SABER: A computational method for identifying active sites for new reactions

Geoffrey R. Nosrati and K. N. Houk*

Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095-1569

Received 2 December 2011; Revised 1 February 2012; Accepted 8 February 2012

DOI: 10.1002/pro.2055

Published online 28 February 2012 proteinscience.org

Abstract: A software suite, SABER (Selection of Active/Binding sites for Enzyme Redesign), has been developed for the analysis of atomic geometries in protein structures, using a geometric hashing algorithm (Barker and Thornton, *Bioinformatics* 2003;19:1644–1649). SABER is used to explore the Protein Data Bank (PDB) to locate proteins with a specific 3D arrangement of catalytic groups to identify active sites that might be redesigned to catalyze new reactions. As a proof-of-principle test, SABER was used to identify enzymes that have the same catalytic group arrangement present in *o*-succinyl benzoate synthase (OSBS). Among the highest-scoring scaffolds identified by the SABER search for enzymes with the same catalytic group arrangement as OSBS were L-Ala D/L-Glu epimerase (AEE) and muconate lactonizing enzyme II (MLE), both of which have been redesigned to become effective OSBS catalysts, demonstrated by experiments. Next, we used SABER to search for naturally existing active sites in the PDB with catalytic groups similar to those present in the designed Kemp elimination enzyme KE07. From over 2000 geometric matches to the KE07 active site, SABER identified 23 matches that corresponded to residues from known active sites. The best of these matches, with a 0.28 Å catalytic atom RMSD to KE07, was then redesigned to be compatible with the Kemp elimination using RosettaDesign. We also used SABER to search for potential Kemp eliminases using a theozyme predicted to provide a greater rate acceleration than the active site of KE07, and used Rosetta to create a design based on the proteins identified.

Keywords: enzyme design; active site; protein substructure analysis; active site redesign

Introduction

The rational design of enzymes for new reactions is one of the greatest challenges in computational biology. There have been four major recent successes for the design of enzymes that are capable of accelerating reactions not found in nature. Nevertheless, many hurdles must be overcome to make the rational design process robust and general.^{1–6}

The most successful computational enzyme designs thus far involve catalysis of a Kemp elimina-

tion, a retro-aldol reaction, a stereoselective Diels-Alder reaction, and nitric oxide reduction.^{3–6} The first three enzymes were designed using the “inside-out” approach, in which active sites are designed by quantum mechanical methods; transition states with catalytic functionality from appropriate side chains of amino acids are optimized to find a suitable geometrical arrangement that is predicted to accelerate the reaction. This is referred to as a theoretical enzyme, or theozyme.⁷ Subsequently, a protein structure is designed that is predicted to fold to form this active site. Because the *de novo* design of folds is not routinely feasible, the Rosetta programs developed by the Baker laboratory are used to find a suitable fold into which the theozyme can be incorporated.^{8,9} RosettaMatch is used to determine whether the theozyme can be grafted into one or more of the scaffolds in a scaffold library. This must be achieved with low energy conformations of the side chains involved in

Grant sponsors: National Institute of General Medical Science; Grant number: GM 75962; Grant sponsor: National Institutes of Health; Grant numbers: T32 GM008496-16; Grant sponsors: Department of Defense, Defense Advanced Research Projects Agency.

*Correspondence to: K. N. Houk, Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095-1569. E-mail: houk@chem.ucla.edu

the theozyme. Following this step, RosettaDesign is then used to fill in the remaining side chains in the active site around the theozyme, optimizing protein packing and transition state binding.⁹

A critical step in enzyme design is the proper placement of the catalytic residues. The importance of the positioning of the residues in the active site has been discussed extensively and is clearly a feature of proficient enzymes.¹⁰ Many examples have been described in the literature: Warshel has proposed that active site preorganization and electrostatic stabilization of the transition state are the principal factors controlling enzyme catalysis.^{11–13} Preorganization involves the correct spatial positioning of catalytic groups. Hilvert and coworkers have demonstrated that mutating a catalytic Glu residue to an Asp in the 34E4 Kemp eliminase catalytic antibody has a significant (>2 kcal/mol transition state destabilization) effect on catalysis, indicating the need for precise placement of catalytic groups.¹⁴ A recent investigation of serine esterases has shown that their active sites are preorganized into geometries that allow the reaction to be carried out with a minimal rearrangement of catalytic residues in the many steps of the catalytic cycle. These geometries are very close to the optimum geometries computed using quantum mechanics.¹⁵ Significant deviations from the optimum catalytic arrangement of residues are generally not found in nature; computational tests have been carried out on a wide variety of enzymes to show that evolution leads to active sites with optimum catalytic distances, according to comparisons with quantum mechanical calculations.¹⁶

For the reasons discussed above, the selection of a scaffold that can support correctly positioned and oriented catalytic groups is an essential feature of enzyme design. In addition, an ideal scaffold should provide an environment such that the pK_a values of catalytic acids or bases and the reactivities of nucleophilic or electrophilic groups are optimal. Finally, the designed active site must be sufficiently isolated from bulk aqueous solvent to allow for efficient catalysis to take place in an optimum environment.

Mutations distant from the active site can also have a significant impact on enzyme catalysis.¹⁷ In addition, dynamic effects, such as loop motions, may also play a key role in determining the catalytic power of a designed enzyme, as suggested by molecular dynamics simulations of designed retroaldolases.¹⁸ However, whether dynamical motions coupled to the reaction coordinate impact enzyme catalysis is still a subject of considerable debate.^{19–22} All of these factors contribute to the enormous challenge inherent in rational enzyme design, as even minor changes can have significant effects on catalysis. As these effects are difficult to predict, one way to limit their impact is to keep the number of mutations required for active site redesign to a minimum.

The use of active sites that require minimal engineering to catalyze new reactions has been the subject of two recent reviews by the Hilvert, Gerlt, and Babbitt laboratories.^{23,24}

We have explored an alternative to grafting catalytic groups into a protein scaffold. Instead, we search for a natural protein that has the necessary catalytic groups appropriately positioned for catalysis of the desired reaction. Essentially, we are searching for active sites that might be promiscuous for a target reaction, or might be engineered to be promiscuous for the target reaction with mutations to the noncatalytic residues in the active site.

This approach involves searching the public repositories of protein structures, such as the Protein Data Bank (PDB), to determine if there are enzymes that have active site residues in a geometry capable of catalyzing the reaction of interest. If a protein with the necessary preorganized catalytic groups is found, mutations necessary to accommodate the new transition state, and to provide additional stabilizing groups, can then be incorporated through protein design programs such as RosettaDesign, Dezymer, or Phoenix, or by experimental approaches such as directed evolution.^{8,25–27} This approach can also have the benefit of providing an active site with catalytic groups in an environment that provides suitable pK_a values for catalysis without extreme redesign.

We have developed a method for the identification of active sites with geometries suitable for a desired reaction. This procedure uses the program we have developed, Selection of Active/Binding sites for Enzyme Redesign (SABER), and a Catalytic Atom Map (CAM), a truncated model of the ideal geometry of the key catalytic residues in the active site. SABER is used to search the PDB for known enzymes with CAM-like geometric arrangements of functional groups.

The program combines geometric searches of the PDB to locate CAM-like geometries in known protein structures and informatics tools to evaluate the suitability of these geometries to function as catalytic sites. We have validated this methodology using a known example from the literature in which the activity of one enzyme, *o*-succinyl benzoate synthase (OSBS), has been transplanted into two other known enzymes. We have also used SABER in conjunction with RosettaDesign to computationally design a new enzyme that is closely related to a previously designed and active Kemp eliminase, KE07.

Results

SABER as a tool for enzyme design

SABER is described in detail in the Materials and Methods section. Briefly, this program differs from other enzyme design approaches in terms of how the active site residues are located and identified.

SABER is used to locate protein structures that have their catalytic groups arranged in an optimal geometry for the catalysis of the target reaction. The program rapidly searches and scores the entire PDB, providing a very large pool of potential active or binding sites for redesign. The scoring functions of SABER identify the most promising sites for redesign, based both on the geometric placement of the proposed catalytic residues and whether or not they are part of a known active or binding site and have predicted pK_a values of catalytic residues appropriate for the target reaction. We describe how SABER can identify the best set of active site redesign candidates, which we term “predesigns,” from a pool of thousands of geometric matches to the CAM.

Using CAMs to select active sites capable of catalyzing new reactions

The Gerlt and Babbitt laboratories have altered the active sites of two enzymes with different catalytic functions to perform the reaction catalyzed by OSBS. The two enzymes redesigned to perform the OSBS reaction were L-Ala D/L-Glu epimerase (AEE) and muconate lactonizing enzyme II (MLE). The key catalytic residues were unchanged by the mutations and catalysis was verified by experiment.²⁸ We used this example to test the ability of SABER to locate catalytic functionality in cases where experimentalists have already proven the effectiveness of a modified enzyme to catalyze a reaction not found in the native enzyme. A CAM based on the catalytic residues of OSBS was used to search the PDB for scaffolds that have their catalytic groups arranged appropriately to catalyze this reaction.

OSBS, AEE, and MLE all share geometrically similar active sites, with three conserved Asp residues, used to bind Mg^{2+} , and two conserved Lys residues.²⁸ The essential Mg^{2+} stabilizes an enediolate intermediate in each enzyme; however, the reaction catalyzed by each wild-type enzyme is quite different. AEE catalyzes the epimerization at the alpha carbon of its target amino acids. MLE catalyzes the lactonization of muconate, and OSBS catalyzes a dehydration to form *o*-succinyl benzoate. These reactions are depicted in Figure 1. It has been demonstrated experimentally that OSBS activity can be introduced into both AEE and MLE, through mutations to their active sites, even though neither of these enzymes shows any native OSBS activity.^{28–30} These mutations leave the key catalytic residues intact, but optimize the rest of the active site for the new substrate and transition state.

L-alanine/D-glucose epimerase

It has been demonstrated experimentally that wild-type AEE does not catalyze the OSBS reaction.²⁸ However, this enzyme contains a similar active site, as indicated by the CAM search. OSBS activity was installed by Gerlt *et al.* in the *E. coli* AEE via a single

Figure 1. The reactions catalyzed by *o*-succinyl benzoate synthase (OSBS), L-alanine/D-glucose epimerase (AEE), muconate lactonizing enzyme (MLE), chloromuconate lactonizing enzyme (CI-MLE), and mandelate racemase (MR).

mutation, changing Asp297 to a glycine. This single residue change improves k_{cat} for the OSBS reaction from undetectable to $2.5 \times 10^{-3} \text{ sec}^{-1}$.²⁸ This is a modest acceleration compared with a wild-type OSBS enzyme, which has a k_{cat} of 24 sec^{-1} in *E. coli*.²⁸ However, it still represents an acceleration of approximately seven orders of magnitude over the uncatalyzed reaction. Further improvements in k_{cat} were achieved with two additional mutations (I19F and R24W), which improved it to $1.6 \times 10^{-1} \text{ sec}^{-1}$, which is a 10^9 -fold rate enhancement versus the uncatalyzed reaction.^{29,30}

Muconate lactonizing enzyme

MLE has two catalytic lysines and three catalytic carboxylates arranged very similarly to the active site of OSBS. This enzyme has been re-engineered by Gerlt and coworkers to perform the OSBS reaction via a single amino acid change. In this case, MLE II from *Pseudomonas sp. P51* was able to catalyze the OSBS reaction after changing Glu323 to a glycine.²⁸ This single residue change increased k_{cat} from undetectable to 1.5 sec^{-1} , compared to the activity of wild-type *E. coli* OSBS, at 24 sec^{-1} . This represents an approximately 10^{10} -fold enhancement of the rate versus the background reaction. Unlike the AEE case, additional mutation experiments to improve the rate have not been published. However, this single amino acid change produces an enzyme that is within an order of magnitude of wild-type activity.

Figure 2. The OSBS active site from crystal structure 1FHV. The atoms used in the Catalytic Atom Map are shown as spheres. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

We used these examples to test the effectiveness of SABER at identifying suitable candidates for active site redesign. We searched all structures in the PDB90 data set with a resolution ≤ 2.0 Å to locate proteins with arrangements of atoms matching the CAM of the OSBS active site. A five atom map was constructed for the target active site to represent the three carboxylic acids and two lysines in the OSBS active site. This is shown in Figure 2. These atoms were constrained by atom type, residue type, and interatomic distances.

The three carboxylate ligands for Mg^{2+} are defined by the three oxygen atoms in the CAM. These oxygen atoms must be from an aspartate or glutamate (PDB atom codes OD1, OD2, OE1, and OE2), and the nitrogen atoms must be from lysine residues. The CAM specifies that the only nitrogen matches must be lysine ϵ -amino nitrogens, as naturally occurring OSBS enzymes use lysine exclusively in this role. The match radius for each atom was set at 2.0 Å.

All of the SABER predesigns located using the OSBS CAM where the RMSD was ≤ 0.6 Å were ana-

Figure 3. The top predesigns identified by the OSBS catalytic atom map (rendered as spheres). Predesigns by color: MLE (cyan, PDB code: 1MUC), AEE (magenta, 1JPM), Cl-MLE (green, 1NU5), MR (purple, 2QDE), and MR (orange, 2RDX). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figure 4. The Kemp elimination catalyzed by the KE07 designed enzyme. In the enzyme, B indicates a residue functioning as a general base, while HA indicates a residue functioning as a hydrogen bond donor.

lyzed. The search generated five predesigns within this RMSD range, summarized in Table I and shown in Figure 3. As there were no available high resolution structures for an L-Ala/D-Glu epimerase in the PDB90 data set, one structure of this enzyme (PDB code: 1JPM) was manually analyzed, for completeness.

All of the predesigns identified within the specified RMSD range were known active site homologs of OSBS. Each of the predesigns catalyzes one of the following three reactions: the OSBS reaction, L-alanine/D-glutamate epimerization, or muconate lactonization. The 1NU5 predesign is very similar to MLE, but this enzyme prefers chloromuconate as a substrate, instead of muconate. Two of the entries, 2RDX and 2QDE, were deposited in the PDB without full papers to describe them. Both of these proteins are described as mandelate racemase/muconate lactonizing enzyme family proteins in their PDB entries, so it is reasonable to expect that their active sites would be compatible with the OSBS reaction. The geometry of the active site for each of these proteins is in good agreement with the CAM for OSBS, lending further credence to this proposal.

This analysis indicates that the OSBS CAM is able to identify other enzymes that could be modified to be catalysts for the OSBS reaction. Two of the enzymes identified, MLE and AEE, have already been demonstrated experimentally to be capable of catalyzing the OSBS reaction. A key point is that the redesigned forms of both MLE and AEE leave the catalytic residues that match the OSBS CAM intact. The changes made in these enzymes to allow for catalysis of the OSBS reaction only create a binding pocket that can accommodate the new substrate.

Using CAMs to search for active sites geometrically similar to KE07

To evaluate the use of SABER in selecting active sites for redesign, we also performed a search for proteins that could potentially catalyze a Kemp elimination of 5-nitrobenzisoxazole, shown in Figure 4. We based the CAM on the active site of KE07, an enzyme designed specifically for this reaction by the Baker and Houk laboratories.³ KE07 used a Glu

Figure 5. The designed active site for KE07, featuring the catalytic base (Asp/Glu) and a Lys hydrogen bond donor. Catalytic atoms are rendered as spheres. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

residue as a general base and a Lys H-bond donor to catalyze the Kemp elimination, as shown in Figure 5. We sought to identify enzymes with similar active site geometries in the PDB, and then use Rosetta to redesign their active sites to be compatible with the Kemp transition state.

The CAM in Figure 5 was used for the SABER search. The atoms from the general base were allowed to match on the corresponding atoms from either Asp or Glu residues. Otherwise, the search tolerances were identical to the previous SABER search, with the match radius for each catalytic atom set at 2.0 Å.

The search generated 2259 predesigns with RMSD values ≤ 0.4 Å of the KE07 CAM. Two of the predesigns had an ActiveSiteScore of 2, while another 21 had an ActiveSiteScore of 1. These should be the most promising choices for redesign, because the active sites have already been annotated in the Catalytic Site Atlas (CSA). There were an additional 286 predesigns that were identified by BindingSiteFinder as having heteroatoms near the residues identified by the CAM, which were also potential candidates for redesign. The breakdown of the SABER predesigns is summarized in Figure 6.

Figure 6. Analysis of the results from the PDB search using the KE07-based catalytic atom map in Figure 5. Data is shown for predesigns with an RMSD ≤ 0.4 . [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

The top predesign with a predicted catalytic base pK_a in the range of 5–9 from each category (ActiveSiteScore = 2, ActiveSiteScore = 1, and ActiveSiteScore = 0 with a BindingSiteFinder flag) is shown in Table II.

As a proof-of-principle test of SABER's ability to identify scaffolds for redesign, the best predesign from the ActiveSiteScore = 2 category, PDB code 1X0L, was redesigned to accommodate the Kemp transition state using the RosettaDesign program. 1X0L is an isocitrate dehydrogenase enzyme from *T. thermophilus*. As shown in Figure 7, the catalytic residues from the KE07 crystal structure and those residues identified by SABER in the 1X0L crystal structure are in very similar orientations.

The starting structure for RosettaDesign was generated by superimposing the 5-nitrobenzisoxazole Kemp elimination transition state from the KE07 design on the 1X0L active site. The transition state was placed in the orientation identified during the SABER analysis. This resulted in a large number of steric clashes in the active site, as shown in Figure 8(A). After a single round of constrained optimization, redesign, and repacking, RosettaDesign was able to incorporate the Kemp elimination transition state into the 1X0L active site with 9 mutations of noncatalytic residues. This compares favorably with the original KE07 design in the 1THF scaffold, which required 13 mutations. Most importantly, the

Table I. Active Sites with Appropriate Catalytic Group Positioning for the OSBS Reaction, Identified by the 1FHV Catalytic Atom Map for *o*-Succinyl Benzoate Synthase

PDB code	Protein	RMSD	Residues identified (Mg^{2+}) ^a	Residues identified (Lys)
1FHV	<i>o</i> -succinyl benzoate synthase	0.00	ASP161, GLU190, ASP213	LYS133, LYS235
1MUC	Muconate lactonizing enzyme	0.26	ASP198, GLU224, ASP249	LYS169, LYS273
1NU5	Chloromuconate lactonizing enzyme	0.33	ASP194, GLU220, ASP245	LYS165, LYS269
2QDE	Mandelate racemase/MLE	0.38	ASP195, GLU221, ASP246	LYS167, LYS270
1JPM ^b	L-alanine/D-glucose epimerase	0.43	ASP191, GLU219, ASP244	LYS162, LYS268
2RDX	Mandelate racemase/MLE	0.52	ASP193, GLU218, ASP241	LYS165, LYS265

^a These residues compose the carboxylate triad predicted to coordinate with the Mg^{2+} atom.

^b There was no structure of AEE available with a resolution ≤ 2.0 Å, so this predesign was included manually to demonstrate the geometric match with the active site of this enzyme.

Table II. Active Sites Identified by the CAM Search as Having Catalytic Groups Arranged in a Geometry Similar to that of the KE07 Design

PDB	RMSD	ActiveSiteScore	BindingSiteFinder	Base pK_a	Residues
1X0L	0.28	2	None	5.1	Asp204, Lys171
1F0L	0.30	1	Yes	5.1	Glu189, Lys162
2C14	0.22	0	Yes	7.2	Asp131, Lys229

The top predesign from each category (ActiveSiteScore = 2, ActiveSiteScore = 1, ActiveSiteScore = 0, and BindingSiteFinder = yes) is shown.

catalytic residues are from the original scaffold and thus did not require any changes to the protein structure to be in the optimal position for catalysis of the new reaction. The redesigned active site of 1X0L is shown in Figure 8(B), with the steric clashes in the active site removed during the course of redesign.

As a final validation of SABER's usefulness in the enzyme design process, one more Kemp elimination design was produced. Examination of Figure 4 shows that the ideal placement for the hydrogen bond donor should have it oriented toward the oxygen atom of the nitrobenzisoazole transition state. However, as can be seen in Figure 5, the KE07 design has the hydrogen bond donor oriented toward the nitrogen atom. We used SABER and a CAM based on a theozyme calculated in our laboratory with a more ideal orientation of the H-bond donor to search the PDB for active sites that might be readily redesigned to function as more proficient Kemp eliminases. This theozyme, like the KE07 active site, employs a carboxylate (Asp/Glu) as the catalytic base, but uses an alcohol (Ser/Thr/Tyr) as the hydrogen bond donor, a motif found in other Kemp designs from reference 3. This theozyme is shown in Figure 9. It should be noted that unlike the KE07 active site, this theozyme places the carboxylate base such that both oxygen atoms can interact with the transi-

tion state. This catalytic arrangement for the Kemp elimination has been calculated to be the most efficient using quantum mechanics.³¹

Other than the use of a new CAM, the SABER search was performed in the same manner as with the KE07-based CAM. The SABER analysis resulted in 2603 predesigns with an RMSD ≤ 0.4 Å. Of these 2603 designs, 94 had ActiveSiteScore values > 0 . The best predesign candidate was 1TRB, a thioredoxin reductase protein from *E. coli*. This predesign had a GeometryScore of 0.13, an ActiveSiteScore of 1, and a predicted catalytic base pK_a of 5.8, with Glu160 serving as the proposed catalytic base and Ser138 as the H-bond donor. As with the KE07-based design, the RosettaDesign software was used to redesign the FAD-binding site of 1TRB to accommodate the Kemp elimination transition state. The results of this enzyme redesign are shown in Figure 10. The 1TRB predesign required 6 mutations in the FAD-binding site region of the protein in order to accommodate the new transition state.

Discussion

The PDB now contains over 70,000 structures. This represents an enormous wealth of data to draw upon when pursuing new enzyme designs. With SABER, it is possible to screen the entire PDB for geometric matches to a CAM, and also to identify the predesigns that are part of known active sites. This program provides a method for the rapid identification of scaffolds that have catalytic residues placed in a theozyme-like geometry within an active site. Since there will be no need to position the catalytic residues and make mutations to hold them in place, this approach has the potential to significantly reduce the number of mutations required to modify a scaffold to carry out a new reaction.

In this article, we have demonstrated a new technique, catalytic atom mapping, for selecting active sites for rational, minimalistic redesign. We have validated this technique using a known active site redesign example in the literature, the conversion of MLE and AEE into OSBSs. SABER was able to correctly identify the ideal scaffolds for redesign from a very large dataset. In addition, we have used the catalytic atom mapping methodology to identify active sites that could be remodeled to match that of a computationally design enzyme for a Kemp elimination, KE07. We have demonstrated that designs based on the active sites identified by SABER are

Figure 7. Superposition of the 1X0L active site residues identified by SABER and the KE07 designed active site. The KE07 design is shown in green, while the 1X0L residues are shown in blue. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Figure 8. The active site of 1X0L, redesigned to catalyze the Kemp elimination. The catalytic residues and transition state are shown in yellow. The catalytic base (Asp204) and the hydrogen bond donor (Lys171) are unchanged from the original isocitrate dehydrogenase protein scaffold. A: The original 1X0L active site, with the residues that clash with the Kemp elimination transition state shown in magenta. B: The 1X0L active site after redesign, showing the 9 mutations from RosettaDesign in magenta. Mutations: R118G, A169G, V203G, T222G, L225T, L226G, D228G, I229G, L230A (1X0L numbering). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

complementary to the designs based on the standard Rosetta method. Studies are currently in progress to synthesize and evaluate enzymes designed using this methodology.

Materials and Methods: Computational Details

SABER overview

Each section describes an element of the SABER program and how it integrates into the overall methodology for searching the PDB for active sites compatible with new reactions. Each protein identified during the search process is called a “predesign,” to indicate a protein structure with appropriately placed catalytic groups, but requiring additional modification to the active site to accommodate a new transition state. An

example of the output of a typical SABER redesign is listed in Table III and described later.

Search protocol

This method first involves the creation of a CAM from a 3D geometrical arrangement of atoms determined from a quantum mechanical model or from an enzyme crystal structure. The CAM is then used to scan protein structures in the PDB for similar arrangements of atoms.

A CAM is set of coordinates of atoms that are involved in the chemistry of the reaction, termed “catalytic atoms”; this can also involve other atoms that orient the position of the catalytic atoms. For example, in a serine protease, the minimal catalytic atom definition of the serine nucleophile is the γ -oxygen atom of the alcohol in the side chain. To fix the orientation of the γ -oxygen, the β -carbon can also be included. The remainder of the triad could be defined using the ring nitrogens of the histidine residue and one or both of the carboxylate oxygens from the acid residues. A more extended CAM would include the oxyanion hole nitrogens of the NH hydrogen bond donors. The CAM can be based on a known active site geometry from experiment, or on a theoretical model of an active site, such as a theozyme.⁷

CAM parameters

The CAM has a number of parameters that can be set for each atom. In addition to the geometric constraints, each atom in the CAM can be specified to match with only certain residue and/or atom types during the search. For example, a catalytic atom defined by the trypsin γ -oxygen from serine can be set to match with only other serine γ -oxygens, or can be set to match with either a serine γ -oxygen, or another potentially nucleophilic oxygen, or even a cysteine γ -sulfur. The latter case would be useful if

Figure 9. The theozyme used in the SABER search to locate predesigns with a more ideal geometry for catalysis of the Kemp elimination. Note the alternative arrangement of the catalytic base vs. that shown in Figure 5. The atoms used in the catalytic atom map are shown as spheres. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figure 10. The active site of 1TRB, redesigned to catalyze the Kemp elimination. The catalytic residues and transition state are shown in yellow. The catalytic base (Glu160) and the hydrogen bond donor (Ser138) are unchanged from the original thioredoxin reductase protein scaffold. A: The original 1TRB active site, with the residues that clash with the Kemp elimination transition state shown in magenta. B: The 1TRB active site after redesign, showing the mutations from RosettaDesign in magenta. Mutations: C135G, A136T, T137G, D139H, Y163G, Q294T (1TRB numbering). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

the CAM were intended to identify both serine and cysteine proteases. The radius of each atom in the map is set individually, making it possible to set a different tolerance for geometric deviation from the CAM for individual atoms.

Once the CAM is created, the Jess algorithm is used to search the PDB via geometric superposition.³² The Jess algorithm searches through the protein structures one by one, finding 3D matches to the arrangement of atoms, defined by a CAM. Jess uses a geometric hashing function to maximize the superposition of the CAM and the atoms in the proteins being searched.³² It is possible to have multiple matches to the CAM in a single protein structure. The Jess output consists of the atoms involved in the match, as well as the RMSD of the match to the CAM and the maximum deviation from the CAM.

A Jess search of the entire PDB will typically yield thousands to hundreds of thousands of geometric matches to the CAM. This is especially true when loose tolerances are set for the catalytic atoms or when a

small number of catalytic atoms are used. The additional steps in SABER identify the geometric matches that are most likely to be of interest, that is, the matches that are in known binding sites or active sites.

ActiveSiteFinder and score

ActiveSiteFinder determines if the identified catalytic residues are part of an active site already contained in the CSA, which lists the known or putative active site residues for a given enzyme.³³ The CSA has been compiled by the Thornton group, and provides a publicly available database of known active site residues for proteins in the PDB. The database contains active site residue information from sequence alignments as well as from experimental data in the literature. If the identified catalytic residues are part of an active site in the CSA, the active site residues from the CSA are compared to the residues that matched the CAM. The ActiveSiteScore is the number of catalytic atoms that are part of a known active site in the CSA.

Table III. An Example of a Predesign Generated by SABER, Along with the Scoring Functions used to Identify Optimal Active Sites for Redesign

Feature	Description	Example
PDB	PDB code for protein that contains matching residues	1EUA
RMSD	RMSD vs. catalytic atom map (Å)	0.44
Maximum displacement	Largest individual deviation from catalytic atom map (Å)	0.69
Protein name	Name listed in the Protein Data Bank	KDPG aldolase
Match residues	The residues identified in the protein as matching the catalytic atom map	GluA45, LysA133
ActiveSiteScore	Number of residues matching the catalytic atom map that are listed as active site residues in the Catalytic Site Atlas entry for the protein	2
BindingSiteFinder	The closest PDB-style heteroatom within 5 Å of the residues matching the CAM, along with the distance between the two atoms	Lys 133 NZ Pyr 2103 C2 1.23 Å
pK _a	PROPKA pK _a values for the match residues. Also gives the PROPKA classification for each residue as buried or solvent-exposed.	GluA45, buried. pK _a = 6.0 LysA133, buried. pK _a = 9.0

For example, if the residues identified by the CAM are: Asp100 Lys120 His140, and the known active site residues in the CSA are: Asp100 Lys120 Glu160 His180, the ActiveSiteScore is 2. Two of the residues in the predesign, Asp100 and Lys120, have been annotated as part of an active site in the CSA. In the case of proteins for which no active site data are available, the score is 0.

BindingSiteFinder

BindingSiteFinder is an alternative method for locating active/binding sites in proteins that have not been identified in the CSA. This program analyzes the PDB file for every predesign and searches for nonwater PDB heteroatoms within 5 Å of the residues matching the CAM. If any such heteroatoms are present, BindingSiteFinder will display the name of the ligand, as well as the identity of the heteroatom closest to the residues identified by the CAM. The idea is that these heteroatoms will be from inhibitors, substrate analogs, or cofactors present in the crystal structure, and this will show that the identified site is capable of binding a substrate.

Prediction of pK_a values

PROPKA, from the Jensen group, is used to predict pK_a values for all of the residues identified as matching the CAM.^{34,35} When acid–base catalysis is used, this gives information about the protonation states of the proposed catalytic groups.

References

- Kaplan J, DeGrado WF (2004) De novo design of catalytic proteins. *Proc Natl Acad Sci USA* 101: 11566–11570.
- Nanda V, Rosenblatt MM, Osyczka A, Kono H, Gethahun Z, Dutton PL, Saven JG, DeGrado WF (2005) De novo design of a redox-active minimal rubredoxin mimic. *J Am Chem Soc* 127:5804–5805.
- Rothlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS, Baker D (2008) Kemp elimination catalysts by computational enzyme design. *Nature* 453:190–195.
- Jiang L, Althoff EA, Clemente FR, Doyle L, Rothlisberger D, Zanghellini A, Gallaher JL, Betker J, Tanaka F, Barbas CF III, Hilvert D, Houk KN, Stoddard BL, Baker D (2008) De novo computational design of retro-aldol enzymes. *Science* 319:1387–1391.
- Yeung N, Lin YW, Gao YG, Zhao X, Russell BS, Lei L, Miner KD, Robinson H, Lu Y (2009) Rational design of a structural and functional nitric oxide reductase. *Nature* 462:1079–1082.
- Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, St Clair JL, Gallaher JL, Hilvert D, Gelb MH, Stoddard BL, Houk KN, Michael FE, Baker D (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* 329:309–313.
- Tantillo DJ, Chen J, Houk KN (1998) Theozymes and compuzymes: theoretical models for biological catalysis. *Curr Opin Chem Biol* 2:743–750.
- Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein structure prediction using rosetta. *Method Enzymol* 383:66–93.
- Zanghellini A, Jiang L, Wollacott AM, Cheng G, Meiler J, Althoff EA, Rothlisberger D, Baker D (2006) New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci* 15:2785–2794.
- Fersht A (1999) Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding, Vol. XXI. New York: W.H. Freeman, 631 p.
- Warshel A (1978) Energetics of enzyme catalysis. *Proc Natl Acad Sci USA* 75:5250–5254.
- Warshel A (2003) Computer simulations of enzyme catalysis: methods, progress, and insights. *Annu Rev Biophys Biomol Struct* 32:425–443.
- Warshel A, Sharma PK, Kato M, Xiang Y, Liu H, Olson MH (2006) Electrostatic basis for enzyme catalysis. *Chem Rev* 106:3210–3235.
- Seebeck FP, Hilvert D (2005) Positional ordering of reacting groups contributes significantly to the efficiency of proton transfer at an antibody active site. *J Am Chem Soc* 127:1307–1312.
- Smith AJ, Muller R, Toscano MD, Kast P, Hellinga HW, Hilvert D, Houk KN (2008) Structural reorganization and preorganization in enzyme active sites: comparisons of experimental and theoretically ideal active site geometries in the multistep serine esterase reaction cycle. *J Am Chem Soc* 130:15361–15373.
- Dechancie J, Clemente FR, Smith AJ, Gunaydin H, Zhao YL, Zhang X, Houk KN (2007) How similar are enzyme active site geometries derived from quantum mechanical theozymes to crystal structures of enzyme-inhibitor complexes? Implications for enzyme design. *Protein Sci* 16:1851–1866.
- Lassila JK, Keeffe JR, Kast P, Mayo SL (2007) Exhaustive mutagenesis of six secondary active-site residues in *Escherichia coli* chorismate mutase shows the importance of hydrophobic side chains and a helix N-capping position for stability and catalysis. *Biochemistry* 46: 6883–6891.
- Ruscio JZ, Kohn JE, Ball KA, Head-Gordon T (2009) The influence of protein dynamics on the success of computational enzyme design. *J Am Chem Soc* 131: 14111–14115.
- Olsson MH, Parson WW, Warshel A (2006) Dynamical contributions to enzyme catalysis: critical tests of a popular hypothesis. *Chem Rev* 106:1737–1756.
- Nashine VC, Hammes-Schiffer S, Benkovic SJ (2010) Coupled motions in enzyme catalysis. *Curr Opin Chem Biol* 14:644–651.
- Kamerlin SC, Warshel A (2010) At the dawn of the 21st century: is dynamics the missing link for understanding enzyme catalysis? *Proteins* 78:1339–1375.
- Boekelheide N, Salomon-Ferrer R, Miller TF III (2011) Dynamics and dissipation in enzyme catalysis. *Proc Natl Acad Sci USA* 108:16159–16163.
- Toscano MD, Woycechowsky KJ, Hilvert D (2007) Minimalist active-site redesign: teaching old enzymes new tricks. *Angew Chem Int Ed Engl* 46:3212–3236.
- Gerlt JA, Babbitt PC (2009) Enzyme (re)design: Lessons from natural evolution and computation. *Curr Opin Chem Biol* 13:10–18.
- Vizcarra CL, Zhang N, Marshall SA, Wingreen NS, Zeng C, Mayo SL (2008) An improved pairwise decomposable finite-difference Poisson-Boltzmann method for computational protein design. *J Comput Chem* 29: 1153–1162.

26. Brustad EM, Arnold FH (2011) Optimizing non-natural protein function with directed evolution. *Curr Opin Chem Biol* 15:201–210.
27. Hellinga HW, Richards FM (1991) Construction of new ligand binding sites in proteins of known structure. I. Computer-aided modeling of sites with pre-defined geometry. *J Mol Biol* 222:763–785.
28. Schmidt DM, Mundorff EC, Dojka M, Bermudez E, Ness JE, Govindarajan S, Babbitt PC, Minshull J, Gerlt JA (2003) Evolutionary potential of (beta/alpha)₈-barrels: Functional promiscuity produced by single substitutions in the enolase superfamily. *Biochemistry* 42: 8387–8393.
29. Vick JE, Schmidt DM, Gerlt JA (2005) Evolutionary potential of (beta/alpha)₈-barrels: in vitro enhancement of a “new” reaction in the enolase superfamily. *Biochemistry* 44:11722–11729.
30. Vick JE, Gerlt JA (2007) Evolutionary potential of (beta/alpha)₈-barrels: stepwise evolution of a “new” reaction in the enolase superfamily. *Biochemistry* 46: 14589–14597.
31. Na J, Houk KN, Hilvert D (1996) Transition state of the base-promoted ring-opening of isoxazoles. Theoretical prediction of catalytic functionalities and design of haptens for antibody production. *J Am Chem Soc* 118: 6462–6471.
32. Barker JA, Thornton JM (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics* 19:1644–1649.
33. Porter CT, Bartlett GJ, Thornton JM (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* 32:D129–D133.
34. Bas DC, Rogers DM, Jensen JH (2008) Very fast prediction and rationalization of pK(a) values for protein-ligand complexes. *Proteins: Struct Funct Bioinformatics* 73:765–783.
35. Olsson MHM, Sondergaard CR, Rostkowski M, Jensen JH (2011) PROPKA3: consistent treatment of internal and surface residues in empirical pK(a) predictions. *J Chem Theory Comput* 7:525–537.