



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Quantum mechanical approaches to in silico enzyme characterization and drug design

J. P. Nilmeier, J. L. Fattebert, M. P. Jacobson, C. Kalyanaraman

January 18, 2012

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Quantum mechanical approaches to *in silico* enzyme characterization and drug design

The Big Picture

The astonishing, exponentially increasing rates of genome sequencing has led to one of the most significant challenges for the biological and computational sciences in the 21st century: assigning the likely functions of the encoded proteins. Enzymes represent a particular challenge, and a critical one, because the universe of enzymes is likely to contain many novel functions that may be useful for synthetic biology, or as drug targets. Current approaches to protein annotation are largely based on bioinformatics. At the simplest level, this annotation involves transferring the annotations of characterized enzymes to related sequences. In practice, however, there is no simple, sequence based criterion for transferring annotations, and bioinformatics alone cannot propose new enzymatic functions.

Structure-based computational methods have the potential to address these limitations, by identifying potential substrates of enzymes, as we and others have shown[1]. One successful approach has used *in silico* 'docking' methods, more commonly applied in structure-based drug design, to identify possible metabolite substrates. A major limitation of this approach is that it only considers substrate binding, and does not directly assess the potential of the enzyme to catalyze a particular reaction using a particular substrate. That is, substrate binding affinity is necessary but not sufficient to assign function. A reaction profile is ultimately what is needed for a more complete quantitative description of function.

To address this rather fundamental limitation, we propose to use quantum mechanical methods to explicitly compute transition state barriers that govern the rates of catalysis. Although quantum mechanical, and mixed quantum/classical (QM/MM), methods have been used extensively to investigate enzymatic reactions, the focus has been primarily on elucidating complex reaction mechanisms. Here, the key catalytic steps are known, and we use these methods quantify substrate specificity. That is, we bring the power of quantum mechanics to bear on the problem of annotating enzyme function, which is a novel approach.

Although it has been clear to us at the Jacobson group for a some time that enzyme specificity may be encoded in transition states, rather than simply substrate recognition, the main limitation has always been computational expense. Using a hierarchy of different methods, we can reduce the list of plausible substrates of an enzyme to a small number in most cases, but even identifying the transition states for a dozen plausible substrates requires significant computational effort, beyond what is practical using standard QM/MM methods.

The technologies championed by our collaborators at LLNL have breakthrough potential in this respect, by combining linear scaling quantum mechanics with impressive parallelization. This technology, combined with our extensive knowledge of the enzyme superfamilies we study, including their reaction mechanisms, makes QM/MM a practical approach to solving the most challenging enzyme function assignment problems, i.e., those that have been refractory to other methods. We wish to stress that the expertise of our collaborators at LLNL

is equally important. The lead investigator on the project, Dr. Chakrapani “CK” Kalyanaraman, with nearly a decade of experience on enzyme function assignment, will spend considerable time working directly (on site) with our collaborators. If this project is as successful as we believe it will be, there is also significant potential to use our preliminary results to obtain long-term funding from the NIH or DOE.

For this project, we have chosen two enzyme superfamilies which we have used as ‘model systems’ for functional assignment. The enolase superfamily is a large group of α - β barrel enzymes with highly diverse substrates and chemical transformations. Despite decades of work, over a third of the superfamily remains unassigned, which means that the remaining cases are by definition difficult to assign. We have focused focus on acid sugar dehydratases, and have considerable expertise on the matter. We are also interested in the isoprenoid synthase superfamily, which is of central interest to the synthetic biology community, because these enzymes are used by nature to create complex rare natural products of medicinal value. The most notable example of this is the artemisinin, an antimalarial compound that is found in trace amounts in the wormwood root. From the standpoint of enzyme function assignment, these enzymes are intriguing because they use a small number of chemically simple substrates to generate, potentially, tens of thousands of different products. Hence, substrate binding specificity is only a small part of the challenge; the key is determining how the enzyme directs the carbocation chemistry to specific products. These more complex modeling approaches clearly require quantum mechanical methods.

For both enzyme superfamilies, the Jacobson lab has close collaborations with enzymology groups who can test the predictions, as well as a longstanding collaboration with structural biologist Steven Almo, who has determined the structures of many enzymes in complex with substrates suggested by Jacobson’s group.

The problem of predicting enzyme substrate specificity is closely related to the problem of engineering enzymes. Here, the enzymes we work with are certain to be proficient catalysts for some substrate, and the challenge is ‘merely’ identifying the substrate. These methods, however, could be applied to a similar challenge the related challenge of modifying enzymes to catalyze a desired reaction. This approach could be a longer term goal, and would build from the knowledge and technology developed for studying the known systems.

Aim 1: *Understanding substrate specificity in enolase superfamily enzyme glucarate dehydratase using a quantum mechanical method.*

The Jacobson research group at UCSF is currently collaborating with the enzymology group of Professor John Gerlt at University of Illinois Urbana Champaign on assigning functions to enolase superfamily enzymes. Enolase superfamily enzymes have two domains, a mixed α/β N-terminal domain and $(\beta/\alpha)_7\beta$ -barrel C-terminal domain[2]. While the N-terminal domain acts as a capping domain, the catalytic machinery is located in the barrel domain. Although the enolase superfamily enzymes catalyze a wide variety of overall reactions, all of these reactions share a common initiation step, which is the abstraction of a proton to form an enolate anion intermediate, which can then lead to various products. These enzymes have a highly charged active site with a variety of residues that can perform the proton abstraction

and one or two bivalent metal ions (usually Mg^{2+}) that stabilize the enolate anion. Some of the many diverse substrates of enolase superfamily enzymes include amino acid derivatives, xenobiotic-derived species such as mandelate, and acid sugars such as glucarate[3-4], rhamnonate, fuconate[5], talarate, galactarate[6], mannionate[7] and tartrate[8].

Sugar dehydratase enzymes are highly stereo- and regio- specific. For instance, glucarate dehydratase (GlucD) enzyme takes D-glucarate or its epimer at the 5-th carbon, L-idarate, as the substrate and produces 5-keto-4-deoxy glucarate as the product[4]. Even a small change in chirality at one chiral center of the 6-carbon sugar can dramatically alter the kinetic efficiency (k_{cat}/K_M) by as much as 3 orders of magnitude[5]. The binding site of the product bound structure of the GlucD enzyme is shown in figure 1. The catalytic and specificity determining residues are all either highly charged or polar, and there is a bivalent cation. The presence of a metal in the binding site. The complex and highly polarized electrostatic environment necessitates a more complex treatment than the standard point charge based molecular-mechanics energy functions that are typically used. A full reaction-profile analysis based on mixed quantum mechanics/molecular mechanics (QM/MM) approach, whereby by critical residues in the binding site can be treated by quantum mechanics and the remaining part can be modeled using molecular mechanics, is appropriate. Preliminary calculations performed in the Jacobson lab considered just the metal ion, substrate and 3 polar residues in the active site (Thr103, Tyr150 and Asn341), quantum mechanically and the remaining part using molecular mechanics. Thr103 was identified as important for specificity determination in accordance with the experiment. However, this study also suggested that

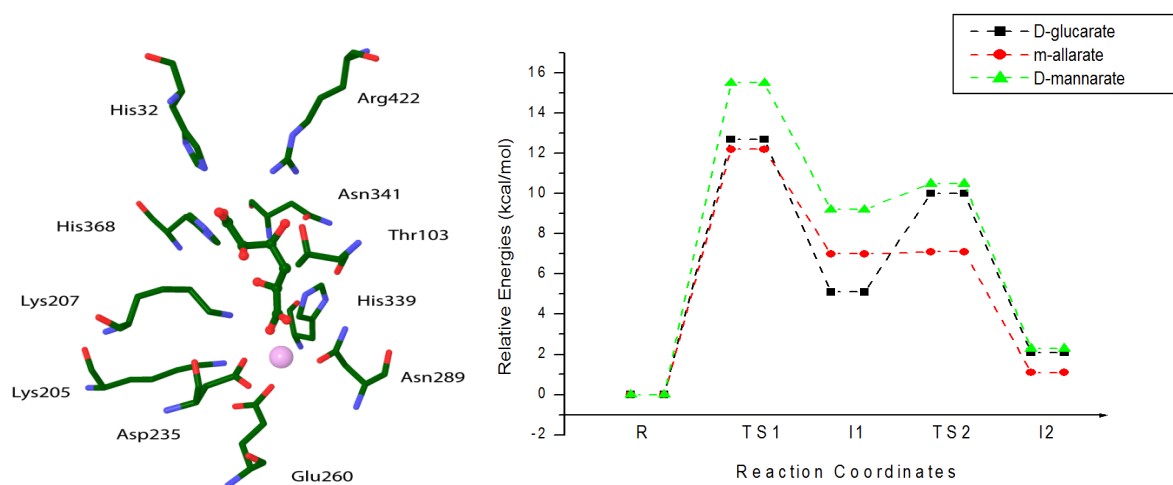


Figure 1. Left: Active site of glucarate dehydratase (GlucD) enzyme is shown. The product, 5-keto-4-deoxy glucarate, is present in the active site (ball-and-stick representation). Right: Preliminary calculations using standard QM methods, that required many CPU-years of computational expense. (Boxue Tian and Frank Wallrapp). Note that the reaction profile clearly excludes D-mannarate as a substrate based on the high energy transition state.

treating additional residues quantum mechanically might be necessary to fully understand how polar and charged residues in the binding site determine the substrate specificity. In this work, we propose to treat *all* residues in the first shell of the binding site, the metal ion and

the sugar metabolite quantum mechanically, with the remaining portion modeled classically. First, we will validate the model by explaining the experimentally determined specificity by calculating the reaction profile (i.e., barrier height) for all 6-carbon dicarboxylic acid sugars, followed by an enzyme engineering study of what possible mutations can change the specificity. The Jacobson lab also has collaborations with the Gerlt lab, who can provide laboratory measurements to test the predictions of the study.

Aim 2. Deciphering the mechanism of product specificity in terpene synthases

Terpenes are structurally and stereochemically diverse natural products. To date, over 55,000 terpenoid compounds have been described. These compounds are all produced from basic 5-carbon building blocks, isopentenyl diphosphate (IPP) and dimethyl allyl diphosphate (DMAPP), by terpene synthases [9-10]. These precursors can be coupled in head-to-tail fashion in chain elongation reactions to produce polyisoprenoid diphosphates which can be cyclized to generate single or polycyclic terpenes. The precursors can also be coupled in non head-to-tail (or irregular) fashion to yield cyclobutyl, branched or cyclopropyl products[11]. The terpene synthases act as a template for which the flexible isoprenoid substrate(s) bind with proper orientation and conformation so that when the reactive substrate carbocation is generated, a series of carbon-carbon formation steps is facilitated. The template also functions as a chaperone in subsequent steps of the reaction cascade. The reaction is eventually terminated when the reactive carbocation is quenched by proton elimination or hydroxylation. While some terpene synthases such as aristolochene synthase from *Aspergillus terreus* produce only one product[12], other enzymes promiscuously produce multiple products. For instance, the enzyme γ -humulene synthase produces 52 distinct products of which γ -humulene is only 29% [13].

Four years ago, Poulter and co-workers [11, 14] created a chimera of two isoprenoid synthases from *Chrysanthemum cineralefolium* (chrysanthemum) and *Artemisia tridentata* (sagebrush) and demonstrated how all four reactions, namely chain elongation, cyclobutanation, branching and cyclopropanation could be elucidated from a common mechanism involving same reaction intermediates based on the dissociative electrophilic alkylation mechanism for chain elongation[14] (see figure 2). The key feature of the mechanism is a series of carbocation intermediates related by internal rearrangements.

We hypothesize that it is the stabilization of these intermediates, which differ geometrically and electrostatically, that determines which classes of products will be formed. This hypothesis is consistent with the ability to alter the product profile by a relatively small number of aminoacid substitutions in the binding site [11, 14]. In addition, the position of the catalytic base that removes protons and forms the final products to terminate the reaction is also a major determinant of product specificity.

We will build models of these chimeric enzymes and use the QM/MM methods to map out the reaction profiles of the various carbocation intermediates, and the barriers separating them. The Jacobson lab has an ongoing collaboration with Poulter group at University of Utah characterizing isoprenoid superfamily enzymes, and can measure changes in specificities experimentally to verify the predictions.

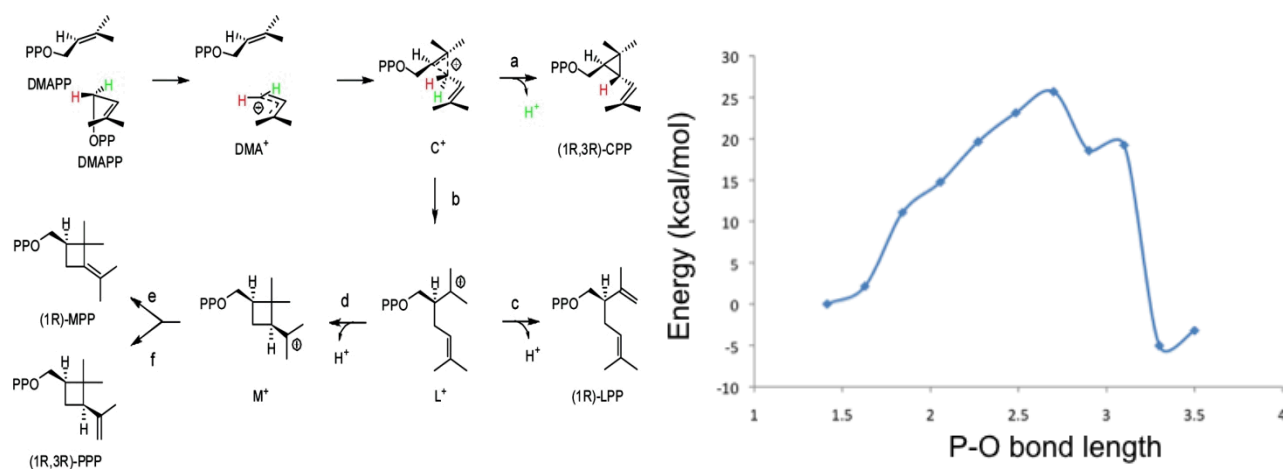


Figure 2. Left: Common reaction mechanism involving the chain elongation reaction intermediate, C^+ , undergoing cyclopropanation (1R, 3R-CPP), branching (1R-LPP) and cyclobutane (1R-MPP or 1R, 3R-PPP) products [14]. Right Preliminary computational results on the first step of the reaction, removal of the diphosphate to form the first carbocation intermediate, in the simpler case of a chain elongation enzyme.

Aim 3: Using quantum mechanics to study enzymatic catalysis at large scale

The systems in Aims 1 and 2 represent a class of computational problems that require quantum mechanical treatment, particularly since the transition state of a reactive system is of central interest. Quantum systems are notoriously difficult to study at the larger scale, since the complexity of the calculation typically scales as the number of atoms cubed, which can often limit the size of a calculation to under 100 atoms. We wish to use a quantum code developed at LLNL that scales linearly with the number of atoms, allowing for very large systems (greater than 500 atoms) to be studied. This, combined with the substantial computational resources at LLNL provides a unique opportunity to study enzymatic systems at unprecedented scales.

MGmol is a parallel first-principles molecular dynamics code designed for high-performance simulation of large molecular systems and developed in the Center for Applied Scientific Computing at LLNL by Jean-Luc Fattebert. It solves the equations of Density Functional Theory (DFT) using the Finite Difference method. MGmol scales to thousands of processors on high-end parallel platforms. It is a unique tool in the sense that its computational cost scales linearly with the number of electrons, because of a novel algorithm that uses localized orbitals to represent the electronic structure [15], while achieving the accuracy of the more widely used (and computationally costly) plane wave codes.

It uses standard norm-conserving pseudopotentials, enabling straightforward incorporation of new atomic types from the literature. It uses a multigrid algorithm to efficiently solve for the long-range coulomb interactions and accelerate the iterative solver. The solver [16], developed and optimized to run specifically with the MGmol code, is a modern and efficient

approach. The code can easily manage periodic and nonperiodic boundary conditions, and is implemented C++/MPI, with the parallel linear algebra library ScaLapack [17]. Data and restart files are saved in the HDF format, which is designed handle massive datasets. MGmol runs on Linux clusters and BGL architectures, providing flexibility to run at very large scales, including the soon to be installed BGL architecture Sequoia, which be the world's most powerful petascale computing system once complete.

While the code is well suited for many applications, we have invested considerable effort in modeling biomolecular systems. A new methodology was recently developed to model the electrostatic environment allowing for biological sub-systems to be embedded so that Quantum Mechanics simulations can be performed in realistic protein electrostatic environments[18]. While this method is novel, it is comparable to the hybrid method more widely described as QM/MM, and enables quantum calculations of molecular systems made of several hundred atoms, which is far greater than typical QM/MM calculations. To facilitate transition state calculations, the Nudged Elastic Band (NEB) algorithm was also recently implemented.

Recent first-principles molecular dynamics simulations of biological systems of 612 quantum mechanical atoms were carried out on LLNL Linux clusters. Using 31 nodes (~363 processors), one converged molecular dynamics step can be completed in 2 minutes (wall clock time) with MGmol for that system. Typical geometry optimization calculations of such systems can be carried out in a few hours.

Graduate Training and Engagement

Reference:

1. Kalyanaraman, C., K. Bernacki, and M.P. Jacobson, *Virtual screening against highly charged active sites: identifying substrates of alpha-beta barrel enzymes*. Biochemistry, 2005. **44**(6): p. 2059-2071.
2. Babbitt, P.C. and J.A. Gerlt, *Understanding enzyme superfamilies. Chemistry As the fundamental determinant in the evolution of new catalytic activities*. J Biol Chem, 1997. **272**(49): p. 30591-4.
3. Gulick, A.M., et al., *Evolution of enzymatic activities in the enolase superfamily: crystal structure of (D)-glucarate dehydratase from Pseudomonas putida*. Biochemistry, 1998. **37**(41): p. 14358-68.
4. Hubbard, B.K., et al., *Evolution of enzymatic activities in the enolase superfamily: characterization of the (D)-glucarate/galactarate catabolic pathway in Escherichia coli*. Biochemistry, 1998. **37**(41): p. 14369-75.
5. Yew, W.S., et al., *Evolution of enzymatic activities in the enolase superfamily: L-fuconate dehydratase from Xanthomonas campestris*. Biochemistry, 2006. **45**(49): p. 14582-97.
6. Yew, W.S., et al., *Evolution of enzymatic activities in the enolase superfamily: L-talarate/galactarate dehydratase from Salmonella typhimurium LT2*. Biochemistry, 2007. **46**(33): p. 9564-77.

7. Rakus, J.F., et al., *Evolution of enzymatic activities in the enolase superfamily: D-Mannonate dehydratase from Novosphingobium aromaticivorans*. Biochemistry, 2007. **46**(45): p. 12896-908.
8. Yew, W.S., et al., *Evolution of enzymatic activities in the enolase superfamily: D-tartrate dehydratase from Bradyrhizobium japonicum*. Biochemistry, 2006. **45**(49): p. 14598-608.
9. Christianson, D.W., *Unearthing the roots of the terpenome*. Curr Opin Chem Biol, 2008. **12**(2): p. 141-50.
10. Kellogg, B.A. and C.D. Poulter, *Chain elongation in the isoprenoid biosynthetic pathway*. Curr Opin Chem Biol, 1997. **1**(4): p. 570-8.
11. Thulasiram, H.V., H.K. Erickson, and C.D. Poulter, *Chimeras of two isoprenoid synthases catalyze all four coupling reactions in isoprenoid biosynthesis*. Science, 2007. **316**(5821): p. 73-6.
12. Felicetti, B. and D.E. Cane, *Aristolochene synthase: mechanistic analysis of active site residues by site-directed mutagenesis*. J Am Chem Soc, 2004. **126**(23): p. 7212-21.
13. Steele, C.L., et al., *Sesquiterpene synthases from grand fir (Abies grandis). Comparison of constitutive and wound-induced activities, and cDNA isolation, characterization, and bacterial expression of delta-selinene synthase and gamma-humulene synthase*. J Biol Chem, 1998. **273**(4): p. 2078-89.
14. Thulasiram, H.V., H.K. Erickson, and C.D. Poulter, *A common mechanism for branching, cyclopropanation, and cyclobutanation reactions in the isoprenoid biosynthetic pathway*. J Am Chem Soc, 2008. **130**(6): p. 1966-71.
15. Fattebert, J.-L. and F. Gygi, *Linear Scaling First-Principles Molecular Dynamics with Plane-waves Accuracy*. Physical Review B, 2006. **73**: p. 115124.
16. Fattebert, J.L., *Accelerated Block Preconditioned Gradient method for large scale wave functions calculations in Density Functional Theory*. Journal of Computational Physics, 2010. **229**(2): p. 441-452.
17. Choi, J., et al. *ScaLAPACK: A scalable linear algebra library for distributed memory concurrent computers*. 1992: IEEE.
18. Fattebert, J.L., et al., *Quantitative assessment of electrostatic embedding in Density Functional Theory calculations of biomolecular systems*. Journal of Chemical Theory and Computation, 2009. **5**(9): p. 2257-2264.