.7	156.7	5.87	6.13	5.58	5.05	4.43
HUM 5templates	81.9	149.3	4.05	4.45	3.75	3.10	1.08
HUM1	82.6	149.5	4.11	4.43	3.76	3.09	0.93
HUMFINAL	86.5	163.1	4.07	4.38	3.72	3.07	0.99
RAT	86.2	151.7	4.06	4.36	3.69	3.08	1.02

[a]Percentage of residues in most favoured regions of the ϕ–ψ Ramachandran plot.
[b]This analysis was done using unsolvated structures.
[c]The RMSD values were obtained for the main chain of the residues that were aligned without any gaps, in a total of 1336 atoms.

substrates were fitted so that the amine was as close as possible to the heme. After a rough docking of the substrate, we performed a short minimization of the substrate, the heme and the side chains of the residues located up to 10 Å from the substrate; again we have used CHARMM. Subsequently, the whole structure was submitted to a second minimization using 500 steps of steepest descent followed by a conjugate gradient, also using a 12 Å non-bonded cut-off distance.

3. Results and discussion

The strategy used to build the model for the human CYP1A2 resulted in a good quality structure, both stereochemically (table 2) and according to literature information (table 3). In table 2 we list the scores obtained by the quality checks performed for the models in the several stages of refinement. They are: (a) the percentage of the residues found in the most favoured regions of the ϕ–ψ Ramachandran plot (Ramachandran plot %), (b) the compatibility of the models obtained with their sequences (3D profile score), and (c) the main-chain root mean square deviation (RMSD) for conserved regions. It can be seen that the number of residues in the allowed regions of the Ramachandran plot increases with the refinement of the human CYP1A2 model. The final model has a good Ramachandran score of 86.5%, much better than the one for the most homologous template, the quimeric CYP2C5-2C3, and close to the scores for the better resolution bacterial templates. Additionally, the overall fold and side-chain packing of the models provided by the Quanta Profile analysis (3D profile score) is better for the last stage of the refinement of the human CYP1A2 model, closer to the values obtained for the templates. We also show the RMSD values for the main-chain atoms of the correctly aligned amino acids. For sequence identities around 30%, the RMSD value for the main-chain atoms should be around 1.5 Å [15]. As can be seen in table 2, the RMSD value between the model and the templates improves during refinement. Using a rational selection of the structural information provided by the model built with five templates (HUM 5templates), and the model resulting from the class II template alignment (HUM 3templates), we ended up with a structure that is generally better related to all the templates than the two models that originated it.

As mentioned briefly in section 2.2, although the RMSD between the last stage of the refinement and the most homologous template is slightly higher when compared to the previous stage, it is lower relative to all the other templates. This may correspond to the correction of a structure that might have been slightly biased towards the rabbit cytochrome structure, as this template was much more homologous to the human CYP1A2 than any of the others and therefore had a relatively high weight in the automatic modelling process.

It is worth mentioning that the RMSD values in table 2 are not representative of the most conserved regions. As can be seen in figure 5, the overall (all residues)

Table 3. Important residues in the active site of the human CYP1A2, according to the literature.

Residue	Type of experimental data	Experimental result	Explanation according to model
Thr183	QSAR [11]	Involved in ligand docking	Interacts with a water molecule which is important for ligand orientation in the active site
Phe186	Experimental mutation studies [13, 27]	Mutants reduced MeIQ mutagenicity/reduced MeIQ catalysis	Phe186 limits the mobility of the ligand on the active site and the orientation of the polar groups
Asp280	Experimental mutation studies [13, 27]	Mutants show significantly lower activity/inefficient with MeIQ	This residue is involved in ligand docking
Thr281	Experimental mutation studies [27] QSAR [11]	Mutants display lower catalysis of MeIQ than wild type enzyme Important in ligand binding	Thr281 interacts with the water molecule bounds to both the heme and the ligand
Val282	Experimental mutation studies [27]	Important residue in ligand binding	Val282 side chain points towards the protoporphyrin
Leu342	QSAR and experimental mutation studies [11]	Mutation to Thr resulted in a practically null activity	Leu342 side chain points into the active site and seems to exert a structural role towards Phe341 which is important in ligand orientation
Ile346	Experimental mutation studies [13, 27]	Mutations to Pro or Thr gave much reduced activities	Close to heme ligands Arg416 and His348

QSAR, quantitative structure–activity relationship.

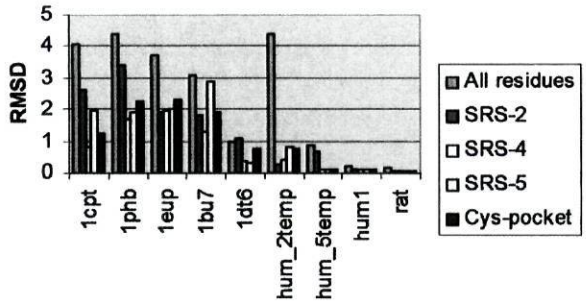


Figure 5. Comparison of the RMSD values between the main-chain atoms of the residues of specific regions versus all residues (aligned with no gap) of all structures and the final human CYP1A2 model.

RMSD is much higher than the ones for SRS-4 and the Cys-pocket. This is one of the reasons why it is reasonable to use very few homologous templates in the building of a model; here we are very interested in a good preview of the active site, whose secondary structure is reasonably well conserved through evolution. So, although we cannot fully trust the loop region, we can rely on the information related to the active site.

This modelling procedure has shown the importance of the use of a large number of templates for the building of a better homology model, as long as the secondary structure information is carefully taken into account in the sequence alignment, as it is better conserved through evolution than the amino acid sequence. Furthermore, the analysis of models built automatically using different alignments, together with the crossing of the available information on the protein of interest and the templates chosen, may allow the selection of the better generated parts of each model into a single, more realistic model.

Figure 6 shows the overall view of the models built for the human and rat CYPs after the docking of the substrate. One can observe the secondary structure elements and the relative position of the heme and the substrates.

In addition to the previously presented quality checks, the docking of MeIQ and MeIQx on to the active site of the models also demonstrates their good quality. As can be seen in table 3 and figure 7, some experimental evidence obtained with mutation and QSAR studies is in agreement with the results obtained after the docking of the two substrates. Figure 8 shows how the substrate recognition sites (SRSs) are positioned relative to the heme and the substrate binding position.

There is a striking difference between the active sites of the human and the rat CYP1A2, which is the existence of the $Glu279_{rat}$ residue in the rat active site and the $Asp280_{hum}$ in the same relative position in the human active site. In the human complexes, the side

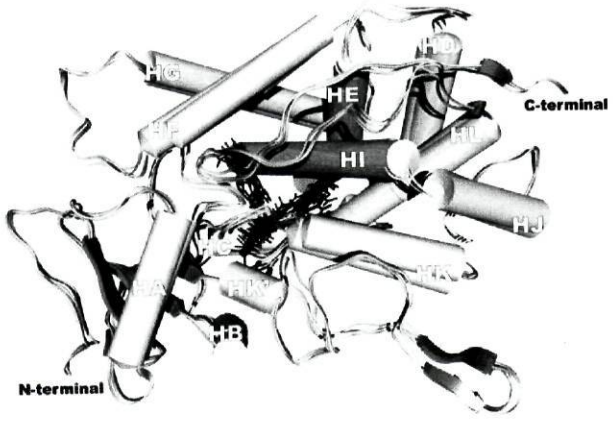


Figure 6. Overall view of the rat and human CYP1A2 models complexed with the two substrates, MeIQ and MeIQx.

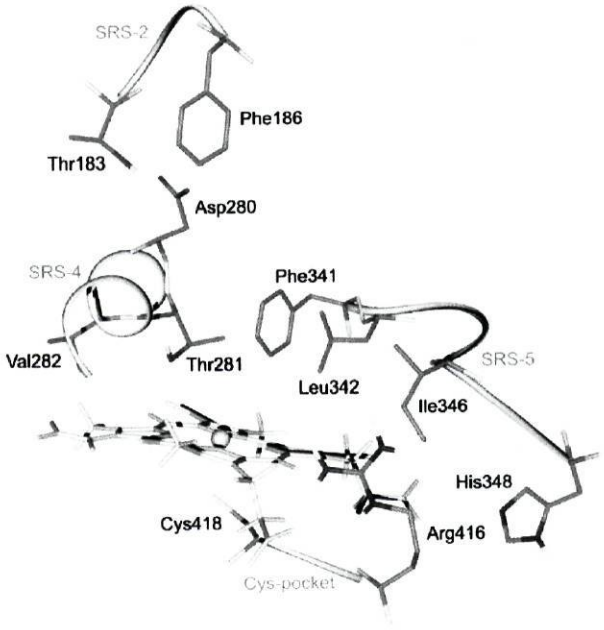


Figure 7. Relative position of the residues mentioned in table 3.

chain of $Asp280_{hum}$ is involved in an H-bond with a water molecule, which also interacts with the very close $Thr183_{hum}$, and this interaction is important for keeping the substrate in place. In the rat active site, the residue corresponding to the $Thr183_{hum}$ is a $Ser183_{rat}$, which is not involved in the direct docking of the substrate, similar to $Glu279_{rat}$, which is flipping around and not really involved in any H-bonding, although its conformation influences the substrate's position relative to the heme. This could explain why the catalytic efficiency for N-oxidation of MeIQx is much higher for human CYP1A2 than for the rat homologue [3], as the substrate is held in place with the 2-amino group as close as possible to the iron atom of the heme, with the help of

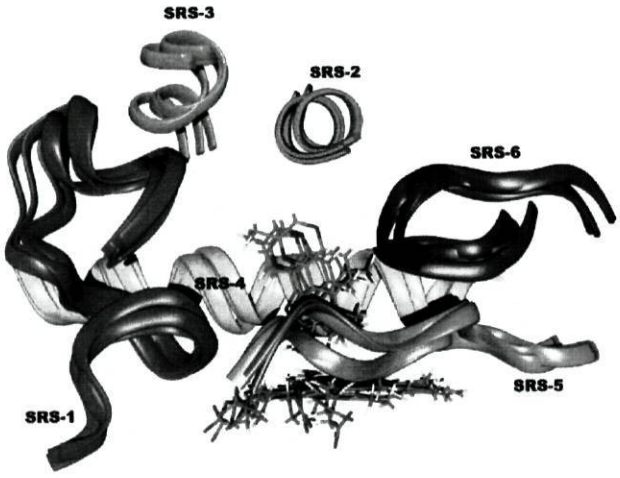


Figure 8. SRS positioning relative to the heme and substrate binding position.

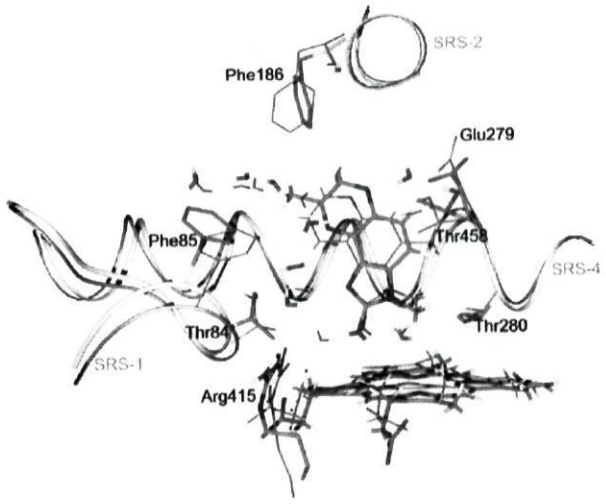


Figure 10. Rat CYP1A2 model active site docked with the MeIQ (grey) and MeIQx (black).

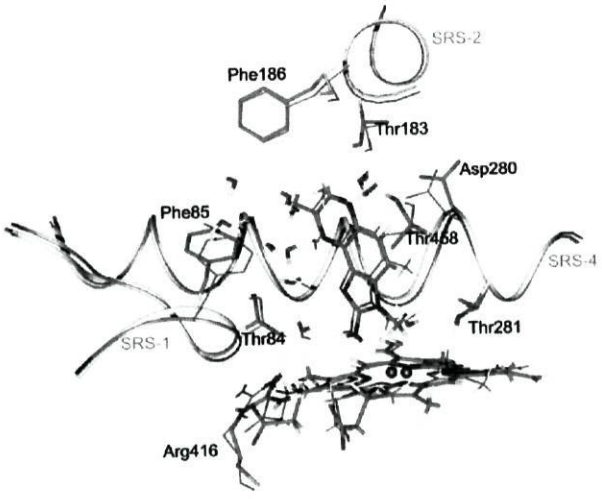


Figure 9. Human CYP1A2 model active site docked with the MeIQ (grey) and MeIQx (black).

some interaction with $Asp280_{hum}$. Figures 9 and 10 show views of the modelled human and rat CYP1A2 active sites, respectively, docked with the substrates MeIQ and MeIQx.

The bulkier side chain of $Glu279_{rat}$ actually causes some steric hindrance to the positioning of the substrate in the vertical orientation relative to the heme, which would be preferred for placing 2-amino group close to the heme. This fact can be related to some experimental results. It seems clear that the products resulting from the rat CYP1A2 action on MeIQx are the outcome of the oxidation of the central ring of the aromatic ring system of the substrate. This implies a horizontal orientation of the substrate towards the heme. On the other hand, the products resulting from the human CYP1A2 action on the same substrate have modifications on the rings at the extremities, implying a vertical orientation of the substrate towards the heme. One could speculate on the possibility that the different length of the carboxylate side chain between $Glu279_{rat}$ and $Asp280_{hum}$ is responsible for these differences in normal substrate orientation. Another role probably played by these residues has been brought up by a previous work, which related the side-chain conformation changes on the P450cam $Asp251_{cam}$ residue, corresponding to $Glu279_{rat}/Asp280_{hum}$, with a possible proton-shuttle between their solvent accessible neighbours and $Thr252_{cam}$ ($Thr281_{hum}/Thr280_{rat}$) [40], but at this stage it is not possible to relate this with the difference in catalysis rate.

Another difference is in the number of water molecules that fit the active site of each model. The rat model can fit a total of ten water molecules when complexed with MeIQ and eight water molecules when complexed with MeIQx. The human complexes of both substrates have ten water molecules each. These water molecules form a sort of H-bonded water chain that links $Arg416_{hum}/Arg415_{rat}$ to the extreme of the active site pocket, close to the F-G loop, the probable substrate entrance place. $Arg290_{cam}$, corresponding to $Arg416_{hum}/Arg415_{rat}$, has been previously suggested [41] to be involved in a functional water channel that takes away water molecules from the active centre towards a water cluster located on the thiolate side of the heme, using a change in the side chain of the arginine from the initial (stable) side-chain conformation to another rotamer (metastable). As the substrate lies almost perpendicular to the entrance tunnel, it forms a barrier between the incoming water molecules and the iron atom from the heme. These water

molecules stabilize the orientation of the substrate towards the heme and interact with the polar groups from the substrates (the amines) and the polar amino acid side chains from the active site of both models. In fact, several threonines play an important role in stabilizing the substrate through H-bonds with the water molecules that interact with it. Besides $Thr183_{hum}$, which was already mentioned, $Thr84_{hum}$ and $Thr84_{rat}$ interact with the amine groups closer to the heme, N1 and the 2-amine group (the oxidation of this amine group will produce the carcinogenic product). Additionally, residues $Thr281_{hum}/Thr280_{rat}$, both corresponding to $Thr252_{cam}$, have been considered important as it is towards these residues that the water molecule that is initially bound to the heme moves. This is in agreement with the existence of an internal water channel that also involves $Glu427_{hum}/Glu426_{rat}$, which has been suggested to withdraw water molecules from the active site as the substrate enters it [40]. Located on the edge of SRS6 (see figure 8), $Thr458_{hum}/Thr458_{rat}$ are also a part of the H-bonded chain that holds the substrate in place. This 'arm' seems to have some mobility, and it seems to regulate the overall volume of the active site. This means it could have a different conformation on a substrate-free structure.

Besides the residues whose importance was already focused on by previous research (table 3), the models evidence the relevance of some other active site residues. $Phe85_{hum}$ and $Phe85_{rat}$ and $Phe341_{hum}/Phe340_{rat}$, together with $Phe186_{hum}$ and $Phe186_{rat}$, limit the ligand mobility by stereochemical hindrance and electrostatic group repulsion. They are expected to be very important for optimal ligand orientation in the active site. $Phe186_{hum}$ and $Phe186_{rat}$ are responsible for the different orientations of the planes of the two ligands observed, as they clash with the methyl group from C8 on MeIQx and push it over to Phe85 and Thr458. Two other non-polar amino acid residues flank the substrate in the active site, $Leu342_{hum}/Val341_{rat}$ and $Ile346_{hum}/Ile345_{rat}$. $Leu342_{hum}/Val341_{rat}$ is located next to $Phe341_{hum}/Phe340_{rat}$, restraining the phenyl group movements along with the substrate mobility in this area. $Ile346_{hum}/Ile345_{rat}$ is very close to $Arg416_{hum}/Arg415_{rat}$ and $His348_{hum}/His347_{rat}$, both involved in heme binding through interaction with its propionate groups.

When we compare the model presented here with the previously reported model made for the human CYP1A2 [1] using all the templates except for the rabbit chimeric 2C3-2C5, several differences are readily noticed. Although in the previous model the two very conserved regions of SRS-4 and the Cys-pocket present RMSD values for the main-chain atoms in a range of 0.5 to 1.6, other regions that define the active site show significant deviation. On SRS-1 there is an extra secondary structure element, a helix corresponding to HB′ that is seen in some of the bacterial templates. The alignment corresponding to this region is quite different for the two models, and the use of 1DT6 as a template (which does not have this secondary structure element in this region) resulted in a very different conformation for this location. Furthermore, the region that includes SRS-2 and extends from the end of HE to the beginning of HG (thus including the probable region for the entrance of the substrate in the active site) shows very high RMSD values when compared to our model, mostly because of a different alignment which resulted on a different location in space of SRS-2 relative to the active site; this reflects differences in the docking of substrates, for example the fact that the amino acid side chain of Asn182 shows H-bond interactions with the substrate MeIQ in the previous model whereas in our model it is not directly involved in any contact with this substrate. Another major difference is located in SRS-6, which in our model is a very important region for the binding of the substrate. In the previously reported model, this region is more buried in the active site, probably limiting the movements of the substrate. This difference seems to be a result of the very different N-terminal that the previous model presents, which pushes SRS-6 into the active site. This is not surprising, as it is a region involved in membrane binding, and therefore the use of the mammalian rabbit chimeric 2C3-2C5 resulted in a different conformation for this site than that predicted, using as templates proteins that are not membrane bound. These differences in structure take place not only because of different local alignments, but mostly because of the use of the now available rabbit CYP2C3-2C5 X-ray structure, which has undoubtedly contributed in a very positive way to the building of a new generation of mammalian CYP models.

The models we have presented should be good tools for studying further the behaviour of CYP1A2 in human and rat as, after following a rational modelling procedure they end up confirming experimental data from various sources. However, they have to be used only as working hypotheses, challenged against further experimental data, and modified or updated as soon as additional experimental information is available.

We thank the NFCR (National Foundation for Cancer Research) Centre for Drug Discovery, University of Oxford, UK, for financial support, and the FCT (Fundação para a Ciência e Tecnologia) for a doctoral scholarship for Rute Fonseca.

References

[1] De Rienzo, F., Fanneli, F., Menziani, M. C., and De Benedetti, P. G., 1999, *J. comput.-aided molec. design*, **14**, 93, and references therein.

[2] WILLIAMS, P. A., COSME, J., SRIDHAR, V., JOHNSON, E. F., and MCREE, D. E., 2000, *Molec. Cell*, **5,** 121, and references therein.
[3] LANGÖET, S., WELTI, D. H., KERRIDUY, N., FAY, L. B., HUYNH-BA, T., MARKOVIC, J., GUENGERICH, F. P., GUILLOUZO, A., and TURESKY, R. J., 2001, *Chem. Res. Toxicol.*, **14,** 211.
[4] TURESKY, R. J., CONSTABLE, A., FAY, L. B., and GUENGERICH, P., 1999, *Cancer Lett.*, **143,** 109.
[5] GOTOH, O., 1992, *J. biol. Chem.*, **267,** 83.
[6] FELTON, J. S., KNIZE, M. G., HATCH, F. T., TANGA, M. J., and COLVIN, M. E., 1999, *Cancer Lett.*, **143,** 127, and references therein.
[7] GARNER, R. C., LIGHTFOOT, T. J., CUPID, B. C., RUSSELL, D., COXHEAD, J. M., KUTSCHERA, W., PRILLER, A., ROM, W., STEIER, P., ALEXANDER, D. J., LEVESON, S. H., DINGLEY, K. H., MAUTHE, R. J., and TURTELTAUB, K. W., 1999, *Cancer Lett.*, **143,** 161, and references therein.
[8] SINHA, R., and ROTHMAN, N., 1999, *Cancer Lett.*, **143,** 189, and references therein.
[9] IBA, M. M., and FUNG, J., 2001, *Biochem. Pharmacol.*, **62,** 617.
[10] WEI, C., CACCAVALE, R. J., KEHOE, J. J., THOMAS, P. E., and IBA, M. M., 2001, *Cancer Lett.*, **171,** 113.
[11] BOBBIS, A. R., LYNCH, A. M., MURRAY, S., DE LA TORRE, R., SOLANS, A., FARRÉ, M., SEGURA, J., GOODERHAM DAVIES, D. S., 1994, *Cancer Res.*, **54,** 89.
[12] TURESKY, R. J., CONSTABLE, A., RICHOZ, J., VARGA, N., MARKOVIC, J., MARTIN, M. V., and GUENGERICH, F. P., 1998, *Chem. Res. Toxicol.*, **11,** 925.
[13] LOZANO, J. J., PASTOR, M., CRUCIANI, G., GAEDT, K., CENTENO, N. B., GAGO, F., and SANZ, F., 2000, *J. comput.-aided molec. Design*, **14,** 341.
[14] JOSEPHY, P. D., BIBEAU, K. L., and EVANS, D. H., 2000, *Env. molec. Mutagenesis*, **35,** 328.
[15] CHOTHIA, C., and LESK, A. M., 1986, *EMBO J.*, **5,** 823.
[16] SEVRIOUKOVA, I. F., LI, H., ZHANG, H., PETERSON, J. A., and POULOS, T. L., 1999, *Proc. Natl. Acad. Sci. USA*, **96,** 1863.
[17] BODDUPALLI, S. S., HASEMAN, C. A., RAVICHANDRAN, K. G., LU, J.-Y., GOLDSMITH, E., DEISENHOFER, J., and PETERSON, J. A., 1992, *Proc. Natl. Acad. Sci. USA*, **89,** 5567.
[18] WILLIAMS, P. A., COSME, J., SRIDHAR, V., JOHNSON, E. F., and MCREE, D. E., 2000, *J. inorg. Biochem.*, **81,** 183.
[19] KIRTON, S. B., BAXTER, C. A., and SUTCLIFFE, M. J., 2002, *Adv. Drug Delivery Rev.*, **54,** 385.
[20] DAI, R., PINCUS, M. R., and FRIEDMAN, F. K., 2000, *Cell. molec. Life Sci.*, **57,** 487.
[21] LEWIS, D. F. V., 2002, *J. inorg. Biochem.*, **91,** 502.
[22] SZKLARZ, G. D., and PAULSEN, M. D., 2002, *J. biomolec. Struct. Dyn.*, **20,** 155.
[23] ALTSCHUL, S. F., MADDEN, T. L., SCHAFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W., and LIPMAN, D. J., 1997, *Nucleic Acids Res.*, **25,** 3389.
[24] http://genopole.toulouse.inra.fr/blast/blast.html
[25] BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N., and BOURNE, P. E., 2000, *Nucleic Acids Res.*, **28,** 235.
[26] http://www.pdb.org/
[27] BAIROCH, A., and APWEILER, R., 2000, *Nucleic Acids Res.*, **28,** 45.
[28] THOMSON, J. D., HIGGINS, D. G., and GIBSON, T. J., 1994, *Nucleic Acids Res.*, **22,** 4673.
[29] http://www.ebi.ac.uk/clustalw
[30] http://www.accelrys.com
[31] THOMPSON, J. D., GIBSON, T. J., PLEWNIAK, F., JEANMOUGIN, F., and HIGGINS, D. G., 1997, *Nucleic Acids Res.*, **25,** 4876.
[32] KELLEY, L. A., MACCALLUM, R. M., and STERNBERE, M. J. E., 2000, *J. molec. Biol.*, **299,** 499.
[33] PARIKH, A., JOSEPHY, P. D., and GUENGERICH, F. P., 1999, *Biochemistry*, **38,** 5283.
[34] SALI, A., and BLUNDEN, T. L., 1993, *J. molec. Biol.*, **234,** 779.
[35] LASKOWSKI, R. A., MACARTHUR, M. W., MOSS, D. S., and THORNTON, J. M., 1993, *J. appl., Crystallogr.*, **26,** 283.
[36] BOWIE, J. U., LÜTHY, R., and EISENBERG, D., 1991, *Science*, **253,** 164.
[37] GUEX, N., and PEITSCH, M. C., 1997, *Electrophoresis*, **18,** 2714.
[38] http://www.expasy.ch/spdbv/
[39] BROOKS, B. R., BRUCCOLERI, R. E., OLAFSON, B. D., STATES, D. J., SWAMINATHAN, S., KARPLUS, M., 1983, *J. comput. Chem.*, **4,** 187.
[40] GERBER, N. C., and SLIGAR, S. G., 1992, *J. Am. chem. Soc.*, **114,** 8742.
[41] OPREA, T. I., HUMMER, G., and GARCIA, A. E., 1997, *Proc. Natl. Acad. Sci. USA*, **94,** 2133.

1.2. Computational insight into anti-mutagenic properties of CYP1A flavonoid ligands

Medicinal Chemistry, 2005, *1*, 355-360

Computational Insight into Anti-mutagenic Properties of CYP1A Flavonoid Ligands

Rute da Fonseca[&], Michele Marini[‡], André Melo[&], Maria Cristina Menziani[‡] and Maria João Ramos*[,&]

[&]*Departamento de Química, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal and* [‡]*Dipartamento di Chimica, Università di Modena and Reggio E., Via Campi 183, 41100 Modena, Italy*

Abstract: Cytochrome P450 1A (CYP1A) is a subclass of enzymes involved in the biotransformation of heterocyclic amines present in cooked red meat to carcinogenic compounds. Anti-cancer properties have long been associated with flavonoids, and some compounds of this class have been shown to interact directly with CYP1A2. The understanding of this interaction is the purpose of this work. As the number of experimentally tested molecules is limited, two complementary methods in terms of information provided, are proposed for the study of protein-inhibitor interaction as alternatives to a QSAR analysis, using quantum mechanics as well as molecular mechanics.

Key Words: Binding free energy, cancer, CYP1A2, cytochrome P450, flavonoids, molecular mechanics, quantum mechanics.

INTRODUCTION

Cytochrome P450 (CYP) is a large class of enzymes which represents 12% of all the human cytochrome. These enzymes are responsible for transforming xenobiotic substances into products easier to remove from the body. Other important roles of CYPs are the synthesis of steroid hormones thromboxane, cholesterol and bile acid and the degradation of endogenous compounds such as fatty acids, retinoic acids and steroids [1].

Despite the benefits of their actions, CYPs also produce toxins and mutagenic compounds. For example, some heterocyclic amines (HAs), which can be found in significant amounts in meat cooked at high temperatures, are N-hydroxylated by CYP class 1A enzymes (CYP1A) to toxic mutagens. These compounds can cause DNA damage, which results in the formation of tumors in a variety of tissues in several species [2-4]. Flavonoids reduce the risk of DNA damage by competing with the HAs for binding to the CYP1A active site [5-10], and therefore inhibiting its damaging catalytic activity.

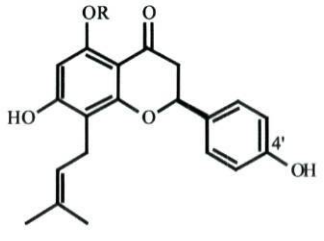
CYPs are involved in interactions with flavonoid compounds in at least three ways: (i) flavonoids induce the biosynthesis of several CYPs; (ii) enzymatic activities of CYPs are modulated (inhibited or stimulated) by these compounds; and (iii) flavonoids are metabolized by several CYPs [9].

In humans, the CYP1A inhibition mechanism by flavonoids is very complex, involving a large number of chemical and biological aspects. For a detailed description of such mechanism, it is essential to understand the molecular recognition of flavonoids by CYP1A at molecular level.

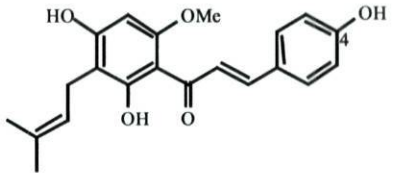
*Address correspondence to this author at the Departamento de Química, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal; E-mail: mjramos@fc.up.pt

1573-4064/05 $50.00+.00

We have carried out a theoretical study of the anti-mutagenic properties of 8–Prenylnaringenin (8-PN), Isoxanthohumol (IX) and Xanthohumol (XN) (see Fig. (**1**)). These compounds occur naturally in hops and beer and their inhibitory power towards human CYP1A2 (hCYP1A2) metabolism has been tested experimentally [11]. Using this data, it is possible to make both a qualitative and semi-quantitative evaluation of the ligands inhibitory potency based on structural data. By analyzing the electrostatic potential of all the ligands, we can indicate which features are most likely to be related to a higher inhibitory power, either because these are involved in molecular recognition processes or in stabilization at the active site.



R = H for 8-Prenylnaringenin (**8-PN**)
R = CH3 for Isoxanthohumol (**IX**)



Xanthohumol(**XN**)

Fig. (1). Flavonoids used in this study.

This last characteristic can be further explored by the assessment of the interaction energy between specific residues in the active site and the ligand obtained by means of enzyme-ligand complex calculations [12,13]. Actually, it is possible to compare the affinities of different ligands towards the same enzyme by calculating appropriated thermodynamic quantities. In this work, two approaches have been used to correlate the inhibitory power determined experimentally with calculated properties, using both explicit and continuum descriptions of the solvent.

MATERIALS AND METHODS

The ligands' structures were modeled using InsightII [14]. The S diastereoisomer conformation on carbon 2 was used for 8-PN and IX; this is the diastereoisomer naturally occurring in hops [15].

Molecular Electrostatic Potentials

The ligands were then minimized at the B3LYP/6-31G* level, using Gaussian98 [16]. The molecular electrostatic potentials for the optimized geometries have been generated using the CUBEGEN utility available in Gaussian98 [16]. The molecular electrostatic potential (values ranging from –56 kcal/mol to 56 kcal/mol) was mapped onto 0.002 e/bohr^3 electron density surface with Molekel 4.2 [17].

Dockings

The ligands were docked onto the active site of hCYP1A2 [12], using GOLD [18]. GOLD provides two scoring functions for analyzing the results of a search for the best binding of a ligand (allowing its total conformational flexibility) in an active site cavity by using a genetic algorithm: GoldScore, the original scoring function, optimized for predicting ligand binding positions, and ChemScore, derived from regression against ligand-receptor binding free energies [18]. We have used both of these scoring functions and the standard default settings, to get the best 5 results of 50 docking runs for each ligand. In order to get a reliable orientation, we have used a distance constraint of 5 Å between the heme's iron and the oxygen on carbon C4' in 8-PN and IX, and the correspondent carbon C4 in XN, as this seems to be the potential site of oxidation of flavonoid type structures by members of the cytochrome P450 family [19,20]. Visualization of docking results was performed with DS ViewerLite from Accelrys [14].

Stabilization Energies of Complexes

The best results of the docking were then energy minimized using CHARMM27 [21]. We have used the complete model solvated with a 9 Å layer of TIP3P explicit water molecules. Several steps of minimization have been done. First, the docking position of the ligand has been optimized together with the side-chains of the aminoacids. Next, the backbone of the enzyme was relaxed, with harmonic constrains of 2 kcal.mol^{-1} on the residues located more than 20 Å away from the ligand. In all steps, harmonic constraints have been applied to the water molecules, which are located between 6 and 9 Å away from the enzyme, to prevent the solvent from evaporating. The ligands were also geometrially optimized inside a 20 Å sphere of TIP3P explicit water molecules. Harmonic constraints have been applied to the water molecules, which are located more than 15 Å away from the centroid of the system, resulting in the same number of constrained molecules for the three ligand-water complexes.

Using the structures solvated with water molecules, the hCYP1A2/flavonoids complexes stabilization energy, $\Delta E_{aq}^{stab}(E:L)$, has been calculated as follows:

$$\Delta E_{aq}^{stab}(E:L) = E_{aq}(E:L) - E_{aq}(E) - E_{aq}(L) \quad (1)$$

$E_{aq}(E)$ is the energy of the isolated solvated enzyme, and $E_{aq}(L)$ is the energy of the isolated solvated flavonoid. Within a molecular mechanics formalism, the stabilization energy can be partitioned as

$$\Delta E_{aq}^{stab}(E:L) = \Delta E_{\text{int}}(E:L) + \Delta E_{conf}(E:L) + \Delta E_{solv}(E:L) \quad (2)$$

where is $\Delta E_{\text{int}}(E:L)$ interaction energy between the enzyme and the flavonoid within the hCYP1A2:flavonoid complex, and $\Delta E_{conf}(E:L)$ corresponds to the energy needed to change the conformation of the enzyme and ligand from the free to the complexed form. $\Delta E_{solv}(E:L)$ is the difference in energy correspondent to the interactions with the solvent between the free and the complexed form of both enzyme and ligands.

In this work, the most significant specific interfragment interactions responsible for the stabilization for the hCYP1A2/flavonoid complexes have been evaluated using the INTER utility available in CHARMM27 [21].

Binding Free Energies of the Complexes

The binding free energy of the same complexes calculated, combining the use of molecular mechanics and a classical continuum solvation approach. The binding free energy can be defined as [22]:

$$\Delta G_{bind} = G_{aq}(E:L) - \left[G_{aq}(E) + G_{aq}(L)\right] \quad (3)$$

where $G_{aq}(E:L)$, $Gaq(E)$ and $Gaq(L)$ correspond to Gibbs free energies. The Gibbs free energy of a generic species is [22]:

$$G_{aq} = G_{gas} + G_{solv} \quad (4)$$

In equation [4], G_{gas} is the binding free energy of the species in gas phase and G_{solv} is its solvation free energy. G_{gas} corresponds to the sum of the energy of the species in gas phase ($E_{gas} = \Delta E_{\text{int}}(E:L) + \Delta E_{conf}(E:L)$) with the correspondent entropic contribution. The latter can be regarded as non-differential when comparing two complexes, as it refers to the process of association of similar ligands to the same protein [22,23] and will not be considered. The solvation entropic effects are included in the continuum method used to solvate our systems.

G_{solv} can be partitioned into polar and non-polar terms:

$$G_{solv} = G_{solv}^{polar} + G_{solv}^{nonpolar} \quad (5)$$

The polar solvation free energy was calculated by solving the Poisson-Boltzmann equation with the Delphi program [24]. The nonpolar component was calculated using the solvent accessible surface area as follows [25]:

$$\Delta G_{solv}^{nonpolar} = \gamma A + b \quad (6)$$

where A is the solvent accessible surface area of the species calculated in InsightII [14] using a 1.4 Å radius probe. γ and *b* are 0.00542 kcal.mol^{-1} and 0.92 kcal.mol^{-1}, respectively [25].

All the energy values will be presented as normalized with respect to the weakest ligand, XN.

RESULTS AND DISCUSSION

In this study, we present different approaches to the ligand binding efficiency prediction, when the number of molecules tested experimentally is too small for a QSAR type of study.

The experimental inhibitory activity data available measures the degree of inhibition of the mutagenesis of 2-amino-3-methylimidazo-[4,5-f]quinoline (IQ) in a Salmonella Assay induced by a concentration of 10μM of flavonoid. The experimental data was obtained by *in vitro* analysis with human DNA recombinant CYP1A2, guaranteeing the absence of interference of other enzymes in the process [11].

A first step towards the understanding at molecular level of the experimental results was the characterization of the molecular electrostatic recognition pattern of the flavonoids. These represent the electrostatic potential profile of the ligand, which allows the identification of enzyme-ligand interaction sites based on the concept of electrostatic complementarity. These interactions may be important as molecular recognition features when the ligand approaches the enzyme, and for electrostatic complementarities between the docked ligand and the active site.

The optimized geometries of the ligands obtained at B3LYP/6-31G* level are presented in Fig. (**2**), together with the respective molecular electrostatic potentials mapped onto an electron density surface. It can be observed that all the flavonoids present common patterns in the electrostatic potential surface. In fact, the most negative potentials are located on the oxygen atoms of carbonyl, hydroxyl and methoxy groups. These will be hydrogen bonding spots, as can be seen in Fig. (**3**), which will stabilize the ligands in the complex (see Table **1**). One marked difference between the two best inhibitors and XN is the presence of strong negative potential spots all around the molecule in the latter, while for the first two, these spots are located only in one side, which is the upper side in Fig. (**2**). This will be important because the lower side will be facing Phe_{125}, which shows a highly stabilizing van der Waals interaction with the ligands (the van der Waals term represents over 90% of the total interaction energy for this residue).

Next, an appropriate docking of the inhibitors into the active site using an automated approach was done, followed by a geometry optimization of the complexes using molecular mechanics. The results of this procedure were used in two ways with the aim of relating the experimentally determined inhibitory power of the flavonoids with an appropriate thermodynamic quantity. One of them implies the use of explicit water molecules to solvate ligand and enzyme. This procedure enables the description of interactions for individual atoms or sets of atoms. A detailed description of the specific interactions responsible for the stability of the enzyme-inhibitor complexes has been carried out with the aim of designing possible anti-mutagenic compounds in future works. Using this approach, the stabilization energy of the complexes was calculated, based on the molecular mechanics energy obtained for the solvated

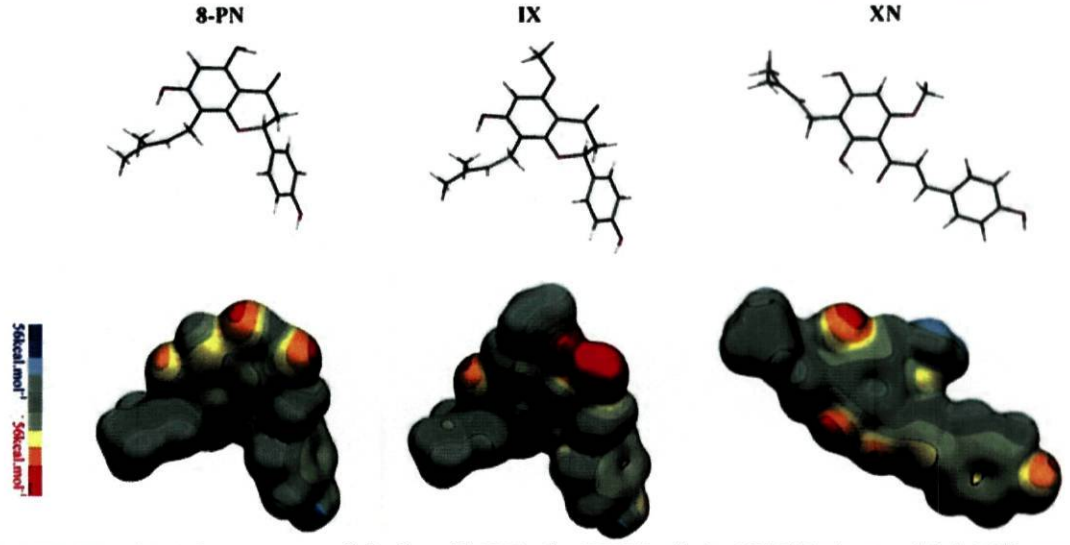


Fig. (2). 8-PN, IX and XN after geometry optimization with DTF using B3LYP with the 6-31G* basis set and their MEPs mapped onto a 0.002e/borh3 electron density surface.

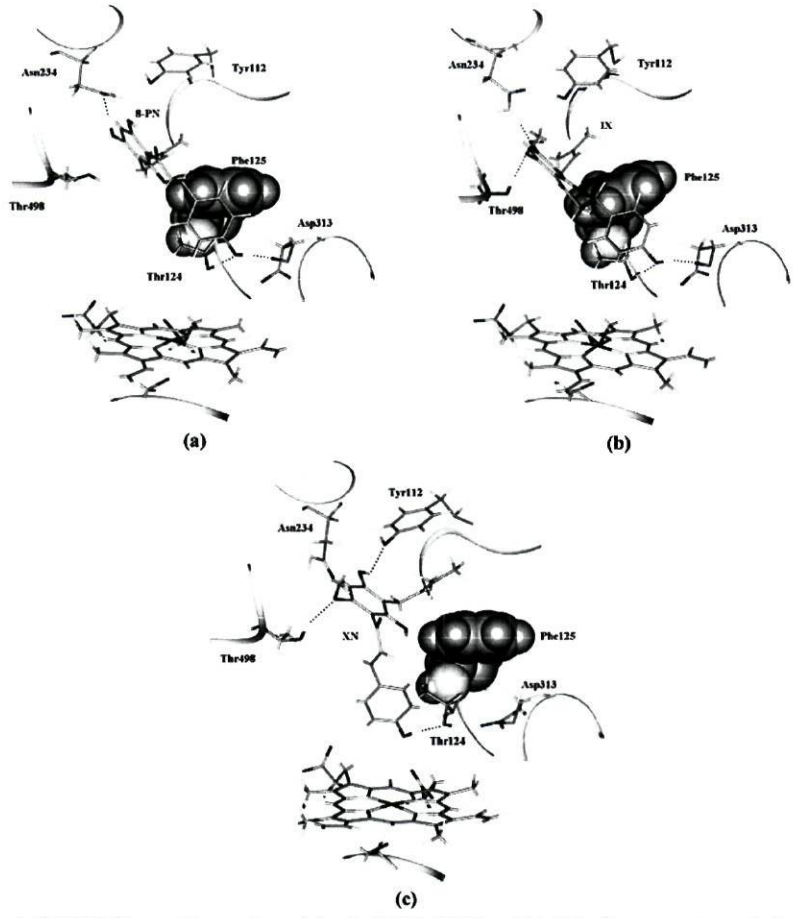


Fig. (3). Active site of the hCYP1A2/flavonoid complexes (a) with 8-PN, (b) IX and (c) XN after geometry optimization. The residues which present the most significant interaction energies with the ligands (see Table **1**) are shown. Hydrogen bonding is indicated by grey lines.

complexes and ligands. This quantity was well correlated with the experimentally determined inhibitory power of the flavonoids. The main interactions of the complexes can be analyzed using the decomposition of $E_{int}(E:L)$ into the appropriate components (Table **1**).

Considering the interactions between the flavonoids and the residues of hCYP1A2, it can be observed that all three flavonoids share a large number of specific interactions (Tyr_{112}, Thr_{124}, Phe_{125}, Asp_{313}, Asn_{234} and Thr_{498}, see Table **1**). These interactions represent 50% of the total interaction energy between the flavonoids and the rest of the system ($E_{int}(E:L)$). Additionally, the interactions involving Val_{227}, Ala_{230}, Ser_{231}, Ile_{386}, Leu_{497} have Met_{499} have a significant role in the stabilization of the complexes, representing around 20% of the total $E_{int}(E:L)$. A very strong interaction common to the three ligands corresponds to a hydrogen bond between the side chain of Thr_{124}, deeply buried in the active site and the hydroxyl substituent in carbon C4 for 8-PN and IX or carbon C4' in XN (see Fig. (**3**)). This substituent is further stabilized with an interaction with the oxo group from the backbone of Asp_{313} in 8-PN and IX. This interaction is not present for the complex with XN, and instead, this compound presents the highest single aminoacid interaction energy contribution to $E_{int}(E:L)$, with a residue located at the entrance of the active site, Thr_{498}.

Moreover, XN is disfavored in relation to the other two, for an already mentioned strong stabilizing interaction with Phe_{125}, common to all the ligands. The ligands structure is wrapped around the side-chain of this residue, with the prenyl substituent almost parallel to one side of the aromatic ring and the phenyl substituent coming down to the other face of the ring (see Fig. (**3**)), towards the interaction with Thr_{124} and Asp_{313}. Although the fit of XN in the active site leads to a higher total $E_{int}(E:L)$ (see Table **2**), this does not necessarily imply a higher stabilization of this ligand inside the enzyme. Stabilizing interactions that are related to a higher residence time such as those with the residues buried in the cavity Thr_{124} and Asp_{313}, or the anchor-like effect of the interaction with Phe_{125}, are less important to the total $E_{int}(E:L)$ when compared to the other two ligands.

Another disadvantageous property of the complex between hCYP1A2 and XN is the high energy cost of the conformational changes that both ligand and enzyme undergo to form the complex when compared to the other two (see Table **2**). This component of the stabilization energy of the hCYP1A2/flavonoid complexes is correlated

Table 1. Most Significant Interaction Energies (in Bold, >5kcal.mol^{-1}) Between Individual Residues of the Human CYP1A2 Active Site and 8-PN, IX and XN

	8-PN		IX		XN	
Aminoacid residue	E_{int} (kcal.mol^{-1})	Interaction	E_{int} (kcal.mol^{-1})	Interaction	E_{int} (kcal.mol^{-1})	Interaction
Tyr 112	**-5.0**	3,5A electrostatic interaction	**-5.4**	3A electrostatic interaction	**-6.3**	H-bond
Thr 124	**-6.3**	H-bond	**-10.0**	H-bond	**-10.5**	H-bond
Phe 125	**-5.0**	Ligand folds around its side-chain	**-5.8**	Ligand folds around its side-chain	-3.1	Ligand folds around its side-chain
Asn 234	**-10.5**	H-bond	**-5.5**	H-bond	-2.2	
Asp 313	**-8.5**	H-bond	**-7.4**	H-bond	-3.8	
Thr 498	-1.4		**-7.8**	H-bond	**-14.6**	H-bond

with the experimentally determined inhibitory activities, and will be determinant for the overall stabilization of the complex. The large methoxy group that distinguishes structure IX from the best inhibitor's will be responsible for its lower inhibitory power because of the enzyme's conformational rearrangement term, although it actually contributes to a docking position that provides an extra stabilizing hydrogen bond interaction when compared to 8-PN (see Fig. (**3**)). XN is disfavored in all stabilization energy components.

The second method described here deals with the solvent as a continuum. In this case, it is possible to calculate the binding free-energy of the complexes, and this was also correlated with the inhibitory power of the flavonoids. There was a full agreement between the relative $\Delta G_{bind}(E:L)$ and the experimental inhibition results (see Table **3**).

These values were strongly dependent on the gas phase energy component. In fact, $\Delta G_{solv}(E:L)$ actually favored XN in relation to the other ligands, but $\Delta E_{gas}(E:L)$ was decisive for the final $\Delta\Delta G_{bind}(E:L)$ value.

CONCLUSION

We have presented two different approaches complementary in terms of information provided, to study the ligand binding problem. One used an atomistic description of the solvent, and allowed us to calculate the stabilization energy of the complexes. The other includes the use of continuum methods in a classical approach, allowing us to calculate the correspondent binding free-energies. Both correlated with the experimental results for inhibition of hCYP1A2. The inhibitory power is strongly correlated with the conformational rearrangement energies. There are also some specific structural features of the ligands contributing to a higher binding energy. One is the presence of small electronegative groups bound to the atoms that sit between C8 and the oxo group of flavanone-like structures which participate in stabilizing hydrogen bond type interactions with Tyr_{112}, Asn_{234} and Thr_{498}. The prenyl tail in the compounds helped stabilizing the complex through non-polar interactions with Phe_{125}, which seems to be an important residue in the fitting of the ligands in the active site. The hydroxyl substituent in C4 in 8-PN and IX, or carbon C4' in XN should contribute for a longer residence time, as it is involved in hydrogen bonding with residues which are buried in the active site, Thr_{124} and Asp_{313}.

It is a noteworthy fact that the flavonoids studied establish strong stabilizing interactions with both Thr_{124} and

Table 2. Stabilization Energy of the Complexes Between Human CYP1A2 and 8-PN, IX and XN. All the Values are Normalized with Respect to the Weakest Ligand, XN

hCYP1A2 complex	%inhibition	$\Delta\Delta E_{int}(E:L)$ (kcal.mol^{-1})	$\Delta\Delta E_{conf}(E:L)$ (kcal.mol^{-1})	$\Delta\Delta E_{solv}(E:L)$ (kcal.mol^{-1})	$\Delta\Delta E_{stab}(E:L)$ (kcal.mol^{-1})
8-PN	94	6.6	-132.1	-416.4	-541.9
IX	84	1.1	-110.2	-211.3	-320.3
XN	48	0	0	0	0

Table 3. Binding Free Energies of the Complexes Between Human CYP1A2 and 8-PN, IX and XN. All the Values are Normalized with Respect to the Weakest Ligand, XN

hCYP1A2 complex	%inhibition	$\Delta\Delta E_{gas}(E:L)$ (kcal.mol^{-1})	$\Delta\Delta G_{solv}(E:L)$ (kcal.mol^{-1})	$\Delta\Delta G_{bind}(E:L)$ (kcal.mol^{-1})
8-PN	94	-125.6	69.4	-56.1
IX	84	-109.0	96.8	-12.2
XN	48	0	0	0

Val_{227}, which have been shown to be important for the maintenance of the catalytic activity of hCYP1A2 [26,27]. This constitutes an extra validation of the results obtained in the present work.

ACKNOWLEDGEMENTS

We thank the FCT (Fundação para a Ciência e Tecnologia) for a doctoral scholarship for R.F. and the NFCR (National Foundation for Cancer Research) Centre for Drug Discovery, University of Oxford, U.K., for financial support. M.M. was supported by a bilateral Erasmus agreement.

REFERENCE

[1] Guengerich, F. P. *Chem. Res. Toxicol.*, **2001**, *14*, 611.
[2] Boobis, A. R.; Lynch, A. M.; Murray, S.; de la Torre, R.; Solans, A.; Farre, M.; Segura, J.; Gooderham, N. J.; Davies, D. S. *Cancer Res.*, **1994**, *54*, 89.
[3] Turesky, R. J.; Constable, A.; Richoz, J.; Varga, N.; Markovic, J.; Martin, M. V.; Guengerich, F. P. *Chem Res Toxicol.*, **1998**, *11*, 925.
[4] Garner, R. C.; Lightfoot, T. J.; Cupid, B. C.; Russell, D.; Coxhead, J. M.; Kutschera, W.; Priller, A.; Rom, W.; Steier, P.; Alexander, D. J.; Leveson, S. H.; Dingley, K. H.; Mauthe, R. J.; Turteltaub, K. W. *Cancer Lett.*, **1999**, *143*, 161.
[5] Zhai, S.; Dai, R. K.; Friedman, F. K.; Vestal, R. E. *Drug Metab. Dispos.*, **1998**, *26*, 989.
[6] Zhai, S.; Dai, R.; Wei, X.; Friedman, F. K.; Vestal, R. E. *Life Sci.*, **1998**, *63*, 119.
[7] Bear, W. L.; Teel, R. W. *Anticancer Res.*, **2000**, *20*, 3609.
[8] Hodek, P.; Trefil, P.; Stiborová, M. *Chemico-Biological Interactions*, **2002**, *139*, 1.
[9] Lee, H.; Yeom, H.; Kim, Y. G.; Yoon, C. N.; Jin, C.; Choi, J. S.; Kim, B. R.; Kim, D. H. *Biochem. Pharmacol.*, **1998**, *55*, 1369.
[10] Edenharder, R.; Rauscher, R.; Platt, K. L. *Mutat. Res.*, **1997**, *379*, 21.
[11] Miranda, C. L.; Yang, Y. H.; Henderson, M. C.; Stevens, J. F.; Santana-Rios, G.; Deinzer, M. L.; Buhler, D. R. *Drug Metab. Dispos.*, **2000**, *28*, 1297.
[12] Fonseca, R.; Menziani, M. C.; Melo, A.; Ramos, M. J. *Molecular Physics*, **2003**, *101*, 2731.
[13] De Rienzo, F.; Fanelli, F.; Menziani, M. C.; De Benedetti, P. G. *J. Comput. Aided Mol. Des.*, **2000**, *14*, 93.
[14] http://www.accelrys.com .
[15] Heller, W. *Acta Horticulturae*, **1994**, *381*, 46.
[16] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery Jr, J. A.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Rega, N.; Salvador, P.; Dannenberg, J. J.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J. Gaussian 98, Revision A.11.2. **2001**. Pittsburgh PA, Gaussian, Inc.
[17] Portmann, S.; Luthi, H. P. *Chimia*, **2000**, *54*, 766.
[18] Jones, G.; Willet, P.; Glen, R. C.; Leach, A. R.; Taylor, R. *J. Mol. Biol.*, **1997**, *267*, 727.
[19] Otake, Y.; Walle, T. *Drug Metab. Dispos.*, **2002**, *30*, 103.
[20] Kuffel, M. J.; Schroeder, J. C.; Pobst, L. J.; Naylor, S.; Reid, J. M.; Kaufmann, S. H.; Ames, M. M. *Mol. Pharmacol.*, **2002**, *62*, 143.
[21] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. J. *Comp. Chem.*, **1983**, *4*, 187.
[22] Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. *Acc. Chem. Res.*, **2000**, *33*, 889.
[23] Ramos, M. J.; Fernandes, P. A. *Curr. Comp-Aided Drug Des.*, **2004**, submitted.
[24] Rocchia, W.; Alexov, E.; Honig, B. *J. Phys. Chem. B,* **2001**, *105*, 6507.
[25] Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.*, **1994**, *98*, 1978.
[26] Liu, J.; Ericksen, S. S.; Sivaneri, M.; Besspiata, D.; Fisher, C. W.; Szklarz, G. D. *Arch. Biochem. Biophys.*, **2004**, *424*, 33.
[27] Parikh, A.; Josephy, P. D.; Guengerich, F. P. *Biochemistry*, **1999**, *38*, 5283.

Received: 01 December, 2004 Accepted: 10 March, 2005

1.3. Molecular interactions between human CYP1A2 and flavones derivatives

Medicinal Chemistry, 2006, *2*, 000-000

Molecular Interactions Between Human Cytochrome P450 1A2 and Flavone Derivatives

Rute da Fonseca[1], André Melo[1], Francesco Iori[2], Maria Cristina Menziani[2], Maria João Ramos[1,*]

[1]*Departamento de Química, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal;* [2]*Dipartamento di Chimica, Università di Modena, Via Campi 183, 41100 Modena, Italy*

Abstract: Activation by human cytochrome P450 1A2 (hCYP1A2) of heterocyclic amines is assumed to trigger of a number of carcinogenic processes. In this work, a group of natural inhibitors of human cytochrome P450 1A2 reported in literature has been theoretically analysed. These consist of flavone hydroxylated derivatives, natural compounds that exist in plants and associated products. Different theoretical/computational tools were used to describe the specific molecular interactions between these compounds and hCYP1A2. Based on this analysis, a method is proposed for helping the selection of specific molecular features that enhance protein-inhibitor interaction.

Key Words: Flavonoids, cytochrome P450, CYP1A2, stabilization energy.

INTRODUCTION

CYPs are ubiquitous enzymes that undertake an important role in detoxification of the organism, by oxidizing xenobiotic substances such as pesticides and food additives to excretable products. They are also involved in the biosynthetic pathways of endogenous compounds such as fatty acids, retinoic acid and steroids. One negative side effect of their oxidative action is the activation of carcinogenic compounds to reactive mutagens [1-3]. For that reason, the understanding of the inhibition of this type of CYPs is a necessary step towards lowering the possibilities of such carcinogenic process to occur.

Different flavonoids have already been shown to lower the carcinogenic activity of human CYP1A2 (hCYP1A2) [4]. Flavonoids are plant secondary metabolites responsible for odor, taste and coloration. They are ubiquitous in various constituents of the human diet such as vegetables, fruit, tea and red wine. Their high antioxidant activity has been associated with prevention against diseases caused by oxidative damage and their pharmacological relevance includes also anti-inflammatory and antiviral action [4-7]. In humans, flavonoids interact with homologous enzymes in particularly with CYP1A1/1A2 isoforms [5-9]. This gives rise to yet another beneficial role attributed to these phytochemicals – inhibition of CYP1A1/1A2 activation of promutagens.

In this work, the hCYP1A2 inhibitory potency variation of a series of six flavonoids has been studied (see Fig. (**1**)). The experimental data we will be using concerns the inhibitory strength of different flavone derivatives scaled according to the correspondent IC50 values concerning the inhibition of methoxyresorufin demethylase (MROD) activity in microsomes containing c-DNA expressed hCYP1A2 [8]. The use of a set of ligands with the same basic structure and a high rigidity allows for their inhibitory strength to be correlated with a simple substitution pattern, such as the number of hydroxyl groups and their position in the conjugated rings.

The six flavonoids were shown to be competitive inhibitors towards hCYP1A2 [8]. Competitive inhibition takes place when both substrate and inhibitor compete for binding into the same active site. Geometrical characteristics, such as shape, volume and surface, involved in enzyme-ligand complementarities, have an important role in this process. The flavones used in this study are very similar from a geometrical point of view and other differential characteristics should be selected to discriminate the inhibitory potency within this group. The analysis of the electrostatic potential pattern of the ligands together with both the inhibitor-enzyme complex geometry and energy for the series of flavones offers a number of clues on the physical properties that best contribute to their inhibitory potency over hCYP1A2.

METHODS

The initial structures of the flavone derivatives studied in this work have been modelled in InsightII [10] using the crystallographic structure of 3,5,7-trihydroxyflavone taken from Cambridge Database [11]. All these structures have been subsequently optimized using the Gaussian98 package [12], at the B3LYP/6-31G* level.

Molecular electrostatic potentials (MEPs) of the ligands were generated at the B3LYP/6-31G* level. We used MOLEKEL [13] to map the electrostatic potential onto an electron density surface of 0.002 electrons/bohr^3 (generally used, corresponding to about 95% of the electronic charge) and to draw three-dimensional electrostatic potential isosurfaces. The latter are used to predict long-range interactions [14-16].

The structure for human CYP1A2 used for docking the flavone derivatives is a homology model built as described in

*Address correspondence to these authors at the Departamento de Química, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal; E-mail: mjramos@fc.up.pt

1573-4064/06 $50.00+.00

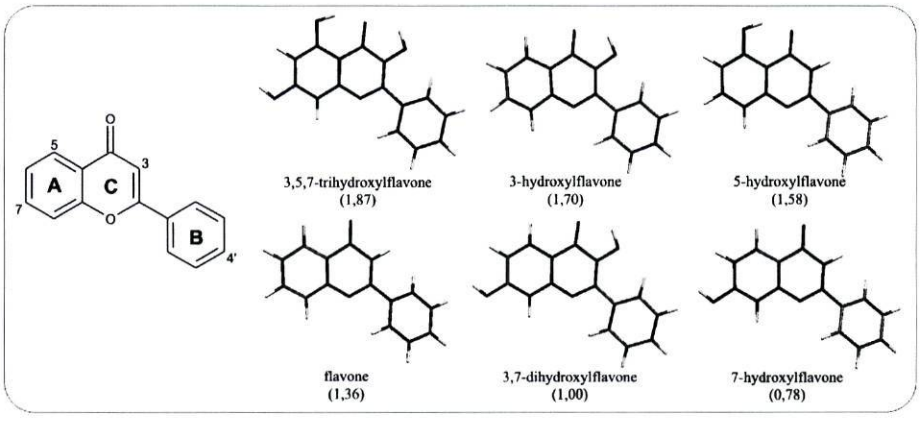


Fig. (1). The flavone derivatives studied by Zhai *et al.* - in parenthesis are the correspondent values for –log[IC50(M)].

Fonseca *et al.* [17]. The final geometry was used for automated docking of the flavone derivatives with GOLD [18], using the standard default settings. The best 5 results of 50 docking runs for each ligand were analyzed. A distance constraint of 6 Å between the most likely carbon atom to be oxidized, C4', and the iron atom [9, 19, 20] was used. Subsequently we refined the docking results. First, the docking position was geometry optimized using the complete solvated model, keeping the protein fixed, with programme CHARMM and corresponding force field [21]. Then, an energy minimization was performed allowing the side-chains to move together with the ligands. The united atom force-field parameters and 12 Å non-bonded cut-off distance were used. Solvent was treated explicitly by using an 8 Å layer of TIP3P water. Harmonic constrains were applied to the water molecules located more than 5 Å away from the enzyme. Minimizations used 500 steps of steepest descent followed by conjugate gradient.

After determining the most stable docking configurations, the affinity of a flavonoid to an enzyme active site can be evaluated by the correspondent stabilization energy. This has been calculated as follows:

$$\Delta E^{Stab}(cyp:flv) = E(cyp:flv) - [E(cyp) + E(flv)] \quad (1)$$

where, *E(cyp : flv)*, *E(cyp)* and *E(flv)* are the total energies of the hCYP1A2:ligand complex, the enzyme and the flavonoid, respectively. The stabilization energy will be presented as normalized with respect to the weakest ligand:

$$\Delta\Delta E^{Stab}(cyp:flv) = E(cyp:flv) - E(cyp:flv\_weakest) - [E(flv) - E(flv\_weakest)] \quad (2)$$

The interaction energies in the optimized complexes were determined with the INTER utility available in CHARMM [21]. This quantity, generally named ΔE^{inter}, is one of the components of the stabilization energy:

$$\Delta E^{Stab}(cyp:flv) = \Delta E^{Inter}(cyp,flv) + \Delta E^{rearr}(cyp) + \Delta E^{rearr}(flv) \quad (3)$$

In equation [3], ΔE^{rearr} *(cyp)* and ΔE^{rearr} *(flv)* are the conformational rearrangement energies for the hCYP1A2 enzyme and the flavonoid, respectively. Energy of this type is associated with the transition from the optimized geometry of the correspondent fragment to the characteristic geometry assumed by this species in the rearranged complex *CYP:flv*. In the same equation, ΔE^{rearr} *(cyp,flv)* is the interaction energy between the enzyme and the flavonoid within the same rearranged complex. This quantity can be calculated for any given pair of fragments of the complex (enzyme/ligand, heme/ligand, etc). We will present this quantity normalized with respect to the weakest ligand, $\Delta\Delta E^{inter}$.

RESULTS AND DISCUSSION

Electrostatic Potentials

Molecular recognition processes are usually involved in the approach of the ligands to the active site entrance and its subsequent binding. In this context, the electrostatic pattern recognition has been demonstrated to have a crucial role [14, 22-24]. By observing the molecular electrostatic pattern it is possible to detect the potential sites for H-bonding and other noncovalent interactions formation, which could be very important for a correct orientation of the inhibitor inside the enzyme [14-16, 24]. The local minima in the potential surface corresponds to areas, which are susceptible of electrophilic attack while the regions predisposed to nucleophilic interactions are only recognizable when displayed at a certain distance from the nucleus (the highest positive peak of electrostatic potential in the molecule), which is why we have displayed the MEP mapped onto a molecular surface.

Fig. (**2**) shows the electrostatic potential mapped onto an electron density isosurface (0.002 electrons/bohr^3) of the flavone hydroxylated derivatives studied experimentally by Zhai *et al.*, those of the new ligands drawn in this study, and of one aminoflavone substrate. The regions of lowest electrostatic potential are in red and the peaks of positive electrostatic potential are in blue. The regions of lower electrostatic potential seem to carry the features which will be determinant for molecular recognition and are represented separately in the same figure. The aminoflavone derivative is presented as a model for the features that are important for enzyme-ligand complementarity, as it is a molecule that has a high specificity for hCYP1A2 [20].

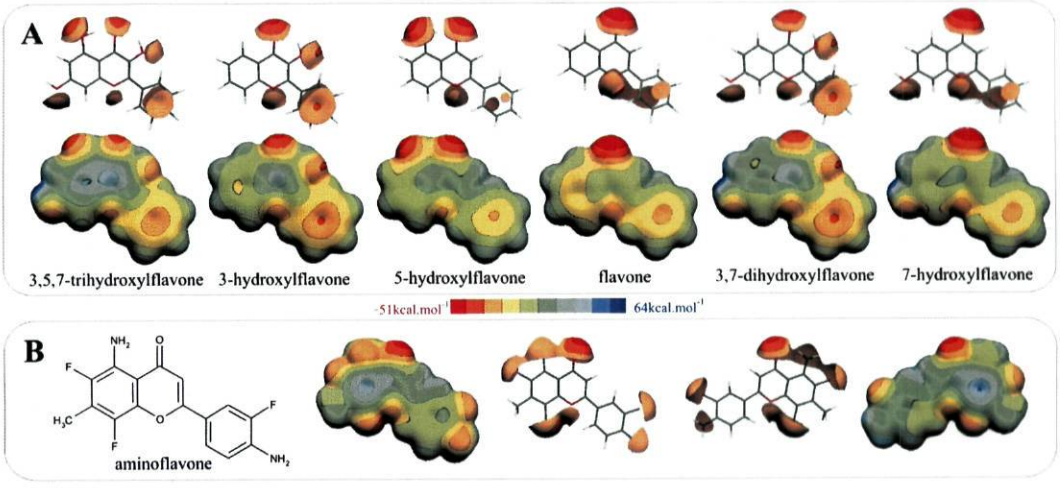


Fig. (2). Electrostatic potential mapped onto an electron density isosurface (0.002 electrons/bohr3). The regions of lowest electrostatic potential are represented separately in the second row for each particular set of compounds. (a) Compounds studied experimentally by Zhai *et al.* (b) Specific aminoflavone substrate for hCYP1A2.

The fact that the B ring from these flavones has no electronegative substituent and that phenyl rings have a very low reactivity will definitely contribute to their inhibitory properties. The common feature between the best inhibitor of the set and the aminoflavone is the spreading of the lowest potential peak between the oxo group and the hydroxyl groups on carbons C3 and C5 in the flavones and also for the fluoride bound to carbon C6 in the aminoflavone. This delocalization seems to be a molecular recognition feature related to an increase in inhibitory power by means of specific interaction with the active site of hCYP1A2. If we compare 3-hydroxylflavone and 5-hydroxylflavone with flavone, and 3,5,7-trihydroxylflavone with 5-hydroxylflavone/3-hydroxylflavone/3,7-dihydroxylflavone, this seems obvious. If this negative potential area is wide, meaning there are hydroxyl substituents on both C3 and C5, then another negative potential spot subsequent to C5 will contribute to a higher inhibitory power. Actually, in the aminoflavone, there is a methyl group in C7, correspondent to a positive potential area, followed by a fluoride group, which corresponds to negative potential. This combination seems to favor a good fit in the active site, and should be related to this compound's specificity towards hCYP1A2.

The electrostatic potential isosurfaces at 7 kcal mol^{-1} (white) and -7 kcal mol^{-1} (grey) for the flavone hydroxylated derivatives are represented in Fig. (**3**). Also in these maps we can see that the spreading of the negative density towards the A ring correlates with higher inhibition power. Here it is more evident that the best inhibitors have, on the A ring, a small positive density spot in between two negative regions. In the aminoflavone substrate, the same spot corresponds to a methyl substituent between the two fluoride substituents on positions 6 and 8 and in 3,5,7-trihydroxylflavone it is related to the hydrogen atom on C6.

In the following discussion it will be shown how besides playing a role in molecular recognition these features are involved in complex stabilization.

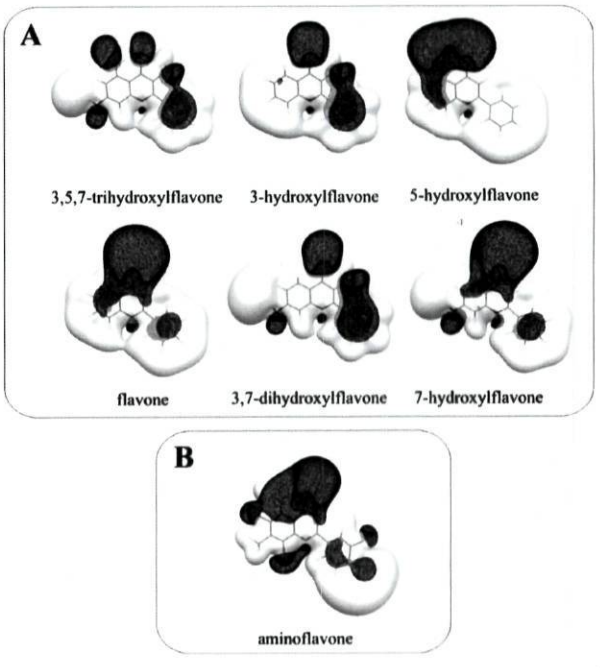


Fig. (3). Electrostatic potential isosurfaces at 7 kcal mol^{-1} (white/transparent) and -7 kcal mol^{-1} (grey/chickenwire) for the flavone hydroxylated derivatives. (a) Compounds studied experimentally by Zhai *et al.* (b) Specific aminoflavone substrate for hCYP1A2.

Docking

The results of docking optimization are shown schematically in Fig. (**4**) and (**5**).

The main points for electrostatic interaction between ligand and enzyme are the hydrogen bonds between the side-chains of Thr_{498}, Tyr_{112} and Asn_{234} and both the hydroxyl substituents of the flavone derivatives and the oxo group on

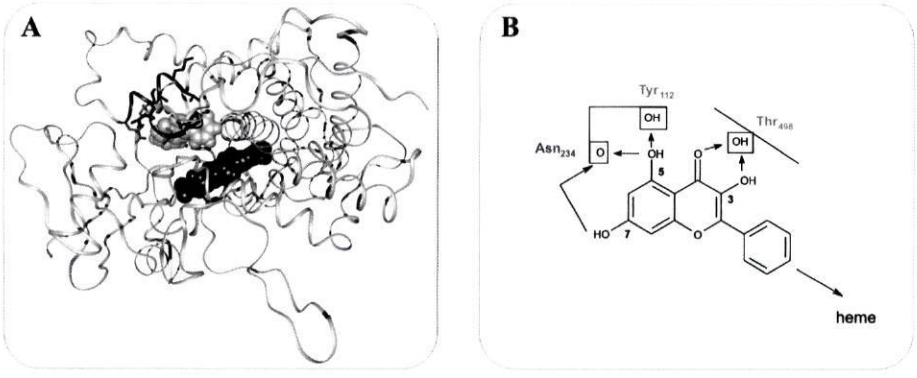


Fig. (4). (a) C-alpha trace of the model of hCYP1A2 with 3-hydroxylflavone docked in the active site. The trace in black corresponds to the entrance of the active site area and the side-chains of the corresponding residues are shown in stick. The side-chains of the aminoacids residues that make H-bonds with the different flavone derivatives are shown in colors: green for Tyr112, blue for Asn234 and Thr498 is colored in red. The side-chains of Val227, Glu228, Ala230, Ser231, Gly233, Leu497, and Met499 are in black. The heme is represented in black/CPK and 3-hydroxylflavone is represented in white/CPK. The hydrogen bonds formed between the different flavone derivatives and the enzyme are shown schematically in (b).

the carbon atom C4 (see Fig. (**4**)). The interaction between the oxo group on C4 and Thr_{498} is a common feature in all the flavone derivatives complexes. This substituent represents the most electronegative area on the ligands MEP and should be a major spot for molecular recognition. The fact that Thr_{498} is located close to the surface of the active site (see Fig. (**4**)) supports this idea.

Also Asn_{234} and Tyr_{112} are located on the top of the active site, and together with Thr_{498} should be one of the first residues of the active site to make contact with the ligand. The spreading of the negative electrostatic potential of the oxo group on C4 towards the hydroxyl groups in ring A, which seems to be a common feature among the best inhibitors, is also related to a higher stabilization of the ligands by hydrogen bonding of the latter and the neighbouring aminoacid sidechains of Asn_{234} and Tyr_{112} as can be seen in Fig. (**5**). The water molecules, which are located close to the entrance of the active site are also shown. It is noteworthy that the ligands, which have an hydroxyl group on C7 and none on C5 present a less buried docked conformation with the substituent on C7 leaning towards the solvent molecules.

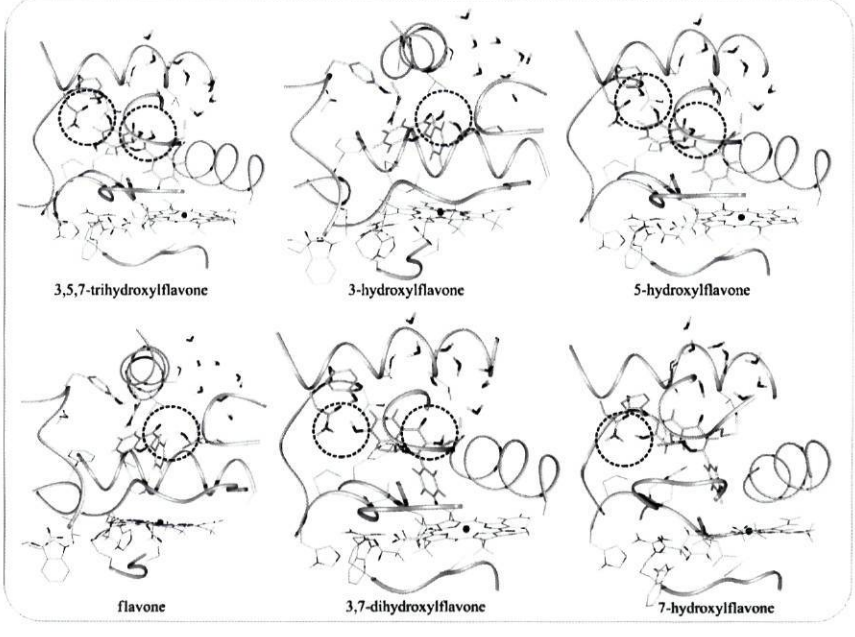


Fig. (5). Flavone hydroxylated derivatives docked in the active site of hCYP1A2. The circles show the H-bonds schematically represented in Fig. (4b).

Table 1. IC_{50} Values for the Inhibitors Used in this Work Relative to Their Activity Over the *O*-demethylation of Methoxyresorufin by hCYP1A2 Determined by Zhai *et al.*, and the Stabilization Energies of Their Complexes with Human CYP1A2. Quantity $\Delta\Delta E^{Inter}$, Which is the Part of the Total Energy that Shows the Interaction Between Individual Fragments of the Complex, is Presented for the Interactions Between the Flavone Derivatives and hCYP1A2 ($\Delta\Delta E^{Stab}$(cyp:flv)), Thr_{498} ($\Delta\Delta E^{Inter}(Thr_{498}, flv)$), and for the Group of Residues Located on the Entrance of the Active Site Shown in Fig. (4) ($\Delta\Delta E^{Inter}$(actsite_top, flv)). All Values are Normalized with Respect to the Weakest Inhibitor, 7-hydroxylflavone.

Flavonoid (-log[IC_{50}(M)])	$\Delta\Delta E^{Stab}$(cyp:flv) (kcal mol^{-1})	$\Delta\Delta E^{Inter}(Thr_{498}, flv)$ (kcal mol^{-1})	$\Delta\Delta E^{Inter}$(actsite_top, flv) (kcal mol^{-1})
3,5,7-triOHflavone (1.87)	-43.5	-8.0	-6.4
3-OHflavone (1.70)	-40.6	-5.9	-5.4
5-OHflavone (1.58)	-27.4	-5.9	-3.5
Flavone (1.36)	-13.5	1.4	3.1
3,7-diOHflavone (1.00)	-0.9	2.3	1.2
7-OHflavone (0.78)	0.0	0.0	0.0

Stabilization in the Active Site

The stabilization energy of the hCYP1A2/flavone derivatives complexes after geometry optimization is shown in Table **1**, together with its more relevant components and the corresponding IC_{50} values for the inhibitors. It can be seen that the trend observed for the IC_{50} values is maintained for the stabilization energy.

This is also true for some of the components of the stabilization energy, such as the part correspondent to the interaction energy between the flavones and the aminoacid residues at the top of the active site (see values for $\Delta\Delta E^{Inter}$ *(cyp, flv)* in Table 1 and Fig. (**4**)). A stabilizing interaction in this area of the active site favours the best inhibitors (which confirms the importance of the oxo group on the C4 atom), particularly the interaction with Thr_{498}. Notice that for the two weaker inhibitors, the previously described less buried docking position result in a lower enzyme-ligand energetic complementarity with respect to the residues located on the top of the active site.

Using this approach, we have also calculated the stabilization energy for the molecule 5,7-dihydroxylflavone, which has been shown to be a more potent inhibitor than flavone [7]. We obtained a value of -33.2 kcal mol^{-1}, confirming the correlation between the inhibitory potency of flavonoid inhibitors and the stabilization energy.

CONCLUSIONS

In this work, a set of theoretical tools for analysing enzyme/inhibitor association were presented. The hCYP1A2 inhibition by flavone hydroxylated derivatives has been studied using several approaches. The MEPs study showed that the negative potential located around the oxo group in C4 is important for enzyme-ligand complementarity, and it becomes more evident when it is spread over the substituents in ring A. The study of specific interactions between the enzyme and the ligands related this molecular recognition feature with a stabilizing interaction resulting from hydrogen bonding between the substituents in the A and C rings of the ligands and the aminoacids located at the top of the active site. The hydroxyl substituents on C3 and C5, which surround the oxo group on C4, have an important role in the fitting of the ligand in the active site. This involves a stabilizing interaction with Thr_{498}, which is strong in the best inhibitors of the group. The existence of a negative potential area from C5 to C7 improved the stabilization of the complexes in case there was an hydroxyl substituent in C3. As far as molecular complementarity is concerned, a positive potential between these two negative potential areas increases specificity. Different aspects of the molecular relationship between hCYP1A2 and flavone derivatives were covered in this way. This type of approach should help in the refinement of the binding properties of specific classes of inhibitors.

ACKNOWLEDGMENTS

We thank the FCT (Fundação para a Ciência e Tecnologia) for a doctoral scholarship for R. F and the NFCR (National Foundation for Cancer Research) Centre for Drug Discovery, University of Oxford, U.K., for financial support.

REFERENCES

[1] Boobis, A. R.; Lynch, A. M.; Murray, S.; de la Torre, R.; Solans, A.; Farre, M.; Segura, J.; Gooderham, N. J.; Davies, D. S. *Cancer Res.*, **1994**, *54*, 89-94.

[2] Turesky, R. J.; Constable, A.; Richoz, J.; Varga, N.; Markovic, J.; Martin, M. V.; Guengerich, F. P. *Chem. Res. Toxicol.*, **1998**, *11*, 925-936.

[3] Garner, R. C.; Lightfoot, T. J.; Cupid, B. C.; Russell, D.; Coxhead, J. M.; Kutschera, W.; Priller, A.; Rom, W.; Steier, P.; Alexander, D. J.; Leveson, S. H.; Dingley, K. H.; Mauthe, R. J.; Turteltaub, K. W. *Cancer Lett.*, **1999**, *143*, 161-165.

[4] Bear, W. L.; Teel, R. W. *Anticancer Research*, **2000**, *20*, 3609-3614.

[5] Hodek, P.; Trefil, P.; Stiborová, M. *Chemico-Biological Interactions*, **2002**, *139*, 1-21.

[6] Heller, W. *Acta Horticulturae*, **1994**, *381*, 46-73.

[7] Tsyrlov, I. B.; Mikhailenko, V. M.; Gelboin, H. V. *Biochimica et Biophysica Acta-Protein Structure and Molecular Enzymology*, **1994**, *1205*, 325-335.

[8] Zhai, S.; Dai, R. K.; Friedman, F. K.; Vestal, R. E. *Drug Metab. Dispos.,* **1998**, *26*, 989-992.
[9] Otake, Y.; Walle, T. *Drug Metab. Dispos.,* **2002**, *30*, 103-105.
[10] http://www.accelrys.com
[11] www.ccdc.cam.ac.uk
[12] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery Jr, J. A.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Rega, N.; Salvador, P.; Dannenberg, J. J.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J. Gaussian 98, Revision A.11.2. 2001. Pittsburgh PA, Gaussian, Inc.
[13] Portmann, S.; Luthi, H. P. *Chimia,* **2000**, *54*, 766-770.
[14] Narayszabo, G.; Ferenczy, G. G. *Chemical Reviews,* **1995**, *95*, 829-847.
[15] Politzer, P.; Murray, J. S. *Reviews in Computational Chemistry,* **1991**, *2*, 273-312.
[16] Portela, C.; Afonso, C. M.; Pinto, M. M.; Ramos, M. J. *J. Comput. Aided Mol. Des.,* **2003**, *17*, 583-595.
[17] Fonseca, R.; Menziani, M. C.; Melo, A.; Ramos, M. J. *Molecular Physics,* **2003**, *101*, 2731-2741.
[18] Jones, G.; Willet, P.; Glen, R. C.; Leach, A. R.; Taylor, R. *Journal of Molecular Biology,* **1997**, *267*, 727-748.
[19] Knaggs, A. R. *Natural Product Reports,* **2003**, *20*, 119-136.
[20] Kuffel, M. J.; Schroeder, J. C.; Pobst, L. J.; Naylor, S.; Reid, J. M.; Kaufmann, S. H.; Ames, M. M. *Molecular Pharmacology,* **2002**, *62*, 143-153.
[21] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comp. Chem.,* **1983**, *4*, 187-217.
[22] Murray, J. S.; Politzer, P. In *Molecular Orbital Calculations for Biological Systems*; Sapse, A., Ed.; Oxford University Press: **1998**; pp 49-84.
[23] Narayszabo, G. In *Molecular Interactions*; Sheiner, S., Ed.; John Wiley & Sons Ltd: **1997**; pp 335-350.
[24] Fonseca, R.; Marini, M.; Melo, A.; Menziani, M. C.; Ramos, M. J. *Medicinal Chemistry,* **2005**, *1*, 355-360.

Received: 27 October, 2005 Revised: 28 September, 2005 Accepted: 29 September, 2005

1.4. Structural divergence and adaptive evolution in mammalian cytochromes P450 2C

Structural divergence and adaptive evolution in mammalian cytochromes P450 2C

Rute R. da Fonseca, Agostinho Antunes, André Melo, Maria João Ramos

REQUIMTE, Departamento de Química, Faculdade de Ciências, Universidade do Porto

Rua do Campo Alegre, 687, 4169-007 Porto, Portugal

ABSTRACT

Cytochromes P450 (CYPs) comprise a superfamily of enzymes involved in various physiological functions, including the metabolism of drugs and carcinogenic compounds present in food, making them of great importance for human health. The possibility that CYPs could be broadening or changing substrate specificity in accordance to the high diversity of xenobiotics compounds environmentally available suggests that their metabolic function could be under adaptive evolution. We evaluated the existence of functional divergence and signatures of selection on mammalian genes from the drug-metabolizing CYP2 family. Thirteen of the sites found to be functionally divergent and the eight found to be under strong positive selection occurred in important functional domains, namely on the substrate entrance channel and within the active site. Our results provide insight into CYPs evolution and the role of molecular adaptation in enzyme substrate-specificity diversification.

Keywords

Positive selection; enzyme; 3D-structure; functional divergence

INTRODUCTION

Cytochromes P450 (CYPs) can be found in almost all organisms and present a high intra and interspecies diversity (Werck-Reichhart and Feyereisen 2000). One of their tasks is the oxidation of xenobiotics compounds to facilitate their excretion from the organism (Guengerich 2001; Omiecinski et al 1999). These include drugs and carcinogenic compounds present in food (Anzenbacher and Anzenbacherova 2001), which explains their importance for pharmaceutical research.

The possibility that the CYPs performing such tasks could be broadening or changing substrate specificity in accordance to environmental changes suggests that their metabolic function could be under adaptive evolution. In fact, mosquito CYPs seem to be involved in insecticide resistance (Ranson et al 2002). Different evolutionary pressures on particular sites of CYPs should be revealed by the analysis of the rate of substitutions occurring within coding regions.

Protein functional divergence after gene duplication can be tested by calculating the replacement rates that occur in the originated subfamilies (Gaucher et al 2001). Likelihood ratio tests (LRTs) can be use to assess whether the rates differ between the subfamilies and/or are accelerated in one/all of them (Gu 2003; Knudsen et al 2003; Knudsen and Miyamoto 2001).

Nonsynonymous substitutions may influence the fitness of an individual or population. Thus, adaptive molecular evolution may cause the nonsynonymous substitution rate (d_N) to be higher than the synonymous rate (d_S), with the ratio ω (d_N/d_S) being higher than 1 (Yang et al 2000). The methods that provide statistical measures of such mutation rates fail, however, in evaluating the physicochemical importance of the correspondent amino acid changes and its consequences for the protein function (McClellan et al 2005). Further analyses are necessary to determine if there is a statistically significant change in the amino acid properties at particular sites. CYPs chemical and structural variations in areas that are in contact with the ligands (substrate recognition sites, SRSs (Gotoh 1992)) will have consequences on the size, shape, and chemical characteristics of both substrates and products.

In this study, we assessed functional divergence among mammalian CYP2C and used two types of methods to determine if these genes are under adaptive evolution: (i) a gene level approach, testing for functional divergence and positive selection using statistical methods and (ii) a protein level approach, evaluating statistically significant physicochemical amino acid changes and verifying the impact of the previous analyses on the protein three-dimensional (3D) structure. So far, positive selection has been mostly associated with protein recognition domains, such as those involved in immune response and reproduction (for a list of examples see Table S1 in supporting

information; henceforth all supporting information will be indicated by an S preceding the correspondent numbering). We retrieved unequivocal evidence that diversifying selection is acting on CYP's active site, thus providing insight to understand rapid enzyme function diversification.

MATERIALS AND METHODS

Sequences and structures

The CYP superfamily is classified into families and subfamilies (> 40% or 55% amino acid sequence identity, respectively) (Nelson et al 2004). The enzymes used in this study belong to the largest and most diverse of CYP families, CYP2 (Omiecinski et al 1999). Its members are responsible for the metabolism of a variety of different pharmaceutical agents and represent more than 20% of the human liver total CYPs content (Omiecinski et al 1999). Mutations occurring both within and outside the active site modulate CYP2s activities, as shown for the polymorphisms occurring in human enzymes (Table S2). The nucleotide sequences of the CYP2C enzymes used in the phylogenetic analyses were retrieved from GenBank (refer to Table S3 for species names and GenBank accession numbers).

Two datasets were analysed, one containing 37 sequences (large dataset) and another containing 12 sequences (small dataset) (Table S3 for details), with the purpose of testing the reliability of the dataset expansion/contraction. Both datasets included the enzymes for which 3D structures are available from the Protein Data Bank [human proteins CYP2C8 and CYP2C9 and rabbit enzymes CYP2B4 and CYP2C5 (PDB codes: 1PQ2, 1OG5, 1SU0 and 1NR6, respectively). Sequences were aligned with ClustalX (Thompson et al 1994) and edited with SeaView (Galtier et al 1996) (the alignment is provided as supporting information). Gaps were removed from further analyses. The numbering of sites was made according to the alignment without gaps. Correspondence between the PDB sequence numbering and that of the amino acids mentioned in the results is presented in Table S4.

A sliding window analysis was used as implemented in SWAAP 1.0.2 (Pride 2000) to obtain the synonymous and nonsynonymous mutation ratios computed using the Nei and Gojobori method (Nei and Gojobori 1986), and both amino acid and nucleotide similarity plots for the CYPs with available X-ray structure.

Maximum likelihood phylogenetic trees were built for the two datasets using PAUP 4.0b10 (Swofford 1998) after determining the optimal model of sequence substitution with Modeltest 3.04 (Posada and Crandall 1998) (Fig. S2). The detection of possible recombination and gene

conversion events was evaluated using GENECONV (Sawyer 1999). All the default settings were used except for the mismatch penalty (set to 1) and the option to analyze only silent sites.

To test for functional divergence, detection of different evolutionary rates was done on a site specific basis according to the Knudsen and Miyamoto (Knudsen and Miyamoto 2001) likelihood ratio test (LRT). Hypothesis stating the occurrence of divergence type I (when the amino acid configuration is highly conserved in one subfamily but highly variable in the other) or II (when the amino acid site is under similar functional constraints in both subfamilies, but the amino acid properties that are being selected are different between the two) are tested against a null hypothesis that considers that a single rate describes the evolution for the site in question (Knudsen et al 2003). In case the site presents a single rate of evolution, another LRT tests whether this rate is different from the average, presenting conserved or accelerated evolution (Knudsen et al 2003). The LRT procedure estimated the branch lengths using the Jones, Taylor and Thornton (JTT) model (Jones et al 1992). At each point in the alignment, the method performs the LRT for the significance of a rate shift at a given point in the phylogeny using 10,000 replicates (Knudsen and Farid 2004). The 5% cutoffs from the simulations are then compared with those calculated with the real data. If the value for the real data is higher that that of the simulations ($\Delta U > 0$), then the rate change hypothesis fits the data significantly better than the corresponding null hypothesis (Knudsen et al 2003).

Positive selection analyses: gene level approach

Evidences for positive selection on CYPs were tested using different codon substitution models, which differ in how d_N/d_S ratio varies along sequences, as implemented in PAML 3.14b (Yang 1997). To test whether sites exist where $\omega > 1$, we used a likelihood-ratio test comparing two probabilistic models of variable ω ratios among sites, the simpler of which does not allow sites with $\omega > 1$ and the more general which does (Wong et al 2004). Henceforth, codons will be referred to as sites. In this study we compared M1a and M7, models that do not allow $\omega > 1$, with M2a and M8, which allow $\omega > 1$ (respectively), as these were shown to be more suitable in detecting positive selection (Yang et al 2005). The level of significance of such test is calculated as twice the difference of the likelihood scores estimated by each model ($2\Delta \ln L$) and the null distribution of these results can be approximated by a χ^2 distribution with the number of degrees of freedom calculated as the difference in the number of estimated parameters between models (Wong et al 2004; Yang 2000). Codon sites under positive selection were identified using the Bayes empirical Bayes (BEB) calculation of posterior probabilities for sites classes (Yang et al 2005) that analyses the sites under positive selection identified by models M2a and M8. Models

that allow heterogeneity in the d_N/d_S ratio among lineages were also tested. The simplest model is the one-ratio model M0, that assumes only one value of ω. The most general model is the free-ratio model, which assumes as many ω parameters as the number of branches in the tree (Yang 1998). Furthermore, we tested the two-ratio model, which allows predefined lineages to have a different ω value from the rest of the tree.

Selective constrains were further analysed using a sliding-window maximum parsimony approach in SWAPSC (Fares 2004). This method infers a statistically optimum codon-window size using simulated sequence alignments, and applies the Kimura-based model of Li (Li 1993). The sliding window size significance is then tested based on the deviations of the observed nonsynonymous or/and synonymous nucleotide substitutions from the expectation under neutrality (Fares et al 2002b; Fares et al 2002a).

Selection analyses: protein level approach

We first used the method implemented in TreeSAAP (Woolley et al 2003) that calculates the goodness-of-fit between an observed distribution of physicochemical changes inferred from a phylogenetic tree and an expected distribution based on the assumption of completely random amino acid replacement expected under the condition of selective neutrality. We have looked particularly at the amino acid changes correspondent to positive-destabilizing selection, which implicates a radical physicochemical variation in order to determine which sites have had their function deeply changed. The 31 physicochemical properties evaluated by TreeSAAP were subdivided according to their effect on protein properties (Table S5). Subsequently, we performed protein structure analyses based on CYP 3D structures. The available X-ray structures for CYP2 family enzymes were superimposed and the root mean square deviation values (RMSD) for the backbone Cα atoms were determined with InsightII (Accelrys ©).

RESULTS AND DISCUSSION

Selection analyses

The first evidence of selection in CYPs was retrieved by the SWAAP sliding window approach (Fig. 1A). Very low values of d_N/d_S were observed for the heme binding areas. This is consistent with the fact that these areas are associated with the catalytic oxidative mechanism, which is common to all CYPs (Poulos 1995). By contrast, the substrate binding areas of the active site presented some of the higher values of d_N/d_S. These include SRS-1, SRS-3, SRS-2, SRS-5 and SRS-6 (Fig. 1B).

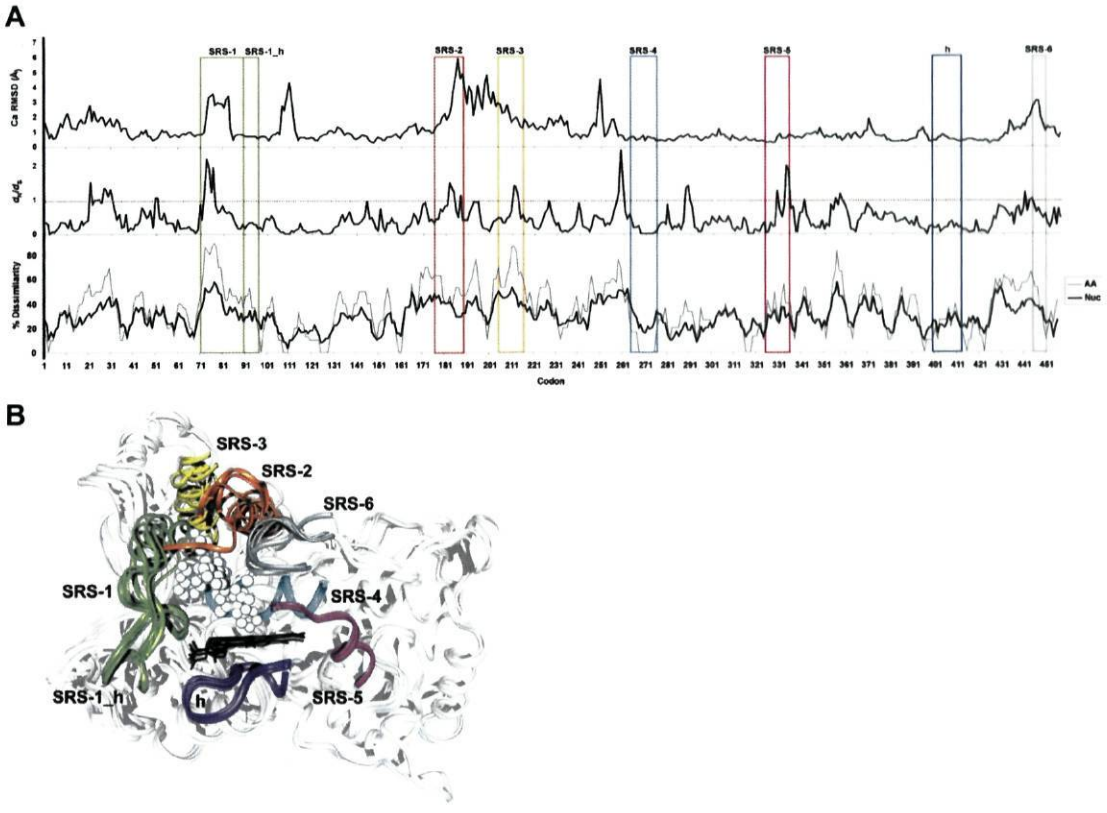


Fig. 1. (a) Results of SWAAP sliding window analysis together with the RMSD and the sequence dissimilarities measured for the four CYPs with available X-ray structure (AA: amino acid; Nuc: nucleotide). The values presented are averaged. (b) Superimposed X-ray structures of mammalian cytochomes P450 hCYP2C8, hCYP2C9, rCYP2B4 and rCYP2C5 shown as ribbon drawings; the active site area is detached and the atoms of the ligands are represented as white spheres. The different substrate recognition areas (SRSs) are shown together with the heme binding areas (SRS-1_h and h).

We used maximum likelihood trees generated with the TVM+I+G (large set) and TVM+G (small set) models of sequence substitution (Fig. S2) as frameworks for the phylogenetic based selection approaches. The large set shows some saturation in the third codon position (Fig. S3), but the phylogenetic reconstruction removing this position has little effect on the tree topology (results not shown). No statistically significant evidence of recombination or gene conversion events has

been detected among the CYP sequences studied for both datasets (samples using the human, rat, rabbit and golden hamster sequences independently were also tested retrieving similar results).

For the LRT analysis two clusters were defined in the ML tree (A and B; Fig. S2) based on substrate specificity divergence (Fig. S1) and phylogenetic relationships. Thirty-one sites were found to exhibit functional divergence type I and/or II (Table S4; Fig. S4A). Thirteen are located within and in adjacent areas of the active site, three in the N-terminal domain and the remaining in surface areas (Fig. S5). The other sites were further divided into those that presented a slower or a faster rate than the average. Substrate binding regions SRS-1, SRS-2, SRS-3 and SRS-6 present a high incidence of sites with accelerated evolutionary rates, while the heme area (h) shows a high percentage of conserved sites (Fig. S4B).

LRTs performed in PAML indicated that both the small and the large dataset presented evidence of positive selection. Log-likelihood values for the models implemented in PAML showed that model M8 was the one that best fitted the data (Table 1).

Table 1. PAML results. Log-likelihood values and parameter estimates obtained for the two data sets of CYP2C enzymes with the site models implemented in PAML.

Model	Small set	Large set	2Δ(ln*L*)	Small set	Large set
M1a	p_0 = 0.68325 (p_1 = 0.31675) ω_0 = 0.12240 ln*L* = -7627.0	p_0 = 0.71162 (p_1 = 0.28838) ω_0 = 0.15153 ln*L* = -19278.0	M1a *vs* M2a	16 ($p = 3E^{-04}$)	30 ($p = 3E^{-07}$)
M2a	p_0 = 0.67491 p_1 = 0.29992 (p_2 = 0.02517) ω_1 = 0.12794 ω_2 = 3.19454 ln*L* = -7618.7	p_0 = 0.70276 p_1 = 0.26733 (p_2 = 0.02991) ω_1 = 0.15605 ω_2 = 2.13829 ln*L* = -19262.8			
M7	p = 0.36064 q = 0.63479 ln*L* = -7629.8	p = 0.44815 q = 0.84807 ln*L* = -19218.7	M7 *vs* M8	32 ($p = 1E^{-7}$)	64 ($p = 1E^{-14}$)
M8	p_0 = 0.94177 p_1 = 0.05823 q = 1.10020 p = 0.50127 ω = 2.25021 ln*L* = -7613.4	p_0 = 0.90625 p_1 = 0.09375 q = 1.68177 p = 0.61964 ω = 1.43833 ln*L* = -19186.7			

This model showed that 5% of the sites were under positive selection (ω = 2.25) in the small set and 6% in the large set (ω = 1.44). The BEB analysis identified six sites under positive selection in the small set with posterior probability (PP) $\geq$ 95%, five of which were located within the substrate binding regions SRS-1 (74, 78), SRS-3 (214), SRS-4 (263) and SRS-6 (444), and one with PP $\geq$ 90% located within SRS-4 (267). Within the large dataset the same sites were observed with PP $\geq$ 95%, plus two others located in SRS-3 (208, 211) (Table 2; Fig. S5). There was no

significant evidence that particular lineages in the dataset were under positive selection (Table S6).

Table 2. Sites found to be under positive selection in CYP2C enzymes by PAML. The Bayes Empirical Bayes (BEB) posterior probabilities are shown for sites with PP > 0.95 (in bold) detected with models M2a and M8 for the two datasets. Additional data for these sites, provided by the TreeSAAP (number of radical changes in amino acid properties for each site) and LRT (sites that present an evolutionary rate different that average) analyses, is presented. Sites in grey boxes are within the active site.

	PAML (BEB)				TreeSAAP		LRT
	Small Set		Large set		Small Set	Large set	Type (ΔU)
Sites	M2a	M8	M2a	M8			
74	0.95	**0.98**	**0.97**	**0.99**	23	32	Fast (9.8)
78	*0.93*	**0.98**	1.00	**1.00**	3	35	Fast (4.4)
196	0.89	**0.97**	-	-	7	1	-
208	-	0.56	*0.91*	**0.96**	2	13	Fast (10.9)
211	0.57	0.81	**0.96**	**0.98**	8	19	Fast (37.4)
214	*0.91*	**0.97**	*0.94*	**0.97**	24	28	Fast (8.2)
263	*0.93*	**0.98**	0.52	*0.90*	7	22	Fast (6.8)
267	0.77	*0.90*	0.86	**0.95**	21	23	Fast (11.9)
444	*0.94*	**0.98**	1.00	**0.99**	12	37	Fast (20.0)

Further evidence of selection within the active site was obtained with SWAPSC (Fig. 2), suggesting that 12% of the windows analysed were under molecular adaptation in the small dataset and 28% in the large dataset. Negative selection was the predominant character, particularly within areas involved in heme binding (SRS-1_h and h). Substrate binding areas presented a high percentage of windows with accelerated rates of non-synonymous nucleotide substitutions and saturation. Both accelerated rates of nucleotide substitutions and actual positive selection were predominant characteristics in windows belonging to SRS-3 and SRS-6. Positive selection in SRS-2 was also detected in the large dataset. SRS-1, which have had two sites under positive selection assigned with PAML, showed a high number of windows under accelerated rate of non-synonymous nucleotide substitutions and saturation. A small percentage of positive selection was detected in the large dataset.

The existence of evolutionary stress on SRS-1 and SRS-3 is quite interesting, as the interface between the two has been designated as the most likely substrate entrance channel in mammals (Schleinkofer et al 2005), i.e. a structural domain directly involved in the first contact with the substrate. The total saturation for the large dataset was 4.9% and the 1.6% for the small dataset. This indicates that positive selection results could be inflated for the large dataset, which is why we have only considered those sites that present strong signals of positive selection in both datasets.

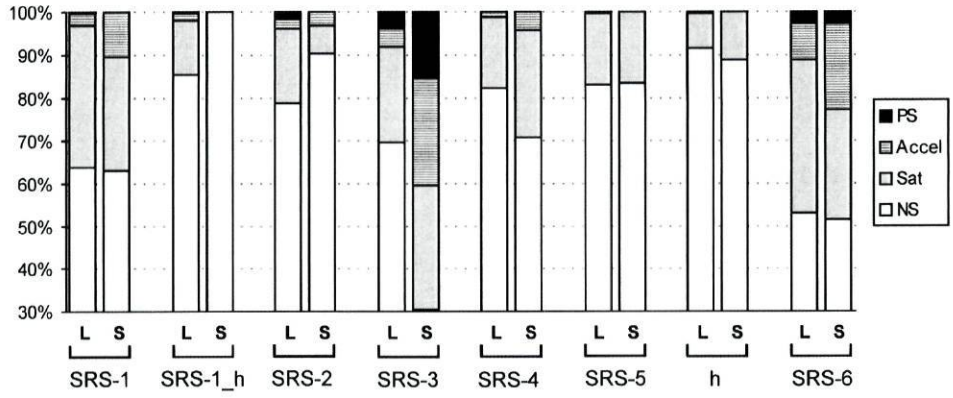


Fig. 2. SWAPSC results presented as the percentage of non-neutral sliding windows within each sequence interval (Sat: saturated; Accel: accelerated mutation rate; PS: positive selection; NS: negative selection) for the large (L) and small (S) datasets.

The TreeSAAP protein level results showed that some of the positively selected sites previously detected by PAML had a high number of radical changes of their amino acid properties (Table 2; Fig. S6). This is the case of sites 74, 214, 263, 267 and 444. Notice that sites 205 and 405 which show a high degree of radical physicochemical modifications in both datasets were not detected by PAML. The number of radical physicochemical modifications occurring in the active site areas, reinforced previous analyses, with the highest values being observed in SRS-1, SRS-3 and SRS-6 (Fig. 3; Table S7).

Overall, our results highlight the importance of using complementary approaches at both the gene and protein level to undertake a thorough evaluation of molecular adaptation.

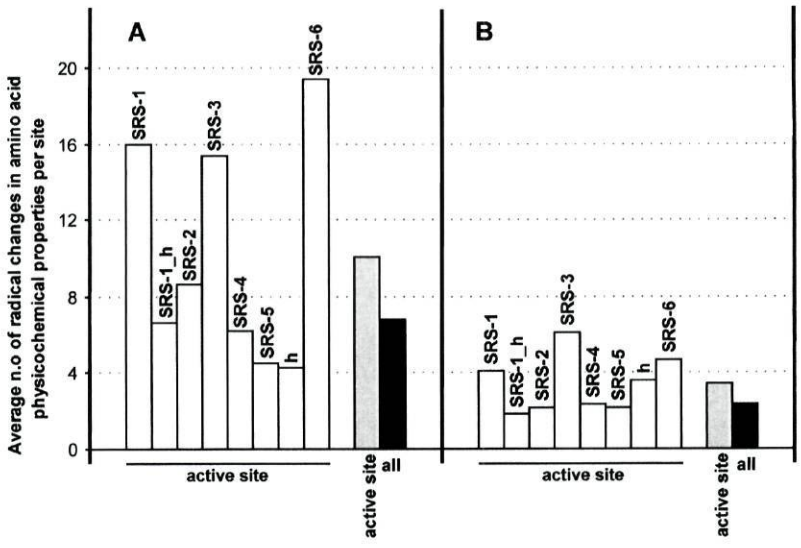


Fig. 3. Distribution of positive-destabilizing amino acid properties per site within the active site regions (AVE: average number of cases per designated area). a) Large dataset; b) Small dataset.

Enzyme Structure and Function versus Adaptation

The ultimate evidence that adaptive molecular evolution was acting on CYPs substrate specificity was provided by the analysis of the specific mutations that directly influence the enzyme structure and function. From the analysis of available X-ray structures, it was possible not only to infer the local consequences of a mutation event but also to determine the possible effects of mutations in related enzymes without a known 3D structure.

Since mammalian Cytochromes P450 are membrane proteins anchored by their N-terminal helix (domain not included in this study), functional divergence occurring in the surface N-terminal domain (Fig. S4) could be interfering with the interaction of the enzyme with the membrane, that has been suggested to be important for activity towards some kind of substrates (Schleinkofer et al 2005). The RMSD variation among the CYP2 enzyme structures (Fig. 1A) was high in the substrate binding regions SRS-1, SRS-2, SRS-3 and SRS-6 (Fig. 1B), a fact shown to be correlated with the occurrence of positive selection. We confirmed the great importance of site 74 modifications, as the correspondent amino acid residue sets the beginning of SRS-1, and its modification causes a kink on the backbone drawing, which is likely to be responsible for the high structural variability of this area (Fig. 4A).

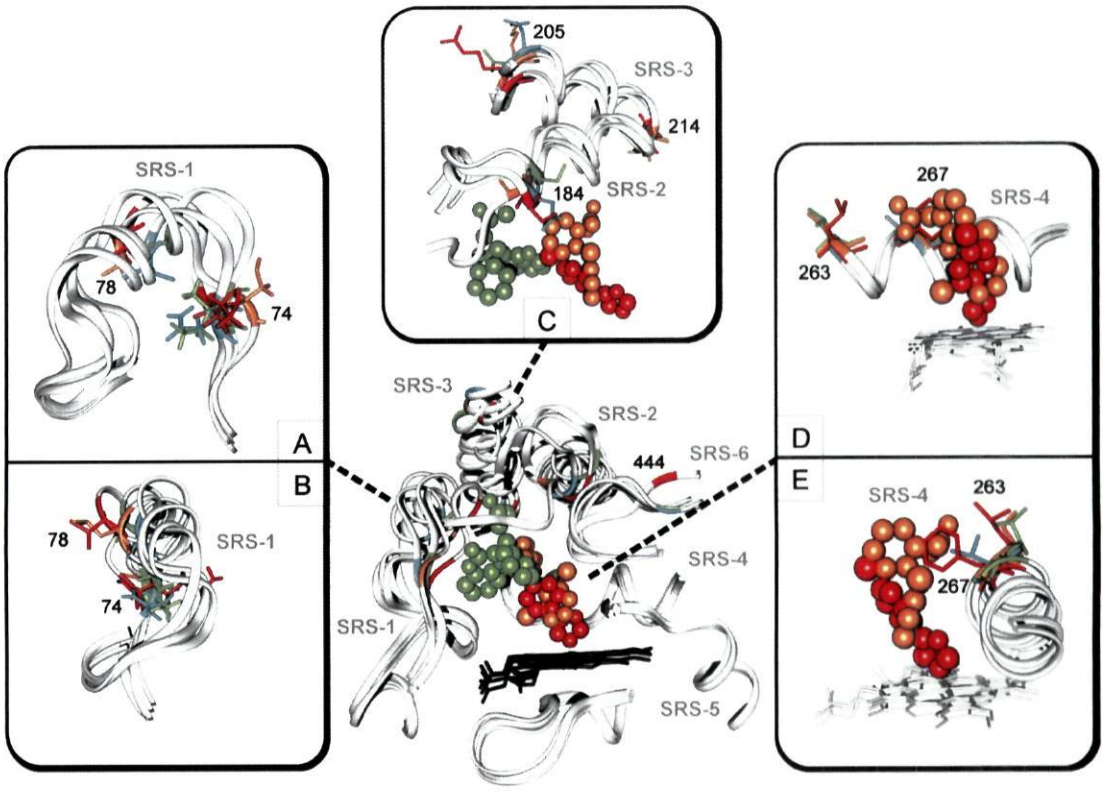


Fig. 4. Sites detected to be under positive selection on the substrate binding areas of mammalian CYPs (blue - hCYP2C8; green - hCYP2C9; red - rCYP2B4; orange - rCYP2C5). The non-hydrogen atoms of the ligands are represented as spheres.

Several amino acid properties have changed in site 74 (the amino acids varying were His, Ser, Ile, Asn, Val and Arg, all with very different chemical and physical characteristics), which justified the different preferences for the backbone drawing and orientation of the side-chain. In the middle of SRS-1, site 78 was in a place where the structural divergence is high. Changes in the preferred orientation of the amino acid side-chains enhance this disorder (Fig. 4B). Furthermore, it is possible to observe that Ser in hCYP2C8 turned to the outside and Leu (CYP) and Ala (CYP) turned inwards (Fig. 4B). The adjacent area, SRS-1_h, in contact with the carboxylate groups from the heme, showed preservation of the local charge and especially of the Trp residue that directly binds one of the heme propionates.

In SRS-2 there was a high structural variability. The highest deformation in the backbone was observed in rCYP2C5, which had SRS-2 drawn into the active site cavity, with consequences in its volume and interaction with the ligands (Fig. 4C). This is the result of having a Gly in site 184, found to be a type I divergence site. In SRS-3, the site 205 indicated by the TreeSAAP analysis, was within an area implicated in substrate recognition – the variation in chemical properties will thus have direct consequences on such event. Still in SRS-3, site 214 corresponds to amino acids where the side-chains turned to the inside of the active site cavity. Large fluctuations in side-chain volume and polarity, such as that from a Met (hCYP2C9) to a Thr (hCYP2C8) and to a Ile (rCYP2C5) will change the packing of the substrate when already inside the active site (Fig. 4C). Site 209 (type I divergence), changing its characteristics from positively charged to polar, has its side-chain turned to the outside and could be interfering with approaching substrates.

In SRS-4, sites 263 and 267 are buried in the active site and any fluctuations in size and polarity will modulate the fitting and orientation of the substrate in the active site. Concerning site 263, although not many differences were observed in the X-ray structures, the CYP2C alignment (provided as supporting information) indicated that this site can have very different residues varying such as a Val to a Trp, and therefore have implications on the orientation of the ligands (Fig. 4D). The impact of amino acid changes in site 267 is clearly represented in Fig. 4E. It is possible to see how the side-chain of the amino acid Phe in rCYP2B4 overlaps the ligand correspondent to rCYP2C5, meaning that there will be a modulation of the substrate binding mode very close to the site of oxidation, with several consequences, namely on the stereochemistry of the products. Site 265, indicated as a mixed type I & II divergence, has its side-chain turned away from the active site cavity, but the changes in side-chain size (Val to Ile to Met) can interfere with the packing of the surrounding area.

Finally, sites 444 and 449 are located in SRS-6, which is interacting directly with SRS-2. The side-chain of these amino acid in this site may influence, therefore, the folding of that adjacent area.

When functional divergence sites are located outside the active site areas its influence will likely be related with global protein folding or substrate recognition. Indeed, many of the polymorphisms observed for the human enzymes (Table S2) that change the enzyme's activity (increasing/decreasing or inactivating it) are located outside the SRSs, and therefore mutations within SRS areas are definitely not the only responsible for changes in the enzyme's substrate specificity.

Understanding the detected functional divergence is not as straightforward as inferring functional diversification from positive selection events. Also, the functional divergence results provided by the LRT analysis should be interpreted with caution, due to the low confidence phylogenetic support of the clades defined, and the scarcity of metabolic data on the various CYP2C enzymes. However, available experimental information regarding substrate specificity provided some insight to unravel the effects of functional divergence of rabbit and human CYP2C isoforms. Rabbit CYP2C1/2 (cluster A) hydroxylates laurate omega-1 (Laethem and Koop 1992; Uno et al 1993) (Fig. S1), a long-chain fatty acid, while the CYP2C4/5 (Johnson et al 1987; Williams et al 2000) (cluster B) metabolize the 21-hydroxylation of progesterone, a steroid (Fig. S1). The two human enzyme CYP2C8 (cluster A) and CYP2C9 (cluster B) show differences both at the level of substrate recognition and size of the active site (larger for the former) (Totah and Rettie 2005). Drugs metabolized by both enzymes vary in size and structure, with those metabolized by CYP2C8 being large, mildly acidic, basic or neutral (e.g. paclitaxel is a typical substrate of CYP2C8; Fig. S1) (Totah and Rettie 2005), while most of those metabolized by CYP2C9 being weak acids (e.g. S-warfarin is specifically metabolized by CYP2C9; Fig. S1) (Miners and Birkett 1998). However, some substrates are shared by enzymes from the different clusters and species. This is true for many drugs metabolized by both human CYP2C8 and CYP2C9, such as fluvastavin (used to treat hypercholesterolemia and to prevent cardiovascular disease) and ibuprofen (a nonsteroidal anti-inflammatory drug), although with varied metabolic rates (Miners and Birkett 1998) (Fig. S1), and also steroids (namely oral contraceptives) (Delaforge et al 2005; Sandberg et al 2004; Zhou et al 2004) (Fig. S1). Also, these enzymes, together with the mouse CYP2C enzymes and the rabbit enzymes CYP2C1/2 take part in the metabolism of arachidonic acid (Laethem and Koop 1992; Luo et al 1998; Rifkind et al 1995; Wang et al 2004) (Fig. S1). Notwithstanding, future enzyme-activity studies regarding substrate specificity would be important to fully understand these functional divergence results.

CONCLUSIONS

Hence, we have shown Cytochromes P450 functional diversification exhibits signatures of functional divergence and molecular adaptation. The combined use of protein structural information and that of statistical approaches at both gene and protein level demonstrated to be a successful methodology to unravel signatures of natural selection, which should be followed in future studies. Gene duplication likely played an important role in functional diversity, with different functional constraints acting on different sites of the duplicate genes. Additionally, positive selection is influencing the CYP's structure and function, enabling these enzymes to phenotypically adapt and acquire a myriad of substrate affinities, in a continuous process of molecular evolution. This provides an unusual example of functional divergence related with the variation of enzyme substrate specificity which so far has been detected in only a few cases of positive selection reported cases (for a list of examples see Table S1). The knowledge that specific amino acids located in CYP2Cs active sites present a high mutability rate is also an interesting asset for pharmacological sciences, given the fact that many drugs are designed to be metabolized by these enzymes.

ACKNOWLEDGMENTS

We thank the FCT (Fundação para a Ciência e Tecnologia) for a doctoral scholarship (SFRH/BD/7089/2001) for R.F. and the NFCR (National Foundation for Cancer Research) Centre for Drug Discovery, University of Oxford, U.K., for financial support.

REFERENCES

Anzenbacher,P., Anzenbacherova,E., 2001. Cytochromes P450 and metabolism of xenobiotics. Cellular and Molecular Life Sciences 58, 737-747.

Delaforge,M., Pruvost,A., Perrin,L., Andre,F., 2005. Cytochrome P450-mediated oxidation of glucuronide derivatives: Example of estradiol-17 beta-glucuronide oxidation to 2-hydroxyestradiol-17 beta-glucuronide by CYP2C8. Drug Metab Dispos 33, 466-473.

Fares,M.A., 2004. SWAPSC: sliding window analysis procedure to detect selective constraints. Bioinformatics 20, 2867-2868.

Fares,M.A., Barrio,E., Sabater-Munoz,B., Moya,A., 2002a. The evolution of the heat-shock protein GroEL from Buchnera, the primary endosymbiont of aphids, is governed by positive selection. Molecular Biology and Evolution 19, 1162-1170.

Fares,M.A., Elena,S.F., Ortiz,J., Moya,A., Barrio,E., 2002b. A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. Journal of Molecular Evolution 55, 509-521.

Galtier,N., Gouy,M., Gautier,C., 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput. Appl. Biosci. 12, 543-548.

Gaucher,E.A., Miyamoto,M.M., Benner,S.A., 2001. Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. PNAS 98, 548-552.

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

Gu,X., 2003. Functional divergence in protein (family) sequence evolution. Genetica 118, 133-141.

Guengerich,F.P., 2001. Common and uncommon cytochrome P450 reactions related to metabolism and chemical toxicity. Chem Res Toxicol 14, 611-650.

Johnson,E.F., Barnes,H.J., Griffin,K.J., Okino,S., Tukey,R.H., 1987. Characterization of A 2Nd Gene-Product Related to Rabbit Cytochrome-P-450-1. J. Biol. Chem. 262, 5918-5923.

Jones,D.T., Taylor,W.R., Thornton,J.M., 1992. The Rapid Generation of Mutation Data Matrices from Protein Sequences. Computer Applications in the Biosciences 8, 275-282.

Knudsen,B., Farid,N.R., 2004. Evolutionary divergence of thyrotropin receptor structure. Molecular Genetics and Metabolism 81, 322-334.

Knudsen,B., Miyamoto,M.M., 2001. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proceedings of the National Academy of Sciences of the United States of America 98, 14512-14517.

Knudsen,B., Miyamoto,M.M., Laipis,P.J., Silverman,D.N., 2003. Using evolutionary rates to investigate protein functional divergence and conservation: A case study of the carbonic anhydrases (vol 164, pg 1261, 2002). Genetics 165, 453.

Laethem,R.M., Koop,D.R., 1992. Identification of rabbit cytochromes P450 2C1 and 2C2 as arachidonic acid epoxygenases. Molecular Pharmacology 958-963.

Li,W.H., 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36, 96-99.

Luo,G., Zeldin,D.C., Blaisdell,J.A., Hodgson,E., Goldstein,J.A., 1998. Cloning and expression of murine CYP2Cs and their ability to metabolize arachidonic acid. Archives of Biochemistry and Biophysics 357, 45-57.

McClellan,D.A., Palfreyman,E.J., Smith,M.J., Moss,J.L., Christensen,R.G., Sailsbery,A.K., 2005. Physicochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome b proteins. Molecular Biology and Evolution 22, 437-455.

Miners,J.O., Birkett,D.J., 1998. Cytochrome P4502C9: an enzyme of major importance in human drug metabolism. British Journal of Clinical Pharmacology 45, 525-538.

Nei,M., Gojobori,T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3, 418-426.

Nelson,D.R., Zeldin,D.C., Hoffman,S.M., Maltais,L.J., Wain,H.M., Nebert,D.W., 2004. Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. Pharmacogenetics 14, 1-18.

Omiecinski,C.J., Remmel,R.P., Hosagrahara,V.P., 1999. Concise review of the cytochrome P450s and their roles in toxicology. Toxicol. Sci. 48, 151-156.

Posada,D., Crandall,K.A., 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14, 817-818.

Poulos,T.L., 1995. Cytochrome P450. Curr Opin Struct Biol 5, 767-774.

Pride,D.T. SWAAP - A tool for analysing substitutions and similarity in multiple alignments. 2000.

Ranson,H., Claudianos,C., Ortelli,F., Abgrall,C., Hemingway,J., Sharakhova,M.V., Unger,M.F., Collins,F.H., Feyereisen,R., 2002. Evolution of supergene families associated with insecticide resistance. Science 298, 179-181.

Rifkind,A.B., Lee,C., Chang,T.K.H., Waxman,D.J., 1995. Arachidonic-Acid Metabolism by Human Cytochrome P450S-2C8, Cytochrome P450S-2C9, Cytochrome P450S-2E1, and Cytochrome P450S-1A2 - Regioselective Oxygenation and Evidence for A Role for Cyp2C Enzymes in Arachidonic-Acid Epoxygenation in Human Liver-Microsomes. Archives of Biochemistry and Biophysics 320, 380-389.

Sandberg,M., Johansson,I., Christensen,M., Rane,A., Eliasson,E., 2004. The impact of CYP2C9 genetics and oral contraceptives on cytochrome P4502C9 phenotype. Drug Metab Dispos 32, 484-489.

Sawyer,S.A. GENECONV: A computer package for the statistical detection of gene conversion. Distributed by the author, Department of Mathmatics, Washington University in St. Louis, available at http://www.math.wustl.edu/~sawyer. 1999.

Schleinkofer,K., Sudarko, Winn,P.J., Lüdemann,S.K., Wade,R.C., 2005. Do mammalian cytochrome P450s show multiple ligand access pathways and ligand channelling? EMBO reports 6, 584-589.

Swofford,D.L., 1998. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachussetts.

Thompson,J.D., Higgins,D.G., Gibson,T.J., 1994. Clustal-W - 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.

Totah,R.A., Rettie,A.E., 2005. Cytochrome P4502C8: Substrates, inhibitors, pharmacogenetics, and clinical relevance. Clinical Pharmacology & Therapeutics 77, 341-352.

Uno,T., Imai,Y., Nakamura,M., Okamoto,N., Fukuda,T., 1993. Importance of Successive Prolines in the Carboxy-Terminal Region of P450-2C2 and P450-2C14 for the Hydroxylase-Activities. Journal of Biochemistry 114, 363-369.

Wang,B., Sanchez,R.I., Franklin,R.B., Evans,D.C., Huskey,S.E.W., 2004. The involvement of CYP3A4 and CYP2C9 in the metabolism of 17 alpha-ethinylestradiol. Drug Metab Dispos 32, 1209-1212.

Werck-Reichhart,D., Feyereisen,R., 2000. Cytochromes P450: a success story. Genome Biol 1, reviews3003.1-3003.9.

Williams,P.A., Cosme,J., Sridhar,V., Johnson,E.F., McRee,D.E., 2000. Microsomal cytochrome P450 2C5: comparison to microbial P450s and unique features. J Inorg Biochem 81, 183-190.

Wong,W.S.W., Yang,Z., Goldman,N., Nielsen,R., 2004. Accuracy and Power of Statistical Methods for Detecting Adaptive Evolution in Protein Coding Sequences and for Identifying Positively Selected Sites. Genetics 168, 1041-1051.

Woolley,S., Johnson,J., Smith,M.J., Crandall,K.A., McClellan,D.A., 2003. TreeSAAP: Selection on Amino Acid Properties using phylogenetic trees. Bioinformatics 19, 671-672.

Yang,Z.H., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer Applications in the Biosciences 13, 555-556.

Yang,Z.H., 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular Biology and Evolution 15, 568-573.

Yang,Z.H., 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. Journal of Molecular Evolution 51, 423-432.

Yang,Z.H., Wong,W.S.W., Nielsen,R., 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Molecular Biology and Evolution 22, 1107-1118.

Yang,Z., Nielsen,R., Goldman,N., Pedersen,A.M.K., 2000. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. Genetics 155, 431-449.

Zhou,S., Chan,E., Lim,L.Y., Boelsterli,U.A., Li,S.C., Wang,J., Zhang,Q., Huang,M., Xu,A., 2004. Therapeutic drugs that behave as mechanism-based inhibitors of cytochrome P450 3A4. Curr. Drug Metab 5, 415-442.

SUPPORTING INFORMATION
- Figures -

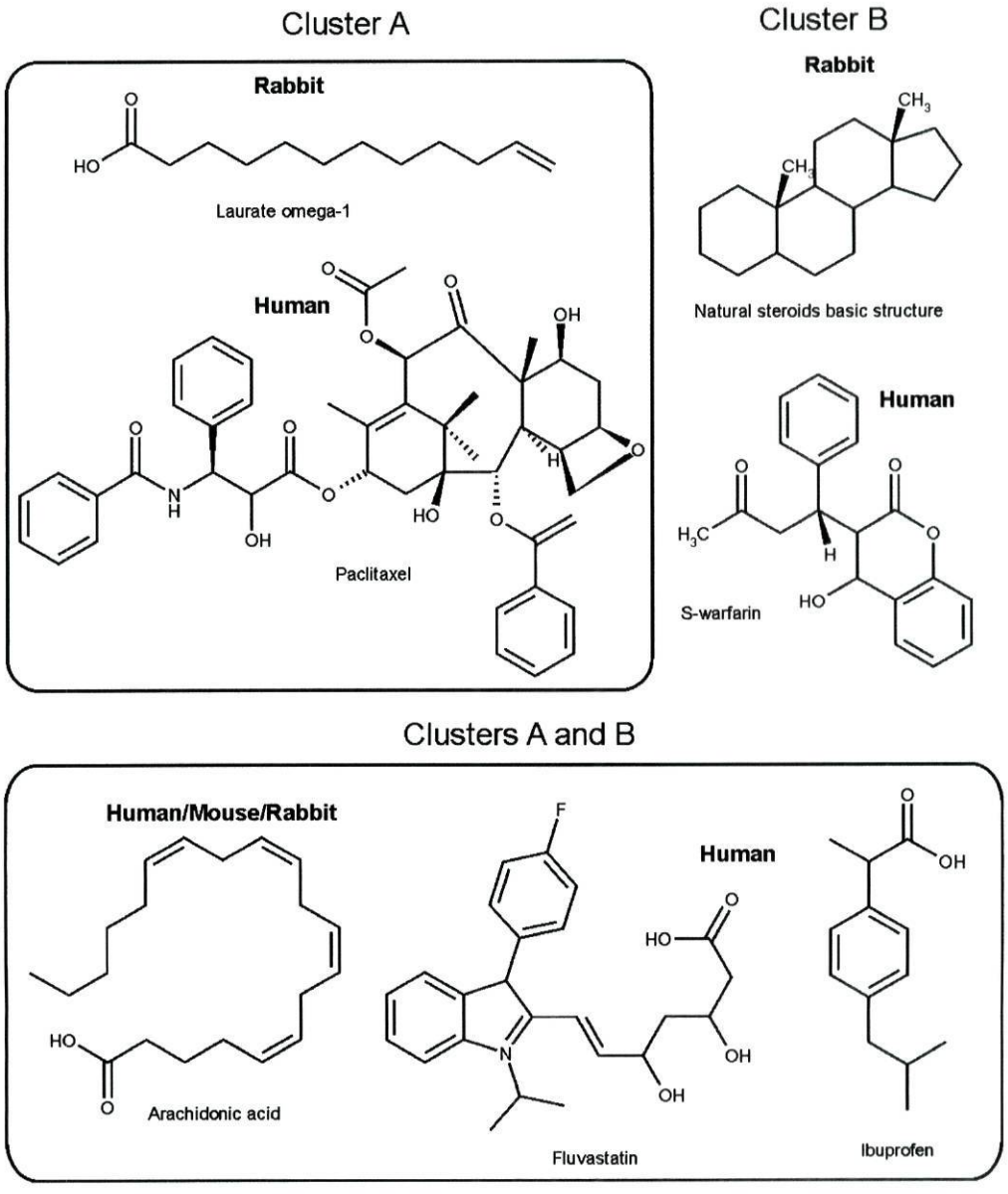


Fig. S1. Examples of substrates metabolized by the CYP2C enzymes of the clusters A and B depicted in Fig. S2.

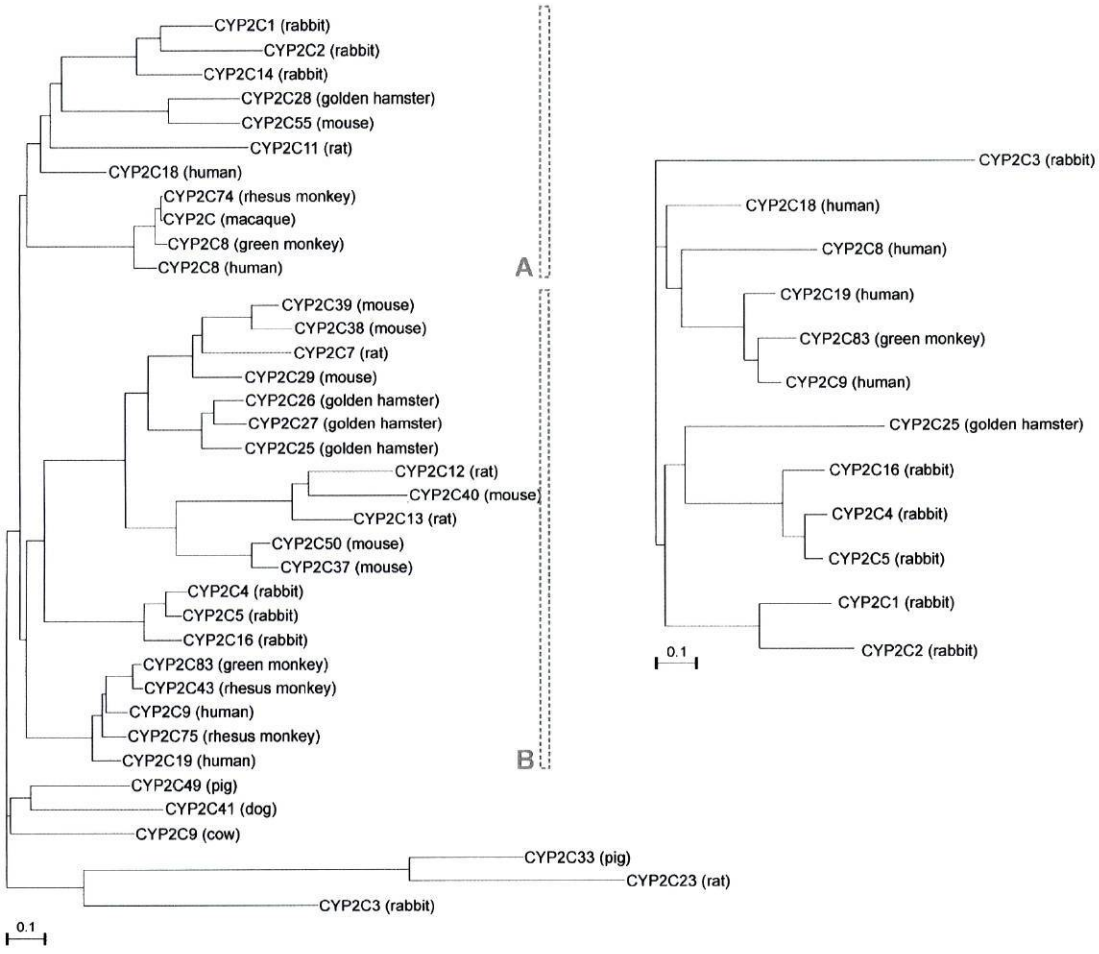


Fig. S2. On the left is the maximum likelihood tree obtained for the large dataset using the TVM+I+G model of sequence substitution and on the right is the maximum likelihood tree obtained for the small dataset using the TVM+G model of sequence substitution. Maximum-likelihood estimates of branch lengths were obtained under the "free-ratios" model implemented in PAML, which assumes an independent ω (d_N/d_S) ratio for each branch in the tree.

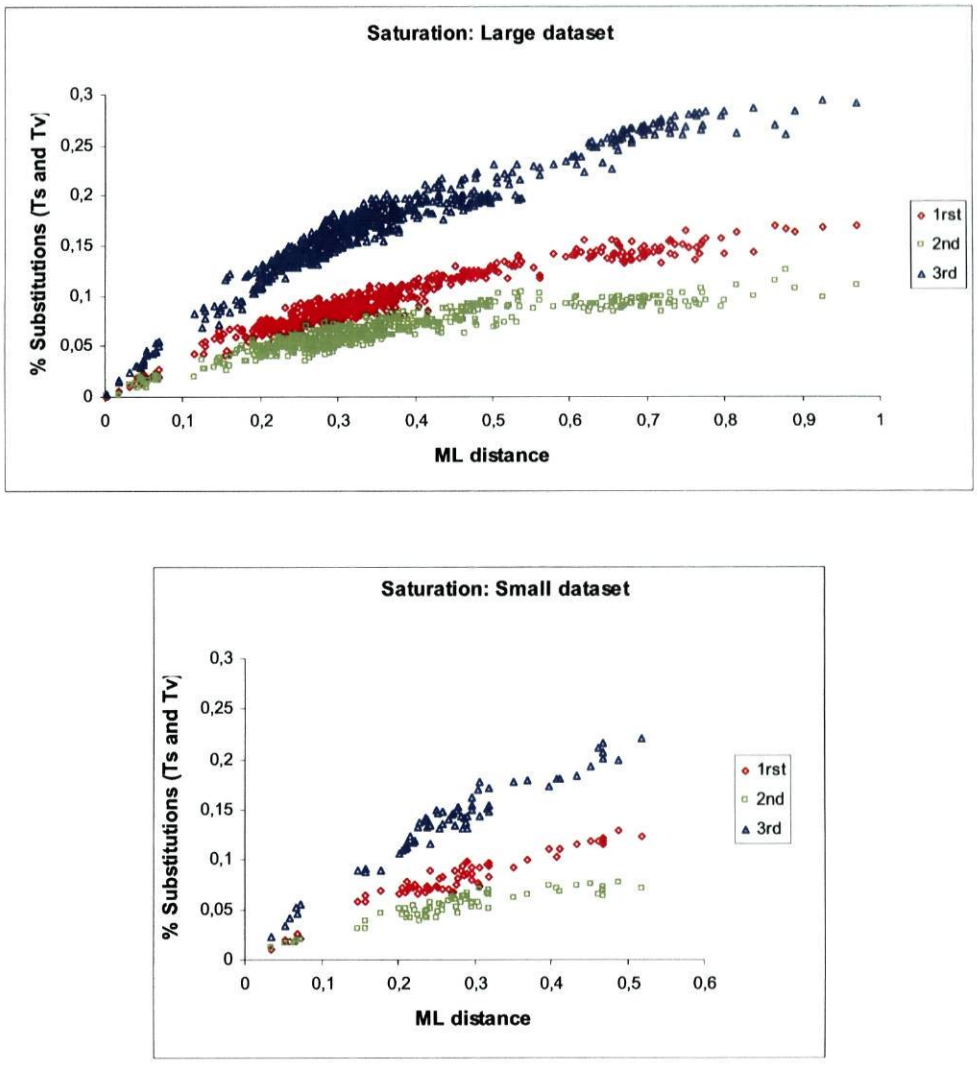


Fig. S3. Saturation plots of transition and transversion calculated through pairwise sequence comparisons. Maximum-likelihood distances calculated using the substitution model that best fitted the data are plotted against the percent number of substitutions in the first, second and third codon positions.

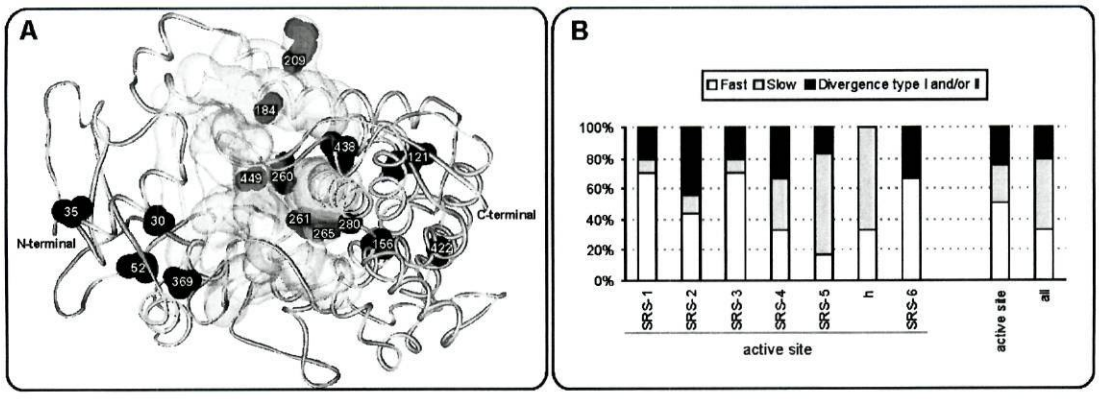


Fig. S4. LRT comparative analysis of the A and B clusters of CYP2C enzymes (Fig. 1). (a) Type I and/or II sites that have a $\Delta U > 2$ are mapped onto the X-ray structure of human CYP2C9 (pdb code: 1OG5); sites are presented in black and active site regions are represented by a transparent white surface. (b) Distribution of the sites that either present divergence type I or/and II or a single rate of evolution which is different from the average (slower or faster evolution) within the active site regions.

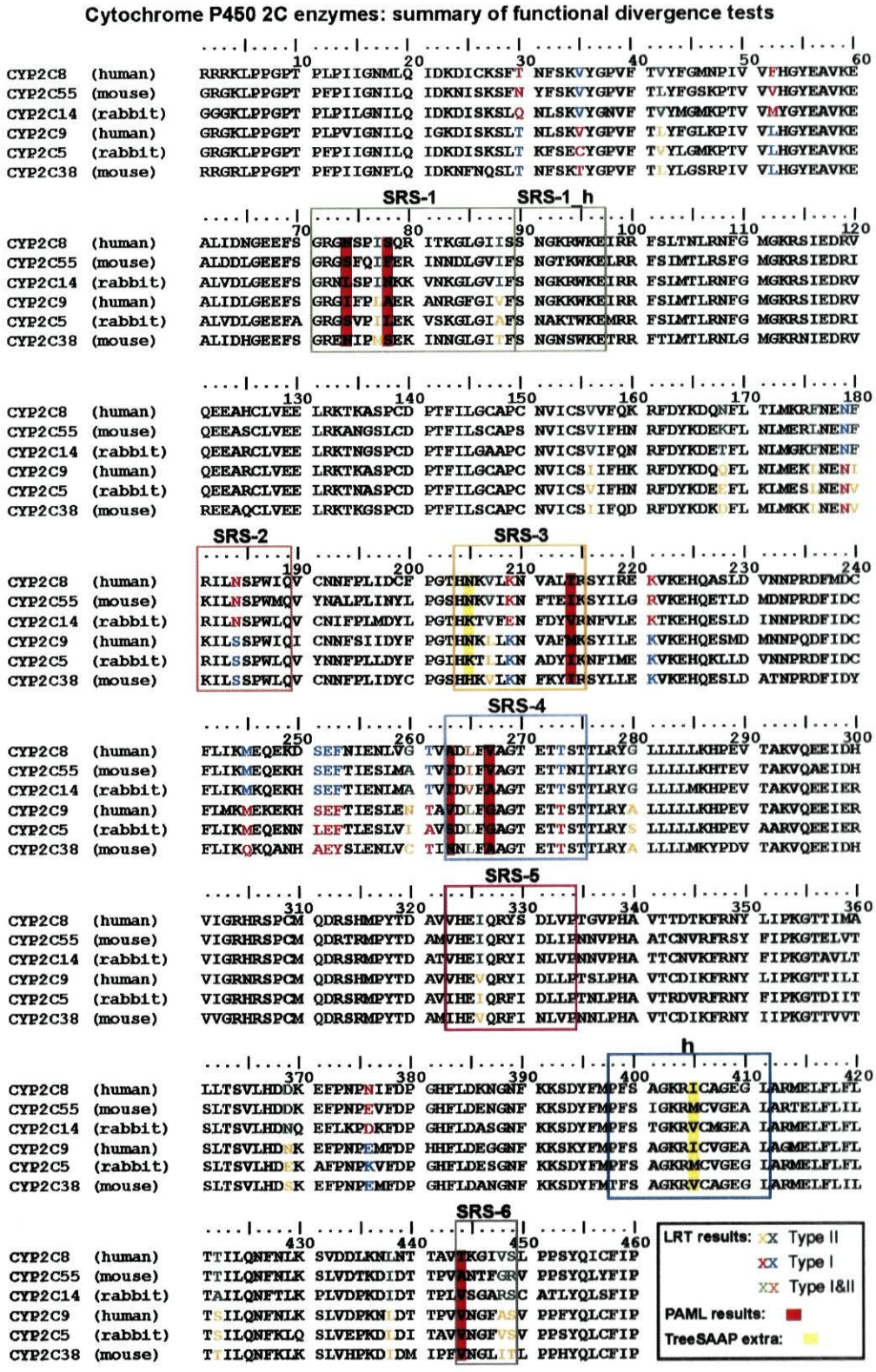


Fig. S5. Summary of the results obtained for positive selection analysis: sample alignment with six of the CYP2C sequences used in this study.

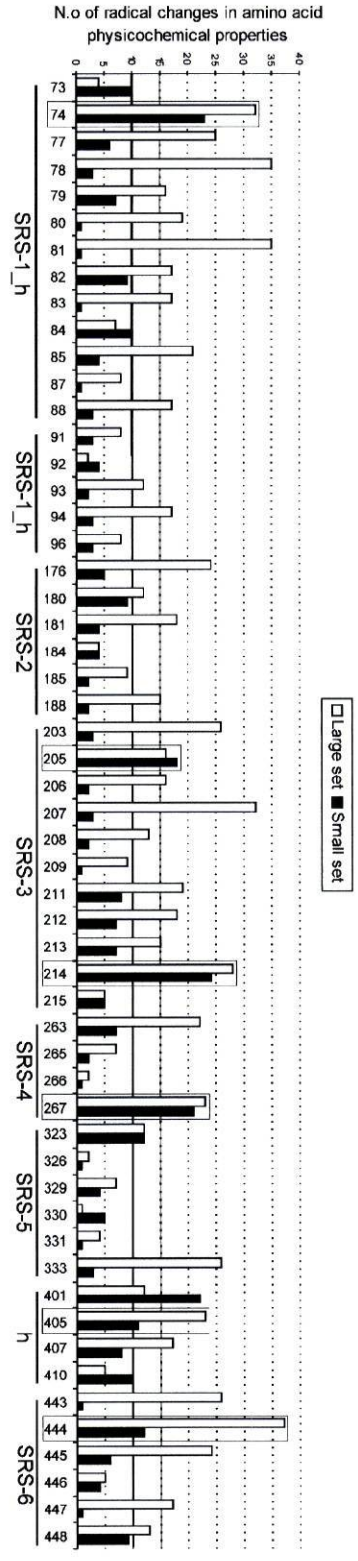


Fig. S6. TreeSAAP results for the residues within the active site areas.

SUPPORTING INFORMATION
- Tables -

Table S1. Examples of proteins under positive selection (* positive selection acting on an enzyme's active site).

Category (number of cases)	Gene/protein	Functional domain under selection/ Function	Lineage	Methods	Ref.
Host-pathogen interactions (27)					
	α-defensins	Active antimicrobial peptide region	Mammals	PAML SWAPSC	(1)
	Cytidine deaminase APOBEC3G	Antiviral enzyme	Primates	Others	(2)
	β-defensins	Small cationic peptides with a broad spectrum of antimicrobial activity	Humans	PAML others	(3)
	CD2 (T-cell and natural killer cell-surface protein)	Extracellular domain	Mammals	PAML	(4)
	CD45 (leukocyte common antigen)	Involved in B- and T-cell maturation and activation	Primates	PAML	(5)
	CD59	Cell-surface protein	Primates	Others	(6)
	Class II MHC – major histocompatibility complex	Cell-surface glycoprotein - supply peptides to helper T cells which then stimulate immune response			(7)
	Eosinophil cationic protein (ECP)	Anti-pathogen function	Human Primates	Others	(8)
	Glycophorin A	Expressed on the erythrocyte surface	Humans Primates	PAML Others	(9, 10)
	HA gene	surface	Influenza virus	AdaptSite	(11)
	HLA – human leukocyte antigen	Antigen recognition sites	Human	AdaptSite	(11)
	Immunoglobulin A	Hinge region, cleaved by pathogenic bacteria proteases	Primates	PAML Others	(12)
	Interleukin-2 (*IL2*)	Cytokine involved in induction and regulation of the immune response in mammals	Mammals	PAML Others	(13, 14)
	Killer cell inhibitory receptors	Extracellular domains	Human	Others	(15)
	Lysozyme	Digestion of bacteria	Primates	PAML Others	(16)
	Omp85, outer membrane protein	Surface loops – most likely to interact with host immune response	Gram-negative bacteria	PAML SWAPSC	(17)
	Protein I	Interaction with host cells/immune system	*N. gonorrhoeae*	PAML TreeSaap	(18)
	Pyrin	Possibly interacts pathogens or moleculaes of the host immune system	Primates	PAML	(19)
	R gene	Proteins that trigger resistance responses possibly by recognizing pathogens gene products	Wild Tomato	PAML Fisher's exact test	(20)
	RB gene product	Recognizes pathogen associated molecular patterns	Potato	PAML	(21)
	RH blood group genes		Humans Primates	PAML	(22)
	Transmembrane protein	Outside of transmembrane region – where interaction with host occurs	Parasitic rickettsiaceae bacteria	PAML	(23)
	TRIM5α gene product	Interaction of unknown type with viruses	Human Primates	PAML	(24)
	Transferrin	Bacteria interaction surface areas	Vertebrates	PAML	(25)
	Toll-like receptor 4	Cell surface receptor that recognizes pathogen associated molecular patterns	Bovine Cattle	PAML	(26)
	VP2 capsid proteins	Interact with host immune response	Carnivore parvovirus	PAML	(27)
	Whole genome		HIV-1	AdaptSite PAML	(11, 28, 29)
Reproduction (12)					
	Chorionic gonadotropin	Critical signal in establishing pregancy	Humans Primates	PAML	(30)
	Chromodomain protein Y (*CDY*)	Involved in sprematogenesis	Primates	Others	(31)
	Female fertilization glycoproteins ZP2 and ZP3	Zona pellucida – sperm-egg interaction	Mammals	PAML	(32)
	Fertilin	Sperm-egg adhesion domains	Mammals	PAML Others	(33)
	Protamine 1	Influence sperm's morphology	Mammals	Others	(9)
	Protamine 2	Influence sperm's morphology	Mammals	Others	(9)
	Sperm associated calcium channel 1 (*CATSPER1*)	Sperm motility	Primates	Others	(34)

	Sperm bindin	Egg-surface receptor recognition	Sea urchin	Others	(35)
	Sperm-ligand zonadhesin	Solvent accessible residues of MAM domain	Mammals	PAML HyPhy CRANN Fisher's exact test	(36)
	Sperm lysin	Binds to VERL molecules of the egg vitelline envelope (fibrous molecules that loose cohesion and allow sperm to enter)	Abalone	PAML Others	(7, 37-40)
	Sperm protein associated with the nucleus on the X chromosome (*SPANX*) genes	Cancer/testis specific antigens	Human	Others	(41)
	Transition protein 2	Necessary for spermatogenesis	Mammals	Others	(9)
Others (17)					
	Alanine:glyoxylate aminotransferase	Mitochondrial targeting sequence	Mammals	PAML	(42)
*	Aldehyde oxidase	Oxidation of aldehydes into acids	Various	PAML	(43)
	Angiogenin	Tumor growth promoter	Primates	Others	(44)
	ASPM – abnormal spindle-like microcephaly associated	Size of human brain	Human	Fisher's exact test	(45)
	Chalcone synthase	Biosynthesis of flavonoids		PAML	(46)
	Cytochrome B		Cetaceans	Treesaap Others	(47)
	Cytochrome C oxidase subunits	Part of the mitochondrial respiratory chain	Humans Primates	Others	(6, 48)
	MADS-box genes (transcriptional regulators)	Changes in coding region leading to phenotypic variation	Plants	PAML	(49)
	Microcephalin	Regulates human brain development	Humans	Others	(50)
	morpheus gene family	Unknown function	Humans Primates	Others	(51)
	MAS-related genes (*MRG*)	Modulation of nociception (sensitivity and/or selectivity to peptide ligands such as opioids)	Primates	Others	(52)
*	Nitrilase	Unknown	Bacteria	PAML	(53)
	Opsin genes	Related to the modulation of absorvance properties of visual pigments	Cichlid fish	PAML	(54)
	Proteorhodopsin	Close or at retinal binding pocket (change wavelength of light absorption)	Marine bacterium	PAML	(55)
	Red and green opsins	Color vision genes	Human Primates	Others	(56)
	Taste receptor	Extracellular regions – presumably involved in tastant-binding	Mammals	PAML Others	(57)
	Tumor suppressor BRCA1 gene	Maintenance of genomic integrity, including recombinational and transcription-coupled DNA repair and in transcription regulation	Primates	PAML	(16)
	Vomeronasal receptor-like	Extracellular domain	Primates	PAML	(58)

Reference List

1. Lynn, D. J., Lloyd, A. T., Fares, M. A. & O'Farrelly, C. (2004) Molecular Biology and Evolution 21, 819-827.
2. Zhang, J. & Webb, D. M. (2004) Hum. Mol. Genet. 13, 1785-1791.
3. Semple, C. A., Rolfe, M. & Dorin, J. R. (2003) Genome Biol 4, R31.
4. Lynn, D. J., Freeman, A. R., Murray, C. & Bradley, D. G. (2005) Genetics 170, 1189-1196.
5. Filip, L. C. & Mundy, N. I. (2004) Mol. Biol. Evol. 21, 1504-1511.
6. Osada, N., Kusuda, J., Hirata, M., Tanuma, R., Hida, M., Sugano, S., Hirai, M. & Hashimoto, K. (2002) Genomics 79, 657-662.
7. Yang, Z. H. & Swanson, W. J. (2002) Molecular Biology and Evolution 19, 49-57.
8. Zhang, J., Rosenberg, H. F. & Nei, M. (1998) PNAS 95, 3708-3713.
9. Wyckoff, G. J., Wang, W. & Wu, C. I. (2000) Nature 403, 304-309.
10. Baum, J., Ward, R. H. & Conway, D. J. (2002) Mol. Biol. Evol. 19, 223-229.
11. Suzuki, Y. & Gojobori, T. (1999) Molecular Biology and Evolution 16, 1315-1328.
12. Sumiyama, K., Saitou, N. & Ueda, S. (2002) Mol. Biol. Evol. 19, 1093-1099.
13. Zelus, D., Robinson-Rechavi, M., Delacre, M., Auriault, C. & Laudet, V. (2000) Journal of Molecular Evolution 51, 234-244.
14. Zhang, J. Z. & Nei, M. (2000) Molecular Biology and Evolution 17, 1413-1416.
15. Hughes, A. L. (2002) Mol Phylogenet Evol 25, 330-340.
16. Yang, Z. H. & Nielsen, R. (2002) Molecular Biology and Evolution 19, 908-917.
17. Fitzpatrick, D. A. & McInerney, J. O. (2005) Journal of Molecular Evolution 60, 268-273.
18. Perez-Losada, M., Viscidi, R. P., Demma, J. C., Zenilman, J. & Crandall, K. A. (2005) Mol. Biol. Evol. 22, 1887-1902.
19. Schaner, P., Richards, N., Wadhwa, A., Aksentijevich, I., Kastner, D., Tucker, P. & Gumucio, D. (2001) Nat Genet 27, 318-321.
20. Caicedo, A. L. & Schaal, B. A. (2004) PNAS 101, 17444-17449.
21. Song, J., Bradeen, J. M., Naess, S. K., Raasch, J. A., Wielgus, S. M., Haberlach, G. T., Liu, J., Kuang, H., Austin-Phillips, S., Buell, C. R. et al. (2003) PNAS 100, 9128-9133.
22. Kitano, T. & Saitou, N. (1999) Journal of Molecular Evolution 49, 615-626.
23. Jiggins, F. M., Hurst, G. D. D. & Yang, Z. H. (2002) Molecular Biology and Evolution 19, 1341-1349.
24. Sawyer, S. L., Wu, L. I., Emerman, M. & Malik, H. S. (2005) PNAS 102, 2832-2837.
25. Ford, M. J. (2001) Molecular Biology and Evolution 18, 639-647.
26. White, S. N., Taylor, K. H., Abbey, C. A., Gill, C. A. & Womack, J. E. (2003) PNAS 100, 10364-10369.

27. Shackelton, L. A., Parrish, C. R., Truyen, U. & Holmes, E. C. (2005) PNAS 102, 379-384.
28. Travers, S. A. A., O'Connell, M. J., McCormack, G. P. & McInerney, J. O. (2005) Journal of Virology 79, 1836-1841.
29. de Oliveira, T., Salemi, M., Gordon, M., Vandamme, A. M., van Rensburg, E. J., Engelbrecht, S., Coovadia, H. M. & Cassol, S. (2004) Genetics 167, 1047-1058.
30. Maston, G. A. & Ruvolo, M. (2002) Molecular Biology and Evolution 19, 320-335.
31. Dorus, S., Gilbert, S. L., Forster, M. L., Barndt, R. J. & Lahn, B. T. (2003) Hum. Mol. Genet. 12, 1643-1650.
32. Swanson, W. J., Yang, Z., Wolfner, M. F. & Aquadro, C. F. (2001) PNAS 98, 2509-2514.
33. Civetta, A. (2003) Molecular Biology and Evolution 20, 21-29.
34. Podlaha, O. & Zhang, J. (2003) PNAS 100, 12241-12246.
35. Metz, E. C. & Palumbi, S. R. (1996) Molecular Biology and Evolution 13, 397-406.
36. Herlyn, H. & Zischler, H. (2005) Mol Phylogenet Evol.
37. Swanson, W. J. & Vacquier, V. D. (2002) Nat Rev Genet 3, 137-144.
38. Galindo, B. E., Vacquier, V. D. & Swanson, W. J. (2003) PNAS 100, 4639-4643.
39. Lee, Y. H., Ota, T. & Vacquier, V. D. (1995) Molecular Biology and Evolution 12, 231-238.
40. Yang, Z. H., Swanson, W. J. & Vacquier, V. D. (2000) Molecular Biology and Evolution 17, 1446-1455.
41. Kouprina, N., Mullokandov, M., Rogozin, I. B., Collins, N. K., Solomon, G., Otstot, J., Risinger, J. I., Koonin, E. V., Barrett, J. C. & Larionov, V. (2004) PNAS 101, 3077-3082.
42. Birdsey, G. M., Lewin, J., Cunningham, A. A., Bruford, M. W. & Danpure, C. J. (2004) Molecular Biology and Evolution 21, 632-646.
43. Rodriguez-Trelles, F., Tarrio, R. & Ayala, F. J. (2003) PNAS 100, 13413-13417.
44. Zhang, J. Z. & Rosenberg, H. F. (2002) Molecular Biology and Evolution 19, 438-445.
45. Evans, P. D., Anderson, J. R., Vallender, E. J., Gilbert, S. L., Malcom, C. M., Dorus, S. & Lahn, B. T. (2004) Hum. Mol. Genet. 13, 489-494.
46. Yang, J., Gu, H. Y. & Yang, Z. H. (2004) Journal of Molecular Evolution 58, 54-63.
47. McClellan, D. A., Palfreyman, E. J., Smith, M. J., Moss, J. L., Christensen, R. G. & Sailsbery, A. K. (2005) Molecular Biology and Evolution 22, 437-455.
48. Jobson, R. W., Nielsen, R., Laakkonen, L., Wikstrom, M. & Albert, V. A. (2004) PNAS 101, 18064-18068.
49. Martinez-Castilla, L. P. & Alvarez-Buylla, E. R. (2003) PNAS 100, 13407-13412.
50. Wang, Y. q. & Su, B. (2004) Hum. Mol. Genet. 13, 1131-1137.
51. Johnson, M. E., Viggiano, L., Bailey, J. A., Abdul-Rauf, M., Goodwin, G., Rocchi, M. & Eichler, E. E. (2001) Nature 413, 514-519.
52. Choi, S. S. & Lahn, B. T. (2003) Genome Res. 13, 2252-2259.
53. Podar, M., Eads, J. R. & Richardson, T. H. (2005) Bmc Evolutionary Biology 5, 42.
54. Spady, T. C., Seehausen, O., Loew, E. R., Jordan, R. C., Kocher, T. D. & Carleton, K. L. (2005) Molecular Biology and Evolution 22, 1412-1422.
55. Bielawski, J. P., Dunn, K. A., Sabehi, G. & Beja, O. (2004) PNAS 101, 14824-14829.
56. Zhou, Y. H. & Li, W. H. (1996) Mol. Biol. Evol. 13, 780-783.
57. Shi, P., Zhang, J. Z., Yang, H. & Zhang, Y. P. (2003) Molecular Biology and Evolution 20, 805-814.
58. Mundy, N. I. & Cook, S. (2003) Molecular Biology and Evolution 20, 1805-1810.

Table S2. Human CYPs haplotypes where polymorphisms occur in the coding regions (information available at http://www.imm.ki.se/CYPalleles/). In bold are represented sites located in the active site regions.

	Allele	Protein change	Enzyme activity	References
CYP2C8	*CYP2C8*2*	I269F		1
	*CYP2C8*3*	R139K; K399R		1
	*CYP2C8*4*	I264M	None	2
	*CYP2C8*5*	Frameshift		3
	*CYP2C8*6*	G171S		4
	*CYP2C8*7*	R186X	None	4
	*CYP2C8*8*	R186G	Decreased	4
	*CYP2C8*9*	K247R		4
	*CYP2C8*10*	K383N		4
CYP2C9	*CYP2C9*2A*	R144C	Decreased	5; 6; 7; 8; 9
	*CYP2C9*2B*	R144C	Decreased	9
	*CYP2C9*2C*	R144C	Decreased	9
	*CYP2C9*3A*	**I359L**	**Decreased**	9; 10; 11; 12; 13; 14; 15
	*CYP2C9*3B*	**I359L**	**Decreased**	9; 15
	*CYP2C9*4*	**I359T**		16
	*CYP2C9*5*	**D360E**	**Decreased**	17; 18
	*CYP2C9*6*	Frame shift	None	19
	*CYP2C9*7*	L19I		20
	*CYP2C9*8*	R150H	Increased	20
	*CYP2C9*9*	H251R		20
	*CYP2C9*10*	E272G		20
	*CYP2C9*11A*	R335W	Decreased	9; 20; 21
	*CYP2C9*11B*	R335W	Decreased	9
	*CYP2C9*12*	P489S	Decreased	22
	*CYP2C9*13*	L90P		23
	*CYP2C9*14*	R125H	Decreased	23; 24
	*CYP2C9*15*	S162X	None	23; 24
	*CYP2C9*16*	**T299A**	**Decreased**	23; 24
	*CYP2C9*17*	P382S		23; 24
	*CYP2C9*18*	**I359L; D397A**	**Decreased**	23; 24
	*CYP2C9*19*	Q454H		23; 24
	*CYP2C9*20*	G70R		23
	*CYP2C9*21*	P30L		25
	*CYP2C9*22*	N41D		25
	*CYP2C9*23*	V76M		25
	*CYP2C9*24*	**E354K**		26
CYP2C19	*CYP2C19*1B*	I331V	Normal	27
	*CYP2C19*1C*	I331V	Normal	28
	*CYP2C19*2A*	splicing defect; I331V	None	29
	*CYP2C19*2B*	E92D; splicing defect; I331V	None	30
	*CYP2C19*2C* *(CYP2C19*21)*	A161P, splicing defect, I331V		31
	*CYP2C19*3A*	W212X; I331V	None	32
	CYP2C19*3B *(CYP2C19*20)*	W212X; D360N; I331V		31
	*CYP2C19*4*	GTG initiation	None	33

	codon; I331V		
*CYP2C19*5A*	**R433W**	**None**	34; 35
*CYP2C19*5B*	I331V; **R433W**	**None**	35
*CYP2C19*6*	R132Q; I331V	None	30
*CYP2C19*7*	splicing defect	None	36
*CYP2C19*8*	**W120R**	**None/Decreased**	36
*CYP2C19*9*	R144H; I331V		28
*CYP2C19*10*	P227L; I331V		28
*CYP2C19*11*	R150H; I331V		28
*CYP2C19*12*	I331V; X491C; 26 extra aa	Unstable	28
*CYP2C19*13*	I331V; R410C		28
*CYP2C19*14*	L17P; I331V		28
*CYP2C19*15*	I19L; I331V		28
*CYP2C19*16*	R442C		37
*CYP2C19*18*	R329H; I331V		31
*CYP2C19*19*	S51G; I331V		31

Reference List

1. Pharmacogenetics. 2001 Oct;11(7):597-607
2. Biochem Pharmacol. 2002 Dec 1;64(11):1579-89
3. Drug Metab Pharmacokinet. 2002;17(4):374-7
4. Drug Metab Dispos. 2005 May;33(5):630-6
5. Pharmacogenetics. 1994 Feb;4(1):39-42
6. Pharmacogenetics. 1997 Jun;7(3):203-10
7. Blood. 2004 Apr 15;103(8):3055-7.
8. Drug Metab Dispos. 2004 May;32(5):484-9
9. Pharmacogenetics. 2004 Dec;14(12):813-22
10. Pharmacogenetics. 1996 Aug;6(4):341-9
11. Arch Biochem Biophys. 1996 Sep 15;333(2):447-58
12. Lancet. 1999 Feb 27;353(9154):717-9
13. Pharmacogenetics. 1999 Feb;9(1):71-80
14. Pharmacogenetics. 2000 Mar;10(2):95-104
15. Clin Pharmacol Ther. 2001 Aug;70(2):175-82
16. Pharmacogenetics. 2000 Feb;10(1):85-9
17. Mol Pharmacol. 2001 Aug;60(2):382-7
18. Clin Pharmacol Ther. 2004 Aug;76(2):113-8
19. Pharmacogenetics. 2001 Dec;11(9):803-8
20. Pharmacogenetics. 2004 Aug;14(8):527-37
21. JAMA. 2002 Apr 3;287(13):1690-8
22. Pharmacogenetics. 2004 Jul;14(7):465-9
23. Clin Pharmacol Ther. 2004 Sep;76(3):210-9
24. J Pharmacol Exp Ther. 2005 Dec;315(3):1085-90
25. Clin Pharmacol Ther. 2005 May;77(5):353-64
26. Herman *et al.*, manuscript in preparation
27. Arch Biochem Biophys. 1995 Oct 20;323(1):87-96
28. Pharmacogenetics. 2002 Dec;12(9):703-11
29. J Biol Chem. 1994 Jun 3;269(22):15419-22
30. J Pharmacol Exp Ther. 1998 Sep;286(3):1490-5
31. Drug Metab Pharmacokinet. 2005 Aug;20(4):300-7
32. Mol Pharmacol. 1994 Oct;46(4):594-8
33. J Pharmacol Exp Ther. 1998 Jan;284(1):356-61
34. J Pharmacol Exp Ther. 1997 Apr;281(1):604-9
35. Pharmacogenetics. 1998 Apr;8(2):129-35
36. J Pharmacol Exp Ther. 1999 Aug;290(2):635-40
37. Drug Metab Pharmacokinet. 2004 Jun;19(3):236-8

Table S3. Sequences and species used in the study. The large dataset contained all the sequences and the small dataset contained the asterisked enzymes.

Common name	Species name	Enzyme name	GenBank Acession no.
cow	*Bos taurus*	CYP2C9	XM_612374
dog	*Canis familiaris*	CYP2C41	AF016248
green monkey	*Cercopithecus aethiops*	CYP2C8; CYP2C83*	DQ022200; DQ022201*
human	*Homo sapiens*	CYP2C18*; CYP2C19*; CYP2C8*; CYP2C9*	M61853*; M61854*; NM_000770*; NM_000771*
macaque	*Macaca fascicularis*	CYP2C	S53046
rhesus monkey	*Macaca mulatta*	CYP2C43; CYP2C74; CYP2C75	AB212264; AY635462; AY635463
golden hamster	*Mesocricetus auratus*	CYP2C25*; CYP2C26; CYP2C27; CYP2C28	X63022*; D11435; D11436; D11437
mouse	*Mus musculus*	CYP2C29; CYP2C37; CYP2C38; CYP2C39; CYP2C40; CYP2C50; CYP2C55	D17674; NM_010001; NM_010002; AF047726; AF047727; NM_134144; AK008580
rabbit	*Oryctolagus cuniculus*	CYP2C1*; CYP2C2*; CYP2C3*; CYP2C4*; CYP2C5*; CYP2C14; CYP2C16*	K01522*; M19137*; D26152*; J02716*; M55664*; D00190; M29968*
rat	*Rattus norvegicus*	CYP2C7; CYP2C11; CYP2C12; CYP2C13; CYP2C23	BC097939; BC088146; BC089790; J02861; U04733
pig	*Sus scrofa*	CYP2C33; CYP2C49	NM_214414; AB052258

Table S4. Correspondence of the sites numbering to the sequence numbers of the PDB Databank files. (D = Functional divergence; P = positive selection)

Location	Property	Site	rabbit CYP2B4	rabbit CYP2C5	human CYP2C8	human CYP2C9
	D	**30**	56	55	55	55
	D	**35**	61	60	60	60
	D	**42**	68	67	67	67
	D	**52**	78	77	77	77
(SRS-1)	P	**74**	100	99	99	99
(SRS-1)	D	**77**	103	102	102	102
(SRS-1)	P	**78**	104	103	103	103
(SRS-1)	D	**88**	114	113	113	113
	D	**156**	182	181	181	181
	D	**168**	194	193	193	193
(SRS-2)	D	**176**	202	201	201	201
(SRS-2)	D	**179**	205	204	204	204
(SRS-2)	D	**180**	206	205	205	205
(SRS-2)	D	**184**	210	209	209	209
(SRS-3)	P	**205**	232	231	231	231
(SRS-3)	D	**207**	234	233	233	233
(SRS-3)	D	**209**	236	235	235	235
(SRS-3)	P	**214**	241	240	240	240
	D	**221**	248	247	247	247
	D	**245**	272	271	271	271
	D	**251**	278	277	277	277
	D	**252**	279	278	278	278
	D	**253**	280	279	279	279
	D	**260**	287	286	286	286
	D	**261**	288	287	287	287
(SRS-4)	P	**263**	293	289	292	292
(SRS-4)	D	**265**	295	291	294	294
(SRS-4)	P	**267**	297	293	296	296
(SRS-4)	D	**273**	303	299	302	302
	D	**280**	310	306	309	309
(SRS-5)	D	**326**	356	352	355	355
	D	**369**	399	395	398	398
	D	**376**	406	402	405	405
hemo	P	**405**	435	431	434	434
	D	**422**	452	448	451	451
	D	**438**	468	464	467	467
(SRS-6)	P	**444**	474	470	473	473
(SRS-6)	D	**448**	478	474	477	477
	D	**449**	479	475	478	478

Table S5. Division of the 31 physicochemical properties evaluated by TreeSaap according to their effect on protein properties.

Effect on protein properties	TreeSaap physicochemical properties
structural	α-helical tendencies; β-structure tendencies; coil tendencies; helical contact area; power to be at the middle of the α-helix; turn tendencies; average number of surrounding residues; bulkiness; compressibility; mean RMS fluctuation displacement; molecular volume; partial specific volume
chemical	buriedness; chromatographic index; equilibrium constant; hydropathy; isoelectric point; long-range nonbonded energy; normalized consensus hydrophobicity; polarity; polar requirement; short-range and medium-range nonbonded energy; refractive index; solvent accessible reduction ratio; surrounding hydrophobicity; thermodynamic transfer hydrophobicity; total nonbonded energy
others	composition; molecular weight; power to be at the C-terminal; power to be at the N-terminal

Table S6. PAML results for the branch models. A: likelihood-ratio tests that compare the one ratio model with Model A (in which the ω values of the branches that showed ω > 1 in the free-ratios were a free parameter); B: likelihood-ratio tests that compare the lnL of Model A where the ω values of the branches that showed ω > 1 in the free-ratios were a free parameter with a test were the same branches had their value ω = 1.

	A			B		
	MbranchX vs M0			MbranchX vs MbranchX-w1		
	2(DlnL)	df	*p*	2(DlnL)	df	*p*
free-ratios	154,1	73	9,98E-08			
branchA	3,3	1	6,8E-02	0,5	1	0,46
branchB	3,6	1	5,7E-02	0,5	1	0,50
branchC	5,0	1	2,5E-02	0,7	1	0,41
branchD	6,3	1	1,2E-02	1,9	1	0,17
branchE	1,4	1	2,4E-01	0,5	1	0,50
branchF	7,6	1	5,9E-03	1,6	1	0,20
branch-all	27,3	6	1,3E-04	5,5	6	0,49
branch-all-1omega	26,8	1	2,3E-07	4,9	1	0,03

Table S7. Number of radical changes in 31 amino acid properties provided by TreeSAAP for all sites (AVE = average).

		Large set					Small set				
	n.o sites	Struct	Chem	Others	All	AVE per site	Struct	Chem	Others	All	AVE per site
SRS-1	19	130	133	40	303	15.9	39	32	8	79	4.2
SRS-1_h	8	20	26	7	53	6.6	4	7	4	15	1.9
SRS-2	14	59	43	19	121	8.6	14	12	5	31	2.2
SRS-3	13	76	98	26	200	15.4	35	32	13	80	6.2
SRS-4	13	30	33	18	81	6.2	18	10	3	31	2.4
SRS-5	12	28	15	11	54	4.5	9	6	11	26	2.2
heme	14	25	24	11	60	4.3	32	7	12	51	3.6
SRS-6	7	66	55	15	136	19.4	14	15	4	33	4.7
active site	100	434	427	147	1008	10.1	165	121	60	346	3.5
total	460	1194	1415	519	3128	6.8	418	486	183	1087	2.4

3. Concluding Remarks

The results presented in this thesis shed some light into the complex and wide biological universe of Cytochrome P450 enzymes.

The initial choice to use CYP1A2 as a case study was made within the goals of the agency that financed our project, NFRC:

"The National Foundation for Cancer Research (NFCR) was founded in 1973 to support cancer research and public education relating to the prevention, early diagnosis, better treatments and ultimately, a cure for cancer. NFCR promotes and facilitates collaboration among scientists to accelerate the pace of discovery from bench to bedside. NFCR is committed to ***Research for a Cure*** *– cures for all types of cancers."*

CYP1A2 enzyme has been implicated in the development of tumours in various tissues, and its carcinogenic activity is related to the activation of heterocyclic amines, compounds present in cooked red meat. This activity was the first issue focused in this PhD work.
Homology models for the rat and human enzymes were built and their interaction with two heterocyclic amines was evaluated. The models were thoroughly refined so that they presented a good stereochemistry and explained mutation experimental data. The difference in the metabolites produced by the two isozimes was shown to be related with differences in the active sites. Point mutations were found in key locations in the active site: $Glu318_{rat}/Asp320_{human}$ in SRS-4 and $Ser222_{rat}/Thr223_{human}$ in SRS-2 (X-ray numbering of positions $Glu279_{rat}/Asp280_{human}$ in SRS-4 and $Ser183_{rat}/Thr183_{human}$ referred to in the results of Article I, respectively). These deeply influenced the binding to the HAs in a way that explained differences in the catalytic efficiency of both enzymes towards the same substrates. Such comparative analysis of human and rat enzymes at molecular level is very useful for pharmacological essays, as it is important to be aware of how different are the animal models and the human systems of study.

This human model was further used in ligand binding studies involving flavonoids, natural inhibitors of CYP1A2. These studies intended to provide information on the molecular features that are mostly responsible for the inhibition of the enzyme, in order to use the flavonoids as scaffolds in future drug design and development strategies.
In both studies we obtained a correlation between calculated stabilization energies and experimentally determined binding constants. Both flavone derivatives and

flavanones/chalcone ligands are bulkier than the previously studied HAs and interact with different amino acid residues in the active site.

An evaluation of the molecular electrostatic pattern of the ligands allowed the withdrawal of the first assumptions on what characteristics are important for inhibition. One of such characteristics, the location of a large negative electrostatic potential region on flavones opposite the atom that should be closer to the heme, was shown to be related to the most important electrostatic interactions stabilizing these ligands. These are held by residues located on the top of the active site: Tyr112, Asn234 and Thr498. The larger flavanones/chalcone ligands share such interactions and have some extra at the bottom of the active site, with Thr124 and Asp313. The different electrostatic interactions established by each ligand were the main contribution to the differential inhibition of CYP1A2 by the groups of flavonoids studied. The flavanones/chalcone ligands further exhibit a non-polar stabilizing interaction with Phe125 through their prenyl tail. This amino acid residue also seems to be important in the docking of these partially flexible ligands, as they literally wrap around its side-chain in the active site.

In the last part of this PhD, the experience gained in the previous studies in exploring protein structure and function was used to dig into one very important aspect of the Cytochrome P450 family as a whole: its diversity in substrate specificity. The starting point was the fact that this superfamily is involved in the metabolism of various xenobiotics compounds, such as drugs and carcinogenic compounds. Bearing this in mind, an attempt was made to verify whether the variation in these environmentally available and always-diverse compounds was causing accelerated molecular changes in the CYPs that handle them. The study was based in the use of statistical methods that detect if the fixation of amino acid mutations is being driven by adaptive evolution. CYP2 family, the biggest, was the target of this research, as besides playing an important role in drug disposal there are four available X-ray structures for mammalian enzymes belonging to this family.

Functional divergence and positive selection were indeed found to be interfering with CYP2C subfamily molecular evolution. Positive selection is mostly affecting the active site areas responsible for substrate binding, which presented high rates of mutation in all analyses. The heme binding pocket was always found to be extremely conserved, as it was expected from the conserved catalytic mechanism of CYPs. The evaluation of the mutation rates was done in parallel with a physicochemical assessment of the amino acid changes, and their impact on the available 3D structure of CYP2 enzymes.

In the end, a thorough and unusual approach of exploring the relationship between molecular evolutionary events and protein structure and function was described, presenting an important contribution for future evolutionary biology studies. Furthermore, this was the first study ever presenting evidences that molecular adaptation is driving CYPs evolution. Such information is of great interest for biochemists and clinical researchers, as well as evolutionary biologists.

In summary, this PhD work consisted on the use of various computational chemistry approaches, quantitative and qualitative, from quantum to molecular mechanics, from protein homology modelling to docking and classical electrostatics, ending with an insightful application of computational genomics methodologies. The methods were used complementarily and were successful in providing various types of clinically/biologically relevant information regarding the broad Cytochrome P450 superfamily.

This knowledge can be used in future studies combined with other approaches to explore the pharmacological profile of CYPs. Using the flavonoid structure as an initial scaffold, one can built different inhibitors using the characteristics that were shown to favour inhibition. Molecular dynamics can be used to better characterize the conformation of the molecular complexes formed between the enzyme and the different inhibitors. Eventually, the binding energies can be used to choose the best set of newly designed inhibitors. As far as the molecular evolution studies are concerned, other CYP families can be studied using the same methodology. It would be particularly interesting to compare the results obtained for a group of enzymes that is mainly involved in xenobiotic metabolism and another that metabolizes a narrow range of compounds or a specific substrate.