Virtual screening against highly charged active sites: Identifying substrates of alpha-beta barrel enzymes†

Chakrapani Kalyanaraman, Katarzyna Bernacki and Matthew P. Jacobson*

Department of Pharmaceutical Chemistry

University of California San Francisco

San Francisco, CA 94143-2240

* To whom the correspondence should be addressed. Telephone: (415) 514-9811. Fax: (415) 514-4260. E-mail: matt@cgl.ucsf.edu

Suggested running title: Virtual metabolite screening

†This work was supported by NSF Grant 0302445, start-up funds provided by HHMI Biomedical Research Support Program grant #5300246 to the UCSF School of Medicine, NIH grant GM071790, and by an NSF Postdoctoral Fellowship in Interdisciplinary Informatics to K.B.

‡Abbreviations: MR, mandelate racemase; GlucD, glucarate dehyratase; MLE, muconate lactonizing enzyme; MAL, methylaspartate ammonia lyase; AEE, alanine-glutamate epimerase; OSBS, ortho-succinyl benzoate synthase; KEGG, Kyoto encyclopedia of genes and genomics; PLOP, protein local optimization program.

Abstract

We have developed a virtual ligand screening method designed to help assign enzymatic function for alpha-beta barrel proteins. We dock a library of ~19,000 known metabolites against the active site and attempt to identify the relevant substrate based on predicted relative binding free energies. These energies are computed using a physics-based energy function based on an all-atom force field (OPLS-AA) and a Generalized Born implicit solvent model. We evaluate the ability of this method to identify the known substrates of several members of the enolase superfamily of enzymes, including both holo and apo structures (11 total). The active sites of these enzymes contain numerous charged groups (lysines, carboxylates, histidines, and one or more metal ions), and thus provide a challenge for most docking scoring functions, which treat electrostatics and solvation in a highly approximate manner. Using the physics-based scoring procedure, the known substrate is ranked within the top 6% of the database in all cases, and in 8 of 11 cases, within the top 1%. Moreover, the top-ranked ligands are strongly enriched in compounds with high chemical similarity to the substrate (e.g., different substitution patterns on a similar scaffold). These results suggest that our method can be used, in conjunction with other information including genomic context and known metabolic pathways, to suggest possible substrates or classes of substrates for experimental testing. More broadly, the physics-based scoring method performs well on highly charged binding sites, and is likely to be useful in inhibitor docking against polar binding sites as well. The method is fast (<1 minute per ligand), due largely to an efficient minimization algorithm based on the Truncated Newton method, and thus can be applied to thousands of ligands within a few hours on a small Linux cluster.

Computational ligand screening ("virtual screening" or "docking") is widely used in structure-based drug design projects to rapidly and inexpensively identify lead compounds (*1*-*3*). Here, we consider a different application of docking methods, namely to assist in the identification of possible substrates of an enzyme, when the function of the enzyme is unknown. We expect this capability to become increasingly important as the "production phase" of structural genomics efforts gets under way. These projects are expected to generate thousands of protein structures, with the ultimate goal of providing structural representatives of the majority of protein families. However, knowing the structure of a protein does not always uniquely or unambiguously suggest its function. Although the term "function" can encompass a broad range of meanings, here we refer specifically to the reaction(s) that an enzyme can catalyze *in vivo*.

The alpha-beta barrel enzymes, a subset of which is considered in this paper, pose a particular challenge for functional annotation. The basic alpha-beta barrel scaffold of eight parallel strands forming a barrel, flanked by eight helices, is known to catalyze a highly diverse set of reactions. The SCOP database (*4*) enumerates 26 superfamilies, and each of these is functionally diverse. For example, the enolase superfamily, which provides the focus for this work, likely contains over 1000 members based on currently available sequences; the 12 currently characterized functions, associated with nine experimentally solved structures, likely represent only a subset of the total functional diversity of the superfamily (*5*-*7*). Thus, for this large class of proteins, function must be assigned using information other than the known or predicted tertiary structure. For many alpha-beta barrel enzymes, the identity and spatial arrangement of the active site residues has been correlated with the function (and in some cases, with a proposed reaction mechanism), but it is not simple to extrapolate such knowledge to enzymes of unknown function.

The approach we are developing to address this problem can be described as "virtual metabolite screening". The central idea is to dock a database of metabolites (small molecules known or suspected to be involved in various enzymatic pathways *in vivo*) into the enzyme's active site, and then rank the ligands according to their estimated binding affinity. Binding affinity alone is not likely to be sufficient to uniquely identify an enzyme's substrate among a database of metabolites, because it does not describe the ability of the enzyme to catalyze a reaction on the substrate (k_{cat} would be much more difficult to estimate by computational means). In addition, the substrates of enzymes may not bind as tightly as many inhibitors, potentially making it difficult to distinguish true substrates from false positives. Nonetheless, we expect that relative binding affinity, if it can be estimated with sufficient accuracy, will provide a useful first filter for identifying possible substrates. A relatively short list of metabolites would then be subjected to further scrutiny for plausibility. For example, although the enolase superfamily enzymes catalyze a wide variety of overall reactions, they all share the common reaction initiation step, that is to abstract a proton α- to the carboxylate functional group that co-ordinates to the metal ion (*8*). Finally, an even shorter list can be subjected to experimental testing.

Current docking programs use a variety of scoring functions to estimate binding affinity, including force-field based, knowledge based, and empirical scoring functions (*9*). Several studies have been carried out in the past to assess the quality of these scoring functions (*10, 11*). For efficiency reasons, energy terms including van der Waals interactions, electrostatics, and solvation are approximated in these scoring functions. These approximations contribute to the large numbers of false positives observed in most docking calculations.

The alpha-beta barrel enzymes we consider pose a challenge for docking scoring functions because the active sites contain many charged and polar groups. The substrate binding sites of the enolase superfamily members typically contain 2 lysines, 3 or more carboxylate groups (from Asp/Glu residues), and at least one metal ion, in addition to histidines and other polar residues. Thus, these binding sites are unusually polar, much more so than a large majority of binding sites targeted for rational drug design. The binding affinity of a ligand in such an active site involves a complex interplay of strong electrostatics and solvation. The binding of the ligand necessarily reduces or eliminates the solvent accessibility of many of the charged and polar groups (desolvation), resulting in a large reduction of the solvation free energy, which must be compensated by favorable electrostatic and other interactions.

Most docking scoring functions, which treat electrostatics and solvation only very approximately, would seem poorly suited for screening against such highly polar active sites. However, our goal here is not to benchmark docking methods for virtual screening against alpha-beta barrel enzymes, but rather to develop a robust method that can assist in assigning function to the thousands of such enzymes for which no function is known, and cannot be reliably assigned by other methods. We have done so by combining docking calculations using the program Glide (*12-14*) with a physics-based rescoring procedure of our own design. The rescoring procedure uses an all-atom force field (OPLS-AA) (*15, 16*) and a Generalized Born implicit solvent model (*17*). Thus, the energy function used is similar to that used in sophisticated approaches such as free energy perturbation (FEP) and molecular-mechanics Poisson-Boltzmann surface area (MM-PBSA) (*18*). However, these methods are computationally expensive and thus are generally

applied only to a small number of ligands. Our method provides only crude estimates of entropic contributions to binding free energies, relies on the docking program to identify the correct binding pose, and uses only one configuration to estimate the binding affinity. Conceptually similar efforts have been described in the literature (*19-23*). One distinguishing feature of the method described here is its speed, only ~45 seconds per ligand on a single processor, which allows us to rescore thousands of ligands within a few hours on a small Linux cluster. The speed is due largely to the use of a fast minimization algorithm (as opposed to molecular dynamics or Monte Carlo used in FEP and MM-PBSA), based on the Truncated Newton method and optimized for use with Generalized Born solvent models.

The key conclusion of this study is that the physics-based rescoring method makes it possible, with a few exceptions, to identify the substrate for the enzymes we study within the top 1% of the database hits (within the top 0.1% in many cases). This conclusion holds for apo as well as *holo* structures, which is encouraging for future work aimed at identifying the function of alpha-beta barrel enzymes whose function is currently unknown. It is also encouraging to note that the rescoring procedure is quite robust. In cases where the docking program itself yields good results, they are preserved by the rescoring. In the majority of cases, however, the results from the docking program are improved significantly, and often dramatically, by the rescoring procedure.

METHODS

The computational approach involves two steps. First, we dock a metabolite ligand library against the enzyme active site using the docking program Glide (*12-14*). We then rescore the top 25% of the docking results using the physics-based method we have developed. In this section, we describe the ligand library preparation, docking and rescoring protocols.

Metabolite Ligand Library

The Kyoto Encyclopedia of Genes and Genomics (KEGG) database contains metabolite ligands, toxins, inhibitors and pollutants (*24*). We obtained 10,004 compounds from the KEGG database. After filtering out entries with unspecified chemical groups (listed as "R"), polymers, and monoatomic ions, 8,182 compounds remained. Many of these lacked proper chiral center definitions (i.e., chiral centers listed as "undefined"). For this reason, we used the Daylight (*25*) software *chiralify* to enumerate up to 16 chiral forms for each ligand. The resultant database contained 19,007 compounds. The 3D structures of these compounds were obtained using the ZINC database pipepline (Shoichet laboratory, UCSF), which use the OpenEye (*26*) software *omicron* for 1D to 3D conversion.

Docking

Glide (*12-14*) is used for flexible docking. Glide uses an expanded version of the ChemScore (*27*) empirical scoring function. The protein receptor was prepared using standard Glide procedures, i.e., by adding hydrogens and specifying correct charge and atom type for the metal ions. Any charged residues farther than 20 Å from the binding site were neutralized. For holo structures, we used the *impref* utility script to perform a restricted minimization of residues

around the ligand atom such that the heavy atoms do not move beyond 0.3 Å RMSD from their crystallographic positions. This minimization facilitates orienting the receptor heavy atoms so that they can make proper hydrogen bond and van der Waals interactions with ligand atoms. For holo structures, the binding pocket is identified using the bound ligand, and for apo structures, we specify the active site residues based on information gathered from the literature. After completing the receptor preparation step, the ligands were docked using the "standard precision" Glide algorithm.

Rescoring

The top 25% of the ligands from the docking results were subjected to rescoring; we considered only the best scoring pose for each ligand. In this work, we have found that Glide's top ranking binding pose reproduced the known crystallographic poses quite well in all but one of the holo cases, i.e., within 1 Å root mean square deviation (RMSD).

Protein Local Optimization Program (PLOP) (*28, 29*), a software package developed and maintained in our group, was used for the physics-based rescoring procedure. For each ligand, the protein-ligand complex ($E_{Lig\text{-}Prot}$), the free protein (E_{prot}), and the free ligand (E_{Lig}) were all subjected to energy minimization in implicit solvent (Generalized Born). Note that the energy of the free protein (E_{prot}) is constant and thus does not affect our estimates of *relative* binding affinities. For the minimization of the protein-ligand complex, the protein atoms were held fixed, i.e., only the ligand was allowed to minimize. This choice yielded good results, and improved computational efficiency. However, the method can be trivially modified to allow the receptor to relax as well, thus enabling of small amounts of "induced fit".

The isolated ligand and the protein energies are subtracted from the energy of the ligand-protein complex energy to calculate the relative binding energy (E_{RBE})

$$E_{RBE} = E_{Lig\text{-}Prot} - E_{Prot} - E_{Lig} + 0.5925 \cdot N_{Rot.\ Bond}$$

where $N_{Rot.\ Bond}$ is the number of rotatable bonds in the ligand. The last term in the above energy expression is a penalty term, kT per rotatable bond, to account crudely for the loss of ligand internal entropy due to binding. A similar term is used in Glide (*12-14*) and other docking programs.

We used this relative binding energy to rank order the ligands. Note that both ligand and receptor desolvation are accounted for in this procedure, and the minimization of the ligand outside the receptor provides an estimate of the internal strain energy. However, since entropy terms are not properly accounted for in our formalism, the relative binding energy computed here cannot be compared to the experimental (absolute) binding free energy.

As mentioned above, PLOP was used for the energy minimizations. PLOP uses the OPLS-AA all atom force field with the Surface Generalized Born implicit solvent model (*28, 29*). The minimization algorithm is based upon the truncated-Newton (TN) method, specifically the TNPACK method of Schlick and co-workers (*30-32*). The algorithm has been accelerated by approximately a factor of 3 by applying simple multi-scale ideas, in which the short-range forces are updated more frequently than the long-range ones (M. P. Jacobson, K. Zhu, and R. A. Friesner, in preparation). In addition, the algorithm has been optimized for minimizations in generalized Born implicit solvent, which requires only 2-3 times greater computational expense

compared to minimization in vacuum. On average, it took ~15 s to minimize the ligand in the protein in implicit solvent. Therefore, thousands of ligands can be refined in a matter of few hours on a small Linux cluster, which is roughly comparable to the time required for the docking.

Chemical Similarity

We have used the program QikSim (*33*) to obtain quantitative measures of chemical similarity. This program computes the Tanimoto coefficient based on a series of descriptors that include both the numbers of common functional groups present, and "whole molecule" descriptors such as dipole moment and surface area. Compounds that are chemically similar to the substrate or product will have a Tanimoto coefficient near 1 and those that are dissimilar will have 0 (*34*). In the plots of chemical similarity vs. ligand rank, the data shown have been smoothed to emphasize the trends.

RESULTS

Here we analyze the results in detail for 11 representative structures of enzymes in the enolase superfamily, 6 of them holo and 5 apo. We focus on three distinct types of analysis:

1. In cases where a holo structure is available, we compare the predicted ligand poses with the co-crystallized ligand. In many cases, the co-crystallized ligand is either the product or an inhibitor, and is generally chemically very similar, but not identical to, the substrate. When a holo structure is not available, we can still analyze whether the predicted pose is qualitatively consistent with the proposed reaction mechanism. Plots of the substrate poses are shown in Figure 1.

2. We determine the enrichment of the known substrate and product relative to the other ~19,000 ligands in our preparation of the KEGG metabolite database. That is, we analyze whether it would be possible to use virtual metabolite screening to help identify the substrate and/or product, and thus infer a likely reaction carried out by the enzyme, if it were not known. In the case of holo structures, this establishes the suitability of the physics-based rescoring procedure for distinguishing ligands likely to bind to the active site from a large database of metabolites. Virtual metabolite screening against apo structures is more representative of how the method would be used in practice to help assign function to uncharacterized enzymes. Table 1 summarizes the ranks of the known substrates and products after docking with Glide and after rescoring the results with the physics-based energy function.
3. Although the results after rescoring typically rank the known substrate very highly relative to the other metabolites in the KEGG database (often within the top few tenths of 1% of the database), typically a few tens of ligands rank higher. Are these "false positives" entirely spurious, or do they also provide useful clues concerning the possible substrate? We have found that, after applying the rescoring procedure, the top ranked compounds, on average, have high chemical similarity to the known substrate, i.e., many of them share a common sub-structure with the substrate. We use the program QikSim (*33*) to obtain quantitative measures of chemical similarity, specifically the Tanimoto coefficient. The Tanimoto coefficients are computed for each ligand with respect to the known substrate and plotted against the ligand's rank before and after rescoring. The results, after applying a smoothing function to reduce the noise, are shown in Figure 2, and demonstrate that the results after rescoring show good enrichment of ligands that are

chemically similar to the known substrate in the top few percent of the database. This result suggests that the top-ranked false-positives can provide useful clues concerning the possible substrate/function. A few of these highly ranked compounds might even be turned over by the enzyme, i.e., that the top-ranked compounds may suggest possible "promiscuous" reactions that could be tested experimentally. The chemical similarity plots prior to rescoring (docking only) are more mixed, with good correlation between chemical similarity and rank in some cases, and anti-correlation or no correlation in others. The rescoring procedure also improves the enrichment of compounds with the "proton alpha to a carboxylate" substructure required for the half-reaction (formation of an enolate intermediate) that defines the enolase superfamily (Figure 3).

Overview of enolase superfamily

Members of the enolase superfamily typically have two domains, an N-terminal capping domain and a C-terminal $(\alpha|\beta)_8$ – barrel domain. The binding pocket is located in the barrel domain, while residues in the specificity determining N-terminal domain also contact the substrate. Among the enolase superfamily enzymes, residues in the binding pocket are well conserved (*8*). They all require a doubly charged metal ion in the center, except enolase subgroup members, which require two doubly charged metal ions. Depending on the residues that coordinate the metal ion and the residues that participate in the proton abstraction reaction, enolase superfamily members are classified into mandelate racemase (MR), muconate lactonizing enzyme (MLE) and enolase subgroups (*8*). Several families exist in a given subgroup. Enzymes in the same family tend to have higher sequence identity than those in different families. Different family members within a subgroup usually have 10-25% sequence identity. We have selected 11 members from

the enolase superfamily consisting of both holo (with co-crystallized ligand) and apo (without a co-crystallized ligand) structures.

Mandelate racemase (MR)

Members of the MR family catalyze the racemization of mandelic acid (*35, 36*). Lys166 and His297 have been identified as the general bases responsible for abstracting α-proton from the S-Mandelic and R-Mandelic acid respectively (*8*). The enolate anion thus produced is stabilized by the metal ion and Glu317 (*37*).

We have tested our algorithm on both holo and apo structures. The holo structure (PDB ID 1mdr, organism *P. putida*, 2.1 Å resolution) is co-crystallized with an inhibitor, S-atrolactic acid (*36*). The reactive α-proton is replaced by a methyl group in the inhibitor. The presence of the bulky methyl group pushes the side chain of the catalytic Lys166 slightly away from the position we expect that it would occupy with the substrate bound. His297 forms a hydrogen bond with Asp270, and for this reason we protonated His297 at both the δ and the ε nitrogen in all calculations. In addition, Glu317, which provides hydrogen bonding support to the enolate anion intermediate, is protonated during all calculations.

S- and R-mandelate ranked 1228 and 1719 respectively using Glide, corresponding to the top 6.5% and 9.0% of the total database. After rescoring the top 25% of the database (4996 ligands), the ranks of S- and R-mandelate improved to 77 (0.41%) and 140 (0.74%) respectively. Similarly, the co-crystallized inhibitor, S-atrolactic acid, ranked 675 (3.6%) after docking, and enriched to 36 (0.19%) after rescoring. The binding pose of S-mandelate, after rescoring, is

shown in Figure 1a along with the co-crystallized inhibitor. The heavy atom RMSD between the docked pose and the co-crystallized ligand's pose is 0.47 Å. The binding pose is consistent with the reaction mechanism, in that the proton on the α-carbon is directly facing the catalytic residue Lys166 for the proton abstraction to occur. The oxygen atoms that are co-ordinated to the metal ion are however slightly different from the crystal structure. In the crystal structure, a caboxylate oxygen and the oxygen from the hydroxyl group of the ligand are co-ordinated to the metal ion, whereas in our docked pose, both carboxylate oxygens of the ligand are co-ordinated to the metal ion.

We have also docked to the apo structure of MR from *P. putida* (PDB ID 2mnr, 1.9 Å resolution) (*35*). The binding pocket was defined based on the residues that interact with the ligand in the holo structure. S- and R-mandelate ranked 2299 (12.1%) and 1758 (9.3%), respectively, after the docking phase. After rescoring, ranks of these ligands improved to 122 (0.65%) and 732 (3.9%) respectively. It is interesting to note that one stereoisomer is enriched significantly over the other. The all atom force field with implicit solvation based rescoring enriches the known substrate in both holo and apo structures to be within the top 1% of the database.

The chemical similarity plots in Figures 2a and 2b reveal that the rescoring procedure enriches not only the substrate but also compounds that are similar to the substrate, for both the holo and apo structures. Understanding how compounds chemically similar to the substrate bind in the binding pocket may provide clues to determine if these enzymes can promiscuously catalyze functions of other enzymes.

Glucarate Dehydratase (GlucD)

GlucD catalyze the conversion of D-glucarate or L-idarate to 5-keto-4-deoxy-(D)-glucarate (KDG) (*38*). Lys207 and His339 are responsible for abstracting the α-proton from L-idarate and D-glucarate respectively (*5, 38*). Three histidine residues are present in the binding pocket: His32, His339 and His368. His339 makes a salt-bridge interaction with Asp313. The other two histidines are also in close proximity to carboxylate residues. Therefore, we protonated the δ and the ε nitrogen of all three histidine residues during the docking and rescoring calculations. The protein structure considered in this family had a competitive inhibitor, 4-deoxy-D-glucarate, bound in the active site (PDB ID 1ecq, 2.0 Å resolution) (*38*). In this case, the docking program performed excellently in the absence of rescoring. Specifically, the ranks following Glide docking are 5 (0.03%) for D-glucarate, 40 (0.21%) for L-idarate and 8 (0.04%) for KDG. After rescoring, their ranks improved slightly to 4 (0.02%), 15 (0.08%), and 1 (0.005%) respectively. The predicted pose for the D-glucarate, shown in Figure 1c, is in good agreement with the proposed mechanism, whereby the α-proton is positioned close to the Lys207 for abstraction to form the enolate anion intermediate.

Consistent with the similarity of the ranks of the known ligands before/after rescoring, the chemical similarity analysis, shown in Figure 2c, reveals that the enrichment of ligands similar to the known substrates does not change significantly upon rescoring. This is not the case with the majority of the other cases. We have not identified any clear reason why the Glide scoring works much better in this case than in the others.

Muconate Lactonizing Enzyme-I (MLE-I)

Members of this family perform the cycloizomerization of cis-cis muconate to muconolactone (*39*). No holo crystal structure is available in this family. We have studied a high-resolution apo crystal structure (PDB ID 1muc, 1.85 Å resolution) (*40*). The ligand binding pocket is defined based on the residues that participate in the reaction, Lys167, Lys169, Lys273 and Glu327, as well as the residues that co-ordinate to the metal ion, Asp198, Glu224 and Asp249 (*8*). After docking using Glide, the substrate was ranked 3679 (19.4%) and the product was ranked 1933 (10.2%). After rescoring, the substrate and the product ranks improved to 1020 (5.4%) and 56 (0.29%), respectively. We believe that the relatively poor performance of the substrate relative to the product may be due to the docking algorithm generating an incorrect pose for the substrate. Nonetheless, the strong enrichment of the product, and ligands that are chemically similar to it (Figure 2d), would arguably provide important clues to the function of the enzyme, if it were not known. The binding pose of the product, muconolactone, along with active side residues is shown in Figure 1d.

Methyl Aspartate Ammonia Lyase (MAL)

MAL catalyzes the reversible β-elimination of ammonia from L-threo-(2S,3S)-3-methyl aspartic acid to yield mesaconic acid (*8, 41*). This enzyme catalyzes the reaction with the 3R substrate 38 times slower than with the 3S substrate (*8, 42*). Thus, MAL is a stereo selective and not a stereo specific enzyme. Lys331 acts as the general base to abstract the proton α to the carboxylate group that coordinates the metal ion. We have performed virtual metabolite screening with both a holo structure, in which the substrate was bound in the active site (PDB ID 1kkr, 2.1 Å resolution) (*41*), and an apo (PDB ID 1kko, 1.33 Å resolution) (*41*) structure.

Six side chains were missing in the structure of the holo enzyme; none of these residues is located near the binding pocket, and we did not build them before docking, but did reconstruct them prior to rescoring using PLOP. Glide ranked the substrate, L-threo-(2S,3S)-3-methyl aspartic acid, for the holo structure as 1723 (9.1%), and the rescoring procedure brought the substrate to 198 (1.0%).

In the apo enzyme structure, the metal ion, which is required for enzymatic function, was missing. We therefore devised a strategy for placing the ion in the binding site, based on its position in other enolase superfamily member. Specifically, we used a structure-based alignment algorithm (combinatorial extension) (*43*) to align the MAL apo enzyme to the MLE-I apo enzyme (PDB ID 1muc), and then copied the metal ion coordinates from the PDB. Although we could have used the MAL holo enzyme structure to obtain the ion's coordinates, this would not represent a generally applicable strategy, i.e., in the context of attempting to assign function for a functionally uncharacterized enzyme with no holo structure available. The choice of MLE-I for obtaining the metal ion coordinates is arbitrary; the position of the metal ion is extremely well conserved in structural alignments of enolase superfamily members, and thus the absence of a metal ion in an apo crystal structure is not a serious impediment to performing virtual metabolite screening.

After adding the metal ion, the protein was prepared for docking using the procedure outlined in Methods. Glide ranked the substrate 1119 (5.9%). The rescoring procedure improved the rank to 9 (0.05%). That is, the results for the apo structure, after rescoring, are actually significantly better than the results for the holo structure in this case.

The substrate binding poses for the holo and the apo structures are shown in Figures 1e and 1f. The co-crystallized substrate is also shown in Figure 1e. The RMSD difference between the co-crystallized substrate and the binding pose predicted by docking and rescoring calculations is 1.16 Å. The proton α-to the carboxylate is still present near the base Lys331. A similar binding pose is also observed for the apo structure as shown in Figure 1f.

The chemical similarity plots for holo and apo structures are shown in Figures 2e and 2f respectively. In case of the holo structure, both Glide and rescoring show similar trends. However, for the apo structure, the results after Glide docking do not show any correlation between chemical similarity and ligand rank (or perhaps a slight anti-correlation). However, the rescoring procedure enriches substrate-like compounds significantly.

Alanine-Glutamate Epimerase (AEE)

The L-Ala-D/L-Glu epimerases belong to the muconate lactonizing enzyme subgroup and catalyze the epimerization of a component of the murein peptide substrate (*44, 45*). Recently, a holo structure of the complex of L-Ala-L-Glu with the AE epimerase from Bacillus subtilis has been published (*46*). We have docked to the holo structure (PDB ID 1tkk), consisting of 359 residues. Residues Lys160, Lys162, Asp191, Glu219, Asp244, Lys268, and Asp321 define the active site. The carboxylate oxygens of Asp191, Glu219, and Asp244 form close interactions with the Mg^{2+} ion. The negatively charged side chains of Asp321 and Asp323 form hydrogen bonds with the amino group on the Ala residue of the substrate (*45*). The δ and ε

nitrogens of His223 and His309 were protonated, because they form hydrogen bonds with the carboxylate group of Asp225 and the carbonyl group of Leu352, respectively.

Most dipeptides, including Ala/Glu, are not present in the KEGG database. For this case, we added to the KEGG database all 400 dipeptides formed from the standard L-amino acids. In the holo case, the docking program ranked the L-Ala-L-Glu substrate at 174 (0.92%) and the L-Ala-D-Glu product at 116 (0.61%). After rescoring the top 25% of the database, the rank of the substrate and product improved to 89 (0.47%) and 80 (0.42%), respectively. The docked pose superimposed on the holo structure is shown in Figure 1g. The docked ligand conformation agrees quite well with the co-crystalized ligand.

Two apo crystal structures of the L-Ala-D/L-Glu epimerases are available, from *E. coli* and *Bacillus subtilis*. We have used the structure of the epimerase YkfB from *Bacillus subtilis* (PDB ID 1jpm, Chain B, 2.25 Å resolution) consisting of 366 amino acids. The structure contains several disordered residues (Lys20, Lys165, Lys207, Lys349, and Leu359) modeled as Ala residues. We have used PLOP to predict the conformations of these side chains prior to docking. The His residues were treated as in the holo structure. The docking program ranked the substrate L-Ala-L-Glu at 3232 (17.0%) and the product L-Ala-D-Glu at 2519 (13.2%). After rescoring the top 25% of the database, the rank of the substrate and product improved to 934 (4.9%) and 1281 (6.7%) respectively. The docked conformation of the substrate is shown in Figure 1h.

The rescored rank for the epimerase is relatively poor compared to the results for the holo structure and the other apo enzymes that we have studied. This relatively low rank is due to a

major structural change that occurs upon substrate binding. Loop residues 14-30, which are believed to be responsible for substrate specificity (*44, 45*), close around the substrate upon binding. The open conformation of the apo loop and the closed conformation of the holo loop are shown in Figures 4a and 4b, respectively. When the apo and holo structures are superimposed, the overall RMSD is around 0.3 Å with this only major displacement occurring in the above mentioned loop. In the holo structure, loop residue Arg24 forms hydrogen bonds with the carboxylate group of the Glu side chain in the substrate, and residue Lys162 interacts with carbonyl oxygen of Ala and carboxylate oxygens of Glu in the substrate as well as Asn193. In the apo structure, the loop is displaced by approximately 12 Å, eliminating any interactions with the substrate. Also, the side chain of Lys162 is rotated by ~6 Å away from the active site. These conformational changes appear to lead to the incorrect docked pose in the apo structure. Thus, significant structural changes such as loop movement close to the active site could prove to be a major challenge in determining substrates and functionality using apo structures. We are in the process of using side chain and loop prediction algorithms in addition to the docking/rescoring framework presented in the paper to address these issues.

.

The chemical similarity plots for the holo and apo epimerase structures are shown in Figures 2g and 2h, respectively. Again, the trend between chemical similarity to the known substrate and ligand rank is more favorable after applying the rescoring protocol, which strongly enriches compounds similar to the substrate.

Ortho-Succinyl Benzoate Synthase (OSBS)

OSBS catalyzes the syn-elimination of H_2O from the substrate 2-succinyl-6-hydroxy-2, 4 – cyclohexadiene-1-carboxylate (1R, 6R) to yield 2-succinyl benzoate (*47*, *48*). Residue Lys133 abstracts the α-proton from the substrate, and the resulting enolate ion intermediate is stabilized by both the metal ion and Lys235 (*48*). We have performed virtual metabolite screening on the product bound holo structure of OSBS (PDB ID 1fhv, 1.77 Å resolution) (*47*) and an apo structure (PDB ID 1fhu, 1.65 Å resolution) (*47*).

For the holo structure, the substrate ranked 1166 (6.1%) and the product ranked 926 (4.9%) after the docking phase. After rescoring, the ranks of the substrate and product improved to 39 (0.21%) and 8 (0.04%) respectively. Interestingly, we also docked a substrate analog with an 'S' configuration instead of 'R' at position 1. Its rank improved from 996 to 102 after rescoring. If this compound binds to the enzyme, neither the α-proton abstraction nor the syn-elimination of water can take place. The rescoring procedure clearly brings the true substrate (1R, 6R) ahead of the diastereomer (1S, 6R).

The metal ion required for catalysis, $Mg^{2+,}$ was missing in the apo structure. Using a procedure identical to that used for the apo MAL structure, we structurally aligned the apo OSBS structure to the apo structure of MLE-I, and copied the metal ion coordinates. After the usual receptor preparation steps, Glide predicted ranks for the substrate and product to be 1871 (9.8%) and 3492 (18.4%) respectively. The rescoring protocol dramatically enriched the product to 57 (0.29%), and the substrate to 993 (5.2%).

The predicted poses of the substrate OSB are shown for the holo and apo OSBS structures in Figures 1i and 1j respectively. In the case of the holo structure, we have also shown the co-crystallized product in Figure 1i. The heavy atom RMSD is 0.64 Å. In addition, the reactive α-proton in both holo and apo structures is positioned near the base Lys133.

The chemical similarity plots for the holo and the apo structures are shown in Figures 2i and 2j. In both cases, there is a much better correlation between chemical similarity to the substrate and rank after rescoring.

Enolase

Members of the enolase subfamily catalyze the reversible dehydration of 2-phospho-D-glycerate to form phosphoenolpyruvate in the glycolytic pathway (*49*). [Note that "enolase" refers to both to the entire superfamily, as well as a subset of the enzymes in that superfamily.] Unlike the MR and MLE subfamily members, the enolase subfamily members bind two metal ions in the active site, which are coordinated by carboxylate groups from the protein, as well as the carboxylate and the phosphate group of the ligand in the holo structure. We have performed virtual metabolite screening using a structure with the competitive inhibitor, phosphonoacetohydroxamate, bound in it (PDB ID 1ebg, 2.1 Å resolution) (*49*). The substrate and the product were ranked 8 (0.04%) and 11 (0.06%) by Glide. After rescoring the top 25% of the docked database, rank of the substrate and the product changed to 30 (0.16%) and 17 (0.09%) respectively, a very slight decrease. Nonetheless, it is encouraging that, in cases where the docking program ranks the known substrate very highly, the rescoring procedure retains very high ranks. The binding pose of the substrate is compared to the co-crystallized inhibitor in

Figure 1k. The heavy atom RMSD is 0.6 Å. The α-proton is favorably located near the base Lys345. The chemical similarity plot, shown in Figure 2k, reveals that the results both before and after rescoring greatly enrich compounds similar to the substrate in the top few percent of the ranked list.

DISCUSSION

The first key conclusion of this study is that the docking algorithm generally performs well in predicting the pose of the substrate and product. The key criteria for assessing the poses are RMSD, when a holo structure is available, and the positioning of the alpha proton for abstraction by the relevant catalytic residue. Calculation of the RMSD for substrates, when it is possible at all, is slightly complicated by the fact that the co-crystallized ligand is generally either the product of the reaction or an inhibitor. Nonetheless, in all cases where a holo structure is available (6 total), the RMSD over common atoms is ~1 Å or better, and the α-proton is appropriately positioned for abstraction. Although no holo structure is available for MLE-I, the predicted pose of the substrate is almost certainly incorrect based on the known chemistry (the two carboxylate groups in the predicted pose are positioned pointing away from each other, which would prevent the cyclization reaction from occurring). It is interesting to note that the substrate in this case had the worst rank out of the 11 cases, 19% after docking and 5% after rescoring. We suspect that the incorrect pose is related to conformational changes in the receptor upon substrate binding; when we performed docking with reduced van der Waals radii, we could create poses that seemed consistent with the proposed mechanism, albeit with poor energies. An incorrect pose also seems to limit enrichment of the known substrate using the apo OSBS structure, as well.

The overall high quality of the predicted poses is important for two reasons. First, as suggested by the relatively poor result with MLE-I, the rescoring procedure, as currently implemented, relies on the docking algorithm to predict the pose with reasonable accuracy. The ligand minimization performed during the rescoring permits optimization of the hydrogen bonding interactions and other relatively small (but energetically important) conformational changes, but does not result in qualitative changes in the pose. Second, the predicted substrate binding poses, if sufficiently accurate, could be used to help form hypotheses concerning the enzymatic reactions carried out by an enzyme (e.g., what might the product or mechanism be?), given knowledge of or a predicted substrate. All members of the enolase superfamily extract a proton alpha to a carboxylate, and this knowledge can be used to help further screen the binding hits, after rescoring.

The second key conclusion is that the physics-based rescoring procedure is critical to the robustness of our method, and permits the known substrates (and products) to be ranked very highly out of the 20,000 metabolites in our version of the KEGG ligands database. In 8 out of 11 total cases, the rescoring procedure succeeded in ranking the known substrate in the top 1% of the ranked list, in contrast to only 3 out of 11 prior to the rescoring. The cases where the docking scoring function performed well retained very high ranks after rescoring.

The worst result, MLE-I, ranked the known substrate at ~5% (the docking algorithm alone ranked the known substrates in the top 5% in only 3 out of 11 cases). Although this is hardly a disastrous result, it is worth noting that the product ranked very highly in this case, within the top

0.5% (a similar result was found for the apo structure of OSBS). In general, we found that the known products ranked highly, and it may be possible to use potential products in the hit lists to create hypotheses about the possible reactions carried out by the enzymes, although this is not as straightforward as simply prioritizing possible substrates for testing.

Perhaps unsurprisingly, the results for the holo structures are better on average than those for the apo structures. The results for the holo structures primarily measure the ability of our scoring function to distinguish known substrates from other metabolites, which are assumed not to be substrates (although a small number could in principle represent substrates in promiscuous reactions). The physics-based rescoring performs excellently in this test, with the known substrate in all six cases ranking in the top 1%. It is worth pointing out that this is not simply a "redocking" exercise but rather more akin to "cross-docking", because the co-crystallized ligands are, in all but one case, not substrate but either inhibitor or product.

The apo cases are more relevant to our ultimate goal of using virtual metabolite screening to assist in assigning function to alpha-beta barrel enzymes. The results, although not as strong as those for the holo structures, are generally encouraging. In one case, MAL, the results for the apo structure are better than those for the holo structure. In all cases, the substrate ranks within the top ~5% of the ranked list. Thus, at a minimum, we should be able to eliminate ~95% of the metabolites from the database based on virtual screening, without further improvements in the methods. This may be adequate for some purposes; other available information (including operon context, sequence-based clustering, chemical similarity analysis as discussed below, and the half-reaction associated with the superfamily) can likely be used to reduce the probable

substrates further. The worst case is MLE-I, which is discussed above. We reiterate that the product ranked highly in this case (and the OSBS apo case), and relatively poor enrichment of the substrate is probably due to an incorrect pose. The second worst case is AEE. In this case and for OSBS, we hypothesize that a large loop motion associated with ligand binding, which is not currently modeled by our method, precludes better enrichment. We believe that it will be possible to further improve results on these cases by sampling the protein receptor during the rescoring stage, rather than leaving it rigid. Although this is beyond the scope of this work, we provide further comments on this possibility, currently under testing, in the Conclusions.

A third major conclusion is that, after the rescoring procedure, the top-ranked compounds tend to be chemically similar to the known substrate, as shown in Figure 2 (this is not always the case with the standard-precision Glide scoring function). More broadly, a strong correlation exists between chemical similarity and ligand rank after rescoring. For this reason, we believe that the top ranked ligands, other than the substrate, provide useful qualitative information about the likely chemical nature of the substrate (e.g., size and functional groups likely to be present). A few of the top-ranked ligands may even prove to be "promiscuous" substrates upon experimental testing.

Finally, the rescoring procedure appears to be capable of capturing selectivity. The known substrates of the 7 enzymes considered here all contain carboxylate groups, as well as a proton at the carbon α to the carboxylate, which is abstracted in the half-reaction that defines the enolase superfamily. Compounds with this substructure are strongly enriched in the top few percent of the ranked ligand list, especially after rescoring, as shown in Figure 3. However, the method

also shows the ability to identify the correct ligand containing this substructure from others. In Table 2, we have gathered the ranks for each known substrate obtained after docking and rescoring using each of the enzyme structures. The columns of the table make it possible to assess whether a given ligand scores better against the enzyme for which it is the known substrate than against the other enzymes. The rows make it possible to assess whether the known substrate for a particular enzyme outranks the known substrates for other enzymes. By both criteria, the results after rescoring show strong evidence of capturing selectivity, i.e., the right ligand for the right enzyme. The only problematic case is the apo MLE enzyme. Although the known substrate scores better against MLE than against any of the other enzymes, the D-glucarate and S-mandelate ligands outscore it. The results before rescoring (i.e., using the docking scoring function) show very little ability to capture selectivity, with GlucD and enolase being the only exceptions.

CONCLUSION

We have developed a physics-based method for rescoring protein-ligand complexes generated by a docking program, and applied it to virtual metabolite screening against a diverse set of alpha-beta barrel enzymes in the enolase superfamily, which have highly charged binding sites. We conclude by briefly commenting on the strengths of, and possible improvements to, our approach.

In general, the rescoring method appears to be highly robust, improving the rank of the known substrates significantly in a large majority of cases; the only exceptions are cases where the docking program performs excellently to begin with. We attribute this success to the treatment of electrostatics and solvation in our energy function, which consists of the OPLS-AA force field and a Generalized Born implicit solvent model. Thus, the rescoring method accounts for desolvation of both the protein and ligand upon binding, which would be very difficult to account for in grid-based scoring functions used in high-throughput docking. We believe this to be critical for studying the enolase and other alpha-beta barrel enzymes, in which the active sites generally contain a large number of charged groups (and the substrates are frequently charged as well). Initial tests of our method on virtual inhibitor screening against polar binding sites have also demonstrated significant improvements in enrichment (N. Huang, C. Kalyanaraman, J. Irwin, B. K. Shoichet, and M. P. Jacobson, in preparation).

The other major strength of the method described here is its speed. The average computational cost of rescoring a protein-ligand complex in this work was ~45 s, on a recent-generation single processor PC. Thus, tens of thousands of complexes can be rescored on a small cluster with relatively modest computational expense, in contrast to more sophisticated but expensive physics-based methods such as MM-PBSA and FEP. The speed of the method is made possible by a highly efficient minimization algorithm, based on the Truncated Newton method, in Generalized Born solvent. In fact, the minimization itself requires only ~15 s on average, with the remaining time associated with loading the protein, assigning parameters, etc. Further algorithmic optimization will reduce this computational overhead. This speed enables our

method to be used with large ligand libraries, such as those used in most virtual inhibitor screening applications.

One requirement for the success of our method is correct assignment of protonation states on both the protein receptor and ligands. In this work, we manually assigned protonation states to histidines in the binding site, based on their hydrogen bonding partners, but we can envision automated assignment of protonation states, e.g., using algorithms based on continuum electrostatics. In tests where we set the histidine protonation states incorrectly, the results of the rescoring were almost invariably worse, sometimes dramatically (data not shown). This is not surprising, because we include full electrostatics in the rescoring.

One major limitation of our method is the treatment of entropic losses associated with ligand binding. We crudely account for the loss of internal ligand entropy by using a simple penalty based on the number of rotatable bonds. Translational and rotational entropy losses are not accounted for at all. Although we attempt to reproduce only relative and not absolute binding free energies, we nonetheless expect that improved treatment of entropic losses would improve the enrichment of binders by our rescoring method.

Finally, all rescoring results presented here treated the receptor as entirely rigid. Relaxing this constraint could potentially improve results in cases where nontrivial conformational changes occur upon ligand binding. The simplest approximation would be to simply allow residues in the binding site to minimize along with the ligand (which requires only modest increases in computational expense); early results suggest that this strategy can improve results on docking to

apo structures. More elaborate rescoring methods can include, e.g., rotamer searches for side chains in the binding site, to deal with larger conformational changes.

Acknowledgments

We thank Patsy Babbitt (UCSF) and John Gerlt (UIUC) for introducing us to this problem and helping to guide our work; Brian Shoichet, John Irwin, Niu Huang, Elaine Meng, Brian Tuch and Robert Rizzo (UCSF) for many helpful discussions and critical technical assistance with docking; and Schrödinger, Inc. for use of and assistance with Glide and QikProp. MPJ is a member of the Scientific Advisory Board of Schrödinger, Inc.

References

1. Shoichet, B. K., McGovern, S. L., Wei, B., and Irwin, J. J. (2002) Lead discovery using molecular docking, *Curr Opin Chem Biol 6*, 439-46.
2. Brooijmans, N., and Kuntz, I. D. (2003) Molecular recognition and docking algorithms, *Annu Rev Biophys Biomol Struct 32*, 335-73.
3. Jorgensen, W. L. (2004) The many roles of computation in drug discovery, *Science 303*, 1813-8.
4. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J Mol Biol 247*, 536-40.
5. Gerlt, J. A., and Babbitt, P. C. (2001) Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies, *Annu Rev Biochem 70*, 209-46.
6. Babbitt, P. C. (2003) Definitions of enzyme function for the structural genomics era, *Curr Opin Chem Biol 7*, 230-7.
7. Gerlt, J. A., and Raushel, F. M. (2003) Evolution of function in (beta/alpha)8-barrel enzymes, *Curr Opin Chem Biol 7*, 252-64.
8. Babbitt, P. C., Hasson, M. S., Wedekind, J. E., Palmer, D. R., Barrett, W. C., Reed, G. H., Rayment, I., Ringe, D., Kenyon, G. L., and Gerlt, J. A. (1996) The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids, *Biochemistry 35*, 16489-501.

9. Bissantz, C., Folkers, G., and Rognan, D. (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations, *J Med Chem 43*, 4759-67.

10. Stahl, M., and Rarey, M. (2001) Detailed analysis of scoring functions for virtual screening, *Journal of Medicinal Chemistry 44*, 1035-1042.

11. Wang, R., Lu, Y., and Wang, S. (2003) Comparative evaluation of 11 scoring functions for molecular docking, *J Med Chem 46*, 2287-303.

12. Glide. (2003), Schrodinger Inc. New York, NY.

13. Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., and Shenkin, P. S. (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy, *J Med Chem 47*, 1739-49.

14. Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., and Banks, J. L. (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening, *J Med Chem 47*, 1750-9.

15. Jorgensen, W. L., Maxwell, D. S., and TiradoRives, J. (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids, *Journal of the American Chemical Society 118*, 11225-11236.

16. Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., and Jorgensen, W. L. (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides, *Journal of Physical Chemistry B 105*, 6474-6487.

17. Ghosh, A., Rapp, C. S., and Friesner, R. A. (1998) Generalized born model based on a surface integral formulation, *Journal of Physical Chemistry B 102*, 10983-10990.

18. Kollman, P. (1993) Free-Energy Calculations - Applications to Chemical and Biochemical Phenomena, *Chemical Reviews 93*, 2395-2417.

19. Hoffmann, D., Kramer, B., Washio, T., Steinmetzer, T., Rarey, M., and Lengauer, T. (1999) Two-stage method for protein-ligand docking, *Journal of Medicinal Chemistry 42*, 4422-4433.

20. Zou, X. Q., Sun, Y. X., and Kuntz, I. D. (1999) Inclusion of solvation in ligand binding free energy calculations using the generalized-born model, *Journal of the American Chemical Society 121*, 8033-8043.

21. Verkhivker, G. M., Bouzida, D., Gehlhaar, D. K., Rejto, P. A., Arthurs, S., Colson, A. B., Freer, S. T., Larson, V., Luty, B. A., Marrone, T., and Rose, P. W. (2000) Deciphering common failures in molecular docking of ligand-protein complexes, *Journal of Computer-Aided Molecular Design 14*, 731-751.

22. Wu, G., Robertson, D. H., Brooks, C. L., and Vieth, M. (2003) Detailed analysis of grid-based molecular docking: A case study of CDOCKER - A CHARMm-based MD docking algorithm, *Journal of Computational Chemistry 24*, 1549-1562.

23. Floriano, W. B., Vaidehi, N., Zamanakos, G., and Goddard, W. A., 3rd. (2004) HierVLS hierarchical docking protocol for virtual ligand screening of large-molecule databases, *J Med Chem 47*, 56-71.

24. Goto, S., Okuno, Y., Hattori, M., Nishioka, T., and Kanehisa, M. (2002) in *Nucleic Acids Res* pp 402-4.

25. Daylight. Los Altos, CA.

26. OpenEye. Santa Fe, NM.

27. Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., and Mee, R. P. (1997) Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes, *J Comput Aided Mol Des 11*, 425-45.

28. Jacobson, M. P., Kaminski, G. A., Friesner, R. A., Rapp, C. S.,. (2002) Force Field Validation Using Protein Side Chain Prediction, *J. Phys. Chem. B106*, 11673.

29. Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J., Honig, B., Shaw, D. E., and Friesner, R. A. (2004) A hierarchical approach to all-atom protein loop prediction, *Proteins 55*, 351-67.

30. Schlick, T., and Fogelson, A. (1992) Tnpack - a Truncated Newton Minimization Package for Large-Scale Problems .1. Algorithm and Usage, *Acm Transactions on Mathematical Software 18*, 46-70.

31. Xie, D. X., and Schlick, T. (1999) Efficient implementation of the truncated-Newton algorithm for large-scale chemistry applications, *Siam Journal on Optimization 10*, 132-154.

32. Schlick, T., and Overton, M. (1987) A Powerful Truncated Newton Method for Potential-Energy Minimization, *Journal of Computational Chemistry 8*, 1025-1039.

33. QikSim. (2003).

34. Willett, P., Barnard, J. M., Downs, G. M.,. (1998) Chemical Similarity Searching, *J. Chem. Inf.Comput. Sci. 38*, 983.

35. Neidhart, D. J., Howell, P. L., Petsko, G. A., Powers, V. M., Li, R. S., Kenyon, G. L., and Gerlt, J. A. (1991) Mechanism of the reaction catalyzed by mandelate racemase. 2.

Crystal structure of mandelate racemase at 2.5-A resolution: identification of the active site and possible catalytic residues, *Biochemistry 30*, 9264-73.

36. Landro, J. A., Kallarakal, A. T., Ransom, S. C., Gerlt, J. A., Kozarich, J. W., Neidhart, D. J., and Kenyon, G. L. (1991) Mechanism of the reaction catalyzed by mandelate racemase. 3. Asymmetry in reactions catalyzed by the H297N mutant, *Biochemistry 30*, 9274-81.

37. Mitra, B., Kallarakal, A. T., Kozarich, J. W., Gerlt, J. A., Clifton, J. G., Petsko, G. A., and Kenyon, G. L. (1995) Mechanism of the reaction catalyzed by mandelate racemase: importance of electrophilic catalysis by glutamic acid 317, *Biochemistry 34*, 2777-87.

38. Gulick, A. M., Hubbard, B. K., Gerlt, J. A., and Rayment, I. (2000) Evolution of enzymatic activities in the enolase superfamily: crystallographic and mutagenesis studies of the reaction catalyzed by D-glucarate dehydratase from Escherichia coli, *Biochemistry 39*, 4590-602.

39. Babbitt, P. C., Mrachko, G. T., Hasson, M. S., Huisman, G. W., Kolter, R., Ringe, D., Petsko, G. A., Kenyon, G. L., and Gerlt, J. A. (1995) A functionally diverse enzyme superfamily that abstracts the alpha protons of carboxylic acids, *Science 267*, 1159-61.

40. Helin, S., Kahn, P. C., Guha, B. L., Mallows, D. G., and Goldman, A. (1995) The refined X-ray structure of muconate lactonizing enzyme from Pseudomonas putida PRS2000 at 1.85 A resolution, *J Mol Biol 254*, 918-41.

41. Levy, C. W., Buckley, P. A., Sedelnikova, S., Kato, Y., Asano, Y., Rice, D. W., and Baker, P. J. (2002) Insights into enzyme evolution revealed by the structure of methylaspartate ammonia lyase, *Structure (Camb) 10*, 105-13.

42. Goda, S. K., Minton, N. P., Botting, N. P., and Gani, D. (1992) Cloning, sequencing, and expression in Escherichia coli of the Clostridium tetanomorphum gene encoding beta-methylaspartase and characterization of the recombinant protein, *Biochemistry 31*, 10747-56.

43. Shindyalov, I. N., and Bourne, P. E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein Eng 11*, 739-47.

44. Schmidt, D. M., Hubbard, B. K., and Gerlt, J. A. (2001) Evolution of enzymatic activities in the enolase superfamily: functional assignment of unknown proteins in Bacillus subtilis and Escherichia coli as L-Ala-D/L-Glu epimerases, *Biochemistry 40*, 15707-15.

45. Gulick, A. M., Schmidt, D. M., Gerlt, J. A., and Rayment, I. (2001) Evolution of enzymatic activities in the enolase superfamily: crystal structures of the L-Ala-D/L-Glu epimerases from Escherichia coli and Bacillus subtilis, *Biochemistry 40*, 15716-24.

46. Klenchin, V. A., Schmidt, D. M., Gerlt, J. A., and Rayment, I. (2004) Evolution of Enzymatic Activities in the Enolase Superfamily: Structure of a Substrate-Liganded Complex of the l-Ala-d/l-Glu Epimerase from Bacillus subtilis(,), *Biochemistry 43*, 10370-8.

47. Thompson, T. B., Garrett, J. B., Taylor, E. A., Meganathan, R., Gerlt, J. A., and Rayment, I. (2000) Evolution of enzymatic activity in the enolase superfamily: structure of o-succinylbenzoate synthase from Escherichia coli in complex with Mg2+ and o-succinylbenzoate, *Biochemistry 39*, 10662-76.

48. Klenchin, V. A., Ringia, E. A. T., Gerlt, J. A., and Rayment, I. (2003) Evolution of enzymatic activity in the enolase superfamily: Structural and mutagenic studies of the

mechanism of the reaction catalyzed by o-succinylbenzoate synthase from Escherichia coli, *Biochemistry 42*, 14427-14433.

49. Wedekind, J. E., Poyner, R. R., Reed, G. H., and Rayment, I. (1994) Chelation of serine 39 to Mg2+ latches a gate at the active site of enolase: structure of the bis(Mg2+) complex of yeast enolase and the intermediate analog phosphonoacetohydroxamate at 2.1-A resolution, *Biochemistry 33*, 9333-42.

Tables

Table 1. Ranking of known substrates and products after docking and rescoring.

Enzyme	Substrate	Rank after Docking (%)	Rank after Rescoring (%)	Product	Rank after Docking (%)	Rank after Rescoring (%)
MR						
Holo	S-mandelate	6.5	0.41	R-mandelate	9.0	0.74
Apo		12.1	0.65		9.3	3.9
GlucD						
Holo	D-glucarate	0.03	0.02	5-keto,4-deoxy-D-glucarate	0.04	0.005
MLE-I						
Apo	cis-cis muconate	19.4	5.4	Mucono-lactone	10.2	0.29
MAL						
Holo	L-threo(2S, 3S) 3-methyl aspartate	9.1	1.0	Mesaconic-acid	10.7	0.08
Apo		5.9	0.05		14.1	2.2
AEE						
Holo	L-Ala-L-Glu dipeptide	0.92	0.47	L-Ala-D-Glu dipeptide	0.61	0.42
Apo		17.0	4.9		13.2	6.7
OSBS						
Holo	2-succinyl-6-hydroxy 2,4-cyclo hexadiene 1-carboxylate	6.1	0.21	2-succinyl benzoate	4.9	0.04
Apo		9.8	5.2		18.4	0.3
Enolase						
Holo	2-Phospho-glycerate	0.04	0.16	Phospho-enol pyruvate	0.06	0.09

Table 2. Rank (in percent) of the known substrate of a particular family compared against other family members, after docking (a) and rescoring (b). The entry ">25" signifies that the ligand ranked lower than the top 25% after the docking phase, and thus was not subjected to rescoring. The Ala-Glu epimerase (AEE) is not included in these results because the amino acid dipeptides are not part of the standard KEGG LIGANDS library.

(a)

	S-mande-late	D-glucar-ate	cis-cis muconate	L-threo (2s,3s)-3-methyl aspartate	2-succinyl-2-hydroxy-2,4-cyclohexadiene-1-carboxylate	2-phospho-glycerate
MR (holo)	6.5	12.8	22.6	24.7	10.7	6.5
MR (apo)	12.1	22.9	>25	>25	12.6	7.6
GlucD (holo)	>25	0.03	>25	3.9	>25	0.93
MLE-I (apo)	3.8	0.69	19.4	23.7	16.0	6.8
MAL (holo)	4.9	0.22	10.6	9.1	5.1	2.7
MAL (apo)	7.1	11.6	>25	5.9	>25	>25
OSBS (holo)	12.9	11.5	13.3	25.7	6.1	11.7
OSBS (apo)	>25	>25	>25	>25	9.8	>25
Enolase (holo)	3.7	5.7	>25	1.9	>25	0.04

(b)

	S-mande-late	D-glucar-ate	cis-cis muconate	L-threo (2s,3s)-3-methyl aspartate	2-succinyl-2-hydroxy-2,4-cyclohexadiene-1-carboxylate	2-phospho-glycerate
MR (holo)	0.41	4.5	21.1	9.3	18.8	21.9
MR (apo)	0.65	9.8	>25	>25	12.0	19.8
GlucD (holo)	>25	0.02	>25	5.0	>25	8.6
MLE-I (apo)	3.2	3.3	5.4	4.1	23.1	23.9
MAL (holo)	6.2	11.0	22.2	1.1	23.8	22.6
MAL (apo)	0.78	4.1	>25	0.05	>25	>25
OSBS (holo)	2.1	6.5	18.7	7.0	0.21	12.3
OSBS (apo)	>25	>25	>25	>25	5.2	>25
Enolase (holo)	21.2	1.4	>25	1.1	>25	0.16

Figure Captions:

Figure 1. Substrate binding pose for (a) MR holo, (b) MR apo, (c) GlucD holo, (d) MLE-I apo, (e) MAL holo, (f) MAL apo, (g) AEE holo, (h) AEE apo, (i) OSBS holo, (j) OSBS apo and (k) enolase holo structures. Residues that participate in catalysis and coordinate to the metal ion are also shown. For holo enymes, we also show the co-crystallized ligand (green). The metal ion present in the binding pocket is either Mg^{2+} or Mn^{2+}.

Figure 2. Chemical similarity (Tanimoto Coefficient) as a function of % of database for (a) MR holo, (b) MR apo, (c) GlucD holo, (d) MLE-I apo,(e) MAL holo, (f) MAL apo, (g) AEE holo, (h) AEE apo, (i) OSBS holo, (j) OSBS apo and (k) enolase holo structures. The Tanimoto coefficient decreases from 1 as the chemical similarity decreases. Chemical similarity is defined by descriptors that include both the numbers of common functional groups and whole molecule descriptors such as dipole and volume. Enrichment of compounds that are chemically similar to the known substrate after docking (blue line) and rescoring (red line) are shown. The results have been smoothed to decrease noise and emphasize the overall trends.

Figure 3. Percentage of compounds with a hydrogen atom at the carbon α to a carboxylate group, which is required for the half-reaction that defines the enolase superfamily. The results have been smoothed to decrease noise and emphasize the overall trends.

Figure 4. Conformation of loop residues 14-30 in AEE in the (a) open (apo) and (b) closed (holo) forms.

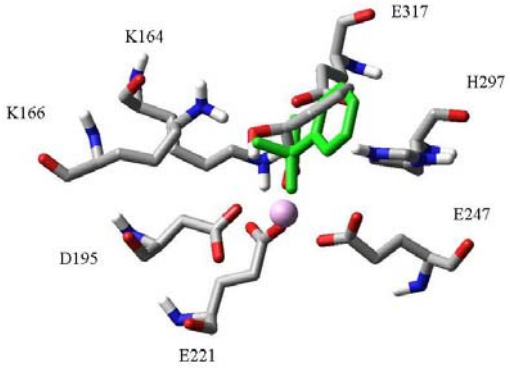

Fig 1a) Holo MR (1MDR)

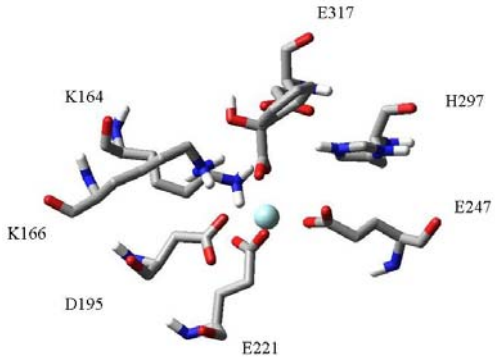

Fig 1b) Apo MR (2MNR)

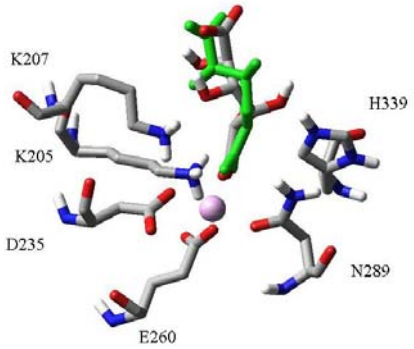

Fig 1c) Holo GlucD (1ECQ)

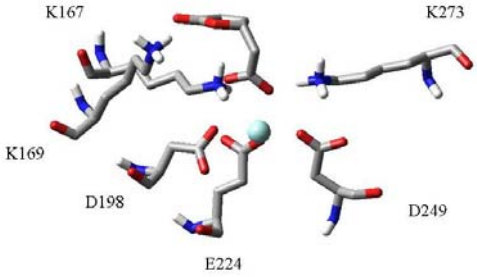

Fig 1d) Apo MLE-I (1MUC)

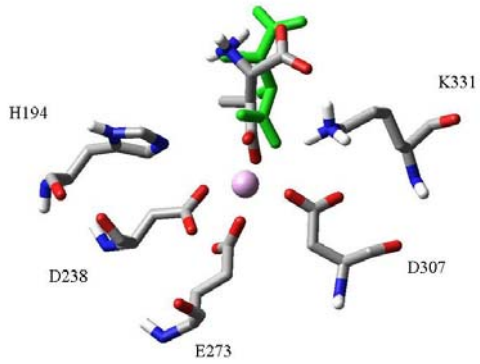

Fig 1e) Holo MAL (1KKR)

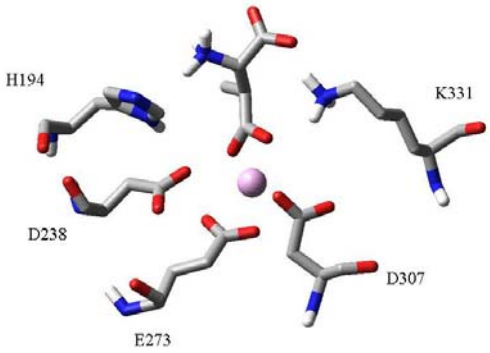

Fig 1f) Apo MAL (1KKO)

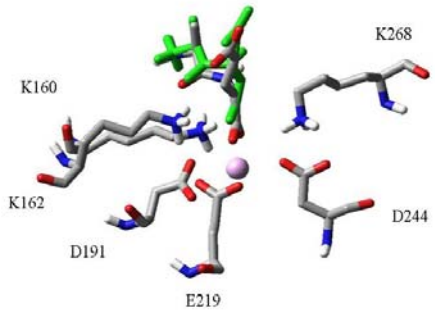

Fig 1g) AEE Holo (1TKK)

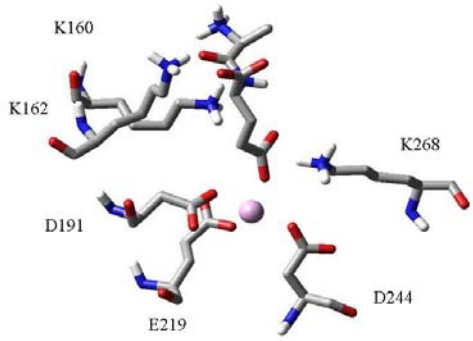

Fig 1h) AEE Apo (1JPM)

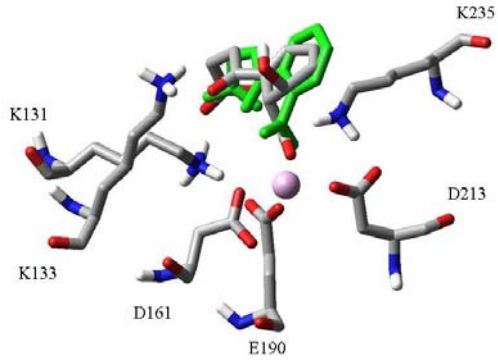

Fig 1i) Holo OSBS (1FHV)

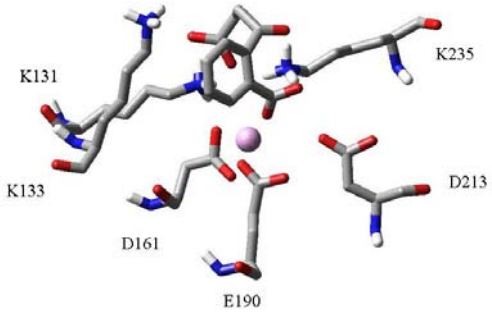

Fig 1j) Apo OSBS (1FHU)

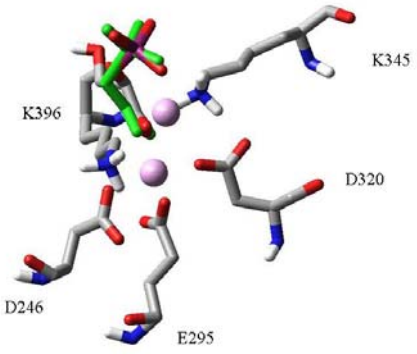

Fig 1k) Holo Enolase (1EBG)

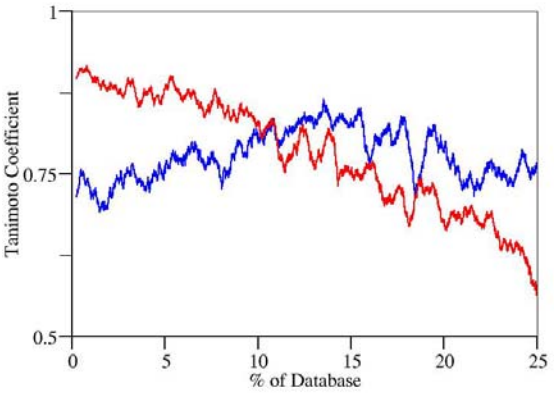

Fig 2a) Holo MR (1mdr)

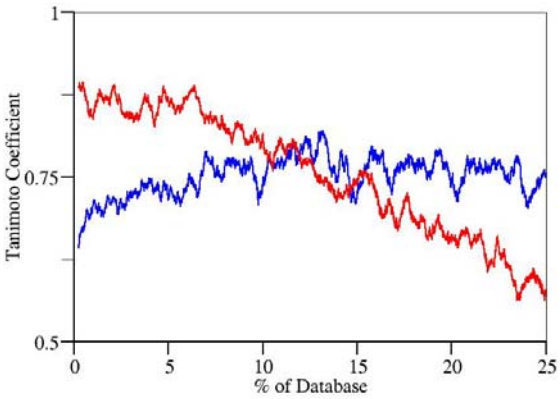

Fig 2b) Apo MR (2mnr)

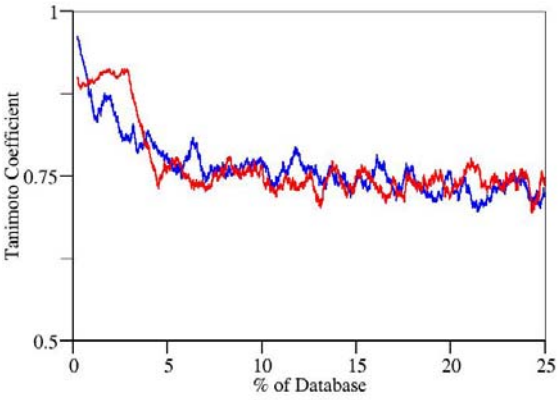

Fig 2c) Holo GlucD (1ecq)

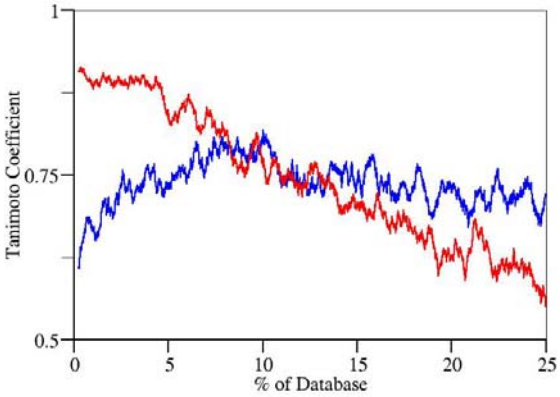

Fig 2d) Apo MLE-I (1muc)

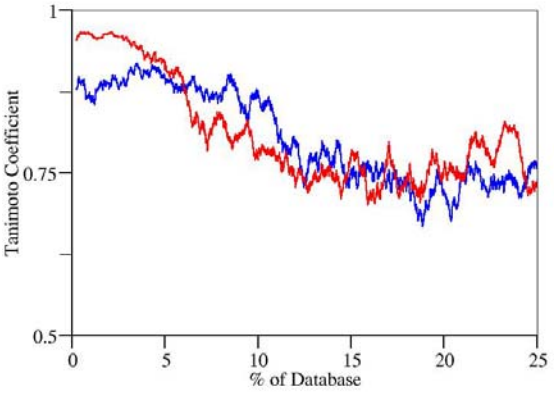

Fig 2e) Holo MAL (1kkr)

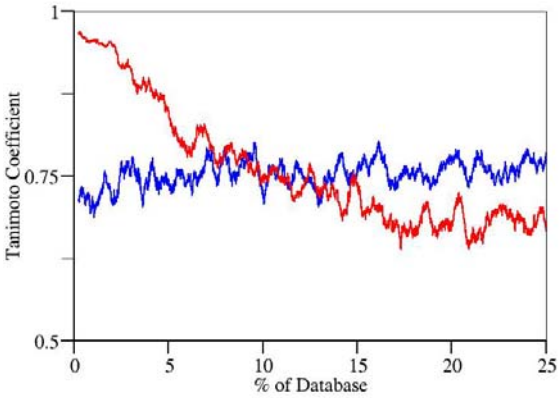

Fig 2f) Apo MAL (1kko)

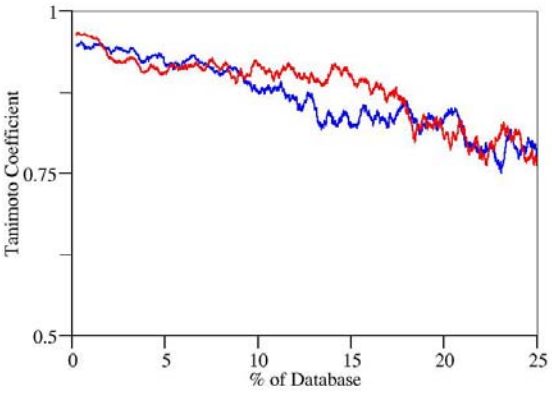

Fig 2g) AEE Holo (1tkk)

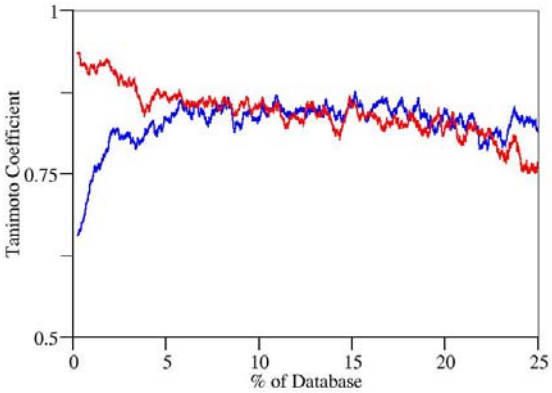

Fig 2h) AEE Apo (1jpm)

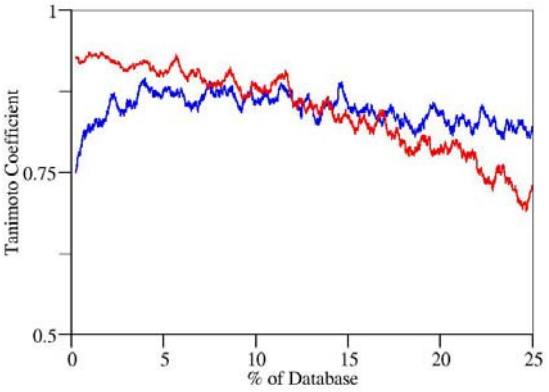

Fig 2i) Holo OSBS (1fhv)

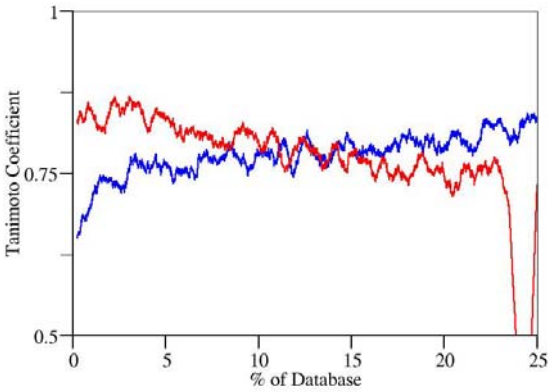

Fig 2j) Apo OSBS (1fhu)

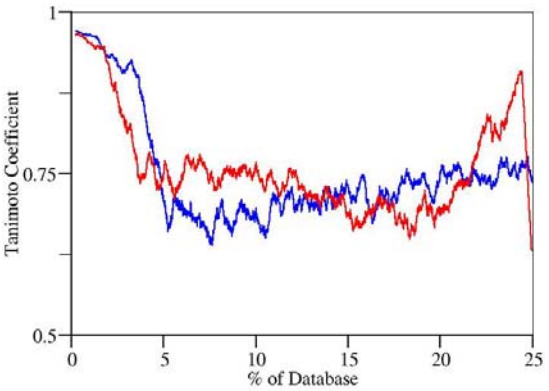

Fig 2k) Holo Enolase (1ebg)

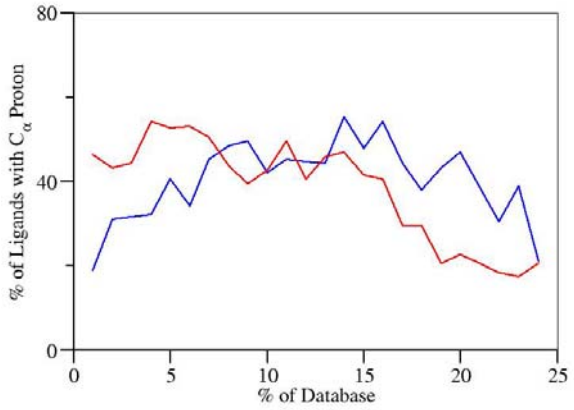

Fig 3a) Holo MR (1mdr)

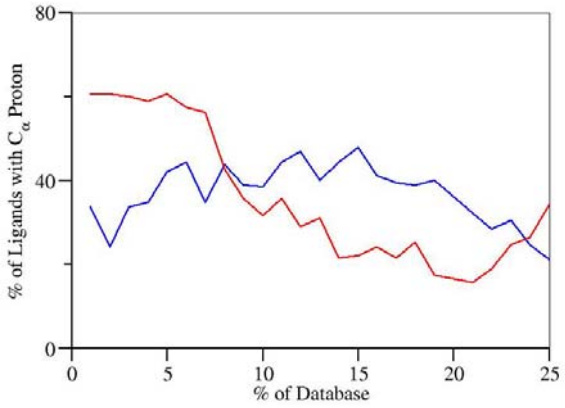

Fig 3b) Apo MR (2mnr)

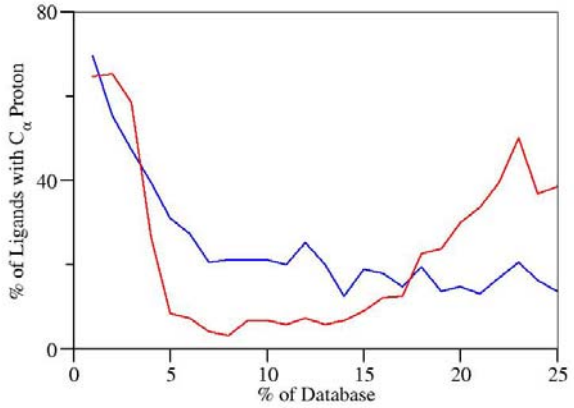

Fig 3c) Holo GlucD (1ecq)

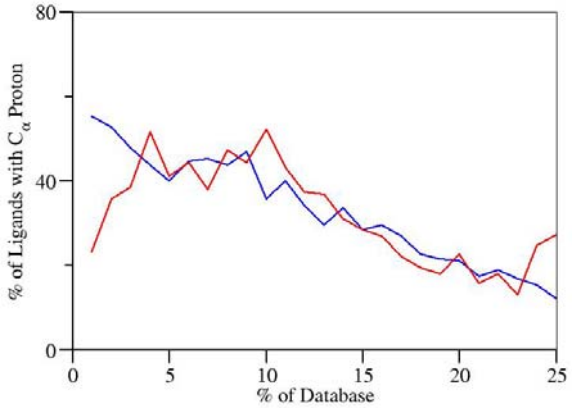

Fig 3d) Apo MLE-I (1muc)

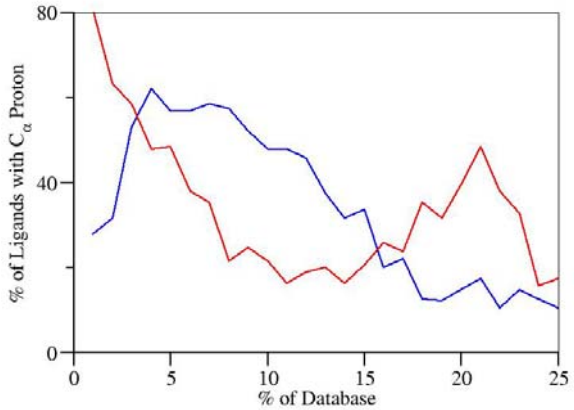

Fig 3e) Holo MAL (1kkr)

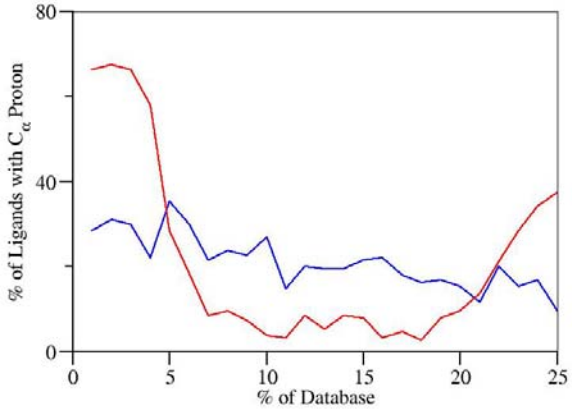

Fig 3f) Apo MAL (1kko)

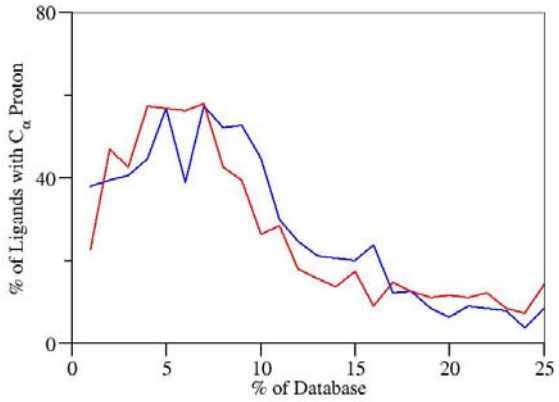

Fig 3g) AEE Holo (1tkk)

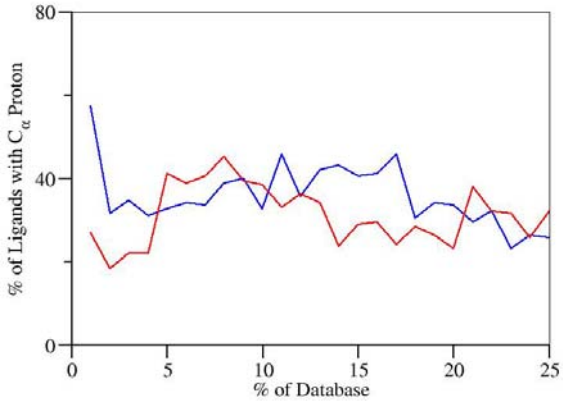

Fig 3h) AEE Apo (1jpm)

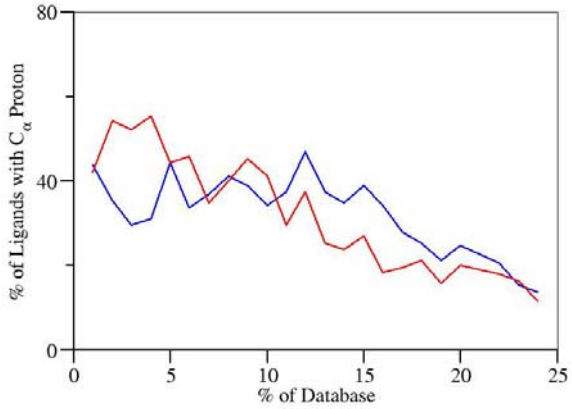

Fig 3i) Holo OSBS (1fhv)

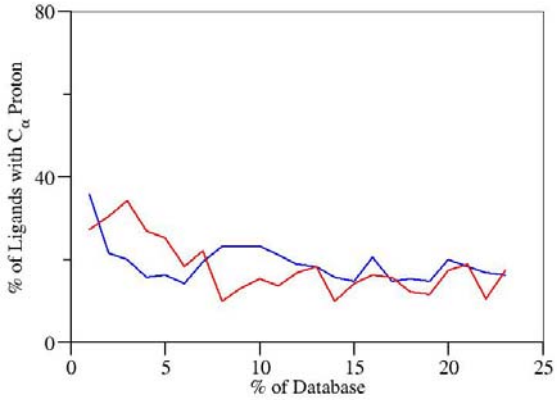

Fig 3j) Apo OSBS (1fhu)

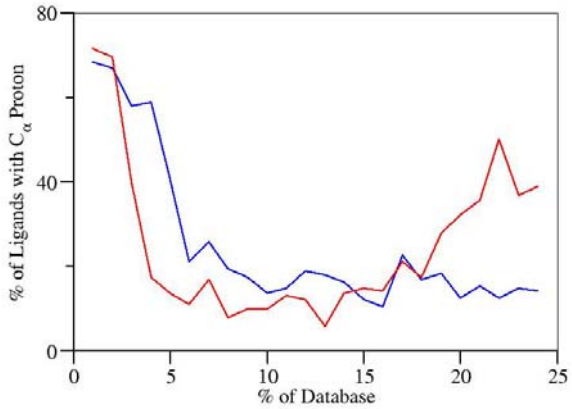

Fig 3k) Holo Enolase (1ebg)

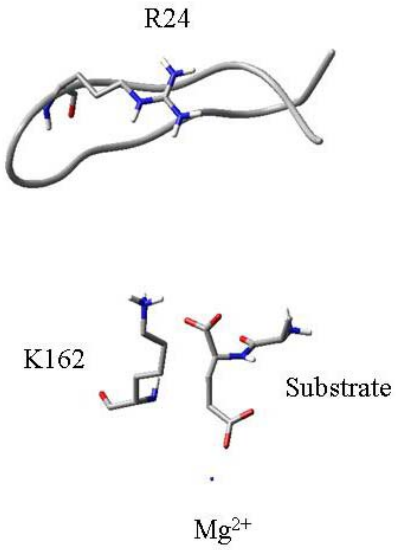

Fig 4a) AEE Apo (1jpm)

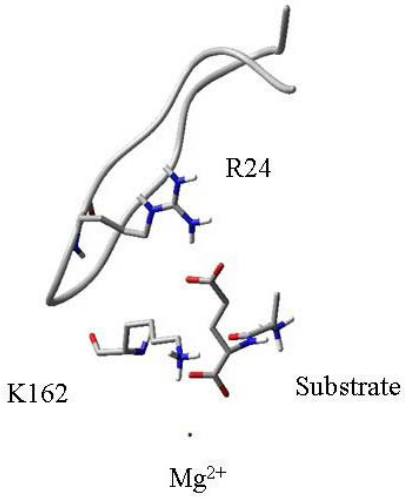

Fig 4b) AEE Holo (1tkk)