Title

Molecular docking for substrate identification: the short-chain dehydrogenases/reductases

Running Head

Substrate identification using docking

Authors & Affiliations

Angelo D. Favia[*1], Irene Nobeli[2], Fabian Glaser[1] and Janet M. Thornton[1]

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

[*] Corresponding author (phone number:+44 (0) 1223 492545)

[1] EMBL–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

[2] Randall Division of Cell and Molecular Biophysics, New Hunt's House, King's College London, Guy's Campus, London, SE1 1UL, UK

Abstract

Protein ligand docking has recently been investigated as a tool for protein function identification, with some success in identifying both known and unknown substrates of proteins. However it remains challenging to identify a protein's substrate when cross-docking a large number of enzymes and their cognate ligands. To explore a more limited yet practically important and timely problem in more detail, we have used docking for identifying the substrates of a single protein family with a remarkable substrate diversity, the short-chain dehydrogenases/reductases (SDRs).

We examine different protocols for identifying candidate substrates for 27 SDR proteins of known catalytic function. We present the results of docking more than 900 metabolites from the human metabolome to each of these proteins together with their known cognate substrates and products, and we investigate the ability of docking to a) reproduce a viable binding mode for the substrate and b) to rank the substrate highly among the dataset of other metabolites. In addition, we examine whether our docking results provide information about the nature of the substrate, based on the best-scoring metabolites in the dataset. We compare two different docking methods and two alternative scoring functions for one of the docking methods, and attempt to rationalise both the successes and failures.

Finally, we introduce a new protocol, whereby we dock only a set of representative structures (medoids) to each of the proteins, in the hope of characterising each binding site in terms of its ligand preferences, with a reduced computational cost. We compare the results from this protocol with our original docking experiments and we find that although the rank of the representatives correlates well with the mean rank of the clusters they belong to, a simple structure-based clustering is too naïve for the purpose of substrate identification. Many clusters comprise ligands with widely varying affinities for the same protein and, hence important candidates can be missed, if a single representative is used.

Introduction

Every year structural genomics initiatives generate many new protein structures, with one aim being to provide representatives for as many families as possible[1-5]. The function of some of these proteins is known before their structure is solved, but for many proteins the 3D structure represents a fundamental source of information to improve our understanding of the molecular mechanism of their biochemical role. As a first level of analysis, the sequence similarity to characterise proteins has been shown to be a very powerful but sometimes insufficient method to infer a protein's function. Indeed, different aminoacid sequences are capable of generating proteins sharing similar folds but this does not always imply a related function[6]. For this reason, taking into account only geometrical features seems to be a simplistic way of looking at the complex biochemistry beyond the structural data, although it is obviously useful for assigning a given protein to an already defined family. Structure-based prediction of enzymatic capabilities has become an important issue and many research groups are working, worldwide, with the ultimate goal of providing a suitable, *in silico* protocol that, given an experimental structure, can lead to a reliable prediction of its function.

For years, docking has been a powerful tool for medicinal chemists, allowing the rapid and inexpensive identification of a pool of potential protein inhibitors[7-9]. Lately, docking simulations have been used, in the field of structural biology, with a different purpose: the identification of plausible substrates of proteins of unknown function[10-13]. The main idea is to dock a database of potential metabolites into a protein's binding site and then rank them on the basis of their calculated binding affinities. Ideally, the cognate ligand is expected to be found among the top ranked hits. Due to the approximation of the simulation, the calculated binding affinity by itself is unlikely to identify a protein's cognate partner among a crowd of candidates, but this approach can still be useful to experimentalists, narrowing down the number of molecules that need to be tested. The goal is that *in vitro* tests would only need to be conducted for a relatively small pool of compounds (i.e. the top *in silico* hits), and that these compounds would somehow capture the essential physicochemical features of the true unknown ligand. A number of docking algorithms are nowadays available. Several published papers have compared state-of-the-art software on the basis of their ability to reproduce experimentally determined complexes[14-19]. Unfortunately, the quality of the results depends on the biochemical nature of the molecules involved, so that different algorithms can, in turn, work better for different cases.

In this study, we tested the ability of different docking protocols to find the real substrate of several members of a specific protein superfamily, the short-chain dehydrogenases/reductases (SDRs). SDRs form a large family of NAD(P)H-dependent enzymes, with about 70 genes found in the human genome. Despite low levels of sequence identity (usually 16-30%), members of this family show a highly conserved α/β folding pattern (Rossmann fold), with largely superimposable peptide backbones[20-23]. A catalytic tetrad comprising serine, tyrosine, lysine and asparagine residues constitutes the active site. The final step of the catalysis is mediated by the nicotinamide ring of NAD(P)H, which acts as an electron carrier[21]. SDRs represent a challenging test-case for computational simulations, since the broad substrate specificity calls for a docking protocol that is capable of distinguishing and assessing several different types of chemical interactions. Twenty-seven SDR proteins were studied using two different programs, Autodock[24] and Glide[25]. In addition, since molecular mechanics-based docking programs generally tend to underestimate the electrostatic and solvation contributions to the binding, and following previously published studies, a physics-based rescoring procedure of the docking poses was also applied[11,12].

A generally accepted protocol in these types of studies is to test a large number of molecules in order to explore as much as possible the *n*-dimensional ligand physicochemical space, but here we also consider whether a comparable coverage could be achieved by docking a smaller number of molecules. In order to answer this question, results from docking a database of 922 human metabolites were compared to those obtained from docking a set of 115 dissimilar molecules, selected as their representatives. This selection of human metabolites lowers the number of docking simulations to be performed in order to obtain the maximum amount of information about the nature of a protein's cognate ligand with minimum computation. Our approach shares some similarity with that suggested by Shoichet and co-workers[26]. Their approach involved defining families of ligands sharing the same scaffold and docking all members of all families to a given protein, but only ranking the best-scoring member of each ligand family. This method aimed at improving diversity in the top ranks, by allowing only one member of each family in the ranked list. Our proposed method also clusters ligands into families prior to docking, but our aim is different. First of all, we test the hypothesis that similar ligands will bind with similar energies, even if we do not impose a strong positional constraint on the orientation of the ligand scaffold, as was the case in the previous study. In addition, we are not trying to enrich the diversity of the top-ranking compounds, as we are interested here in function identification, not the discovery of novel inhibitors. Our aim is to test the feasibility and usefulness of this approach, which we believe may work better in the case of the natural substrates. Our argument is the following: Small changes in a random ligand structure can have a significant effect on the energy of binding. However, proteins evolved from a common less specialised ancestor, may be capable of binding, and even acting on, several members of the same ligand family. This would require a certain lack of sensitivity to small structural differences, allowing our approach to work better for these cognate ligands than it would for unrelated ligands. We recognise that it is questionable whether reducing the number of docking simulations is useful at a time when software and hardware allow us to screen millions of compounds in a reasonable amount of time, but we think there are two reasons why screening initially only representatives may be useful. One is that, should this approach be successful, it could be copied by experimental laboratories, as an affordable way of systematically screening proteins against the same set of ligands. The second reason is a smaller dataset of ligands would allow more sophisticated and time demanding simulations to be carried out, thus making more reliable the calculation of the binding affinities. We note that screening families of ligands is not likely to be a substitute to screening large datasets, but a complementary approach that can potentially help us to understand better what features of related molecules are responsible for their different binding properties.

Results

1.The dataset

1.1. The proteins

Twenty-seven SDR structures from the Protein Data Bank (PDB)[27] were selected for docking following the criteria described in the Methods. The PDB codes, protein names, selected chain, EC classification, organism, substrate types and the presence of a substrate-like ligand bound in the X-ray structures are listed in Table 1. Of the 27 proteins, 13 are from bacterial organisms and 14 from eukaryotes (including 3 human and 2 mouse proteins), with no examples from archaea. Approximately half of the enzymes in this study were available as ternary complexes (protein complexed with a cofactor and a substrate-like ligand). This dataset has representatives from three EC classes, with the majority of the proteins being oxidoreductases (EC class 1), two being lyases (class 4) and three being isomerases (class 5). By far the most common type is oxidoreductases acting on the CH-OH group of donors (EC 1.1).

The SDR proteins are known to be evolutionarily related (as confirmed by their structural similarity) but their pairwise percentage sequence identities vary widely. For example, a comparison of the chains 1geg:A[28] (acetoin (diacetyl) reductase from *K. Pneumoniae*) and 1orr:A[29] (CDP-tyvelose-2-epimerase from *Salmonella*) shows that a global alignment is not possible for this pair, and even a local alignment results in only 22 identical residues over a relatively short alignment of 113, from a total of 255 and 338 residues. Both these proteins are single-domain, so the sequence differences are genuine rather than due to the presence of an extra domain. At the other end of the spectrum we have a pair of very closely related sequences, 1b14:A[30] and 1mg5:A[31] (both alcohol dehydrogenases from fruit fly), whose global alignment yields an 82% sequence identity. A multiple sequence alignment of all 27 proteins using MUSCLE[32] yields the phylogram in Figure 1. The tree in this figure is annotated by the EC numbers of these proteins (branch colour and text colour) and by their substrate similarities (coloured box on the right), and it shows that in our dataset the lyases (in orange) and isomerases (in pink) are phylogenetically close, and these two classes are further away from the oxidoreductases (in green and blue). 1kc3 is an exception, as an oxidoreductase whose substrate belongs to the same type of substrate recognised by the lyases and isomerases in our dataset.

Most SDRs are single-domain proteins, exhibiting the ubiquitous Rossmann fold, where a beta-sheet is flanked by a number of helices. The N-terminal part of the domain is used to bind the NAD(P)H cofactor which sits at the top of the beta-sheet, whereas the C-terminal part is involved in the binding of the substrate[20]. As expected from the wide substrate specificity of these proteins, the N-terminal parts of the structure superimpose very well, whereas most of the variation is seen in the C-terminal parts of the domain (see Figure 2a-b). In our dataset there are 19 proteins that are single-domain and 6 proteins that are made up of two domains (1bsv[33], 1ek6[34], 1kc3[35], 1ker[36], 1keu[36], and 1udc[37] – all bind nucleotide sugars and the second domain in all is a UDP-galactose 4-epimerase, domain 1 (CATH number 3.90.25.10)). The remaining proteins (1orr[29] and 1w4z[38]) are not yet classified in CATH[39].

1.2 The substrates and products

Schematic representations of the substrates for each of the 27 proteins used in this study are shown in Figure 3 (diagrams of the products are available as Supplementary Information). A quick look through the structures reveals that they can be broadly grouped into the following types: a) a group of steroids (1ahh, 1equ, 1e6w_c) and a molecule sharing some resemblance with steroids (1ja9), b) a group of relatively small and polar molecules, mainly primary aliphatic alcohols (1b14, 1e6w_a, 1geg, 1h5q, 1mg5, 1pr9), c) a group of coenzyme A derivatives (1bvr, 1c14, 1d7o, 1e6w_b, 1edo, 1q7b, 1w4z), d) a group of nucleotide sugars (1bsv, 1ek6, 1kc3, 1ker, 1keu, 1orr, 1udc) and e) others that cannot easily fit into the above classification (1ae1, 1cyd_a, 1cyd_b, 1ipf, 1iy8, 1sep). This classification is based on the presence of well-known chemically important groups, or obvious physicochemical properties, such as size and polarity, and, as such, it allows a general understanding of the types of molecules involved in this study at a qualitative level. An alternative quantitative way of clustering the substrates uses the similarity matrix of their pairwise fingerprint-based scores (see Methods for details). A dendrogram representing this matrix is shown in Figure 4 and is annotated by the group the substrate belongs to (coloured box on the left) and the EC number of the protein (coloured box on the right). Although some mixing of the groups defined above is observed in this dendrogram (both due to the arbitrary way the groups were defined, and due to the limitations of the method for assessing similarity), overall a good clustering of these groups is shown. Neither the manual nor the automatic approaches are perfect and, indeed, a unique way of assessing small molecule similarity that will always be satisfactory does not exist, thus some overlap between classes is inevitable.

1.3 The metabolite representatives

Although nowadays it is common practice to dock large databases of ligands to receptors, we believe it is worth testing the hypothesis that only a number of selected representatives is enough to obtain reasonably accurate information regarding the preferred type of substrates for a given enzyme, or more generally the preferred binders for a protein receptor. The reasoning behind this plan was that experimental measurements of binding affinity are still limited by the availability and cost of buying the compounds that have to be screened. Without any previous knowledge, one would have to test every single available metabolite for binding affinity. This is clearly impractical, if not impossible, and hence, here we are testing computationally the alternative of using a selection of ligands whose structures represent a variety of chemical classes found in human metabolic pathways.

One hundred and fifteen dissimilar molecules selected as representatives of human metabolites (see Methods for details) were used in this study to probe the nature of the binding site of the 27 SDR proteins. A list of all 115 KEGG identifiers corresponding to these molecules is provided as Supplementary Information. Figure 5 shows that this set of representatives provides good coverage of most areas of the human metabolome, and justifies their selection for use in this study.

In addition to the metabolite representatives, we also docked all 922 human metabolites in our dataset for comparison. All ligands were processed using the program LigPrep [40].

2. Results from the docking experiments

This section examines the results of our docking experiments as a method for identifying likely substrate candidates. We performed docking runs using the following datasets: a) We docked the substrates and products of all proteins in our dataset to each protein, b) we docked 115 metabolite representatives to all the proteins in our dataset and c) we docked all small molecules from the set of human metabolites (922) to all proteins in our dataset.

The results are presented as follows:

a) We start by assessing the quality of the docking research method, using as a benchmark the information we have about the binding of the cognate substrate to its receptor protein. For this we use the set of docking results corresponding to available structures of the complexes of the substrates/products (or their analogues) with their corresponding enzymes, and ask whether our experiments have resulted in plausible positions for the substrates in the binding site;

b) We ask whether the scoring function was successful in identifying the natural substrate among a pool of other SDR substrates and a larger pool of diverse ligands. We also present results from an alternative scoring function and compare the two sets;

c) We examine the specificity of recognition from both the substrate and the protein side;

d) We examine the top-scoring ligands and assess whether docking can offer clues about the properties of the substrate using common physicochemical descriptors;

e) We present the results of docking a set of representative structures of human metabolites to our SDRs, and examine the usefulness of these docking experiments in limiting the number of compounds that need to be screened *in silico* or *in vitro*.

Substrates and products which didn't belong to the starting human metabolites dataset were eventually added to the distribution. Unless otherwise specified, the results in the following sections refer to Glide SP scores and docking poses constrained with a 4 Å distance filter (to reject docked poses too far from the NAD(P)H C_4 atom) (see Figure 2c).

2.1 Does docking work with substrates/products?

2.1.1 Do natural substrates dock in the active site?

All but two of the tested cognate substrates dock in the active site of their partner proteins, and pass our distance filter (see Methods), i.e. are close enough to the NAD(P)H cofactor to represent plausible docked positions. The exceptions are substrate "a" of carbonyl reductase (1cyd[41]), and the substrate of alcohol dehydrogenase (1b14[30]). Since it is a well known fact that proteins can undergo conformational change upon ligand binding[42-45], one of the criteria leading to the protein dataset selection was the presence, where possible, of a substrate-like molecule bound (see Methods). In fact, X-ray structures represent only an average snapshot of a protein's multiple conformations and if the natural substrate or, at least, a molecule that mimics it, has not been co-crystallized with the macromolecule, the chances that the side chains at the binding site are in a 'ready-to-host' conformation are quite low. Unfortunately complexes of the protein with both the cofactor and the substrate or substrate analogue are not commonly available. The alcohol dehydrogenase in our set (1b14) was crystallised as a binary complex and the carbonyl reductase (1cyd) protein was crystallized as a complex with a molecule that does not resemble the enzyme's natural substrate. In the case of 1b14 allowing just the side chains of the binding site to move, using the Induced Fit Docking (IFD) protocol (see Methods)[46], results in a good pose for the substrate (see Figure 6a). The spatial rearrangement of few side chains (namely I145, I183 and L206) creates the right environment for the natural substrate to be docked properly. The same methodology does not work for 1cyd, although this may simply be because for this complex a much larger rearrangement of the binding site is needed than is normally allowed during the IFD method. However, since we found in the literature an alternative, much smaller, substrate (referred to here as substrate "b") that docks close to the NADP$^+$ molecule, this has been used in the rest of this work. Neither of the two crystal structures contains a ligand similar to the substrates we docked, although it is worth mentioning that there are other structures in our dataset (of different proteins) where a ligand similar to the substrate is lacking, but where the substrate can be well docked without any conformational changes needed in the binding site. Thus, we were able to obtain docked complexes involving the substrate for all 27 proteins.

In the case of natural substrates, it also makes sense to ask whether the orientation of the substrate in the binding site is the correct one. Although for small substrates this is less important (taking into account the limitations of docking, small errors in orientation should be acceptable), for large substrates it is significant, as the part of the molecule that is involved in catalysis may in fact be completely outside the active site. To summarise the results for the top ranking poses of both substrates and products, we monitor the distances between the C_4 carbon atom of the nicotinamide ring (see Figure 2c) and the atom of the ligand where we expect the chemical reaction to take place (where two or three atoms are involved, the shortest distance is taken into account). We find that in well over half of the cases (22 out of 27) this distance is less than 5 Å. This distance cannot be an accurate quantitative measure of the success of docking, but it indicates in a qualitative manner whether the top ranking solution is a reasonable approximation to a productive mode or not. In the majority of cases (19 of 27) we find an acceptable pose even for very large substrates, and in some cases (4 of 27) even though the substrate docks in a reverse orientation, the product is docked well. Some of the cases, where neither the substrate nor the product are in a correct orientation are simply very flexible molecules, and their conformational space was probably not adequately explored during docking. There is no correlation between the docking scores and the distance of the substrate/product from the NAD(P)H structure, i.e. the estimated binding affinity is not indicative in this case of the quality of the docking pose. A detailed table showing the results of this analysis is available as Supplementary Information.

2.1.2 Are the natural substrates/products recognised among other SDR substrates and among other ligands?

As the pie chart in Figure 7 shows, in 18 out of 27 cases, the substrate or product is ranked in the top 5% of all 922 ligands docked (human metabolites + substrate and product of the studied protein). In 3 cases, the natural substrate or product is ranked between the 5% and 7% (5-10% interval is shown in Figure 7) of the distribution, while in 2 cases it can be found within the top 10% and 25% of the distribution. The few remaining cases, where neither the substrate nor the product score well, are examined below.

a) 1b14[30] (Alcohol dehydrogenase from *Drosophila lebanonensis*). The flexible loop of residues 186 to 191 is disordered, and the right conformation is probably important for substrate binding. To test this hypothesis an Induced Fit Docking run was performed on a limited dataset of molecules. Final results show that the natural substrate docks only if some conformational rearrangements occur (see Figure 6a). As previously mentioned, the 1b14 X-ray structure lacks a molecule bound capable of mimicking the natural substrate. In such cases, residues at the binding site tend to extend as much as possible their side chains while looking for favourable interactions, thus lowering the space available for a ligand to be docked.

b) 1h5q[47] (Mannitol dehydrogenase from *Agaricus bisporus*). According to the crystallographers who solved this protein structure, all attempts to crystallise the enzyme with its substrate failed, most likely due to the high *Km* value. Instead, they used the program Autodock to model the substrate in the binding site. Our own best pose from Glide shows a good orientation (Figure 6b), forming the hydrogen bonds reported by the crystallographers, but the rank is low (223 out of 922+61). As the score seeks to estimate the affinity and is our main criterion here for identifying candidate substrates, it is expected that our approach will necessarily fail in cases like this, where the substrate has a low affinity for the enzyme.

c) 1bvr[48] (Enoyl-acp reductase (InhA) from *M. tuberculosis*) and 1pr9[49] (Human L-xylulose reductase). Inspection of the molecular features of the top 10 binders of each protein shows that these do not resemble at all the features of the natural substrates (see Figure 10). In such cases, obviously, the docking protocol used here, is not capable of properly accounting for the physicochemical interactions involved in the docking process.

2.1.3 Comparison of alternative docking/scoring methodologies

A docking calculation is composed of 2 main parts: a search method and a score assignment. These two steps are closely related and equally important. In order to verify how our results vary with the methodology used we tested a few alternative docking and scoring approaches, including:

a) docking and scoring using the Autodock software[24];

b) docking and scoring using the Glide software[25];

c) rescoring our Glide SP poses with a physics-based scoring function (using the Prime software[50]) that calculates an approximation to the free energy of binding by combining the energy of the complex (E_{PL}) with the energy of the ligand on its own (E_L) and the energy of the protein (E_P) on its own (minimising the ligand in the free state and in the complex, but keeping the protein rigid);

$$E_{bind} = E_{PL} - E_L - E_P$$

d) rescoring our Glide SP poses with the same physics-based scoring function but allowing the side chains in the active site to move during minimisation of the complex.

For this comparison we used only the smaller dataset of 176 representatives (115 human metabolites representative + each protein substrate and product). Figure 8 compares the results of the four methods, and although the dataset is rather small, the Glide poses scored with the Glide SP (standard precision) scoring scheme give the best results in this case. However, Figure 9a shows that in some cases the rank of the substrate improves dramatically as a consequence of using a physics-based scoring function, although the reverse is true for other examples. Indeed, this function has shown superior performance in the past [11,12], but we think it is more likely to work where the binding sites are charged or highly polar, and where solvation energies need to be better taken into account in order to properly estimate the binding affinity. The natural substrates for 1h5q, 1iy8 1pr9 and 1mg5 (shown in yellow in Figure 9a) are indeed small hydrophilic molecules whose ranks improve drastically by means of the rescoring procedure. Interestingly the substrate of 1ahh, which is the only steroidal molecule to have an acidic function, is the only member of its group whose rank improves after the rescoring procedure. The variety of chemical interactions involved in the ligand-protein docking process for this SDR dataset represents an extremely challenging test for scoring functions. With the generally hydrophobic binding sites of the SDRs, it is not surprising that an empirical scoring function like the Glide SP score performs usually better.

2.1.4 Specificity in substrate/enzyme recognition

Since our proteins are evolutionarily related, it is interesting to examine the level of specificity in the recognition of the enzyme by the substrate and vice versa. Although it is possible to use the ranks from each distribution of scores to assess specificity, these distributions contain relatively few observations, especially when considering the recognition of the protein by the substrate (where we only have 27 proteins). This makes estimates of the probabilities of these scores less reliable. To overcome this problem, we normalise each distribution of scores and pool all of them into one large distribution (see Methods for details), thus allowing us to estimate better the *p* values associated with the ranking of the cognate partners. We create two large distributions: one for all scores of 176 ligands docked against each protein, and the other for all 27 scores associated with each substrate being docked against each of the set of 27 proteins. Clearly since the protein pool is very small and the proteins are related, the estimates from that distribution will be less reliable than estimates from the distribution of all ligand scores docked against each protein.

In Figure 9b we plot the *p* values for each protein and substrate of a cognate pair, as calculated from these two large distributions. The substrate *p* value tells us how well the substrate recognises its cognate protein from a pool of 27 SDR proteins (the lower the *p* value the better the specificity). The protein *p* value tells us how well the protein recognises its cognate substrate from a pool of a maximum of 176 ligands (61 substrates and products and 115 metabolite representatives). In this plot 14 of the cognate pairs have both substrate and protein *p* values less than 0.25, 17 pairs have a substrate *p* value less than 0.25, and 19 pairs have a protein *p* value less than 0.25. For 5 pairs neither the substrate, nor the protein *p* value is less than 0.25, and these are cases where docking clearly cannot help us identify the true cognate pair. However, in four out of five cases (1h5q, 1q7b, 1geg, 1iy8, exception is 1b14), and particularly, in the case of 1q7b, the cognate product is scored better than the substrate (only the results for substrates are shown in this plot). A closer look at the figure also reveals that our small substrates (1geg, 1iy8, 1mg5, 1pr9) all tend to have poor protein *p* values, i.e. it is difficult for the protein to identify a small cognate ligand amongst a set of large molecules that can also fit the binding site (obviously this is not true if large ligands are excluded due to volume restrictions within the active site). This fact reflects a common problem with scoring functions, namely the penalising of very small molecules that can only achieve a limited number of interactions in a complex. The polar groups in these ligands are also likely to be responsible for their worse ranking, since solvation is not properly accounted for with empirical scoring functions such as the one used in Glide.

In summary, we have shown so far that a) we can dock the substrate in the active site, close to the NAD(P)H molecule and in most cases in an orientation close to what would be expected for catalysis, b) we can rank, in the majority of cases, the cognate substrate or product among the top 5% of all ligands docked, and c) in a more stringent test, where the substrate needs not only to be ranked well among other ligands, but also to be selective for its cognate protein, we have shown that reasonable dual (i.e. both for the protein and the substrate) selectivity is observed for more than half of our complexes, and for a further one-third of the dataset, selectivity is observed for either the protein or the substrate. Encouragingly, if the current approach is followed, the success of docking in discriminating the substrate does not depend strongly on the size, polarity, number of rotational bonds in the ligand, or the identity of the co-crystallised ligand in the PDB structure.

2.2 Do top-scoring ligands tell us anything about generic binding site preferences?

A general answer to this question can be found by visual inspection of Figure 10. Ligands are here represented by means of 8 well-known 1D molecular descriptors, including the number of aromatic atoms (“a.aro.”), the fraction of rotatable bonds (“b.rotR”), the number of rings (“rings”), the molecular weight (“Weight”), the formal charge (“FCharge”), the Lipinski acceptor and donor counts (“lip. acc.” and “lip. don.”, respectively) and the log octanol/water partition coefficient (“logP o.w.”). The descriptors were calculated using the Molecular Operating Environment (MOE) package from the Chemical Computing Group Inc[51]. For each studied protein (rows), the descriptors for the natural substrate (first column) along with the top 10 Glide hits (remaining columns) are shown as “star” plots drawn with the software R.

The examined cases can be divided into three groups:

a) Top 10 hits and real substrate are similar;

b) Top 10 hits are similar to each other but different from the real substrate;

c) Top 10 hits and real substrate are diverse unrelated molecules.

Ideally (case a), in each case, the top ranked ligands are expected to be similar and to share with the protein's natural ligand some molecular features (e.g. 1ae1 and 1geg). On the other hand, the worst possible *scenario* one could face (case c) is that of having diverse unrelated molecules as top hits, that achieve similar estimated binding energies, despite the fact that they don't have common functional groups (e.g. 1iy8 and 1orr). In the middle (case b) we have cases where the top 10 hits look alike, but their properties differ from those of the real substrate (e.g. 1bvr and 1pr9). In such cases, it would be difficult for the docking protocol used here to recognise the cognate substrate, or at least allow us to use the top-ranking ligands as clues to the nature of the substrate. We believe that this is made especially difficult by the multi-substrate specificities of at least some of the members of the SDR family, and we expect that the same approach using different, more selective proteins, would likely be more informative.

We have also looked more generally at the structural similarity (in 2D) of the top-ranking ligands to the cognate substrate. Previous studies on different proteins[11] have found a strong correlation between the ligand rank and its similarity to the natural substrate. We find some correlation for some of our proteins, but generally the correlation was low (The plots are available as Supplementary Information). We believe that our results are different from previous studies for a variety of reasons: a) Many of our proteins have large hydrophobic binding sites, where many ligands can bind relatively easily. In the case of smaller, highly charged binding sites one can expect that only few types of ligands can be docked successfully. b) The equivalent plots in other studies have been "smoothed", whereas we are showing the raw scores. Smoothing can eliminate many of the ligands that have low substrate similarity but a good rank, or vice versa, as long as the majority of ligands that have low substrate similarity are also ranked poorly. We think it is more helpful to show the raw scores, as in reality, there will always be ligands that have a low rank but a high substrate similarity. c) Our dataset is a lot smaller, so one can imagine that there are not enough cases of ligands that are similar to the substrate. However, we have also docked the whole of the KEGG Ligand dataset (approx 20000 molecules) to our proteins and the corresponding plots look remarkably similar to what we obtained by docking only 922 human metabolites (data not shown). We conclude that it is indeed likely to obtain metabolites resembling the natural substrate at the top ranks, but there are also many cases where this will not be true. In addition, we point out that such a correlation of the similarity with docking rank is only really possible for top ranking compounds. After a certain rank, there is no reason why the similarity to the substrate should keep on falling with falling rank.

2.3 Can we use representatives of structural classes to reduce the number of compounds screened?

The basic premise to be tested in using ligand class representatives (hereafter also called medoids) for docking (rather than all the molecules in a dataset), is whether the representative's docking profile is similar to that of all members of its class. Clearly due to the many subtleties in intermolecular interactions and their associated energies, this premise can never be wholly true, but the question is whether it is a reasonable and useful approximation. In the following paragraphs we test metabolite representatives in docking.

2.3.1 Do members of a cluster rank similarly to the cluster representative?

We have used in this study a standard molecular similarity measure to cluster our dataset of small molecule metabolites and select a representative from each cluster. There is no unique way of determining how similar two molecules are, but selected properties can be chosen to depict specific behaviours. Where molecular interactions are concerned, and more specifically recognition of a protein, the relevant properties are not known *a priori*, and reflect the nature of the binding site (which itself can change due to conformational rearrangements of the amino acids involved in binding).

Whatever similarity measure we use, our expectation is that members of a cluster will rank similarly to the cluster representative. In practise, however, we see a diversity of behaviours. Figure 11 shows the distribution of binding scores from docking our dataset of 922 molecules against each of 27 protein sites. The members of three distinct clusters (selected in order to depict the general behaviour of the 115 clusters) have been highlighted in blue colour, and the score of the medoid for the particular cluster is highlighted in red. Ideally (Figure 11a), in each case, the distribution of energies within a cluster is expected to be narrow and centred around the energy value of its medoid. Clearly, due to the intrinsic limits of the clustering process and to the differences in chemical properties of the binding sites, this is not always achievable (Figure 11c). Figure 11b shows what can be found in between these two extremes. These plots show that members of a cluster often have similar scores (to each other and to the representative), but there are clearly cases where this is not true. One of course should keep in mind that it does not make sense to compare the scores in detail. What we are looking for is whether the score of the cluster representative falls in the same quartile as that of the members of the cluster. Our results show that this is often but not always the case.

A quantitative approach to assessing the usefulness of the cluster representatives in docking can be based on a comparison of the ranking of clusters based on the ranking of their representatives, and the ranking of clusters based on the mean rank achieved by all members of this cluster. We use the root-mean-square-error (RMSE) as a measure of the difference in the rankings from docking simulations against each protein in our dataset:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N_{clusters}} (r_{med,i} - r_{mean,i})^2}{N_{clusters}}}$$

N is the number of clusters, $r_{med,i}$ is the rank of the medoid of cluster *i*, and $r_{mean,i}$ is the mean rank of all members of cluster *i*. This number, calculated for each protein in the dataset, represents the error associated with the use of the medoid instead of all members of a cluster. If calculated on the overall distribution, the RMSE values could be biased by the presence of low affinity binders. This is because differences between mean and medoid ranks of the same magnitude affect equally the calculated RMSE value, regardless of the locations of the ranks in the distribution in which they occur, while, for our purposes, the difference between low-scoring compounds is expected to be less significant than the difference between a high-scoring compound and a low-scoring one. Hence, here we only consider cases where the mean rank is less than 231 (25% of 922). In this case, RMSE values for the ranks calculated for each protein vary from 28 (for 1mg5) to 112 (for 1ahh) depending on the protein. It is worth noting that, large RMSE values cannot be ascribed just to the structural diversity of certain clusters, since, clusters performed differently, depending on the protein. Thus, it is difficult to single out cluster representatives that perform well and those that do not (indeed there is no significant correlation between this error and how “tight” a cluster is, as measured by the average dissimilarity of the molecules within the cluster). The actual correlation between the mean ranks for a cluster and the representative ranks for the same cluster for each protein of the studied dataset is on average 0.84. The high correlation values show that, overall, medoids were able to depict the general trend of the larger distribution they were chosen to represent. On the other hand, the relatively high RMSE values indicate a large diversity of ranks within a cluster.

2.3.2 Can we use representatives instead of the whole dataset?

Since not all the proteins in our dataset are of human origin, the relevant cognate ligand was not always found in the human metabolome dataset. In these cases (12 out of 27), the medoid most similar to the ligand was considered to be its representative. The pie chart in Figure 12 shows the results obtained by docking only 115 structural representatives. In 66% of the cases the medoid, representing the real substrate for a given protein, was found in the top 25% of the distribution, while in 80% of the cases it was found in the top 60% (25-50% window is actually shown in the pie chart). These results are considerably worse than what we would obtain by docking every single ligand and retaining the top 25% of hits. Hence, it would be hard to justify the use of representatives in this case. Although we find that the medoids capture at least some of the binding properties of the members of a cluster, confirming the observation that on average, structurally similar ligands tend to bind in a similar fashion[52], it is also true that within each cluster the diversity is such, that important binders can be missed when representatives are used.

Discussion/conclusions

Lately, mostly due to structural genomics initiatives, functional annotation is not proceeding *au pair* with the number of newly released X-ray structures. Since the functional information that can be deduced from the amino acid sequence alone is limited, structure-based prediction of the function of a protein is an important challenge. Here we have presented different docking protocols for identifying candidate substrates of 27 proteins with a known catalytic function from the large short-chain dehydrogenases/reductases family. In addition, pros and cons of using representative molecules for docking, instead of a whole set of plausible metabolites, in an attempt to save computational time, have been analysed.

In this study, for 2/3 of the proteins the real substrate or product was ranked within the top 5% of the entire ligand dataset using Glide, which performed slightly better than the other methodologies employed. Most of the few cases, where neither the substrate nor the product are ranked in the top 25% of the distribution, were improved by using a physics-based rescoring procedure. This is not surprising, since this kind of protocol has already been proven to be extremely helpful in the past[11,12], especially when desolvation of both protein and ligand upon binding and polarisation effects play a major role. Among the factors that can affect the quality of a docking simulation, the presence of a ligand that resembles the natural substrate bound in the starting structure seemed to play an important role (as expected). In our dataset, nearly half of the proteins were crystallised as complexes with ligands similar to the natural ones, which is the best *scenario* one could hope for when dealing with docking algorithms. As a general result, the natural substrate was appropriately docked and well-scored within the binding site, regardless of the nature of the starting structure, but we note that in two out of the four failed cases (substrate or product badly scored, or not docked at all) the crystal structure lacked a ligand bound similar to the natural one. The Induced Fit Docking protocol, which gives full flexibility to selected protein residues, helped, partly, to overcome this problem, but it can only work when major structural rearrangements are not needed. In addition, the Induced Fit Docking is not a practical solution for large-scale docking calculations, as allowing flexibility of the side-chains considerably increases the degrees of freedom, and subsequently the required computational time.

Another question we attempted to answer with this study was whether a small set of representative compounds could give us similar information on the function of a protein as that obtained by a large dataset. To our knowledge, this is the first time that this question is being addressed, although the study of Shoichet and co-workers[26] shares some similarities with ours, in that they also used only ligand family representatives when ranking their docking scores, in an attempt to increase the diversity of the top-ranking ligands. We have found that the ranks (and energies) of our naïvely constructed structural representatives correlate well with the mean ranks of the clusters they represent. However, there is great variation in the quality of the clusters, which is reflected in an equally varied distribution of ranks and energies of the compounds within a cluster. Where this distribution is not tight, the representative of the cluster clearly cannot cover the range of binding affinities observed for members of the cluster. In a way, one should expect these results: small structural changes on a molecular scaffold often lead to ligands with widely varying affinities for a receptor. This, in fact, may be more pronounced in real binding experiments, because the approximations in the scoring functions of *in silico* calculations may often overlook many small differences in the structure. However, very similar ligand structures are still, on average, more likely to exhibit similar binding patterns, as compared with very different structures, and this is what we tried to test here, with some success. Although progress in the hardware and software technology has made it possible to dock large datasets against a single protein, we believe that docking representative ligands may still have a role to play. This is because a) this approach could be directly tested with *in vitro* experiments of the same scale, giving us a glimpse of the chemical space accessible to a binding site, and b) it would allow us to routinely screen against a large set of proteins (rather than a single target), and possibly using computational protocols that are a lot more time consuming but also more accurate. This study has shown that a representative set based on simplistic 2D structure-based clustering will not generally work well enough, and so we are currently investigating alternative approaches to this problem.

However we note that a small dataset of structurally diverse ligands might be used to explore the nature of the binding sites of proteins, regardless of what their function may be. Although a single ligand may interact very differently with two binding sites and still achieve similar docked scores for both, it is increasingly unlikely that this will be the case if many ligands are used to probe these binding sites. It is then possible to assume that the profile of scores against each protein can be used to cluster the proteins. Such clusters may highlight similarities that are not obvious from comparisons of the sequences or structures of these proteins. We are currently investigating the usefulness of this approach for clustering sets of proteins that are likely to share similar sets of ligands

In our study we chose to dock the substrates and products of reactions catalysed by our SDR proteins, as these were readily available in databases and the literature. Two research groups have recently independently achieved interesting results by docking high-energy intermediates to protein structures[10,53]. This clever approach relies on the fact that proteins catalysing a reaction lower the activation energy of that reaction by stabilising the transition state. These reaction intermediates have very different electrostatic properties compared with the substrate or product, and this apparently increases the chance of successful identification of the target molecule/function. A careful study of this approach using a diverse set of proteins will be interesting to explore its general applicability. We think such a protocol cannot be readily adopted in a family of proteins like the SDRs, where the space of catalytic reactions that need to be covered is so broad, as to prohibit the automatic construction of all possible hypothetical transition states in preparation for docking.

In conclusion, we have found that the SDR family of proteins is a particularly challenging case for cognate ligand identification. This family achieves a broad substrate target range through extensive sequence variation, but also because the binding site incorporates a flexible loop. Such large scale flexibility remains a challenge for ligand docking, and renders function prediction for this family especially difficult. Whilst the results we present here are encouraging, the procedure we followed can only achieve limited accuracy. Some progress can be made using new software that accounts for protein residue mobility[46,54,55], at the price of an increased computational time, but nontrivial structural rearrangements are still difficult to predict. Perhaps more promising are new protocols that attempt to simulate the dynamics of the process of protein-ligand recognition, starting from the moment the ligand enters the binding site[56,57]. In our future work we will use the protocols developed here to study the biological role of the many new structures of human SDRs that have become available through the work of the Structural Genomics Consortium (SGC Oxford: http://www.sgc.ox.ac.uk), including experimental validation.

Materials and methods

The dataset

The proteins, substrates and products

The 27 SDR proteins in our dataset were selected from all available SDR proteins in the PDB using the following criteria:

a) The protein should be biochemically characterised, and its substrate(s) known;

b) The cofactor corresponding to the EC reaction associated with that protein should be present in the structure;

c) No residues should be missing from the active site;

d) The sequence should have no mutations introduced;

e) No protein should be included twice in the dataset.

Twenty seven proteins satisfied these criteria at the time of selection (April 2005) and are listed in Table 1.

Information on the substrates and products of these proteins was taken either from the literature corresponding to each PDB structure, or if that was not available, then the reaction information available in the KEGG database[58] or BRENDA database[59] was used. In the case of acyl-carrier-protein (ACP) -binding SDRs, we substituted the ACP moiety with a suitable coenzyme A derivative, for which we had evidence from either the literature or the BRENDA database that it is also a substrate (although it may only be so *in vitro*). Two-dimensional diagrams for the substrates and products were manually constructed and protonated. Three-dimensional structures for these were obtained using the online version of CORINA[60] (this web-accessible version is now only available as a demo on the internet site: www.mol-net.com).

The dataset of human metabolites and selection of representatives

To create the dataset of human metabolites, all 1131 small molecules listed in KEGG's human pathway compound files (version available in February 2005) were extracted. As the pathway maps of KEGG show only substrates and products (e.g. cofactors are generally missing from these maps), this list may not cover the complete human metabolome from KEGG. From this list we have additionally removed any molecules that contain generic “R” groups, or are polymers with an unspecified number of repeats. This results in a final dataset of 931 small molecules (see Supplementary Information for a list of KEGG codes). The connectivity tables for these molecules were taken directly from KEGG with no further manual inspection. All molecules used as ligands in docking experiments were prepared using the LigPrep program (i.e. hydrogens and charges were added at physiological pH, and a 3D optimised structure was calculated for each molecule). Of the 931 molecules, some are too big or have too many rotational bonds and are rejected by the docking programs, leaving 922 molecules that could be docked.

Small molecule similarity calculations

Small molecule comparisons for the clustering of the human metabolome were performed using the hashed-fingerprint algorithm available in the Chemistry Development Kit library of Java classes[61]. We have used fingerprints of 1088 bits and bond paths were calculated for up to 7 atoms. The Tanimoto similarity score was used to estimate the pairwise similarity of all pairs of molecules.

For the comparison of the substrate of a protein to each molecule docked to that protein (for the similarity vs. docking rank plots available as Supplementary Information) we used the JKlustor part of the chemistry software available from ChemAxon[62]. More specifically, we used the “generatemd” command to generate 1024-bit hashed fingerprints for each molecule, and the “compr” command to obtain a Tanimoto score for each pair of fingerprints.

Clustering of the human metabolites

Clustering was performed using the Partitioning Around Medoids (PAM) method as implemented in the R suite of statistical software[63] on the CDK fingerprint-based pairwise dissimilarity matrix. This method looks for a set of representative objects (medoids) among the observations of the dataset, around which the clusters of the data are built. The best set of representatives minimises the dissimilarities of objects in a cluster to the medoid of that cluster. An advantage of this method is that it can offer a simple way of choosing the “best” number of clusters, based on what is known as silhouette information. Each observation is assigned a silhouette width, which expresses how well it is clustered. The average silhouette width for all clusters is then an indication of how well the observations are clustered for a given number of clusters. Ideally, the optimum number of clusters would be the one with the highest average silhouette width.

We have used the silhouette width values to guide us in the selection of the number of clusters. In addition, we have inspected the clusters, removing a few of them (metals, water and oxygen, and some xenobiotic compounds included in the KEGG metabolic network), and adding some metabolites that were clearly outliers in their own clusters, and hence not well represented by the cluster medoids. The final number of metabolite representatives docked was 115 (see Supplementary Information for a list of KEGG codes).

Docking calculations

Autodock

Autodock 3.05 was used to dock 115 medoids plus all protein substrates and products, giving a total of 176 small molecules, to each of the 27 SDR proteins. Water molecules present in the crystal structures were removed, as well as inhibitors, natural substrates, or products. Cofactors were retained. Pre-processing of the proteins was done using the scripts available within the Autodock suite and its graphical user interface, ADT[64]. Ligand preparation was done with the help of the PRODRG software[65].

For each docking calculation, a box was defined around the active site, and affinity maps for each of the atom types present in the ligand dataset were calculated. The box was centred on the centre of mass of all the ligands found co-crystallised with the studied proteins, after the proteins were structurally superimposed on their C alpha atoms. The box contained 126 points along the 3 axes, spaced at 0.250 Å. Docking simulations were performed using a modified version of the genetic algorithm, with 2-point crossover and random mutations, for the global search, and an optimized version of the Solis and Wets algorithm[66] for the subsequent local minimization. All the docking parameters were set to their standard values except for the number of search runs to be done (i.e. docking poses obtained), that was set to 100. The ligand internal electrostatic contribution was also considered in the calculation. For each ligand in the dataset, only poses containing at least one atom within 4 Å from the C_4 atom of the nicotinamide ring of the cofactor were retained (see Figure 2c). Different ligands were compared and ranked on the basis of their estimated best free energy of binding.

Glide calculations

We used the Glide program from the Schrödinger software suite[67] to perform docking of all our small molecule datasets to each of the 27 SDR proteins. The protein receptors were prepared in Maestro. All waters were removed, hydrogens were added, and the grid was centred manually in each case, using the nicotinamide C_4 atom to guide the position of the centre in each case. The size of the grid was 39 Å in each direction. In all cases, we imposed a distance constraint, forcing at least one atom of each ligand to be within 4 Å of the C_4 atom in the nicotinamide ring. We did not dock ligands with more than 200 atoms or 35 rotatable bonds.

We allowed only the best pose for each ligand to be reported while docking the human metabolites dataset. When docking the representatives, we kept the ten best poses and subsequently ranked the ligands by their Glide SP scores after selecting one pose either using the best Emodel energy (which takes into account the internal energy of the ligand) or the best Glide SP score (which is more suitable for comparing the affinities of different ligands). Results from the two rankings were similar and we report the best Glide SP-selected energy, wherever we refer to results for the representatives alone. The van der Waals energies of the ligand atoms with partial charge less than 0.15 were scaled by 0.8 to soften the effect of large repulsions and allow for possible errors in the crystal structure coordinates (this is a standard and recommended procedure in Glide).

Induced Fit calculations

In the few cases where the substrate either did not dock in the binding site or it was not scored well, we performed induced fit calculations using Schrödinger's Induced Fit Docking [46], in order to "relax" the side chains of protein residues in the binding site and achieve a better fit between the substrate and its partner protein.

The Induced Fit Docking works by initially mutating to alanine the residues in the binding site that are suspected to block the binding of the substrate in the apo structure, and by docking ligands using a softened van der Waals potential. For each pose that is kept, a Prime[50] energy minimisation is performed which allows the side chains to be optimised for that pose. Once the receptor represents an induced fit structure, Glide is used to redock the ligands and finally ligand poses are scored using a combination of the Prime energy and Glide SP score.

Due to the Prime calculations, the Induced Fit Docking is considerably more time consuming than a simple Glide docking run, and hence we have only used it where Glide had failed and it was obvious that there was a problem with the starting receptor structure (such as 1b14 and 1cyd proteins). In addition, we have not included the induced fit results in any table or figure that also includes results from standard rigid-protein docking. We have only performed these calculations in addition to the rigid-protein ones to demonstrate that some of the problems in our docking results stem from side chains blocking the binding site in the apo structure.

Acknowledgements

The authors wish to thank Drs. Eric Blanc and Tim Massingham for advice on the statistics in this study and Drs. Rafael Najmanovich and Udo Oppermann for useful discussions.

Funding

A.D.F. visited the European Bioinformatics Institute with a Marie Curie Training Site Fellowship and is currently funded by the Structural Genomics Consortium (SGC) and the European Molecular Biology Laboratory (EMBL). The SGC is a registered charity (Number 1097737) funded by the Wellcome Trust, GlaxoSmithKline, Genome Canada, the Canadian Institutes of Health Research, the Ontario Innovation Trust, the Ontario Research and Development Challenge Fund, the Canadian Foundation for Innovation, Vinnova, the Swedish Strategic Research Foundation, the Knut and Alice Wallenberg Foundation, and the Karolinska Institutet.

I.N. gratefully acknowledges the UK Medical Research Council for a Training Fellowship in Bioinformatics (2001-2005) and the Royal Society for a Research Grant (2005/R2) that partly financed the purchase of licenses for the programs Glide and Prime.

TABLES

PDB code	Protein Name	Chain	EC	Organism	Substrate
1ae1*	Tropinone reductase-I	B	1.1.1.206	Thornapple	OTH
1ahh*	7 α-hydroxysteroid dehydrogenase	A	1.1.1.159	*E. Coli*	STER
1b14*	Alcohol dehydrogenase	A	1.1.1.1	Fruit fly	OTH
1bsv*	GDP-fucose synthetase	A	1.1.1.271	*E. Coli*	NUCS
1bvr	Enoyl reductase	A	1.3.1.9	*M. Tuberculosis*	COA
1c14*	Enoyl reductase	A	1.3.1.9	*E. Coli*	COA
1cyd*	Carbonyl reductase	A	1.1.1.184	Mouse	OTH
1d7o*	Enoyl reductase	A	1.3.1.9	Oilseed rape	CoA
1e6w	3-Hydroxyacyl-CoA dehydrogenase II	A	1.1.1.35	Norway rat	SMP/COA/STER
1edo*	3-Oxoacyl reductase 1	A	1.1.1.100	Oilseed rape	COA
1ek6	Udp-galactose 4-epimerase	A	5.1.3.2	Human	NUCS
1equ	Estradiol 17-beta-dehydrogenase 1	A	1.1.1.62	Human	SMP
1geg	Acetoin(diacetyl) reductase	A	1.1.1.5	*K. Pneumoniae*	SMP
1h5q*	Mannitol dehydrogenase	A	1.1.1.138	Mushroom	SMP
1ipf*	Tropinone reductase-II	A	1.1.1.236	Thornapple	OTH
1iy8	Levodione reductase	A	1.1.1.-	*C. Aquaticum*	OTH
1ja9	Tetrahydroxynaphthalene reductase	A	not assigned	Rice fungus	STER
1kc3	dTDP-4-dehydrorhamnose reductase	A	1.1.1.133	Salmonella	NUCS
1ker	dTDP-glucose 4,6-dehydratase	A	4.2.1.46	Streptococcus	NUCS
1keu	dTDP-glucose 4,6-dehydratase	A	4.2.1.46	Salmonella	NUCS
1mg5	Alcohol dehydrogenase	A	1.1.1.1	Fruit fly	SMP
1orr	CDP-tyvelose-2-epimerase	A	5.1.3.-	Salmonella	NUCS
1pr9*	L-xylulose reductase	A	1.1.1.10	Human	SMP
1q7b*	3-Oxoacyl reductase	A	1.1.1.100	*E. Coli*	COA
1sep	Sepiapterin reductase	null	1.1.1.153	Mouse	OTH
1udc	UDP-glucose 4-epimerase	null	5.1.3.2	*E. Coli*	NUCS
1w4z*	Ketoacyl reductase	A	1.3.1.-	*E. Coli*	COA

Table 1. The dataset of 27 SDR proteins used in this study presented in alphabetical order of their PDB codes. Asterisks indicate the lack of a substrate-like molecule bound in the X-ray structure.
The abbreviations used are:

- For the Substrate column: STER – Steroid (or steroid like) molecules, NUCS – Nucleotide Sugars, COA – Coenzyme A derivatives, SMP – Small, polar molecules, OTH – Others (which don’t fit in any of the previous groups).

FIGURES

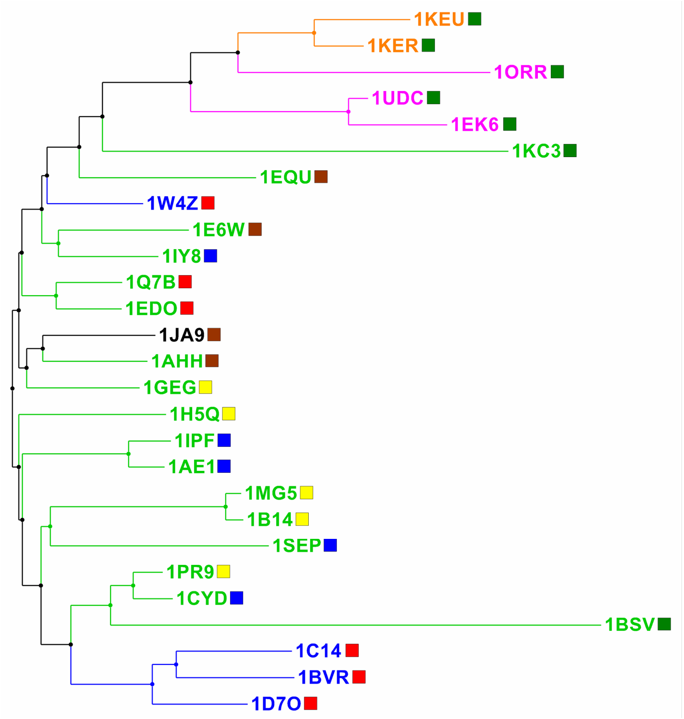


Figure 1

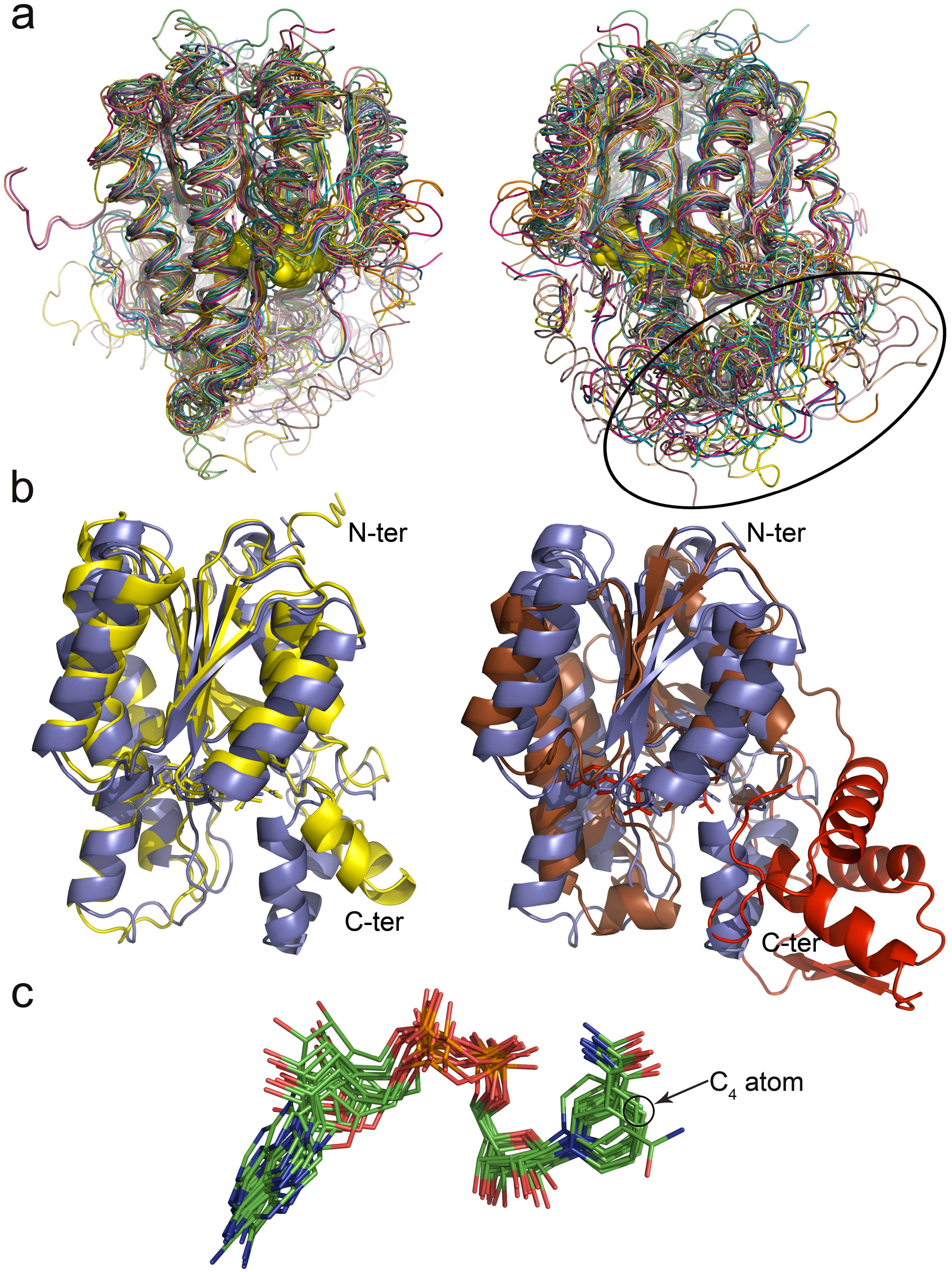


Figure 2

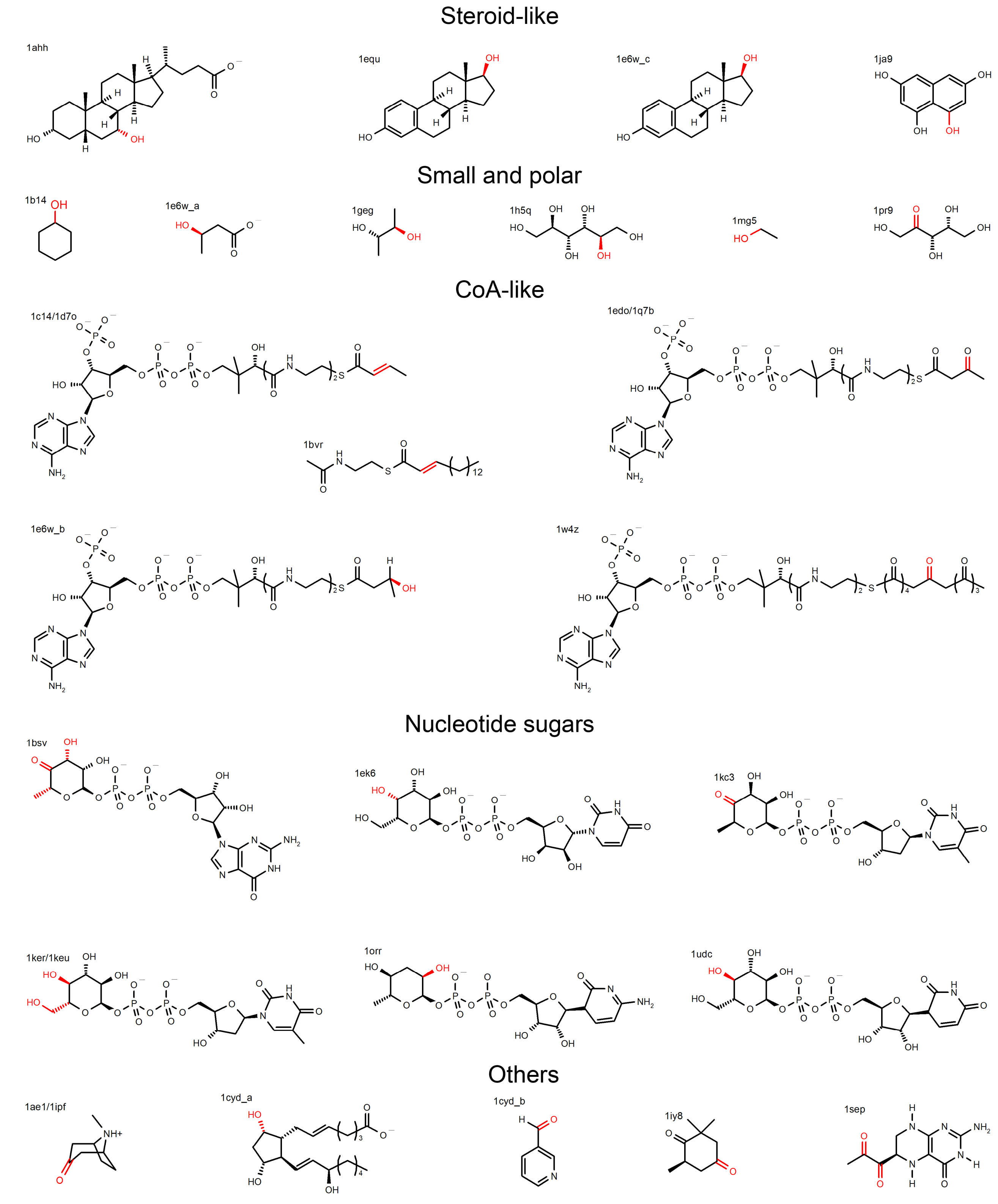

Steroid-like
1ahh
1equ
1e6w_c
1ja9
Small and polar
1b14
1e6w_a
1geg
1h5q
1mg5
1pr9
CoA-like
1c14/1d7o
1edo/1q7b
1bvr
1e6w_b
1w4z
Nucleotide sugars
1bsv
1ek6
1kc3
1ker/1keu
1orr
1udc
Others
1ae1/1ipf
1cyd_a
1cyd_b
1iy8
1sep

Figure 3

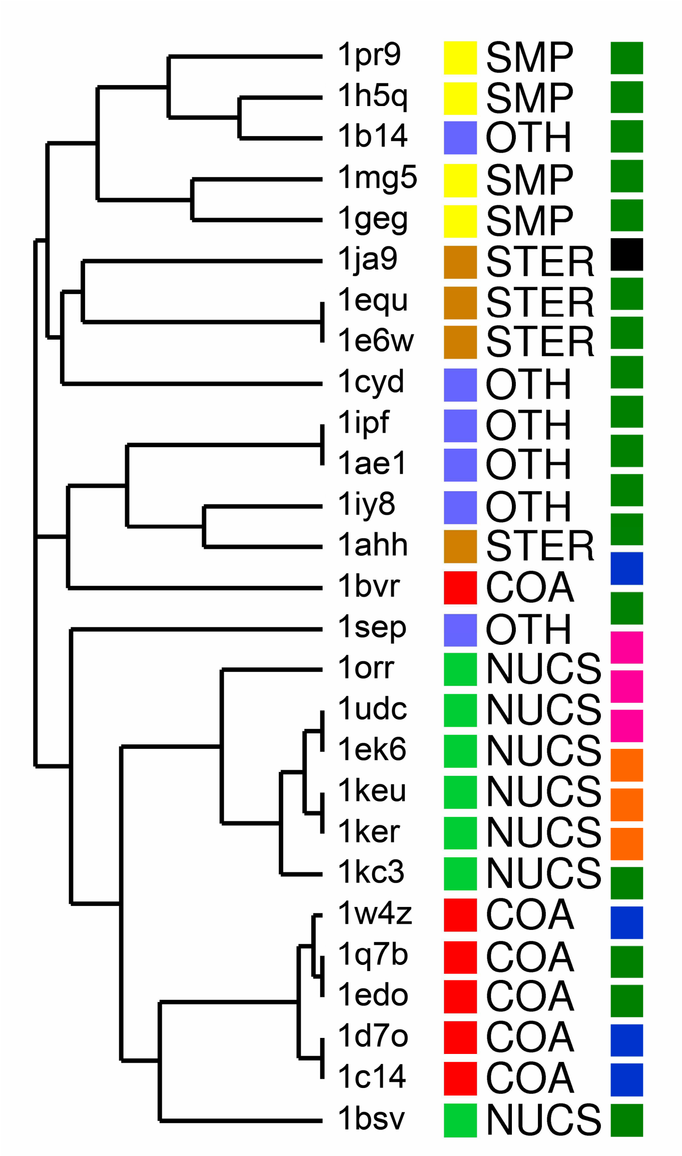

1pr9 SMP
1h5q SMP
1b14 OTH
1mg5 SMP
1geg SMP
1ja9 STER
1equ STER
1e6w STER
1cyd OTH
1ipf OTH
1ae1 OTH
1iy8 OTH
1ahh STER
1bvr COA
1sep OTH
1orr NUCS
1udc NUCS
1ek6 NUCS
1keu NUCS
1ker NUCS
1kc3 NUCS
1w4z COA
1q7b COA
1edo COA
1d7o COA
1c14 COA
1bsv NUCS

Figure 4

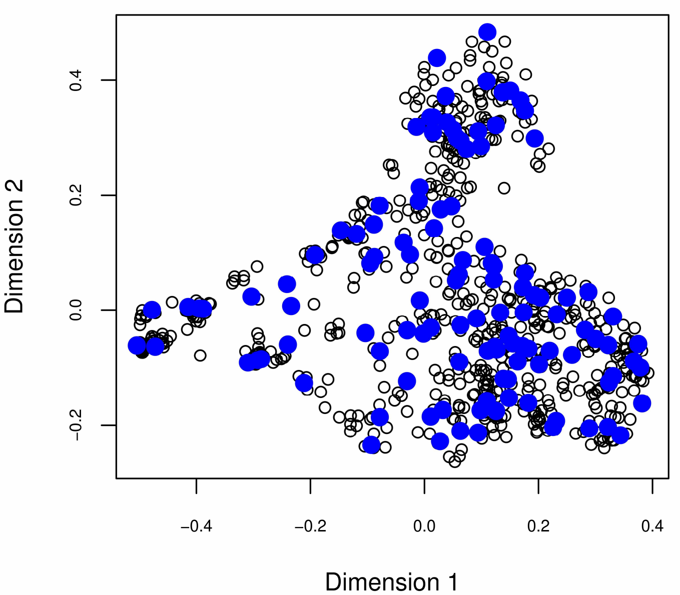


Figure 5

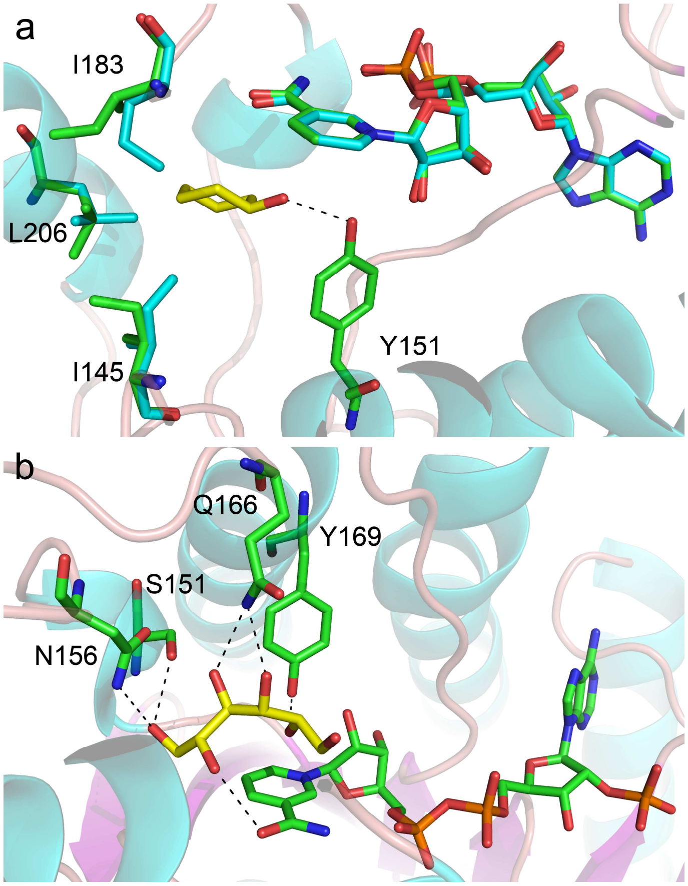


Figure 6a-b

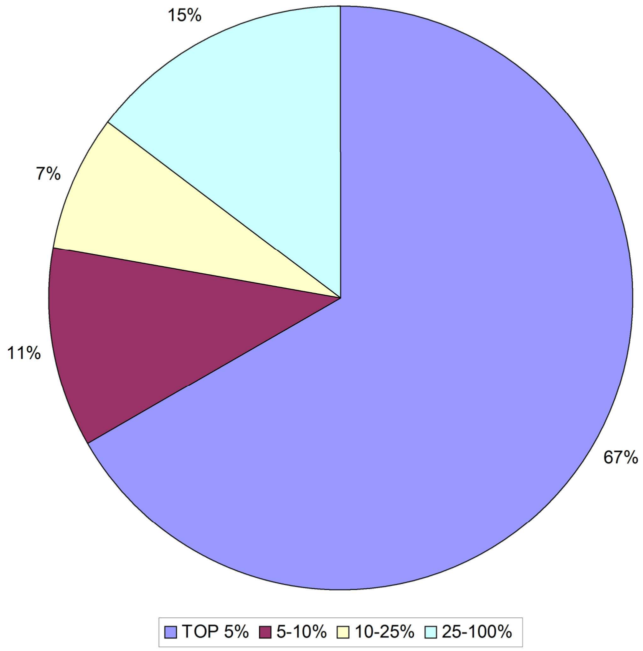


Figure 7

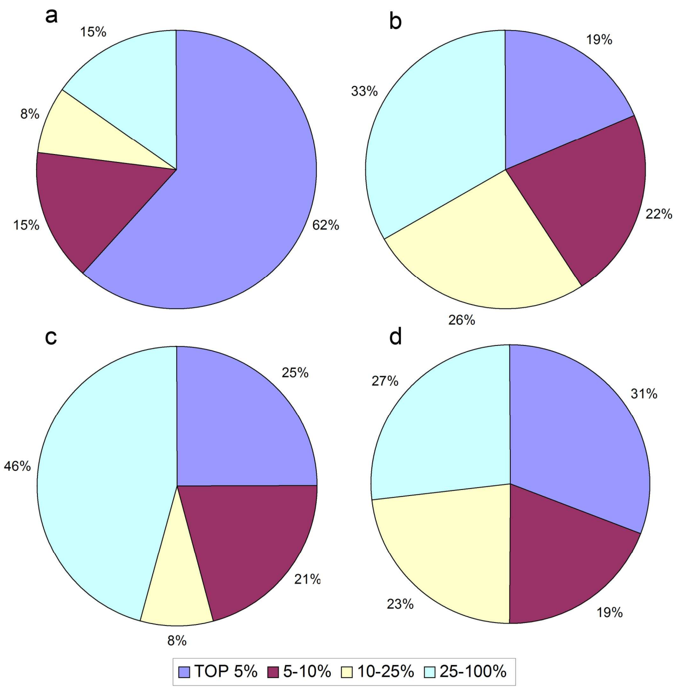


Figure 8

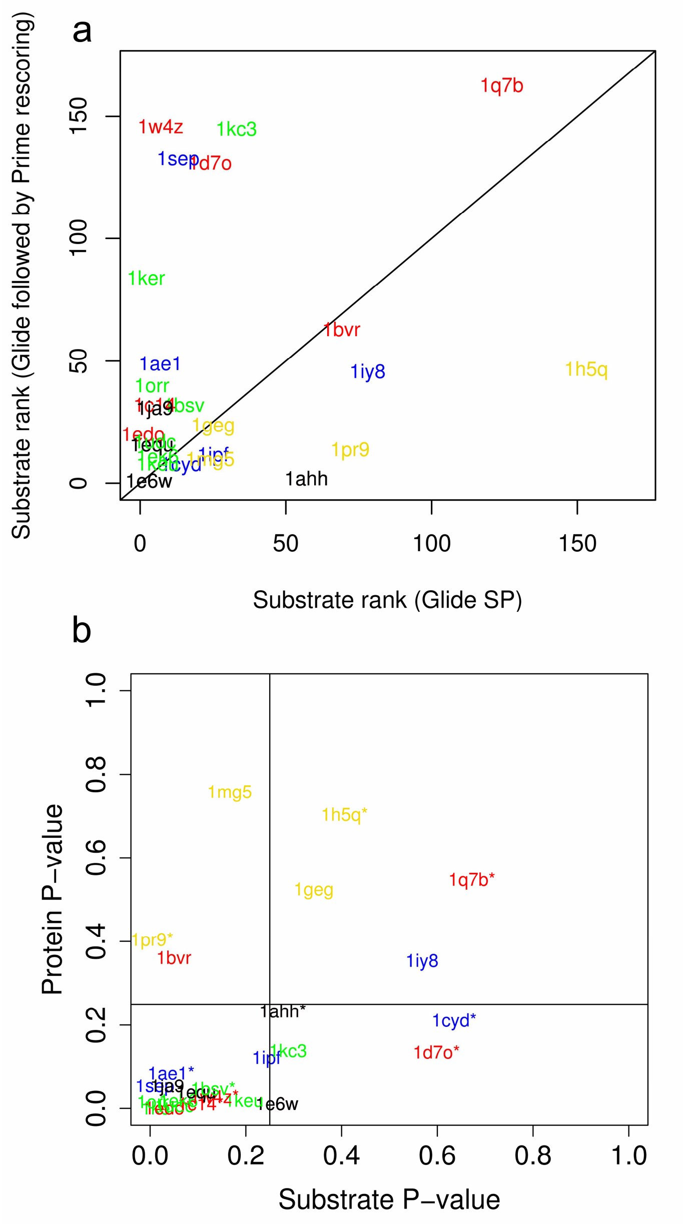

a
Substrate rank (Glide followed by Prime rescoring)
Substrate rank (Glide SP)
0
50
100
150
1q7b
1w4z
1kc3
1sep
1d7o
1ker
1bvr
1ae1
1iy8
1h5q
1orr
1c14
1bsv
1ja9
1geg
1edo
1pr9
1equ
1udc
1ek6
1ipf
1cyd
1mg5
1e6w
1ahh
b
Protein P−value
Substrate P−value
0.0
0.2
0.4
0.6
0.8
1.0
1mg5
1h5q*
1q7b*
1geg
1pr9*
1bvr
1iy8
1ahh*
1cyd*
1ipf
1kc3
1d7o*
1ae1*
1sep
1ja9
1equ
1bsv*
1w4z*
1keu
1e6w
1c14
1edo

Figure 9a-b

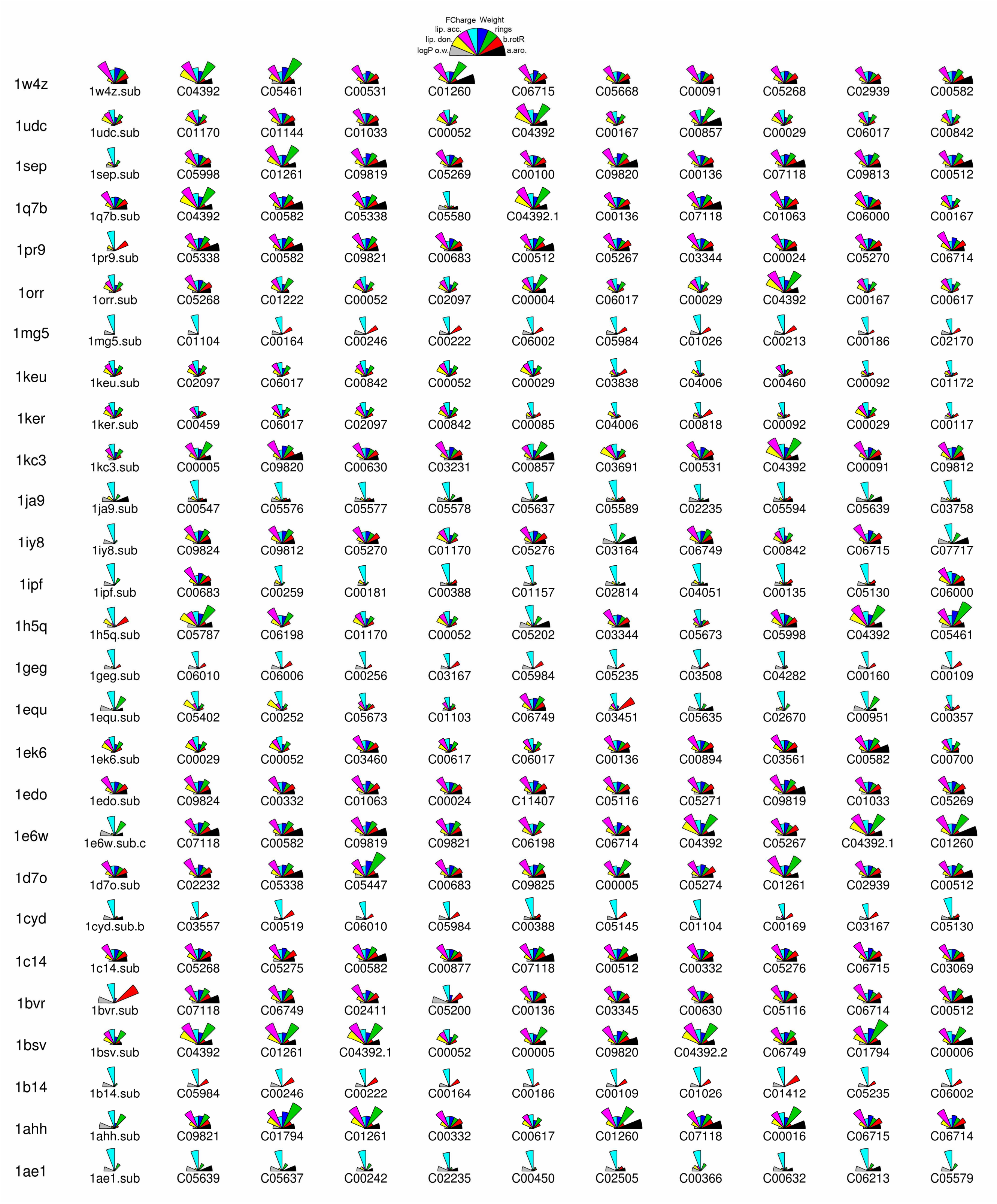

FCharge
Weight
lip. acc.
rings
lip. don.
b.rotR
logP o.w.
a.aro.
1w4z 1w4z.sub C04392 C05461 C00531 C01260 C06715 C05668 C00091 C05268 C02939 C00582
1udc 1udc.sub C01170 C01144 C01033 C00052 C04392 C00167 C00857 C00029 C06017 C00842
1sep 1sep.sub C05998 C01261 C09819 C05269 C00100 C09820 C00136 C07118 C09813 C00512
1q7b 1q7b.sub C04392 C00582 C05338 C05580 C04392.1 C00136 C07118 C01063 C06000 C00167
1pr9 1pr9.sub C05338 C00582 C09821 C00683 C00512 C05267 C03344 C00024 C05270 C06714
1orr 1orr.sub C05268 C01222 C00052 C02097 C00004 C06017 C00029 C04392 C00167 C00617
1mg5 1mg5.sub C01104 C00164 C00246 C00222 C06002 C05984 C01026 C00213 C00186 C02170
1keu 1keu.sub C02097 C06017 C00842 C00052 C00029 C03838 C04006 C00460 C00092 C01172
1ker 1ker.sub C00459 C06017 C02097 C00842 C00085 C04006 C00818 C00092 C00029 C00117
1kc3 1kc3.sub C00005 C09820 C00630 C03231 C00857 C03691 C00531 C04392 C00091 C09812
1ja9 1ja9.sub C00547 C05576 C05577 C05578 C05637 C05589 C02235 C05594 C05639 C03758
1iy8 1iy8.sub C09824 C09812 C05270 C01170 C05276 C03164 C06749 C00842 C06715 C07717
1ipf 1ipf.sub C00683 C00259 C00181 C00388 C01157 C02814 C04051 C00135 C05130 C06000
1h5q 1h5q.sub C05787 C06198 C01170 C00052 C05202 C03344 C05673 C05998 C04392 C05461
1geg 1geg.sub C06010 C06006 C00256 C03167 C05984 C05235 C03508 C04282 C00160 C00109
1equ 1equ.sub C05402 C00252 C05673 C01103 C06749 C03451 C05635 C02670 C00951 C00357
1ek6 1ek6.sub C00029 C00052 C03460 C00617 C06017 C00136 C00894 C03561 C00582 C00700
1edo 1edo.sub C09824 C00332 C01063 C00024 C11407 C05116 C05271 C09819 C01033 C05269
1e6w 1e6w.sub.c C07118 C00582 C09819 C09821 C06198 C06714 C04392 C05267 C04392.1 C01260
1d7o 1d7o.sub C02232 C05338 C05447 C00683 C09825 C00005 C05274 C01261 C02939 C00512
1cyd 1cyd.sub.b C03557 C00519 C06010 C05984 C00388 C05145 C01104 C00169 C03167 C05130
1c14 1c14.sub C05268 C05275 C00582 C00877 C07118 C00512 C00332 C05276 C06715 C03069
1bvr 1bvr.sub C07118 C06749 C02411 C05200 C00136 C03345 C00630 C05116 C06714 C00512
1bsv 1bsv.sub C04392 C01261 C04392.1 C00052 C00005 C09820 C04392.2 C06749 C01794 C00006
1b14 1b14.sub C05984 C00246 C00222 C00164 C00186 C00109 C01026 C01412 C05235 C06002
1ahh 1ahh.sub C09821 C01794 C01261 C00332 C00617 C01260 C07118 C00016 C06715 C06714
1ae1 1ae1.sub C05639 C05637 C00242 C02235 C00450 C02505 C00366 C00632 C06213 C05579

Figure 10

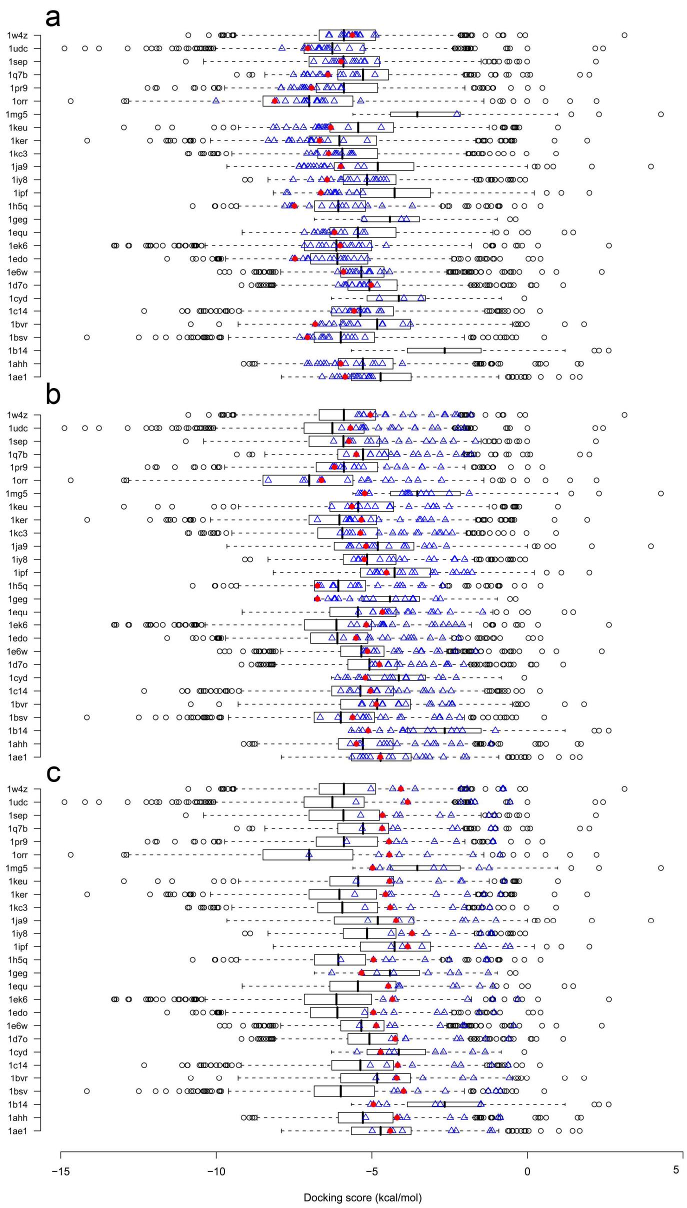

a
b
c
1w4z
1udc
1sep
1q7b
1pr9
1orr
1mg5
1keu
1ker
1kc3
1ja9
1iy8
1ipf
1h5q
1geg
1equ
1ek6
1edo
1e6w
1d7o
1cyd
1c14
1bvr
1bsv
1b14
1ahh
1ae1
−15
−10
−5
0
5
Docking score (kcal/mol)

Figure 11

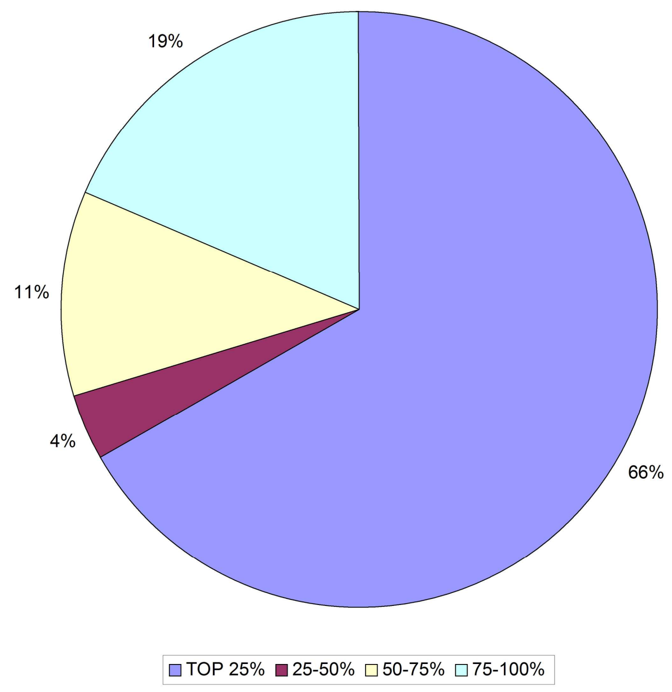


Figure 12

Figure Captions

1. Phylogram based on the multiple sequence alignment of the 27 protein sequences in our dataset. The dendrogram was obtained using ArboDraw[68]. Proteins are annotated by their EC classification (coloured text and branches) and substrate similarity (right box, colour code refers to Figure 4). Proteins sharing the first two EC levels are annotated in the same colour (green = EC 1.1, blue = EC 1.3, orange = EC 4.2 pink = EC 5.1, black = EC not assigned yet).

2. The structure of the SDR family of proteins. a) Superposition of all 27 SDR protein structures based on their C alpha carbons, two different views (obtained by a 180° rotation on the main axis) are shown. The C terminal part of the proteins, where most of the variation occurs, is highlighted. Cofactors are depicted as sphere model, coloured in yellow. Multiple alignment was performed with the Sheba software[69]. b) On the left, two superimposed single-domain proteins from our dataset are shown: PDB codes 1ae1 (in slate blue) and 1ahh (in gold yellow). On the right: the single domain protein 1ae1 (in slate blue) is compared with a two-domain protein from our dataset (1bsv). The two domains are coloured in brown (CATH 3.40.50.720) and red (CATH 3.90.25.10). N-terminals and C-terminals are highlighted. The C-terminal parts of the two domains show a lot more variation than the N-terminal parts, which bind the coenzymes. Cofactors are shown as stick models, coloured according to their parent protein. c) Superposition of all 27 NAD(P)H coenzymes (C_4 atom of the nicotinamide ring is highlighted) derived by superimposing the SDR protein structures on their protein C alpha carbons.

3. Schematic representations of the natural substrates of the 27 studied SDRs. Chemical groups involved in the catalytic process are highlighted in red. (where two proteins share the same substrate, all the functional groups involved were highlighted).

4. Clustering of the SDR substrates using a fingerprint-based similarity matrix. Molecules are annotated by the group they belong to (coloured box on the left) and by the EC of the proteins (box on the right, colour code refers to Figure 1). In the case of 1e6w and 1cyd, substrates “c” and “b” were used respectively. The substrates are annotated using a broad classification (from visual inspection) into five classes: steroids/steroid-like (STER), nucleotide sugars (NUCS), coenzyme A derivatives (COA), small & polar (SMP), and “other” (OTH). The substrate of 1ahh is a steroid but contains no aromatic ring. The substrate of 1bsv contains a larger nucleic acid base (guanine) compared with the remaining nucleotide sugar substrates.

5. The method of multidimensional scaling (as implemented in the R software suite[63]) was used to project the human metabolome in two dimensions (using the symmetric similarity matrix of all pairwise fingerprint-based scores). In this plot, each metabolite is represented by a black circle, and metabolites with similar 2D structures are expected to be close in space. The blue filled circles are the 115 metabolite representatives selected using clustering, and, as seen in the figure, they cover reasonably well the 2D projection of all metabolites.

6. a) Movement of side chains in biding site of protein (1b14) as a result of induced fit docking. Alcohol dehydrogenase from *drosophila lebanonensis* binding site, before (cyan) and after induced fit docking (green). Top ranked pose for the 1b14 substrate is shown in yellow. Residues I183, L206, I145 Y151 and NAD^+ cofactor are highlighted, depicted as stick models. The dashed line is drawn between the atoms likely to be involved in a hydrogen bond interaction. b) Top ranked pose for the 1h5q substrate, within the 1h5q protein binding site, is shown in yellow. Residues S151, N156, Q166, Y169 and $NADP^+$ cofactor are depicted as stick models. The dashed lines are drawn between the atoms likely to be involved in hydrogen bond interactions.

7. Quality of substrate recognition: Pie chart for substrate or product ranks (whichever was best) for 27 proteins. 922 ligands (human metabolome) were docked and rank of cognate substrate or product is shown, according to docking scores. The quarter of the distribution in which the substrate or product belongs to is reported.

8. Comparison of cognate ligand recognition (substrate or product ranks) for 4 different docking/scoring protocols. In each case we have docked 176 ligands and we summarise the results by binning the substrate or product ranks (whichever was best). a) Glide SP, b) Autodock, c) Glide SP poses rescored with Prime (protein was kept rigid during minimisation of the complex), d) Glide SP poses rescored with Prime and allowing the protein some flexibility during minimisation of the energy of the complex. Glide SP clearly performs better for our dataset.

9. a) Comparison of the substrate ranks resulting from the Glide SP score (x-axis) and the rescoring of the Glide poses with Prime (y-axis, allowing minimisation of the protein binding site). In case of multiple substrates, only the best-scoring substrate is shown. The protein 1b14 is not in the plot as the substrate pose did not pass our Glide distance filter. b) Substrate *p* values vs. protein *p* values for the 27 cognate pairs of enzyme-substrate complexes. The substrate *p* value is an estimate of how well the substrate recognises its protein partner, based on the docking scores of this substrate against each of 27 proteins.

The protein *p* value is an estimate of how well the protein recognises its substrate partner, based on the docking scores of 176 ligands docked to that protein. The Glide scores used to prepare this plot were the best Glide SP scores from a list of 10 candidate high-scoring poses. We have also tried to select the best pose using Glide' s E_{model} energy, but we found that the results were slightly worse in this case. Labels of both plots are coloured accordingly to the following scheme: black for steroid-like molecules, yellow for small and polar molecules, red for CoA-like compounds, green for nucleotide sugars and blue for the remaining molecules. Asterisks indicate the lack of a substrate-like molecule bound in the starting X-ray structure.

10. The function "stars" in R[63] was used to depict as “star plots” the relative magnitude of 8 1D descriptors of the top 10 Glide hits for each of the 27 enzymes. The same descriptors are also shown for the substrate (first column) for comparison. The following descriptor values were calculated using the MOE descriptor calculation software[51]: Number of aromatic atoms (a.aro.), fraction of rotatable bonds (b.rotR), number of rings in the molecule (rings), molecular weight (Weight), sum of formal charges (FCharge), Lipinski acceptor count (lip.acc.), Lipinksi donor count (lip.don.), log octanol/water partition coefficient (logP o.w.). All descriptors have been scaled between 0 and 1, as required by the stars plotting function. Scaling was applied to values of each descriptor within the set of all human metabolites + all substrates, so the scale used is the same for all rows in this plot.

11. Box-and-whisker plots for the distribution of Glide SP scores resulting from docking 922 small molecule ligands to 27 SDR proteins. Each box-and-whisker plot is created using the distribution of all available scores for a given protein binding site (not necessarily 922, as some molecules fail to dock, depending on the protein). The scores of a given cluster of ligands are then annotated on top of this distribution as blue triangles, and the score of the medoid (where available), is shown as a red bullet. We only show three (randomly selected) clusters in this figure (from a total of 115). This plot was created using R[63].

12. Usefulness of medoids: The pie chart summarises the results obtained by docking only 115 structural representatives to each of the 27 SDR proteins, and then reporting the quarter of the distribution of scores to which the representative closest to the real substrate belongs to. For example, in 66% of cases, the structural representative closest to the cognate substrate is ranked in the top 25%, according to the Glide SP scores.

References

1. Chance MR, Bresnick AR, Burley SK, Jiang JS, Lima CD, et al. (2002) Structural genomics: a pipeline for providing structures for the biologist. Protein Sci 11: 723-738.
2. Stevens RC, Yokoyama S, Wilson IA (2001) Global efforts in structural genomics. Science 294: 89-92.
3. Brenner SE (2001) A tour of structural genomics. Nat Rev Genet 2: 801-809.
4. Burley SK (2000) An overview of structural genomics. Nat Struct Biol 7 Suppl: 932-934.
5. Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, et al. (1999) Structural genomics: beyond the human genome project. Nat Genet 23: 151-157.
6. Watson JD, Laskowski RA, Thornton JM (2005) Predicting protein function from sequence and structural data. Curr Opin Struct Biol 15: 275-284.
7. Jorgensen WL (2004) The many roles of computation in drug discovery. Science 303: 1813-1818.
8. Brooijmans N, Kuntz ID (2003) Molecular recognition and docking algorithms. Annu Rev Biophys Biomol Struct 32: 335-373.
9. Shoichet BK, McGovern SL, Wei B, Irwin JJ (2002) Lead discovery using molecular docking. Curr Opin Chem Biol 6: 439-446.
10. Hermann JC, Ghanem E, Li Y, Raushel FM, Irwin JJ, et al. (2006) Predicting substrates by docking high-energy intermediates to enzyme structures. J Am Chem Soc 128: 15882-15891.
11. Kalyanaraman C, Bernacki K, Jacobson MP (2005) Virtual screening against highly charged active sites: identifying substrates of alpha-beta barrel enzymes. Biochemistry 44: 2059-2071.
12. Bernacki K, Kalyanaraman C, Jacobson MP (2005) Virtual ligand screening against Escherichia coli dihydrofolate reductase: improving docking enrichment using physics-based methods. J Biomol Screen 10: 675-681.
13. Macchiarulo A, Nobeli I, Thornton JM (2004) Ligand selectivity and competition between enzymes in silico. Nat Biotechnol 22: 1039-1045.
14. Cummings MD, DesJarlais RL, Gibbs AC, Mohan V, Jaeger EP (2005) Comparison of automated docking programs as virtual screening tools. J Med Chem 48: 962-976.
15. Cole JC, Murray CW, Nissink JW, Taylor RD, Taylor R (2005) Comparing protein-ligand docking programs is difficult. Proteins 60: 325-332.
16. Perola E, Walters WP, Charifson PS (2004) A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. Proteins 56: 235-249.
17. Bursulaya BD, Totrov M, Abagyan R, Brooks CL, 3rd (2003) Comparative study of several algorithms for flexible ligand docking. J Comput Aided Mol Des 17: 755-763.
18. Wang R, Lu Y, Wang S (2003) Comparative evaluation of 11 scoring functions for molecular docking. J Med Chem 46: 2287-2303.
19. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, et al. (2006) A critical assessment of docking programs and scoring functions. J Med Chem 49: 5912-5931.
20. Persson B, Kallberg Y, Oppermann U, Jornvall H (2003) Coenzyme-based functional assignments of short-chain dehydrogenases/reductases (SDRs). Chem Biol Interact 143-144: 271-278.
21. Oppermann U, Filling C, Hult M, Shafqat N, Wu X, et al. (2003) Short-chain dehydrogenases/reductases (SDR): the 2002 update. Chem Biol Interact 143-144: 247-253.
22. Kallberg Y, Oppermann U, Jornvall H, Persson B (2002) Short-chain

dehydrogenases/reductases (SDRs). Eur J Biochem 269: 4409-4417.

23. Kallberg Y, Oppermann U, Jornvall H, Persson B (2002) Short-chain dehydrogenase/reductase (SDR) relationships: a large family with eight clusters common to human, animal, and plant genomes. Protein Sci 11: 636-641.
24. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, et al. (1998) Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19: 1639-1662.
25. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, et al. (2006) Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. J Med Chem 49: 6177-6196.
26. Su AI, Lorber DM, Weston GS, Baase WA, Matthews BW, et al. (2001) Docking molecules by families to increase the diversity of hits in database screens: computational strategy and experimental evaluation. Proteins 42: 279-293.
27. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235-242.
28. Otagiri M, Kurisu G, Ui S, Takusagawa Y, Ohkuma M, et al. (2001) Crystal structure of meso-2,3-butanediol dehydrogenase in a complex with NAD+ and inhibitor mercaptoethanol at 1.7 A resolution for understanding of chiral substrate recognition mechanisms. J Biochem (Tokyo) 129: 205-208.
29. Koropatkin NM, Liu HW, Holden HM (2003) High resolution x-ray structure of tyvelose epimerase from Salmonella typhi. J Biol Chem 278: 20874-20881.
30. Benach J, Atrian S, Gonzalez-Duarte R, Ladenstein R (1998) The refined crystal structure of Drosophila lebanonensis alcohol dehydrogenase at 1.9 A resolution. J Mol Biol 282: 383-399.
31. Benach J, Winberg JO, Svendsen JS, Atrian S, Gonzalez-Duarte R, et al. (2005) Drosophila alcohol dehydrogenase: acetate-enzyme interactions and novel insights into the effects of electrostatics on catalysis. J Mol Biol 345: 579-598.
32. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.
33. Somers WS, Stahl ML, Sullivan FX (1998) GDP-fucose synthetase from Escherichia coli: structure of a unique member of the short-chain dehydrogenase/reductase family that catalyzes two distinct reactions at the same active site. Structure 6: 1601-1612.
34. Thoden JB, Wohlers TM, Fridovich-Keil JL, Holden HM (2000) Crystallographic evidence for Tyr 157 functioning as the active site base in human UDP-galactose 4-epimerase. Biochemistry 39: 5691-5701.
35. Blankenfeldt W, Kerr ID, Giraud MF, McMiken HJ, Leonard G, et al. (2002) Variation on a theme of SDR. dTDP-6-deoxy-L- lyxo-4-hexulose reductase (RmlD) shows a new Mg2+-dependent dimerization mode. Structure 10: 773-786.
36. Allard ST, Beis K, Giraud MF, Hegeman AD, Gross JW, et al. (2002) Toward a structural understanding of the dehydratase mechanism. Structure 10: 81-92.
37. Thoden JB, Hegeman AD, Wesenberg G, Chapeau MC, Frey PA, et al. (1997) Structural analysis of UDP-sugar binding to UDP-galactose 4-epimerase from Escherichia coli. Biochemistry 36: 6294-6304.
38. Hadfield AT, Limpkin C, Teartasin W, Simpson TJ, Crosby J, et al. (2004) The crystal structure of the actIII actinorhodin polyketide reductase: proposed mechanism for ACP and polyketide binding. Structure 12: 1865-1875.
39. Orengo CA, Pearl FM, Thornton JM (2003) The CATH domain structure database. Methods Biochem Anal 44: 249-271.
40. LigPrep, version 2.0, Schrödinger, LLC, New York, NY, 2005.

41. Tanaka N, Nonaka T, Nakanishi M, Deyashiki Y, Hara A, et al. (1996) Crystal structure of the ternary complex of mouse lung carbonyl reductase at 1.8 A resolution: the structural origin of coenzyme specificity in the short-chain dehydrogenase/reductase family. Structure 4: 33-45.
42. Gunasekaran K, Nussinov R (2007) How different are structurally flexible and rigid binding sites? Sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding. J Mol Biol 365: 257-273.
43. Hammes GG (2002) Multiple conformational changes in enzyme catalysis. Biochemistry 41: 8221-8228.
44. Joseph D, Petsko GA, Karplus M (1990) Anatomy of a conformational change: hinged "lid" motion of the triosephosphate isomerase loop. Science 249: 1425-1428.
45. Koshland DE, Jr. (1963) Correlation of Structure and Function in Enzyme Action. Science 142: 1533-1541.
46. Sherman W, Day T, Jacobson MP, Friesner RA, Farid R (2006) Novel procedure for modeling ligand/receptor induced fit effects. J Med Chem 49: 534-553.
47. Horer S, Stoop J, Mooibroek H, Baumann U, Sassoon J (2001) The crystallographic structure of the mannitol 2-dehydrogenase NADP+ binary complex from Agaricus bisporus. J Biol Chem 276: 27555-27561.
48. Rozwarski DA, Vilcheze C, Sugantino M, Bittman R, Sacchettini JC (1999) Crystal structure of the Mycobacterium tuberculosis enoyl-ACP reductase, InhA, in complex with NAD+ and a C16 fatty acyl substrate. J Biol Chem 274: 15582-15589.
49. El-Kabbani O, Ishikura S, Darmanin C, Carbone V, Chung RP, et al. (2004) Crystal structure of human L-xylulose reductase holoenzyme: probing the role of Asn107 with site-directed mutagenesis. Proteins 55: 724-732.
50. Prime, version 1.6, Schrödinger, LLC, New York, NY, 2007.
51. Molecular Operating EnVironment; Chemical Computing Group Inc. http://www.chemcomp.com/.
52. Bostrom J, Hogner A, Schmitt S (2006) Do structurally similar ligands bind in a similar fashion? J Med Chem 49: 6716-6725.
53. Tyagi S, Pleiss J (2006) Biochemical profiling in silico--predicting substrate specificities of large enzyme families. J Biotechnol 124: 108-116.
54. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WT, et al. (2007) Diverse, high-quality test set for the validation of protein-ligand docking performance. J Med Chem 50: 726-741.
55. Huey R, Morris GM, Olson AJ, Goodsell DS (2007) A semiempirical free energy force field with charge-based desolvation. J Comput Chem 28: 1145-1152.
56. Borrelli KW, Vitalis A, Alcantara R, Guallar V (2005) PELE: Protein energy landscape exploration. A novel Monte Carlo based technique. Journal of Chemical Theory and Computation 1: 1304-1311.
57. Gervasio FL, Laio A, Parrinello M (2005) Flexible docking in solution using metadynamics. J Am Chem Soc 127: 2600-2607.
58. Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. Nucleic Acids Res 30: 42-46.
59. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, et al. (2004) BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res 32: D431-433.
60. Gasteiger J, Rudolph C, Sadowski J (1990) Automatic generation of 3D-atomic coordinates for organic molecules. Tetrahedron Comp Method 3: 537-547.
61. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, et al. (2003) The Chemistry

Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. J Chem Inf Comput Sci 43: 493-500.

62. JChem 3.2, 2006 (http://www.chemaxon.com).

63. R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

64. Sanner MF (1999) Python: A programming language for software integration and development. Journal of Molecular Graphics & Modelling 17: 57-61.

65. Schuttelkopf AW, van Aalten DM (2004) PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. Acta Crystallogr D Biol Crystallogr 60: 1355-1363.

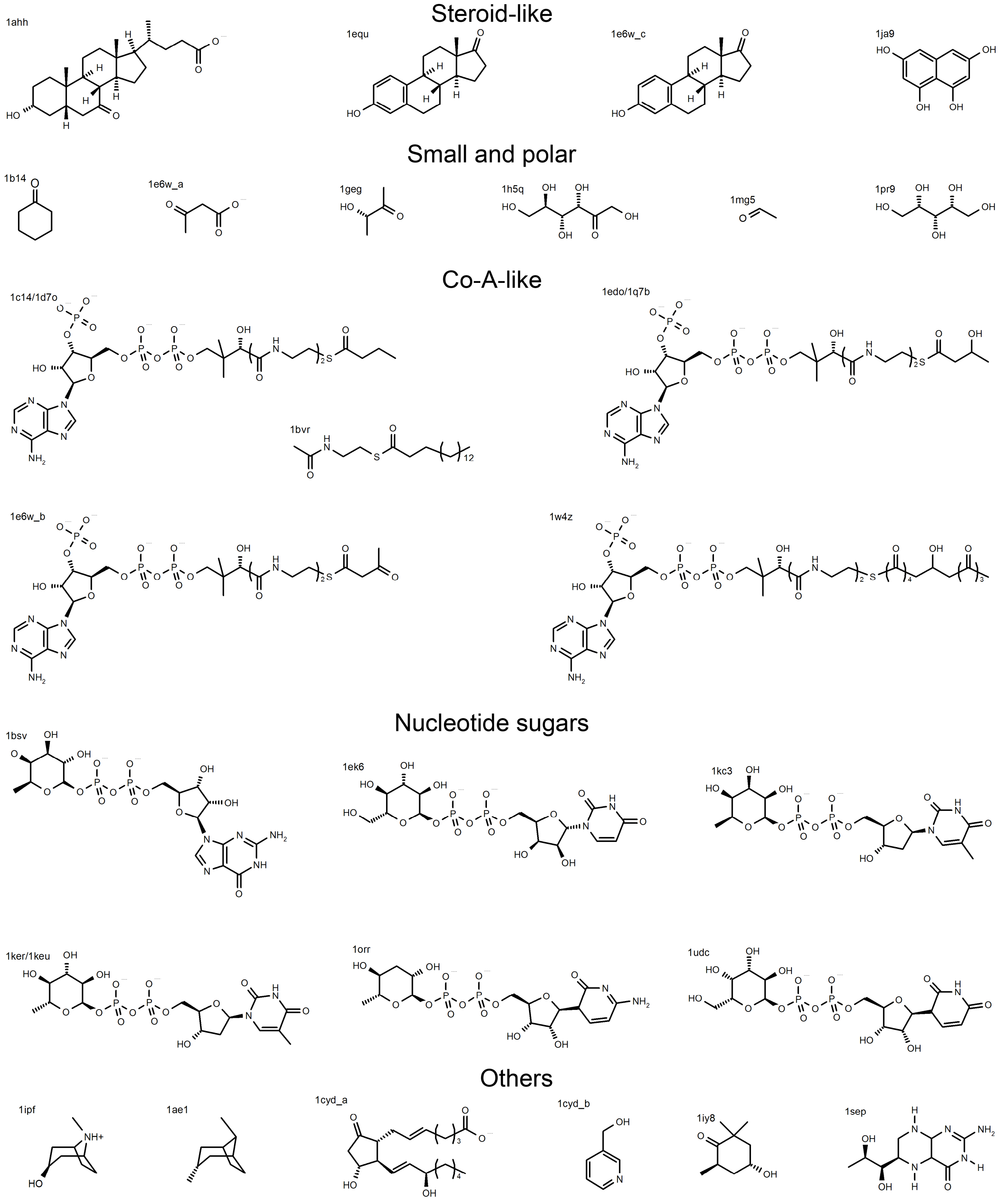
66. Solis FJ, Wets RJB (1981) Minimization by Random Search Techniques. Mathematics of Operations Research 6: 19-30.

67. Schrödinger Suite 2006 Induced Fit Docking protocol; Glide version 4.0, Schrödinger, LLC, New York, NY, 2005; Prime version 1.5, Schrödinger, LLC, New York, NY, 2005.

68. ArboDraw, 2007 (http://dunbrack.fccc.edu/ArboDraw).

69. Jung J, Lee B (2000) Protein structure alignment using environmental profiles. Protein Eng 13: 535-543.

SUPPLEMENTARY INFORMATION



SI-1 Schematic representations of the natural products of the 27 studied SDRs.

C00006	C00962	C06099									
C00009	C01026	C07717									
C00020	C01134	C07725									
C00021	C01161	C07731									
C00025	C01176	C07733									
C00033	C01181	C11136									
C00035	C01197	C11356									
C00051	C01346	C11431									
C00052	C01384	C11432									
C00061	C01412										
C00067	C01697										
C00068	C02166										
C00082	C02222										
C00085	C02470										
C00094	C02505										
C00106	C02571										
C00112	C02723										
C00136	C02727										
C00153	C03267										
C00158	C03360										
C00164	C03557										
C00180	C03758										
C00183	C03824										
C00187	C04006										
C00197	C04244										
C00231	C04281										
C00233	C04376										
C00250	C04805										
C00256	C04895										
C00262	C05119										
C00319	C05202										
C00346	C05300										
C00386	C05404										
C00438	C05419										
C00440	C05462										
C00455	C05480										
C00469	C05485										
C00486	C05503										
C00515	C05552										
C00519	C05579										
C00523	C05583										
C00565	C05635										
C00581	C05639										
C00606	C05651										
C00628	C05655										
C00645	C05766										
C00655	C05804										
C00763	C05811										
C00780	C05827										
C00792	C05850										
C00822	C05890										
C00899	C05932										
C00906	C05959										

SI-2 List of the 115 KEGGS identifiers corresponding to the human representatives (medoids) used in this study.

PDB – substrate (S)/product (P)	Distance of C4 to reaction centre	Glide Score	Rank among representatives (176)	rank % among representatives	Rank among all human metabolites (922)	rank% among all human metabolites
1ae1 - S	3.28	-6.54	7	3.98	27	2.93
1ae1 - P	3.71	-6.42	10	5.68	39	4.23
1ahh - S	6.43	-6.08	57	32.39	223	24.19
1ahh - P	7.11	-5.16	118	67.05	525	56.94
1b14 - S	Not docked					
1b14 - P	Not docked					
1bsv - S	17.28	-8.7	15	8.52	521	56.51
1bsv - P	5.73	-8.97	9	5.11	39	4.23
1bvr - S	10.98	-5.48	69	39.20	278	30.15
1bvr - P	10.15	-5.17	82	46.59	361	39.15
1c14 - S	4.69	-9.4	5	2.84	28	3.04
1c14 - P	3.95	-8.75	10	5.68	29	3.15
1cyd - S_a	Not docked					
1cyd - P_a	Not docked					
1cyd - S_b	4.26	-5.33	14	7.95	49	5.31
1cyd - P_b	4.45	-5.67	10	5.68	72	7.81
1d7o - S	3.71	-6.32	24	13.64	102	11.06
1d7o - P	3.17	-7.85	4	2.27	23	2.49
1e6w - Sa	5.49	-4.35	144	81.82		0.11
1e6w - Pa	5.46	-5.79	62	35.23		0.11
1e6w - Sb	4.77	-8.18	3	1.70	18	1.95
1e6w - Pb	6.64	-9.06	1	0.57	5	0.54
1e6w - Sc	11.10	-6.68	18	10.23		0.11
1e6w - Pc	11.13	-6.67	19	10.80		0.11
1edo - S	3.27	-10.71	1	0.57	5	0.54
1edo – P	4.09	-9.48	3	1.70	122	13.23
1ek6 - S	3.37	-10.91	6	3.41	17	1.84
1ek6 - P	3.36	-10.46	11	6.25	30	3.25
1equ - S	3.08	-7.77	4	2.27	10	1.08
1equ - P	3.07	-7.36	8	4.55	26	2.82
1geg - S	3.52	-4.32	25	14.20	126	13.67
1geg - Pa	2.95	-6.4	4	2.27		0.11
1geg – Pb	3.00	-7.11	1	0.57	1	0.11
1h5q – S	3.09	-4.72	153	86.93	795	86.23
1h5q - P	3.26	-6.01	109	61.93	852	92.41

1ipf - S	7.49	-5.84	25	14.20	171	18.55
1ipf - P	4.28	-6.69	2	1.14	16	1.74
1iy8 – S	5.01	-5.49	78	44.32	351	38.07
1iy8 – P	5.50	-6.77	18	10.23	57	6.18
1ja9 - S	3.88	-7.75	5	2.84	37	4.01
1ja9 - P	5.15	-7.74	6	3.41	49	5.31
1kc3 - S	18.62	-7.32	33	18.75	215	23.32
1kc3 - P	3.96	-7.36	31	17.61	310	33.62
1ker - S	4.59	-13.95	2	1.14	2	0.22
1ker – P	3.53	-12.7	4	2.27	3	0.33
1keu - S	3.41	-9.49	6	3.41	8	0.87
1keu – P	3.29	-11.16	3	1.70	4	0.43
1mg5 - S	3.46	-0.73	24	13.64	150	16.27
1mg5 - P	3.47	-4.74	13	7.39	33	3.58
1orr – S	12.87	-12.14	4	2.27	11	1.19
1orr – P	3.75	-10.86	14	7.95	231	25.05
1pr9 – S	3.99	-6.03	72	40.91	371	40.24
1pr9 – P	3.54	-5.54	88	50.00	427	46.31
1q7b – S	16.26	-4.93	124	70.45	782	84.82
1q7b – P	16.05	-8.48	1	0.57	4	0.43
1sep - S	4.21	-8.48	13	7.39	53	5.75
1sep – P	4.03	-7.86	23	13.07	127	13.77
1udc – S	4.90	-12.41	5	2.84	5	0.54
1udc – P	3.82	-10.5	15	8.52	20	2.17
1w4z – S	3.07	-8.74	7	3.98	36	3.90
1w4z - P	Not docked					

SI-3 A qualitative measure of the docking poses. For each protein, substrate and product binding poses (column 1, pdb code-S and -P, respectively) are analyzed in terms of distance between the C4 (see Figure 2c) and the reaction centre, of Glide score, actual rank and % rank among representatives and among all human metabolites (columns from 2 to 7).

C00001	C00073	C00143	C00239	C00363	C00519	C00683	C01026	C01230	C02305
C00002	C00074	C00144	C00242	C00364	C00523	C00687	C01031	C01235	C02325
C00003	C00075	C00146	C00243	C00365	C00524	C00689	C01033	C01236	C02336
C00004	C00077	C00147	C00245	C00366	C00526	C00696	C01035	C01243	C02373
C00005	C00078	C00148	C00246	C00376	C00527	C00700	C01036	C01245	C02406
C00006	C00079	C00149	C00248	C00378	C00530	C00705	C01042	C01259	C02411
C00007	C00081	C00152	C00249	C00379	C00531	C00719	C01044	C01260	C02442
C00008	C00082	C00153	C00250	C00385	C00532	C00735	C01051	C01261	C02465
C00009	C00083	C00154	C00252	C00386	C00534	C00750	C01054	C01262	C02470
C00010	C00084	C00155	C00253	C00387	C00535	C00751	C01060	C01272	C02501
C00011	C00085	C00158	C00255	C00388	C00544	C00762	C01061	C01284	C02505
C00013	C00086	C00159	C00256	C00398	C00546	C00763	C01063	C01312	C02512
C00014	C00089	C00160	C00258	C00402	C00547	C00780	C01079	C01344	C02515
C00015	C00091	C00163	C00259	C00407	C00555	C00785	C01081	C01345	C02538
C00016	C00092	C00164	C00262	C00408	C00559	C00788	C01083	C01346	C02571
C00018	C00093	C00166	C00267	C00410	C00565	C00792	C01089	C01353	C02576
C00019	C00094	C00167	C00268	C00415	C00570	C00794	C01094	C01380	C02593
C00020	C00095	C00168	C00270	C00417	C00575	C00804	C01096	C01384	C02637
C00021	C00096	C00169	C00272	C00418	C00577	C00811	C01097	C01412	C02642
C00022	C00097	C00170	C00275	C00427	C00579	C00818	C01103	C01419	C02646
C00023	C00099	C00178	C00279	C00429	C00581	C00819	C01104	C01494	C02670
C00024	C00100	C00179	C00280	C00430	C00582	C00822	C01107	C01528	C02714
C00025	C00101	C00180	C00286	C00437	C00583	C00831	C01110	C01596	C02723
C00026	C00103	C00181	C00294	C00438	C00584	C00836	C01120	C01598	C02727
C00029	C00104	C00183	C00295	C00439	C00588	C00842	C01134	C01613	C02734
C00031	C00105	C00184	C00299	C00440	C00590	C00847	C01136	C01674	C02759
C00032	C00106	C00185	C00300	C00445	C00601	C00857	C01137	C01693	C02814
C00033	C00108	C00186	C00301	C00446	C00603	C00864	C01143	C01697	C02835
C00035	C00109	C00187	C00307	C00447	C00606	C00870	C01144	C01724	C02888
C00036	C00111	C00188	C00311	C00448	C00617	C00877	C01149	C01762	C02918
C00037	C00112	C00189	C00314	C00449	C00620	C00881	C01152	C01780	C02934
C00041	C00114	C00191	C00315	C00450	C00624	C00882	C01157	C01794	C02939
C00042	C00116	C00197	C00319	C00455	C00627	C00894	C01159	C01801	C02946
C00043	C00117	C00198	C00322	C00458	C00628	C00899	C01161	C01829	C02985
C00044	C00118	C00199	C00325	C00459	C00630	C00900	C01164	C01832	C02990
C00047	C00119	C00206	C00327	C00460	C00631	C00906	C01165	C01888	C03028
C00048	C00120	C00208	C00328	C00468	C00632	C00909	C01168	C01921	C03069
C00049	C00121	C00212	C00329	C00469	C00636	C00921	C01169	C01944	C03087
C00051	C00122	C00213	C00330	C00472	C00637	C00931	C01170	C01953	C03090
C00052	C00123	C00214	C00332	C00473	C00639	C00937	C01172	C01962	C03150
C00053	C00124	C00217	C00334	C00475	C00642	C00942	C01176	C01996	C03164
C00054	C00127	C00219	C00337	C00483	C00643	C00944	C01177	C02043	C03167
C00055	C00128	C00221	C00341	C00486	C00645	C00951	C01179	C02094	C03205
C00058	C00129	C00222	C00345	C00487	C00647	C00954	C01181	C02097	C03221
C00059	C00130	C00224	C00346	C00490	C00655	C00956	C01185	C02110	C03227
C00061	C00131	C00227	C00352	C00491	C00664	C00957	C01189	C02140	C03231
C00062	C00132	C00230	C00353	C00492	C00665	C00962	C01197	C02165	C03232
C00063	C00134	C00231	C00355	C00499	C00668	C00978	C01204	C02166	C03263
C00064	C00135	C00232	C00356	C00500	C00669	C00986	C01211	C02170	C03267
C00065	C00136	C00233	C00357	C00504	C00670	C01005	C01213	C02191	C03287
C00067	C00137	C00234	C00360	C00506	C00672	C01013	C01220	C02198	C03373
C00068	C00140	C00235	C00361	C00512	C00673	C01019	C01222	C02218	C03406
C00072	C00141	C00236	C00362	C00515	C00674	C01024	C01227	C02222	C03410
C03344	C04352	C05268	C05457	C05596	C05830	C06142	C08060	C02232	C03415
C03345	C04373	C05269	C05458	C05598	C05831	C06144	C08061	C02235	C03428
C03360	C04376	C05270	C05460	C05619	C05832	C06145	C08062	C02291	C03440

C03451	C04554	C05279	C05472	C05640	C05849	C06199	C09824		
C03453	C04555	C05280	C05473	C05641	C05850	C06206	C09825		
C03460	C04640	C05284	C05474	C05642	C05852	C06207	C09880		
C03465	C04644	C05285	C05475	C05643	C05889	C06212	C09884		
C03479	C04677	C05290	C05476	C05647	C05890	C06213	C11131		
C03492	C04734	C05293	C05477	C05648	C05893	C06241	C11132		
C03508	C04751	C05294	C05478	C05651	C05894	C06452	C11133		
C03546	C04760	C05298	C05479	C05653	C05898	C06459	C11134		
C03557	C04778	C05299	C05480	C05655	C05899	C06548	C11135		
C03561	C04805	C05300	C05485	C05656	C05901	C06604	C11136		
C03564	C04823	C05302	C05487	C05657	C05921	C06606	C11356		
C03569	C04853	C05332	C05488	C05659	C05922	C06607	C11405		
C03582	C04874	C05335	C05489	C05660	C05923	C06608	C11407		
C03594	C04895	C05338	C05490	C05665	C05925	C06644	C11419		
C03680	C05100	C05345	C05497	C05668	C05931	C06645	C11421		
C03684	C05110	C05356	C05498	C05673	C05932	C06649	C11422		
C03691	C05111	C05378	C05499	C05674	C05933	C06650	C11425		
C03722	C05116	C05379	C05500	C05686	C05935	C06651	C11429		
C03758	C05118	C05381	C05501	C05688	C05936	C06674	C11431		
C03765	C05119	C05382	C05502	C05689	C05938	C06675	C11432		
C03771	C05122	C05394	C05503	C05691	C05946	C06676	C11433		
C03772	C05125	C05396	C05504	C05692	C05947	C06711			
C03785	C05127	C05399	C05512	C05695	C05951	C06714			
C03793	C05130	C05400	C05520	C05696	C05956	C06715			
C03794	C05135	C05401	C05527	C05697	C05959	C06749			
C03824	C05138	C05402	C05528	C05698	C05966	C07083			
C03838	C05139	C05403	C05539	C05699	C05983	C07084			
C03912	C05140	C05404	C05548	C05711	C05984	C07086			
C03917	C05141	C05418	C05552	C05766	C05985	C07096			
C03972	C05145	C05419	C05565	C05768	C05993	C07097			
C04006	C05172	C05437	C05576	C05775	C05998	C07098			
C04041	C05176	C05439	C05577	C05787	C05999	C07112			
C04043	C05200	C05444	C05578	C05791	C06000	C07113			
C04051	C05202	C05445	C05579	C05800	C06001	C07114			
C04063	C05235	C05446	C05580	C05801	C06002	C07118			
C04076	C05258	C05447	C05581	C05802	C06006	C07271			
C04079	C05259	C05448	C05582	C05803	C06008	C07715			
C04185	C05260	C05449	C05583	C05804	C06010	C07717			
C04244	C05262	C05450	C05584	C05805	C06017	C07718			
C04256	C05263	C05451	C05585	C05811	C06054	C07724			
C04257	C05264	C05452	C05587	C05812	C06055	C07725			
C04281	C05265	C05453	C05588	C05823	C06099	C07731			
C04282	C05266	C05454	C05589	C05827	C06114	C07733			
C04295	C05267	C05455	C05594	C05828	C06124	C07734			
C04392	C05271	C05461	C05634	C05838	C06148	C09332			
C04405	C05272	C05462	C05635	C05839	C06157	C09812			
C04409	C05273	C05467	C05636	C05841	C06178	C09813			
C04424	C05274	C05469	C05637	C05842	C06196	C09819			
C04468	C05275	C05470	C05638	C05843	C06197	C09820			
C04546	C05276	C05471	C05639	C05844	C06198	C09821			

SI-5 List of the 931 KEGGS identifiers corresponding to the human metabolites used in this study.