Chapter IV

The computational design and molecular dynamics simulation analysis of multiple Kemp elimination enzymes

All of the experimental work described in this chapter was performed in the Mayo laboratory. Gert Kiss of the Houk lab (UCLA) carried out the molecular dynamics simulations.

Abstract

Crystallographic and molecular dynamics (MD) simulation analysis of our inactive Kemp elimination (KE) design HG-1 suggested that possible reasons for the inactivity lay in the flexible, solvent-exposed active site. In an effort to construct a more hydrophobic, rigid active site in the same protein scaffold, we placed the catalytic residues farther into the barrel of the protein, resulting in an active KE enzyme. Using similar computational methods, three other active KE designs were also identified in two scaffolds that had been previously shown to be amenable to computational design procedures for this reaction. We also show that MD analysis of these designs is able to accurately differentiate between active and inactive designs, suggesting a novel *in silico* screening step for the computational enzyme design process.

Introduction

Most of the expense and time associated with computational enzyme design is in the experimental evaluation of designs predicted by the software. After a protein sequence is predicted to have activity for a given reaction, the corresponding DNA sequence must be ordered from a commercial source or constructed from scratch, the heterologous expression of the protein must be optimized and a purification scheme must be developed. Each of these steps has the potential for unforeseen problems unique to the system of interest, which may take weeks to overcome.[1] Indeed, much work has been dedicated to overcoming challenges associated with heterologous protein expression including optimizing gene sequences, using cell-free extracts, and employing eukaryotic expression systems.[1,2]

Before activity assays can be carried out, the substrate must be purchased or chemically synthesized and purified. Only then can the expressed protein can be assayed for activity and stability and characterized structurally. At a minimum, the entire process starting from the computationally predicted protein sequence can take a month and cost thousands of dollars in synthetic genes, oligonucleotides, and reagents.

Recent successes in *de novo* computational enzyme design have relied in part on the synthesis and screening of a large number of putative enzymes.[3,4] Although these methods did result in active enzymes, the time and costs associated with their synthesis and screening make the process extremely inefficient and point to inaccuracies in the design process. *De novo* enzyme design with our protein design software, ORBIT, and the recently developed software called Phoenix, has also been challenging (see Chapter

III and Appendix A), indicating that we also have undetermined problems in our design procedure.

Two strategies could help us address the inefficiencies in computational enzyme design: (1) Use the protein design cycle, mentioned in Chapter I and Chapter III, to analyze the inactive or less active designs and modify the design procedure accordingly. (2) Identify sequences from the design process that are likely to be inactive so that experimental time and money are focused on active enzymes.

Here, we attempt to use both strategies in the context of the Kemp elimination (KE). First, we use information from the analysis of a previous inactive KE design to inform the direction of new designs for the KE in the *Thermoascus aurantiacus* xylanase (TAX) scaffold. This cyclic design procedure resulted in a highly active KE enzyme that appears to accelerate the reaction with a 10^6-fold rate increase.

Second, we created a small set of KE enzymes, which were analyzed by a new MD simulation procedure to observe dynamics of the transition state (TS) structure, explicit solvent, and the catalytic residues within the active site. These analyses correctly predicted the presence or absence of activity in most of the designs. A method of this type could serve as an initial screen to help us eliminate false-positives from the design process.

The inability of our design procedure to reliably predict active sequences and the success of the MD simulations in making that distinction point to several limitations of our design procedure. Both ORBIT and Phoenix use fixed backbone structures to represent the scaffold and neither can currently model dynamics or explicit solvation. However, both of these factors are modeled in MD simulations. The combination of

fixed-backbone computational enzyme design and MD simulation could combine the best features of both procedures to improve the accuracy and efficiency of the computational enzyme design process.

Materials and methods

Active site placement

The protein design software Phoenix, which is similar to ORBIT, was used for these designs.[5,6,7]

For the active site search in design HG-2, all non-Pro positions within 5 Å of D127 in 1GOR that point into the barrel of TAX were designated design positions (residues 17, 42, 44, 81, 83, 129, 170, 172, 209, 234, 236, 237, 265, and 267). These positions were allowed to sample all conformations of Gly, Phe, Trp, Ser, Thr, and Tyr. Position 127 was defined as the only catalytic position and was allowed to sample all conformations of Asp. As in the design of Chapter III, a backbone-independent conformer library was used.[8] The library of TS poses was generated through the targeted ligand placement as discussed in Chapter II and Chapter III using the geometric variations in Table 3-2. The other catalytic contacts were enforced using the geometric constraints described in Chapter III (Table 3-3) except that the additional hydrogen bond contact to the base (Table 3-4) was not required. The catalytic contacts identified in this design were D127, W44, and S265.

The active site placement for the designs in scaffolds 1THF and 1A53 were carried out in a similar manner. The design positions and the required contacts for these

designs are shown in Table 4-1 along with the actual residues chosen as catalytic contacts in the active site search.

Active site repacking

For the repacking calculation, the initial TS position was taken from the active site chosen after the active site search. From this initial position, the TS structure was translated ± 0.4 Å in x, y, and z in 0.2 Å steps and rotated 10° in each direction in 5° steps. The geometric constraints from Tables 3-3 and 3-4 were applied to enforce the contacts between the TS and each of the three catalytic residues identified in the active site search.

For design HG-2, residues 42, 21, 81, 83, 84, 125, 130, 172, 234, 236, and 267 were designated design positions and were allowed to sample all conformations of all residues except Pro and Cys. A second shell of float positions was designated around the design positions (residues 16, 17, 46, 47, 50, 79, 87, 90, 170, 207, 209, 239, 275, and 276), which were allowed to sample all conformations of the wild-type residue at that position. Positions 44, 127, and 265 were allowed to sample all rotamers of Trp, Asp, and Ser, respectively. The residues that were designated as design and float residues for the calculations in scaffolds 1THF and 1A53 are shown in Table 4-1.

An occlusion-based solvation potential was applied with scale-factors of 0.05 for nonpolar burial, 2.5 for nonpolar exposure, and 1.0 for polar burial.[7] Other standard parameters were applied as in Lassila *et al.* and a backbone-independent conformer library was used to represent side chain flexibility.[8] As in the active site search, sidechain-TS interaction energies were biased to favor those contacts that satisfy the

geometries in Table 3-3 and Table 3-4. Sequence optimization was carried out with FASTER,[9,10] and a Monte Carlo-based algorithm[11,12] was used to sample sequences around the minimum energy conformation from FASTER (FMEC).

Gene synthesis and cloning

Genes for the designed proteins were designed and constructed as described in Chapter III. The DNA sequence for a factor Xa cleavage site and a His_6-tag added to the C-terminus of all genes. Closely related sequences (e.g., wild-type TAX, HG-1, and HG-2) were synthesized from a common oligonucleotide set with new oligonucleotides introduced as needed for differences in the gene sequence. Site-directed mutagenesis of the genes to create point mutations is also described in Chapter III.

Protein expression and purification

For initial activity screening, 200 μL LB/ampicillin starter cultures were inoculated with a single BL-21 (DE3) *E. coli* colony and grown at 37°C overnight with shaking. The entire starter culture was used to inoculate 5 mL of Overnight Express Instant TB media (Novagen) with ampicillin in 24-well culture plates (Whatman). The plates were fitted with Bugstopper Venting Capmats (Whatman) and incubated for 3 hours at 37°C with shaking. Expression was continued at 18°C overnight with shaking. The cells were harvested by centrifugation and then washed with 1 mL phosphate buffered saline. Cell pellets were frozen on dry ice for 30 min, then thawed at 4°C and resuspended in 400 μL lysis buffer (50 mM sodium phosphate, pH 8.0, 300 mM NaCl, 2.5 mM imidazole, 1x CelLytic B (Sigma-Aldrich), 1 mg/mL lysozyme (Sigma-Aldrich),

and 1 U Benzonase endonuclease (Merck)). The lysate was centrifuged at 5000 × g for 20 min at 4°C and crude protein samples were taken from the supernatant.

For further purification of 5 mL cultures, His-Select plates (Sigma-Aldrich) were equilibrated with 600 μL of equilibration buffer (50 mM NaPO$_4$ pH 8.0, 300 mM NaCl). The crude supernatant was applied to the plate wells and the columns were washed with 600 μL of wash buffer (50 mM NaPO$_4$ pH 8.0, 300 mM NaCl, 5 mM imidazole). The purified protein was eluted in 500 μL elution buffer (50 mM NaPO$_4$ pH 8.0, 300 mM NaCl, 250 mM imidazole).

For larger-scale production, proteins were expressed in 1 L LB/ampicillin cultures of BL-21 (DE3) *E. coli* as described in Chapter III except that the expression was induced at 18°C for up to 18 hours. The proteins were purified using 1 mL of Ni-NTA resin (Qiagen) and the elution buffer was exchanged with 50 mM sodium citrate, 150 mM NaCl pH 5.5 or 25 mM HEPES, 100 mM NaCl pH 7.25.

Protein concentration determination

Protein concentrations were measured by UV absorbance after protein denaturation in 8 M guanidinium hydrochloride for 10 min with a dilution of at least 10x. The extinction coefficient of each protein at 280 nm was calculated based on the number of Trp and Tyr residues in the sequence (Table 4-2).

Protein characterization

Circular dichroism (CD) and mass spectrometry (MS) were carried out as in Chapter III. CD spectra and thermal denaturation curves were obtained with 10 μM protein in 25 mM HEPES, 100 mM NaCl pH 7.25 or 25 mM MES, 100 mM NaCl pH 5.5.

KE activity screening

5-nitrobenzisoxazole (5-NBZ) was synthesized and purified as described in Chapter III. The reactions were carried out in a total volume of 200 μL in black 96-well microtiter plates with clear bottoms (Greiner) at 27°C. The reaction contained 20 μL of the lysate supernatant and 1 mM 5-NBZ diluted in 25 mM HEPES, 100 mM NaCl pH 7.25. The acetonitrile concentration was 2% for all reactions. The production of the phenolate product (Figure 1-2) was monitored by an increase in absorbance at 380 nm using a Safire[2] microplate reader (Tecan).

Kinetic measurements

Kinetic parameters were determined by monitoring phenolate production at 380 nm ($\varepsilon = 15,800$ cm^{-1}M^{-1})[13] using a Shimadzu UV-1601 spectrophotometer equipped with a temperature-controlled cell holder. A 50 mM stock of 5-nitrobensoxazole was made in acetonitrile and stored at −20°C. Assays were carried out with 5 μM protein in 25 mM HEPES, 100 mM NaCl pH 7.25 at 27°C. Assays contained substrate concentrations between 16 μM and 1 mM and the acetonitrile concentration was kept constant at 2% of the 1 mL reaction volume. Initial reaction rates were determined from the linear portion

of the reaction progress curve and were corrected for the initial rate contribution for the buffer catalyzed (no protein) reaction.

Kaleidagraph was used to fit the data to the Michaelis-Menten equation (Equation-1) to determine the kinetic parameters k_{cat} and K_m:

$$v_0 = \frac{k_{cat} \cdot [E]_T \cdot [S]_0}{K_m + [S]_0}$$
Equation-1

where v_0 is the initial reaction rate, $[E]_T$ is the total enzyme concentration and $[S]_0$ is the initial substrate concentration. In the cases where the enzyme was not saturated because of limited substrate solubility, the data were fit to a line to determine k_{cat}/K_m.

pH profiles

pH profiles were determined in 96-well microtitre plates by monitoring the absorbance increase at 380 nm with a Safire[2] microplate reader (Tecan). The final concentration of protein was 10 μM in a 200 μL total volume. Each reaction was carried out in 25 mM buffer (pH 5.4: MES, pH 6.6 and 7.0: potassium phosphate, pH 7.25: HEPES, pH 8.0 and 9.0: Tris) and 100 mM NaCl. A dilution series of 5-NBZ stocks was made in acetonitrile, resulting in substrate concentrations between 0.39 mM and 50 mM. 4 μL of each substrate stock was added to each reaction well. Initial rates were calculated from the first 400 sec of measurements. The initial rates of the buffer-catalyzed reaction (no protein) were subtracted from the protein-catalyzed initial rates. A pathlength of 0.67 cm was assumed for a volume of 200 μL in the 96-well plates.

Crystallography

Crystallization screens were set up by the Caltech Molecular Observatory. Crystals for HG-2 were obtained in three conditions at 20°C and a protein concentration of 9.5 mg/mL: (1) 100 mM MES, pH 6.5, 1.6 M magnesium sulfate (Figure 4-5C); (2) 100 mM HEPES, pH 7.5, 20 mM magnesium chloride, 22% (w/v) polycrylic acid 510 sodium salt; (3) 100 mM HEPES, pH 7.5, 1.26 M ammonium sulfate. Crystals for HG-2/K50A were obtained at 20°C in 200 mM ammonium sulfate, 30% (w/v) PEG 8000, 9.9 mg/mL protein.

Molecular dynamics (MD) simulations

MD simulations were carried out as described in Chapter III, Materials and Methods.

Results and Discussion

Second-generation designs

The design calculations for the second-generation designs were carried out using Phoenix, a computational protein design software package similar to ORBIT, which has also been developed in the Mayo lab.[5,6] Like ORBIT, Phoenix utilizes geometric-constraint-based methods to generate libraries of ligand poses including targeted ligand placement and rotation/translation;[8] stochastic search methods, including FASTER and Monte Carlo, are used for sequence optimization.[10]

The second-generation designs were carried out in the same scaffold as HG-1 (see Chapter III) with identical geometric constraints. A wild-type carboxylate residue in the TAX scaffold, D127, was identified as being in a promising position to serve as the catalytic base. This residue is deeply buried inside the core of the protein in an area that appeared to have enough room for the TS and other active site residues. The area surrounding D127 is adjacent to the wild-type binding pocket but is located farther into the barrel of the $(\alpha/\beta)_8$ scaffold (Figure 4-1). One major difficulty associated with using an active site that is not located in the natural binding pocket of a scaffold is potential destabilization of the protein by disruption of important interactions in the core to make room for the TS and catalytic residues.

The active site chosen in the active site search consisted of D127 as the general base, W44 as the π-stacking residue and S265 as the hydrogen bond donor (Figure 4-2A). This active site displays good stacking between W44 and the TS, and good hydrogen bond geometry between D127 and the TS H3. In addition, the oxazole ring is pointed into the back of the active site pocket and is well shielded from solvent. The nitro substituant on the TS is solvent-exposed and may be in a position to interact favorably with the wild-type residue K50.

Active site repacking resulted in a sequence that is 12 mutations away from wild-type TAX (Table 4-1, Figure 4-2B). Overall, the residues surrounding the TS and catalytic residues are hydrophobic and are predicted to pack well around the TS structure. In addition to the required polar catalytic contacts, the wild-type residue K50 is predicted to make a favorable interaction with the TS nitro group.

MD analysis of the design, which was done blind to the experimental results, indicated that the active site undergoes a conformational rearrangement that results in two distinct states with active site RMSD values of 1.85 and 2.28 Å from the initial configuration, respectively (Figure 4-3 and Figure 4-4A). This rearrangement results from changes in the secondary structure of the protein due to removal of stabilizing interactions (especially between D127 of β-1 and R81 of β-2) to make room for the active site. As a result of this change in secondary structure, D127 is able to make a favorable hydrogen bond interaction with the backbone carbonyl of G81 about 10 ns into the simulation, drawing it into an alternative conformation that coincides with the transition between the two states.

State 1 is more structurally similar to the original design than state 2 and many of the designed contacts are maintained. In state 2, the substrate has rotated 90° from its initial position, losing the stacking contact with W44 and the electrostatic contact with K50 (Figure 4-3). However, the base-substrate contact is maintained in both states as well as during the transition (Figure 4-4B), due in part to the flexibility of the methionine-rich active site. The designed hydrogen bond contact with S265 is not observed in either state, but the positioning of this residue may force the substrate closer to the base in state 1.

The simulation also shows that the binding site residues pack well around the substrate, preventing the invasion of any explicit solvent molecules into the active site to interact with the base. The lack of water in the active site suggests that the pKa of the base may be appropriately elevated to facilitate abstraction of the substrate proton. In

combination with the well-defined base-substrate contact, these analyses suggest that this design will be active.

A single Ni-NTA affinity chromatography purification step was sufficient to produce about 10 mg of pure HG-2 from 2 liters of culture (Figure 4-5A) and the mass was confirmed with electrospray MS (Figure 4-5B). CD analysis of HG-2 showed that the 12-fold mutant is folded with similar secondary structure to wild-type TAX (Figure 4-6A). However, HG-2 is significantly destabilized with respect to the wild-type protein (Figure 4-6B).

As predicted by the MD simulations, kinetic analysis of HG-2 indicates that the design is an active catalyst for the KE, far above the rate of the buffer catalyzed reaction (Figure 4-7A). The wild-type TAX scaffold was determined to have no activity for the KE. Substrate saturation was never reached for this design, so the kinetic constants k_{cat} and K_m could not be determined reliably. k_{cat}/K_m was determined to be 122 $M^{-1}s^{-1}$ from the slope of the initial rate versus initial substrate concentration plot, making the efficiency of this enzyme comparable to the best KE enzymes designed by Röthlisberger et al.[3] Knockout mutations of the base (D127N) and the hydrogen bond donor (S265A) show a significant decrease in activity compared to HG-2, with the base knockout losing almost all activity (Figure 4-7B). The loss of activity upon mutation of the putative catalytic residues indicates that these residues are, in fact, important for catalysis and support the proposed designed mechanism for this enzyme.

The pKa of the base was observed to be significantly elevated from the solvent-exposed pKa of aspartate (~ 7 versus ~ 4) (Figure 4-8A). The pH profile also showed a significant decrease in activity at high pHs, which may indicate the presence of a second

ionizable group in the active site with a high pKa or simply the unfolding of the protein at high pH. CD analysis ruled out the latter possibility, showing that the protein remains folded at pHs up to 9.0 under the conditions of the assays (25 mM HEPES, 100 mM NaCl pH 7.25, 27°C) (Figure 4-8B). Further investigation into the cause of the decrease in activity of HG-2 at high pH will be necessary.

Third-generation designs

In an attempt to increase activity of HG-2, point-mutants of HG-2 were selected based on the MD simulations. G81A was chosen to rigidify the local secondary structure to prevent the backbone hydrogen bond to D127 (Figure 4-3). However, MD analysis of HG2/G81A indicates that though the secondary structure of the loop may be stabilized somewhat, the base can adopt an alternate conformation, preventing interactions with the substrate. Subsequent experimental analysis of this mutant showed that it was less active than the original design HG-2 (Figure 4-9, Table 4-1).

MD analysis of HG-2 suggested that S265 was acting to push the substrate closer to the base and prevent water from entering the active site. The S265T mutation was chosen because of the slightly larger volume of Thr, which may provide better packing around the substrate. In addition, this mutation also reverts to the wild-type residue at this position. MD analysis of HG2/S265T predicted that this variant would be more active than the original design as the substrate-base interaction is maintained just as well as in HG-2, but no active site rearrangement occurs. Experimental evaluation showed that although this design has a slightly lower k_{cat} than HG-2, the K_m is also lower,

resulting in a k_{cat}/K_m that is about three fold higher than HG-2 (Figure 4-9, Table 4-1) or any of the KE enzymes from Röthlisberger *et al.*[3]

Additional designs

With one successful KE design in hand, we tested Phoenix's ability to recapitulate the active sites of enzymes with known activity that had been designed by Röthlisberger *et al.*[3] Starting with the scaffolds and base positions for KE07, KE10, and KE59, we applied our enzyme design methods to generate active sites within the binding pockets of *Thermotoga maritima* imidazoleglycerolphosphate synthase (PDB: 1THF)[14] and *Sulfolobus solfataricus* indole-3-glycerolphosphate synthase (PDB: 1A53)[15]. The methods used were the same as those described above for HG-2 except that lysine was allowed to be the hydrogen bond donor catalytic contact to fully access some of the designs.

Starting with the scaffold and catalytic base position from KE59 (1A52, E231), we used targeted ligand placement to generate ligand poses and required a π-stacking residue and a hydroxyl hydrogen bond donor in addition to the base contact as defined in Table 3-2. The sequence containing catalytic residues from KE59 (W110, S131, E231) was identified in the top 20 active site sequences from the design calculation. The catalytic residues for this design were predicted to overlay well with the catalytic residues of KE59 (Figure 4-10A). The repacking calculation introduced noncatalytic active site residues that differed from those in KE59, resulting in 1A53-1, an 8-fold mutant from KE59. However, this design showed good stacking interactions and an additional hydrogen bond donor contact to the base (S210) (Figure 4-10B).

MD analysis predicted that this design would be inactive because the catalytic base, E231, rearranges readily to form stable interactions with nearby S210 and S211 preventing productive interaction with the substrate (Figure 4-13A, B). In addition, the active site was not well packed around the substrate, allowing the substrate to adopt alternate binding modes (Figure 4-13C). The inactivity of this design was confirmed by experimental KE activity screening (Figure 4-11).

Starting with the scaffold and catalytic base position from KE10 (1A53, E178), targeted ligand placement was used to generate ligand poses. As seen in KE10, the only contacts required in the active site search were a π-stacking and base contact. Phoenix identified the catalytic residues of KE10 (E178 and W210) through the active site search. In addition, a second π-stacking contact was identified (W110), essentially sandwiching the substrate in the active site (Table 4-1, Figure 4-10C). The active site was repacked around the base and double π-stacking contacts, resulting in 1A53-2, a 12-fold mutant from the wild-type sequence and a 9-fold mutant from KE10 (Table 4-1, Figure 4-10D).

The MD analysis of this design showed that the binding pocket contributes to a significant reorientation of the substrate over the course of the simulation, at first maintaining the base-substrate contact (Figure 4-14A, B, C, E). Towards the end of the simulation, the flexible active site allows water molecules to enter and the substrate diffuse away from the base (Figure 4-14D, F). Because of the sharp distance distribution of the substrate-base contact (Figure 4-14E), the design was predicted to have some activity. However, it was not expected to be highly active because of the late intrusion of solvent and diffusion of the substrate. This design was determined to be active by experimental KE activity screening (Table 4-3, Figure 4-9).

In addition to the active site of 1A53-2 found in the KE10 recapitulation calculation, an alternate active site was identified with a base at E157 and π-stacking contacts at W110 and W210 (Table 4-1, Figure 4-10E). Active site repacking resulted in the design 1A53-3, which is an 11-fold mutant from the wild-type scaffold (Figure 4-10F).

MD analysis predicted 1A53-3 to be active because the active site configuration is maintained over the course of the simulation (Figure 4-15). Soon after the simulation begins, substrate rotates slightly from the initial configuration, but the base maintains contact with both the substrate and the π-stacking (W209) residue. In addition, only one water molecule is able to enter the active site and interact with the base (Figure 4-15B). Experimental evaluation of 1A53-3 confirmed the MD prediction of activity for this design (Table 4-3, Figure 4-9).

Two strategies were implemented in the recapitulation of KE07. In the first, a base, π-stacking residue, and a Lys hydrogen bond donor are required. However, Röthlisberger *et al.* found that the removal of the Lys hydrogen bond donor from the active site of KE07 actually increased activity because this residue may interact directly with the base, lowering the pKa.[3] Thus, in the second strategy, we required a hydroxyl hydrogen bond donor instead of the Lys. Starting with the base E101 in the scaffold 1THF, Phoenix identified W50 and K222 as the catalytic residues in the first calculation (Figure 4-12A) and W50 and S201 as the catalytic residues in the hydroxyl hydrogen bond donor case (Figure 4-12C). In both of these designs, W128 is identified as an additional π-stacking contact. Repacking of the E101/W50/W128/K222 active site resulted in the design 1THF-1, which is a 13-fold mutant from wild type and a 9-fold

mutant from KE07. In this design, the base is anchored by S78 and the substrate is held in place by the two Trps (Figure 4-12B). Repacking the E101/W50/W128/S201 active site resulted in the design 1THF-2, which is a 10-fold mutant from the wild type and a 10-fold mutant from KE07.

MD analysis of 1THF-1 indicated that the active site maintains its overall configuration with respect to the initial structure during the simulation (Figure 4-16A, B). The base-substrate contact is less well maintained than the previously described active designs (Figure 4-16C) because the base alternates its interactions between the substrate and solvent molecules in the pocket. As in the crystal structure of KE07, K222 appears to interact directly with the base, perhaps lowering the pKa.[3] Because the base-substrate distance seemed appropriate for catalysis, this design was predicted to be active, although less active than KE07. For 1THF-2, the overall configuration is the same as the initial structure (Figure 4-17A, B), and it also has an extremely sharp base-substrate distance distribution that is characteristic of previous active designs (Figure 4-17C). The simulation indicated that the substrate moved about 2 Å deeper into the pocket than was predicted in the design, preventing S201 from maintaining its contact to the substrate. However, in multiple KE designs, the hydrogen bond donor contact has been shown not to be critical for catalysis.[3] As a result, this design was predicted to be more active than 1THF-1. Experimental evaluation could not detect any activity in 1THF-1 but substantial activity was detected in 1THF-2 (Table 4-3, Figure 4-9), agreeing with the MD prediction.

We were able to recapitulate the placement of the active site residues in all three of the previously designed KE catalysts. However, the final sequences that were

generated by Phoenix were very different from the existing active designs. The sequences of the active enzymes found by Röthlisberger *et al.* are not necessarily the optimal sequences with respect to activity; each is just one of the many possible active sequences for a given set of catalytic residues. Despite our repacked sequences being at least eight mutations away from the active design, 3 out of 5 of our designs showed significant KE activity. In some cases, our sequences were more active than the corresponding sequence that we were trying to recapitulate and in other cases, our sequences were less active, or lacked activity completely. Starting with the initial sequence from computational enzyme design, it has been previously demonstrated that KE activity can be significantly improved through many rounds of directed evolution;[3] however, we have shown here that activity can also be optimized by additional rounds of design based on structural or MD analysis. In both cases, computational enzyme design served to generate an active starting point for further optimization by arranging critical catalytic residues in productive orientations in an environment amenable to the chemical reaction. This initial step is necessary for directed evolution, which generally cannot generate enzymatic activity *de novo*.[16]

MD analysis versus experiment

When compared with experimental results, MD analysis successfully predicted activity in HG-2, 1THF-2, 1A53-2, and 1A53-3, as well the inactivity of 1A53-1 (Table 4-4). In addition, the MD analysis was able to predict relative levels of activity in point mutants of HG-2. In the case of KE enzymes, the main criteria for prediction of activity versus inactivity were solvent inaccessibility of the active site, stable binding of the

substrate, and a stable substrate-base contact within hydrogen bonding distance. The success of these criteria in predicting activity agrees with the original work of Kemp *et al.* on the decomposition of benzisoxazoles, which shows that nonpolar solvents enhance the rate of the reaction significantly relative to water,[13,17] and more recent work, which shows that the nonpolar microenvironment of the binding pocket and the positioning of the base in the catalytic antibody and serum albumins are important.[18,19]

The criteria that are used to predict activity will likely vary for other reaction types and will help to elucidate the subtle requirements for a successful catalyst. However, determination of the best criteria for MD analysis may not be straightforward, especially for reactions that have been less well studied; MD evaluation of positive and negative controls for similar reactions will likely be necessary.

Conclusions

In this work, we used Phoenix to successfully design four active KE enzymes in three different inert scaffolds, demonstrating the applicability of Phoenix to *de novo* computational enzyme design. Of these enzymes, HG-2/S265T showed a catalytic efficiency that was 3 times higher than any other computationally designed KE enzyme. The success of the Phoenix KE designs is due, in large part, to iterative analysis and redesign of inactive enzymes as a part of the protein design cycle discussed in Chapters I and III. The analysis of inactive designs provides valuable information as to the deficiencies of our design procedure, allowing us to correct these problems to produce active enzymes.

We also show that MD simulations of these enzymes can predict the presence or absence of activity and, in the case of point mutants of HG-2, the relative level of activity can also be predicted. Once the MD methods are optimized and fully automated, these types of analyses could be used as a pre-screen to help eliminate any false-positives that may result from the design process, allowing us to focus our experimental efforts on designs that are most likely to be active. That the results of our computational protein design procedure can be assisted by the addition of MD analysis points to three major deficiencies in our design procedure: the reliance on a fixed backbone scaffold, discrete sidechain rotamers, and implicit solvation, all of which help make our design calculations tractable. The MD analysis of computationally designed enzymes can serve as an immediate solution to these problems until new methods that address these deficiencies are fully integrated.

While the activity of these Phoenix designs is encouraging, the enzymes described here still do not demonstrate rate accelerations or efficiencies close to those of natural enzymes. In addition, the K_ms of most of the computationally designed enzymes for this reaction so far have been above 1 mM, suggesting that the active sites are not optimized for substrate binding. Crystal structures of the designs described here are currently being pursued to identify possible improvements in the active site structure for better substrate binding in fourth-generation designs. In addition, directed evolution has previously been shown to be an effective supplemental strategy for significantly enhancing both the k_{cat} and K_m of designs with low-level activity.[3]

Even though the complete recapitulation of the active sites of previous KE designs was not possible, the redundancy of the active designs with respect to the base

position and scaffold demonstrate the ability of computational protein design to provide multiple correct answers to *de novo* enzyme design problems, yielding multiple starting points for further optimization by directed evolution. *In vitro* evolution studies are currently underway in the Hilvert lab at the Swiss Federal Institute of Technology to optimize the activity of some of these designs, and the results of these experiments should help us to better understand this model system and the limitations of computational enzyme design.

Acknowledgements

References

1. Gustafsson, C.; Govindarajan, S.; Minshull, J., Codon bias and heterologous protein expression. *Trends Biotechnol.* **2004**, *22*, 346-353.

2. Higgins, S. J.; Hames, B. D., *Protein Expression: A Practical Approach*. Oxford University Press: 1999.

3. Rothlisberger, D.; Khersonsky, O.; Wollacott, A.; Jiang, L.; Dechancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. A.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D.; Baker, D., Kemp elimination catalysts by computational enzyme design. *Nature* **2008**, *453*, 190-195.

4. Jiang, L.; Althoff, E. A.; Clemente, F. R.; Doyle, L.; Röthlisberger, D.; Zanghellini, A.; Gallaher, J. L.; Betker, J. L.; Tanaka, F.; Barbas, C. F.; Hilvert, D.; Houk, K. N.; Stoddard, B. L.; Baker, D., De novo computational design of retro-aldol enzymes. *Science* **2008**, *319*, 1387-1391.

5. Allen, B. D. *Development and validation of optimization methods for the design of protein sequences and combinatorial libraries*. California Institute of Technology: Pasadena, CA, **2009**.

6. Nisthal, A.; Allen, B. D., (*Manuscript in preparation*). **2009**.

7. Chica, R.; Allen, B. D., (*Manuscript in preparation*). **2009**.

8. Lassila, J. K.; Privett, H. K.; Allen, B. D.; Mayo, S. L., Combinatorial methods for small-molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 16710-16715.

9. Desmet, J.; Spriet, J.; Lasters, I., Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins* **2002**, *48*, 31-43.

10. Allen, B. D.; Mayo, S. L., Dramatic performance enhancements for the FASTER optimization algorithm. *J. Comput. Chem.* **2006**, *27*, 1071-1075.

11. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087-1092.

12. Voigt, C. A.; Gordon, D. B.; Mayo, S. L., Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* **2000**, *299*, 789-803.

13. Casey, M.; Kemp, D.; Paul, K.; Cox, D., Physical organic chemistry of benzisoxazoles. I. Mechanism of the base-catalyzed decomposition of benzisoxazoles. *J. Org. Chem.* **1973**, *38*, 2294-2301.

14. Lang, D.; Thoma, R.; Henn-Sax, M.; Sterner, R.; Wilmanns, M., Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. *Science* **2000**, *289*, 1546-1550.

15. Hennig, M.; Darimont, B. D.; Jansonius, J. N.; Kirschner, K., The catalytic mechanism of indole-3-glycerol phosphate synthase: crystal structures of complexes of the enzyme from *Sulfolobus solfataricus* with substrate analogue, substrate, and product. *J. Mol. Biol.* **2002**, *319*, 757-766.

16. Shao, Z.; Arnold, F. H., Engineering new functions and altering existing functions. *Curr. Opin. Struct. Biol.* **1996**, *6*, 513-518.

17. Kemp, D.; Casey, M., Physical organic chemistry of benzisoxazoles. II. Linearity of the Broensted free energy relation for the base-catalyzed decomposition of benzisoxazoles. *J. Am. Chem. Soc.* **1973**, *95*, 6670-6680.

18. Hu, Y.; Houk, K. N.; Kikuchi, K.; Hotta, K.; Hilvert, D., Nonspecific medium effects versus specific group positioning in the antibody and albumin catalysis of the base-promoted ring-opening reactions of benzisoxazoles. *J. Am. Chem. Soc.* **2004**, *126*, 8197-8205.

19. Seebeck, F. P.; Hilvert, D., Positional ordering of reacting groups contributes significantly to the efficiency of proton transfer at an antibody active site. *J. Am. Chem. Soc.* **2005**, *127*, 1307-1312.