



W&M ScholarWorks

---

Dissertations, Theses, and Masters Projects

Theses, Dissertations, & Master Projects

---

2003

## Evolution and Analysis of a Catalytically Effective Model Enzyme: The Importance of Active Site Orientation and Tuned Conformational Fluctuations

G. S. Blair Williams  
*College of William & Mary - Arts & Sciences*

Follow this and additional works at: <https://scholarworks.wm.edu/etd>

 Part of the [Biochemistry Commons](#)

---

### Recommended Citation

Williams, G. S. Blair, "Evolution and Analysis of a Catalytically Effective Model Enzyme: The Importance of Active Site Orientation and Tuned Conformational Fluctuations" (2003). *Dissertations, Theses, and Masters Projects*. Paper 1539626409.

<https://dx.doi.org/doi:10.21220/s2-vvne-4n36>

This Thesis is brought to you for free and open access by the Theses, Dissertations, & Master Projects at W&M ScholarWorks. It has been accepted for inclusion in Dissertations, Theses, and Masters Projects by an authorized administrator of W&M ScholarWorks. For more information, please contact [scholarworks@wm.edu](mailto:scholarworks@wm.edu).

EVOLUTION AND ANALYSIS OF A CATALYTICALLY  
EFFECTIVE MODEL ENZYME:

The importance of active site orientation and tuned conformational  
fluctuations.

---

A Thesis

Presented to

The Faculty of the Department of Chemistry

The College of William and Mary in Virginia

In Partial Fulfillment

Of the Requirements for the Degree of

Master of Arts

---

by

G. S. Blair Williams

2003

# APPROVAL SHEET

This thesis is submitted in partial fulfillment of the  
requirements for the degree of

Master of Arts

A handwritten signature in black ink, appearing to read "G. S. Blair Williams", written over a horizontal line.

G. S. Blair Williams

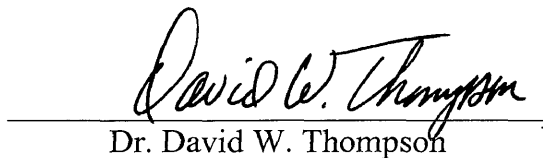
Approved, May 2003

A handwritten signature in black ink, appearing to read "Carey K. Bagdassarian", written over a horizontal line.

Dr. Carey K. Bagdassarian

A handwritten signature in black ink, appearing to read "David E. Kranbuehl", written over a horizontal line.

Dr. David E. Kranbuehl

A handwritten signature in black ink, appearing to read "David W. Thompson", written over a horizontal line.

Dr. David W. Thompson

# TABLE OF CONTENTS

	Page
Acknowledgments	iv
List of Figures	vi
Abstract	vii
Chapter 1. Introduction	2
Part I	
Chapter 2. Molecular Dynamics (MD)	7
2.1    Molecular Dynamics Design	7
2.2    Molecular Dynamics Tests & Stability	10
Part II	
Chapter 3. Genetic Algorithm (GA)	15
3.1    Genetic Algorithm Design	15
3.2    Genetic Algorithm and MD Coupling	18
3.3    Genetic Algorithm Results	19
Part III	
Chapter 4. Analysis and Normal Mode Approximation	21
4.1    Early Analysis	22
4.2    Data Mining	24
4.3.1  Design of Normal Mode Approximation	28
4.3.2  Design of Normal Mode Analysis Program	31
4.3.3  Normal Mode Analysis Results	34
Chapter 5. Discussion	38
Bibliography	39
Vita	42

## ACKNOWLEDGMENTS

I would like to first and foremost thank my mentor Carey Bagdassarian. He has given me the invaluable opportunity to realize that my true passion lay in scientific research. He has demonstrated a great deal of patience and guidance throughout the course of this work.

I would also like to thank Aftab Hossain -- without his extremely well designed molecular dynamics simulation and programming guidance this project would not have been possible. Most instrumental was the insight from Dr. David Kranbuehl, our collaborator and co-author, as well as very fruitful discussions with Dr. Michael Trosset, and Dr. Brian Space. These insights and discussions were very important in shaping both the course and content of this work.

Finally I cannot forget to thank my parents whose support has always been fundamental in my accomplishments. I must also thank Danae, for her infinite patience in my often research related conversation topics and very unusual work hours.

## LIST OF FIGURES

Figure	Page
1. Three Dimensional Layout of MD simulation	9
2. Equilibration of potential energy with time for the stiffest and loosest individuals	11
3. Catalytic fitness as a function of 10 different initial conditions	12
4. Number of chemical events within bins of $1.3 \times 10^4$ time-steps for four individuals (fittest, least fit, all loose, all stiff).	13
5. Gaussian distribution of Positions in X, Y, and Z directions	13
6. Performance of the Gas	19
7. Average catalytic fitness as a function of the number of 0-subunits	24
8. Scaled C-S distance as a function of the number of 0-subunits	25
9. Scaled C-S distance as function of catalytic fitness	27
10. Geometry of a simple normal mode example	28
11. Relationship between the number of beneficial vectors found in low frequencies to catalytic fitness	35
12. Relation of "catalytic score" to catalytic fitness	36

## ABSTRACT

The work included in this thesis is in three major parts. First, in order to study the function of long-range many-atom motions on catalytic efficacy, we designed and implemented a molecular dynamics simulation used to simulate the varying degrees of conformational freedom that amino acid residues exhibit when in different tertiary structures within an enzyme. Second, we designed the coupling of this molecular dynamics engine to a specialized genetic algorithm with the goal of “evolving” catalytically more effective fluctuations by modifying, through the process of selection, recombination, and mutation consistent with Darwinian evolution, the arrangement of stiff, intermediate, and loose interactions. Third, the study of this “evolution”--using various data mining techniques as well as a normal mode approximation--is presented.

Approximately 24,000 different model enzymes are created for study. The least “catalytically fit” enzyme manages only 16 chemical events, while the fittest boasts 253. A normal mode approximation lends insight into how low frequency modes generate and maintain beneficial conformational fluctuations. Furthermore, point mutations far from the active-site are shown to have a significant detrimental impact on catalytic fitness, which reinforces the belief that effective catalysis requires long-range globally correlated fluctuations.

EVOLUTION AND ANALYSIS OF A CATALYTICALLY  
EFFECTIVE MODEL ENZYME:

The importance of active site orientation and tuned conformational  
fluctuations.



## Chapter 1

### 1.1 Introduction

The primary questions driving this work are Why did enzymes evolve to be so large? and What role does this global structure have on conformational fluctuations?. Enzymes contain hundreds of amino acid groups folded into a complex three-dimensional structure. However, the business portion of this protein structure, the active site, is very small relative to the overall size of the enzyme. The role of the active site in transition state stabilization is well appreciated as the hallmark for the amazing rate accelerations during enzyme catalysis.<sup>1,2,3</sup> It is also well appreciated that large-scale domain motions are important to an enzyme's ability to capture and sequester substrate within the active site.<sup>4,5,6</sup> These large domain structures of the enzyme are not simply motionless once the substrate is in the active site. In fact, the roles of atomic scale conformational fluctuations during reaction are not fully understood and are the focus of recent studies.<sup>7,8,9</sup>

Global correlated thermal fluctuations have been proposed to couple with the reaction coordinate thus improving catalysis.<sup>10</sup> More specifically the enzyme-substrate complex's three-dimensional structure could have evolved to favor catalytically beneficial global motions while restricting those motions considered "useless" or "stray."<sup>11</sup> Additional literature further indicates that global fluctuations have an influence on catalysis, specifically that residues distal from the active site may facilitate the linkage of substrate to catalyst.<sup>12,13</sup> In fact, recent results suggest that motion from long range residues enhances the crossing of the chemical reaction

barrier and further support a dynamical role of the protein even during catalysis.<sup>14</sup>

While the cause of this motion is not identified, we believe that our model may lend insight into how these correlated motions are generated and maintained.

This work attempts to “tune” these conformation fluctuations in order to evolve a model enzyme population that grows in catalytic efficacy. Experimental and computational results have shown that amino acid groups located in different tertiary domains of a protein exhibit differing degrees of conformational freedom and can be studied and predicted using molecular dynamics.<sup>15</sup> Further experimental studies have been performed on conformational freedom induced by ligand binding.<sup>16</sup> We simulated these varying degrees of freedom using a spatial distribution of “stiff” and “loose” domains arranged to form a “toy” enzyme, which is then evaluated using molecular dynamics. A genetic algorithm then operates on this arrangement of domains via selection, crossovers, and mutations as it attempts to improve catalytic function.

Alder and Wainwright created the MD simulation concept nearly fifty years ago.<sup>17</sup> While many different “flavors” of MDs exist today with modifications for different applications and improvements, the universal concept is simple, using Newtonian force calculations to predict the motion of objects. Lattice models have been used in other protein modeling applications including investigations in protein dynamics and protein folding.<sup>18,19,20,21</sup> Computational limitations make a “true” simulation of an enzyme nearly impossible since time considerations would only allow the MD to run for only a few nanoseconds of simulated time. Since the chemical events of interest usually occur on a millisecond timescale the enzyme must be simplified. By focusing on only the different degrees of conformation freedom demonstrated by amino acid residues in different tertiary structures, “realistic”

simulation of large-scale fluctuations is possible. The simpler model allows the MD simulation to run for much longer time periods allowing the capture of numerous chemical events. Currently, a single MD run takes approximately 4 minutes with our computational capability. Because of this severe bottleneck, running all possible combinations of “stiff” and “loose” domains would take approximately 3 million days. For this reason the MD must be run with some form of optimization method, in this case, a custom genetic algorithm.

The genetic algorithm (GA) used in this work was designed specifically for the project and possessed some unique modifications. These modifications were designed to improve and even maximize the efficiency of the GA. GAs attempt to simulate evolution by selecting solutions with a probability based on their relative fitness and act on those solutions in an attempt to create better solutions.<sup>22</sup> “GAs are stochastic methods which enforce the survival of the fittest paradigm of evolution along with the genetic propagation of characteristics.”<sup>22</sup> This method of optimization has emerged from being just a concept to a very useful tool to computational chemists for optimization and molecular design.<sup>22</sup> While quantification of our GA’s efficiency would be difficult, it was able to create a 15-fold increase in fitness within 12,000 generations. Since two individuals are created each generation this is a total population of 24,000 individuals generated from over a billion possible unique individuals allowed by the size of the system. This means the GA created only 0.002% of the possible individuals to find what we now believe to be a potential maximum and minimum fitness for the system.

For the purpose of better understanding the fluctuations present in the enzyme population a normal mode analysis was created. The anharmonic and distance related properties of the chemical barrier present in the active site of the model force this analysis to exist only as an approximation. Even with this limitation, however, the analysis gave us extremely useful insights into what types of fluctuations are present and also how they are generated and maintained.

Part I  
Molecular Dynamics

## Chapter 2

### Molecular Dynamics Simulations

#### 2.1 Molecular Dynamics Engine Design

Our Molecular Dynamics simulation is a simple lattice based model, allowing control of all variables in the system. The three dimensional layout of the lattice is shown in figure 1. The model enzyme consists of 168 thermally fluctuating subunits. These thermally fluctuating subunits are surrounded a stationary shell of residues, called phantoms (P), which serve to maintain the three dimensional shape of the enzyme. Those labeled "N" in figure 1 are known as neutrals since the genetic algorithm does not modify them. The genetic algorithm does, however, operate on those subunits known as dynamics labeled "D" in figure 1, as well as the Catalytic and Substrate residues labeled "C" and "S" respectively. Each "amino acid" in the system has its own unique properties, including block number, global position, local position (scaled by equilibrium amino distance), velocity, type (either stiff or loose), and finally category (either Phantom (P -- stationary), Neutral (N), Dynamic (D -- type can change), Catalytic(C ) or Substrate(S) ). The MD also requires information about the different spring properties (governed by the amino acid type); these vary from loose, medium, to stiff. Thus two stiff domains interacting have a stiff spring between them while two loose domains contain a loose spring. Interactions between a stiff domain and a loose domain would utilize a medium spring. Run length and time step length must also be provided.

The molecular dynamics engine, hereafter referred to as MD, has several "modes" requiring additional information or producing additional output:

- 1) Detailed Mode:
  - a. Detailed information including position, velocity, temperature, and energies are outputted at defined intervals along the run length.
- 2) RMS Mode:
  - a. Root mean square values for all non-phantom blocks are written to a file.
- 3) Enzyme Mode:
  - a. Catalytic (C) and Substrate (S) residues have a unique non-spring interaction between them taking the form of a potential energy barrier (slope and size must be provided).
  - b. Also a fitness value is provided at the end of the run consisting of the number of times C and S overcome the potential energy barrier.

Once all information is provided to the MD it must first be run for an equilibration run in order to allow the initial displacements to equilibrate into a Gaussian distribution of velocities and stable 'target' temperature. These new velocities are the provided for the equilibrated run and the MD is run for a much longer length. Current MD settings are as follows:

Time step:	1e-14s	
Equilibration run:	1e-11s	
Equilibrated run:	1.3e-09s	
Target Temperature	298K	
Equilibrium spring length:	1.523e-10m	(carbon-carbon single bond)
Mass of amino acid:	1.9943e-026 kg	(single carbon atom)
Loose spring constant:	2 J/m <sup>2</sup>	
Medium spring constant:	25 J/m <sup>2</sup>	
Stiff spring constant:	50 J/m <sup>2</sup>	
Number of dimensions:	3D	
Size of system:	450 amino acids	
Number of Dynamics:	30	
Active site(C&S) location:	Center	
Barrier Force:	2e-10 J/A	
Outer Barrier Force Range:	1.0 times equilibrium distance	
Inner Barrier Force Range:	0.7 times equilibrium distance	

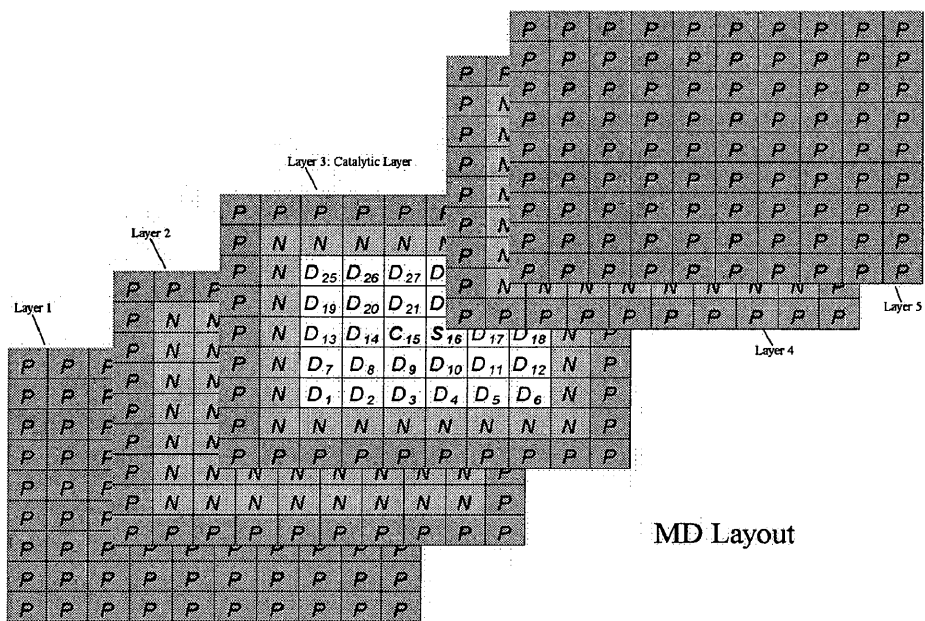


Figure 1: Three Dimensional Layout of MD simulation



## 2.2 Molecular Dynamic Engine Tests and Stability

Extensive preliminary tests were conducted to make sure all calculations were correct and the MD was performing as expected. Additional stability tests were performed to support this conclusion. First we made sure that the time step was sufficiently small enough to allow adequate capture of motion. Using the largest  $k$  value contained in the system, equation 1 yields the characteristic frequency of the system. Inverting this value gave us the characteristic time step of the system. Dividing this characteristic time step by ten creates a molecular dynamics time step capable of capturing system motion.

$$\nu = \frac{1}{2\pi} \sqrt{\frac{k}{m}}$$

Equation 1

Next the length of the equilibration run must be checked. Once the system has reached equilibrium the potential energy of the system will have stabilized in its normal oscillatory behavior. Figure 2 shows the stabilization of the potential energy with the passage of time. The system appears to be equilibrated after around 100 steps equaling  $5e-12$  seconds, so our equilibration run length of  $1e-11$  seconds is adequate to allow for equilibration.

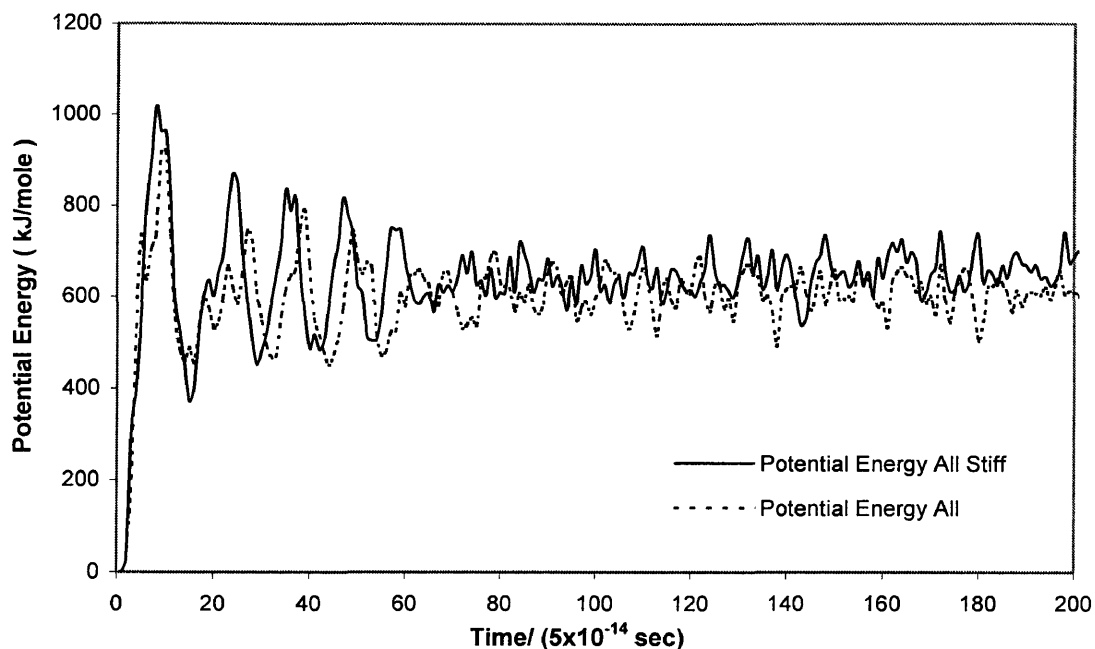


Figure 2: Equilibration of potential energy with time for the stiffest and loosest individuals.

The MD should also be stable to the initial configuration of displacements used to create motion in the system. Three individuals were run ten times with random initial positions for each amino acid and the resulting range of 'hits' with an average error of seven percent was well within our goal of ten percent error as shown in figure 3.

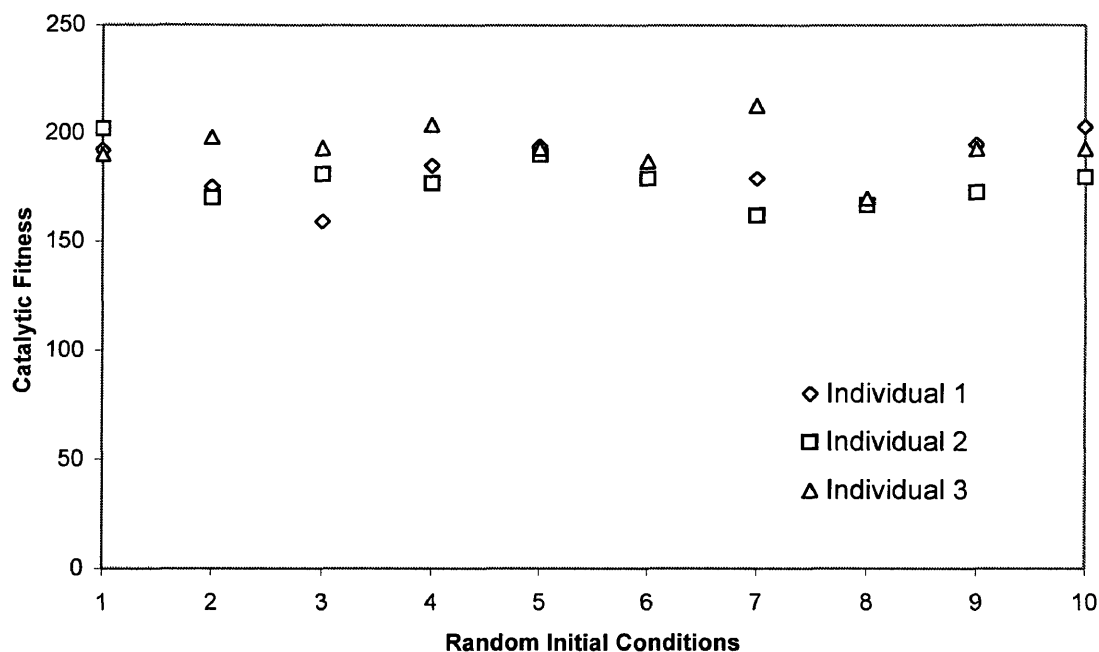


Figure 3: Catalytic fitness as a function of 10 different initial conditions.

Finally the rate at which the chemical events occur should also be stable across the entire length of the run. Figure 4 shows the two homogenous individuals (all stiff and all loose), the best, and the worse individual's chemical events divided into ten bins of time. These results demonstrate that chemical events occur at regular intervals for all four individuals and the "hit stability" was actually surprisingly high. Also a histogram (figure 5) of an atom's positions through time yields a bell curve--indicating that the atom's average position is the center of the block--as expected.

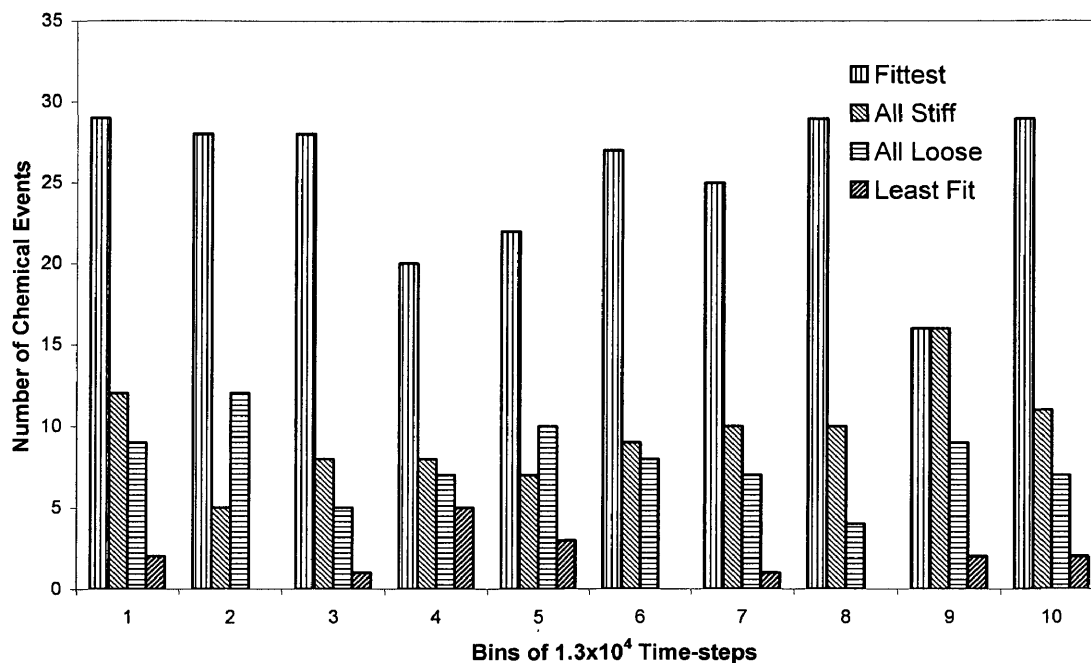


Figure 4: Number of chemical events within bins of  $1.3 \times 10^4$  time-steps for four individuals (fittest, least fit, all loose, all stiff).

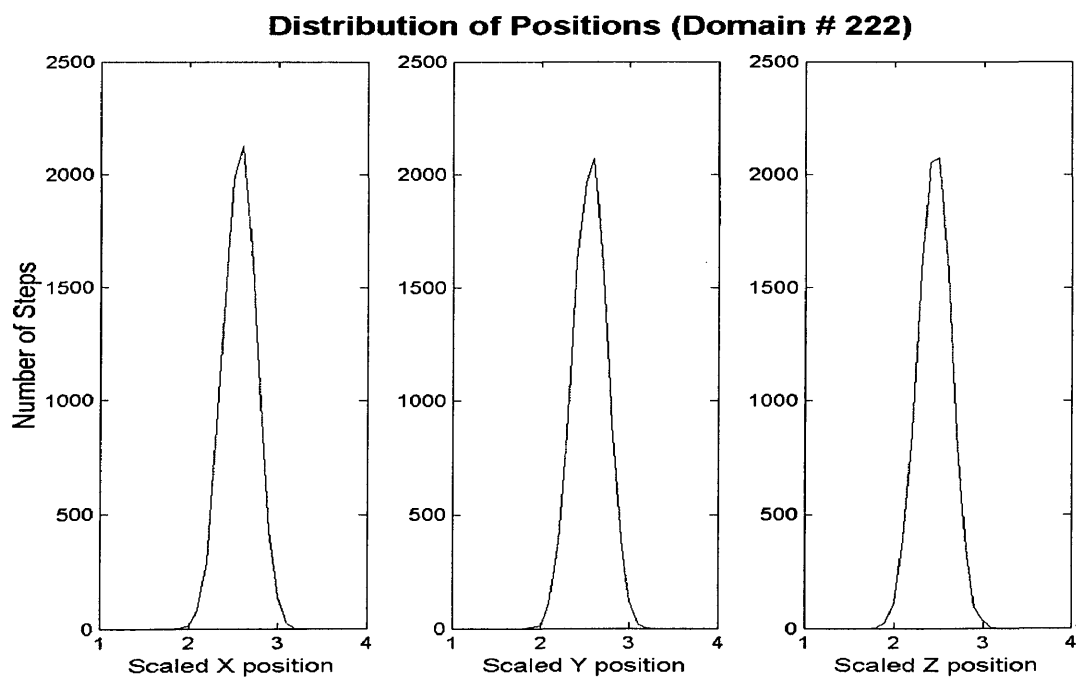


Figure 5: Gaussian distribution of Positions in X, Y, and Z directions

**Part II**  
**Genetic Algorithm**

## Chapter 3

### Genetic Algorithm

#### 3.1 Genetic Algorithm Design

Every Genetic Algorithm, abbreviated GA from hence forth, consists of three basic steps. First the “parents” must be selected based on fitness, then they are mated via sequence crossover and possible mutation, and finally any unique offspring must be evaluated for fitness (number of times C and S surmount the potential energy barrier) and added to the population. “In GAs natural selection occurs by choosing solutions with a probability proportional to their relative fitness values by some scheme.”<sup>22</sup> The methods for doing this are almost as infinite as the imagination can allow, however, we believed that certain methods could lead to a more “efficient” GA.

For our purposes we used a 30 bit ‘binary’ GA, meaning that each individual is characterized by a sequence of 30 binary digits. With this sequence size the potential number of possible individuals is very large ( $2^{30} = 1,073,741,823$ ) reinforcing the need for efficiency. For increased efficiency the following initial population was used; containing all the possible ‘traits’ at every location (heterogenous individuals) as well as the two extremes (homogenous sequences).

```
Heterogeneous individual #1: 010101010101010101010101010101
Heterogeneous individual #2: 101010101010101010101010101010
Homogeneous individual #3: 000000000000000000000000000000
Homogeneous individual #4: 111111111111111111111111111111
```

First “potential parents” must be selected from the initial population. For our purposes we used tournament selection where a specified number (i.e. 10) of

individuals are randomly selected be members of a “potential parent population.”

Of these randomly chosen members the fittest two become the “mating” parents.

This method avoids the affects of a few large fitness members towards those of simply ‘above average’ fitness. This scheme is also more of a static selection procedure versus the other well-known “roulette wheel scheme” resulting in a more constant probability of selection across multiple cycles. The tournament selection method seems the best for our needs since it decreases the likelihood that the same highly fit parents will be continually selected each cycle, thus preventing the chance that the GA will settle into a false optimum.

Next, the “mating” parent’s sequences are crossed, creating two new “offspring” sequences. During each mating, existing parts or sequences are crossed to create new and possibly fitter individuals. The mating of parents presents many possible means to increase GA efficiency. We used a modified single point crossover driven by catalytic coefficients. Crossovers occur at only one single point and the location is biased by the catalytic coefficients assigned to each site. If a crossover at a particular site results in a large change in fitness then the coefficient was increased and vice versa if only a small fitness change occurs. The higher the catalytic coefficient of a site the greater the chance it would be selected as the crossover site. Catalytic coefficient driven (single point) crossovers allowed for small modifications to the parent individuals at catalytically important sites. These small modifications are more desirable than larger multiple site changes since our simple molecular dynamics engine design is very sensitive to even small system changes. Thus small modifications allowed the GA to ‘move’ smoothly though the fitness gradient. These small changes did, however, increase the probability for the GA to settle into local optima. Various techniques involving how the catalytic factors are updated as well as

how the crossover site was selected attempted to minimize the changes of a false optimum. Mutations at random locations were also used in order to allow the introduction of "traits" not present in the initial population.

Step three is to calculate the fitness of the new sequences (run the MD) and add the sequences to the general population if they are unique (i.e. not previously generated). Finally steps 1, 2, and 3 are simply repeated until the desired fitness or final population is achieved.



### 3.2 Genetic Algorithm and Molecular Dynamics Coupling

The genetic algorithm operates only on the binary sequence of numbers, which is interpreted by the MD as different types of amino acids. A '0' in the sequence represents a stiff domain while a '1' represents a loose domain. Before any MD can be run the input files required by it must be created. Thus the GA first creates the files and inserts the dynamic sequence of amino acids into to the three-dimensional structure of the lattice layout. Since every 'mock' enzyme is unique to its dynamic sequence it is thereby identified by that binary sum of the binary dynamic sequence, hereafter, identified as the enzyme ID.

Using this unique identifier a file system for the GA-MD coupling is created. For each individual generated by the GA, a folder is created named by its enzyme ID number. The GA-MD coupling then places the input and initialization files required by the MD inside this folder. For clarity all files used or created by the MD use a prefix of a lowercase 'md' and suffix of the enzyme id number and similarly those used or created by the GA use a lowercase 'ga' prefix. So, for example "mdEquilibratedRunInfo-1073741823.txt" contains run info for the individual with all loose domains and "gaMasterLog.txt" contains log information pertinent to the GA. A detailed log of the GA was created so that if the GA was interrupted at any point in the run it could be restarted without loss of information. While the GA is a stochastic method and need not necessarily be deterministic, the lost of information from time intensive MD runs would be detrimental. For further time optimization this coupling was designed to call and run two MDs at a single time in order to make most efficient use of our dual Intel Xeon processor system.

### 3.3 Genetic Algorithm Results

Keeping track of the maximum, average, and minimum fitness for the population through time serves to quantify the genetic algorithm's evolutionary progress. The GA by constantly selecting for fitter individuals drives the population towards increased fitness. Also by selecting for the less fit individuals the GA can be driven backwards to obtain a potentially least fit individual. These two types of GA are hereafter referred to as "forward" and "reverse" GAs respectively. The performance of these two optimization programs is shown in figure 6.

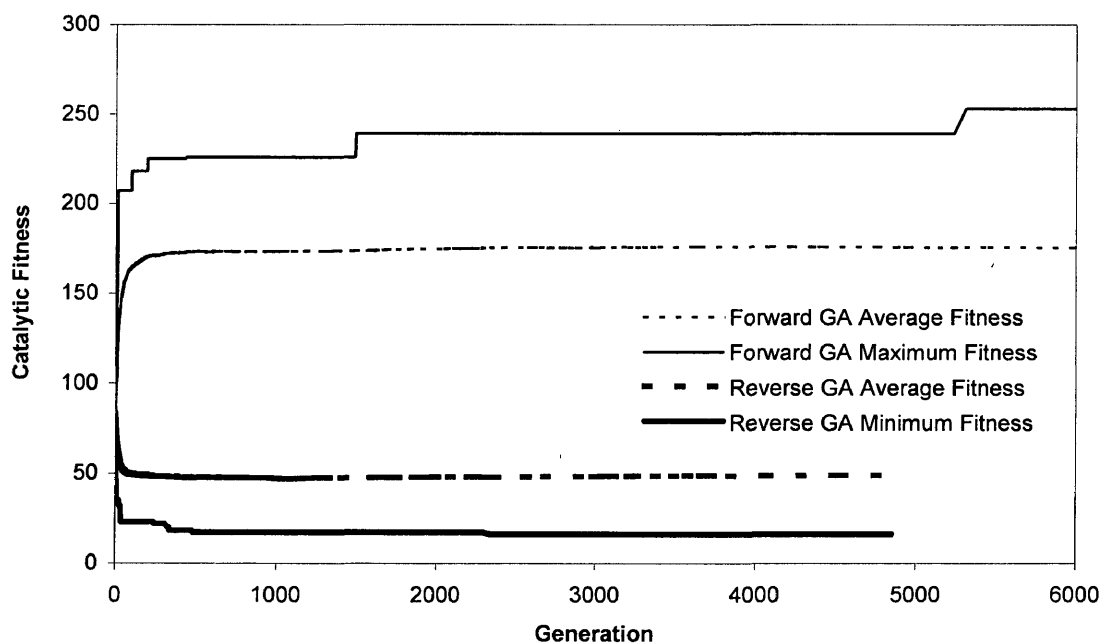


Figure 6: Performance of the GAs

Not shown in figure 6 is the minimum fitness for the forward GA or the maximum fitness for the reverse GA, however, it is important to note that the forward GA never created an individual less fit than the least fit individual present in the initial population and contrarily the reverse GA never found a more fit individual. This is powerful demonstration of the efficiency of the GA, showing its ability to move effectively through the “fitness gradient” and not waste computational time by creating very undesirable individuals.

## **Part III**

### **Analysis and Normal Mode Approximation**

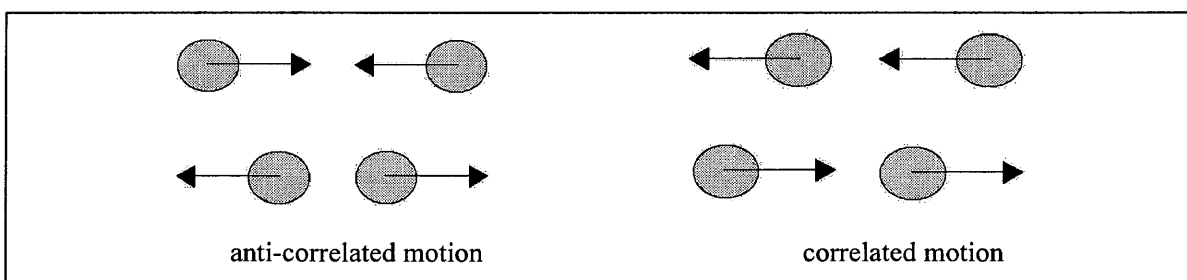
## Chapter 4

### Analysis and Normal Mode Approximation

#### 4.1 Early Analysis

Early analysis of catalytic efficiency consisted of using cross-correlation calculations, Fourier transforms, parametric plots, and even graphical movies of the enzyme's motion.

Cross-correlation yielded some interesting results and eventually we could successfully differentiate a bad enzyme (<50 hits) consisting of mostly loose domains from a good enzyme (>200 hits) consisting of mostly stiff domains. We believed that the symmetry was broken only in the x direction; a situation caused by replacing the spring (in the x direction) between the catalytic and substrate site with by a potential energy barrier. Using this assumption we analyzed in detail uncoupled velocities in the x direction, hoping to see an increase of anti-correlated velocities between the catalytic (C) and substrate residues (S) demonstrating an increased probability of a 'catalytic event' followed by the subsequent necessary outward 'breathing' motion.



However, trying to perceive a quantitative difference between an enzyme with average catalytic efficiency (~100 hits) and a very good enzyme proved near impossible.

Fourier transform showed that, unsurprisingly, enzymes with more stiff domains operated at higher frequencies than those with more loose domains. Enzyme efficiency, however, seems to be a result of the geometric placement of different domains since an enzyme with 19 stiff domains could be as mediocre as 136 hits or as good as 253 hits. Thus a mediocre and good enzyme, possessing similar frequencies, would thus produce very similar frequency spectrums despite the large difference in catalytic efficacy. We needed to gain more insight into characteristics that existed within the enzyme population.

## 4.2 Extensive Data Mining

This followed the early analysis in order give us some direction in our search instead of constant probes in the dark. Three main data mining programs were designed, the first was created to give us a plot of average number of hits vs. the number of stiff domains (figure 7).

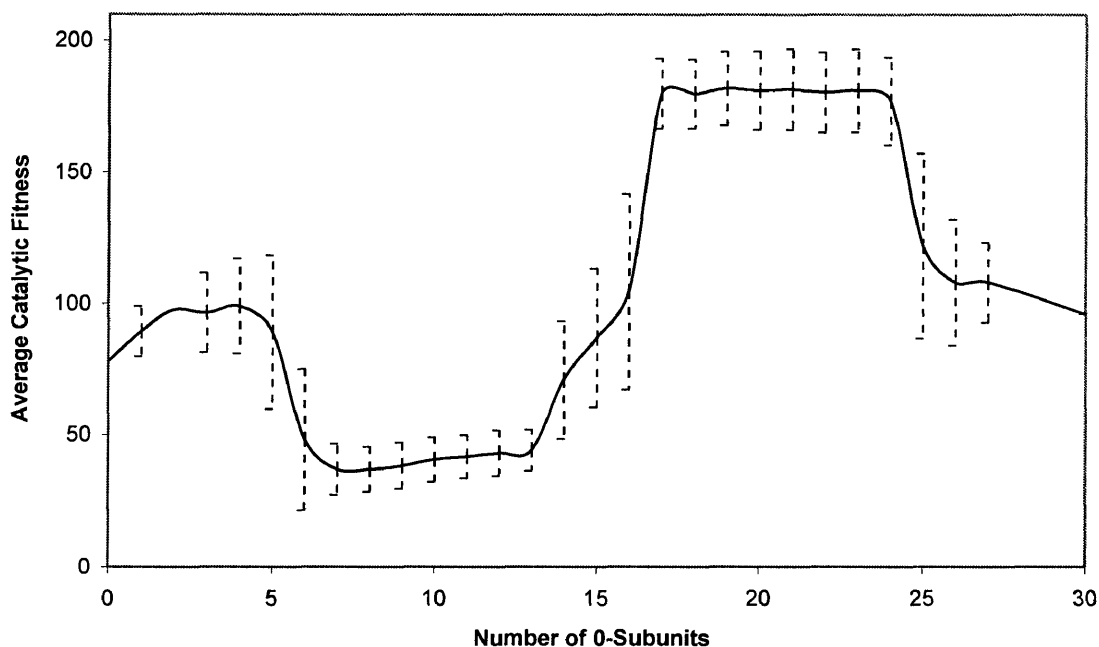


Figure 7: Average catalytic fitness as a function of the number of 0-subunits.

This plot told us several very important things, primarily, that the increase in catalytic efficacy is not just a factor of increasing the number of stiff domains and consequentially increasing the frequency but instead depends on an arrangement of domains creating beneficial global fluctuations. Secondly, it provided a visualization that optimum regions for both poor and good enzymes exist away from the two extremes of all stiff and all loose as well as not simply a 50/50 ratio of the two. Furthermore, the graph also showed us that there existed a large range in the

number of stiff domains that can produce either a very good or very bad enzymatic activity. This served to reinforce the conclusion that it is a combination of geometry, fluctuations, and frequency that is required for optimum results.

The next program was designed to provide insight into an unexpected phenomenon that had occurred. The introduction of a barrier force acting only as repulsive force (not containing the opposing restorative force present in a spring) had resulted in a stretching of the average C-S distance. Data showed that these values now ranged from 1.4 to 1.2 times the equilibrium distance rather than the expected 1.0.

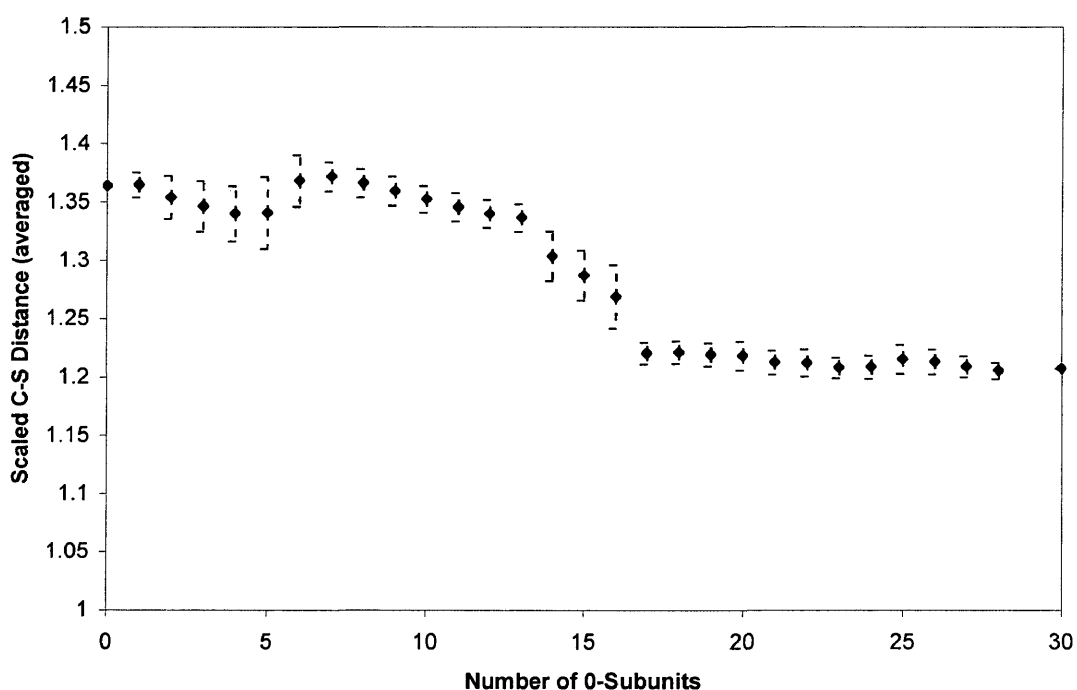


Figure 8: Scaled C-S distance as a function of the number of 0-subunits.

It is apparent from figure 8 that there is a region (>16 stiff) where the distance remains the same and once you get below this approximate 50/50 ratio the enzyme becomes loose enough to allow 'slack' from the potential energy barrier to be taken



up by the dynamic blocks. Even mostly stiff individuals still have an average scaled C-S distance of 1.2, the remaining 0.2 of 'slack' that exists even with all stiff domains was seen to be absorbed by the weak perimeter interactions of the phantoms and normals. This is an unexpected result from designing the perimeter forces to be very weak in order to simulate the large external 'thermal' fluctuations present in a real protein. Combining the knowledge obtained from figure 7 & 8 we realized the two plateaus in the first plot correspond to the fairly stable distance regions of the second plot.

The next question to be answered was How does this C-S distance relates to the number of catalytic events?. If a simple decrease in distance were allowing for the increase in catalytic efficacy this would be a fairly uninteresting result. It is important to first mention that, due to the chaotic nature of such a simple molecular dynamics system and our known error of around 7% in hit stability, it seemed reasonable to "bin" those individuals with similar fitness values together and average data over that "bin." For the remainder of this work all charts showing a data set in relation to the number of chemical events, individuals will be separated into bins of 5 hits. With this in consideration, the results of data-mining program are presented in figure 9, which shows the relationship between C-S distance and catalytic fitness.

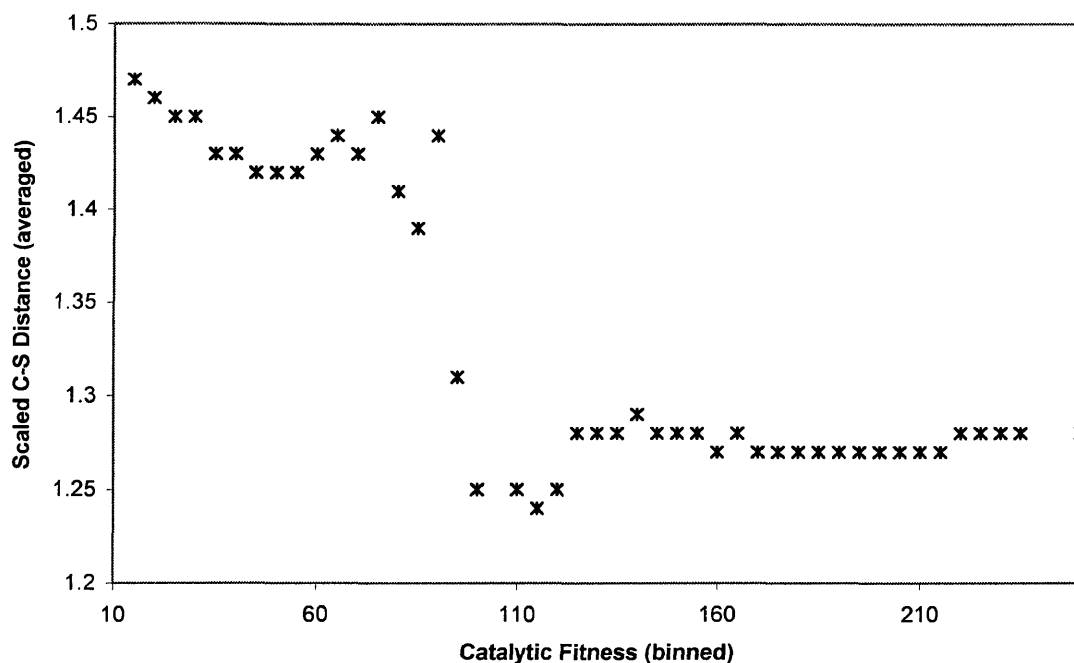


Figure 9: Scaled C-S distance as function of catalytic fitness. Data points are binned into groups of 5 hits and the distance is averaged. Standard deviation varies on this plot from around 0.05 to negligible values.

Interestingly after around 120 hits the average distance between the catalytic and substrate sites seemed to stabilize and level off. Something else must be occurring to allow for the continued increase in the number of chemical events with constant distance.

Another result from the data mining results was the realization that there were actually four different 'categories' of enzymes existing in our genetic algorithms.

Type 1 (best): Good Correlation :: Good C-S Distance	Type 2: Poor Correlation :: Good C-S Distance
Type 3: Good Correlation :: Poor C-S Distance	Type 4 (worst): Poor Correlation :: Poor C-S Distance

We now know that differences between these different types exist, but the detection of this 'good correlation' had still evaded us.

### 4.3.1 Design of Normal Mode Approximation

Since thermal noise made any analysis of position or velocity signals nearly impossible, we decided to attempt a normal mode analysis. Any normal mode analysis would have to exist only as an approximation for several reasons, the primary being the anharmonic nature of the potential energy barrier between the C and S residues as well as its “on” and “off” nature (i.e. it is only active between the equilibrium distance and 0.7 times the equilibrium distance). The normal mode analysis would consist of creating a normal mode matrix, which would be used to find the eigenvalues and corresponding eigenvectors for each system. The first step in this process is to populate the normal mode matrix with the proper values. An entry in the matrix consists of the second derivative of the potential energy equation evaluated at equilibrium, depending on the atom of reference, divided by the mass of the atom. To explain, consider a simple three-dimensional example.

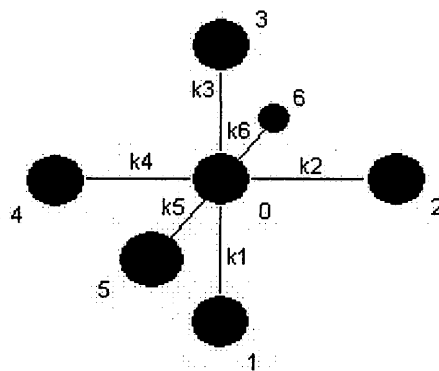


Figure 10: Geometry of a simple normal mode example. Atoms 5 and 6 extend out from the page.

Each atom is represented by a sphere with a corresponding number ranging from 0 to 6 and each potential (spring) is labeled from  $k_1$  to  $k_6$ . The Potential Energy ( $V$ ) of this system would be represented by the following equation.

$$V = \frac{1}{2} \sum_{i=1}^6 k_i [\{(x_i - x_0)^2 + (y_i - y_0)^2 + (z_i - z_0)^2\}^{\frac{1}{2}} - l_{eq}]^2$$

Equation 2: Potential energy of simple system.

With  $l_{eq}$  representing the equilibrium length of the spring between atom  $i$  and atom  $0$ .

In order to allow the second derivative of this equation to be time independent, a transformation is required, so we'll let the following be true;

$$\begin{aligned} x'_i &= x_i - x_{ie} \\ y'_i &= y_i - y_{ie} \\ z'_i &= z_i - z_{ie} \end{aligned}$$

Equation 3: Transformation to include displacement from equilibrium positions.

After substitution the potential energy equation looks like the following.

$$V = \frac{1}{2} \sum_{i=1}^6 k_i [\{(x'_i + x_{ie} - x_0 - x_{0e})^2 + (y'_i + y_{ie} - y_0 - y_{0e})^2 + (z'_i + z_{ie} - z_0 - z_{0e})^2\}^{\frac{1}{2}} - l_{eq}]^2$$

Equation 4: Equation 2 after substitution of terms from equation 3.

Combing  $x_{ie} - x_{0e}$  into a  $dx_{ie}$  term (equilibrium distance in the x direction) and the same substitution for y and z allows further simplification.

$$V = \frac{1}{2} \sum_{i=1}^6 k_i [\{(x'_i - x_0 + dx_{ie})^2 + (y'_i - y_0 + dy_{ie})^2 + (z'_i - z_0 + dz_{ie})^2\}^{\frac{1}{2}} - l_{eq}]^2$$

Equation 5: simplified version of equation 4.

The second derivatives, after substituting 0 for the deviation from equilibrium terms and either 0,  $l_{eq}$ , or  $-l_{eq}$  for the equilibrium distance terms, are simply a combination of  $k$  values for the self terms and a single negative  $k$  value for the mixed terms. The

resulting matrix of entries would be a  $(d*n)X(d*n)$  matrix (with  $n$  being the number of atoms (i.e. 7) and  $d$  being the number of dimensions (i.e. 3) thus a  $21x21$  matrix).

	x0	y0	z0	x1	y1	z1	x2	y2	z2	x3	y3	z3	x4	y4	z4	x5	y5	z5	x6	y6	z6
x0	$k_2+k_4$	0	0	0	0	0	- $k_2$	0	0	0	0	0	- $k_4$	0	0	0	0	0	0	0	0
y0	0	$k_1+k_3$	0	0	- $k_1$	0	0	0	0	0	- $k_3$	0	0	0	0	0	0	0	0	0	0
z0	0	0	$k_5+k_6$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	- $k_5$	0	0	- $k_5$
x1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
y1	0	- $k_1$	0	0	$k_1$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
z1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x2	- $k_2$	0	0	0	0	0	$k_2$	0	0	0	0	0	0	0	0	0	0	0	0	0	0
y2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
z2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
y3	0	- $k_3$	0	0	0	0	0	0	0	0	$k_3$	0	0	0	0	0	0	0	0	0	0
z3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x4	- $k_4$	0	0	0	0	0	0	0	0	0	0	0	$k_4$	0	0	0	0	0	0	0	0
y4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
z4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
y5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
z5	0	0	- $k_5$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$k_5$	0	0	0
x6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
y6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
z6	0	0	- $k_6$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$k_6$

Table 1: Hessian matrix for simple system shown in figure 10.

It is important to remember that this is only a seven-atom system while our system contains 450 atoms, with 168 of them being free to move. The resulting normal mode matrix is considerably more complicated having dimensions of  $504X504$ . Also important to note is that all atoms were connected directly by a single spring in the simple example, however, in the larger system there will be many atoms, which are not directly connected, thus their second derivatives will automatically be zero. Furthermore, since the barrier force present between C and S can not be simulated via normal mode, the  $\delta V_{C_x, S_x}$  (second derivative of the potential energy with respect to C and S in the x dimension) term is set to zero. The justification for this is that on average C and S are outside the effective range of this barrier so when C and S are at their average positions there is zero potential between them.

### 4.3.2 Design of Normal Mode Analysis Program

Performing a normal mode analysis on each individual in a population of over 24,000 by hand would simply be impossible. For this reason a program was designed to create the Hessian matrix for each individual in the GA's population, gather statistics, and report results for the entire population. In order to know anything useful about the motion of the system the eigenvalues and corresponding eigenvectors must be generated from the Hessian matrix. Since loading each individual into a math program such as Matlab or Mathematica would also be so time intensive as to make it impossible, the normal mode program must be linked with a scientific library capable of calculating the eigenvalues and eigenvectors. We chose the GNU Scientific Library or GSL (version 1.2) available from the GNU website (<http://www.gnu.org/software/gsl/>). "The GNU Scientific Library (GSL) is a numerical library for C and C++ programmers. It is free software under the GNU General Public License."<sup>23</sup>

The first step in this analysis was to create the Hessian matrix for every individual in the population. The geometry and spring constants were generated by loading the "mdUnequilibrated-\*.txt" file for each individual. Using the symmetry of the Hessian matrix to our advantage we divided the matrix entries into three classes.

- Class 1: Second derivatives of self terms for example  $\delta^2 V_{x_0, x_0}$  (second derivative with respect to the 0<sup>th</sup> atom in the x dimension). This position's value is simply the sum of all springs in the dimension being considered. So for  $\delta^2 V_{x_0, x_0}$  all the spring in the x direction would be summed.
- Class 2: Second derivatives for mixed terms for example  $\delta^2 V_{x_0, x_1}$  (second derivative with respect to the 0<sup>th</sup> and 1<sup>st</sup> atom in the x dimension). This position's value is simply the negative value of the spring (in the dimension of reference) between the two atoms. C and S are a special case and the mixed term

containing these two blocks in the x direction is set to zero (for explanation see normal mode approximation design section).

- Class 3: Second derivatives for mixed terms for atoms not directly connected by a spring. This position's value is automatically set to zero.

Using these classes as a guideline each entry in the matrix is created and stored in a two dimensional matrix.

Eigenvalues and eigenvectors are obtained for each individual from their respective matrices. Since each matrix is 504x504 there will be 504 eigenvalues each with an corresponding eigenvector containing 504 components. The eigenvalues are stored in a one-dimensional array and the eigenvectors in a two-dimensional array with columns represent which eigenvalues they correspond to and each row representing each atoms x, y, and z components. For the current system the x component of the catalytic residue (C) and the substrate residue (S) are contained in rows 249 and 252 respectively. Vector components for C and S are then analyzed for all eigenvalues. "Good" and "bad" eigenvectors are stored for all frequencies as well as for only the low frequencies. Low frequencies are defined as being the lower half of the frequency spectrum for an average individual. "Good" eigenvectors are those where C and S are anti-correlated or when one (C or S) is moving while the other is stationary. "Bad" eigenvectors are those where C and S are moving in a correlated fashion.

This program also maintains other important statistics for the population.

For each individual the following values are calculated and stored.

- a) Number of catalytic events.
- b) Number of good and bad eigenvectors for low frequencies.
- c) Number of good and bad eigenvectors for all frequencies.
- d) Average C-S distance.
- e) Number of stiff domains.
- f) Rms value for C residue.
- g) Rms value for S residue.

This allows the normal mode analysis to serve as an overall analysis and data-mining program for any important data.



### 4.3.2 Normal Mode Analysis Results

Our results show that the number of beneficial frequencies—“good” correlation—increases with catalytic fitness. Specifically, this increase in beneficial frequencies occurs in the lower frequencies, i.e. those below the dominant frequency.

To clarify, there are 504 eigenvalues each having a corresponding eigenvector. When looking at the Cx and Sx components of the eigenvectors only a very few have significant components. The number of eigenvectors with Cx and Sx components above our threshold (0.00001) remains fairly constant at around 8 vectors from a possible 504. This seemed at first to be somewhat uninteresting, however, realizing that cooperative motion is very often present in low frequencies while random uncoupled motion or “thermal noise” are naturally high frequency, we decided to isolate out low frequency vectors. Once we looked at the number of good correlated vectors present only in low frequencies we notice that there was an increase in the number of these vectors, however, a slightly inconsistent fashion as seen in figure 11.

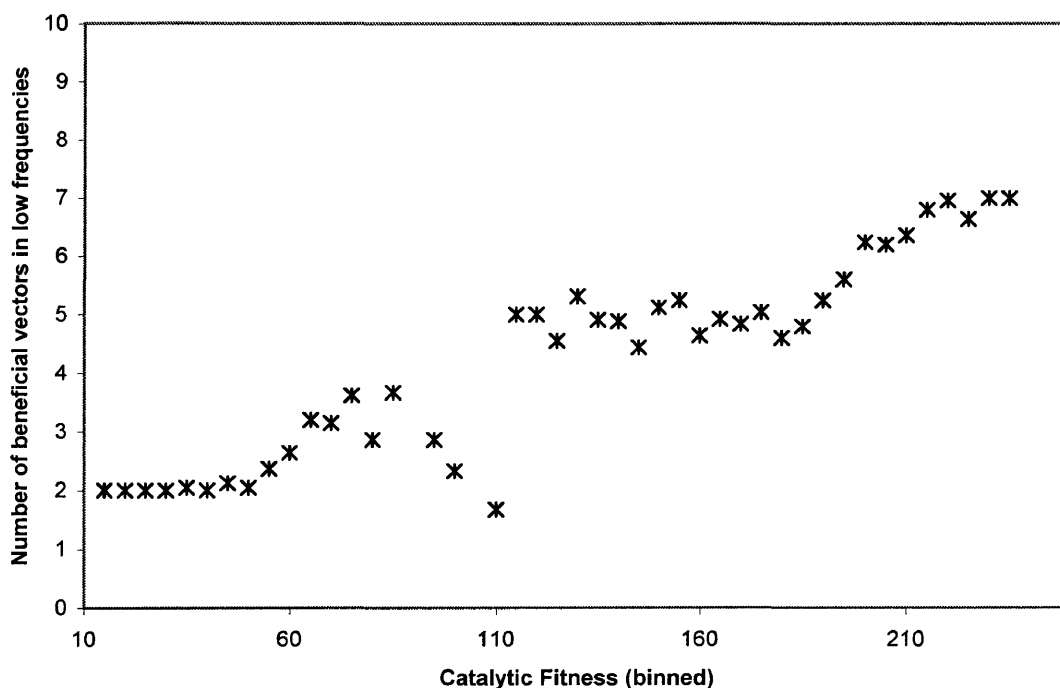


Figure 11: Relationship between the number of beneficial vectors found in low frequencies to catalytic fitness.

While the average number of beneficial eigenvectors present in all frequencies is constant across the entire population, this is not true for specific individuals or even small groups. This led us to believe that normalization was necessary. For any particular group of individuals there are a number of eigenvectors with C and S components (good or bad) above threshold as well as the number of beneficial low frequency eigenvectors. Dividing this number of beneficial low frequency vectors by the number of total vectors above threshold provides a “catalytic score” representing a percent optimum value.

When plotting this catalytic score in figure 12, the relationship between the number of catalytic events and the number of beneficial vectors in low frequencies is

much smoother than the pre-normalization plot (figure 11). For statistical reasons any fitness bins containing less than 5 individuals are removed.

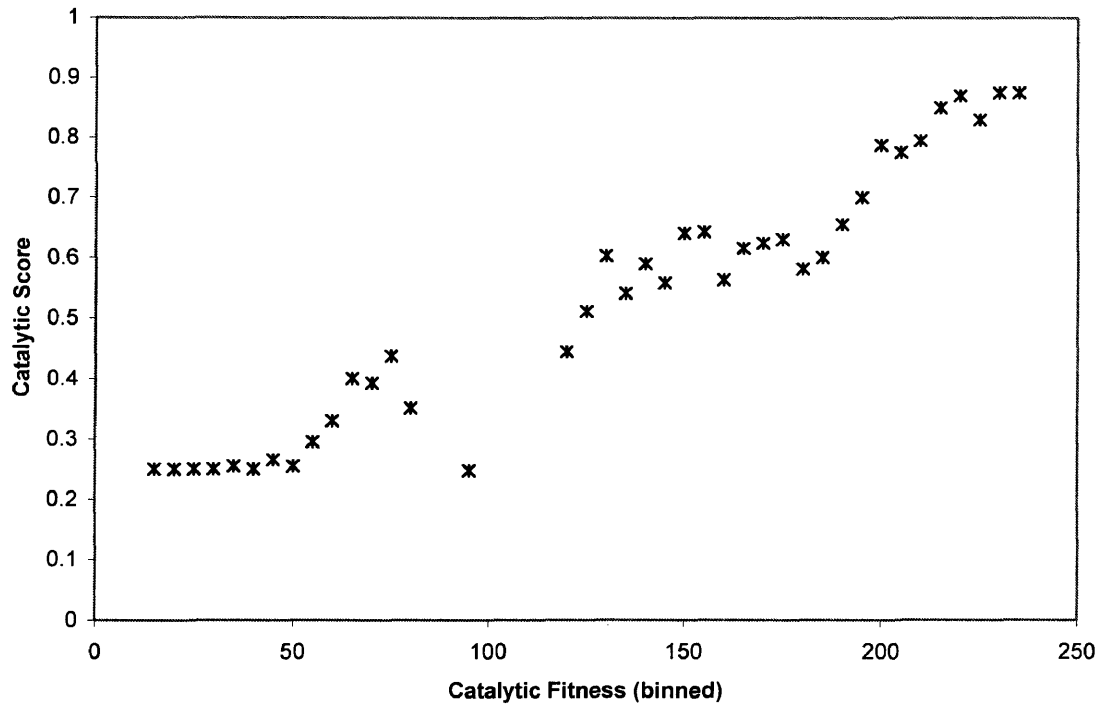


Figure 12: Relation of "catalytic score" to catalytic fitness.

Returning to the original four categories of enzymes discussed in the Data Mining section of chapter four, we can now better understand these four categories of enzymes. The following table shows examples of enzymes that fit the category requirements.

<p style="text-align: center;"><b>Type 1 (best): 253 Hits</b></p> <p>Good Correlation :: Good C-S Distance</p> <p><b>Enzyme Id:</b>            <b>10711652</b></p> <p><b>Catalytic Score:</b>    <b>0.88 or 88%</b></p> <p><b>Scaled C-S Distance:</b> <b>1.27</b></p>	<p style="text-align: center;"><b>Type 2: 112 Hits</b></p> <p>Poor Correlation :: Good C-S Distance</p> <p><b>Enzyme Id:</b>            <b>1024</b></p> <p><b>Catalytic Score:</b>     <b>0.08 or 8%</b></p> <p><b>Scaled C-S Distance:</b> <b>1.25</b></p>
<p style="text-align: center;"><b>Type 3: 92 Hits</b></p> <p>Good Correlation :: Poor C-S Distance</p> <p><b>Enzyme Id:</b>            <b>1073724415</b></p> <p><b>Catalytic Score:</b>    <b>0.92 or 92%</b></p> <p><b>Scaled C-S Distance:</b> <b>1.45</b></p>	<p style="text-align: center;"><b>Type 4 (worst): 16 Hits</b></p> <p>Poor Correlation :: Poor C-S Distance</p> <p><b>Enzyme Id:</b>            <b>1018998707</b></p> <p><b>Catalytic Score:</b>    <b>0.25 or 25%</b></p> <p><b>Scaled C-S Distance:</b> <b>1.45</b></p>

## Chapter 5

### 5.1 Discussion and Future Work

Conformation fluctuations (beneficial low frequencies) and active site orientation (catalytic distance) are key factors to evolving a high probability of chemical events and thus an effective catalyst. These beneficial frequencies and the distance between the catalyst and substrate sub-units are a function of the number and distribution of stiff, intermediate, and loose domains present in the model enzyme. Once the optimum active site orientation or distance is established evolution acts to increase the number of low frequency beneficial modes (i.e. modes where catalyst and substrate have an increased probability of overcoming the potential energy barrier between them). Furthermore, preliminary results show that single point mutations on residues distal from the active site have a profound impact on catalytic fitness, further supporting the belief that these dynamics are a global feature. Future work will entail detailed investigation into how point mutations influence catalytic efficacy and conformation fluctuations.

## BIBLIOGRAPHY

- 
- <sup>1</sup> Radzicka, A. and Wolfenden, R. 1995. Transition state and multistate analog inhibitors. *Meth. Enzymol.* 246: 284-312.
- <sup>2</sup> Mader, M.M. and Bartlett, P.A. 1997. Binding energy and catalysis: The implications for transition-state analogs and catalytic antibodies. *Chem. Rev.* 97: 1281-1301.
- <sup>3</sup> Kraut, J. 1988. How do enzymes work? *Science* 242: 533-540.
- <sup>4</sup> Wierenga, R.K., Borchert, T.V., and Noble, M.E.M. 1992. Crystallographic binding studies with triosephosphate isomerases: Conformational changes induced by substrate and substrate-analogues. *FEBS Lett.* 307: 34-39.
- <sup>5</sup> Greenwald, J., Le, V., Butler, S.L., Bushman, F.D., and Choe, S. 1999. The mobility of an HIV-1 Integrase active site loop is correlated with catalytic activity. *Biochemistry* 38: 8892-8898.
- <sup>6</sup> Illyin, V.A., Temple, B., Hu, M., Li, G., Yin, Y., Vachette, P., and Carter C.W. Jr. 2000. 2.9 Å crystal structure of ligand-free tryptophanyl-tRNA synthetase: Domain movements fragment the adenine nucleotide binding site. *Protein Sci.* 9: 218-231.
- <sup>7</sup> Kohen, A., Cannio, R., Bartolucci, S., and Klinman, J.P. 1999. Enzymatic dynamics and hydrogen tunneling in a thermophilic alcohol dehydrogenase. *Nature* 399: 496-499.
- <sup>8</sup> Ringe, D. and Petsko, G.A. 1999. Quantum enzymology: Tunnel vision. *Nature*. 399: 417-418.
- <sup>9</sup> Balabin, I.A. and Onuchic, J.N. 2000. Dynamically controlled protein tunneling paths in photosynthetic reaction centers. *Science*. 290: 114-117.
- <sup>10</sup> Alper, K.O., Singla, M., Stone, J.L. and Bagdassarian, C.K. 2001. Correlated conformational fluctuations during enzymatic catalysis: Implications for catalytic rate enhancement. *Protein Science*. 10: 1319-1330.

- 
- <sup>11</sup> Noonan, R.C., Carter, C.W. Jr., Bagdassarian, C.K. 2002. Enzymatic conformational fluctuations along the reaction coordinate of cytidine deaminase. *Protein Science*. 11: 1424-1434.
- <sup>12</sup> Cameron, C.E., and Benkovic, S.J. 1997. Evidence for a functional role of the dynamics of Glycine-121 of Escherichia coli dihydrofolate reductase obtained from kinetic analysis of a site-directed mutant. *Biochemistry*. 36: 15792-15800.
- <sup>13</sup> Miller, G.P., and Benkovic, S.J. 1998. Strength of an interloop hydrogen bond determines the kinetic pathway in catalysis by Escherichia coli dihydrofolate reductase. *Biochemistry*. 37: 6336-6347.
- <sup>14</sup> Benkovic, S.J., Rajagopalan, P.T. Ravi, Lutz, S. 2002. Coupling Interactions of distal residues enhance dihydrofolate reductase catalysis: mutational effects on hydride transfer rates. *Biochemistry*. 41:12618-12628.
- <sup>15</sup> Karplus, M. and Petsko, G.A. 1990. Molecular Dynamics Simulations in Biology. *Nature*. 347: 631-639
- <sup>16</sup> Wang, F., Li, W., Emmett, M.R., Hendrickson, C.L., Marshall, A.G., Zhang, Y.-L. Wu, L., and Zhang, Z.-Y. 1998 Conformational and dynamic changes of Yersinia protein tyrosine phosphatase induced by ligand binding and active site mutation and revealed by H/D exchange and electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Biochemistry*. 37: 15289-15299.
- <sup>17</sup> Alder, B.J., and Wainwright, T.E. 1957. Phase transition for a hard sphere system. *J. Chem. Phys.* 27: 1208-1209.
- <sup>18</sup> Klimov, D.K., and Thirumalai, D. 1998. Linking rates of folding in lattice models of proteins with underlying thermodynamic characteristics. *J. Chem. Phys.* 109: 4119-4125.
- <sup>19</sup> Hoang, T.X., and Cieplak, M., 1998. Protein folding and models of dynamics on the lattice. *J. Chem. Phys.* 109: 9192-9196.

- 
- <sup>20</sup> Socci, N.D., Onuchic, J.N., and Wolynes, P.G. 1999. Stretching lattice models of protein folding. *PNAS*. 96: 2031-2035.
- <sup>21</sup> Williams, P.D., Pollock, D.D., and Goldstein, R.A. 2001. Evolution of functionality in lattice proteins. *J. Mol. Graphics Mod.* 19: 150-156.
- <sup>22</sup> Venkatasubramanian, V., and Sundaram, A. 1998. Genetic Algorithms: Introduction and applications. *Encyclopedia of computational chemistry*. 2: 1115-1126.
- <sup>23</sup> Free Software foundation, Inc. Dec 2002. GSL – GNU Scientific Library.  
<http://www.gnu.org/software/gsl/>.



## VITA

### G. S. Blair Williams

Born in Staunton, Virginia, November 2, 1978. Graduated from Wilson Memorial High School and Central Shenandoah Valley Regional Governor's School in Fishersville, Virginia in June 1997. Earned a Bachelor of Science degree in Chemistry from the College of William and Mary, Williamsburg, Virginia in May 2001. A Master of Arts in Chemistry candidate at the College of William and Mary from January 2002 to May 2003. Course requirements and thesis have been completed.