



New methods in predictive metabolism

Scott Boyer* and Ismael Zamora**

*Chemical Computing, Enabling Sciences and Technologies, AstraZeneca R&D, 431 83 Mölndal, Sweden; Tel: +46 31 776 2882; **Lead Molecular Design,s.l., Francesc Cabanes i Alibau, 1–2 2^o-1^a Sant Cugat del Valles, 08190 Barcelona, Spain, Tel: + 34 93 674 90 25

General introduction

Target validation, and particularly in vivo target validation, has always been one of the major bottlenecks in the process of drug discovery. Quickly optimising ligands for their in vivo pharmacokinetic properties is one of the basic requirements for enhancing the efficiency of this notoriously slow and problematic process.

In the process of optimising the pharmacokinetic profile of a series, metabolic stability is a basic requirement. Often, high metabolic clearance is observed and invariably limits the systemic exposure, duration of action and ultimately the usefulness of newly identified ligands for in vivo efficacy testing in the critical early target validation efforts. Addressing metabolic stability problems most often involves identification of the site of metabolism. This information is then used to modify the structure in a way that will eliminate the site or protect the site in some way (steric or electronic). New methods are being developed constantly to identify metabolites with greater speed and accuracy [1], however this is still a tedious process requiring specialized personnel and instrumentation.

Whilst most Drug Discovery projects have to address issues of metabolic stability, reducing metabolic clearance is only one of the goals of predictive metabolism methods. For reducing patient-to-patient pharmacokinetic variation, clearance of any drug-like molecule should ideally be via several pathways, including metabolism and renal and biliary excretion. Likewise, dependence on a single metabolic pathway or protein can also be a source of variation and the metabolic component of clearance should therefore also be as diverse as possible. Thus new tools for predicting metabolism need to address not only issues of metabolic site to repair high clearance but also need to predict routes of metabolism, i.e., iso-

forms. Some added benefits of these kinds of modelling tools are that prodrug design could be facilitated and that *a priori* knowledge of metabolic sites would aid in the design of metabolite structure elucidation experiments.

Site of metabolism prediction has, in the past, taken three basic forms, 1) rule-based systems that have libraries of molecular fragments reported to be involved in metabolic transformations [2, 3, 4] and 2) pharmacophore-based methods [5, 6] in which a site is identified that is consistent with proximity to the cytochrome P450 reactive center and 3) site-based reactivity calculations [7]. These three approaches are not mutually exclusive. Rather, they are complementary and need to be combined to produce more robust models. Here we review our recent work to improve on these established methods using new tools and new approaches to old datasets.

Probabilistic scoring of reactions in the MDL metabolite database

Overview

This section will focus on methods developed to mine useful knowledge about metabolic transformations out of a widely available metabolic transformation database, Metabolite curated and marketed by MDL [8].

The MDL Metabolite database is a collection of metabolic transformations curated from the literature and regulatory documentation. The current release 2001.2 of Metabolite contains well over 57000 individual metabolic transformation records involving over 9000 substrates in all major species. Roughly 45% of the transformation records in Metabolite are from rat, approximately 30% are from human and about 10% are from dog, the remaining approxi-

mately 15% of the records arising from the remaining species. The database reports both positively confirmed metabolites (those established either by both Mass Spectrometry and $^1\text{H-NMR}$ or synthetic standard) and those that are 'putative' (those that are from only mass spectrometric data or are conjectural).

The most common use of Metabolite is the assessment of whether a metabolic transformation will occur, i.e., is this substructure involved in any sort of metabolic reaction? Most commonly the substructure of interest is entered into the search window and the results returned from the database are used to assess the variety of metabolic transformations that particular substructure that could undergo. In working with the database in this manner, it was noticed that certain transformations return a large number of 'hits' and others returned very few. The occurrence in the database could be due to a very common metabolic pathway, a very common substructure, or a very stable, and thus easily identified, metabolite. The atom-based searching methods outlined below will address the first possibility, controlling for the number of times a given substructure occurs in the database.

This atom-based scoring method differs from the rule-based methods in that it is dynamic – depending solely on the database in which it is being used. Thus one could imagine that using these kinds of method, databases like metabolite could be partitioned by species and/or isoform and dynamic reaction frequencies could be mined from these separate databases for comparison. This method also opens the possibility of introducing and mining knowledge from the proprietary biotransformation data within an organization.

Fragment-based searching and scoring methods

The theory of atom scoring originates from the frequency of reported specific reactions at a particular atom. Since metabolic reactions at a particular atom are influenced by the neighbouring 3–4 atoms, a fragment-based approach has been employed. Also, where possible, rings have not been broken in the fragmentation schemes. Once formed, the fragments are subjected to metabolic reactions based on possible metabolic transformations. Possible metabolic reactions can be mined directly from Metabolite or can be obtained from other sources such as Meta (2), METEOR (3), or MetabolExpert (4).

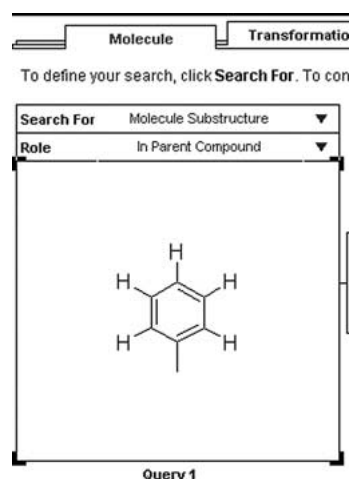


Figure 1. Fragment search window from Metabolite Database.

Once generated, the fragments and reactions are entered into the Metabolite search windows under the following conditions:

Fragment Search: The fragment search is a standard substructure search with the exception that hydrogens are specified explicitly (Figure 1).

Reaction Search: The reaction search searches against the subset of the database retrieved by the Fragment SSS. This is done by simply leaving the fragment in the 'Molecule' window while performing the reaction search (Figure 2).

If one performs this search for hydroxylations in the para, meta and ortho positions, the table of results will be that shown in Table 1. The first column of table one represents the sum of *transformation* occurrences in the database of each hydroxylation event at position i (Σt_i). The second column represents the sum of occurrence of the open phenyl ring (*substrate*) in the database and is similar for all three searches (Σs_i). The third column is then the ratio of the transformation occurrences to the substrate occurrences ($\Sigma t_i / \Sigma s_i$) and finally the forth column is the 'normalized' occurrence ratio. Thus, the normalized occurrence ratio (P_i) can be expressed as Equation 1 where ($\Sigma t_\alpha / \Sigma s_\alpha$) is the highest occurrence ratio in the molecule of interest and is used to normalize all other occurrence ratios for that given compound, in this case for the open phenyl ring, searching the entire metabolite database results in a high occurrence ratio at the para position with lower occurrences at the meta and ortho positions.

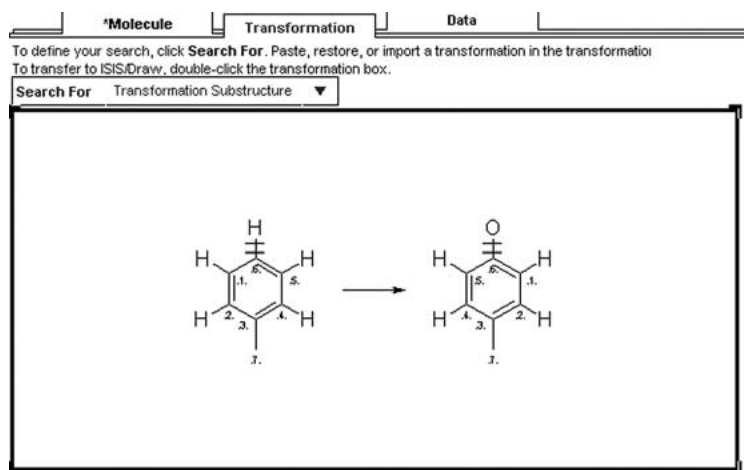


Figure 2. Transformation search window from Metabolite Database.

$$p_i = \frac{\left[\frac{\sum t_i}{\sum s_i} \right]}{\left[\frac{\sum t_\alpha}{\sum s_\alpha} \right]} \quad (1)$$

If one assigns the colors red, yellow and green to atoms with normalized occurrence ratios of > 0.67 , $0.33\text{--}0.67$ and < 0.33 , respectively, these data can then be transferred back onto the atomic positions such that an easily readable map of the database results can be presented. The sites identified by this method are meant to be starting points for discussions with Medicinal Chemists trying to prioritise synthesis. Further automation of this procedure will be required to thoroughly test its predictive power and whether it can distinguish species-specific biotransformation reactions.

An example and conclusions

Using a real drug-like molecule such as H234/09 (Figure 3), the searches and results summarized in tabular form are given in Table 2. Interestingly, this search points out quite clearly that the three methylenes surrounding the tertiary amine are not equivalent (Table 2). Results are summarized by color-coding the positions searched in the database in Figure 4). Indeed, when one looks at the actual phase I metabolic pattern from rat microsomes (Figure 5), the results are of a high enough quality to allow a Medicinal Chemist to assess the most frequently reported sites of metabolism and take steps to modify them, if necessary.

This method is meant as a guide to the Medicinal Chemist facing metabolic clearance issues. It is essen-

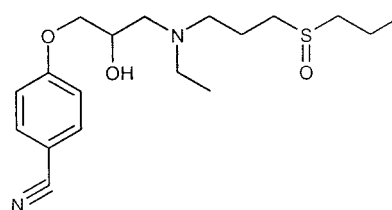


Figure 3. Structure of test example compound, H234/09.

tially an extension of the rule-based systems outlined above [2, 3, 4]. It applies the traditional metabolism rules to a molecule and adds information of the likelihood of the reaction occurring based on its reported frequency in a database. The database does not necessarily have to be Metabolite, but could be an database of internal biotransformation information accumulated within a company.

Site of metabolism prediction: ALMOND pharmacophore approach

Overview

The pharmacophore development for predicting the site of metabolism is reported as a certain geometrical constraints with respect to other chemical features present in the molecule. There are several techniques that could be used to develop a pharmacophore model. The most common ones are: based on a set of known substrates (indirect approach) and based on the protein structure responsible for the metabolic activity (structure based design). In both cases the pharmacophore is built to describe the interaction of a compound with

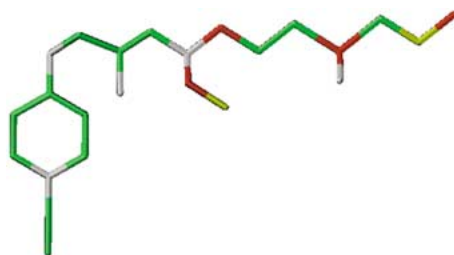


Figure 4. Actual oxidative metabolic pathways for H234/09. Major metabolism routes are denoted with bold arrows.

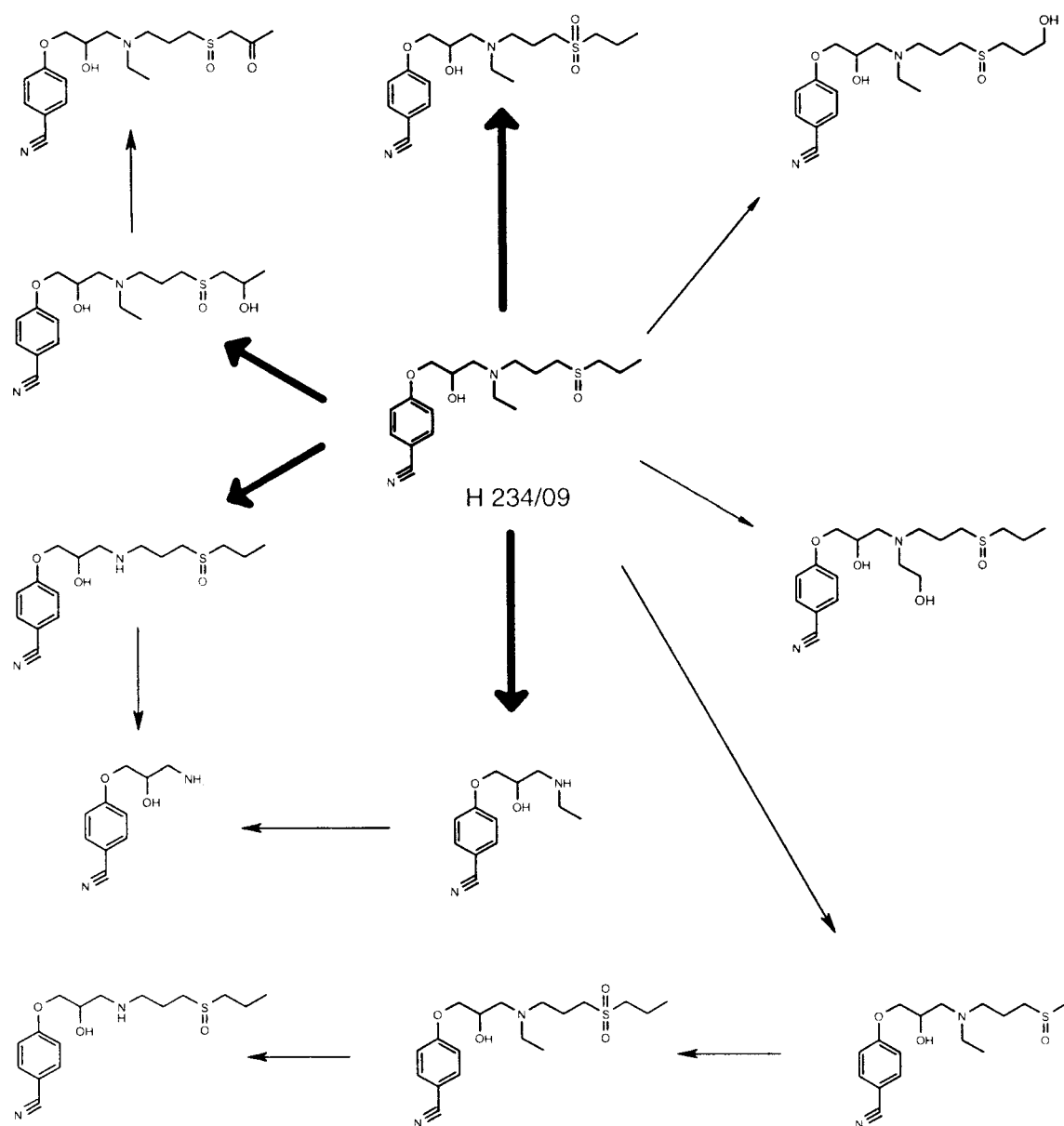

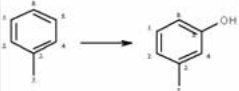
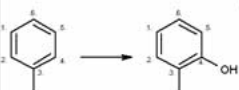


Figure 5. Atomic positions identified in Table 2 color coded by reaction frequency.

Table 1. Example of search routine for positional hydroxylation reactions on toluene substructure.

| Transformation | Transform Occur. | Substrate Occur. | Occur. Ratio | OR _{normalized} |
|---|------------------|------------------|--------------|--------------------------|
|  | 767 | 5859 | 0.1309 | 1.0000 |
|  | 71 | 5859 | 0.012 | 0.0926 |
|  | 34 | 5859 | 0.006 | 0.0443 |

certain protein target. In the case of the metabolic reactions there are great variety of proteins that produces a change in the metabolic clearance of a compound. Nevertheless, the most important family of proteins are the cytochrome P450 which are involve mainly in oxidative process of drug-like compounds. Consequently, the development of pharmacophore models has been focussed in the description of these enzymes.

The indirect method is based on the comparison of the different substrates structures. This comparative analysis can be made by:

1. Considering the different atomic and/or pseudoatomic positions, which represent the interaction pattern of the compound and overlaying all the known substrates, i.e. DISCO procedure
2. Using a representation of the molecular interaction pattern by a field sourcing the compound, followed by statistical analysis, i.e. GRID-GOLPE approach (9). This type of 3D-QSAR techniques also requires the molecular structure alignment.
3. Using a molecular representation were only the internal co-ordinates of interaction pattern are considered, i.e. ALMOND.

These techniques have been applied on the pharmacophore calculation to predict the site of metabolism and/or the affinity of a compounds to certain cytochrome P450 (5,6, 10–12).

The structure based method used the structure of the protein or a model of the structure to develop which is the geometry of the interaction inside the protein with respect to the reactive center. In the case of the cytochrome superfamily of enzymes there are several bacterial structure crystalized, i.e. cytochrome

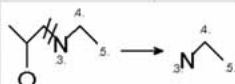
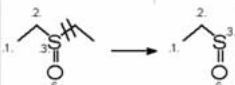
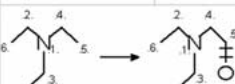
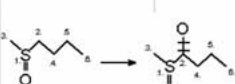
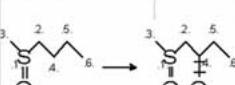

P450 cam. Nevertheless, this structures have low protein similarity and identity with the protein which are relevant for the metabolism in humans, mainly CYP 1A2, CYP 3A4, CYP 2C9, CYP 2D6 and CYP 2E1. Recently, the structure of a mammalian cytochrome P450 2C5 from rabbit was reported (13). This enzyme has a high degree of similarity (> 82%) and identity (> 77%) with the CYP 2C family in humans, although the crystal structure has a relatively low resolution (3 Å) and no substrate has been co-crystalized [9, 14]. Therefore, it can be used as a template in homology modeling for the CYP 2C enzymes. These homology protein models were previously used to the model and predict of competitive inhibition constants and to obtain predictive models and structure-activity information in a selectivity analysis [14].

Here we describe the used of the CYP 2C9 homology model in the site of metabolism prediction based on the protein structure. The working hypothesis consists in the idea that the site of reaction in the binding site in the protein has to be close to the place were the oxidation occurs in the substrate at the time the reaction happens. This hypothesis would not hold in the case that after the first reaction a chemical reorganization takes place. In the literature there are two proposed mechanism to explain the oxidative reaction by cytochrome P 450: In the first mechanism a proton abstraction happen with a radical formation, and in the second mechanism an addition to bond take place. In both cases the site of oxidation in the substrate is close to the oxidative machinery in the enzyme. Therefore, there is a point in the protein structure that corresponds to a point in the substrate. The working hypothesis

Table 2. Summary of oxidation relations searched for H234/09 substructure.

| Transformation | Substrate Occur. | Transform. Occur. | Occur. Ratio | OR _{normalized} |
|----------------|------------------|-------------------|--------------|--------------------------|
| | 27 | 2 | 0.0741 | 0.224 |
| | 27 | 0 | 0.0000 | 0.000 |
| | 74 | 0 | 0.0000 | 0.000 |
| | 74 | 4 | 0.0541 | 0.164 |
| | 2314 | 233 | 0.1007 | 0.305 |
| | 2314 | 104 | 0.0449 | 0.136 |
| | 6476 | 1624 | 0.2508 | 0.760 |
| | 1478 | 83 | 0.0562 | 0.170 |
| | 3425 | 812 | 0.2371 | 0.718 |
| | 2329 | 49 | 0.0210 | 0.064 |
| | 409 | 5 | 0.0122 | 0.037 |

Table 2. Continued.

| | | | | |
|---|------|-----|--------|-------|
|  | 2329 | 49 | 0.0210 | 0.064 |
|  | 409 | 5 | 0.0122 | 0.037 |
|  | 3425 | 438 | 0.1279 | 0.387 |
|  | 122 | 3 | 0.0246 | 0.074 |
|  | 122 | 5 | 0.0410 | 0.124 |
|  | 409 | 135 | 0.3301 | 1.000 |

considers that this point is the oxygen atom in the protein and the hydrogen atom position for the substrate. We are fully aware that this is not exactly the case, but this a simple rule that allow a fast comparison between the structure of the substrate and the protein.

The site of metabolism prediction process comprised three steps:

1. The generation of the pharmacophoric representation in the protein
2. Assessing the pharmacophore of the compound to be predicted
3. Comparison between the pharmacophore of the protein and the substrate.

Step 1. Protein pharmacophore generation

The molecular interaction fields (MIF) in the binding site of the CYP 2C9 were calculated by the GRID methodology [15]. Three MIFs were generated in this analysis: the DRY (hydrophobic), the N1 (amide nitrogen probe hydrogen-bond donor) and the O (carbonyl oxygen probe hydrogen-bond acceptor) probes using a grid step size of 1 Å. The grid box was defined to

cover the binding site. The MIF was generated using the flexible mode in GRID (directive MOVE = 1) [16]. In this mode, some of the amino acid side chains can automatically move depending on the attractive or repulsive interactions with the probe. The side chains flexibility in GRID could mimic the movement of side chains to accommodate different substrates depending on the size, shape and interaction pattern. However, this flexible grid map cannot consider possible movements in the protein backbone.

The MIFs (Figure 6a) were pre-treated by applying a cut out procedure; the regions close to the binding site but not accessible to the substrates were removed from the analysis. However, not all the points provide relevant information. In order to select the most representative ones, the pre-treated MIFs were exported into the ALMOND program [17]. This program selects MIF interaction points (seeds) according to an experimental design technique (D-optimal) which uses the energy of the interaction and the distance between the grid points as selection criteria (Figure 6b). In this

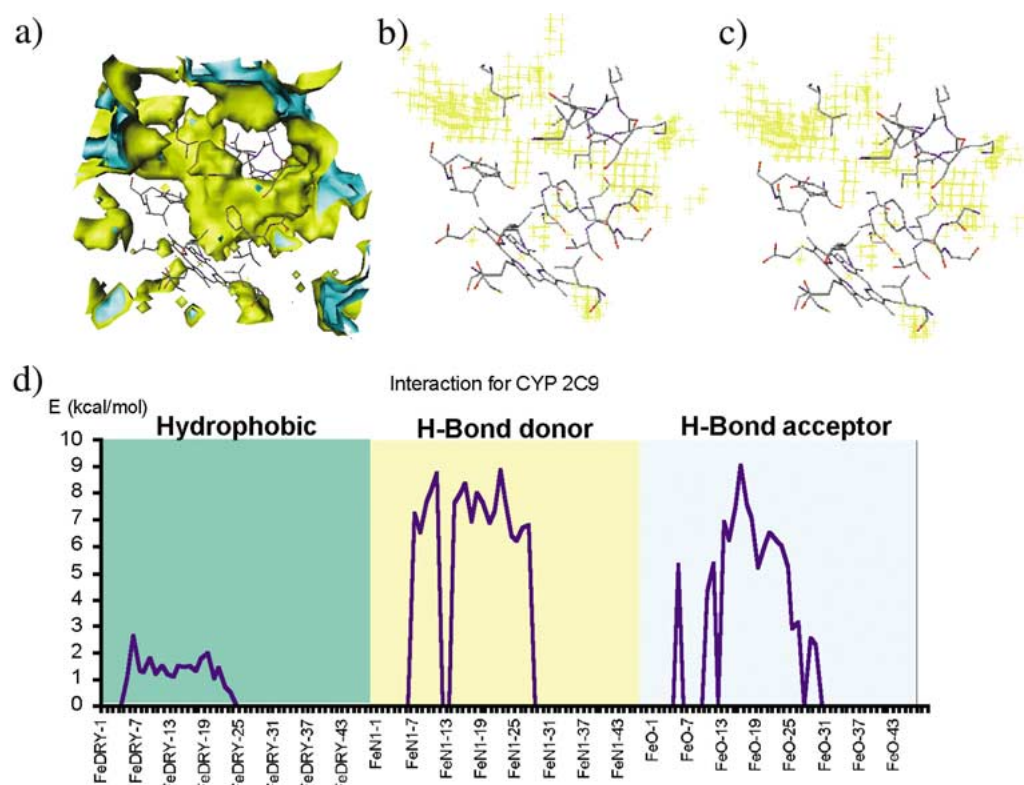


Figure 6. Protein treatment. a) Molecular Interaction Field for CYP2C9 model using N1 probe. b) MIF filtered by cut out option in GOLPE and considering only the negative interactions for the N1 probe. c) Selected points by ALMOND using 1000 seeds and 50% weight for the distance-energy criteria. d) Descriptors calculated for the different interaction fields. e) Representation of the new descriptors in the different bin distances.

study 50% influence of each of these factors and 1000 seeds were considered.

The selected points and the pre-treated fields were then used to calculate a new set of descriptors (Figure 6c). These descriptors transform the interaction energies at a certain spatial position to a binned distance space, similar to the maximum value of the auto-correlogram in the Grid Independent descriptor procedure [17]. In order to calculate this distance, one starting point is always fixed to the oxygen atom attached to the iron atom in the heme complex (the reactive center of the enzyme) and the second point is the selected by the ALMOND program. For a certain bin distance, the maximum energy of interaction corresponding to the selected point is the one used in the description. This set of descriptors is a pharmacophoric representation of the binding site with respect to the fix point, which is the reactive center.

Step 2. Ligand pharmacophore representation

The known substrate atoms were classified into three categories depending on their hydrophobic, hydrogen bond donor or acceptor pattern. The distances between the different atom positions classified using the three interaction criteria were then transformed into binned distances (Figure 7). In this case, the distances between the different hydrogen atoms and the classified atoms were calculated and a value of one or zero was assigned to each bin distance indicating the presence or the absence of such distance in the substrate. One set of descriptors was calculated for each category of atom types: hydrophobic, hydrogen bond acceptor and hydrogen bond donor atoms, giving a fingerprint for each hydrogen atom in the molecule.

Ligand-protein comparison

Once the protein interaction pattern was translated from the co-ordinates to distances from the reactive site in the receptor, and the structure of the ligand was

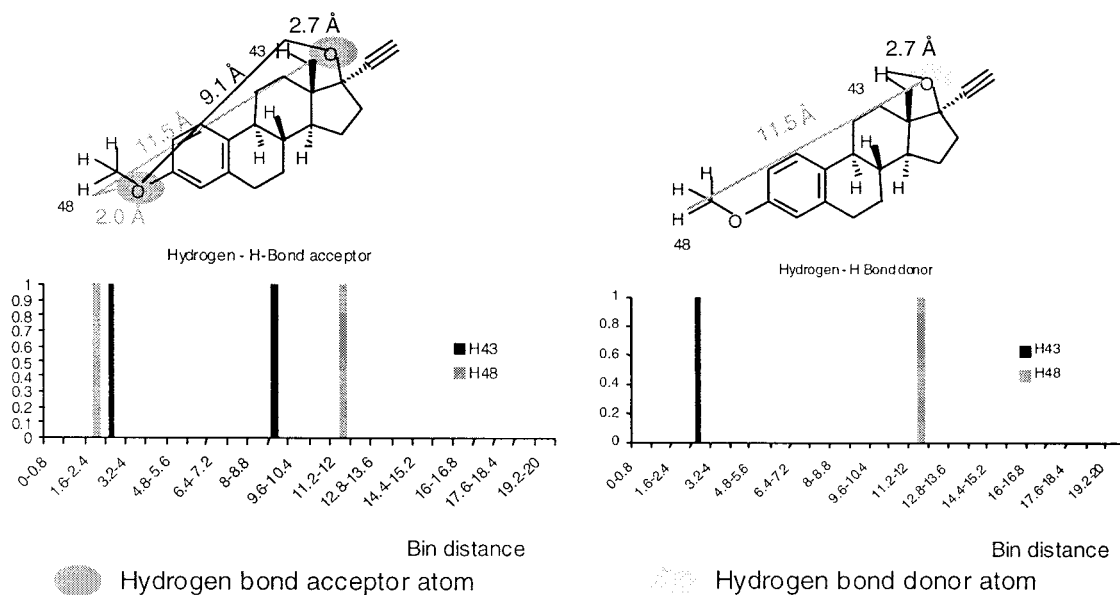


Figure 7. Ligand treatment using a steroid example. Two hydrogen atoms are highlighted: 48 and 43. The hydrogen abstraction is reported at the H48. a) Hydrogen bond acceptor: two atoms are present in this molecule that could be hydrogen bond acceptors. The distances between the hydrogen atoms (43,48) and the hydrogen bond acceptors and the correspondence in the fingerprint are shown. b) Hydrogen bond donor: one atom is present in this molecule that could be a hydrogen bond donor. The distances between the hydrogen atoms (43,48) and the hydrogen bond donor and the correspondence in the fingerprint are shown. The compound conformer was calculated using CONCORD software.

$$\text{Similarity}(\text{H-CYP2C9}) = \frac{\sum_{i=1}^{\text{bin distance}} E_i * I_i}{\sqrt{\sum_{i=1}^{\text{bin distance}} E_i^2} * \sqrt{\sum_{i=1}^{\text{bin distance}} I_i^2}}$$

Total similarity =
Similarity (H-CYP2C9)DRY+
Similarity (H-CYP2C9)N1+
Similarity (H-CYP2C9)O

Equation 1

E_i : Energy of interaction in the protein
 I_i : Presence of the distance in the ligand

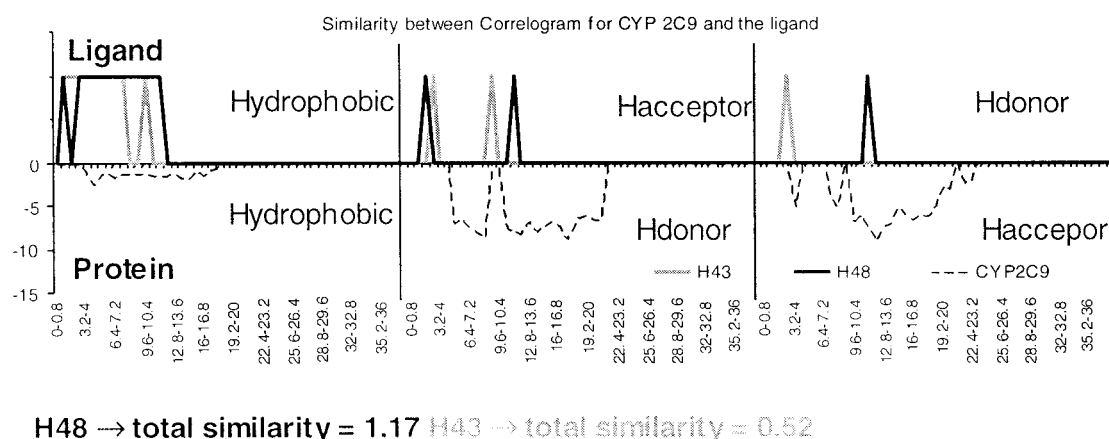


Figure 8. Comparison between the molecule fingerprint and the descriptors generated using the CYP 2C9 homology model and the grid molecular interaction field.

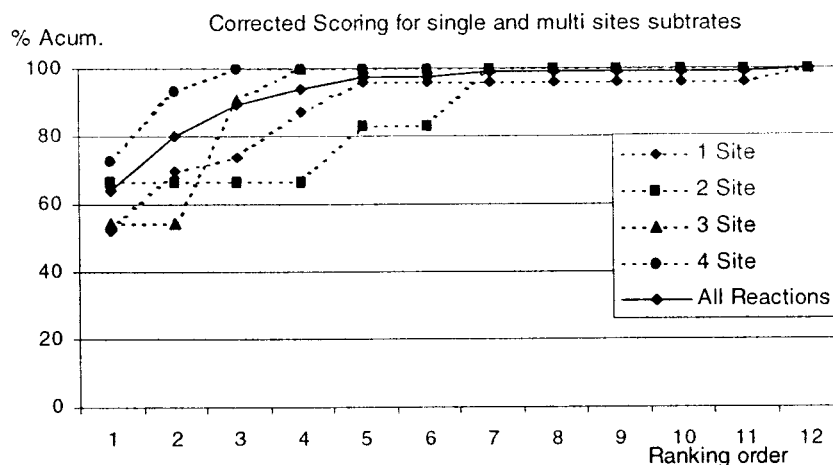


Figure 9. Hydrogen atom ranking for: Average for the 25 random generated conformers.

described as a fingerprint/hydrogen atom, both sets of descriptors were compared (Figure 8). The Carbó similarity index was used for this purpose (Eq. 1-Figure 8) [18]. Three similarity indexes were obtained for each hydrogen atom in a substrate:

1. The hydrophobic interaction in the protein and the ligand - hydrophobic complement.
2. The hydrogen bond donor descriptors for the protein and the hydrogen bond acceptor ones for the substrate.
3. The hydrogen bond acceptor for the protein and the donor from the ligand. Finally, the different hydrogen atoms of the CYP 2C9 substrates were ranked according to the computed total similarity index. Other oxidative metabolic reactions that do not involve a hydrogen atom from the substrate were not considered in the analysis i.e.: oxidative reaction on the lone pairs in a sulphone group or nitrogen atom.

Methodology validation

The 87 metabolic reactions reported in the metabolite database (8), exclusively catalyzed by the CYP 2C9 were used to validate the methodology. This set of reactions includes 43 different substrates: 25 showing one single site of metabolism (Table 1), three compounds present two sites of metabolism (Table 2), four of them with three sites of metabolism (Table 3) and eleven with 4 sites of metabolism (Table 4). Moreover, substrates show a large structural diversity including rigid compounds, e.g. steroids and very flexible ones with more than 10 routable bonds, a wide range of molecular weight and lipophylicity. Only one family of 13 coumarin-like substrates is in this data set.

The predictions are summarized in Figure 9. In more than 50% of all the reactions, the first option selected by the methodology matches the experimental determined one. Moreover, in more than 25% and 15% of the cases, the second and third hydrogen atom respectively, are the ones that fit the experimental one. Therefore, considering the corrected (see section 2nd point in Material and Methods) ranking list for the multiple sites of metabolism, in more than 90% of the cases the methodology predicts the site of metabolism for CYP2C9 within the first three hydrogen atoms selected, independent of the conformer used.

Conclusion

The two methods reviewed here represent the beginnings of what is hoped to be the next generation of computational tools aimed to aid in addressing xenobiotic metabolism. Clearly both methods have limitations and as with most other modelling methods, should be used with as many other methods aimed at giving the same type of information in order to build a 'consensus' prediction, i.e., that most or all methods point to the same result. Consensus predictions are made even more robust if the models included in the predictions use diverse approaches to the same problem.

References

- Clarke, N.J., Rindgen, D., Korfmacher, W.A. and Cox, K.A., *Anal. Chem.*, 73 (2001) 430A–439A.
- Talafous, J., Sayre, L., Mieyal, J. and Klopman, G., *J. Chem. Inf. Comput. Sci.*, 34 (2001) 1326–1333.
- Greene, N. in P.W. Erhart (ed.) *Databases and high-throughput testing during drug design and development*. IUPAC, Research Triangle Park, NC, USA., (1999), pp 289–296.
- Darvas, F., *J. Mol. Biol.*, 6 (1988) 80–85.
- De Groot, M.J., Ackland, M., Horne, V., Alexander, A. and Barry, J., *J. Med. Chem.*, 42 (1999) 4062–4070.
- Lewis, D.F., Dickens, M., Eddershaw, P.J., Tarbit, M.H. and Goldfarb, P.S., *Drug Metab. Drug Interac.*, 15 (1999) 1–49.
- Jones, J.P., Korzekwa, K.R., *Meth. Enzymol.*, 272 (1996) 326–335.
- ISIS BASE, Metabolite database, MDL information system Inc., San Leandro, CA. www.mdli.com.mdl.com
- Afzelius, L., Zamora, I., Ridderström, M., Kalén, A., Andersson, T.B. and Masimirembwa, C., *Mol. Pharmacol.*, 59 (2001) 909–919.
- De Groot, M.J., Ackland, M., Horne, V., Alexander, A. and Barry, J., *J. Med. Chem.*, 42 (1999) 4062–4070.
- Jones, B.C., Hawksworth, G., Horne, V.A., Newlands, A., Morsman, J., Tute, M.S. and Smith, D.A., *Drug Metab. Disp.*, 24 (1996) 260–266.
- Mancy, A., Broto, P., Dijols, S., Dansette, P.M. and Mansuy, D., *Biochemistry*, 34 (1995) 10365–10375.
- Williams, P.A., Cosme, J., Sridhar, V., Johnson, E.F. and McRee, D.E., *Mol. Cell.*, 5 (2000) 121–131.
- Ridderström, M., Zamora, I., Fjåström, O. and Andersson, T.B., *J. Med. Chem.*, 44 (2001) 4072–4081.
- Goodford P.J., *J. Med. Chem.*, 28 (1985) 849–857.
- GRID V.17, Molecular Discovery Ld, 2001 (<http://www.moleculardiscovery.com>).
- Pastor, M., Cruciani, G., McLay, I., Pickett, S. and Clementi, S., *J. Med. Chem.*, 17 (2000) 3233–3243.
- Amat, L. and Carbó-Dorca, R., *J. Comp. Chem.*, 20(9) (1999) 911–920.