LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-TR-218296

Mutations that Cause Human Disease: A Computational/Experimental Approach

P. Beernink, D. Barsky, B. Pesavento

January 20, 2006

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Final Report

Mutations that Cause Human Disease: A Computational/Experimental Approach (03-LW-017)

PI: Peter T. Beernink
Co-investigators: Daniel Barsky, Brad Pesavento

Background and Rationale

International genome sequencing projects have produced billions of nucleotides (letters) of DNA sequence data, including the complete genome sequences of 74 organisms. These genome sequences have created many new scientific opportunities, including the ability to identify sequence variations among individuals within a species. These genetic differences, which are known as single nucleotide polymorphisms (SNPs), are particularly important in understanding the genetic basis for disease susceptibility.

Since the report of the complete human genome sequence, over two million human SNPs have been identified, including a large-scale comparison of an entire chromosome from twenty individuals. Of the protein coding SNPs (cSNPs), approximately half leads to a single amino acid change in the encoded protein (non-synonymous coding SNPs). Most of these changes are functionally silent, while the remainder negatively impact the protein and sometimes cause human disease. To date, over 550 SNPs have been found to cause single locus (monogenic) diseases and many others have been associated with polygenic diseases. SNPs have been linked to specific human diseases, including late-onset Parkinson disease, autism, rheumatoid arthritis and cancer.

The ability to predict accurately the effects of these SNPs on protein function would represent a major advance toward understanding these diseases. To date several attempts have been made toward predicting the effects of such mutations. The most successful of these is a computational approach called "Sorting Intolerant From Tolerant" (SIFT). This method uses sequence conservation among many similar proteins to predict which residues in a protein are functionally important. However, this method suffers from several limitations. First, a query sequence must have a sufficient number of relatives to infer sequence conservation. Second, this method does not make use of or provide any information on protein structure, which can be used to understand how an amino acid change affects the protein.

The experimental methods that provide the most detailed structural information on proteins are X-ray crystallography and NMR spectroscopy. However, these methods are labor intensive and currently cannot be carried out on a genomic scale. Nonetheless, Structural Genomics projects are being pursued by more than a dozen groups and consortia worldwide and as a result the number of experimentally determined structures is rising exponentially. Based on the expectation that protein structures will continue to be determined at an ever-increasing rate, reliable structure prediction schemes will become increasingly valuable, leading to information on protein function and disease for many different proteins.

Given known genetic variability and experimentally determined protein structures, can we accurately predict the effects of single amino acid substitutions? An objective assessment of this question would involve comparing predicted and experimentally determined structures, which thus far has not been rigorously performed. The completed research leveraged existing

expertise at LLNL in computational and structural biology, as well as significant computing resources, to address this question.

Experimental Approaches and Methods

We have used several computational and experimental approaches to examine the relationships between protein sequence, stability and function. Each of the computational and experimental strategies was subdivided into high-throughput and detailed approaches. A schematic of the research strategy is shown in **Figure 1**.

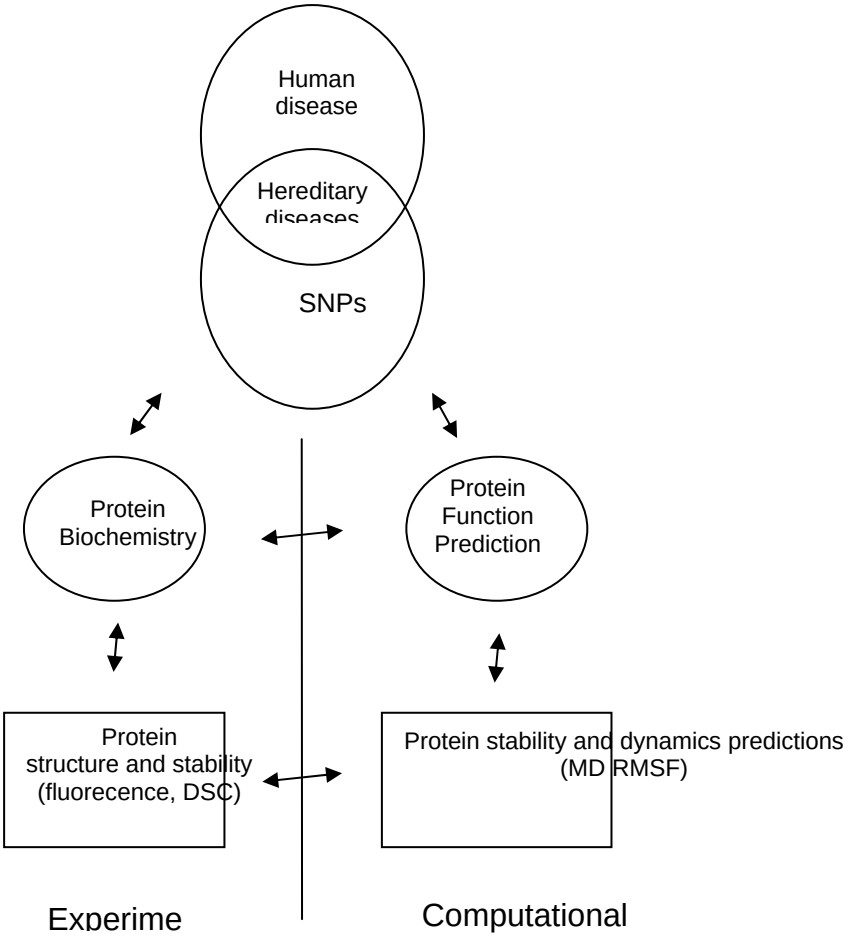


Figure 1. Outline of the complimentary computational and experimental approaches used in the project. A subset of mutations associated with hereditary diseases was the subject of experimental (left) and computational (right) assessments of protein stability.

The choice of proteins for these analyses was based on three criteria: 1) relevance to programmatic interests of DNA repair proteins in BBRP; 2) known associations between mutant gene variants (alleles) and cancer risk; and 3) known three-dimensional structures; and 4) available epidemiological data. The proteins chosen for study were Ape1 (**Figure 2**) and XRCC1 (**Figure 3**), both of which are involved in Base Excision Repair and have been the subject of cross-disciplinary research in BBRP, and p53, which is the protein most commonly mutated in a wide variety of human cancers.

Figure 2. A schematic representation of the human Ape1 DNA repair protein. A ribbon diagram illustrates the overall structure of the protein, the position of residues that are mutated in disease-associated variants are shown in red and the active-site metal ions are shown in yellow.

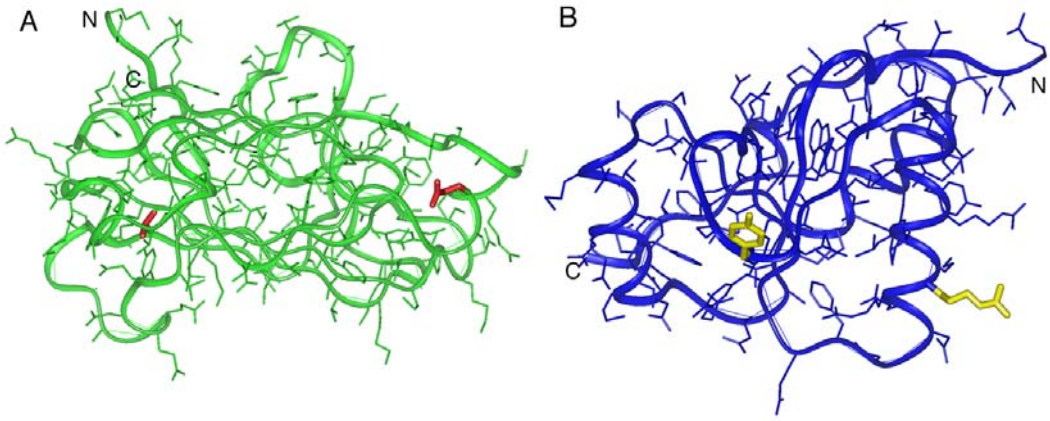


Figure 3. Structures of isolated domains of the human XRCC1 DNA repair protein. The known NMR structure of the N-terminal domain (panel A) and the crystallographic structure of the C-terminal BRCT domain are shown. Variant residues in each domain are depicted in red and yellow, respectively.

Computational Studies

In silico mutagenesis. For each mutation examined, we used the side-chain building program SCWRL, which is freely available, that uses statistical populations of amino acid side chain conformations (rotamers) in known crystal structures. Choice of the most common rotamer was iterated with a bump-energy calculation to identify the most probable side chain conformation that does not result in a steric clash. This step was followed by energy minimization using a molecular mechanics force field, which is implemented in the software package CHARMM. A rapid assessment of the impact of mutations on protein stability was performed with SCWRL.

*MD simulations***.** Molecular Dynamics simulations (MD) involved integrating the classical equations of motions of chemical systems using an empirically derived "ball and spring" model of the molecules. One use of MD was as an exploratory computational experiment to identify motions, interactions, and transitions that would otherwise not be expected or noticed. For example, in previous work we identified several new mutation-dependent interactions within the DNA repair enzyme Ape1, suggesting not only destabilization by reduced substrate-binding affinity. The other major use of MD is as an analytical tool to measure localized perturbations to protein stability.

Experimental Studies

*In vitro mutagenesis***.** For the genes proposed for experimental studies, the coding DNA sequence was obtained from the LLNL IMAGE cDNA collection. Mutant genes will be constructed by the overlap PCR method or by site-specific mutagenesis using the QuikChange Mutagenesis Kit (Stratagene, La Jolla, CA). The mutant genes were screened for protein expression and for rapid assessment of thermal stability using a novel fluorescence-based screen and by the biophysical method of scanning calorimetry for validation of the screen, which are outlined below.

Calorimetric studies. For detailed, quantitative studies of protein thermal stability, proteins were highly purified and analyzed by differential scanning calorimetry. The Ape1 protein was expressed in a native form and the XRCC1 domain fragments were expressed as hexa-histidine tagged fusions, which are readily purified. The stability studies were carried out using a highly sensitive MicroCal VP-DSC differential scanning calorimeter, which we have used in our earlier studies of other disease-causing mutant and engineered protein variants. For these experiments, purified protein (0.5 mg/mL) in phosphate buffer (20 mM $NaPO_4$, 100 mM NaCl, pH 7.0) was scanned from 10-90 °C at a scan rate of 60 °C using a medium feedback (gain) setting. The data were baseline subtracted, normalized for molar concentration and transition midpoints were determined using Origin software (MicroCal Inc., Northampton, MA).

Results

Computational studies

In the computational aspects of the project, we performed both high-throughput calculations and intensive MD simulations. We evaluated the ability to assess protein stability using high-throughput, computational methods on three test proteins with known three-dimensional structures and experimentally determined thermal stabilities, including T4 lysozyme, Staphylococcal nuclease and lactose repressor. For these proteins, single amino acid residue variants were used to examine the agreement between calculated and experimentally determined stability (~225 total variants).

Our initial method employed the SCWRL (Side Chain With Rotamer Library) program to measure only steric clash penalties. Several limitations of SCWRL are: 1) for surface residues, most substitutions do not produce a steric clash; 2) in the core of a protein, changing a residue to one of similar or smaller volume may not result in a penalty, even though the resulting cavity should decrease stability; and 3) solvation effects, which are thought to be critical, are not taken into account. Despite these limitations, when SCWRL was used to assess protein destabilization by core protein residue substitutions in T4 lysozyme, it correctly predicted destabilization in 18/24 cases (75%). Furthermore, SCWRL was able to predict increased stabilization by core residue substitutions in 2/3 cases (66%). Thus, a raw application of SCWRL is of some, limited predictive value. We also evaluated the presence of cavity forming

substitutions as a diagnostic of protein stability. First, buried residues were identified by calculating solvent accessible surface area of mutant residues and then, for buried residues, the internal volume change for the mutation was calculated. We found that the prediction of protein destabilization through the creation of internal cavities roughly correlated with the experimental data for T4 lysozyme and Staphylococcal nuclease. These results may benefit future studies that will employ distinct approaches for assessing potentially destabilizing mutations in the core versus surface mutations, which are probably more dependent on solvent interactions.

For the intensive computational studies, molecular dynamics simulations were carried out on seven naturally occurring variants of a human DNA repair protein, Ape1. Mutations were first built into the protein structure using the normal protein structure as a template. MD simulations identified specific regions of variant proteins that exhibited increased mobility as a result of the substitution. The differences in structure (indicated by position) and dynamics (indicated by color) are shown in **Figure 4**. These thermal motions also can be illustrated along a linear version of the protein sequence (**Figure 5**).

Figure 4. Backbone structure and dynamics of the Ape1 DNA repair protein and its disease-associated variants. The backbone structures are represented by the position of the traces and the rms fluctuations, corresponding to thermal mobility, are indicated in color, with green indicating regions of high mobility and red indicating regions of low mobility.

For five variants of Ape1, three of which are associated with increased cancer risk, we employed a novel analytical scheme, based on the collective motions of all backbone atoms during nanosecond molecular dynamics simulations. A significant advance was the finding that

the average backbone rms fluctuation (RMSF) from MD is highly correlated with global stability (see below).

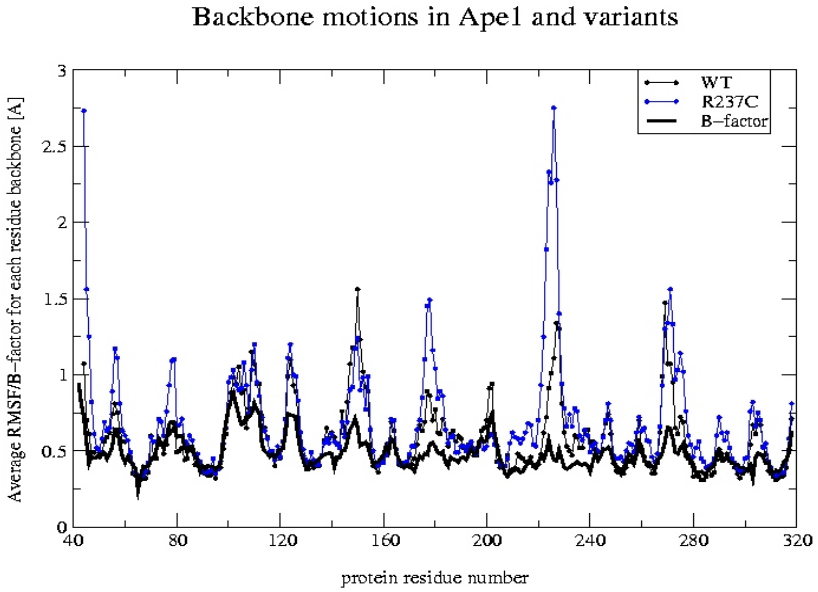


Figure 5. Backbone dynamics of Ape1 variant proteins. The rms fluctuations along the Ape1 backbone plotted versus residue number. More flexible, solvent-exposed loops of the protein are visible as peaks in the rms profile and more stationary regions constitute the baseline of the plot. The normal Ape1 protein (WT) is represented in thin black line, a relatively unstable variant Ape1 protein, R237C, is shown as a blue line and the crystallographic thermal (B) factor is shown as as thick, black line.

To generalize this result, we have extended our initial studies on Ape1 to another biologically important protein with many known cancer-causing variants, the p53 tumor suppressor protein. We conducted simulations of the normal p53 protein and four p53 variants for a minimum of 50 picoseconds. **Figure 6** illustrates differences between the active site structures of normal and variant p53 proteins as determined by MD simulations. These studies were partially completed, and have generated data that will provide preliminary data that will serve as a basis for follow-on funding as well as future studies.

Experimental studies

Construction of variant genes and proteins. Since most SNPs are predicted to affect protein stability, the objective of this component of the project was to evaluate the relationship between protein structure and stability. For a moderate number of human cancer-associated variant proteins, including human Ape1 and XRCC1, we constructed a DNA molecule encoding the variant protein. The proteins were expressed either in cell-free reactions for rapid assessment of protein stability based on a fluorescent reporter system or in *E. coli* for large-scale expression and subsequent protein purification.

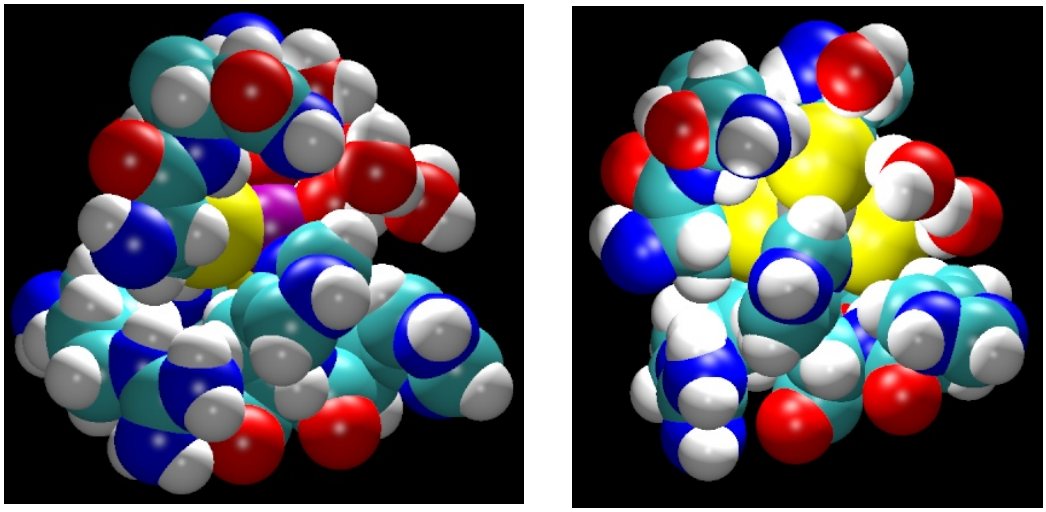


Figure 6. Structural differences in active-site conformations of the core domain of the p53 tumor suppressor protein. Simulations of the normal and C242S variant proteins were conducted for >80 picoseconds. Structures indicate a significantly different conformation in the region of the zinc-finger motif, resulting in exposure of the sulfur atoms (yellow).

High-throughput protein stability screen. We designed and implemented a rapid, protein-stability screen, which is readily scalable to a 96-well microplate format. This screening method is based on a carboxyl-terminal GFP fusion protein, (i.e. a query protein genetically fused to GFP), which allows GFP to report on the stability of the upstream sequence. This method relies on a previously established correlation between thermodynamic protein stability and the *in vivo* half-life. We have demonstrated this method on a destabilized Ape1 variant R237C, which exhibits a 7.3 °C reduction in melting temperature compared to the normal (wild-type) protein. In the fluorescence-based assay, this variant produces an approximate four-fold reduction in fluorescence intensity, which indicates that its *in vivo* stability is significantly reduced.

Calorimentric studies. For Ape1 variant proteins and XRCC1 domain fragments, sufficient amounts of purified proteins were obtained and measurements of their thermal stability were conducted. The results are shown graphically in **Figure 7** and are summarized in **Table 1**. Figure 7 illustrates the relatively high stability of the normal Ape1 protein (blue), which has a transition midpoint (Tm) values near 60 °C. This result contrasted with the relative instability of the disease-associated variant proteins P112L (green) and R237C (violet). Other Ape1 variants also were destabilized, though to a lesser extent (**Table 1**). The experimental results were compared to the predictions obtained through bump energy calculations and MD simulations (see below).

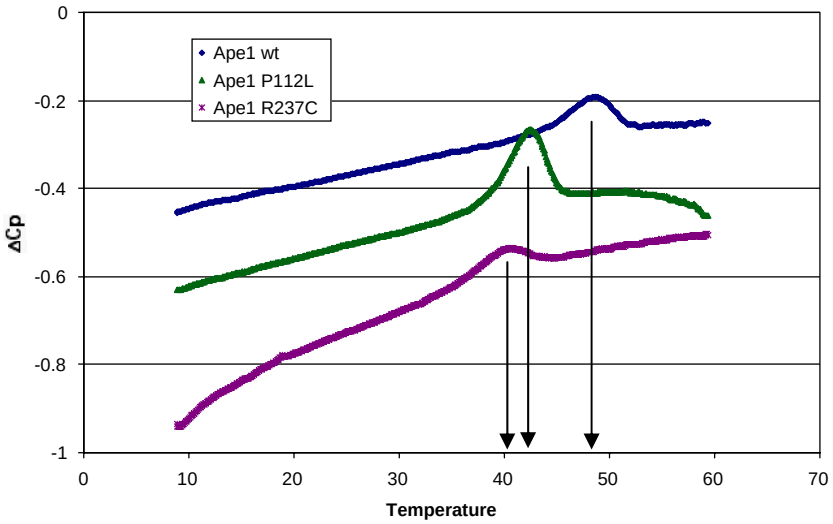


Figure 7. Calorimetric measurements of variant protein stability. The Ape1 DNA repair protein was purified from *E. coli* and scanned in a highly sensitive microcalorimeter. The change in heat capacity (ΔCp) is plotted versus temperature and a peak in the profile corresponds to an endothermic transition. The normal protein unfolds with a transition midpoint temperature (Tm) of 48 °C, whereas the variant proteins unfold with a Tm value of 41-42 °C.

Table 1. Calorimetric transition temperatures of Ape1 variant proteins.

Protein	**T_m[a] (C)**	**n[b]**	**ΔT_m[c]**
Ape1	47.7 (±1.2)	4	---
L104R	41.3	1	-6.6
P112L	42.4 (±0.2)	5	-5.3
E126D	44.1	1	-3.8
D148E	42.8	2	-5.1
R237A	34.9 (±0.8)	3	-12.8
R237C	40.4 (±0.5)	5	-7.3

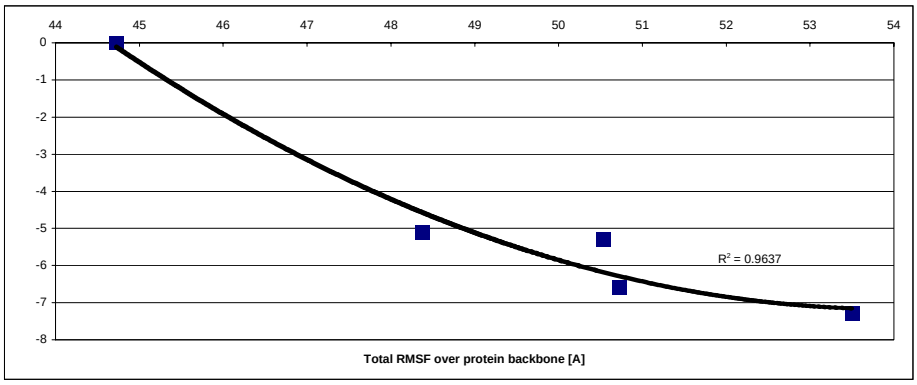
[a] transition midpoint
[b] number of experiments
[c] change in T_m relative to wild-type Ape1

Correlation of computational and experimental studies

Through comparison of our computational and experimental studies, we have shown a close correlation between theory and experiment. The average backbone rms fluctuations were shown to be correlated with the thermal stability of the protein, as indicated by the calorimetric thermal transition temperature (**Figure 8**). This correlation has been further validated using two additional Ape1 variants, which are not depicted in the Figure.

Based on a physical concept known as the Lindemann Criterion[1], we have derived a physical model for this relationship. This result is of considerable significance, since experimental studies are labor-intensive and computational power is continually increasing. Thus for disease-causing proteins of known (or potentially modeled) three-dimensional structure, an evaluation of the effects of sequence variation on the protein is possible. Since many regions in the protein can influence its stability, whereas relatively few contribute directly to functional activity, this type of analysis could identify variants that impact protein stability and indirectly on protein function.

To provide a greater number of data points to support this correlation and to prove that the method is general for proteins, we have extended our earlier studies through simulation of the single most important cancer-causing protein, p53. This protein has a known three-dimensional structure and 36 known variants. We have performed MD simulations on five variants of this protein and have compared the results with those obtained previously on Ape1. Following side-chain building and energy minimization using CHARMM, MD simulations (ca. 10^4 atoms including solvent) are performed for a duration of 1 ns. The MD trajectories are analyzed using CHARMM to obtain the RMS fluctuation for backbone atoms over the entire simulation. The resulting data on 5 or more variants of three different proteins will represent a thorough assessment of this potentially powerful computational tool.



1

$$T = \frac{E}{3\rho k_B} < \Delta x^2 >$$

where E is Young's Modulus, ρ is protein density and x is the RMSF

Figure 8. Correlation of experimentally measured and computationally derived protein stability. Experimental stability (ΔT_m) is plotted on the ordinate and mean backbone RMSF is shown on the abscissa. The R^2 correlation value for a quadratic fit is 0.96.

Outcomes

In summary, we made substantial advances in both computational and experimental components of this project. Through this LDRD project, we have demostrated an association between disease-causing protein variants and thermal stability. Furthermore, we have shown a close correlation between experimentally measured and computationally derived protein stability and thus have developed a novel use of MD in quantifying changes in protein stability due to mutations.

One manuscript based on this work has been published (Beernink, PT et al., J. Biol. Chem. (2005) Aug 26; 280(34): 30206-13. Two additional manuscripts are in preparation (Hadi et al., 2006; Barsky et al., 2006) and will soon be submitted to peer-reviewed journals (projected date early 2006). One grant proposal has been submitted to NIH (on 7/1/04) and two grant proposals have been submitted to DOD Congressionally Directed Medical Research Programs (CDRMP) for breast and prostate cancer-related research. At least one additional grant proposal based on this work is planned.