## REPORT DOCUMENTATION PAGE

AFRL-SR-AR-TR-04-

0572



| 1. AGENCY USE ONLY *(Leave blank)* | 2. REPORT DATE | 3. REPORT TYPE AND DATES COVERED |
|---|---|---|
| | | 01 May 2003 - 31 Oct 2004 FINAL |

**4. TITLE AND SUBTITLE**
(DARPA) Computation Model Optimization for Enzyme Design Applications

**5. FUNDING NUMBERS**
61101E
P957/00

**6. AUTHOR(S)**
Dr Mayo

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
CALIFORNIA INSTITUTE OF TECHNOLOGY
1201 E CALIFORNIA BLVD
MAIL CODE 202-6
PASADENA CA 91125-0600

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
AFOSR/NE
4015 WILSON BLVD
SUITE 713
ARLINGTON VA 22203

**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**
F49620-03-1-0291

**11. SUPPLEMENTARY NOTES**

**12a. DISTRIBUTION AVAILABILITY STATEMENT**
DISTRIBUTION STATEMENT A: Unlimited

**12b. DISTRIBUTION CODE**

**13. ABSTRACT** *(Maximum 200 words)*

The major accomplishments of this project are the development of a two-body-decomposable electrostatic potential energy function that accurately reproduces continuum electrostatic energies computed using the finite difference Poisson-l3oltzmann (PB) method, and the enhanecment of the activity of the naturally occurring E. coil chorismatc mutase (EcCM) enzyme through computational design. Although the stated milestone of creating a novel chorisxnatc mutase (CM) was not achieved, the enhancement of the underlying computational model through the development of the two-body PB method will facilitate the future design of novel protein catalysts.

| 14. SUBJECT TERMS | 15. NUMBER OF PAGES |
|---|---|
| | 16. PRICE CODE |

| 17. SECURITY CLASSIFICATION OF REPORT | 18. SECURITY CLASSIFICATION OF THIS PAGE | 19. SECURITY CLASSIFICATION OF ABSTRACT | 20. LIMITATION OF ABSTRACT |
|---|---|---|---|
| Unclassified | Unclassified | Unclassified | UL |

03-1-0291

**Protein Design Processes Seedling**

**Final Report for Period 5/03-10/04**

# Computational Model Optimization for Enzyme Design Applications

Stephen L. Mayo[1], Leslie F. Greengard[2], and Barry H. Honig[3]

[1]Divisions of Biology and Chemistry and Howard Hughes Medical Institute, California Institute of Technology, Pasadena, CA

[2]Department of Mathematics and Computer Science, Courant Institute of New York University, New York, NY

[3]Department of Biochemistry and Molecular Biophysics and Howard Hughes Medical Institute, Columbia University, New York, NY

## 1. Summary:

The major accomplishments of this project are the development of a two-body-decomposable electrostatic potential energy function that accurately reproduces continuum electrostatic energies computed using the finite difference Poisson-Boltzmann (PB) method, and the enhancement of the activity of the naturally occurring *E. coli* chorismate mutase (EcCM) enzyme through computational design. Although the stated milestone of creating a novel chorismate mutase (CM) was not achieved, the enhancement of the underlying computational model through the development of the two-body PB method will facilitate the future design of novel protein catalysts.

## 2. Two-body Poisson-Boltzmann Method:

Protein design is an exceptionally difficult problem characterized by unique complications. Necessary restrictions such as a fixed protein backbone, discrete side chain conformations (rotamers), and limitation to two-body decomposable potential functions require different considerations of structure/energy relationships than other fields of protein simulation. Until now, damped Coulombic potentials as well as empirical surface area and volume scaling functions have been used to include electrostatic solvation energy in computational protein design. These methods have allowed for the successful design of stable proteins but have been a limiting factor in the rational design of enzymatic activity and molecular recognition, for which polar and charged amino acids are key. To bring protein design energy functions up to date with these new challenges, we have been investigating more sophisticated continuum models for electrostatic solvation. Two related obstacles to improving electrostatic solvation energy functions are the combinatorial explosion in protein design, which requires energy scores for many side chains and pairs of side chains and therefore a very fast energy solver, and the need to

calculate energies in one-body (single side chain) and two-body (pairs of side chains) terms without any knowledge of the rest of the structure. We proposed to use fast perturbation methods for two-body terms, allowing for the computationally lengthy numerical solution to the Poisson-Boltzmann (PB) equation for a large number of side chain pairs. In addition we are investigating the speed and accuracy of various analytical Generalized Born methods. Coupled with strategies for approximating a molecular surface during the design calculation, both of these approaches should allow us to more accurately describe the energy of a protein's charge distribution in the context of its molecular geometry and surrounding solvent. Such improvements in the electrostatic solvation energy model for protein design will have a significant impact in the areas of enzyme design and molecular recognition.

Work on pair-wise decomposable PB calculations has gone well. For example, comparisons of side chain desolvation energies, side chain/backbone screened Coulombic energies, and side chain/side chain screened Coulombic energies computed using full molecular surfaces (X-axis) and a two-body method (Y-axis) show excellent correlations and RMSD errors of 0.31, 0.17, and 0.05 kcal/mol, respectively (Figure 1). These data are from a 10-protein test set; note that most of the points cluster very near the X=Y line such that mainly the outliers are visible.

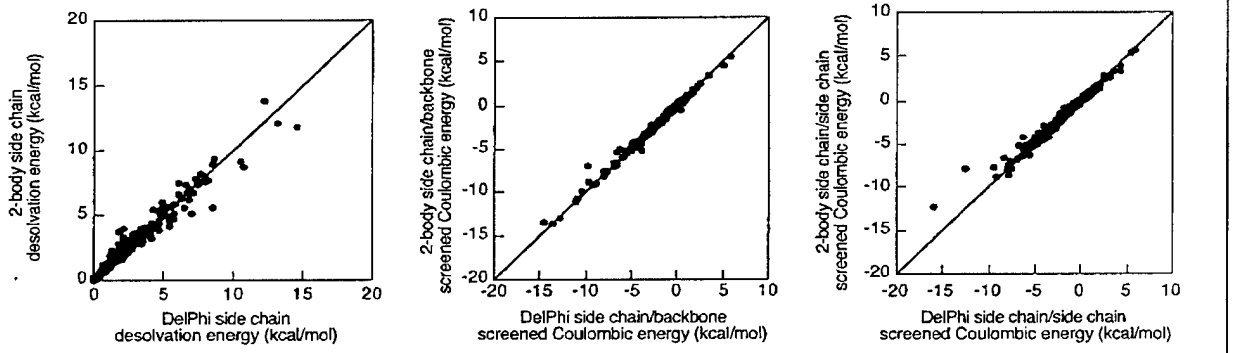



Figure 1. Comparison of exact and two-body PB calculations.

In this case, the two-body method uses a variation of the surface area work of Street and Mayo (*Folding and Design*, 1998) and Wingreen and coworkers (Zhang et al., *Proteins: Structure Function and Bioinformatics*, 2004) where three atom generic side chains are used as surrogates for the actual amino acids. Two-body perturbations are then applied to reconstruct total side chain desolvations.

Exact (i.e., N-body) side chain desolvation energies are computed as illustrated in Figure 2. An unfolded state (reference) solvation energy is computed for a given side chain (rotamer) by charging that side chain and using the side chain and its local backbone to define the protein/solvent interface (Figure 2B). The folded state solvation energy is computed by charging the side chain of interest and defining the protein/solvent interface by using all of the atoms of the protein (Figure 2A). The desolvation energy is then the difference between the folded state and unfolded state solvation energies. Exact side chain/backbone and side chain/side chain screened Coulombic energies are computed in a similar fashion.

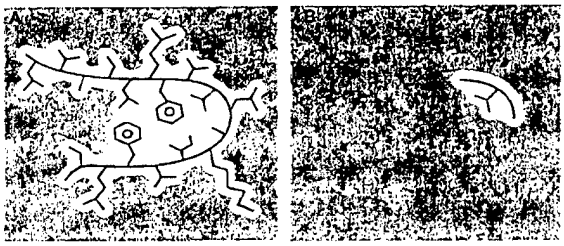


Figure 2. Exact side chain desolvation. The red side chain is the side chain of interest. The white/blue boundary defines the protein/solvent interface.

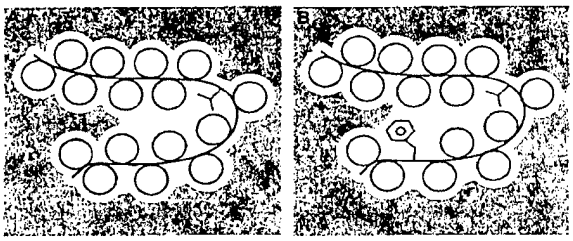


Figure 3. Two-body side chain desolvation. The red side chain is the side chain of interest. The white/blue boundary defines the protein/solvent interface.

For computing two-body side chain desolvation, a one-body folded state solvation energy is computed first by charging only the side chain of interest and using the following components to define the protein/solvent interface: the full protein backbone, the side chain of interest, and generic side chains at all other positions (Figure 3A). The solvation energy for the unfolded reference state is computed as above (Figure 2B).

Two-body perturbations to the one-body solvation energy are then computed by substituting real side chains (rotamers) in turn for each of the generic side chains (Figure 3B) and by subtracting the one-body energy. The total desolvation energy for a side chain by this two-body approach is the sum of the one-body energy and the two-body perturbations minus the solvation energy of the reference state.

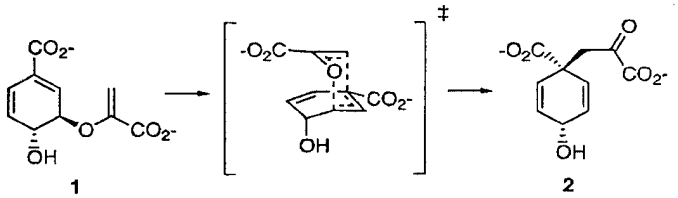
Two-body side chain/backbone screened Coulombic energies are computed in a similar fashion. Two-body side chain/side chain screened Coulombic interactions are computed by charging both side chains and using the full protein backbone, both side chains of interest, and generic side chains at all other positions to define the protein/solvent interface. Achieving additional accuracy for side chain/side chain screened Coulombic interactions would require three-body terms that are disallowed in all significant protein design sequence optimization protocols.

As can be readily appreciated, the time required to compute the two-body perturbations dominates the calculation. Performance improvements that allow this method to be directly used in design calculations are currently being pursued by the Greengard lab at NYU. The general approach will be to recast the PB calculation in terms of a mesh and to utilize the Sherman-Morrison-Woodbury formula to compute the necessary two-body perturbations.

**3. Enhanced Chorismate Mutase:**

The Claisen rearrangement of chorismate (**1**) to prephenate (**2**) is a rare enzyme-catalyzed pericyclic reaction that proceeds through the same mechanism uncatalyzed in solution. Chorismate mutases from various organisms provide rate enhancements of around $10^6$ despite strong

***Scheme 1. Chorismate mutase reaction***



dissimilarity in the X-ray diffraction structures solved to date. The metabolic importance of chorismate as the key branch point in the shikimate pathway has prompted extensive experimental investigation of the chorismate-prephenate rearrangement since the 1960s and has driven complementation experiments to probe the structural determinants of enzyme catalysis. The concerted, unimolecular nature of the rearrangement and the lack of covalent protein interactions have encouraged numerous theoretical studies of the catalyzed and uncatalyzed reactions. In addition, a catalytic antibody showing a rate acceleration ($k_{cat}/k_{uncat}$) of $10^2$ has been isolated. Still, the question of how chorismate mutases achieve rate enhancement is actively debated.

The target objective of our seedling proposal was the recapitulation of CM activity in an *E. coli* periplasmic binding protein. To date, no novel enzymes have been generated, but several computationally designed variants of the wild-type EcCM have produced interesting results. These calculations utilized a QM-derived transition state (TS) structure, which was docked into the known active site by overlaying it with the position of a TS analog in the EcCM crystal structure. A limited rotation/translation search of the TS structure and consideration of 18 residues in the active site region were included. Three designed variants show catalytic efficiency ($k_{cat}/K_M$) at or above the level of the wild-type enzyme. One variant, Ala32Ser (Figure 4), showed catalytic efficiency 60% greater than the wild-type enzyme demonstrating, at a minimum, our ability to perform successful amino acid designs using TS structures in the context of a protein binding site.

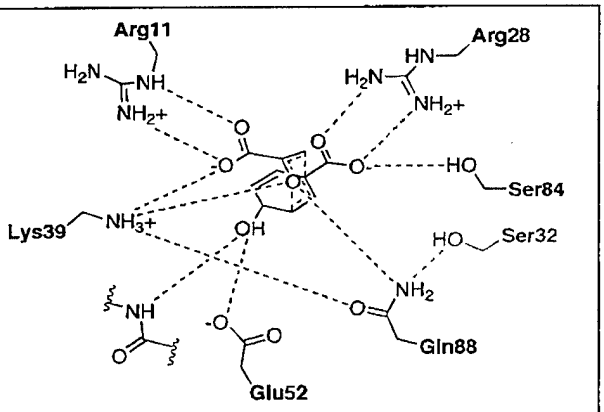


Figure 4. Designed Ala32Ser mutation (red) in *E. coli* chorismate mutase leading to enhanced activity.

**4. Publications:**

Marshall, S.A., Vizcarra C., and Mayo, S.L. 2004. "Electrostatic Models for Protein Design Calculations II: One and Two-Body Decomposable Poisson-Boltzmann Methods." *Protein Science*, submitted.

Lassila, J.K., Keeffe, J.R., Oelschlaeger, P., and Mayo, S.L. 2004. "A Computationally Designed Variant of *Escherichia coli* Chorismate Mutase Shows Enhanced Catalytic Efficiency." *J. Am. Chem. Soc.*, submitted.