



LAWRENCE  
LIVERMORE  
NATIONAL  
LABORATORY

UCRL-ID-147129

# **Computational Biology A Strategic Initiative LDRD**

**D. Barsky, M. Colvin**

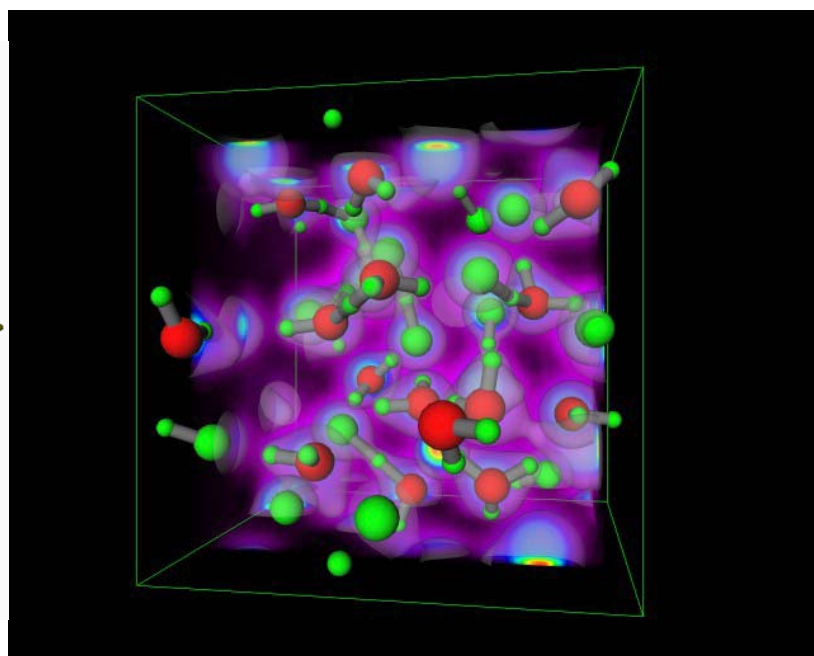
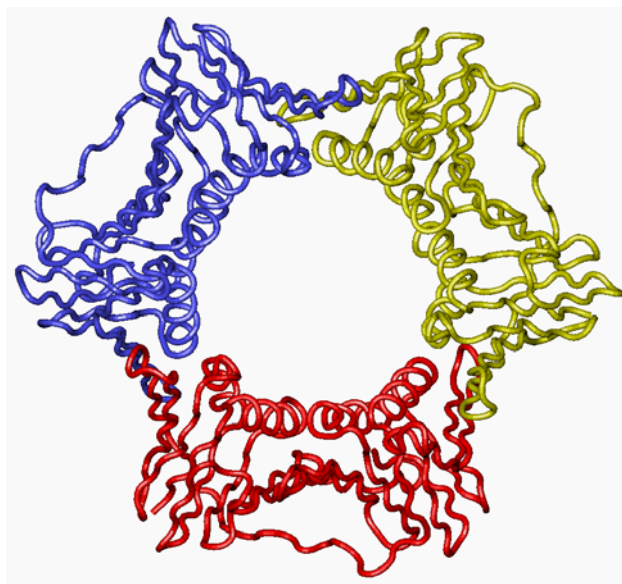
**February 7, 2002**

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express 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 the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

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.

# Computational Biology

A Strategic Initiative LDRD



98-SI-008

Final Report for FY98 – FY00

UCRL-ID-147129

## Table of Contents

<b>1. INTRODUCTION</b>	<b>1</b>
<b>1.1 Relevance to the Laboratory</b>	<b>1</b>
<b>1.2 Research Overview</b>	<b>2</b>
<b>2. ACCOMPLISHMENTS FY98 – FY00</b>	<b>3</b>
<b>2.1 Computational Biochemistry</b>	<b>3</b>
<b>2.1.1. Food Mutagens</b>	<b>4</b>
2.1.1.1. Reaction site on guanine of activated food mutagens	
2.1.1.2. Nature of the immediate chemical precursor to DNA adduction by food mutagens	
2.1.1.3. Quantum chemical studies of cytochrome P450 mechanisms	
2.1.1.4. Development of an active site model for CYP1A2	
2.1.1.5. Computational spectroscopy of food mutagens	
2.1.1.6. Predictions of high potency mutagens	
2.1.1.7. Computational spectroscopy of food mutagens	
<b>2.1.2. DNA-Alkylating Anticancer Drugs</b>	<b>18</b>
2.1.2.1. Elucidation of alternative cyclization products from phosphoramidate mustard	
2.1.2.2. Predictions of activation energies for novel chemical analogs	
2.1.2.3. Quantum chemical simulations of the DNA crosslinking reaction	
2.1.2.4. Molecular dynamics simulations of crosslinking reactions <sup>2</sup>	
<b>2.1.3. Abasic DNA Repair</b>	<b>26</b>
<b>2.1.4. Non-polar DNA Base</b>	<b>29</b>
<b>2.1.5. Parallel Stranded DNA</b>	<b>31</b>
<b>2.1.6. Prediction of Protein Binding Ligands—Computational Docking</b>	<b>32</b>
2.1.6.1. Docking example 1—tetanus toxin (TeNT)	
2.1.6.2. Docking example 2—apuridinic/apyridimic endonuclease 1 (Ape1)	
<b>2.1.7. Solution Structure of the Ligase III BRCT domain</b>	<b>36</b>
<b>2.2 Advanced Computational Chemical Methods</b>	<b>37</b>
<b>2.2.1. Applications of First Principles Molecular</b>	<b>37</b>
2.2.1.1. Liquid water simulations	
2.2.1.2. Ion solvation	
2.2.1.3. Conformational dynamics of dimethyl phosphate	
2.2.1.4. Other applications of first principle molecular dynamics	
<b>2.2.2. Development of First Principles Molecular Dynamics</b>	<b>44</b>
2.2.2.1. Algorithm development	
2.2.2.2. JEEP code development	
<b>2.3 Protein Structure Prediction</b>	<b>47</b>
<b>2.3.1. Homology-based Protein modeling</b>	<b>47</b>
2.3.1.1. External evaluation of our homology-based protein modeling methods	
2.3.1.2 Homology-based structure prediction of DNA sliding clamp proteins	
2.3.1.3 Homology based structure predictions of DNA clamp loading proteins	
<b>2.3.2. Methods Development for Protein Fold Recognition</b>	<b>50</b>
2.3.2.1. Methods for accurate superposition of protein structures	

2.3.2.2. Methods for the Identification of Distant Protein Homologies	
2.4 Computational Gene Discovery	53
3. APPENDIX	55
3.1 Publications—In Print, in Press, or Submitted	55
3.2 Selected Talks, Posters, and Published Abstract	58
3.3 Project Staff	61
3.3.1 LLNL Scientists	61
3.3.2 Students Supported	62
3.4 Software Developed	63
3.5 Outside Grants Funded or Submitted for Continued Funding	64
3.6 Collaborations	65
3.6.1 Internal Collaborations	65
3.6.1 External Collaborations	65

## **1. Introduction**

The goal of this Strategic Initiative LDRD project was to establish at LLNL a new core capability in computational biology, combining laboratory strengths in high performance computing, molecular biology, and computational chemistry and physics. As described in this report, this project has been very successful in achieving this goal. This success is demonstrated by the large number of referred publications, invited talks, and follow-on research grants that have resulted from this project (see appendix). Additionally, this project has helped build connections to internal and external collaborators and funding agencies that will be critical to the long-term vitality of LLNL programs in computational biology. Most importantly, this project has helped establish on-going research groups in the Biology and Biotechnology Research Program, the Physics and Applied Technology Directorate, and the Computation Directorate. These groups include three laboratory staff members originally hired as post-doctoral researchers for this strategic initiative.

### **1.1 Relevance to the laboratory**

Lawrence Livermore Laboratory has a long history of excellence in computational simulation, going back to the earliest days of electronic computers. Notable past achievements at the lab include the computational discovery of two dimensional phase transitions and the development of the first widely used non-linear structural dynamics software. LLNL is presently the world leader in high performance computing and related computational sciences, and is the site for the world's fastest supercomputers. The role of simulation is also a central theme in the LLNL Strategic Plan, which identifies computational simulation as a "critical technology" for 19 of the 22 "major activities" at the lab.

The biological sciences, including molecular biology, genomics, and biotechnology, have also have a growing role at LLNL. Indeed, expanding the laboratory's biology and biotechnology research efforts was one of the primary recommendations of the LLNL *Long Range Strategy Planning Project*. Several LLNL directorates are now involved in biology-related research and the biosciences are expected to play a role in an increasing number of laboratory programs. With the broadening national security missions at the laboratory, especially since the biological attacks in October, 2001, the areas of biological weapons non-proliferation and defense are also emerging as very significant programs at the lab.

Given the laboratory's expertise in the computational sciences and growing involvement in biology, computational biology combines natural strengths of the laboratory in this rapidly emerging new field of biosciences. No field offers a greater combination of challenge and promise for computer simulation than biology. Biochemical processes involve subtle, low-energy reactions and complex interactions of large macromolecules; yet predictive biochemical simulations could lead to breakthroughs with profound impact on human health, environmental protection, engineered biomaterials, and, of great current relevance, to defense against biological threats.

This project has yielded many benefits to the laboratory. Firstly, the research in this project has earned recognition and visibility for the new LLNL capabilities in computational biology, both within the DOE system and in the larger research community. Secondly, the research performed under this project provided the foundation and preliminary data for successful research grant applications to the DOE and to the National Institutes of Health. Thirdly, the expertise in computational docking and protein structure prediction developed under this project has found application in the development of bioweapon detection technologies and in elucidating the virulence mechanisms of potential biowarfare pathogens. Fourthly, this project has brought new scientific talent to the laboratory, including three LLNL staff scientists originally hired as post-doctoral fellows for this project. Finally, the molecular simulation software originally developed with support from this project, in particular the first principles molecular dynamics software, JEEP, has been used in other LLNL applications, ranging from interpreting shock physics experiments to designing new solid state materials.

## **1.2 Research Overview**

The scientific goal of this project was to develop and apply a wide range of computational biology methods to help solve significant biological problems. To this end, during this Strategic Initiative we were involved in a wide range of collaborative biological research projects. The goals of these collaborative studies have included: understanding the mechanisms of enzymes that repair damaged DNA, elucidating the mechanisms of how food-borne mutagens damage DNA, developing anticancer drugs with improved therapeutic effectiveness, and determining the structures of proteins that bind to DNA to modulate a cell's life cycle. These research efforts have produced a large number of refereed publications and

established the necessary expertise and collaborations to provide continued funding for these projects.

These applications depended on the development of new methods for computational biology. Our primary development goal has been to create methods to simulate the mechanisms of biochemical processes using an accurate quantum mechanical description of the molecular interactions. This method are referred to as "first principles molecular dynamics" to distinguish it from less-accurate classical molecular dynamics. This approach was successfully implemented on massively parallel computers and has been used to study a number of chemical systems related to DNA structure and enzyme mechanisms. These simulations are the most accurate ever performed on such biochemical structures and depended on LLNL's TeraFLOP supercomputers. We also developed improved methods for predicting protein structures based on their sequence similarity to proteins of known structure. These new methods depend on the use of structural information derived from the database of experimental protein structures to improve the reliability of fold prediction and to provide full atomic resolution in the resulting models.

In summary, this Strategic Initiative LDRD Project has been successful in creating a new research capability at the laboratory in computational biology. This project has brought together diverse human talent and laboratory resources to apply advanced simulation methods to solve biology questions. The publications, conference presentations and community visibility arising from this project have lead to a sustained LLNL program in the important new discipline of computational biology.

## **2. ACCOMPLISHMENTS FY98 – FY00**

### **2.1 Computational Biochemistry**

The goal of the computational biochemistry has been to collaborate with experimental biologists on projects where advanced chemical simulations could have a significant impact on biological understanding. Over the course of the SI we expanded the number of collaborations and published a large number of papers in life sciences journals, thus demonstrating the effectiveness of joint experimental/theoretical approaches to elucidating biological questions. These projects are described in the following numbered



sections. For each project at least one paper has been published or submitted for publication in the relevant biological literature.

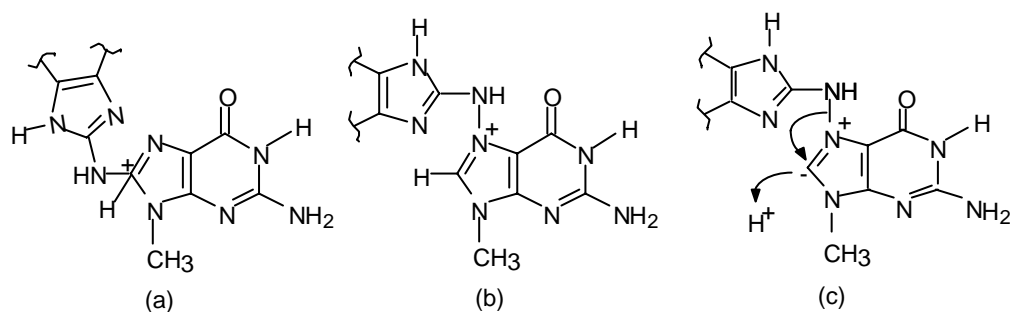
### **2.1.1. Food Mutagens**

The 2-aminoimidazole-azaarene (AIA) compounds formed during the high temperature cooking of meats have been shown to be both mutagenic and carcinogenic. Daily dietary exposure to these compounds may, therefore, represent a significant portion of the cancer risk associated with these foods. The AIA compounds share a common, mutagenically active 2-aminoimidazole group, however they exhibit a 100-million fold range of mutagenic potency. The mechanisms underlying this wide distribution in activity are not known, although quantitative structure activity relationships (QSARs) and preliminary experimental data indicate that at least some of the variation in mutagenic potency results from different rates of metabolic activation. Identifying the factors that modulate mutagenic potency gives us a means to predict the relative human health risks associated with exposure to these food mutagens. In our collaboration with the food mutagen group in the BBRP, we have completed several computational projects that address crucial activation steps in the progression of AIA-induced carcinogenesis. Our roles, described in more detail in the following sections, include 1) performing chemical simulations of the chemical reactions leading to DNA damage 2) using protein structure prediction methods to develop an active site model for the enzyme responsible for initial metabolic activation of the AIA mutagens, 3) modeling of AIA mutagens docking into this model enzyme active site, 4) developing structure-activity relationships for bioflavonoid compounds found in the diet that may reduce cancer risk, 5) using computational modeling to design very high potency mutagens to test mechanistic hypotheses about the food mutagens, and 6) performing quantum chemical simulations to predict spectroscopic properties of food mutagens and their metabolites to assist in analytical studies. This work formed the basis for a new program project for the NIH PO1 Food Mutagen Grant that has been recently funded (see Appendix: Grants)..

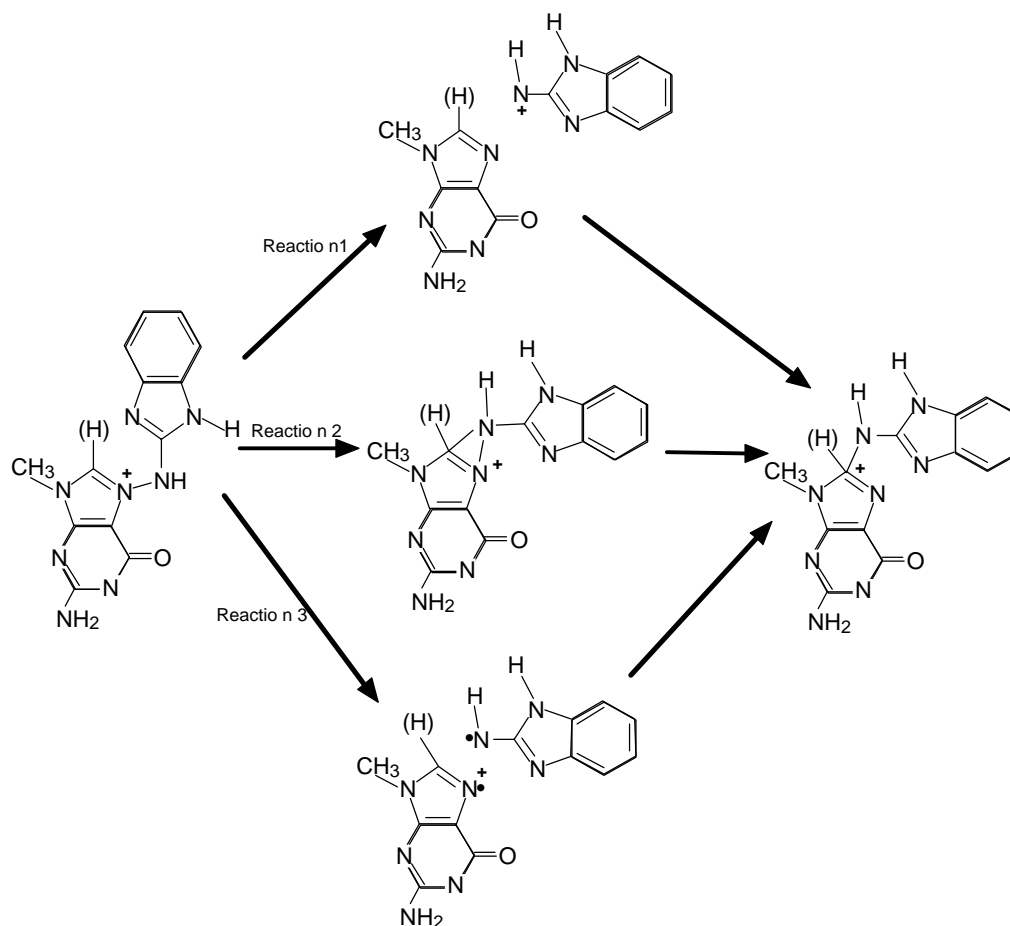
#### **2.1.1.1. Reaction site on guanine of activated food mutagens**

One central question involves the initial binding site on DNA of the nitrenium metabolite that is thought to be the final activated mutagen of the aromatic amine

mutagens. For the AIA food mutagens, the C<sub>8</sub> position on the DNA base guanine is well established as the final attachment site (structure (a) in scheme below). However, some recent studies suggest that the guanine N<sub>7</sub> may be a more likely site for initial attack (structure b), which is then followed by a 1-2 shift to the C<sub>8</sub> position (structure c).



We have used high-level quantum chemical calculations to study this binding reaction for several target compounds, including both small model compounds larger, more biologically relevant chemical structures. We found that for several food mutagens, the guanine C<sub>8</sub>- and N<sub>7</sub>-bound forms are nearly equal in energy, so that either may be formed. Hence, kinetic effects or factors in the DNA environment could favor an initial attack at N<sub>7</sub>. However, even if the N<sub>7</sub> adduct is formed, it will only be biologically relevant if there is an N<sub>7</sub> to C<sub>8</sub> chemical pathway that is lower in energy than simply dissociating back into the original nitrenium and neutral guanine reactants. Otherwise, the N<sub>7</sub> adduct will not be in the direct mutation pathway, and the rate of formation of the final C<sub>8</sub> adduct will depend on the reaction of the nitrenium with the C<sub>8</sub> position. In order to determine if such a low-energy N<sub>7</sub> to C<sub>8</sub> pathway exists, we have studied several chemical pathways connecting the two adduct forms for both small model compounds and guanine-mutagen complexes. For example, the following scheme shows three possible reaction pathways connecting the N<sub>7</sub> to the C<sub>8</sub> binding sites.

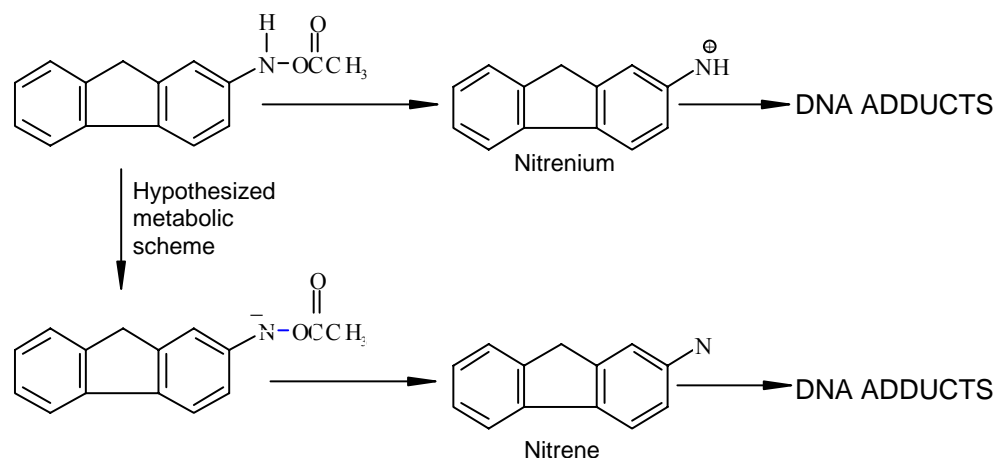


These quantum chemical results indicate that the favored pathway depends strongly on the protonation state of the guanine C8 position (shown in parentheses). If this proton is lost before the 1-2 shift occurs, the radical-pair dissociative pathway (3) is strongly favored over the reverse binding pathway (1), or the concerted reaction (2). If the proton loss occurs after the shift, as is suggested by experiments, then the concerted reaction pathway (2) is favored by a moderate amount of energy over the concerted reaction and is much more favorable than the radical-pair pathway. These results provide information to better explain the relationship between the chemical properties of the food mutagens and the experimentally observed mutagenic potency.

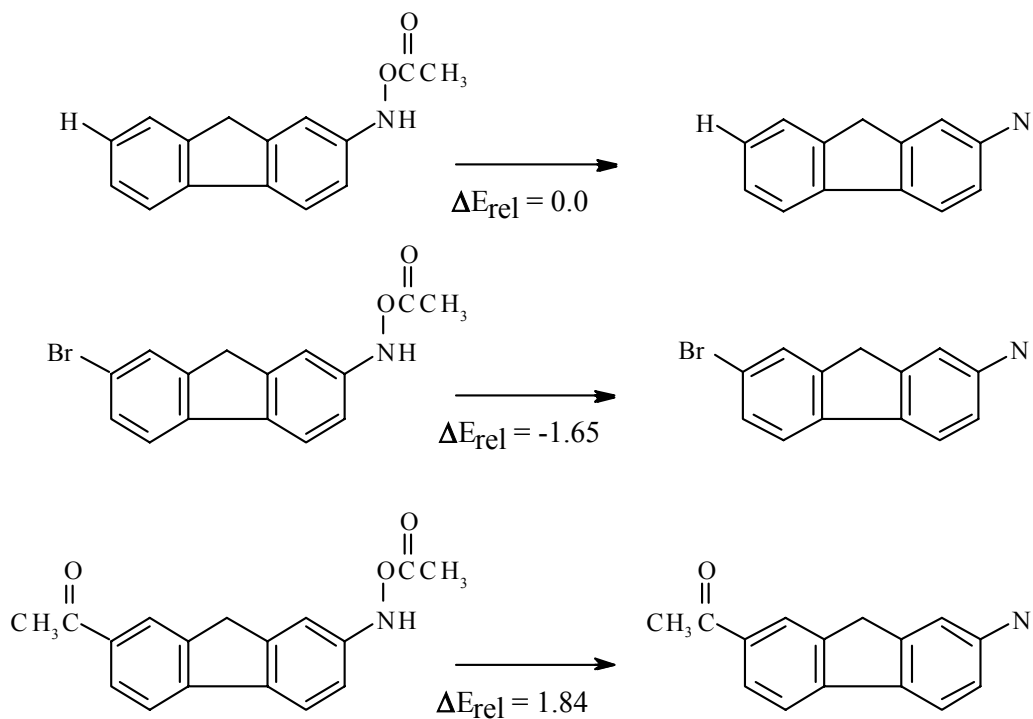
#### 2.1.1.2. Nature of the immediate chemical precursor to DNA adduction by food mutagens

Another aspect of our food mutagen project is the investigation of the actual chemical species leading to DNA mutations. Several lines of evidence indicate that the nitrenium ion is the ultimate metabolite in aromatic amine bioactivation. However,

experimental studies on a series of 7-substituted *N*-acetoxyarylamines generated from *N*-hydroxy-2-acetylaminofluorenes have suggested that, for certain isomers, metabolic activation to an uncharged nitrene species may be a more energetically favorable pathway. In particular, their results have indicated that substitution of an acetyl group at the 7-position of 2-(*N*-acetoxyamino)fluorene increases the electronic stability and results in preferential formation of the nitrene, versus nitrenium species as shown in the following scheme.

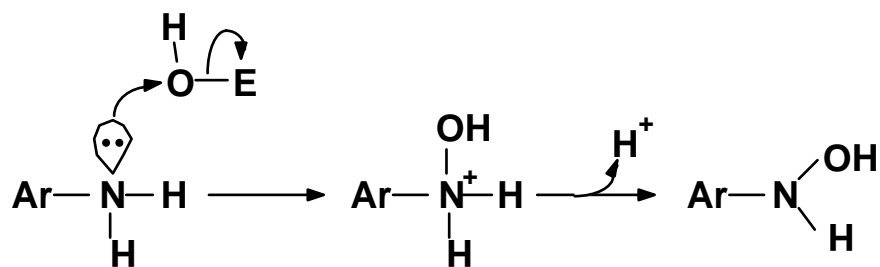


In order to determine the feasibility of this hypothesis, we performed both density functional theory (DFT) and second order Møller-Plesset Perturbation theory (MP2) calculations to compare the relative energies of the nitrene and nitrenium species for three *N*-acetoxyarylamines; unsubstituted 2-(*N*-acetoxyamino)fluorene (AoAF), 7-bromo-2-(*N*-acetoxyamino)fluorene (7-Br-AoAF) and 7-acetyl-2-(*N*-acetoxyamino)fluorene (7-acetyl-AoAF). The calculations indicate that acetyl substitution at the 7-position on AoAF produces nitrene species which may be comparable, but not more energetically stable, than nitrenes produced from unsubstituted or brominated AoAF derivatives. Therefore, in contrast to the experimentally raised hypothesis, these molecular orbital calculations suggest that a nitrene metabolite is not indicated in the bioactivation of 7-acetyl-2-(*N*-acetoxyamino)fluorene. The following scheme shows the computed DFT relative nitrene-formation energies for the different 7-substituted AoAF compounds. The MP2 calculations yielded very similar results.

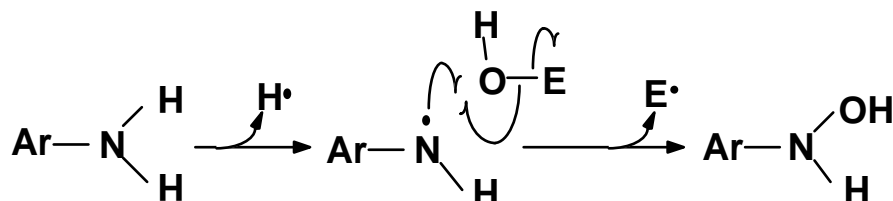


### 2.1.1.3. Quantum chemical studies of cytochrome P450 mechanisms

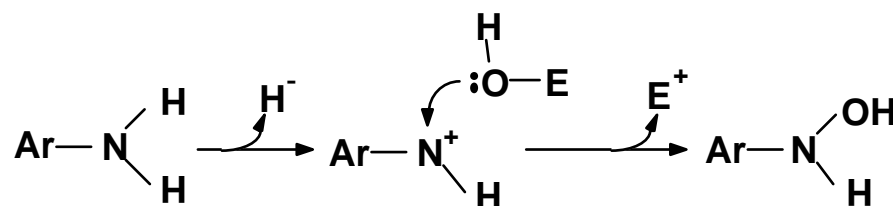
The initial activation step in the mutagenic pathway of the AIA compounds is the enzymatically catalyzed oxidation of the exocyclic amino group at the 2-imidazole position. This mechanism of N-oxidation is not fully understood. In order to explain the measured oxidation rates of the individual substrate structures, it is necessary to understand this chemical mechanism. The three mechanisms that are chemically plausible for this oxidation reaction have been proposed and are shown in the figure below.



(a) Addition/rearrangement mechanism



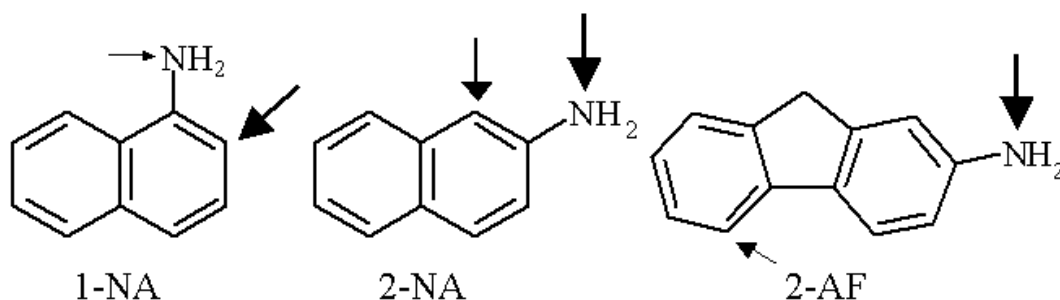
(b) One electron transfer



(c) Two electron transfer

Three proposed mechanisms for CYP1A2 catalyzed oxidation of the HA exocyclic amines. a) The addition-rearrangement mechanism is not consistent with either Huckel or DFT charges. b) The one electron transfer mechanism is consistent with DFT charges that indicates the aminyl radical charge distributions correlate well to experimentally observed N- versus C-oxidation product ratios. c) The two-electron transfer mechanism is consistent with Huckel charge data, but not with the DFT charges.

In order to discern which mechanism occurs, earlier researchers measured the N-oxidation and C-oxidation (ring oxidation) rates of three arylamines that are known to have the same mechanism of action as the AIA food mutagens: 1-naphthylamine (1-NA), 2-naphthylamine (2-NA), and 2-aminofluorene (2-AF). 1-NA was almost exclusively ring oxidized, 2-AF was almost exclusively N-oxidized, and 2-NA exhibited nearly equal amounts of N and ring oxidation (see figure below):



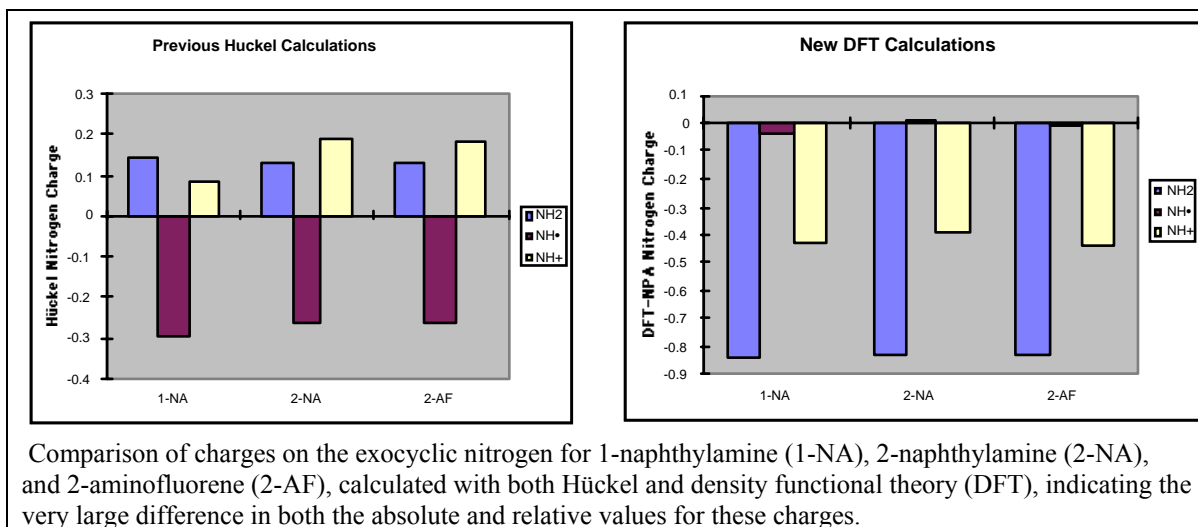
1-naphthylamine (1-NA), 2-naphthylamine (2-NA), and 2-aminofluorene (2-AF). The size of the arrow indicates the relative degree of P-450 catalyzed ring carbon or nitrogen oxidation.

By examining atomic electron charge distributions of the parent compounds and proposed intermediates predicted using Hückel theory, they found that the ratio of atomic charges on the nitrogen and ring carbons in the 1-NA, 2-NA, and 2-AF nitrenium intermediates were correlated with the observed nitrogen versus ring oxidation rates. Thus, they concluded that nitrenium cation intermediates determined product ratios and, therefore, arylamines and related compounds were N-oxidized via a two-electron transfer mechanism.

The validity of these earlier results depends on the accuracy of the chemical simulations employed. Hückel theory has subsequently been shown to be a poor predictor of atomic charges. Hückel theory only accounts for  $\pi$ -electrons and subsequent research using more accurate quantum chemical methods have demonstrated inclusion of the  $\sigma$ -electrons is necessary to achieve qualitatively accurate charge distributions. To overcome this limitation, we have recomputed the atomic charges of the 1-NA, 2-NA, and 2-AF reaction intermediates using much more accurate *ab initio* quantum chemistry methods (specifically Density Functional Theory with Natural Population Analysis (DFT-NPA)) that are well established to yield accurate electronic properties. In the bar graphs below, we have plotted the exocyclic nitrogen charges of the reaction intermediates.

There are significant qualitative differences between the Hückel theory and DFT-NPA charges that indicate the need for high quality quantum chemistry predictions. We find that the atomic charge distribution of the nitrenium cations does not correlate with the observed oxidation rates. Rather, the best match is found with the charge

distributions of the aminyl radicals ( $\text{NH}\bullet$ ), implying that N-oxidation occurs through a one-electron transfer mechanism rather than the earlier proposed two-electron transfer mechanism (b) instead of (c) in the mechanism figure at the top of this section.



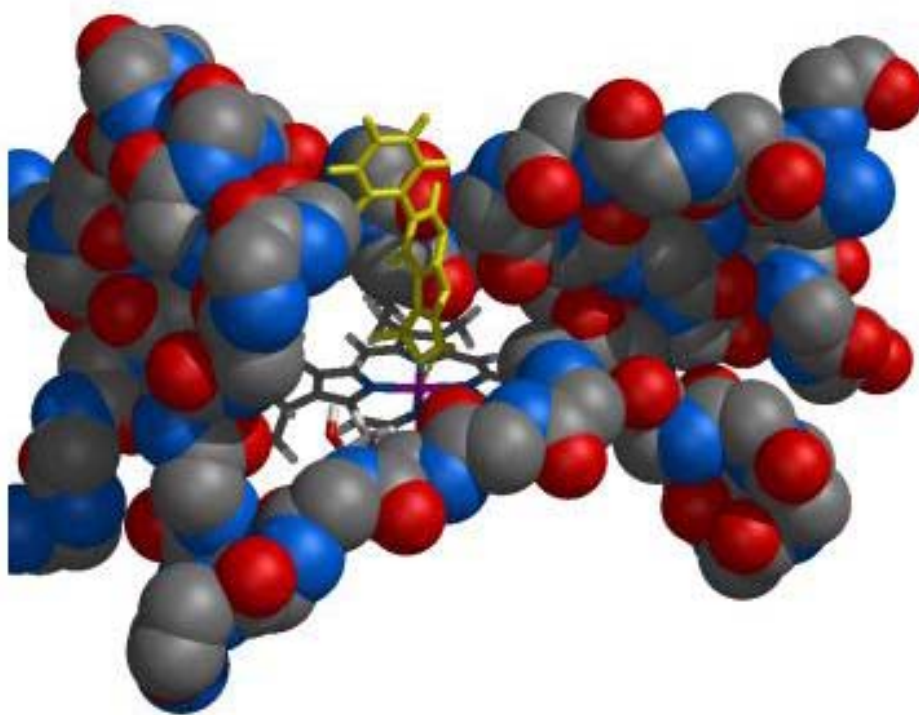
In this project, we have demonstrated that *ab initio* quantum chemical calculations provide more accurate chemical parameters that can be used with greater success in elucidating the chemical mechanisms of biological processes. The results of this study have been submitted to the journal *Mutation Research*.

#### 2.1.1.4. Development of an active site model for CYP1A2

The CYP1A2 isoform of cytochrome P450 is the enzyme that catalyzes the initial oxidation of the AIA food mutagens which ultimately leads to a very reactive nitrogen ester form of the food mutagens. These electrophilic species can interact with DNA and interfere with accurate duplication of the genetic message leading to mutagenesis and carcinogenesis. Earlier work has suggested the initial oxidation reaction modulates the mutagenic potency of these compounds. Thus, in order to understand the parameters that control AIA compound activation and mutagenic activity, it would be useful to directly study the interaction of these mutagens and cytochrome P450. To date, no crystal structure is available for CYP1A2 or any of the other mammalian cytochrome P450 enzymes, so that computer modeling of the CYP1A2 active site was necessary to provide an active site structure for subsequent computational modeling.



There is fairly low overall sequence homology between the human CYP1A2 enzyme and these bacterial enzymes. Nevertheless, by using multiple sequence alignment and homology-based protein structure prediction methods, we have been able to develop several models of the CYP1A2 protein based on available X-ray crystal structures of bacterial P450 enzymes. One of these modeled structures is shown in the figure below, with the food mutagen PhIP docked into the active site. This protein modeling constituted part of the preliminary data for a successful NIH grant that will use these protein models to explain the oxidation activity of known AIA food mutagens. In addition, the model will be used to determine the mechanism of action of chemopreventative agents such as the dietary flavonoid compounds. (See following section).

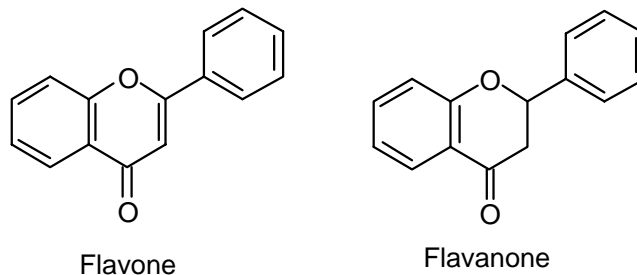


A picture of the P450 active site (shown as spheres) modeled using homology-based protein structure prediction methods with the food mutagen PhIP docked into the active site.

#### **2.1.1.5. Bioflavonoid Inhibitors of P450 enzyme**

In addition to our studies of the AIA food mutagens, we have studied a class of compounds, the bioflavonoids, that are known epidemiologically to reduce cancer risk and cardiovascular disease. The mechanism of the cancer-preventative effect is not known, but

there is evidence that these compounds may inhibit the initial metabolic step of the food mutagens, oxidation by cytochrome P450 1A2. We developed a quantum-chemical Quantitative Structure-Activity Relationship (QSAR) by comparing the predicted properties for thirty bioflavonoid compounds with the experimentally measured inhibition of mutagen oxidation by P450. The two general flavanoids frameworks studied are shown in this scheme, differing only in the placement of a double bond.

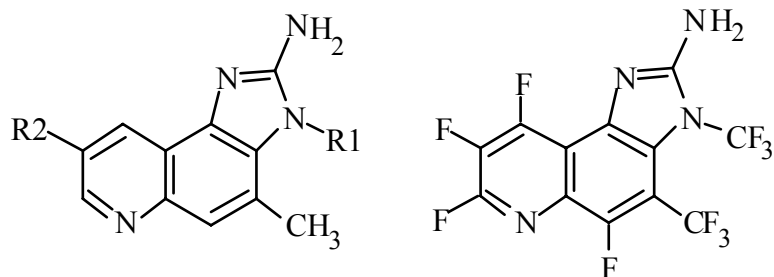


Interestingly, we find a weak correlation between the energy to make the two rings co-planar and the measured inhibitory potential. This is consistent with earlier hypotheses that these compounds must become planar in order to fit into the P450 enzyme active site. We have published a paper describing these results in *Environmental and Molecular Mutagenesis*.

#### 2.1.1.6. Predictions of high potency mutagens

In our earlier QSAR studies of the AIA food mutagens we found that high mutagenic potency is predicted by a low-energy lowest unoccupied molecular orbital (LUMO). However, since no clear mechanism has been established for this relationship, it is not clear if the LUMO energy directly influences potency or if it is simply co-linear with a true, underlying correlation. If the LUMO energy directly influences the mutagenic potency, it should be possible to use predicted LUMO energies to extrapolate the mutagenic potencies for novel AIA derivatives.

Using the reasoning that electron withdrawing groups should lead to lower LUMO energies, we have computationally designed a set of novel analogs of the AIA mutagens, shown in the figure/table below.



R1	R2	E(LUMO)	Pred Log MP
CH3–	H–	2.871	3.31
NH2–	H–	2.913	3.15
NO2–	H–	1.983	6.69
CF3–	H–	2.556	4.51
H–	HCO–	1.972	6.74
H–	NO2–	1.441	8.76
CF3–	HCO–	1.727	7.67
Perfluoro-34miq		1.421	8.84

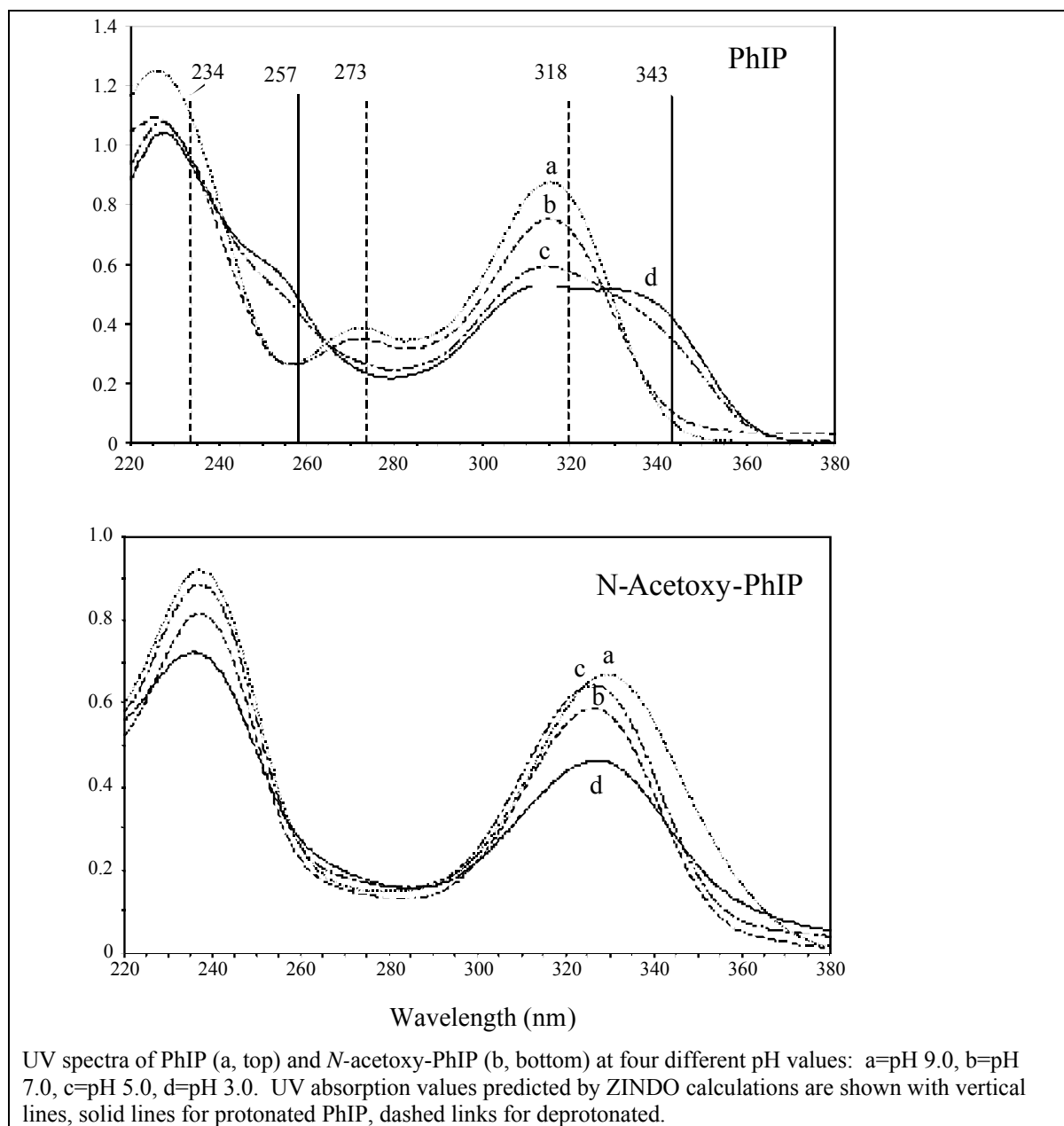
Log MP values for novel AIA analogs extrapolated from LUMO energies (e.v.) using QSAR linear regression fit of Log MP vs. E(LUMO) for 16 AIA compounds.

For each of these compounds we have calculated the optimized structure and LUMO energies using *ab initio* quantum chemical methods. We have then used the linear relation between LUMO energy and log (mutagenic potency) from our published QSAR involving 16 AIA mutagens to extrapolate the potencies for these novel compounds. Several of the proposed compounds are indeed found to have LUMO energies lower than any observed AIA mutagens and have extrapolated potencies between ten- and a thousand-fold greater than the most potent known AIA mutagen, and indeed, would be the most mutagenic compounds ever synthesized. As part of a newly funded NIH grant, we plan to have these compounds synthesized and tested for mutagenic activity. If these exhibit the predicted increased mutagenic potency, this will provide a strong additional validation that the LUMO energies are mechanistically linked to mutagenic activity. In addition, (as pointed out by a member of the Biology and Biotechnology Research Programs External Advisory committee) such highly potent compounds would be extremely valuable test compounds for the experimental studies of the mechanisms of mutagenesis.

#### 2.1.1.7. Computational spectroscopy of food mutagens

We have used the semi-empirical method, ZINDO, combined with higher level quantum chemical structure optimizations to predict the UV absorption spectra of 2-amino-1-methyl-6-phenylimidazo(4,5-*b*)pyridine(PhIP), and its N-acetoxyl metabolite (N-Ac-PhIP) under both high and low pH conditions. In a study of somewhat similar compounds, ZINDO was recently used to calculate the absorption spectra of diazapyrenes and their complexes with adenine, and was found to accurately account for the experimental findings. The next figure shows the experimental UV spectra of both compounds at four pH's (a =pH 9.0, b = pH 7.0, c= pH 5.0 and d= pH 3.0). The PhIP spectrum has superimposed lines indicating the predicted strongest bands for the protonated PhIP (solid lines) and the unprotonated PhIP (dashed lines). The experimentally determined pK<sub>a</sub> for PhIP is 5.7, and corresponds to protonation at the imidazole N3 position. Hence, spectra a and b should correspond to the unprotonated PhIP and spectra c and d to the protonated form. As can also be seen in the figure, the Zindo calculations accurately predict both the number and position of the strong UV absorption bands for both protonation states of PhIP.

Unlike the spectra for PhIP the spectra for N-Ac-PhIP (see figure) did not exhibit a change in shape over the 3.0 - 9.0 pH range tested, suggesting that there was no change in protonation state for N-Ac-PhIP. These results led us to hypothesize that the pK<sub>a</sub> for N-Ac-PhIP was higher than 9.0 and prompted us to perform calculations to obtain theoretical pK<sub>a</sub> values for this compound.



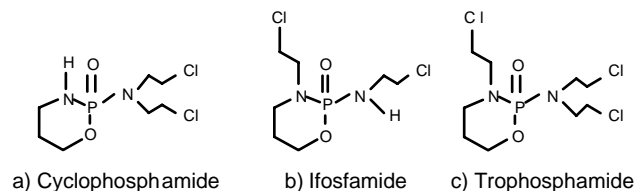
*Ab initio* calculations of protonated versus nonprotonated forms of PhIP and two of its metabolites were used to derive theoretical pKa predictions for *N*-hydroxy-PhIP and *N*-acetoxy-PhIP, using a procedure that we have shown to yield reasonably accurate pKa values. Using the experimentally derived pKa value of 5.7 for PhIP, the following pKa values were derived for the two PhIP metabolites:

	calculated pKa
PhIP	5.7 (expt. reference value)
NHOH-PhIP	5.34
Acetoxy-PhIP	11.46

The unusually high pKa value for *N*-acetoxy-PhIP correlates with the UV spectra for this compound in different pH environments. For PhIP and *N*-hydroxy-PhIP a shift in the spectral shape takes place between pH 5 and 7. This change coincides with the pKa values of these two compounds, 5.7 and 5.3, respectively for the deprotonation of the N3 nitrogen. In contrast, the spectrum of *N*-acetoxy-PhIP shows no analogous change over the 3.0 to 9.0 pH buffer range. This result is consistent with this compound's extremely high predicted pKa value of 11.46. This observation may be explained by formation of a stable hydrogen bonded pseudo 7-membered ring structure between the protonated N3 and to acetoxy oxygen in the *N*-acetoxy-PhIP species, making deprotonation unfavorable. This behavior is similar to that seen for other organic acids that form intramolecular hydrogen bonds. In contrast, protonated forms of PhIP and *N*-hydroxy-PhIP cannot form analogous stable ring structures, yielding lower pKa values. These results demonstrate that predicted spectroscopic data and other chemical properties such as acid constants, can provide insight that allow the explanation of otherwise paradoxical experimental results.

### 2.1.2. DNA-Alkylating Anticancer Drugs

The phosphoramidic mustard anticancer drugs, including cyclophosphamide and ifosfamide (shown below), have been in clinical use for more than forty years and remain the treatment of choice for several forms of cancer.

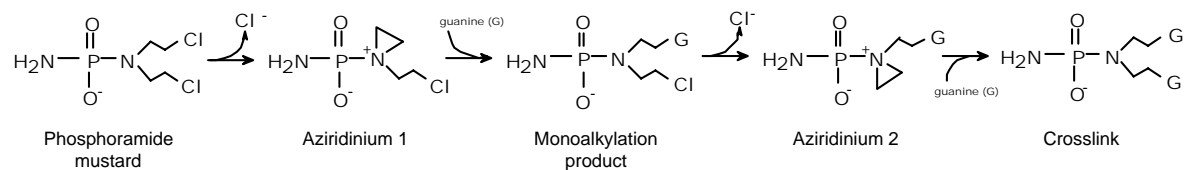


Three phosphoramidic mustard-based anticancer prodrugs. These compounds require metabolic activation to a reactive species that forms DNA crosslinks between guanine bases on opposite DNA strands.

Extensive research on these drugs has revealed many aspects of their pharmacology; however, several important questions about their biological activity remain unanswered and therefore are the topic of much ongoing study. In this project we have performed computational simulations on many different aspects of these important anticancer drugs with the goal of understanding which of their chemical properties are key to their anticancer activity and to use this information to design new chemical analogs of these drugs with improved therapeutic utility.

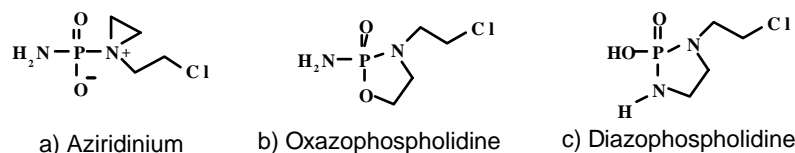
#### 2.1.2.1. Elucidation of alternative cyclization products from phosphoramidate mustard

The phosphoramidate mustard anticancer drugs are known to act by undergoing a series of metabolic transformations leading to the active metabolite phosphoroamide mustard (PM) (see scheme below). The phosphoramidate mustards then undergo an intramolecular reaction to form an aziridinium ion that reacts with the DNA base guanine.



Chemical reactions leading to therapeutically active DNA crosslink as shown for phosphoramidate mustard, the active metabolite of cyclophosphamide. First the phosphoramidate mustard undergoes intramolecular cyclization to form a highly electrophilic aziridinium ion, which subsequently reacts at the guanine N7 position (or other cellular nucleophile). This process is repeated for the second alkylation reaction. Note that P-N bond hydrolysis could occur at any point in this process, but has been shown to preferentially occur for the aziridinylium intermediates.

A potentially important factor in the activity of the phosphoramidate mustards is their ability to undergo other alternative intramolecular reactions to form five membered rings, instead of the therapeutically active aziridinium, as shown in the following scheme.



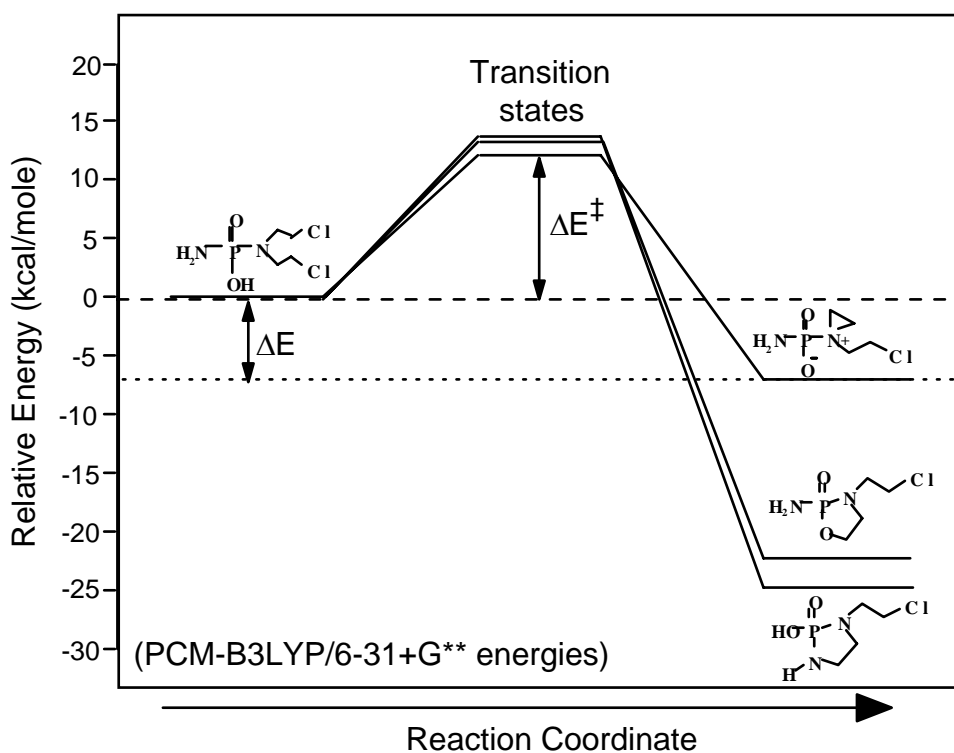
Possible cyclic products arising from intramolecular reactions of phosphoramidate mustard. Structure (a) is the known precursor of DNA damage by the anticancer drug. Structures (b) and (c) are predicted to be lower energy cyclization products.

Such five membered ring products have been reported to arise from the activated analogs to the phosphoramidic mustards and are a slow degradation product of the unactivated cyclophosphamide. Considerations of ring strain would suggest that these five membered rings would be much more stable than the corresponding aziridinium ion. Quantum chemical calculations of the three cyclization products of phosphoramidate mustard bear out this reasoning, showing that the five membered ring forms are more than 20 kcal/mole favored over the aziridinium form. The chemical factors governing the rates and product distributions of such cyclization reactions have been the focus of extensive research, but explanations of the relative rates of intramolecular reactions to form 3- and 5-member rings have proved elusive. The ratio of these cyclization rates,  $k_3/k_5$ , varies over a wide range in studies of model compounds, from greater than 20000 to 0.001. For the model compounds that might be expected to be similar to the phosphoramidic mustards ( $\omega$ -halogenoalkylamines) the 5-member rings are strongly favored.

We have compared the predicted energies and measured hydrolysis rates of a number of cyclic phosphoramidate compounds and found no clear connection between the thermodynamic stability of the compounds and their lifetimes. We have made further progress on this question by predicting both the overall reaction energies ( $\Delta E$ ) and the transition state energies ( $\Delta E^\ddagger$ ). As can be seen in the next figure, the two 5-member phospholidines are thermodynamically considerably more stable than the 3-member aziridinium as indicated by their lower reaction energies. However the reaction barriers to form these three compounds are very similar, with a slightly lower barrier for the aziridinium, suggesting that all three should be produced in nearly equal concentrations. On the basis of these simulations, our collaborators at Duke University have looked for the



formation of 5-membered ring products directly from PM. In a partially polar solvent (DMSO) the 5-member ring species oxazophospholidine (shown in the previous figure) is actually the predominant cyclization product. This result is a significant finding since the formation of such 5-member rings constitutes a non-therapeutic metabolic pathway for these anticancer drugs, and hence, chemical alterations to these compounds that could block this potential pathway would increase their anticancer effectiveness.

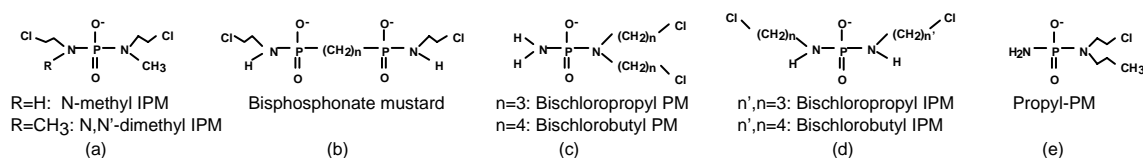


Predicted reaction path for cyclization of phosphoramidate mustard (shown at left) to form 3- or 5-membered ring products. The 3-membered ring is considerably less thermodynamically favored than either of the 5-membered ring products. However, the predicted reaction barriers to form each of the cyclic products is very similar, indicating that equal quantities of these products will be formed at physiological temperatures.

#### 2.1.2.2. Predictions of activation energies for novel chemical analogs

In the therapeutic pathway for these drugs, it is well established that the formation of the aziridinium is the rate limiting step for nearly all of the anticancer drugs in this class. For this reason, any chemical changes that modulate the rate of aziridinium formation should affect the activity of these drugs. We have performed quantum chemical simulations on a series of chemical analogs of phosphoramidate mustard. In particular, we have calculated the reaction energy and transition state barrier to undergo the initial cyclization reaction. We have

performed these preliminary calculations on PM, propyl-PM, the bischloropropyl and bischlorobutyl derivatives of PM, the monochloropropyl and monochlorobutyl derivatives of isophosphoramidate mustard (IPM), and a bisphosphonate mustard that links two halves of an IPM with an alkyl chain (shown below).



Set of phosphoramidic mustard analogs for which quantum chemical simulations have been applied to determine whether they are promising for subsequent synthesis and testing.

Note that the chloropropyl- and chlorobutyl-analogs of PM and IPM can cyclize to form four and five membered rings, respectively. The predicted cyclization energies for the IPM analogs are given in the table below.

The reactions to form the cyclic product for the seven compounds studied are all found to be thermodynamically favored ( $\Delta E_{\text{cyclization}} < 0$ ) and have relative energies consistent with chemical intuition about the effect of ring strain in the cyclic product. The first three mustard compounds in the table cyclize to form three-membered aziridinium rings and show a range of cyclization energies from -6.7 to -11.5 kcal/mole. The cyclization energies for the PM analogs that form 4- and 5-membered ring products are found to be more favored, with  $\Delta E_{\text{cyclization}} = -12.5$  to  $-13.6$  and  $-28.7$  to  $-31.0$  kcal/mole, respectively. These relative energies closely follow the experimentally determined ring strains of cycloalkanes ( $E_{\text{strain}} = 27.6, 26.2,$  and  $6.5$  kcal/mole, for cyclopropane, cyclobutane, and cyclopentane, respectively).

Table: Predicted reaction energies and reaction barriers for the cyclization reactions of PM and several analogs. All energies are given in units of kcal/mole and have been calculated from quantum chemical density functional calculations using the B3LYP gradient-corrected functional and the COSMO solvent model.

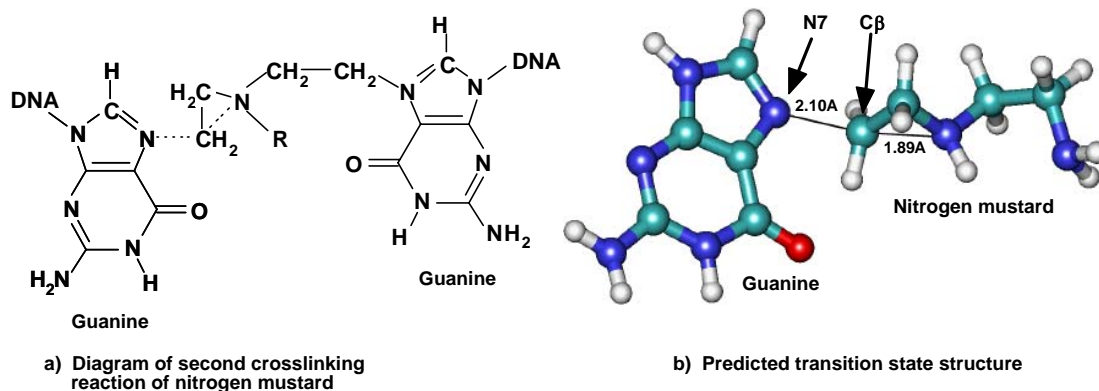
Compound (descriptive name)	$\Delta E_{\text{cyclization}}$	$\Delta E_{\text{cyclization}}^{\ddagger}$
Phosphoramidate mustard(PM)	-6.7	11.2
N-Propyl-N-chloroethyl-phosphoramidate mustard(propyl-PM)	-9.9	10.2
Bischloropropyl-PM (Reaction to form 4-membered ring)	-12.5	14.2
Bischlorobutyl-PM (Reaction to form 5-membered ring)	-28.7	7.9
Chloropropyl-chloroethyl-IPM (Reaction to form 4-membered ring)	-13.6	13.1

Chlorobutyl-chloroethyl-IPM (Reaction to form 5-membered ring)	-31.0	8.4
Bisphosphonate mustard	-7.0	10.8

The activation energies listed in the forth column ( $\Delta E_{\text{cyclization}}^{\ddagger}$ ) are low enough to allow moderate-to-rapid cyclization at room temperature. Hence, these results indicate that all of the proposed compounds should be viable alkylating agents. Using the Arrhenius relationship ( $\text{Rate} \sim e^{-\Delta E^{\ddagger}/kT}$ ) the relative rates of cyclization can be estimated from these activation energies (assuming that the exponential prefactor is the same for all reactions). The PM analogs that can cyclize to form the five membered rings are predicted to be much more reactive than PM, cyclizing about 100 times faster (for PM  $t_{1/2}$  =18 minutes). The analogs that cyclize to form four membered rings are less reactive than PM, cyclizing about 25 time slower. Note that this result is in good agreement with the preliminary experimental results for Bischlorobutyl-PM described in the next section. Hence the proposed analogs bracket the cyclization rate of PM and should therefore experimentally testing these proposed compounds will provide useful data on how varying this activation rate will affect the alkylating potency and crosslinking ability of these compounds. These results also indicate that the non-crosslinking propyl-PM analog should have a half-life very close to that of phosphoramidate mustard. Finally, the bisphosphonate mustard is predicted to have a reaction energy and activation energy very similar to that for PM, indicating that this should be a useful analog for testing the effect of varying linker length on crosslinking efficiency.

#### 2.1.2.3: Quantum Chemical simulations of the DNA crosslinking reaction

As described above, the therapeutic activity of these drugs depend on their ability to ultimately form crosslinks between the two strands of DNA. It has been previously hypothesized that the second alkylation reaction to form an interstrand crosslink would require highly distorting the DNA from its canonical B-form. The degree of distortion required for a specific alkylating agent is likely to provide an estimate of its crosslinking efficiency. In order to accurately quantify the structural constraints of the second alkylation reaction we have completed both quantum chemical studies of the structural properties of the second DNA alkylation reaction of nitrogen mustards leading to the interstrand crosslink (see Figure below).

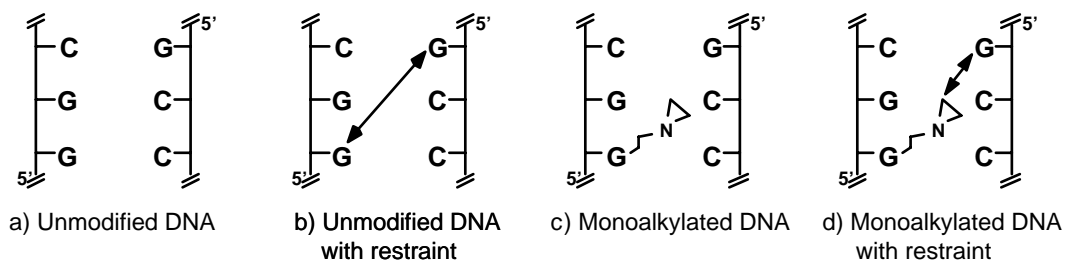


a) Chemical diagram of the interstrand crosslinking reaction. At the point shown, the nitrogen mustard group has already formed a monoalkylation product with the guanine at right, and the second aziridinium ring has already been formed and is in the process of ring opening after nucleophilic attack by the second guanine (at left). b) Transition state structure for guanosine reaction with aziridinium calculated using first principles quantum chemical methods. The bonds being broken and formed are shown with thin black lines; the distances are shown in Angstroms.

We performed this calculation using density functional theory using nor-nitrogen mustard ( $R=H$ ) as the crosslinking agent and replacing the right-hand guanine with a simple amine. The resulting optimized transition state structure is shown in the figure above. As chemical intuition would suggest, this transition state involves a near collinear orientation of the guanine N7 with the C $\beta$  and nitrogen in the aziridinium ring (N7-C $\beta$ -N angle =  $157^\circ$ ). The transition state bond distances are 2.10Å for the N7-C $\beta$  bond being formed and the 1.89Å for the C $\beta$ -N bond being broken. Based on this transition state structure we find the N7-N7 distance to be 6.63Å, which, as expected, is shorter than the 7.4Å length of the fully extended ethyl-amine-ethyl crosslinking group. In the following section we describe how we used the structural constraints from this predicted transition state structure in simulations of DNA undergoing the second crosslink reaction.

#### 2.1.2.4. Molecular dynamics simulations of crosslinking reactions

In order to determine the structural deformation required in DNA to permit the formation of the interstrand 5'GNC crosslink, we performed molecular dynamics simulations on a 12 base pair B-DNA double helix under four different conditions, shown in the following figure.



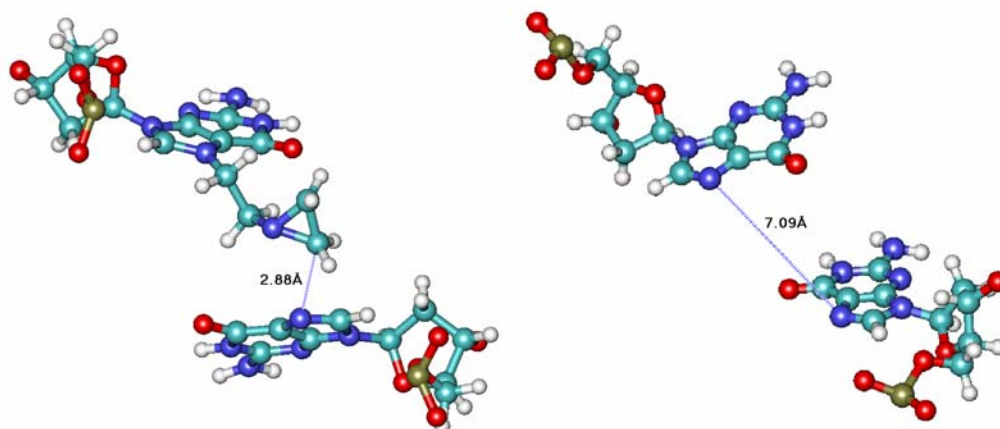
Preliminary molecular dynamics simulations performed on unmodified DNA (a and b) and monoalkylated DNA (c and d), with and without distance restraints related to the configuration necessary for the second alkylation reaction to form the interstrand crosslink.

To be consistent with our earlier simulations of crosslinked DNA, we chose the DNA sequence: 5'-d(CATGTAGGCTAA) with the monoalkylated guanine shown underlined. For these preliminary simulations we used an ethyl-amino-aziridinyl monoalkylation product.

For all four simulations outlined in the figure above, we find that the DNA double helix structure is well preserved during ~100 ps of dynamics simulations. Although there is no simple measure of the integrity of the double helix structure, one useful measure is the preservation of the proper interstrand hydrogen bonds. From monitoring these bonds (defined as having donor-acceptor distances  $< 3.4 \text{ \AA}$  and angles  $< 60^\circ$ ) we find in all four simulations nearly complete preservation of the 29 possible Watson-Crick hydrogen bonds.

The figure below shows snapshots taken at 75 ps. from the molecular dynamics simulations of the DNA oligomers including harmonic distance restraints related to the second alkylation reaction (restraints shown with dotted lines). Note that only the two guanine nucleotides and the mustard crosslink are shown. The picture on the left shows the monoalkylated guanine with a restrained N7-C $\beta$  distance; on the right, the unmodified guanines with a restrained N7-N7 distance.

An interesting result is the very short distance restraint that can be accommodated by rotation around the glycosidic bond and in the phosphodiester backbone without long-range major disruption of the DNA helical structure. Even with a fairly modest applied force, the N7-N7 and N7-C $\beta$  distances closely approach that needed for the final, completed crosslink. However, our results do indicate that a larger degree of DNA distortion is required to achieve the conformation necessary for the second alkylation reaction to take place.



Snapshots taken at 75 ps. from the molecular dynamics simulations of the DNA oligomers including harmonic distance restraints related to the second alkylation reaction (restraints shown with dotted lines). Note that only the two guanine nucleotides and the mustard crosslink are shown. On the left is shown the monoalkylated guanine with a restrained N7-C $\beta$  distance; on the right, the unmodified guanines with a restrained N7-N7 distance.

As described in the previous section we find that the transition state for the second, crosslink-forming alkylation reaction requires a guanine-N7 to aziridinium-C $\beta$  distance of  $\sim 2.1$  Å. To evaluate the degree of distortion necessary to achieve this distance, we ran a series of molecular dynamics simulations of the monoalkylated DNA oligomer using N7-C $\beta$  harmonic distance restraints with increasing force constants. Even for the highest imposed force ( $k=30$  kcal/mole/Å<sup>2</sup>), we find that the average distance is  $\sim 0.7$  Å longer than necessary to allow the second alkylation reaction. Interestingly, for each imposed force constant the N7-C $\beta$  distance rapidly reaches an asymptotic value with very little net drift.

Based on these dynamics results we can draw several conclusions. First, for the mustards with an ethyl-amine-ethyl crosslink (such as PM), significant local distortion of the DNA will be necessary to allow the second alkylation reaction to form the interstrand crosslink. Since these crosslinks are experimentally observed to occur, we plan to extend our dynamics simulations to longer timescales to determine whether the necessary amount of distortion is induced by the monoalkylation as proposed previously. Second, only a relatively small increase in the length of the crosslinking chain,  $\sim 0.7$  Å based on the constrained dynamics described above, would much more easily allow the second alkylation reaction. These results provide key insights into the structural factors of the phosphoramidate mustard anticancer drugs that should be modulated to increase the degree of crosslinking, and therefore the anticancer effectiveness of these drugs.

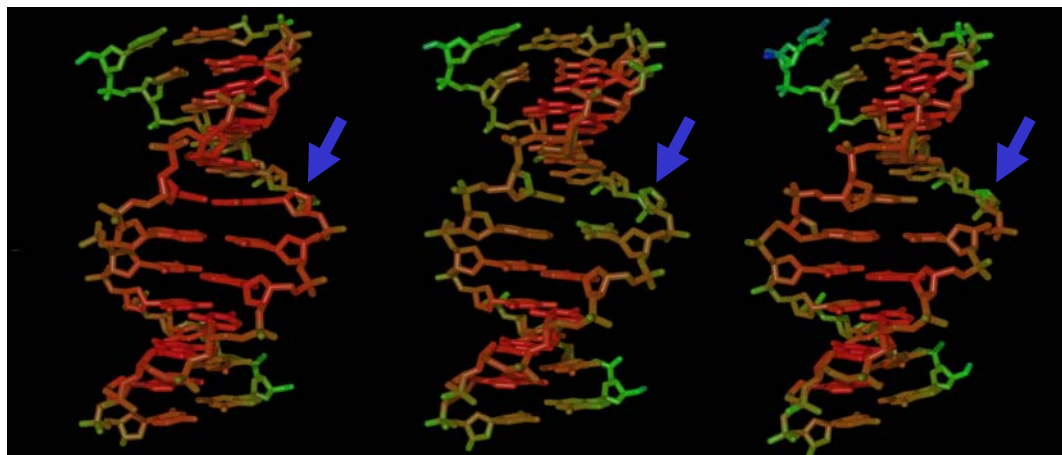
### 2.1.3. Abasic DNA Repair

The study of cellular repair of both spontaneous and induced DNA damage has been the focus of several research groups in the BBRP. We have been collaboratively pursuing this problem “from both ends,” by separately studying the detailed dynamics of an important DNA repair protein—apurinic/apyrimidinic endonuclease (APE1) and the damage, “abasic” DNA, that it repairs. By performing several large-scale simulations of APE1 mutants, we have worked to define the specific intra-protein and protein-DNA interactions of amino acids which have been shown biochemically to play a critical role in recognizing and binding damaged DNA. In work published in the *Journal of Molecular Biology* (Nguyen et al., 2000), we reported on combined experimental biochemical results from Dr. David Wilson's laboratory (BBRP) and results from our simulations of wild-type APE1 to elucidate the metal binding activity of Ape1 and correctly predict specific amino acids that interact with DNA.

We have also completed a molecular dynamics study of abasic DNA, published in *Nucleic Acids Research* (Barsky et al., 2000). Early in 2000, the crystal structure of the APE1-abasic DNA complex was reported in *Nature*, and some important aspects of DNA structure that we predicted were observable therein, as well as our predictions concerning DNA binding amino acids in APE1. In analyses of the simulated dynamics of abasic damaged DNA, the target of the APE1 enzyme, we had found that the abasic site does not form a rigid hole or gap in the DNA, but instead perturbs the canonical structure and induces additional flexibility close to the abasic site. Our results indicated that the sugar can spontaneously flip out into the minor groove, a conformation in which the DNA was observed to bind to APE1.

In the next figure, the average positions of all atoms, from simulations of intact, apyrimidinic, and apurinic simulations are shown, colored by the relative degree of movement, with the abasic sugar denoted by a blue arrow. While abasic sites always destabilize DNA duplexes, the magnitude of this destabilization can vary over a wide range that depends on the sequence context of the abasic site. The unusual base pairing which we reported for one sequence context, was not observed for an alternative sequence. We have observed sequence-dependent degrees of motions within the DNA helix due to the abasic site, and we have hypothesized that this correlates with the sequence dependent helical

stability, an experimentally verifiable prediction with potentially far-reaching applications in the study of molecular stability.



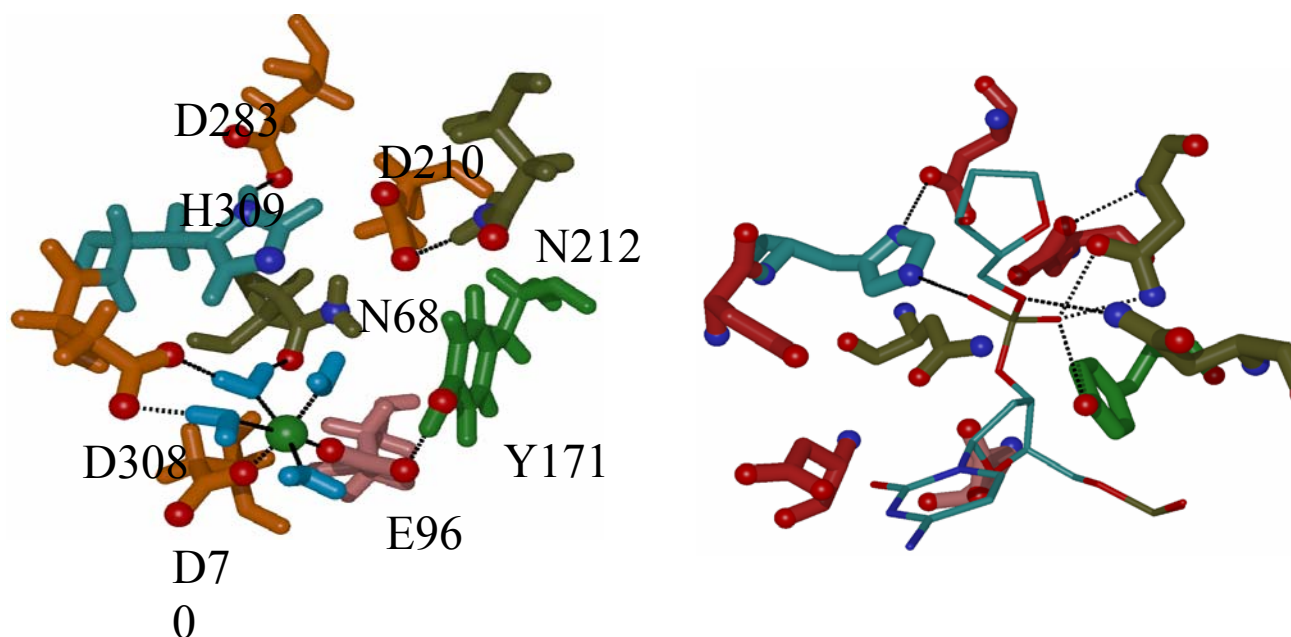
Identical views of average DNA structures from MD simulations where there is no abasic site (left), an abasic site opposite an orphan C base (middle), and an abasic site opposite an orphan G base (right) in otherwise identical dodecamer duplexes. Shown are all non-hydrogen atoms (in capped-sticks), colored by the degree of internal movement such that blue indicates the largest movement (5.0 Å) and red the smallest (0.7 Å). A blue arrow points to the 6th sugar, which is abasic in the middle and right models. Notice that atoms near the abasic site show more movement than the same atoms in the control (left), and the “melted” nature of the region surrounding the C-orphan abasic site is also visible (middle). The average structures were created by performing a root-mean-square (RMS) fit over all atoms for 0.5 ps sampling within a 1.5 ns MD trajectory. The isotropic RMS fluctuations were used to color each atom. The gap created by the missing base was filled mainly by water, and also with neighboring bases when the sequence context makes pairing with the orphan base favorable (Barsky et al., 2000).

We also observed 20 to 30 degree kinks in the abasic DNA, and a compression of the DNA backbone across from the abasic site, which again has been observed in the co-complex. The abasic sugar also underwent an unusual range of sugar puckers which are not normally observed in DNA, and which contradict a long-held belief that the sugars themselves determine the form (e.g., A or B) of the DNA. Finally, we studied the nature of the gap created by the abasic site, showing that it is sometimes water-filled, but that its steric and electrostatic nature does not allow complete filling and may be a source of the duplex destabilization caused by an abasic site.



These observations have not yet been verified experimentally due to the difficulty of observing unbound water by either X-ray crystallography or NMR methods. Beyond the determining the structure of abasic DNA, we have wanted to understand which features make abasic DNA recognizable and amenable to DNA repair. To this end we simulated several new abasic DNA analogues where the abasic site has been replaced by an ethyl (E), or propyl (P), or a lysine-like linkage (Q). The E and P linkages have the interesting property of being recognized and cleaved by APE1, even in the absence of  $Mg^{2+}$ , which APE1 requires for its cleavage activity against regular abasic sites. The Q linkage is neither bound nor recognized by APE1, and results from the simulations have suggested that the linkage, having a charged moiety, tethers the backbone in a way that makes the activated conformation of the DNA energetically unfavorable.

We modeled the protein DNA interactions (below, right) and also performed a 1 ns simulation of the entire APE1 protein, together with  $Mg^{2+}$  in the active site, in order to understand the active site (below, left). These simulations have identified residues involved in  $Mg^{2+}$  coordination, and, secondly, residues on the surface of the protein play a significant role in the protein's binding to DNA, an observation recently verified with the availability of the protein-DNA co-complex.



As mentioned above these results for the DNA and protein simulations were published in separate papers in 2000, and last year we published a compilation and review of this work in *Mutation Research* (Wilson and Barsky, 2001).

#### **2.1.4. Non-polar DNA Base Analogs**

We applied simulations to determine the energetic factors leading to the sequence specific incorporation of DNA nucleotide bases and chemically related analogs. The latter constitute the majority of antiviral compounds, including the HIV drug AZT, and are being studied for treating other diseases including cancer. A complete explanation of the base-pairing requires simulations that accurately include the effects of solvent, local DNA structure, and the structural features of the DNA polymerase active site. In 1997, Kool and coworkers raised important questions about the role of hydrogen-bonding in DNA base pairing by using nonpolar isosteres difluorotoluene (F) and benzimidazole (Z), for thymine (T) and adenine (A), respectively, to demonstrate that pol I (KF exo-) incorporates A and Z specifically and efficiently opposite both F and T, and similarly incorporates F and T opposite A and Z. During this strategic initiative, we have studied these issues using quantum chemical calculations, long time-scale classical MD simulations and the first principles MD methods described in the next section.

Our initial step was to calculate the interaction energy between the natural DNA bases and their non-polar analogs. Several early reports showed that the non-polar isostere of thymine (T), difluorotoluene (F), can partially hydrogen bond with A. By ab initio quantum chemical methods—gas phase (HF/6-31G\* and MP2/6-31G\*\*) and solvent phase (SCI-PCM/HF/6-31G\*)—we showed that the A-F interaction is three times weaker than the A-T interaction in the gas phase, but further show that the A-F interaction is one quarter that of A-T in the solvent phase. Based on the electrostatic solvation energies, we found the non-polar isosteres to be five times less hydrophilic than the natural bases. Of the new base-pairs (F-Z, T-Z, and F-A), only F-A formed an A-T-like arrangement in unconstrained optimizations, yet 2 Å more separated. F-Z and T-Z do not form planar arrangements, and constrained optimizations show that large amounts of energy are

required to make these pairs fit the exact A-T (and G-C) shape. We have published these findings in *J. Biomol. Structure and Dynamics*, 1999.

Although ab initio quantum chemistry using atom-centered basis functions is now widely employed to study molecular interactions, there persists an inherent inaccuracy in the method, known as basis-set superposition error (BSSE). Depending on the size of the basis sets and the level of quantum chemical theory employed, this error can lead to uncertainties as large as the computed interaction energies themselves. A standard way of estimating the BSSE is to compute a Boys-Bernardi counter-poise correction (CPC) with the basis-sets employed in the original calculation, yet the inaccuracy of the estimate itself is not known, and, therefore, the CPC does not provide a rigorous upper bound on the error. Although we have, within this Strategic Initiative, tested the convergence of the CPC for molecular systems small enough to employ huge basis-sets, which render the BSSE arbitrarily small, for larger systems, such as two DNA bases, this has been unfeasible without vast computer resources.

Inherently devoid of BSSE are quantum chemical methods that employ a space-filling plane wave (PW) basis rather than atom-centered basis. Although the PW basis consumes greater computer resources than small atom-centered basis sets (e.g. 6-31G\*\*), a parallelized PW density functional theory program was developed under this strategic initiative. In continuing our study of DNA base-pairing interactions, we used this software to carry out PW calculations for the natural T-A and G-C base-pairs as well as the F-A base-pair analogue (F is difluorotoluene, a DNA base isostere). Our results, published in *Chemical Physics Letters*, 1999, have indicated that the CPC appears to be an underestimate of the BSSE, and in fact the standard methods including CPC do not greatly differ from the PW results.

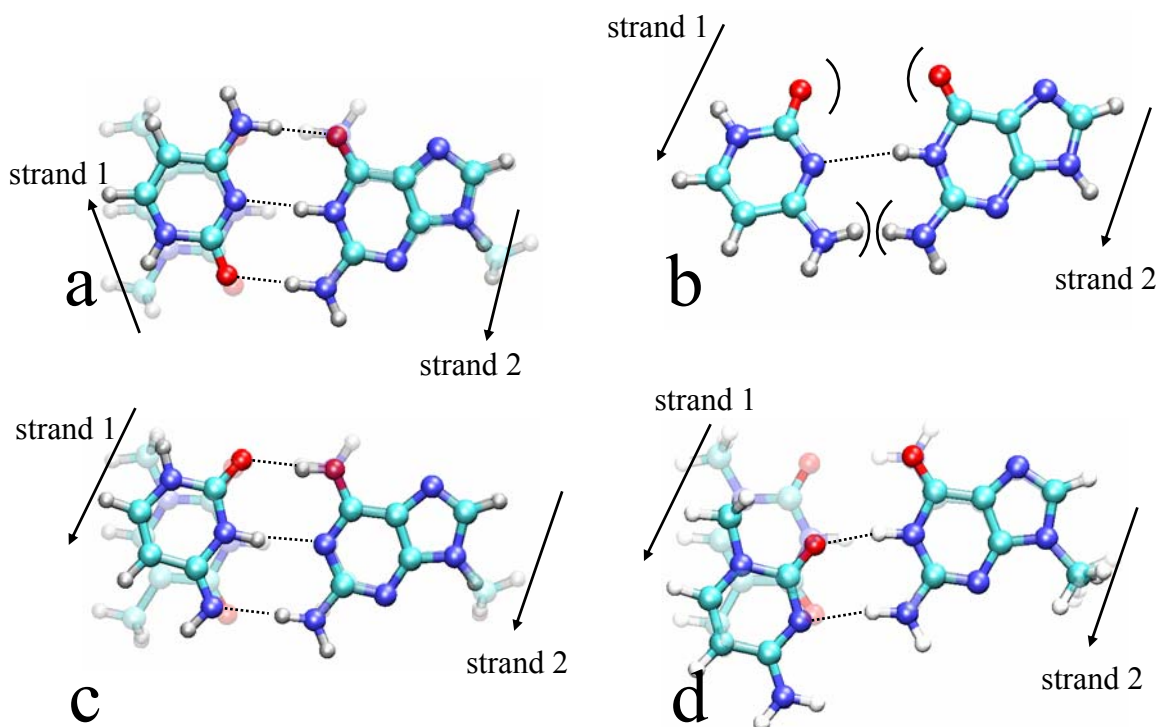
DNA stability and fidelity in DNA replication are essential to ensuring normal functioning and maintenance of an organism, and the energetic factors lending stability to double-helical DNA and leading to selective incorporation of DNA bases in DNA replication are related. Using the nonpolar DNA isostere difluorotoluene (F), in place of thymine (T), we and others performed multi-nanosecond simulations of double helices containing an F-A base pair, and showed that F indeed behaves very much like T, even in

the absence of hydrogen bonds, although it appeared more likely to flip out of the helix.

### 2.1.5. Parallel Stranded DNA

Virtually all naturally occurring DNA helices involve an antiparallel pair of strands (i.e. the 5' end of one strand is paired with the 3' end of the other.) However, parallel-stranded DNA helices are a rare, but important form of DNA organization. These structures have been implicated in forming reverse gyrase intermediates, in single-stranded DNA packing in bacteriophages, and for RNA dimer stabilization in retroviruses. Moreover, they are relevant to the formation of DNA triplexes in anti-sense therapies. A longstanding puzzle is what the role and geometry of G•C base pairs is in parallel-stranded DNA, since a reverse-G•C base pair would appear to be much more destabilizing than it has been experimentally measured to be.

Unlike the A-T pairs that simply form in a “reverse-Watson-Crick” (RWC) orientation, there is no obvious way that reverse G-C pairs can form stable bonds in parallel stranded DNA, considering the dual carbonyl-carbonyl and amino-amino electrostatic clashes (structure b). We and others showed that rare-tautomers can form stable DNA base-pairs, in the RWC geometry (structure c), as was later observed for the closely related cytosine-isocytosine pairs. Including the effects of solvation, however, we found that a “reverse wobble pair” (structure d) is actually much more stable than the rare-tautomer



pair, and thus we have predicted that this form is favored in parallel-stranded DNA. These results have been published in the *Journal of Physical Chemistry*, (2000).

#### **2.1.6. Prediction of Protein Binding Ligands—Computational Docking**

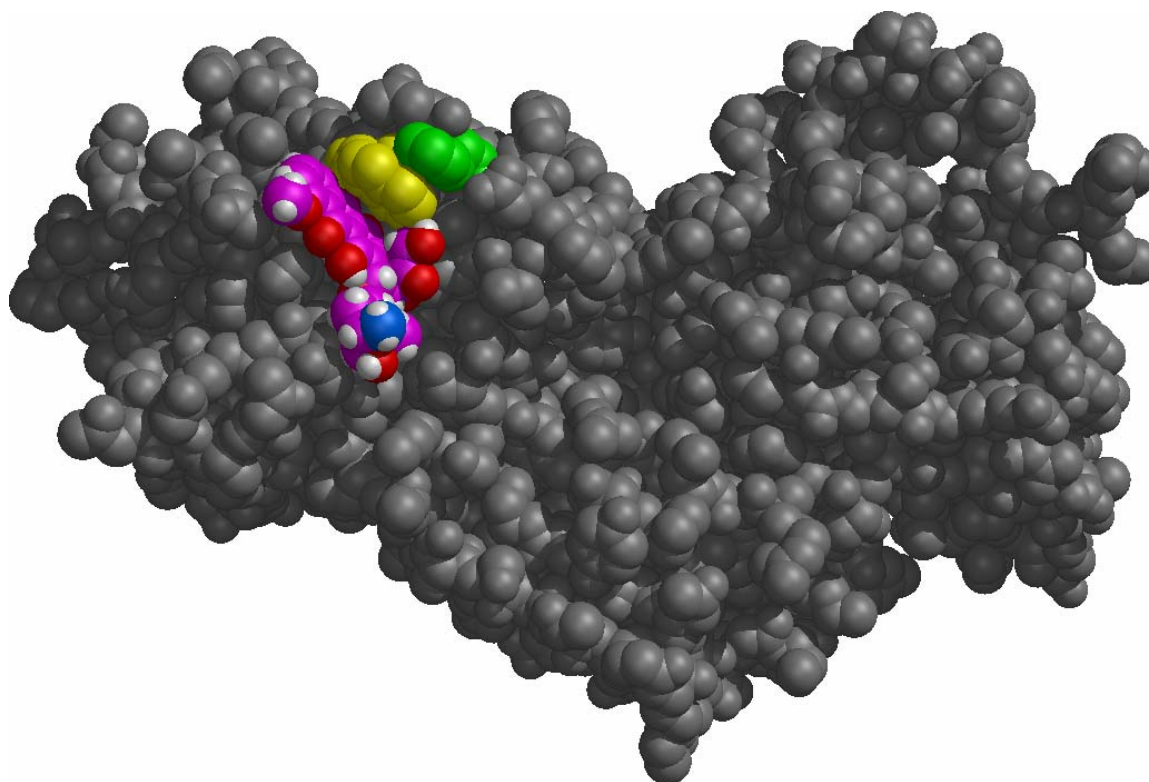
Tetanus toxin belongs to a family of Clostridial protein neurotoxins for which there are no known antidotes. Another closely related member of this family, botulinum toxin, is being used with increasing frequency by physicians to treat severe muscle disorders. Botulinum toxin has also been produced in large quantities by some groups for use in biological weapons. The apurinic/apyrimidinic endonuclease (Ape1) enzyme is an essential enzyme in all higher life forms as an irreplaceable factor in DNA repair. To identify small molecule ligands that might bind to the targeting domain or active sites of these proteins and to facilitate the design of inhibitors and new reagents for their detection, computational docking calculations were used to screen the Available Chemicals Directory for compounds. In the case of the C fragment of tetanus toxin, more than five out of eleven of the predicted ligands were found to bind to the protein.

Recently, computational methods such as docking have been used to speed up the process of drug discovery and inhibitor design by screening large numbers of molecules and predicting whether or not they bind into the active sites of target proteins (Desjarlais et al., 1990; Mao et al., 1998; Olson and Goodsell, 1998; Rutenber et al., 1993). These efforts have met with moderate success in the design of new drugs effective against HIV proteins critical for infection and transmission of the disease, and it is expected that this approach will prove to be generally useful as a first step in the identification of lead (preliminary) compounds that usually bind in the micromolar range. The binding of lead compounds to the protein can be improved by several orders of magnitude by using multiple (2-3) compounds linked together. For the inhibitor to be effective, it needs to recognize specifically the target protein and bind with high affinity. We give two examples of our successes in using docking as a computational screening technique to identify compounds that bind to tetanus toxin C fragment and Ape1.

##### **2.1.6.1 Docking example 1—tetanus toxin (TeNT)**

We used the coordinates obtained from the crystal structure of the C fragment of TeNT (PDB accession code 1A8D), which has been solved to 1.57 Å resolution (Knapp et al., 1998), were used for modeling and docking studies. The solvent accessible surface of the tetanus toxin C fragment was calculated to identify surface pockets as sites for ligand binding. Fifty-two pockets were identified as potential binding sites for small molecules. Four initial sites were selected as potential target binding sites based on the available experimental data about the residues involved in ganglioside GT1B binding, the native substrate. Based on experimental evidence, the site including Trp1289 and proximal to His1293, Site 1, was the first choice for initiating the docking calculations to find a bound ligand which would more likely interfere with ganglioside binding. Following the selection of the site, the computational docking program, DOCK (Ewing and Kuntz, 1997) (UCSF), screened the Available Chemicals Directory (ACD, purchased from MDL, San Leandro, CA) and predicted which molecules would likely bind tightly to the tetanus toxin C fragment. Different orientations of the ligands within the binding site were examined, but different conformations of the ligands were not examined because only rigid docking was performed.

The ligands were each scored and ranked by energy and contact. Though the molecules are ranked based on the scores, the scoring function does not predict the binding affinities. Therefore, the top 1% of scored compounds was visually examined. Specific interactions, such as charge and hydrophobic interactions, were qualitatively noted for each compound. A variety of molecules were chosen to represent the spectrum of available compounds even though peptides dominated the top 1% of compounds. Twenty-nine final compounds were selected as potential binders. One of the ligands, doxorubicin, is shown docked into Site 1 in the figure below. Due to their limited availability, only eleven of these ligands were purchased and tested for binding.



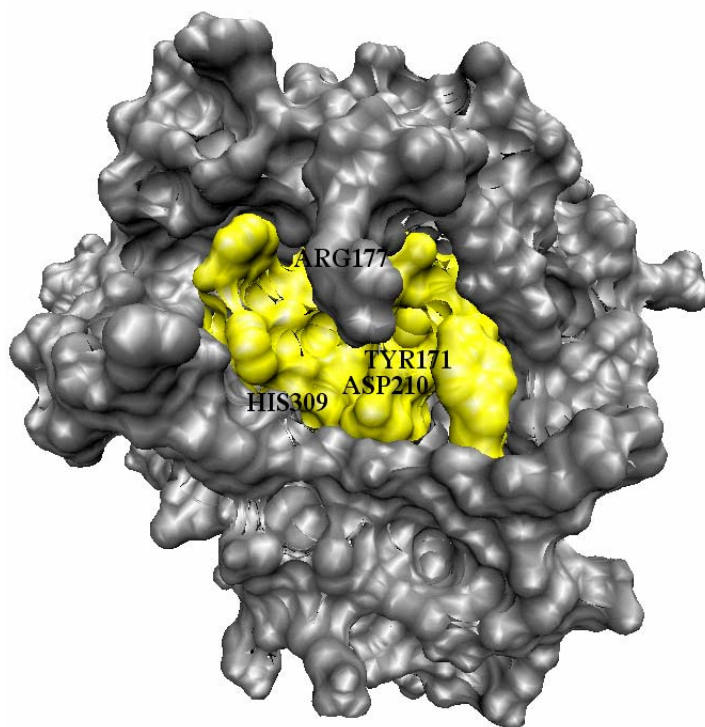
Doxorubicin docked into Site 1 of the heavy fragment of tetanus toxin. The C fragment is shown in black and white. Doxorubicin carbons are shown in magenta; oxygens are in red; nitrogens are in blue; and hydrogens are in white. Doxorubicin is  $\pi$  stacked with Trp 1289 (yellow) and His 1293 (green).

Using electrospray ionization mass spectrometry (ESI-MS) to verify ligand binding, 5 of the 11 (45%) predicted compounds were tested and found to bind to the TeNT C fragment. This success rate of 45% is exceptional particularly since only 10-40% of the predicted compounds identified in other docking studies were shown by Roe and Kuntz to bind to target receptors in the micromolar range. Further testing of these compounds using a competitive liposome binding assay revealed that one of the molecules, doxorubicin, competed with the native ganglioside GT1b for binding to the protein with a dissociation constant of 9.4  $\mu\text{M}$  which is in extremely good agreement with the ESI-MS determined binding constant of 10.6  $\mu\text{M}$ . This work has been published in *Chem. Res. Toxicol.* (Lightstone et al., 2000).

#### **2.1.6.2. Docking example 2—apuridinic/apyridimic endonuclease 1 (Ape1)**

We have also used computational docking to screen for potential ligands of Ape1 from the Available Chemical Database, which consists of >250,000 commercially

available compounds. In brief, using the crystal structure information of Ape1 (Beernink et al., 2001), we first calculated the solvent accessible surface of the entire protein. We then identified potential binding sites within the Ape1 active site. In brief, all compounds from the database were docked in the defined Ape1 active site (see the figure below), and the fit of the ligand was evaluated by energy minimization, also known in the literature as “force-field scoring”. Once all compounds were docked in various orientations and ranked in order of energy (least/minimal to highest), the top 1% of compounds were *visually* evaluated by computer graphics for their chemical likelihood of binding within the Ape1 active site, prior to experimental assessment.

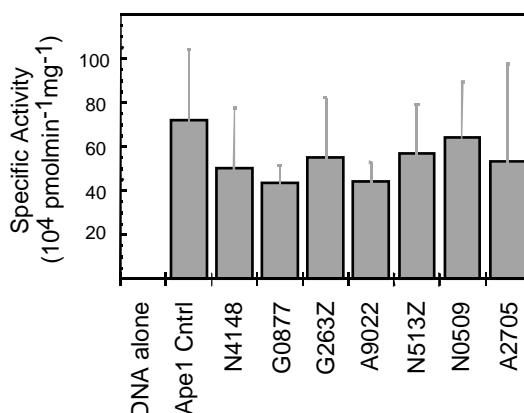


The calculated surface area of Ape1. Shown in yellow are the residues that define the protein's active site. This site contains the amino acids known to be involved in the nuclease function of Ape1. Several key functional amino acids are labeled for reference.

Top candidates from the docking screens above, as determined by (1) potential clinical utility (i.e. solubility, toxicity, and ability to transport/diffusion through cell membranes) and (2) availability and cost, were then tested for their inhibitory effectiveness using established function-based biochemical assays. In brief, these studies involved pre-incubation of the inhibitor with purified Ape1 protein, and subsequent measurement of the



ability of Ape1 to cleave (“repair”) labeled AP-DNA substrates (Wilson III et al., 1995). In one experiment shown in the bar graph below, we found that two (G0877 and A9022) of seven molecules (predicted by computations to be candidate inhibitors) prevented Ape1 incision activity significantly. In total, 7 of 27 (26%) candidate inhibitors (all data not shown) were found to inhibit Ape1 abasic endonuclease activity at least 50%. Further analysis of these chemicals is currently underway.



Candidate Ape1 inhibitors. 30  $\mu$ M of each inhibitor was incubated with 30 pM Ape1 for 20 minutes on ice. Radiolabeled abasic DNA substrate was then added, and this mixture was incubated for 5 minutes at 37°C. The DNA substrate and product were separated by denaturing gel electrophoresis, and the specific activity was calculated as the percent conversion of full-length substrate to incised DNA product. The average and standard deviation of 5 endonuclease assays is shown. “DNA alone” is the negative control containing no Ape1 protein. The Ape1 positive control (“Ape1 Cntrl”) is a reaction without inhibitor, the seven remaining entries in the plot show the results for seven inhibitors predicted by docking.

### 2.1.7. Solution Structure of the Ligase III BRCT domain

BRCT domains are relatively small (~ 100 amino acids), autonomously folded segments of proteins involved in the regulation and recognition processes between two proteins in the cell and are a part of several proteins involved in DNA repair and cell cycle checkpoint mechanisms<sup>12</sup>. The name is derived from the BRCA breast cancer genes and mutations within these domains are considered important in the carcinogenesis process. More than 50 putative BRCT domains have been identified on the basis of sequence similarity alone, a similarity that in some cases is very weak. The homology-based protein structure prediction methods (described below) can play an important role in understanding

the function of BRCT and providing a more rigorous, structure-based criteria for identification of BRCT domains. Until now, however, no BRCT structures were available so that homology-based methods could not be applied. Using NMR, we have determined the 3D solution structure of the human ligase III BRCT domain (L3BRCT) by using NOE derived distance restraints, J-coupling constant data and amide proton exchange data.

A combination of  $^{15}\text{N}$  and  $^{13}\text{C}$ -edited and  $^{13}\text{C}$ -filtered three dimensional experiments carried out on the protein at pH 6.5 and 15 °C yielded proton, nitrogen and carbon chemical shifts for the majority of the 86 residues (See Color Plate 7). The 3D structure of L3BRCT shows structural homology to the x-ray structure of the C-terminal BRCT domain of the repair protein XRCC1.<sup>13</sup> As shown in Color Plate 8, the hydrophobic core consists of a  $\beta$ -sheet, comprised of four parallel  $\beta$ -strands, with two  $\alpha$ -helices packed against one face of the  $\beta$ -sheet. One of the two  $\alpha$ -helices present in the XRCC1 BRCT structure is absent in the structure of the L3BRCT. The L3BRCT is observed to be a symmetric dimer in solution as evidenced by both NOE data and self-diffusion coefficient measurements. This structure will form the basis for both modeling the structures of other BRCT domains, and, combined with the proposed additional NMR experiments, simulations of the interactions between BRCT domains.

## **2.2 Advanced Computational Chemical Methods**

### **2.2.1. Applications of First Principles Molecular Dynamics**

During this Strategic Initiative LDRD Project, we performed first principles molecular dynamics (FPMD) simulations of liquid water, extending these to higher pressures and temperatures. Additionally, we performed FPMD simulations of solvated sodium and magnesium ions and dimethyl phosphate, the chemical structure comprising the DNA backbone.

#### **2.2.1.1. Liquid water simulations**

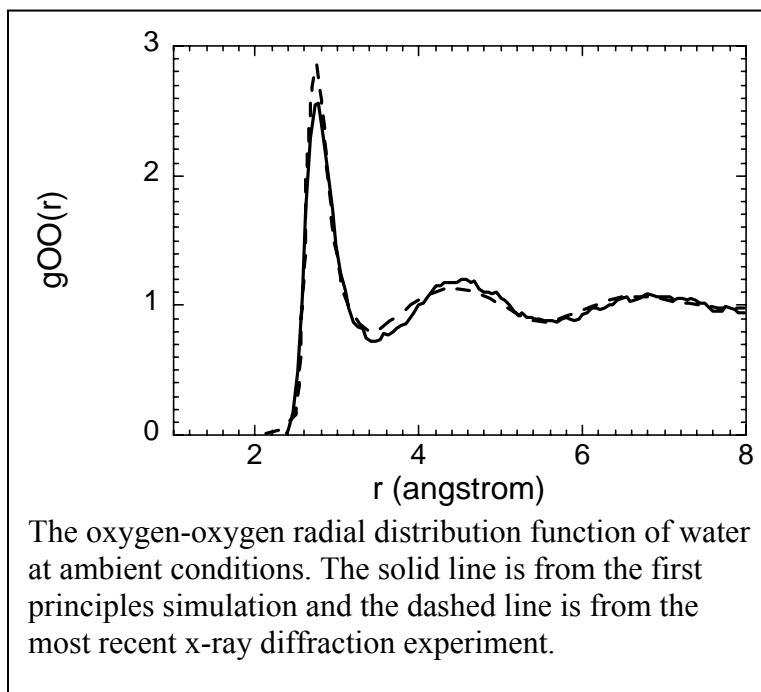
Water has been the subject of numerous experimental and theoretical investigations, given its paramount importance in the physical sciences and its key role in life. Nevertheless, many fundamental questions about the properties of water are yet unanswered. For example, most investigations of the effect of compression on the

microscopic structure of liquid water have been performed at low pressures ( $< 0.1$  GigaPascals (GPa)), and only a small number of experiments have been performed at pressures up to 2 GPa. For pressures larger than 2 GPa, information on the liquid is even more limited; site-site pair correlation functions and structure factors have not been measured, and structural data are only available from simulations based on empirical potentials.

We have investigated the bulk properties of liquid water with FPMD simulations of 54 water molecules in a box with periodic boundary conditions. According to systematic empirical studies of water as a function of the simulation size, 54 water molecules constitutes a sufficiently large sample to reliably model the properties of the liquid state. The nuclear motion was described by Newtonian dynamics, and many body interactions between electrons were described by density functional theory in the local density approximation, and with the PBE generalized gradient approximation. To make a direct comparison with available experimental data, and to determine the accuracy of our theoretical model, we have computed the structural properties of water at ambient condition.

It is often convenient to describe the structural properties of a liquid with pair radial distribution functions,  $g_{\alpha\beta}(r)$ , which represent the probability, relative to a random distribution, of finding

an atom of type  $\beta$  at a distance of  $r$  from an atom of type  $\alpha$ . In the figure above, the oxygen-oxygen radial distribution function  $g_{OO}(r)$  for liquid water obtained from the simulation is shown. Also shown in the figure above, is the  $g_{OO}(r)$  from the latest x-ray

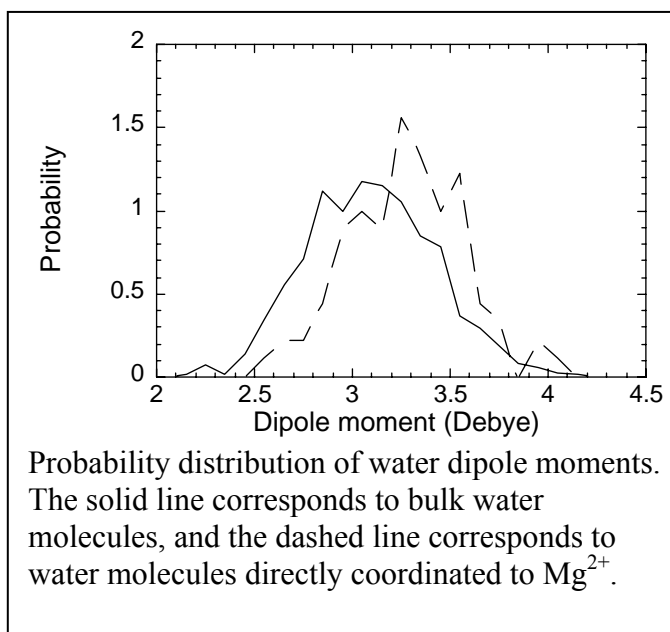


diffraction experiment. Overall, the agreement with experiment is very good. The peak positions in  $g_{OO}(r)$  are essentially identical between the simulation and experiment.

#### 2.2.1.2. Ion solvation

The solvation of ions is a fundamental process that is found in a wide range of biological and chemical systems. In particular, the manner in which water solvates alkali and alkaline earth cations is relevant to systems such as enzymatic catalysis and the structural stability of DNA and RNA. We have performed FPMD simulations of magnesium (II) and sodium (I) in water. By comparison with the available experimental data on these ions in solution, we are able to directly evaluate how well solute-solvent interactions are reproduced in the simulations. In addition, comparisons are made to classical MD simulation models, which clearly demonstrate the advantages of using a highly accurate first principles simulation approach over more traditional simulation models.

We have performed a FPMD simulation of a magnesium cation ( $Mg^{2+}$ ) in liquid water. In agreement with experimental measurements (Caminiti et al., 1977) and classical MD simulations (Szász et al., 1982), we find that  $Mg^{2+}$  prefers to form a stable octahedral complex in water. However, a careful analysis of the first solvation shell around  $Mg^{2+}$  reveals that the water molecules tend to coordinate to  $Mg^{2+}$  in an asymmetric orientation. This orientational property is not reproduced by classical MD simulations (Szász et al., 1982) or in QM calculations of  $Mg^{2+}$ /water clusters (Bock et al., 1994; Pavlov et al., 1998).



In addition to the structural properties of the first solvation shell around  $\text{Mg}^{2+}$ , we have investigated the difference between the electronic structure of the first solvation shell waters and the “bulk” waters (the water molecules outside of the first solvation shell). This analysis was performed by computing maximally localized Wannier functions in a manner similar to the Boys localization procedure that is commonly used in quantum chemistry (Marzari and Vanderbilt, 1997). The localized Wannier functions can be associated with features such as distributions of lone pairs and covalent O-H bonds and can be used to define local dipole moments for each of the water molecules in the simulation. A probability distribution of the water dipole moments in our FPMD simulation is shown in the figure above. As can be seen, the dipole moments of the bulk water molecules form a broad distribution from 2 to 4 Debye, with an average value of 3.1 Debye. For the waters in the first solvation shell around  $\text{Mg}^{2+}$ , the distribution is shifted 0.2 Debye to an average value of 3.3 Debye. These subtle properties of  $\text{Mg}^{2+}$  may prove to be essential in the accurate simulation of chemical reactions where  $\text{Mg}^{2+}$  is known to assist in the hydrolysis of phosphodiesteres, such as in the cleavage of DNA and RNA.

We have also performed FPMD simulations of a sodium cation ( $\text{Na}^+$ ) in water. In agreement with experimental measurements (Skipper and Neilson, 1989), we find that there are on average 5.2 water molecules in the first solvation shell around  $\text{Na}^+$ , and the average separation between  $\text{Na}^+$  and its first solvation shell is 2.49 Å. During the course of the simulation a number of waters exchange between the first and second solvation shell, which indicates that the first solvation shell is soft and flexible. None of the existing classical simulations appear to be able to reproduce this combination of average  $\text{Na}^+$ -O distance, number of waters in the first solvation shell, and exchange of water molecules between the solvation shells. These differences between our first principles and classical potential description (Dietz et al., 1982) of the solvation of  $\text{Na}^+$  were clearly seen in our calculated distribution of first solvation shell tilt angles. The classical MD simulation results in a tilt angle distribution that is strongly peaked at  $\theta=180^\circ$  and decreases to zero at  $\theta < 90^\circ$ . In the FPMD simulation, the tilt angles exhibits a broad, flat distribution that is weakly peaked at  $\theta=180^\circ$  and slowly decreases to zero at  $75^\circ$ . From these simulations, it is evident that including many-body effects beyond pair-wise interactions along with a high-

quality description of the solute-solvent interactions is crucial for describing the aqueous solvation of cations.

### 2.2.1.3. Conformational dynamics of dimethyl phosphate

The dimethyl phosphate anion ( $((\text{CH}_3\text{O})_2\text{P}(=\text{O})\text{O}^-$  or  $\text{DMP}^-$ ) is often used as a simple model of the phosphodiester linkage that is found in the backbone of DNA. The different conformers of  $\text{DMP}^-$ , which are shown in this figure, have been extensively studied with a

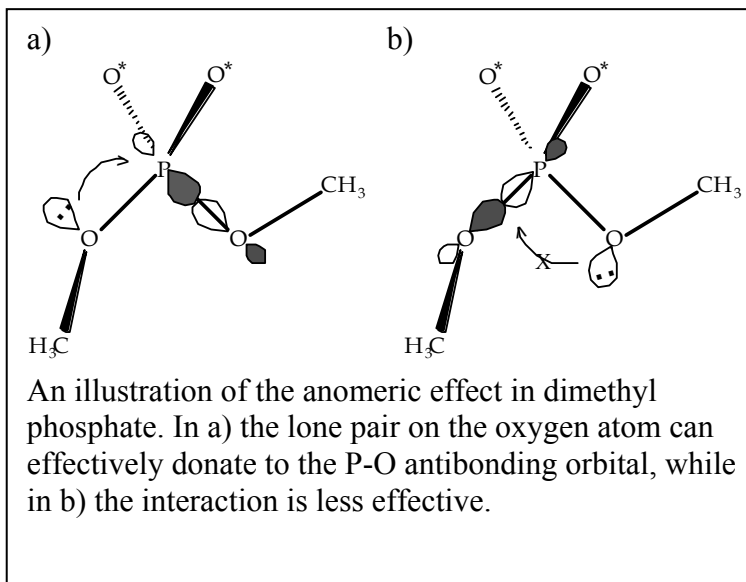
variety of theoretical models ranging from high level QM calculations on gas phase structures (Florián et al., 1996) to classical MD simulations of  $\text{DMP}^-$  in solution

(Jayaram et al., 1988).

The  $-sc$  and  $ap$  notation used in the figure here is the Klyne-Prelog notation

for describing the  $\alpha$  and  $\zeta$  torsion angles ( $s$  stands for syn,  $a$  for anti,  $c$  for clinal, and  $p$  for periplanar) (Saenger, 1984).

From a variety of gas phase QM calculations, the total energies of different conformers of  $\text{DMP}^-$  are arranged in the order  $-sc/-sc < -sc/ap < ap/ap$  (Florián et al., 1996; Liang et al., 1993; MacKerell, 1997; Murashov and Leszczynski, 1999). This ordering is the opposite of what might be expected based on steric arguments and is due to the generalized anomeric effect (Kirby, 1983). More specifically,  $\text{DMP}^-$  has a stabilizing interaction caused by a lone pair located on  $\text{O}_5'$  that partially donates charge to the  $\text{P}-\text{O}_{3'}$  antibonding orbital (and likewise a lone pair on  $\text{O}_{3'}$  can donate to the  $\text{P}-\text{O}_{5'}$  antibonding orbital). As illustrated in the figure (a), when a methyl group in  $\text{DMP}^-$  is oriented in the  $-sc$  conformation, it is possible for the lone pair located on the oxygen to effectively donate to the antibonding orbital. In turn, the  $\text{P}-\text{O}$  bond is lengthened. In the case when the methyl group is in the  $ap$  orientation [(b) in the figure], the lone pair does not effectively donate to



of

the antibonding orbital because it overlaps with alternating positive and negative lobes of the orbital (Liang et al., 1993).

Although the importance of anomeric effects on the conformational stability of DMP<sup>-</sup> is known in the gas phase, it is not clear if these stabilizing interactions persist in the presence of an aqueous environment. Both experimental (Praly and Lemieux, 1987) and theoretical evidence (Cramer and Truhlar, 1993) on related systems indicate that the anomeric effect decreases as the polarity of the solvent increases. In order to better understand how the environment may affect the properties of DMP<sup>-</sup>, we have performed FPMD simulations of DMP<sup>-</sup> in solution with a sodium counterion (see Appendix). In these simulations, structural changes are observed in DMP<sup>-</sup> that are indicative of a decrease in the anomeric effect. For example, the average values of the P-O<sub>3'</sub> and P-O<sub>5'</sub> bond lengths decrease by 0.04 Å in solution as compared to gas phase structures. This decrease in the anomeric effect is caused by a redistribution of charge within DMP<sup>-</sup> that is not generally accounted for in a classical MD simulation model (MacKerell, 1997; Murashov and Leszczynski, 1999). In addition, an interesting mechanism is observed in the FPMD simulation of the -*sc*/-*sc* conformer where the presence of the sodium counterion initiates a change between the -*sc*/-*sc* and -*sc*/*ap* conformers. These simulations demonstrate that FPMD provides an effective means of performing realistic simulations of biologically relevant molecules.

#### **2.2.1.4. Other applications of first principle molecular dynamics**

As it has been developed, the first principles molecular dynamics software has been applied to a number of other chemical systems of wide interest to LLNL programs. These applications have broadened the impact of this project, provided additional validation for the simulation methodology, and helped develop continuing funding sources. Several of these projects are described in the following paragraphs.

In addition to water at ambient conditions, we have also performed a series of first principles molecular dynamics simulations on liquid water at elevated temperatures and pressures. In particular, we have examined to properties of liquid water under a pressure of 10 gigapascal and a temperature of 600 K. Under these extreme conditions, each water molecule is closed packed and surrounded by 12.9 molecules, as opposed to 4.5 neighbors

at ambient conditions. The changes in the atomic structure, which cause a disruption of the hydrogen bond network, are accompanied by sizeable changes in the electronic density as well. Recently, Jon Eggert and Paul Loubeyre at the CEA have performed x-ray diffraction measurements of water in a diamond anvil cell and have found identical changes in the structure of water under pressure.

Hydrogen fluoride is a hydrogen bonded system, which, contrary to water, has a layered structure in the crystalline phase, at ambient density. The liquid state of HF at low temperature and zero pressure is also quite different from liquid water, exhibiting a two dimensional chain-like structure, as opposed to a three dimensional diamond-like network. In order to analyze the main differences between hydrogen bonding in H<sub>2</sub>O and HF, we have carried out a series of simulations in the liquid state, at low and high pressure ( $P=0$ , 10, 100 and 150 GPa) and at several temperatures. Our results indicate that in the non-dissociative regime, the effect of temperature on HF hydrogen bonding at low pressure is similar to that observed in liquid water. Furthermore, while very different at low pressure, liquid water and HF resemble each other at high pressure. At a pressure of 10 GPa, the hydrogen bonding is substantially weakened and the molecules arrange into a close-packed structure in both systems.

Using first principles molecular dynamics, we have studied the structural and electronic properties of liquid deuterium under pressure, in a range of densities relevant to recent experiments with high intensity laser-shocked samples. Our results show that at densities ranging from 5-fold to 7-fold compression, the liquid goes from a recombination regime, where a substantial proportion of atoms form D<sub>2</sub> complexes, to a metallic scattering regime, where diatomic complexes are short-lived and dissociated deuterium atoms predominate. This transition occurs in a continuum fashion. In shock wave experiments the Hugoniot relations, which are conservation laws for mass, momentum, and energy, are often used to determine changes in variables across a shock front as it passes through a material. Our simulations predict a compression in D<sub>2</sub> lower than that found experimentally.

We have used first principles molecular dynamics to investigate the stability and bulk properties of various candidate structures of a new phase of CO<sub>2</sub>, recently synthesized by C.Yoo and collaborators. Total energy results and the comparison of computed and



measured X-ray diffraction patterns point at trydimite as the structure of the newly discovered CO<sub>2</sub>-V phases.

## **2.2.2. Development of First Principles Molecular Dynamics**

### **2.2.2.1. Algorithm development**

In ab initio molecular dynamics simulations, basis sets based on plane wave expansions along with the pseudopotential approximation are the method of choice for solving the Schrodinger equation. The use of plane waves over other types of basis sets, such as Gaussian type functions, is preferred for a number of reasons. Plane waves enable a straightforward convergence in the quality of the basis set. Since plane waves are not a function of the atomic coordinates, forces can be efficiently computed via the Hellman-Feynman theorem. Plane waves do not suffer from basis set superposition error. Lastly, plane waves allow the use of the fast Fourier transform, which is computationally very efficient. Several algorithmic improvements were developed during the course of this project and were integrated into our First-principles simulation software.

### **2.2.2.2. JEEP code development**

The implementation of First-Principles Molecular Dynamics (FPMD) used in this project is the JEEP code. JEEP is a C/C++ implementation of the plane-wave, pseudopotential formalism of FPMD. JEEP was imported to LLNL in April 1998 as a serial code (release 1.2.0) and was parallelized in order to exploit the computational power of the large parallel platforms available at LLNL (e.g. ASCI Blue). A first parallel version based on the Message Passing Interface (MPI) was developed on ASCI Blue in the summer of 1998. ASCI Blue consists of about 300 nodes, each of which comprises four CPUs that share a common memory of about 1.5 GB. This architecture is best exploited using two kinds of parallel programming models: message-passing between nodes, and shared-memory parallelism (using e.g. multiple threads) between CPUs within nodes. Early parallel versions of JEEP used this two-level parallelism by way of vendor-supplied multithreaded libraries. This resulted in near-optimal performance within nodes, since a large part of the computing time in JEEP is spent in numerical libraries performing linear

algebra (BLAS, LAPACK) of Fourier transforms (IBM ESSL library). This approach proved to be efficient, as was demonstrated in a series of large-scale simulations performed on the IBM SP SKY computer in Oct 98. The SKY platform is an assembly of 1536 IBM-SP nodes arranged in 3 sectors communicating by means of a high-performance switch. Sectors consist of 488 nodes interacting through an SP switch. Each node comprises 4 PowerPC 604e CPUs similar to those of ASCI Blue.

The JEEP code was used to perform FPMD simulations of a mixture of hydrogen fluoride and water (HF/H<sub>2</sub>O) on the SKY platform. The simulation included 600 atoms (120 HF + 120 H<sub>2</sub>O) and ran on partitions of 480 and 960 nodes of SKY (1920 and 3840 CPUs) for about 2 weeks of wall-clock time. The results showed the formation of molecular complexes (H<sub>3</sub>O<sup>+</sup>-F<sup>-</sup>, F<sub>2</sub>H<sup>-</sup>) during the simulation. This simulation demonstrated the feasibility of large-scale FPMD simulations of the largest LLNL parallel platforms. A parallel efficiency of 88% was measured by comparing simulations running on 320 nodes (1280 CPUs) and 480 nodes (1920 CPUs). Further optimization of JEEP for runs on such large numbers of CPUs was not pursued given the limited access to large portions (>128 nodes) of ASCI platforms for regular use. Several features needed for calculations relevant to biochemistry were added to JEEP in FY99. New density functionals were implemented (BLYP, Becke+LDA) for comparison with results obtained with atom-centered basis sets (Gaussian code). Data analysis utilities were also written for the extraction of statistical quantities from FPMD trajectories (ion-ion correlation functions, diffusion coefficient).

In FY00, the JEEP code underwent major refactoring, with the aim of reorganizing the code following standard principles of object-oriented design (releases 1.5.x). This redesign considerably reduced class interdependencies, and established a sound basis for further development of the code. Concomitantly, the code was modified to conform to the ISO/C++ Standard, and to take advantage of the functionality of the C++ Standard Template Library (STL) for increased portability. The following new features were added to JEEP:

- Inclusion of holonomic distance constraints in molecular dynamics simulations. This allows us to compute the free energy profile of some simple biochemical

reactions by running simulations in which some interatomic distance related to the reaction coordinate is kept fixed.

- Addition of a preconditioned Steepest Descent optimization algorithm for the calculation of electronic ground states.
- Inclusion of spin-polarization (PBE and LDA functionals). This allows for simulations of simple dissociation reactions such as e.g.  $\text{H}_2\text{O} \rightarrow \text{H}^+ + \text{OH}$ —in vacuum, in which molecular fragments have a finite spin.

The code was also parallelized using OpenMP multithreading directives. This led to a more efficient use of multiple CPUs within nodes, in sections of the code that do not involve the linear algebra or FFT libraries. A visualization program (map3Dv) was developed for the generation of high quality graphical representation of molecules and electronic charge densities. Map3Dv is based on the VTK visualization toolkit, and can be used to generate high-resolution images and movies.

We have also implemented a new scheme for the calculation of the electronic properties of charged molecules. Since the use of a finite set of plane waves amounts to applying periodic boundary conditions, plane wave based calculations are well suited for infinitely repeated periodic systems, such as bulk solids. In disordered systems, such as liquids, the enforced periodicity has only a small effect on bulk properties as long as the simulation cell is of moderate size. For molecular calculations, the situation is not as favorable. In order to eliminate interactions between mirror images, the simulation cell is usually chosen to include extra "vacuum" space around the molecule, which can be computationally expensive. Furthermore, because the Coulomb repulsion energy diverges in periodically repeated charged systems, energy differences between systems with different total charge are ill-defined.

We have implemented a new method that greatly facilitates the calculation of neutral and charged molecules when a plane wave basis set is employed. The method is based on a modified Hamiltonian where the Coulomb potential is smoothly truncated in order to completely remove interactions between cells. By retaining all of the desirable qualities of a plane wave basis, the method is capable of highly accurate as well as efficient calculations on molecular systems.

## **2.3 Protein Structure Prediction**

### **2.3.1. Homology-based Protein modeling**

With the success of the sequencing projects, the structural characterization of the genome-encoded proteins becomes increasingly important. However, the determination of the three-dimensional structure of the genome-encoded proteins that is critical to understanding the role the proteins play in the cell continue to lag far behind.

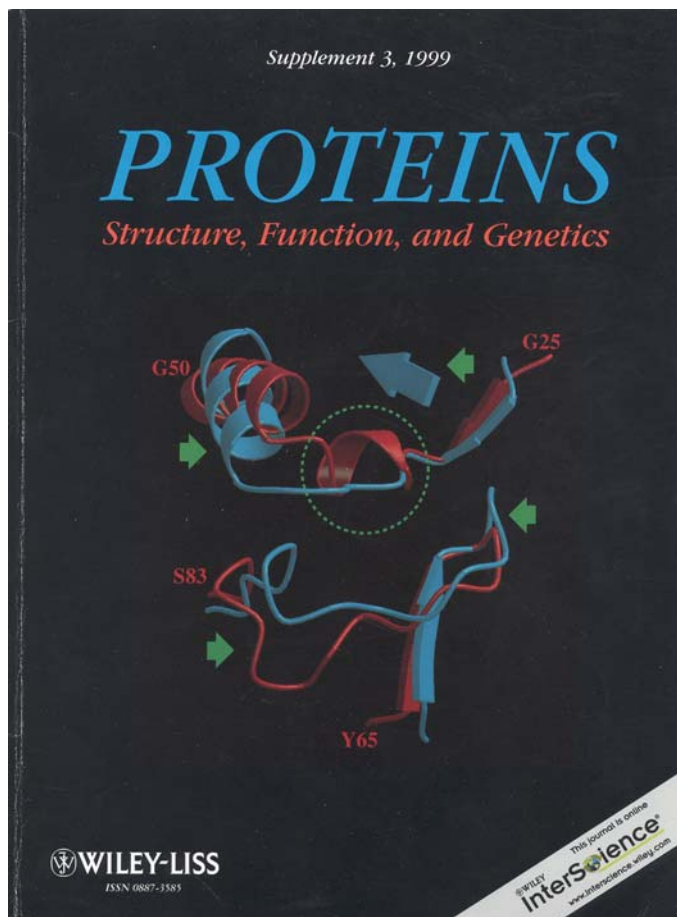
Computational protein structure prediction can and already is contributing significantly to bridge the gap between sequence and structural data. While there are different prediction approaches, comparative modeling among them seems to be the most promising in practical terms as it is able to produce the most accurate and detailed models of protein structures

During the period covered by this report we have been both developing more effective comparative modeling techniques as well as applying these techniques to facilitate understanding of the molecular mechanisms in specific biological processes, such as DNA replication, repair and cell cycle checkpoints.

#### **2.3.1.1. External Evaluation of our homology-based protein modeling methods**

The development of our comparative modeling technique was very favorably stimulated by “blind” mode international experiments on protein structure prediction, called “Critical Assessment of Structure Prediction” or “CASP” in short. We have analyzed prediction data from many different groups to identify major problems as well as to determine contribution of various factors to the accuracy of the models. Based on this analysis we concentrated on improving very important step in comparative modeling – sequence-structure alignment. One of the distinct features of this newly developed method included usage of evolutionary information in the form of multiple sequence alignments to provide both sampling and estimation of the reliability of alignments for particular regions. In conjunction with testing possible alignment variants in the framework of the three-dimensional models, this procedure turned out to be quite effective. For example, during the third experiment of “blind” structure prediction (December of 1998) our method was ranked by independent experts among the top-performing in the field of comparative modeling. Based on the success of our modeling we were invited to publish a description

of our method along with our results in a special issue of *Proteins*, featuring some of our modeling results as a cover story (see the figure below).



Cover of *Proteins Structure Function and Genetics*, 1999 featuring our comparative modeling results.

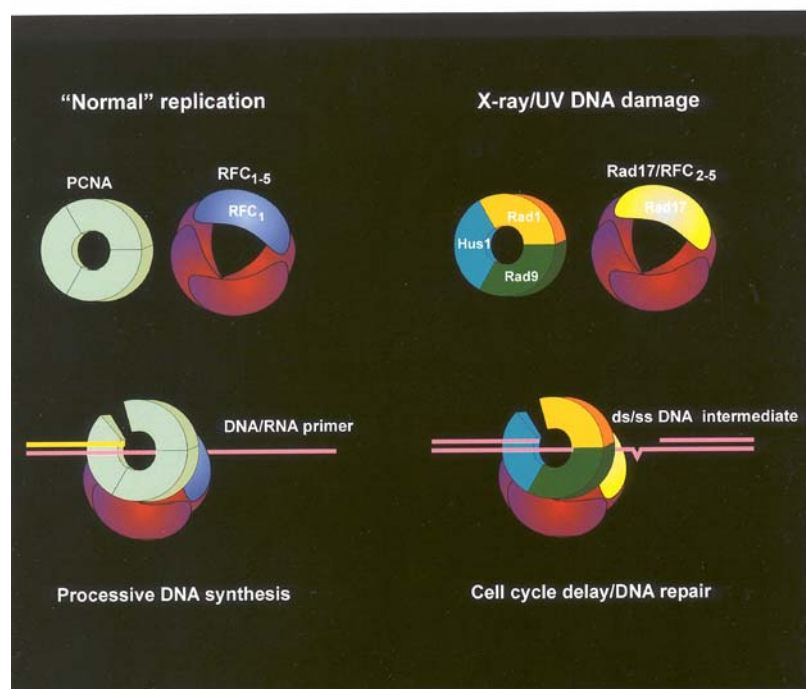
#### **2.3.1.2. Homology-based structure prediction of DNA sliding clamp proteins**

In the area of application of comparative modeling to understand mechanisms of DNA replication, repair and cell cycle checkpoints we contributed by structurally characterizing several protein families involved in these processes. Perhaps even more importantly, we were able to extend the application of the modeling approach from constructing structures of single proteins into modeling entire protein complexes. Initially, we combined fold recognition and comparative modeling to predict the structure of Rad1, a member of cell cycle checkpoint family of proteins, by identifying its similarity with the DNA sliding clamp protein PCNA (Proliferating Cell Nuclear Antigen). During cell division, PCNA forms a ring around the DNA helix and associates with certain DNA polymerases to facilitate processive replication of the genome. Our modeling results were

published in the prestigious journal, *Cell*. Subsequently, we have extended modeling study to several other different proteins functioning in the same pathway as Rad1. Two of these, Rad9 and Hus1, together with predicted structure of Rad1 turned out to be similar to PCNA. By analogy to the mode of PCNA functioning, we also were able to postulate the clamp-like ternary structure for the Rad9, Rad1 and Hus1 protein complex, known to be important for the normal DNA repair function. Our proposed model strongly stimulated biochemical studies of this complex and experimental data that accumulated so far indicates that the modeled structure of the complex is correct.

### **2.3.1.3. Homology based structure predictions of DNA clamp loading proteins**

Another DNA repair protein family, Rad17, has been known to display sequence similarity to clamp loading proteins that are responsible of loading PCNA ring onto DNA allowing for subsequent replication of the genetic material. The functional eukaryotic clamp loading complex, known as replication factor C (RFC), has ATPase activity. ATP binding and subsequent hydrolysis is believed to be associated with the conformational changes of the RFC complex, required to open the PCNA ring, load it onto DNA and seal it. The predicted PCNA-like clamp structure for Rad9-Rad1-Hus1 protein complex (recently termed 9-1-1 complex) and a weak, but significant, sequence similarity to RFC proteins, prompted us to study in detail the Rad17 protein family by molecular modeling. Specifically, we have addressed Rad17 hypothetical ATP-binding function. Of all clamp loading proteins experimentally determined three-dimensional structure was known only for a single protein - *E.coli* polymerase III  $\delta'$  subunit, which is defective in ATP binding. We have identified another protein structure (NSF D2), which has different cellular function, but is able to bind ATP and has the closest structure to that of  $\delta'$  subunit. This structural similarity allowed us to identify regions in  $\delta'$  subunit equivalent to ATP-binding motifs in NSF D2 and the extent of the overall structural conservation. Using series of multiple sequence alignments we were able to map Rad17 sequence regions presumably involved in ATP-binding onto these two three-dimensional structures. The model suggests that Rad17 proteins have a functional nucleotide binding site. It also provides details of predicted Rad17-ATP interaction, suggesting possible roles for individual residues, as well as how ATP-binding/hydrolysis affects conformational changes of the protein.



OXFORD UNIVERSITY PRESS

ISSN 0305-1048 Coden NARHAD

Cover of Nucleic Acids Research, July 1, 2000, featuring our model of DNA replication and repair.

This modeling study allowed us to propose a novel molecular mechanism of DNA repair analogous to the normal DNA replication. In this model, upon extensive DNA damage, the Rad17/RFC complex was predicted to load the 9-1-1 complex (like PCNA) onto damage DNA. The loaded 9-1-1 complex was postulated then to generate a signal to delay cell cycle and also to facilitate DNA repair. At the time this model was featured on the cover of Nucleic Acids Research (see figure above), and by now experimental studies have largely confirmed the predicted mode of action for both of these complexes.

## 2.3.2. Methods Development for Protein Fold Recognition

With the success of the sequencing projects, the structural characterization of the genome-encoded proteins becomes increasingly important. Extensive knowledge of protein structure will significantly aid the investigation of protein function, protein interactions, and biochemical pathways. It will also have a major impact on our understanding of biology, human disease, and eventually on drug design. Experimental determination of structure is inherently time-consuming and costly. At present the structures of less than one percent of proteins that have been sequenced are known, and the rate at which sequence data is accumulating continues to outpace experimental determination of structure by more than two orders of magnitude. Methods of structure modeling and prediction can help close this gap, for example by extending the structural information to proteins within the same sequence family (with methods called homology or comparative modeling) or within the same fold family (fold recognition). However, these techniques require aggressive development to extend the scope of their application and improve their reliability.

#### **2.3.2.1. Methods for accurate superposition of protein structures**

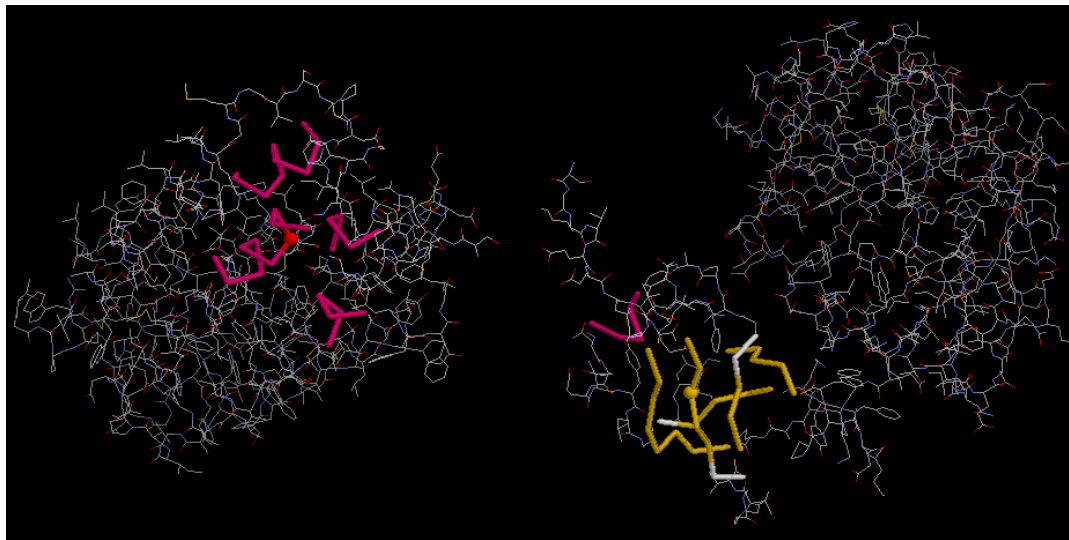
Our new structure superposition algorithm (called LGA for Local/Global Alignment) is based on ideas stemming from local structure classification, and allows for a faster and more reliable identification of regions of similarity among protein structures. Structure superposition in cases of established 1-1 correspondence between residues is a solved problem with a unique best superposition. In the case of different residue sets, or when the best superposition has to be found regardless of the sequence-sequence alignment, the problem is more difficult, with only approximate solutions found in reasonable time. Our new algorithm is based on the idea of local structure classification, and is more robust and faster than other such algorithms we have tested. Applications include (1) Identification of regions of greatest similarity /most extensive similarity between two protein structures, (2) Database-wide classification of similarity relative to the specific structure under consideration. (3) Database-wide organization of structures based on structure homology.

#### **2.3.2.2. Methods for the Identification of Distant Protein Homologies**

In the area of distant homology identification we have further developed and tested three separate techniques of local structure encoding. A newly calibrated measure of



structure similarity allowed for a systematic classification of protein structure on an unprecedented scale. We are currently capable of classifying discrete elements of structure, called local structure descriptors, for the entire database of publicly known protein structures. At this point, and again after further organization of the collected data performed with clustering techniques, it is possible to derive the sequence/structure signal embedded in these data. In addition, a number of analyses of the basic properties of protein structures are now possible.



Two examples of the local structure topology as identified by the descriptor formalism in two different proteins, a helical region (left) and a mostly  $\beta$ -strand region (right). Helices are shown in red, strands in yellow, and coil in white. Similar local structure may be observed in other proteins and likely will be found in protein structures that are not yet solved.

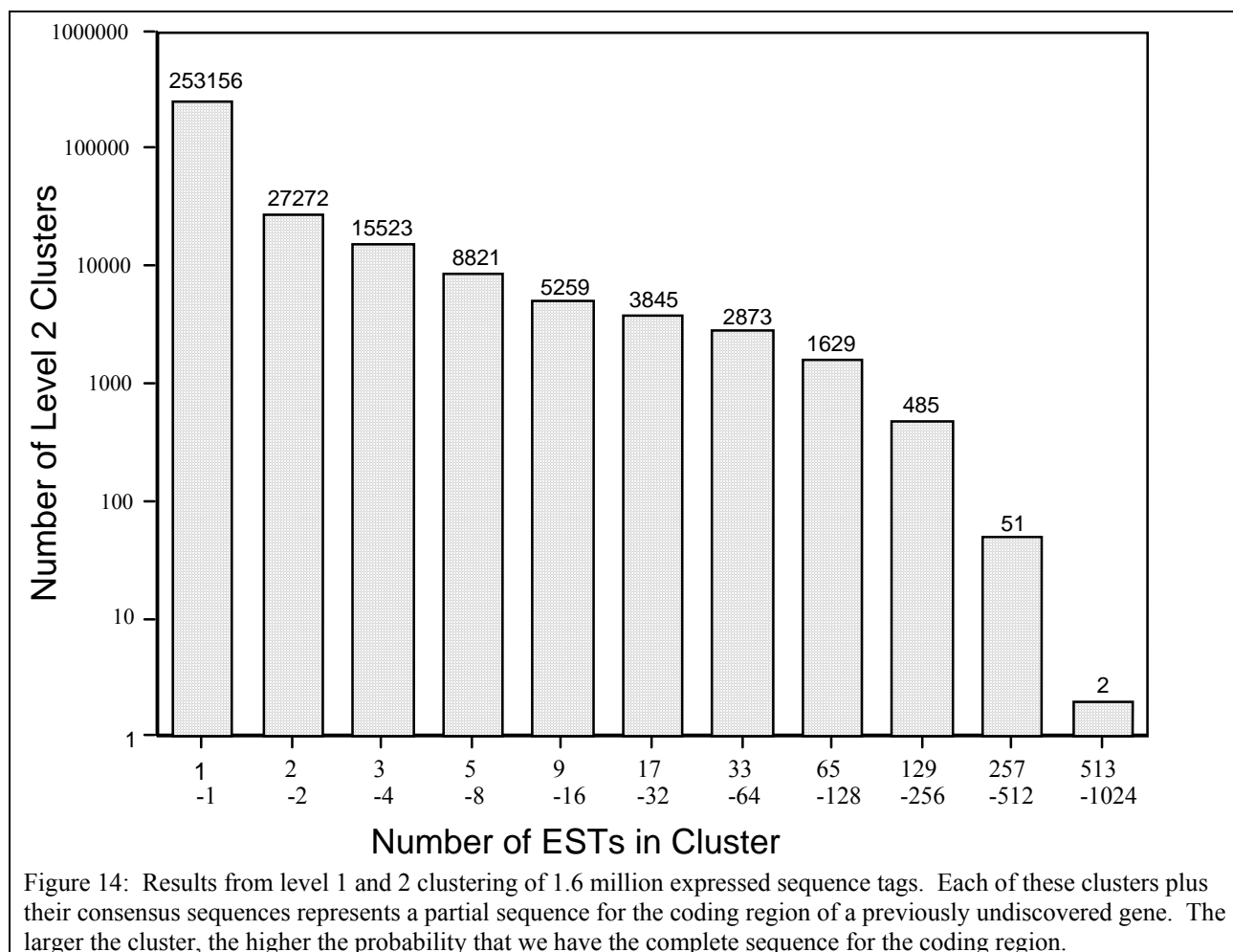
Within our distant homology identification methods development effort we have completed a systematic classification of protein structure on an unprecedented scale of a full all-against-all comparison of known proteins. We have shown that adequate database representation exists for the tightly packed regions of protein structure, sufficient to fully describe each known protein fold, and suggesting that novel, yet uncharacterized, protein structures are built from the same repertoire of small local structure building blocks we call descriptors as shown in the figure above. We have also shown that the sequences associated with each class of descriptors can be generalized to identify known local structures among new protein sequences, leading to the structural characterization of proteins for which little or no structural data are available.

## 2.4 Computational Gene Discovery

There are approximately 100,000 different genes in the human body. However, so far, less than 8,000 of these genes have been discovered (or, more accurately, publicly released). We have used mathematical techniques and computational analysis of unprecedented size to successfully discover significant portions of the DNA sequences for as many as 65,000 previously unavailable human genes. These sequences apply to the “coding region” of the gene, which provides the blueprint for the formation of proteins.

We started with all the ESTs ("expressed sequence tags", short DNA sequences from genes), from Genbank (a publicly accessible Biotechnology Database managed by the NIH) and used the results of an existing LLNL product (IMAGEne-I) to subtract out the ESTs which appear to come from the set of known genes. We then took the remaining ESTs, approximately  $1.4 \times 10^6$  of them and compared them against each other using BLAST sequence similarity searches<sup>22</sup> to form a large number of initial, "Level 1" clusters including approximately 250K singletons, i.e. ESTs that were not associated with any other ESTs. This process is computationally intensive, required approximately 800 cpu-days of computation on ASCI Blue, and would have been impractical without a massively parallel computer.

Information concerning which clusters share ESTs and which ESTs come from the same piece of DNA (called a clone) allowed us to merge the set of Level 1 clusters into a much smaller (and more useful) set of Level 2 clusters each of which contained the maximal sequence information per gene for the given input dataset. The Level 2 clustering algorithm is functionally equivalent to the following steps. Every Level 1 cluster was compared against every other Level 1 cluster and all pairs of clusters that have at least two reasons to merge were combined (see Color Plate 17). This process was then iterated on the remaining clusters until a complete pass with no merges was completed. A consensus sequence (i.e. the DNA sequence most consistent with the overlapping ESTs) of all the sequences in a Level 2 cluster was computed. The resulting Level 2 cluster plus its consensus sequence represents a partial sequence for the coding region of a previously undiscovered gene. The larger the cluster, the higher the probability that the complete sequence for the coding region is present. The results of this clustering are shown in the next figure.



This information is now available to the international research community via the IMAGE web site (<http://bbrp.llnl.gov/imagene/bin/search>). The results of this LDRD subtask, when integrated with the IMAGEne product (funded by the DOE and National Cancer Institute), constitutes a gold mine of data for disease researchers worldwide, affording them a potential shortcut to isolating important genes. The Computational Gene Discovery effort is not static, with updates performed monthly. As more and more ESTs are added to our dataset, the map of sequences for the set of all human genes will become complete, improving the usefulness of this genetic resource.

### 3. APPENDIX

#### 3.1 Publications—In Print, in Press, or Submitted

1. Hatch, F.T., M.G. Knize, M.E. Colvin (2001) Extended QSAR for 80 Aromatic and Heterocyclic Amines: Structural, Electronic and Hydrophobic Factors Affecting Mutagenic Potency *Environmental and Molecular Mutagenesis* (In Press).
2. Williams, R.V., M.E. Colvin, N. Tran, R.N. Warrener, D. Margetic (2000) Exceptionally Pyramidalized Olefins: A Theoretical Study of the Cyclopropenyl Fused Tricycles Tricyclo[3.2.1.0<sup>2,4</sup>]oct-2(4)-ene, Tricyclo[3.2.1.0<sup>2,4</sup>]octa-2(4),6-diene, Tricyclo[3.2.2.0<sup>2,4</sup>]non-2(4)-ene, Tricyclo[3.2.2.0<sup>2,4</sup>]nona-2(4),6-diene *Journal of Organic Chemistry* **65** 562-567.
3. Felton, J.S., M.G. Knize, F.T. Hatch, M.J. Tanga, and M.E. Colvin (1999) Heterocyclic amine formation and the impact of structure on their mutagenicity *Cancer Letters* **143**: 127-134.
4. Fellers, R.S., J.C. Sasaki, and M.E. Colvin (2001) Metabolic Oxidation of Carcinogenic Arylamines by P-450 Monooxygenases: Theoretical Support for the One Electron Transfer Mechanism. *Mutation Research* (Submitted).
5. Colvin, M.E., J.C. Sasaki, and N.L. Tran (1999) Chemical Factors in the Action of Phosphoramidic Mustard Alkylating Anticancer Drugs: Roles for Computational Chemistry *Current Pharmaceutical Design* **5**: 645-663.
6. Tran, N.L., M.E. Colvin (2000) The Prediction of Biochemical Acid Dissociation Constants Using First Principles Quantum Chemical Simulations *Journal of Molecular Structure, Theochem* **532**: 127.
7. Winthrop, M.D., S.J. DeNardo, G. Mirick, L. Kroger, K. Lamborn, C. Venclovas, M.E. Colvin, and G. DeNardo (2001) Preclinical Development of Anti-MUC-1 scFv for Targeted Therapy *Clinical Cancer Research* (Submitted)
8. Colvin, M.E., J.N. Quong (2001) DNA-Alkylating Events Associated with Nitrogen Mustard Based Anticancer Drugs and Metabolic Byproduct Acrolein *Advances in DNA Sequence-Specific Agents Vol. 4*; G.B. Jones, ed., JAI Press, Greenwich. (Submitted)
9. Grossman, J.C., M.E. Colvin, N.L. Tran, S.G. Louie, M.L. Cohen (2001) Aromaticity and Hydrogenation Patterns in Highly Strained Fullerenes. *Chemical Physics Letters* (Accepted pending minor revisions).
10. Leininger, M.L., I.M.B. Nielsen, M.E. Colvin, C.L. Janssen (2001) Complete Basis Set MP2 Binding Energies for Stacked Uracil Dimers *Journal of Physical Chemistry A* (Submitted)
11. Wheelock, C.E., M.E. Colvin, I. Uemura, M.M. Olmstead, J.R. Sanborn, Y. Nakagawa, B.D. Hammock (2002) Use of *ab initio* calculations to predict the biological potency of carboxylesterase inhibitors. *Journal of Medicinal Chemistry* (Submitted).
12. Hatch, F.T., F.C. Lightstone, and M.E. Colvin (2000) QSAR of Flavonoids for Inhibition of Heterocyclic Amine Mutagenicity *Environmental and Molecular Mutagenesis*, **35**, 279.

13. Wilson, III, D.M., and D. Barsky. (2001) The major human abasic endonuclease *ape1*. Formation, consequences, and repair of abasic lesions in DNA. [Review article] *Mutation Res.* **485**: 283–307.
14. Barsky, D., N. Foloppe, S. Ahmadi, D.M. Wilson, III, and A. MacKerell, Jr. (2000) New insights into the structure of abasic DNA from molecular dynamics simulations. *Nucleic Acids Res.* **28**(13): 2613–2626.
15. Barsky, D., E.T. Kool, and M. E. Colvin. (1999) Interaction and Solvation Energies of Nonpolar DNA Base Analogues and their Role in Polymerase Insertion Fidelity *Journal of Biomolecular Structure and Dynamics* **16**: 1119-1134.
16. Barsky, D., and M. E. Colvin. (2000) Guanine-cytosine base pairs in parallel-stranded DNA: An ab initio study of the keto-amino wobble pair versus the enol-imino minor tautomer pair. *J. Phys. Chem. A.* **104**: 8570–8576.
17. Nguyen, L. H., D. Barsky, J. P. Erzberger, and D. M. Wilson, III. (2000) Mapping the protein-DNA interface and the metal binding site of the major human apurinic/apyriminic endonuclease. *J. Mol. Biol.* **298**:447–459.
18. Fellers, R.S., D. Barsky, F. Gygi, and M.E. Colvin (1999) An ab initio study of DNA base pair hydrogen bonding: a comparison of plane-wave versus Gaussian-type functional methods *Chemical Physics Letters* **312**: 548-555.
19. Lightstone, F. C., M. C. Prieto, A. K. Singh, M. C. Piqueras, R. M. Whittall, M. S. Knapp, R. Balhorn, and D. C. Roe, (2000) The Identification of Novel Small Molecule Ligands that Bind to Tetanus Toxin. *Chem. Res. Toxicol.*, **13**, 356.
20. Lightstone, F.C., E. Schwegler, R.Q. Hood, F. Gygi, and G. Galli (2001) A first principles simulation of the hydrated magnesium ion *Chemical Physics Letters* **343**: 549-555.
21. Galli, G., Gygi, F., and Catellani, A. (1999) Quantum Mechanical Simulations of Microfracture in a Complex Material. *Physical Review Letters* **82**:3476-3479.
22. Galli, G. (1998) Tight-Binding Molecular Dynamics for Carbon Systems: Fullerenes on Surfaces. *Computational Materials Science* **12**: 242-258.
23. Galli, G. (2000) Large-Scale electronic structure calculations using linear scaling methods, *Phys. Stat. Sol.* **217**: 231-249.
24. Galli, G., R.Hood, A.Hazi and F.Gygi, (2000) Ab-initio simulations of deuterium under pressure, *Phys. Rev. B* **61**: 909.
25. Galli, G., F. Gygi and A. Catellani, (1999) Wetting silicon carbide with nitrogen: a theoretical study, *Phys. Rev. Lett.* **83**, 2006.
26. Haerle, R., G. Galli and A. Baldereschi, (1999) Structural models of amorphous carbon surfaces, *Appl. Phys. Lett.* **75**: 1718.
27. Krishnan, V.V., M Sukumar, L. M. Gierasch, and M. Cosman. (2000) Dynamics of cellular retinoic acid binding protein I (CRABPI) on multiple time scales with implications for ligand binding. *Biochemistry.* **39**(31): 9119–9129.
28. Pizzagalli, L., A. Catellani, G. Galli, F. Gygi and A. Baratoff, (1999) Theoretical study of the (3x2) reconstruction of beta-SiC, *Phys. Rev. B* **60**: R5129.

29. Schwegler, E., and M. Challacombe (1999) Linear scaling computation of the Fock matrix. III. Formation of the exchange matrix with permutational symmetry. *Theoretical Chemistry Accounts*. **104**: 344.
30. Schwegler, E., and M. Challacombe (1999) Linear scaling computation of the Fock matrix. IV. Multipole accelerated formation of the exchange matrix. *Journal of Physical Chemistry* **111**: 6223.
31. Schwegler, E., G. Galli, and F. Gygi (2000) Water under pressure. *Physical Review Letters* **84**:2429.
32. Schwegler, E., G. Galli, F. Gygi, and R.Q. Hood (2001) Dissociation of water under pressure. *Physical Review Letters* **87**:265501
33. White, J., E. Schwegler, G. Galli, and F. Gygi (2000) Solvation of Na<sup>+</sup> in water from first-principles molecular dynamics. *J. Chem. Phys.* **113**:4668-4673.
34. Schwegler, E., G. Galli, and F. Gygi (2001) Conformational dynamics of the dimethyl phosphate anion in solution *Chemical Physics Letters* **342**: 434-440.
35. Sternberg, M., G. Galli and T.Frauenheim, (1999) NOON—a non-orthogonal localized orbital order-N method, *Computer Physics Communications* **118**: 200-1134.
36. Yoo, C.S., H.Cynn, F.Gygi, G. Galli, V.Iota, M.Nicol, S.Carlson, D. Hausermann, C.Mailhiot, (1999) Crystal Structure of Carbon Dioxide at High Pressure: “superhard” polymeric carbon dioxide, *Phys. Rev. Lett.* **83**, 5527..
37. Thelen, M., C. Venclovas, and K. Fidelis. (1999) A Sliding Clamp Model for the Rec1 Family of Cell Cycle Checkpoint Proteins. *Cell*, **96**:769-770.
38. Venclovas, C., K. Ginalski, and K. Fidelis. (1999) Addressing the issue of sequence-to-structure alignments in comparative modeling of CASP3 target proteins. *PROTEINS: Structure, Function, and Genetics*, Suppl. **3**:73-80.
39. Venclovas, C., A. Zemla, K. Fidelis, and J. Moult. (1999) Some Measures of Comparative Performance on the three CASPs. *PROTEINS: Structure, Function, and Genetics*, Suppl. **3**:231-237.
40. Moult, J., Hubbard, T., Fidelis, K., Pedersen, J. (1999) Critical Assessment of Methods of Protein Structure Prediction (CASP): Round III. *PROTEINS: Structure, Function, and Genetics*, Suppl. **3**:2–6.
41. Zemla, A., C. Venclovas, J. Moult, and K. Fidelis. (1999) Processing and Analysis of CASP3 Protein Structure Predictions. *PROTEINS: Structure, Function, and Genetics*, Suppl. **3**:22-29.
42. Zemla, A., Venclovas, C., Fidelis, K., and Rost, B. (1999) A Modified Definition of Sov, a Segment Based Measure for Protein Secondary Structure Prediction Assessment. *PROTEINS: Structure, Function, and Genetics* **34**:220-223.
43. Venclovas, C., Thelen, M.P. (2000) Structure-based predictions of Rad1, Rad9, Hus1 and Rad17 participation in sliding clamp and clamp loading complexes. *Nucleic Acid Res.*, **28**:2481-2493.
44. Fan, W., K. Fidelis, C. Prange, Z. Wang, and G. Lennon. (2000) A New Zinc Ribbon Gene is Cloned From the Human MHC Class Region. *Genomics*, **63**(1): 139-141.

45. M. Cariaso, Folta, P., Wagner, M., Kuczmarski, T., Lennon, G.(1999) IMAGene I: The Clustering and Ranking of IMAGE EST to Known Genes. *Bioinformatics*, 15(12): 965–973.
46. Critchlow T., K. Fidelis, M. Ganesh, R. Musick, T. Slezak (2000) DataFoundry: Information Management for Scientific Data. *IEEE Transactions on Information Technology in Biomedicine*, 4(1): 52-57

### 3.2 Selected Talks, Posters, and Published Abstracts

1. Interaction and Solvation Energies of Nonpolar DNA Base Analogues. Understanding Polymerase Insertion Fidelity Beyond Watson-Crick Pairing. Barsky, D., Kool, E.T., and Colvin, M.E. *Biophys. J.*, (1999) 76(1):A263. Abstract of a talk presented by D. Barsky Feb. 1999 at the Biophysical Society 43rd annual meeting in Baltimore, MD.
2. Large Scale *Ab Initio* Simulations in Material Science. F. Gygi, Invited presentation to the Centennial Meeting of the American Physical Society, to be held in Atlanta, GA, March 22-26, 1999.
3. Prediction of a Sliding Clamp Structure for the Rec1 Family of Cell Cycle Checkpoint Proteins (poster presentation). Venclovas, C., Thelen, M., and Fidelis, K. Pacific Symposium on Biocomputing, in Mauna-Lani, HI, January 4-9, 1999.
4. Computational Gene Discovery on the LLNL ASCI SST using MPGSS, Kuczmarski, T., Invited presentation to JOWOG 34 - Computing, Joint Working Group - LANL, LLNL, SNL and the United Kingdom Atomic Weapons Establishment (AWE), at Los Alamos National Laboratory, February 11, 1999.
5. Overview of the Evaluation Methods Implemented at the Protein Structure Prediction Center. Fidelis, K. Third Meeting on the Critical Assessment of Protein Structure Prediction (CASP3), Asilomar Conference Center, Pacific Grove, CA, December 13-17, 1998.
6. Quantum Simulations of solids and liquids. Galli, G. Invited Physical Chemistry Seminar, Chemistry Department, UCLA, February 8, 1999.
7. Venclovas, C., Petersen, C. and Fidelis, K. Elaboration of distant homology modeling for P68, a *C. parvum* invasion protein. - A poster presented at the Twelfth Symposium of The Protein Society, San Diego, CA, USA. July 25-29, 1998.
8. Venclovas, C., Ginalski, K., and Fidelis, K.-Comparative modeling: selecting alignments by model-building. Invited talk presented by C. Venclovas at the Third Meeting on the Critical Assessment of Techniques for Protein Structure Prediction., Asilomar, CA, USA. December 13 - 17, 1998.
9. Evaluation Methods Implemented at the Livermore Prediction Center (poster presentation). Zemla, A., Venclovas, C., and Fidelis, K., Hubbard, T., and Moulton,

- J. Third Meeting on the Critical Assessment of Protein Structure Prediction (CASP3), Asilomar Conference Center, Pacific Grove, CA, December 13-17, 1998.
10. Evaluation Methods Implemented at the Livermore Prediction Center (poster presentation). Zemla A., Venclovas C., and Fidelis K., Hubbard T., and Moulton J. Third Meeting on the Critical Assessment of Protein Structure Prediction (CASP3), December 13-17, 1998, Asilomar Conference Center, Pacific Grove, CA.
  11. D. Barsky. Learning to Recognize Damaged DNA. Invited lecture for Theoretical Biophysics seminar series, April 26, 1999, Beckman Institute, Univ. of Illinois, Urbana, Illinois.
  12. D. Barsky, N. Foloppe, A. D. MacKerell, D. M. Wilson, and M. E. Colvin. Abasic DNA: New Observations from Molecular Dynamics Simulations. A talk presented by D. Barsky Feb. 15, 2000 at the Biophysical Society 44th annual meeting in New Orleans.
  13. D. Barsky and M. E. Colvin. An ab initio study of guanine-cytosine base pairs in parallel DNA: Twist, wobble, or tautomerize? A talk presented by D. Barsky at the March 2000 American Chemical Society meeting in San Francisco, CA.
  14. D. Barsky. How abasic DNA works its way into the hand of APE1. a poster presented by D. Barsky at the March 2000 Gordon conference on Mutagenesis and Carcinogenesis, Ventura, CA.
  15. M.E. Colvin. Application of Computational Simulations to the Development of Anticancer Therapeutics. Invited lecture, U.C. Davis Cancer Center, March 23, 2000, Davis, CA
  16. M.E. Colvin. Quantum Chemical Studies of DNA Binding Anticancer Drugs and Environmental Mutagens. Invited lecture, Univ. of Maryland Department of Chemistry, February 18, 2000. College Park, MD.
  17. M.E. Colvin, N.L. Tran, J. Grossman, and C.L. Janssen. Quantum Chemical Studies of C36 and Its Hydrogenation Products. Poster, ACS National Meeting, August 1999, New Orleans, LA.
  18. M. Cosman. Structural and dynamic studies of modified DNA/repair protein complexes and an autoantigen implicated in multiple sclerosis. Dept. of Chemistry, North Dakota State University, December 9, 1999, Fargo, ND and Dept. of Chemistry, U. North Dakota, December 10, 1999, Grand Forks, ND.
  19. R.S. Fellers, J. Sasaki and M.E. Colvin. Metabolic oxidation of carcinogenic arylamines by P-450 monooxygenases: Theoretical support for the one electron transfer mechanism. American Chemical Society, Annual Meeting, March 26-30, 2000, San Francisco, CA.
  20. P. Foltá, IMAGEne: The Clustering, Ranking, and Display of I.M.A.G.E. cDNAs. Invited speaker, Structural and Functional Analysis of Human cDNA international workshop, the Kazusa Institute, 3/23/99. Japan
  21. P. Foltá, T. Kuczmarski, C. Prange. IMAGEne I: The Clustering and Ranking of IMAGE EST to Known Genes (poster presentation) Cambridge Health Institute's Bioinformatics and Genomics Workshop, June '99.



22. Giulia Galli , Society of Engineering Science Symposium on Nanomechanics, University of Texas, Oct 25-27, 1999.
23. Giulia Galli, Sanibel Symposium, Florida, Feb 25-March 3, 2000.
24. Giulia Galli, Workshop on Fifteen years of the Car-Parrinello Method, Minneapolis, MN, March 18-19, 2000.
25. Giulia Galli, 4th Canadian Computational Chemistry Workshop, July 2000.
26. Giulia Galli, CECAM workshop on Compressed Hydrogen, Lyon, France, August 2000.
27. Giulia Galli, SIAM conference on Computational Science and Engineering, Washington D.C., September 2000.
28. F. C. Lightstone, J. H. Satcher, S. M. Lane, M. E. Colvin. A Theoretical Study of the Boron-Nitrogen Dative Bond in Aminomethyl phenyl Boronates. American Chemical Society, Annual Meeting, March 26-30, 2000, San Francisco, CA.
29. F. C. Lightstone, E. Schwegler, R. Q. Hood, F. Gygi, and G. Galli. A First Principles Molecular Dynamics Simulation of Magnesium Ion in Water, American Chemical Society, Annual Meeting, April 1-4, 2001, San Diego, CA.
30. J.C. Sasaki, N.L. Tran, J.S. Felton and M.E. Colvin. Resonance Stabilization Effects on the Formation of Electrophilic Intermediates of Aromatic Amine Mutagens American Chemical Society, Annual Meeting, March 26-30, 2000, San Francisco, CA.
31. J.C. Sasaki, N.L. Tran, M.E. Colvin and J.S. Felton. Investigation of Resonance Stabilization Effects in the Bioactivation of Carcinogenic Aromatic Amines. American Association for Cancer Research, Annual Meeting, April 1-5, 2000, San Francisco, CA.
32. J.C. Sasaki, J.S. Felton, M.E. Colvin and K.A. Fidelis. Modeling of the Human P450 1A2 Active Site; Implications for Bioactivation of the Heterocyclic Aromatic Amine Food Mutagens. Environmental Mutagen Society, Annual Meeting, April 8-13, 2000, New Orleans, LA.
33. Eric Schwegler, Giulia Galli, and Francois Gygi, Solvation dynamics of dimethyl phosphate by first principles molecular dynamics, National American Physical Society meeting, March 2000.
34. Eric Schwegler, Giulia Galli, and Francois Gygi, Water under pressure, American Physical Society meeting, March 2000.
35. Eric Schwegler, Giulia Galli, and Francois Gygi, Water under pressure, American Chemical Society meeting, March 2000.
36. Kevin H. Thornton, V. V. Krishnan and Monique Cosman. Solution Structure of the Ligase III BRCT domain. Experimental Nuclear Magnetic Resonance Conference April 9-14, 2000, Asilomar, CA.
37. Venclovas, C. and Thelen, M. Structure-based predictions of Rad1, Rad9, Hus1 and Rad17 participation in novel sliding clamp and clamp-loading complexes. A poster presented at the FASEB Summer Research Conference: Nucleic Acid Enzymes: Structures, Mechanisms and Novel Applications. Saxton's River, VT, USA, June 17-22, 2000.

### 3.3 Project Staff

#### 3.3.1 LLNL Scientists

Computational Biochemistry			
Staff	Michael Colvin	BBRP	
Post-doc	Daniel Barsky*	BBRP	
Post-doc	Felice Lightstone*	BBRP	
Post-doc	Jennifer Sasaki <sup>†</sup>	BBRP	
Post-doc	Regina Monaco	BBRP	
Advanced computational chemistry			
Staff	Giulia Galli	Physics	
Staff	Francois Gygi	Computations	
Post-doc	Eric Schwegler*	Physics	
Post-doc	Raymond Fellers <sup>‡</sup>	Computations	
Protein Fold-Recognition			
Staff	Krzysztof Fidelis	BBRP	
Staff	Ceslovas Venclovas	BBRP	
Staff	Adam Zemla	BBRP	
Genome Informatics			
Staff	Peg Folta	Computations	
Staff	Tom Kuczmariski	Computations	
Structural Biology			
Staff	Monique Cosman	BBRP	
Post-doc	Kevin Thornton <sup>§</sup>	BBRP	

\* Now senior staff scientists at LLNL

<sup>†</sup> Now a staff scientist at Pfizer Pharmaceuticals

<sup>‡</sup> Now a staff scientist at Applied Biosystems

<sup>§</sup> Now a staff scientist at Berlex

### **3.3.2 Students Supported**

We have mentored two students each summer in the computational biology group.

1. Sarah Ahmadia, Univ. Southern California
2. Ngoc Tran, San Francisco State University
3. Dat Nguyen, University of California, Davis
4. Loan Le Bui, Contra Costa College/University of California, Berkeley
5. Katrine Wilson, University of California, Berkeley
6. James Harrison, California State University

### 3.4 Software Developed

Based on the software developed for this SI LDRD, we are filing two Records of Invention with the LLNL Industrial Partnerships and Commercialization office.

- A. **JEEP**: Version 1.5.2: a parallel ab initio molecular dynamics program, for DEC OSF1, IBM AIX, SunOS. (F. Gygi)
- B. **PONX**: A parallel linear scaling algorithm for the computation of the Hartree-Fock exchange matrix. (E. Schwegler)
- C. **NOON**: A non-orthogonal localized orbital order-N method. (G. Galli)
- D. **IMAGENE CANDIDATE GENE PACKAGE**: The set of scripts and utilities that collectively generate the Level 2 clusters and their consensus sequences. (T. Kuczmarski)
- E. **L2CLUST**: A program that merges Level 1 EST clusters into Level 2 EST clusters. (T. Kuczmarski)
- F. **MPGSS**: Massively Parallel Genomic Similarity Search, a software framework that controls the massively parallel execution of an arbitrary genomic similarity search program. (T. Kuczmarski)
- G. **PSG**: A sophisticated job-control script for running Gaussian 94 and 98 quantum chemical jobs on the LC Cluster computers. (D. Barsky)
- H. **LCS** (Longest Continuous Segments): A program to identify all the longest continuous segments of residues in the prediction deviating from the target by not more than specified CA RMS cutoff, established and implemented within the ACE system. (A. Zemla)
- I. **GDT** (Global Distance Test): A program to identify in the prediction the largest set of residues deviating from the target by not more than specified CA distance cutoff, established and implemented to the ACE system. (A. Zemla)
- J. **LDPS** (Local Descriptors of Protein Structure): A set of programs for (a) assignment of local structure descriptors to amino acid residues in proteins; (b) calculate descriptor similarity function; (c) classification and clustering; and (d) graphical display of local structure and structure similarity maps. A new type descriptors have been developed and included in this package. (A. Zemla)
- K. **LGA**: This program is being developed for structure comparative analysis of two selected protein structures or fragments of protein structures.

### 3.5 Outside Grants Funded or Submitted for Continued Funding

1. M.E. Colvin, J. Sasaki, "The Role of Cytochrome P450 Mediated Oxidation in the Differential Mutagenicity of AIA Mutagens", funded by NIH 2/1/02, PI: M.E. Colvin (Program Project 6 in the PO1 program project "Determining the Carcinogenic Significance of Heterocyclic Amines", PI: J. S. Felton)
2. M.E. Colvin: "Computational Chemistry and Simulation", funded by NIH 2/1/02, PI: M.E. Colvin (Core Component D in the PO1 program project "Determining the Carcinogenic Significance of Heterocyclic Amines", PI: J. S. Felton)
3. M.E. Colvin: "Advanced Molecular Simulations of *E. coli* Polymerase III", funded by DOE, 9/1/01, PI: M.E. Colvin.
4. K.A. Fidelis, A. Zemla, C. Venclovas, "Center for Critical Assessment of Protein Structure Prediction", Funded by NIH 8/1/00, PI: K.A. Fidelis
5. M.E. Colvin: "Molecular Modeling", submitted to NIH 2/1/02, PI: M.E. Colvin (Core Component D in the P01 program project "Translational Development of Multimodality Antibody Therapy", PI: G.L. DeNardo, UC Davis Medical Center)
6. F.C. Lightstone, "Toxin and Virulence Factor Structure/Function Determinations and 'Artificial Antibody' Design, submitted to DOE Chemical and Biological Non-proliferation Program 8/99, P.I. R. Balhorn.
7. F.C. Lightstone, "Design and Synthesis of a Prototype Multidentate Reagent for Toxin Detection", submitted to DARPA 12/13/99, PI: R. Mischak.
8. F.C. Lightstone, "Design and Synthesis of a Multidentate Ligand for Protein Inhibition", STTR submitted to NIH with Enzyme Systems Products 4/15/2000, P.I. R. Balhorn.
9. F.C. Lightstone, "Development of High Affinity Synthetic Ligands for Biological Agent Detection", submitted to DOE Chemical and Biological Non-proliferation Program 8/2000, P.I. R. Balhorn.
10. F.C. Lightstone, "Elucidating Biochemical Reactions by Advanced Simulation," submitted to NIH 2/1/2001 (under revision for resubmission 7/1/02), PI: Felice C. Lightstone
11. D. Barsky, "Impact of Variation in Proteins of Base Excision Repair" submitted to NIH (R01) 10/1/99, PI: David M. Wilson, III
12. D. Barsky, "The Role of DNA Structure in Abasic Site Recognition by Ape1" submitted to NIH (R01) 10/1/2000 (under revision for resubmission 3/1/02), PI: D. Barsky
13. M. Cosman, "The effects of altered DNA structure and dynamics on protein recognition", submitted to NIH 11/1/99, PI: M. Cosman
14. M.E. Colvin and F. Lightstone, "Elucidation of the Activation Mechanisms of Phosphoramidate Mustard Anticancer Drugs, Submitted to NIH 6/1/01 (under revision for resubmittal 7/1/02), PI: M.E. Colvin

15. K.A. Fidelis, M.E. Colvin, D.Barsky, "An Integrated Approach to Genome-Scale, High-Throughput Protein Structure Prediction, Modeling and Simulation: Application to DNA Repair Proteins and their Complexes", submitted to DOE/OBER 5/2/00, PI: K.A. Fidelis
16. D. Barsky, "Mechanisms of Substrate Specificity in the DNA Repair activity of APE1": submitted to NIH (R01 supplementary grant) 3/1/00, PI: D. M. Wilson III

## 3.6 Collaborations

### 3.6.1 Internal Collaborations

The biological applications described in this proposal involve collaborations with experimentalists and theorists at the BBRP and other LLNL directorates. We are presently collaborating with the following investigators:

Investigator	Collaborative Project
Dr. David Wilson, BBRP	Human apurinic endonuclease
Dr. Jim Felton, BBRP	Environmental mutagens
Dr. Ken Turteltaub	Spectral properties of DNA mutagens and DNA adducts
Dr. Karen Dingley	Protein adduction reactions by food mutagens
Dr. Rodney Balhorn, BBRP	Bacterial toxins
Dr. Michael Thelen, BBRP	DNA Repair enzymes
Dr. Steven Lane and Dr. Joe Satcher, Lasers	Glucose Sensors
Dr. Jeffrey Grossman LLNL Directors Fellow	C <sub>36</sub> fullerenes and Quantum Monte Carlo studies of weakly bound systems
Dr. Francis Ree, Physics	High-pressure properties of HF
Dr. Jody White, Physics	Simulations of solvated ions

### 3.6.2 External Collaborations

Since a primary goal of this LDRD project is to create a nationally-recognized center of excellence in computational biology, it is essential to establish collaborations outside of the laboratory. To this end we are actively collaborating with several outside

organizations to develop new computational methods and to perform biological applications. The ongoing collaborations are listed in this table.

Ongoing Collaborations:

Investigator	Collaborative Project
Professor Martin Head-Gordon University of California, Davis	Highly electron correlated calculations of DNA base pairing
Professor Eric Kool University of Rochester	DNA nucleotide mimics
Professor Alexander MacKerell, Jr. University of Maryland	Development of molecular dynamics force fields
Professor Bill Fink University of California, Davis	Simulations of antifreeze proteins.
Professor Susan Ludeman Duke University Cancer Center	Alkylating anticancer drugs
Professor Michael Levitt Stanford University	Dynamics simulations of nucleic acids
Professor Richard Williams University of Idaho	Quantum chemical studies of highly strained aromatic compounds
Professor A.P. Huper University of Basel	Ab initio simulations of liquid CO <sub>2</sub>
Dr. Diana Roe Sandia National Laboratories	Protein-ligand Docking
Prof. John Moulton University of Maryland	Critical assessment of protein structure prediction
Professor Robert Latour Clemson University	Protein-surface interactions
Dr. Karl Sirotkin National Center for Biotechnology Information, NIH	Parallelized BLAST for DNA sequence searches
Professor Bill Fink University of California, Davis	Simulations of antifreeze proteins.
Dr. Harry Tom University of California, Riverside	Picosecond correlated dynamics of water, solvated ions, and solvated-proto-DNA molecules using time-domain terahertz spectroscopy and first principle MD simulations