NEW PROTEIN STRUCTURE PREDICTION METHOD USING INTER-RESIDUE DISTANCES AND A THEORETICAL INVESTIGATION OF THE ISOMERIZATION OF AZOBENZENE AND DISUBSTITUTED AZOBENZENES

By

CHRISTINA R. CRECCA

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2008

To my husband Chris

ACKNOWLEDGMENTS

At the completion of this work, I take great pleasure in acknowledging the people who have supported me over the last few years. I gratefully thank and acknowledge my advisor, Prof. Adrian Roitberg, for his continual guidance, support, understanding, and encouragement. I would also like to thank my committee members Dr. Chang, Dr. Cao, Dr. Fanucci, and Dr. Polfer.

During my time at QTP I have made many great friends. Without their support, I do not think I would have made it. I would like to give special thanks to Andrew, Dan, Georgios, Gustavo, Hui, Joey, Josh, Julio, Kelly, Ken, Lena, Lex, Mehrnoosh, Ozlem, Seonah, Tom, and Yilin. I would also like to thank my family, especially my nieces and nephews, Gabe, Savanna, A. J., and Anna.

I would also like to thank Dr. Eric Deumens for his infinite patience and understanding. I apologize to all the computers that were harmed during this work, particularly Arwen and Cobalt.

Our work was supported in part by DOE contract DE-F602-02ER45995 and a University of Florida Alumni Fellowship. Computer resources were provided by the University of Florida High Performance Computing Center as well as the Large Resource Allocations Committee through grant TG-MCA05S010.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Azo	Unsubstituted azobenzene
Azon	4,4’-diaminoazobenzene
Azonco	N-[4-(4-(Acetylamino)phenylazo)phenyl]-acetamide
$AzoNO_2NH_2$	4,4’-nitro-aminoazobenzene
$AzoNO_2NO_2$	4,4’-dinitroazobenzene
CASP	Critical assessment for techniques in protein structure prediction
CASSCF	Complete active space self-consistent field
CHelpG	Charges from electronic potential
DFT	Density functional theory
EPR	Electron paramagnetic resonance
FRET	Fluorescence resonance energy transfer
GDT	Global distance test
LCS	Longest continuous segment
LCS-5	LCS under 5Å
LGA	Local-global alignment
MS	Mass spectrometry
NOEs	Nuclear overhauser effects
NMR	Nuclear magnetic resonance
PDB	Protein databank
RDC	Residual dipolar couplings
RMSD	Root mean square deviations
SDSL	Site-directed spin labeling
TDDFT	Time dependent density functional theory

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

NEW PROTEIN STRUCTURE PREDICTION METHOD USING INTER-RESIDUE DISTANCES AND A THEORETICAL INVESTIGATION OF THE ISOMERIZATION OF AZOBENZENE AND DISUBSTITUTED AZOBENZENES

By

Christina Crecca

May 2008

Chair: Adrian Roitberg
Major: Chemistry

It is often claimed that knowing a protein's structure is important in understanding its function. The experimental structure determination methods presently available can be costly and time-consuming. This dissertation presents an idea for a fast and inexpensive protein structure prediction method that combines modeling with a minimal set of experimental data. Our method involves three steps: (1) building a decoy set (a set of protein-like structures), (2) measuring inter-residue distances, and (3) comparing the measured distances with those calculated in each decoy. We postulate that structures with a small number of similar inter-residue distances will also have similar three-dimensional structure. We further hypothesize that the minimum number of distances needed to determine structure is much less than the total number of inter-residue distances in the protein. To develop our protocol, we searched the decoy set for target proteins whose structures have been solved experimentally but have not been explicitly included in our decoy set. We simulated experimental data by calculating α-carbon distances from the experimentally determined structures of our target proteins.

We have created a large, generalized decoy set using most of the structures in the Protein Data Bank. This decoy set can be used to study any protein composed of 100 residues or less.

Using this decoy set, we attempted to predict structures for several proteins. We also analyzed the RMSD distributions of the decoys using the search proteins as references and found the distributions to be similar for each protein. Of the nearly five thousand C^{α}-C^{α} distances in a 100 residue protein, knowledge of only twenty-five selected distances will usually result in predicting a reliable model.

In the second part of our study, results are presented for a series of azobenzenes which were studied using ab initio methods to determine the substituent effects on the isomerization pathways. Energy barriers were determined from three-dimensional potential energy surfaces of the ground and electronically excited states. In the ground state (S_0), the inversion pathway was found to be preferred. Results show that electron donating substituents increase the isomerization barrier along the inversion pathway, while electron withdrawing substituents decrease it. The inversion pathway of the first excited state (S_1) showed trans→cis barriers with no curve crossing between the S_0 and S_1. In contrast, a conical intersection was found between the ground and first excited states along the rotation pathway for each of the azobenzenes studied. No barriers were found in this pathway and we therefore postulate that after n→π* (S_1←S_0) excitation, the rotation mechanism dominates. Upon π→π* (S_2←S_0) excitation, there may be sufficient energy to open an additional pathway (concerted-inversion) as proposed by Diau. This pathway is only accessible for unsubstituted azobenzene and 4,4-dinitroazobenzene. Because of the S_0 and S_1 curves crossing on the trans side, the concerted inversion channel explains the experimentally observed difference in trans-to-cis quantum yields between S_1 and S_2 excitations. The concerted inversion channel is not available to the remaining azobenzenes and so they must employ the rotation pathway for both n→π* and π→π* excitations.

CHAPTER 1
INTRODUCTION TO PROTEIN STRUCTURE PREDICTION METHODS

We begin with a brief introduction to protein studies, including general information on protein structure, including methods of structure determination, is provided. We also discuss experimental methods like Electron Paramagnetic Resonance (EPR), Fluorescence Resonance Energy Transfer (FRET), Nuclear Overhauser Effects (NOE) from Nuclear Magnetic Resonance (NMR), and chemical cross-linking with mass spectrometry, all of which can be used to measure distances in proteins. Various methods of protein structure prediction are presented followed by a summary of our proposed method.

1.1 Background Information on Proteins

The building blocks of proteins are the twenty naturally occurring α-amino acids. Each amino acid residue has the same fundamental structure (Figure 1-1) containing a carboxyl group, an amino group, and an α-carbon with an R group attached. Amino acids differ in their R-substituent. Linking the carboxyl and amino groups of adjacent residues forms a peptide bond, thereby joining amino acids in a linear fashion. However, some proteins containing cysteine resides can form disulfide bonds which result in the cross-linking (covalent bonding) of nonadjacent residues.

The amino acid sequence of a protein is encoded by the DNA sequence of a gene and is often referred to as the protein's primary structure (Figure 1-2A). The secondary structure is composed of the regularly repeating local conformations generally stabilized by hydrogen bonds. The most typically seen secondary structural elements are the α-helix and the β-sheet (Figure 1-2B). A single protein may have many regions of differing secondary structure and how these regions relate to one another is described by their tertiary structure (Figure 1-2C). Thus, tertiary structure is the overall shape of a single-chain protein. Stabilizing this structure are many non-

local interactions. For example, to minimize their exposure to water, hydrophobic residues retreat to the protein's core. Salt bridges, hydrogen bonds, and disulfide bonds are also formed to help stabilize the structure. Proteins composed of two or more polypeptide chains may exhibit quaternary structure, which refers to the spatial arrangement of these chains (Figure 1-2D).

1.2 Experimental Methods

1.2.1 Structure Determination

To understand the function of a protein on a molecular level, it helps to have some knowledge of its structure.[1,2] In the late 1950s and early 1960s, the first protein structures were determined via X-ray crystallography. The structure of myoglobin was solved by Sir John Cowdery Kendrew, [3] while Max Perutz gave us the structure of hemoglobin.[4] So important were their discoveries that Perutz and Kendrew were awarded the Nobel Prize in 1962.

1.2.1.1 X-ray crystallography

There are three steps in determining the structure of a protein from X-ray crystallography. First, a suitable crystal must be grown, which is often the most difficult and time consuming step.[5] The quality of the crystal is hard to assess until the diffraction pattern is obtained, but ideally it will be pure and free from imperfections, regular in structure, and suitable in size.

Next, a beam of X-rays is fired at the crystal thereby producing a regular array of reflections (a pattern of spots) that can be seen on a screen behind the crystal. One of the advantages of using crystals is that they are periodic and therefore composed of many repeating unit cells. Constructive interference due to this periodicity serves to amplify the weak scattering of the individual unit cells into a more powerful, coherent reflection. The relative intensities of the reflections are important in determining the arrangement of the molecules in the crystal. They can be recorded using an area detector, charge-coupled device image sensor, or photographic film. After the intensity of each spot is recorded, the crystal is rotated slightly to

produce another set of reflections, whose intensities are also recorded. One image of spots does not provide enough information to determine the molecular structure of the crystal because it only represents a small piece of the full Fourier transform. Therefore, this process is continued for more than a 180° rotation of the crystal. One should also change the axis of rotation at least once to avoid a blind spot in reciprocal space near the rotation axis. Often, several sets of diffraction patterns may be collected.

Finally, the data can be analyzed. The reflections from all the recordings must be indexed by identifying the dimensions of the unit cell. An autoindexing algorithm[6] is usually employed to determine which image peak corresponds to which position in reciprocal space. All the images of all the reflections are converted into a single file containing the Miller index of each reflection along with its intensity. Hundreds of separate images of the crystal are taken at different orientations leading to many symmetry-related reflections being recorded multiple times. One must find which peaks appear in two or more images.

Because the diffraction data is a reciprocal space representation of the crystal lattice, the location of each spot is determined by the size and shape of the unit cell as well as the symmetry in the crystal. The intensity of each spot is proportional to the square root of the structure factor amplitude, which is a complex number that contains information relating to the amplitude and phase of a wave. Both the amplitude and phase must be known in order to generate an electron density map, which is then used to build a starting model of the protein structure. A potential problem in X-ray crystallography is that the phase cannot be directly recorded during the experiment.

The phase problem can be overcome in several ways. (1) *Ab Initio* phasing (direct method) [7,8] is often applied to small proteins by exploiting known phase relationships between specific

groups of reflections to determine the needed phase information. (2) In molecular replacement,[9] the structure of a related protein can be used as a model to determine the orientation and position of the molecules in the unit cell. Phase information is then obtained and used to build an electron density map. (3) Anomalous scattering (multiwavelength anomalous dispersion MAD) [10] involves incorporating anomalously scattering atoms like Selenium into the protein; the scattering is changed in a known way. The position of the anomalously diffracting atoms can be found easily thereby providing their initial phases. (4) Similar to MAD phasing, heavy atoms can be incorporated into the protein. The changes in the scattering amplitudes can be used to determine the phases. MAD phasing with selenomethionine is now more commonly used than heavy atom replacement.[10]

Initial models can be built after the initial phases have been obtained. The models are then used to refine the phases, which are used to further improve the models. The B-factor is an estimate of the thermal motion of the atom and must be included in the phase refinement process. The R-factor is a measure of the agreement between the crystallographic model and the diffraction data, and depends highly on the resolution of the data. It should also be noted that it is not always possible to see every atom in the molecule because the electron density is an average over all molecules in the crystal. Sometimes atoms exist in several conformations causing their electron density to appear smeared. On the other hand, they may appear multiple times in an electron density map.

In summary, the magnitude of the Fourier transform of electron density is found from the multiple recorded intensities of the reflections. The full Fourier transform of the electron density comes from combining the phases and magnitudes. The electron density is converted into the

arrangement of atoms in the crystal and then the determined crystal structure is stored in a publicly accessible database, like the Protein Data Bank (PDB).

1.2.1.2 Nuclear magnetic resonance (NMR)

Structure determination using NMR can be performed in five steps. (1) Prepare the protein solution; unlike X-ray crystallography, Nuclear Magnetic Resonance (NMR) can be performed on samples either in solution or in solid state. (2) Take the NMR measurements. (3) Assign the NMR signals to individual atoms in the molecule. (4) Identify conformational constraints, like distances between hydrogen atoms. (5) Calculate the 3D structure based on experimental constraints.[11]

When a nucleus is placed in a magnetic field, it can exist in one of a small number of allowed orientations (states) with different energy. The nucleus of a hydrogen atom has only two allowed orientations; the magnetic moment of the nuclei can either align parallel to the external magnetic field or point in the opposite direction. Some nuclei will be oriented parallel and others anti-parallel giving rise to a small polarization of nuclear spins and thus a net macroscopic magnetization, which can be manipulated with the appropriate electromagnetic waves. Quantum mechanically speaking, the external magnetic field causes the ground state energy level of the nucleus to split into two spin states (for nuclei with $S= ½$) that have an energy difference of $h\nu$. This energy gap can be measured by applying electromagnetic radiation of frequency ν (usually in the radio-frequency range) causing the nuclei to be excited from the lower energy level to the upper one. This process is classically described as flipping the spin of the nucleus between their two spin states, spin up and spin down. The frequency is typically applied in several pulses, each of which is a few microseconds long, causing the spins of the nuclei to flip. An NMR signal is obtained after perturbing the equilibrium spin states. The signal decays as the system returns to

its equilibrium state. The signal is also called the free induction decay (FID), which is essentially the sum of decaying cosine waves whose frequencies correspond to the resonance frequencies of the nuclei. A Fourier transform of this data yields the NMR frequency spectrum.[12]

Different types of nuclei have vastly different resonance frequencies. Protons (^{1}H), for example, resonate at a frequency four times higher than carbon nuclei (^{13}C) and ten times that of the nitrogen nuclei (^{15}N). Much smaller resonance frequency differences occur between nuclei of the same type. Such variations or chemical shifts are due to the interactions between the nuclei and the surrounding electrons affecting the local magnetic field experienced by a particular nucleus which in turn alters its resonance frequency. It is the chemical shift that allows us to assign protons to different classes. For example, we can distinguish between amide protons and those on methyl groups. The chemical shift is very sensitive to many structural, electronic, magnetic, and dynamic variables and contains a lot of information on the state of the system of interest.

The most important feature of NMR spectroscopy is that individual nuclei interact with the small magnetic fields generated by the spins of the nuclei nearby. The different nuclei can be correlated with one another in the molecule by using these spin-spin interactions. The nuclei interactions are either through-space or through-bond. The through-space interactions are the basis for the nuclear Overhauser effect (NOE), which allows for distance measurements between hydrogen nuclei. The through-bond interactions are called spin-spin coupling or J coupling. Both of these correlations form the basis for the analysis of protein spectra.

To determine the structure of proteins, multidimensional NMR must be used because it provides spectra with improved resolution as well as more easily analyzed correlations. Two-

dimensional NMR was primarily developed by Kurt Wurthrich, who shared the Nobel Prize in chemistry in 2002.

There are four consecutive time periods in multidimensional NMR: excitation, evolution, mixing, and detection. In the excitation period, the nuclear spins are prepared in the desired state. The chemical shifts are then observed during the evolution period τ_1. The spins are correlated with each other during the mixing period and the chemical shift information of one nucleus is transferred to another nucleus whose frequency is measured during τ_2, the detection period. Several experiments are run with successively incremented lengths of τ_1. From this information, a two-dimensional data set is obtained, from which a data matrix $S(\tau_1, \tau_2)$ is generated. The frequency spectrum, $S(\omega_1, \omega_2)$, is obtained from a Fourier transform of $S(\tau_1, \tau_2)$. If two nuclei interact during the mixing time, the interaction will appear as a cross peak in the resonance spectrum at a position corresponding to the resonance frequencies of the two nuclei.

Larger proteins are generally labeled with ^{15}N and ^{13}C, but the preferred nucleus for detection is hydrogen because it is the most sensitive. During the evolution period, the other nuclei, ^{15}N and ^{13}C, can be measured and their information is transferred to the protons for detection. The chemical shift is sensitive to the environment of a nucleus. Thus, multiple copies of the same amino acid in a protein can be distinguished due to the conformation dependent chemical shift. The ^{1}H, ^{15}N, and ^{13}C chemical shifts are known for many 3D NMR structures of proteins and can be used in empirical and semi-empirical correlations with structural parameters.

To assign the spectra, knowledge of the protein sequence is necessary. For large [^{15}N, ^{13}C]-labeled proteins, through-bond correlations across the peptide-bond between sequential amino acids can be used to assign the spectrum. Distance information and/or dihedral angles must be derived from NMR data to calculate the protein's structure. Basic information about

protein structure such as, amino acid sequence, bond lengths, bond angles, chiralities, planar groups, and steric repulsion between non-bonded atom pairs, is used in conjunction with NMR data to do so. The crucial information comes from NOE distance measurements but supplementary dihedral angle constraints can come from through-bond correlations. Chemical shifts can also indicate the type of secondary structure that is present and through-bond interactions can detect hydrogen-bonds. When NOEs are not prevalent, residual dipolar couplings (RDCs) can be used. RDCs are related to the orientation of N-H and C-H internuclear vectors relative to the molecular frame. Using more input constraints in the structure calculation gives rise to higher quality structures.

The final step in this process is to calculate the structure. First a low resolution structure is derived from an unambiguous subset of NOE data. Many computer programs are available for this process and they are divided into two main groups:[13] ones that use inter-atomic distances[14] like DISGEO and DISMAN[15,16] and ones that use torsional bond angles, like DYANA[17] and DIANA[18]. The final result for each type of method is the Cartesian coordinates of the family of structures which satisfy the set of NMR constraints. The experimental constraints do not specify just one unique structure, instead they describe a range of possible values. Also, some distances cannot be determined. Due to these restrictions, an ensemble of structures satisfying the constraints is typically generated by repeating the structure calculation several times. The best ensemble samples all the conformational space the constraints allow.

Restrained molecular dynamics[14,15] and distance geometry[18,19] are two of the most common approaches used in structure generation. Distance geometry is often used to generate an initial structure for molecular dynamics. Using the distance constraints, an error function is minimized, which depends on the sum of all differences between the distance constraint and the actual

distance. In restrained molecular dynamics, energy terms based on the NMR-derived constraints are added to the classical molecular dynamics force fields. Usually, a combination of distance geometry and molecular dynamics is used to calculate the structure of a protein.

1.2.2 Distance Measurements

1.2.2.1 Nuclear overhauser effects NOE) from NMR

Nuclear Overhauser effects are through-space correlations between nearby hydrogen atoms in the protein. Unlike J couplings, the nuclei involved in the NOE can be separated greatly in the protein sequence as long as they are close to each other in space. The NOE results from the transfer of magnetization between spins interacting through their dipoles. The intensity of the NOE is approximately proportional to r^{-6}, where r is the distance between the two interacting nuclei. Because of the dependence on the inverse power of six of the distance, the intensity falls off very quickly with increasing distance. As such, NOEs are only observed for small distances, between 2 – 5 Å[12]. The lower bound is essentially the sum of two hydrogen atomic radii. The NOE distances are also classified as strong, medium, or weak depending on their intensity. The distance ranges corresponding to each type of NOE are usually defined as less than 2.5 Å for strong, 2.5 – 3.5 Å for medium, and greater than 3.5 Å for weak.[20]

1.2.2.2 Electron paramagnetic resonance (EPR)

Electron paramagnetic resonance (EPR) is a magnetic resonance technique used to investigate systems which possess unpaired electrons. This technique has been used in studies of metal centers in metalloproteins as well as reaction-intermediates via spin trapping. The introduction of site-directed spin labeling (SDSL)[21] has allowed EPR to be applied to studies of proteins that do not contain metal atoms. In SDSL, a cysteine residue is introduced into a specific location in the protein sequence, which can then form a disulfide bond with a thiol-specific nitroxide spin label, like methanethiosulfonate (MTSL).

The basic principles of EPR are similar to NMR, except EPR deals with the magnetic moment of an unpaired, free electron instead of nuclear spins. An oscillating magnetic field induces transitions between the two spin states. Energy is absorbed during the transition and the first order derivative of the absorption spectrum is generally recorded.

To measure distances in proteins using EPR, two spin labels must be inserted. If the two spin probes are close enough to each other they will experience dipole-dipole coupling proportional to r^{-3}, the inverse cube of the distance between them.[22,23] EPR can measure a wide range of distances, from 8 – 20 Å for continuous wave techniques and 18 – 80 Å for pulsed techniques.[24]

In continuous wave methods, distance information can be extracted by analyzing line broadening caused by dipolar interactions. To obtain the dipolar interaction, three types of labeled samples are need: a protein labeled in site A, a protein labeled in site B, and a protein labeled in both sites A and B. A spectrum with two spin labels is assumed to be the convolution of the dipolar broadening function and the monoradical spectrum.[25] To separate the dipolar spectrum, a Fourier deconvolution method is used to subtract monoradical contaminants. This yields a pake distribution from which distance information can be obtained.

In pulsed EPR techniques, like double-electron electron resonance (DEER), distance information is obtained by modulating a spin echo at the frequency of the dipolar interaction. Analysis of spin-echo amplitude oscillations reveals such distance information.

There are many benefits to using EPR. The sample preparation is fairly easy compared to other methods. Low temperatures are required but only one type of spin label is needed. EPR often has higher precision than techniques like Fluorescence Resonance Energy Transfer (FRET), which will be discussed in section 1.2.2.3. EPR can be performed with lower sample

volumes and concentrations than NMR and there are no molecular size limitations. EPR can be used to measure medium (5 – 25 Å) to long range distances (25 – 80 Å).

One disadvantage of EPR is that the dynamics of the spin labels are unknown and highly system dependent. To overcome this problem, molecular dynamics simulations can be used to predict the orientation of the labels and their effects on the distance distributions. The experimentally determined distances are those between the spin labels, not the α-carbons, but by predicting the spin probe locations, one may be able to derive inter-residue distance information.

1.2.2.3 Fluorescence resonance energy transfer (FRET)

In fluorescence resonance energy transfer, energy is transferred from an excited state donor (D) to a ground state acceptor (A). Long-range dipole-dipole interactions between the donor and the acceptor cause the energy transfer, which occurs without the appearance of a photon. The rate of energy transfer depends on four things: (1) the spectral overlap between the emission spectrum of the donor and the absorption spectrum of the acceptor, (2) the quantum yield of the donor, (3) the relative orientation of the donor and acceptor transition dipoles, and (4) the distance between the donor and acceptor. In measuring inter-residue distances in proteins, it is this distance dependence of FRET which is exploited.

Proteins can be covalently labeled with a donor, typically a tryptophan, and an acceptor molecule. The distance between the labels is inferred from the efficiency of energy transfer, which can be determined from steady-state measurements of the extent of donor quenching due to the presence of the acceptor. Because such distances can be measured, FRET is often described as a "spectroscopic ruler".[26]

The distance over which energy transfer occurs is similar to the dimensions of proteins. The Förster distance (R_0) is the distance at which FRET is fifty percent efficient and is usually

between 20 – 90 Å, depending on the specific donor and acceptor pair. Transfer fficiency can be expressed in terms of distances (Equation 1-1), decay times (Equation 1-3) and intensities (Equation 1-4). In equation 1-1, R_0 is the Förster distance while r is the distance between the donor and acceptor. The rate of FRET strongly depends on the distance and is inversely proportional to r^6. Any phenomenon that changes the donor-acceptor distance will also cause a change in the transfer rate, which allows, for example, the study of conformational changes in proteins. Energy transfer is assumed to occur if the distance between the donor and acceptor is near the Förster distance and there is enough spectral overlap.

$$E = \frac{R_0^6}{R_0^6 + r^6} \qquad (1\text{-}1)$$

The Förster distance can be calculated from spectral properties of the donor and acceptor molecules as in Equation 1-2. The quantum yield of the donor in the absence of the acceptor is Φ_D, $J(\lambda)$ is the spectral overlap, n is the refraction index of the medium (generally taken to be 1.4 for biomolecules in aqueous solution), A is a constant, and κ^2 is the orientation factor, which describes the relative orientation of the donor and acceptor transition dipoles. A dynamic random average for κ^2 is assumed to be 2/3. The transfer efficiency is measured experimentally via fluorescence intensity and calculated using Equation 1-3 or Equation 1-4. The subscript DA indicates the lifetime (τ_{DA}) or intensity (I_{DA}) of the donor in presence of acceptor while the D subscript represents is the lifetime (τ_D) or intensity (I_D) of the donor in absence of the acceptor.

$$R_o^6 = \frac{A\kappa^2\Phi_D \; J(\lambda)}{n^4} \qquad (1\text{-}2)$$

$$E = 1 - \frac{\tau_{DA}}{\tau_D} \qquad (1\text{-}3)$$

$$E = 1 - \frac{I_{DA}}{I_D} \qquad (1\text{-}4)$$

In summary, FRET can be used to measure long-range distances (20 – 90 Å) in proteins. The actual distances measured are between the donor and acceptor molecules. Distances between two α-carbons, however can be inferred from this data by estimating the position of the labels using molecular dynamics. Some degree of uncertainty is introduced into the measurement when deriving such distances from the experimental data.

1.2.2.4 Chemical cross-linking with mass spectrometry

Chemical cross links are used to connect two polymer chains through covalent bonds. In biochemistry, they are often employed in the study of protein structure, function, and interactions with other proteins. Cross-linkers bind to surface amino acid residues that are near one another in space. This helps to stabilize otherwise weak or transient inter-residue interactions so they can be analyzed. Imidoester cross-linker dimethyl suberimidate[27] and the *N*-hydroxylsuccinamide-ester cross-linker BS^3 (bis(sulfosuccinimidyl) suberate)[28] can both be used as cross-linkers in protein studies. In these cases, lysine's amino group undergoes a nucleophilic attack resulting in a covalent bond between the lysine and the cross-linker. The carbodiimide cross-linker EDC, 1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide, converts carboxyl groups into amine-reactive isourea intermediates that can bind to available primary amines, including lysine residues. The cross-linkers have a known end-to-end distance, which can be taken as the maximum distance

between the two linked residues, as the linkers are usually flexible and can fold over on themselves.[29]

Mass spectrometry (MS) is an analytical tool that has many uses including protein characterization. In general, there are four steps in MS: (1) ionize the sample, (2) separate ions of differing masses, (3) detect the number of ions having each mass produced, and (4) collect and analyze the data. To characterize proteins, two methods can be used. In the "top-down" approach, the intact protein is ionized by electrospray ionization or matrix-assisted laser desorption/ionization and then it is run through a mass analyzer. In the "bottom-up" method, protease enzymes, like trypsin, are used to digest the proteins into smaller peptides, which are then introduced into the mass spectrometer. The identity of the peptide is found through peptide mass fingerprinting or tandem mass spectrometry. It is easy to identify cross-linked residues because the enzymes will not cleave residues containing such cross-linkers.

1.3 Methods of Structure Prediction

Many methods of protein structure prediction are available. The ultimate goal in protein modeling is to predict the structure of a protein using its amino acid sequence. Ideally, the predicted structure will be comparable in accuracy to an experimental structure but the method of deriving the structure will be much faster compared to experiment. Protein modeling is especially useful for proteins that cannot be crystallized for X-ray diffraction or those that are too large to be studied via NMR. The recent interest in genome projects has given rise to enormous amounts of amino acid sequence information, which continues to grow much faster than protein structure data.[30-32] In an attempt to keep up with the demand for protein structures, many scientists are turning to structure prediction methods. We will discuss homology modeling, fold recognition methods, and ab initio methods.

1.3.1 Homology Modeling

The simplest protein structure prediction method is sequence homology, [33,34] which determines the degree of similarity between proteins—one with a known structure and one without**.** There are two basic premises of homology modeling: (1) the protein's structure is determined by its amino acid sequence[35] and (2) over millions of years, a protein's structure is less likely to change than its sequence.[36-38] If two proteins have sequence homology greater than 30%, they are believed to have essentially the same structure[38]. In general, homology modeling can be broken down into seven steps. They are: template recognition and initial alignment, alignment correction, backbone generation, loop modeling, side-chain modeling, model optimization, and model validation.

The first step in homology modeling is to find a template and align the sequences. Sequence alignment programs like BLAST[39] or FASTA[40] compare the target sequence to all sequences that have a known structure and are in the PDB by using two matrices. The first is a residue exchange matrix, which characterizes the probability that two of the twenty amino acids should be aligned. The axes of this matrix are simply the 20 amino acids. The highest values are found along the diagonal representing conserved residues. Exchanges between residues with similar properties, like phenylalanine → tyrosine, have higher scores than those exchanges between very dissimilar residues. In the alignment matrix, the axes correspond to the two sequences being aligned and the elements are the values of the exchange matrix for a particular pair of residues. To find the optimal alignment, the best path through the matrix is found taking care not to use any residue twice. Gaps can be inserted to improve the alignment, but the alignment algorithm will subtract a gap opening penalty. The end result of a BLAST search is a list of hits, which are the modeling template and their corresponding alignments.

In the second step, the alignment can be corrected using programs like CLUSTALW[41] to perform multiple sequence alignments, which use the sequences of other homologous proteins. Such methods are useful for regions of low sequence identity in the original alignment. Multiple sequence alignments can also generate position-specific scoring matrices called profiles, which indicate the residues most likely to be buried in the hydrophobic core and which are on the surface based on the most frequently seen residue exchanges.

The backbone is generated in step 3. The coordinates of the template residues appearing in the alignment are copied. If the template residues are the same as the target, all coordinates of the residue are copied. If not, only the backbone coordinates (N, C^{α}, C, and O) are copied. Multiple template modeling, as performed by programs like Swiss-Model[42], is useful when one template is found to contain errors.

After alignment, there are often gaps from insertions and deletions which change the conformation of the backbone. These changes usually do not occur in regular secondary structural elements, but rather in loop regions and turns. Even without the insertions and deletions, loop conformations often differ between the template and the target. There are two common methods used for loop modeling, knowledge based and energy based. In the first method, the loop conformation is copied from a known loop in the PDB with endpoints that are the same as the residues between which the loop must be inserted. Most major modeling programs can do this.[42-46] The energy based method determines the best loop conformation by minimizing an energy function using Monte Carlo[47] or Molecular dynamics.[48]

The next step is to model the side chains. Usually, this is done using libraries of common rotamers (different conformations) that have been extracted from high resolution X-ray structures. Many such libraries exist.[49-51] Each rotamer is positioned and energy functions are

used to score them. If the residue is conserved, it is easier to copy the coordinates of the entire residues instead of copying the backbone and modeling the side chain. Also, certain conformations of the backbone may prefer certain rotamers, which helps to minimize the search space as the position of one affects the position of its neighbors. Residues in the hydrophobic core generally adopt only one conformation, whereas the more flexible side chains on the surface of proteins may adopt several conformations.

The model optimization step is actually an iteration of rotamer prediction and energy minimization steps. The rotamers are predicted, which changes the backbone, then the rotamers for the new backbone are predicted and so on until convergence is achieved. Molecular dynamics is the preferred method of energy minimization. Some methods restrain the atom positions and/or only employ a few hundred steps of energy minimization. Better, more accurate force fields will help improve model optimization.

The final step in homology modeling is to validate the model. The amount of error can depend on the sequence identity between the target and template. Poor sequence identity (< 25 %) can often lead to very large errors. Such errors can be estimated by calculating the model's energy based on a force field, which checks to see if bond lengths and angles are within normal limits. It is not possible to discern if the protein is folded correctly using force fields alone as misfolded, yet well-minimized models can usually give the same energy as the target structure.[52] Determining normality indices is an additional way to estimate errors. These indices describe how similar certain characteristics are between the model and real structures. They will check properties like the distribution of polar and nonpolar residues in the interior and on the surface of the protein as well as radial distribution functions that can distinguish between good and bad contacts.

In general, sequence homology is quick and computationally inexpensive. A drawback of this method, however, is its inability to detect structural similarities existing between two proteins with very different sequences. Unfortunately, in the protein world such occurrences are still quite common.[53-56] For example, mammalian glycogen phosphorylase and DNA glucosyltransferase have similar shapes but differ greatly in sequence.[54] For these proteins, other methods, like threading and ab initio folding, must be used.

1.3.2 Fold Recognition Methods (Threading)

The basic idea behind fold recognition methods (also called threading or the inverse folding problem), is to determine which of the known protein folds will be most similar to an unknown fold of a new protein knowing only its amino acid sequence. In nature, often two seemingly unrelated proteins may adopt similar folds. It is therefore important for a program to detect structural similarities between proteins with vastly different sequences. Some of these occurrences are the result of divergent evolution; the two proteins are related, but our current sequence analysis methods are not sensitive enough to detect such distant homologies. Convergent evolution, on the other hand, may explain how similar structures can result from proteins having common functional requirements, like binding the same class of substrates. Because only a small number of folds have been found in nature thus far, proteins may have very limited folding space giving rise to similarities between unrelated proteins. This explanation is generally reserved only for single domain proteins.[57]

The first two explanations show that proteins with similar structures sometimes also have similar functions; fold recognition, therefore, might be used as a function prediction tool as well. Usually, the active site, identity of cofactors, and general features of the reaction being catalyzed are highly conserved for enzymes with similar folds.[1] Essentially, for such proteins, function is conserved evolutionarily.

In sequence-based fold recognition, one must first recognize similarities based on sequence and then construct a detailed alignment, which is a residue-by-residue equivalence table between the two proteins. The same methods of sequence alignment were discussed in our section on homology modeling. Fold recognition methods also use position-specific mutation rules derived from the multiple sequence alignment of a homologous family to find distant homologies, even between proteins with less than 25% sequence similarity.

Energy-based fold recognition methods are similar to grid search minimizations. The calculated energy at each grid point is based on known protein structures. This method is also called threading[58,59] because the target sequence is being threaded through or forced to adopt the structure of another protein. Several threading algorithms have been developed, but all follow the same paradigm of sequence alignment, template identification, and alignment building. The same limitations apply to threading as sequence-based fold recognition. If no correct structure exists in the structural database being used, no good models will be built.

Many threading algorithms have been developed over the years. An intuitive approach would be to use a technique that incorporates nonlocal scoring functions. Many approximations are needed to minimize the space of possible alignments. One of the first most successful methods was the Threader algorithm[60]. It used two-level dynamic programming to optimize interaction partners for each pair of aligned residues. Only the strongest interacting residues were considered in this method, which helped reduce its computational cost.

Threading algorithms generally differ in three areas. The first is in their protein model and interaction descriptions. Simplifying the three-dimensional protein structure is one way to speed up energy calculations. Side chains can be simplified by describing them as interaction points, which can be located at C^{α} or C^{β} positions or can encompass the whole side chain. Also, the

energy calculation may only include certain parts of the protein and the interaction energy may or may not be distance dependent. Different algorithms also have various empirical energy parameterizations. Finally, threading methods differ in their alignment algorithms. The threading energy is a nonlocal function based on the alignment between the template structure and the prediction target sequence.

1.3.3 *Ab Initio* Methods

Predicting a protein's native conformation solely from its sequence of amino acids is the basis of *Ab initio* structure prediction. In the early 1970's, Anfinsen suggested that a protein's native conformation corresponds to a global free energy minimum for their sequences, which is commonly referred to as the "thermodynamic hypothesis". He also showed that information needed for a protein to fold is contained in its amino acid sequence[61,62]. It seems logical to assume that given a perfect energy function and the proper computational tools, the native structure can be found. In reality, two problems hinder this method. The conformational space to be searched is huge, while the molecular potentials have limited accuracy.

To reduce the effects of these problems, many methods use reduced representations, simplified potentials, and coarse search strategies[63-66]. In *ab initio* folding, representations of the polypeptide chain are usually simplified in some way. Implicit solvation models are preferred over explicit water molecules. United atom representations are used, where the non-polar hydrogens are drawn into the base of the heavy atoms to which they are bonded. Using the limited set of conformations for each side chain that is most prevalent in the PDB (found in rotamer libraries[51]) can reduce computational cost without loss of predictability. Side chains are also sometimes replaced by locating the side-chain properties at the C^{α} or the centroid of the side

chain. This essentially averages out the side chain degrees of freedom, which speeds up the calculation but also decreases the specificity.

To reduce the size of conformational space to be searched, one can limit the available backbone conformations. Certain local structures prefer certain torsion angle pairs[67-69]. Torsion angles can also be restricted to discrete, commonly seen values by using a small set of phi-psi pairs[70], by choosing pairs from an ideal set based on predicted secondary structure, or by using fragments from known proteins[63,71-73].

There are two types of potentials used to evaluate the free energy of proteins, molecular mechanics potentials and scoring functions. Both classes must be able to properly represent the forces that determine protein conformation. Such forces include solvation, electrostatic interactions between hydrogen bonds as well as ion pairs, covalent bonds, bond angles, dihedral angles, and van der Waals interactions.

For molecular mechanics, the forces needed to determine protein conformation are modeled by using physical based functional forms that have been parameterized from small molecules or quantum mechanical (QM) calculations. Coulomb's law is used to model electrostatic interactions using QM calculations to derive partial charges, while a standard 6-12 potential is usually used to describe van der Waals interactions.

Scoring functions (protein structure-derived potentials), on the other hand, are empirically derived from experimental structures.[74,75] A functional form of the potential is usually not specified but rather the logarithm of probability distribution functions are used to find pseudoenergies. These functions are especially useful when dealing with reduced complexity models as they can represent interactions between side chain centroids after averaging over all possible positions of the non-present atoms.

Molecular dynamics is usually too expensive for the *de novo* generation of protein models using full atom representation. This method, however, has had some success with very small proteins, like the Trp-cage protein.[76,77] Conformational searching is quicker when a coarse sampling of the energy landscape is performed. Methods that take this approach include Metropolis Monte Carlo simulated annealing,[63] simulated tempering,[78] evolutionary algorithms,[72] and genetic algorithms.[79] These methods generally allow for large perturbations in structure in a single move. Because the final structure of a single search may end up being trapped in a local minimum instead of the global minimum, several conformational searches are performed to generate an ensemble of possible structures. Choosing the most native-like structures from the ensemble is difficult and many techniques have been developed to do so.[80-82] As potential functions are improved, identifying the most accurate models will become easier as they will have the lowest free energy. Perhaps the best energy functions for discriminating amongst the possible structures will be ones that combine molecular mechanics potentials with those derived from protein databases.

Two fields in which *ab initio* protein folding might be of great use include genome functional annotation and structural genomics. Many open reading frames have no sequence homology with proteins of known structure and/or function, but links between such proteins may be detected through *ab initio* folding. Structural similarities can be detected by comparing the predicted structures to those in the PDB using a structure-structure comparison tool.[83] One could also look for conserved geometric motifs in these structures.[84] Finally, the predicted structures can be used to make matches to sequence-based motif libraries more sensitive and reliable.

Ab initio structure prediction can be used in structural genomics initiatives as a guide for experimentalists by finding proteins most likely to contain novel folds. A hybrid of *ab initio*

prediction and homology modeling can also be used, if a homology models contains a gap, then it can be filled in by *ab initio* prediction. Combined with a small amount of experimental data, *ab initio* methods can be used in rapid structure determination for proteins whose structures cannot be determined via X-ray or NMR data, like membrane-bound proteins.[85]

1.3.3.1 Rosetta

A specific example of an ab initio folding algorithm is Rosetta (http://www.bakerlab.org/), which is one of the best prediction methods available today. Prediction methods are tested and compared via the Critical Assessment of protein Structure Prediction (CASP) experiments (discussed further in section 1.4). In past CASP experiments,[47,86,87] Rosetta has generated some the top scoring predictions. There are several variants of Rosetta, all of which use sequence information and an energy function to generate protein-like structures. Rosetta has been employed in structure determination using limited experimental constraints,[88,89] *de novo* protein design,[90,91] protein-protein docking,[92] and loop modeling.[93] All methods involve generating a fragment library, piecing the fragments together, clustering the structures by pairwise C^{α} root mean square deviation (RMSD) values, and ranking the representative cluster centers. Incorporating experimental data into the Rosetta method has been successful. RosettaNMR, for example, uses NMR constraints like residual dipolar coupling (RDCs), Nuclear overhauser effects (NOEs), and unassigned chemical shifts (CSs) to restrain certain bond distances and angles to improve the quality of predicted structures.

After the decoy set is generated, some decoys are eliminated using two filters. The contact order filter removes the decoys with low contact order (overly local structures) compared to a test set of proteins. The strand arrangement filter eliminates structures with non paired β-strands

and other nonprotein-like structures. Finally, the decoys are clustered. A representative model from each cluster is chosen and ranked by the size of the cluster it represents.

Discriminating among the decoys is still problematic. Clustering is not always the best option, as the best structures may not be in the most populated cluster. We will present a method involving the generation of decoy sets using Rosetta and discriminating among the decoys using inter-residue distance measurements.

1.3.3.2 Databases to test scoring functions

In developing proper energy functions for use in protein structure prediction, it is important to test the function on a set of computer-generated conformations called decoys to see if the functions can distinguish between the native and non-native-like conformations.[94] Samudrala and Levitt developed many sets of such conformations (the database called Decoys 'R' us located at http://dd.stanford.edu) and made them available to the public to aid in the improvement of scoring functions. The decoys were generated with the intention of fooling the scoring functions; they have similar characteristics of native proteins, but they are not necessarily correct. Decoys have been developed from the following types of methods: molecular dynamics trajectories,[95] crystal structures,[96] conformations with different loop regions,[82] threading of the amino acid sequence onto very different folds,[97] and discrete-state models.[80]

Similar websites have been created to test energy functions for general protein structure prediction (http://prostar.carb.nist.gov) as well as scoring functions specific for fold recognition (http://fold.doe-mbi.ucla.edu). Testing scoring functions on several different decoy sets allows for the exploration of a vast conformational space of proteins, which a single energy function alone might not be able to provide.

A function's performance can be measured in many ways. The simplest method is to look at the RMSD of the best scoring conformation and the native structure. It is also possible to

estimate the probability of choosing the best conformation by chance, RMSD rank of the conformation divided by the total number of conformations. The correlation coefficient of the RMSDs and the scores is also a good method because it uses information about all the conformations in the decoy set.

1.3.4 Distance Geometry

The general goal of distance geometry calculations, is to build model structures that satisfy a set of constraints. This branch of mathematics was developed by Blumenthal while Crippen and co-workers[98-100] were the first to apply these principles to chemical structure problems. Presently, the term distance geometry is used to describe any of the computer programs that convert geometric constraints into three-dimensional molecular coordinates. Distance geometry algorithms are usually fast and can explore a vast conformational space. The structures they generate can be used as input for further refinement methods.

Constraints are usually expressed in terms of an objective function. One way of doing this is to specify a target value for a parameter of interest (e.g. a distance or an angle) and then have the objective function measure deviations from the optimum value. Another way is assign upper and lower bounds on a certain parameter. When boundary conditions are violated, a penalty term is added to the objective function.

Most distance geometry programs have four parts: input preparations, bounds generation and bounds smoothing, embedding, and optimization. Many types of distances can be used as input including holonomic, experimental, and those from secondary structure. Holonomic distances are determined directly from the protein sequence. Templates of each amino acid are made and generally include bond lengths, fixed dihedral distances like those in the peptide bond, and distances involved in rigid structures like aromatic rings. Upper and lower bounds are usually set to +/- 2% of the distance of interest.[100] Experimental distances can be derived from

NOE data. Usually, a 5-Å upper bound is applied while the lower bound is the sum of the appropriate van der Waals radii. Information from secondary structure, like hydrogen bonding constraints, can also be used.

All the distances are stored in a (N*N – 1)/2 symmetric matrix while the bounds matrix is not symmetric. Because only a small number of all the interatomic distances will be found through experiment, other constraints, like the triangle inequality, must be used as well. For the upper bounds, the triangle inequality shows that for three atoms (i, j, k), the distance between i and j (D_{ij}) must be less than the sum of distances from i to k (D_{ik}) and from k to j (D_{kj}). If D_{ij} is greater than $D_{ik} + D_{kj}$, the distance is replaced with the sum. If the sum of the two distances is less than the lower bound, then a triangle violation has occurred. After applying the triangle inequality to the upper bounds, it can be applied to the lower bounds. The overall inequality can be summarized as follows: the upper bound on D_{kj} must be greater than or equal to the sum of the upper bound of D_{ij} and the lower bound of D_{ik} for all atoms i, j, k.

1.3.5 Chemical Cross-Linking with MS

Recently, a technique involving the use of intermolecular cross-linking, mass spectrometry, and sequence threading has been employed in a structure prediction method.[29] Using a lysine-specific cross-linking agent, BS^3 (bis(sulfonsuccinimidyl) suberate, the tertiary structure of (FGF)-2 (bovine basic fibroblast growth factor) was probed. Tripeptide mapping using time-of-flight mass spectrometry was employed to identify the eighteen (Lys-Lys) cross-linking sites and distance constraints were derived from this information. Threading was then used to assign the protein to a family of folds. This method, which requires only a small amount of sample, is fast and easily automated.

The BS^3 cross-linker reacts with the amine groups of Lys and the N terminus. Only one Lys-Lys cross-link per molecule was seen, ensuring the tertiary structure remained unperturbed.

The masses of tryptic peptides were assigned from the mass spectra using the Automated Sequence Assignment Program (ASAP).[29,101-103] The cross-linked residues are identified because Trypsin cleaves at lysines and arginines, but not at BS^3-modified lysines.

To identify the protein fold, a sequence threading program, program 123D,[104] was used to find the twenty best structural models from a database of proteins with at least 30% sequence identity. The models were then ranked by how similar their distances were to the cross-link-derived distance constraints. The threading models were scored according to Equation 1-5 where N is the total number of modes, d_i is the C^{α}-C^{α} distance between the residues in constraint i, and d_0 is the maximum C^{α}-C^{α} through space distance between the BS^3-cross-linked lysines.

$$\sum_{i=1}^{n} \begin{array}{l} 0, \text{ if } d_i \leq d_0 \\ d_i - d_0, \text{ if } d_i > d_0 \end{array} \qquad (1\text{-}5)$$

In their work, approximately N/10 constraints (where N is the number of residues) were found to provide enough distance information to correctly assign the fold of the studied protein. This method can be used to study most proteins if the fold has been previously observed. There are many cross-linkers available that react with other polar groups besides lysine. These cross-linkers may also have different spacer arm lengths and flexibility. This method can also be combined with the other methods discussed in section 1-3.

1.3.6 Our Method

As mentioned previously, the more common experimental structure determination methods are expensive and time consuming. We have employed a creative use of less expensive experimental methods in an attempt to overcome some of the obstacles associated with the more common structure determination methods. We find only a modest decrease in the resolution of the predicted structure. Even low resolution structures, have been demonstrated to provide

insights into protein function.[105] We suggest a method using simple computer algorithms and relatively inexpensive inter-residue distance measurements to generate low resolution models which can be further refined with additional procedures.

We propose a method to predict the unknown structure of protein using a database of protein-like structures. After generating over 8 million decoys, we eliminate the bad ones using inexpensive distance measurements. We will test the following two hypotheses for our method: (1) our decoy set is complete, therefore, a target protein will have similar structure to a member of the decoy set; (2) proteins with a small set of similar inter-residue distances (much smaller than the total number of distances) will have similar overall structure.

Our decoys are derived from structures in the Protein Data Bank (PDB), ensuring that all common protein folds are represented[106] (See Chapter 2). After choosing a target sequence of unknown structure, several C^{α}-C^{α} distances are measured. Experimental techniques like NMR (NOE), EPR, and FRET can be used to measure small (3 – 7 Å), medium (10 – 25 Å), and large (25 – 100 Å) distances respectively. Determining radius of gyration through scattering experiments can also generate useful information. All of these methods generally cost less than X-ray crystallography.

To test the feasibility of our method, we search the decoy set for target proteins whose structures have been solved experimentally but have not been explicitly included in our set. Experimental data is simulated by calculating C^{α}-C^{α} distances from the experimentally determined structures of our target proteins. Those distances are then used as search constraints. The same set of distances are calculated for each of the decoy structures and compared to those measured in the target protein. Structures with several similar α-carbon distances also have similar three-dimensional structure. Our first hypothesis suggests there should be at least one

"surviving" structure in the decoy set while our second hypothesis, if true, guarantees the number of surviving structures to be low. Therefore, our final protein structure predictions are the decoys satisfying the most distance constraints.

Recently, the rate of new protein folds deposited into the PDB has reached a plateau, suggesting that most novel protein folds have already been discovered using the techniques presently available.[107,108] This finding bolsters our assumption that most small (~100 residues), single domain, folded proteins have a structurally similar decoy in our database. We limit the use of our method to proteins containing 100 residues or fewer by generating decoys exactly 100 residues long. Our method is not intended to predict the structure of membrane proteins, as such proteins are not as well represented in the PDB.

1.4 Critical Assessment of Techniques for Protein Structure Prediction (CASP)

The Critical Assessment of Techniques for Protein Structure Prediction (CASP, http://predictioncenter.org/casp7) is a community-wide experiment that allows protein structure prediction groups to test and compare their methods. The goals of CASP are threefold:[109] to determine the abilities and limitations of the current methods; to determine where progress is advancing; to determine where the field is not making progress due to specific bottlenecks. The categories of predictions are always changing slightly from one round of CASP to the next. For example, evaluation of high resolution models was suggested at the CASP6 meeting and implemented in CASP7.

In Chapter 6, we discuss the results of testing our method with CASP targets. The CASP organizers solicit from experimentalists target protein sequences whose structures are close to being determined or have not yet been published. The participants are given only the target sequences and a limited amount of time to use their prediction methods to determine the target protein structures. After analyzing the results, the organizers hold a conference at which the

most successful groups are asked to present their methods. Attendants may also make suggestions for future CASP experiments. There have been seven rounds of CASP since its inception in 1994.

CASP7 included three primary categories of prediction, (1) Tertiary structure predictions, (2) High resolution models, and (3) other predictions. Each category is further divided into automated and human-aided predictions. The "human" predictions can be made using any combination of computational and human methods, but the automated structure prediction servers must be fully automated.

The tertiary structure predictions are further divided into two types, template based modeling and template free modeling. The first group includes the previous categories of comparative modeling, homologous fold based models, and some analogous fold based models. The second group contains models of proteins with new folds (previously unseen) as well as hard analogous fold based models.

The second primary category, high resolution models, is new. It contains a subset of models from the tertiary structure prediction category whose backbones are highly accurate such that the details of active sites, loops, and side chains can be evaluated.

The other prediction category looks at how well predictors were able to define boundaries of structural domains, detect residue-residue contacts, and identify the regions of disorder in the targets. This category also includes predictions of function from structure. Another new facet of evaluations included judging the predictor's ability to discern the best model from their respective decoy set without knowledge of the native structure. Evaluating model refinement is also important as there is much interest in generating models with high accuracy.

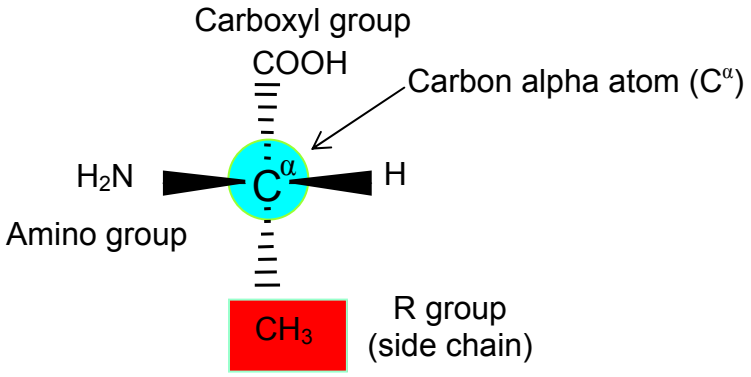


Figure 1-1. Structure of an amino acid (alanine) showing the α-carbon, the R-group (CH_3 for alanine), the amino group, and the carboxyl group. The amino acid is shown in its neutral, non-zwitterion form.

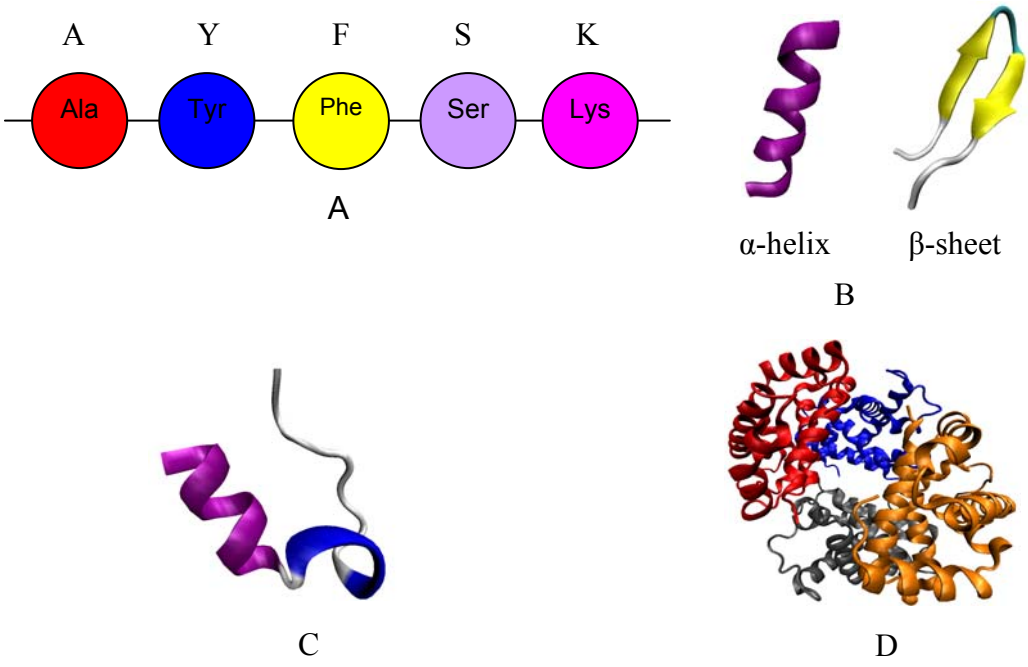


Figure 1-2. Organization of protein structure. A) Primary structure corresponding to the amino acid sequence of: alanine, tyrosine, phenylalanine, serine, lysine. B) Secondary structure: α-helix and β-sheet patterns. C) Tertiary structure of the Trp-cage protein (PDB code 1l2y). D) Quaternary structure of hemoglobin (PDB code 1GZX).

CHAPTER 2
METHODS FOR PROTEIN STRUCTURE PREDICTION

2.1 Decoy Generation

2.1.1 General Decoy Set

All protein structures in the protein databank (PDB) with 100 residues or more (24,561 proteins in all including x-ray and average NMR structures) were used to populate our decoy database. The protein backbones were split into overlapping and running fragments of 100 residues (Figure 2-1) and only the Cartesian coordinates of the α-carbons were stored. A parent protein of more than 100 residues can be segmented into exactly N - 99 overlapping fragments containing 100 residues each, where N is the total number of residues. Thus the first decoy contains the first 100 α-carbons from residues 1 to100, while the final decoy is composed of α-carbons from residues (N - 99) to N. Decoys are named by first listing the PDB code of the parent protein and then the decoy number. If the parent protein is composed of several chains, the chain name is listed after the PDB code. For example, 1m31-a-2 is the second decoy composed of residues 2 – 101 from chain A of PDB code 1m31. Exactly 8,060,245 decoys were generated in this manner, creating a database to find the structure of proteins composed of 100 residues or fewer. Because each decoy is exactly 100 residues long, we disregard the excess terminal residues when searching for shorter proteins.

Several problems were encountered in constructing the decoy set. Some entries in the PDB are missing important atoms or residues causing gaps in the parent protein. Decoys were not generated from the gapped regions of such proteins. The two numbers appearing after some of the PDB codes indicate the parent protein contained a gap and the sequence was renumbered after the gapped region. For example, parent protein 2a6o contained a gap and was divided into two sections, each containing over 100 residues. Decoy 2a6o-2-25 was created from residues 25

– 124 of the second fragment. Multiple positions for a single residue or an entire chain are also frequently seen in PDB entries. We consistently selected the first coordinate set if multiple positions were provided. For multi-chained proteins, the chains were treated as separate entities. Several proteins have multiple PDB entries; no attempt was made to rid the database of redundant structures.

2.1.2 Specific Decoy Set

The Rosetta procedure has been described in depth elsewhere.[63,83,110,111] Briefly, generating decoys with Rosetta requires the initial formation of a fragment library using Robetta [112-114]. Robetta divides the target protein's sequence into fragments of three and nine residues and searches the protein databank (PDB) for the possible structures of these fragments, which represent the range of accessible local structure conformations. These fragments are then pieced together randomly using a Monte Carlo simulated annealing procedure with an energy function that favors hydrophobic burial, paired β-strands, and specific side-chain interactions. Each decoy is evaluated by how well it compares to a protein-like structure based on statistics of known protein structures.

2.2 Decoy Discrimination

As described in section 2.1, decoys are protein-like structures that may or may not look like the target. The goal of the search process is to find a decoy (or small set of decoys) similar in structure to the target protein. Using inter-residue distances from the target protein as constraints, we distinguish between the good decoys (structures with low RMSDs using the target protein as a reference) and the bad decoys (high RMSD structures). These distances can be measured from experiments like nuclear overhauser effects (NOE) in nuclear magnetic resonance spectroscopy (NMR), electron paramagnetic resonance (EPR), and fluorescence resonance energy transfer (FRET), which can measure short (3 – 7Å), medium (10 – 25Å), and

long (25 – 100 Å) distances respectively. Such measurements are not exact; the probes in FRET and EPR are constantly rotating and have a finite size making their exact orientation difficult to predict. Also, the measured distances are between the spin labels, while we simulate the experimental data using the distance between two α-carbons. The distance uncertainties in EPR measurements without considering spin orientation, are estimated to be around 5 Å.[25] All of these uncertainties must be taken into consideration in our search process. While comparing the C^{α}-C^{α} distances of the target protein to those of the decoys, upper and lower bounds are placed on the target protein's distance. A decoy satisfies the constraint only if its distance is within the range, the constraint distance acceptance range.

The acceptance range indicates how much the decoy distance can differ from the target distance and still satisfy the constraint. Smaller ranges, +/- 1 and 2, are too tight; some low RMSD structures do not satisfy many constraints using this range. Also, this range is too small to account for experimental uncertainty. Larger ranges, +/- 10 and +/- 15, are too broad, allowing high RMSD structures to satisfy several constraints. After many trials, a more moderate range of +/- 5 Å was found to yield the best results. This range also compensates for insertions and deletions, as the distance between two consecutive α-carbons is ~ 3.8 Å.

In addition to finding an optimal constraint distance acceptance range, the choice of which distance constraints to use is a key factor in the success of this method. When choosing constraints, it is helpful to initially run the amino acid sequence through a secondary structure prediction method. For our studies we used JPred,[115] a consensus method that gets result from six secondary structure prediction algorithms[116-119] that use evolutionary information from multiple sequences. Based on sequence information, JPred highlights which fragments of the chain are more likely to exist as α-helices and which will be β-sheets. We identify approximate

regions of defined secondary structure to avoid choosing distances between atoms in loop regions, which are highly mobile and less structurally defined, even in fairly similar structures. Therefore, the most effective constraints are distances between atoms in defined areas of secondary structure, like α-helices and β-sheets. These regions are usually more conserved as they often play a significant role in the protein's function.

After using the acceptance range to compare the calculated distances, we scored the decoys by counting the number of constraints each one satisfies. Over a series of trials, we found that twenty-five distances were sufficient to rank the decoys; therefore, our scores range from 0 to 25. Other researchers have found a similar amount of experimental distance information to be necessary in structure prediction.[120,121] This scoring method provides some protection against poor constraint choices (constraints satisfied by high but not low RMSD structures). In our previous trials (Chapter 3), constraints were applied sequentially and decoys not satisfying the constraint were eliminated from the database at each step. The few decoys remaining after several constraint applications were the structure predictions. When a poorly chosen constraint was applied, low RMSD structures were immediately eliminated from the database and were, therefore, unable to become the predicted structures.

The counting method makes the sum of the constraints more important than each one individually, minimizing the effect of a few bad choices. Applying a poor constraint can result in a low RMSD structure having a slightly lower score and a high RMSD structure having a slightly higher score. The effects on the scores are so small that the low RMSD structures are still predictable.

In summary, our search process is divided into 5 steps: (1) use a secondary structure prediction method to identify important distance constraints. (2) Measure distances

experimentally. (3) Calculate the same set of distances for each decoy. (4) Compare each of the target protein's distances with those of the decoys. A decoy satisfies a particular constraint if the two distances are similar within the constraint distance acceptance range (+/- 5Å). (5) Score each decoy by counting the number of constraints it satisfies. We hypothesize that structures with similar α-carbon distances will show similarities in overall structure. The decoys satisfying the most constraints become our structure predictions. While testing our method, we search for a protein of known structure. We can therefore simulate the experimental constraints in step 2 by choosing a set of distances from the native structure of the target protein.

2.3 Choosing Constraints

There are many ways to choose constraints. In our early work (Chapter 3), we attempted to randomly select constraints. Usually, a few of these random constraints involved atoms in loop regions, which is problematic as these regions are often not structurally well defined. To avoid choosing constraints from loop regions, we use a secondary structure prediction method to identify these areas. With this knowledge, we selected constraints between all predicted secondary structure elements (Chapters 4 and 5).

One can also choose constraints in a daisy-chain manner. In this method, each atom in a constraint is also involved with another constraint. For example, let constraint 1 be the distance between residues A and B. Then constraint 2 is composed of residues B and C while constraint 3 is derived from residues C and D. Constraints chosen in this manner are stronger than those randomly chosen because they must obey the triangle inequality.

It is also possible to select a few atoms and use several of the distances between them as the set of constraints. Many implicit constraints are imposed in this manner. Another method is to choose a piece of secondary structure to serve as a reference and have all constraints involve

an atom from this region. This method of choosing constraints has been shown previously to be significantly better than daisy-chaining.[122] In our later work (Chapter 6) we used a combination of these three methods.

2.4 Comparing Results

In the recent Critical Assessment of protein Structure Prediction (CASP) experiments, the Local-Global Alignment (LGA)[123] (http://PredictionCenter.llnl.gov/local/lga) measure has been used to evaluate the prediction results. Although RMSD calculations provide insight into global similarities between protein structures, the LGA method was designed to measure similarities in both global and local structure. This program creates several alignments between the structures of the predicted model and the target to find those regions of the proteins that are most similar to each other. The LGA method has two components, the longest continuous segment (LCS) and the global distance test (GDT). Several iterations of both methods are usually required to find the optimal alignments.

When comparing two protein structures, the LCS procedure searches for the alignment that superimposes the longest section of continuous residues with an RMSD under a specified cutoff. For example, suppose the RMSD cutoff was 5.0 Å and the first three residues of each structure are aligned and had an RMSD 4.0 Å. The program would then align residues 1-4 and compute the RMSD again. If the RMSD was still under 5.0 Å, residue 5 would be included in the calculation otherwise the RMSD between residues 2 – 5 would be computed. This process continues as several alignments are sampled and then the LCS is identified. Any set of residues in the model can be aligned to the target; they do not need to have the same location in each sequence (eg. residues 4, 5, and 6 of the model can be aligned to reside 23, 24, and 25 of the target). Unless otherwise indicated, in this paper we discuss LCS using a cutoff of 5.0 Å and use the abbreviation LCS-5.

In the GDT method, the structures are aligned to find the largest set of residues that differ by less than a selected distance cutoff. The distance cutoffs range from 0.5 – 10.0 Å and are scanned at a 0.5 Å interval. For GDT, the largest set is not necessarily composed of continuous residues. Pairs of residues are selected from both structures and a superposition and RMSD are calculated. The superpositions are used as starting points to generate a list of equivalent residues (carbon-α pairs from both proteins). After aligning the target and model structures, the distances between the equivalent residues are calculated and the number of residue pairs with distances under the cutoff is counted. The residues above the threshold are removed, others are added, and the distances are calculated again. The initial list of equivalent residues is thus iteratively extended to find the largest set of residues that satisfy a specific distance threshold.

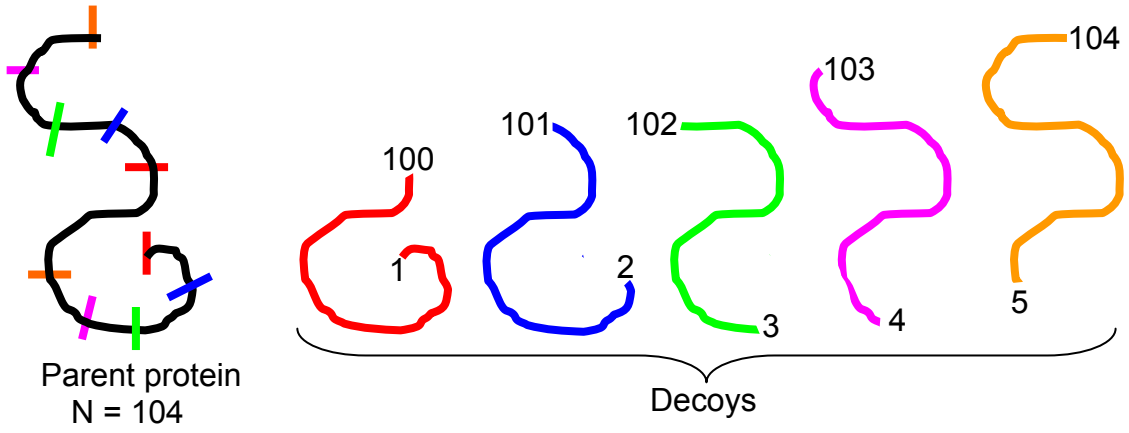


Figure 2-1. An example of how decoys are generated from a single protein. A parent protein with 104 residues (N = 104) can be cut into 5 decoys (104 – 99 = 5), each with exactly 100 residues. The first decoy contains α-carbons 1 – 100 while the final decoy is composed of α-carbons from residues 5 – 104 [(N - 99) – N].

CHAPTER 3
TRIALS AND ERRORS: DEVELOPING THE METHOD

In this chapter method development is discussed. To test the idea for the method, we searched through previously constructed decoy databases that were designed to test scoring functions. We then developed our own decoy set according to the methods set out in Chapter 2. A target protein with a structure known to be in the database was selected to test our search protocol. Finally we predicted the structure of a protein whose native structure was not included in the database.

3.1 Testing the Method on Previously Constructed Databases

We tested our search procedure using Samudrala and Levitt's pre-constructed databases (http://dd.compbio.washington.edu) for four known protein structures (1bba,[124] 1b0n-b,[125] 1ctf,[126] and1dtk[127]), which ranged in length from 31 to 78 residues. For each individual target protein, a unique set of decoys was created. The number of decoys for each target ranged from 216 to 501 (Table 3-1). Standard bond lengths and angles were used to generate the initial structures. The trans configuration was used for all peptide bonds and predefined α-helices and β-sheets were assigned ideal torsion (Φ,Ψ) angles of (-60°, -40°) and (-120°, 150°) respectively. For the remaining residues in the loop regions, a range of random torsion angles were used, -120(±60)º for Φ and 150(±90)º for Ψ.[128]

After obtaining these databases from Samudrala and Levitt's, distance constraints for each protein were chosen from the native structure (Table 3-1). To develop our search protocol, we investigated the effect of using different types of constraints to distinguish the more native-like decoys from the less native-like. We calculated the number of decoys in the database that satisfy particular constraints to identify which distances were common to most decoys and which were

very different. Determining which constraints eliminated the most structures (long or short distances) was investigated by applying the same set of constraints in different orders (randomly, long to short distances, and short to long distances).

3.1.1 The Number of Structures Satisfying Specific Constraints

For each target protein, constraints were chosen by finding the most prevalent residue type and calculating the distances between those residues. For 1bba, Ala was the most common residue but the number of Ala-Ala distances was too small; Tyr-Tyr distances were also included in the constraint list. The number of structures in the appropriate database satisfying each constraint was then determined. We began these studies using a constraint distance acceptance range of +/- 2 Å (we later found this range to be too small when applying the method to our general decoy set). For example, if the distance between the C^{α} of residues A and B in the native structure is 10 Å, a decoy will satisfy that constraint if it has a distance of 8 – 12 Å between the C^{α} of residues A and B. For this test, we were not concerned with the quality of the decoys that satisfy each constraint, only the number. We wanted to find the minimum number of constraints that eliminate all structures in the database except that of the target.

As can be seen in Figure 3-1, care must be taken when choosing constraints. Distances that are present in all of the decoys are not very effective constraints as they eliminate very few structures. Some of the proteins have more complete data sets than others by having decoys that sample a wider range of structural possibilities. As can be seen in Figure 3-1A, constraints 12, 15, and 16 for 1bba eliminate almost all decoys in database. These constraints involve an atom in a highly variable loop region. The same is true for constraints 7, 8, and 16 for 1b0n-b in Figure 3-1B. For 1dtk and 1ctf, (Figure 3-1D, E) most of the constraints are satisfied by fewer than half of the decoys in the database.

3.1.2 The Effects of Applying Constraints in Different Orders

3.1.2.1 Randomly ordered constraints

For 1bba, all Ala-Ala and Tyr-Tyr distances were selected as constraints. These distances ranged from 3.8 – 30.2 Å. A constraint distance acceptance range of +/- 2 Å of the constraint distance was used as the acceptance criterion. The same set of constraints was applied in three different, randomly determined orders. In each of the three trials, many structures are eliminated after applying a single constraint (Figure 3-2) and 5 – 13 constraints were needed to eliminate all but one decoy. The Tyr-Tyr constraints were also found to eliminate more structures in the first step than the Ala-Ala constraints, indicating that the decoys span a wide range of distances at those points. As mentioned previously, some of these constraints involved atoms in a highly mobile or ill-defined loop region. Despite their ability to eliminate many decoys, constraints involving atoms in loop regions may not be the best choices due to their low resolution and high variability between structures which are otherwise quite similar.

3.1.2.2 Same constraints in different order

For each target protein, a list of constraints was chosen and applied in the following orders: (1) long to short distances and (2) short to long distances. For 1b0n-b, all Glu-Glu distances were selected as constraints. While the total set of constraints had distances that ranged from 3.8 – 23.8 Å, the six constraints used in Trial 1 (long to short) ranged from 19.3 – 23.8 Å while the nineteen constraints in Trial 2 (short to long) ranged from 3.8 – 23.8 Å (Figure 3-3A). For 1ctf, all Val-Val distances were chosen as constraints and ranged from 16.4 – 24.7 Å in the Trial 1 and 4.7 – 11.4 Å in the Trial 2 (Figure 3-3B). As seen in Figure 3-3, the number of needed constraints depends on the order of application. In each trial, fewer constraints were needed when the longer distance constraints were used initially. Similar results were seen for 1bba and

1dtk. In each case, the final decoy remaining in the database was the native structure of the target and therefore satisfied all constraints.

Applying longer constraints first eliminates decoys with vastly different overall structures compared to the target protein. Shorter constraints eliminate decoys with differing local structure. We have found that the search time is shortened by first removing structures with great overall differences by applying longer constraints and then applying shorter constraints to fine tune the structure. For 1b0n-b, applying long constraints first (Trial 1, Figure 3-3A) requires only six constraints, much less than the nineteen needed when applying short constraints first (Trial 2, 3-3A). Some constraints do not eliminate any additional decoys resulting in the plateaus seen in Figure 3-3. For 1ctf, application of large constraints first (Trial 1, Figure 3-3B) requires seven constraints, whereas applying small constraints first (Trial 2, Figure 3-3B) requires thirteen. It was also found that changing the constraint distance acceptance range from +/- 2 Å to +/- 1 Å did not change the number of distances required to find the correct structure.

3.2 Developing a Search Protocol Using a Structure Known to be in Our Database

Our general decoy set was generated as described in the Chapter 2. To test our search procedure using our database, we chose 2ezm,[129] Cyanovirin-N (an HIV inactivating protein). The database contains two decoys generated from the 101 residue 2ezm target protein. Our decoy set is spiked—the correct structure is definitely present because the parent protein met all the requirements to be included in the decoy generation process.

In Trial 1, constraints were chosen with distances ranging from 10.1 – 24.7 Å. They were also selected so that the atoms were within 8 – 59 residues of each other in the sequence. For Trial 2, constraints were chosen in the distance range of 5.2 – 17.9 Å. The constraint atoms were also required to be within 3 – 10 residues of each other, much closer than those constraints used in Trial 1. The constraint distance acceptance range was set to +/- 2 Å as used in previous trials.

Due to the changes in constraints, Trial 2 needed more than twice the number of constraints needed in Trial 1. From these results we can conclude that it is more efficient to consider longer distances between atoms several residues apart when choosing constraints

In each trial we were able to narrow down our search to the three structures shown in Figure 3-4, 2ezm-1,[129] 2ezn-1,[129] and 1iiy-1.[130] Their structures are virtually indistinguishable because they are all from the same HIV inactivating protein. The PDB entries differ as follows: 2ezn represents an ensemble of NMR structures, 2ezm is only the mean NMR structure, and 1iiy is the mean NMR structure with a ligand. When shorter distance constraints were used and restrictions were placed on the number of residues apart the two atoms were allowed to be, the number of constraints needed increased from eight in Trial 1 to twenty in Trial 2.

3.3 Developing a Search Protocol Using a Structure Not in Our Database

Next, a search of the database was performed to find a protein whose structure was not included in developing the database. Target protein 1b4c[131] is a homodimer of S100 beta subunits, each 92 residues in length. It has been classified as a metal-binding protein. Due to the nature of the database, the structure of only one chain was chosen as the search target.

3.3.1 Constraint Distance Acceptance Ranges: +/- 2 Å and +/- 4 Å

Our first task was to find upper and lower bounds to use as the constraint distance acceptance range. We started with +/- 2 Å as in previous trials (section 3.1, 3.2). Constraint distances were chosen to be between 11.0 Å and 25.6 Å with an average distance of 20.6 Å. Seven constraints were required to eliminate all but twenty-five structures. Three decoys satisfied eight constraints and can be found in Figure 3-5. The RMSDs for the decoys satisfying seven and eight constraints are listed in Table 3-2. The RMSD for the top three structures was 14.4 Å. The parent proteins for these three decoys are all related; 1kv7[132] and 1n68[133] are the multi-copper oxidase (CueO) and 1pf3[133] is the M441L mutant of the same protein. The 1pf3

decoy found in the search is a fragment of the protein which does not contain the mutation. These three decoys are nearly identical because of the redundant nature of the database.

The average RMSD for the top twenty-five decoys was found to be 13.3 Å. Applying the final constraint apparently removed some of the better (lower RMSD) structures. This search was unable to predict the correct secondary structure. The native structure of 1b4c has five α-helices, this search found decoys with four β-sheets and one short α-helix.

The RMSDs found in Trial 1 indicate that the three decoys remaining in the database are not reliable predictions. Reva et. al.[134] showed that a structure with an RMSD of less than 6.0 Å is a successful prediction for small proteins. If the acceptance range is too tight, low RMSD decoys may not satisfy all constraints. In order to improve our results, the constraint distance acceptance range was increased from +/- 2 Å to +/- 4 Å. Thirteen constraints were required to eliminate all but four decoys. The remaining structures can be found in Figure 3-6. The parent proteins of these decoys are all dehydrogenases; 1h0h[135] is a formate dehydrogenase while 1nek[136] and 1nen[136] are succinate dehydrogenase. Increasing the distance range improved the quality of the final structures. A better prediction of secondary structure is made as the decoys are found to have four α-helices and only two small β-sheets. The RMSDs of 12.2 Å and 13.3 Å are slightly better than +/-2 Å distance range used in Trial 1, but they are still out of range for this method to be considered a success.

3.3.2 Calculation of All RMSDs

Due to the high RMSD values of the final structures found in previous trials, we calculated the RMSDs for all the decoys using the native structure of 1b4c as the reference to determine if any "good" (low RMSD) decoys existed in our database. The distribution of the RMSDs in Figure 3-7 shows that most of the structures are within 12 –20 Å. We found 353 structures that

have RMSDs less than 7 Å and 85 with RMSDs less than 6 Å. The structures with the best RMSDs can be found in Table 3-3 and are depicted in Figure 3-8.

It was found that the good structures were eliminated during our search procedure because some of the chosen constraint atoms are in loop regions which differ greatly among the structures. Other distance constraints were between atoms that were shifted slightly in the sequence due to insertions and deletions. Figure 3-9 shows an example of how insertions and deletions can hinder successful predictions. The two proteins in the figure differ only in their loop regions, the α-helical sections are highly conserved giving rise to a very small RMSD. Using the distance between residues 10 and 13 as a constraint, the present searching method would eliminate the black decoy as an improbable structure because the extra residue in the loop region adds to the length of the distance of interest. To overcome this problem, we increased the acceptable distance range from +/- 4 Å to +/- 12 Å.

3.3.3 Constraint Distance Acceptance Range of +/- 12 Å and +/- 12 Å → +/- 10 Å

Using similar constraints as in the previous trials (sections 3.3.1, 3.3.2) but with a constraint distance acceptance range of +/- 12 Å, seventy-five constraints were required to find the top 1,163 decoys. The structures with the seven lowest RMSDs calculated previously were in the final decoy set. The finding of the seven lowest RMSD structures showed our method to have some promise, however, without knowing the structure *a priori*, it would be extremely difficult to distinguish between the seven good and over one-thousand bad decoys. It can therefore be concluded that this constraint distance acceptance range of +/- 12 Å is too large to adequately eliminate the least likely decoys.

In order to eliminate more structures, a constraint distance acceptance range of +/-12 Å was employed for the first twenty-five constraints and +/-10 Å was used for the next twenty-five

constraints. After these fifty constraints were applied, only 620 structures remained. The seven lowest RMSD structures were once again among the remaining decoys. Unfortunately, 620 structures is still a rather large number and 50 constraints are far too many for this method to be cost effective.

3.3.4 Block of Distances

Instead of comparing one distance to the native structure at a time, in this trial a block of distances +/- 2 residues from the distance of interest was compared. For example, if the experimental data indicated that residues 10 and 20 were 15 Å apart, we would calculate all the distances between residues 8, 9, 10, 11, 12 and 18, 19, 20, 21, 22. These twenty-five distances make up the block for each constraint. For each block, the distance range, maximum, and minimum were calculated. A decoy satisfied the constraint if the native structure distance was found to be in the distance range (max + 2Å, min - 2Å). Two small restrictions were placed on the constraints: (1) the distance constraints ranged from 15.1 – 35.2 Å; (2) the atom pairs were between 9 and 77 residues apart in the sequence.

Twenty-five constraints were required to find the top 943 structures, of which only 79 had RMSDs less than 7.0 Å. Application of 32 constraints resulted in five remaining structures found in Table 3-4 with their RMSDs. This method showed some improvement over the method used in section 3.3.3. Of the final structures remaining in the database, only one was found to have an RMSD greater than the cut off for it to be considered a successful prediction. The four good structures can be found in Figure 3-8 and the higher RMSD decoy, from parent protein 1mka[137] can be found in Figure 3-10. The decoy from 1mka shares very few structural similarities with 1b4c. It has two β-sheets as well as two α-helices that do not align well with 1b4c. A large distance range for the constraint acceptance requirements and poorly chosen

constraints may explain why this decoy was not eliminated during the search process. The large distance range also requires too many constraints making this method computationally expensive.

3.3.5 Vary the Order of Constraint Application

As seen during the initial testing of our method (see section 3.1), the results of each trial depend greatly on the order of application of the constraints. Several trials using the same constraints in different order were performed. The constraint distance acceptance range was set to +/- 5 Å.

The first set of constraints used the same order as that used in the previous trial (section 3.3.4). Eighteen constraints were satisfied by seventeen decoys with RMSDs ranging from 10.8 – 15.2 Å as found in Table 3-5. None of the low RMSD structures were found. It was discovered that the lowest RMSD decoy satisfied 24 of the 25 constraints. One of the atoms in the unsatisfied constraint is in the middle of a loop region. It is known that these regions have much flexibility giving rise to very different conformations, even in otherwise similar proteins. We placed this constraint at the end of the list and performed the trial again. Seven decoys were found to remain after 21 constraints. The RMSDs of the top seven decoys can be found in Table 3-5. This method was able to find two of the best decoys, but it found five high RMSD decoys as well. A final trial was performed using another order of the constraints. Twenty-one constraints were required to eliminate all but six decoys. The same two low RMSD decoys were found as in the previous order. The final trial found four high RMSD decoys which were different than those found earlier.

Initially it was assumed that upon varying the order of the application of the constraints, the low RMSD decoys would remain in the database more often than the less probable ones. As long as the bad constraint was placed at the end of the list, the two lowest RMSD decoys were

always be found. Because *a priori* one would not know if a bad constraint was being used, this method is not as effective as we would like it to be.

3.3.6 Count the Number of Satisfied Constraints for Each Decoy

In order to remove the dependence on the order of constraint application, we counted the number of constraints each decoy satisfied. We assumed the decoys that satisfied the most constraints would have the lowest RMSDs. We performed a trial using the same constraints as those used in the previous trial with a distance range of +/- 5 Å. It was found that four decoys satisfy twenty-five constraints. The RMSDs of these structures are found in Table 3-6. Two of the lowest RMSD decoys were found along with two rather high RMSD decoys.

A slightly different set of constraints were selected that include only distances between atoms that are involved in secondary structure, not the loop regions. The distances were chosen to be between 11.9 – 26.9 Å. Four decoys were found to satisfy these twenty-five constraints. They have the lowest RMSDs in the database (Table 3-6, Trial 2). As seen with the previous set of constraints, half of the decoys in the database satisfy ~11 constraints. This data is shown in Figure 3-11 and is remarkably similar to that obtained using the other, different set of distance constraints.

3.4 Determination of an Average RMSD Distribution

Because the four target proteins had similar RMSD distributions, we wanted to determine a random average RMSD—the RMSD of two randomly chosen structures in the decoy set. We calculated the RMSD of each decoy in the database using other decoys as references. The five reference decoys can be found in Figure 3-12. As explained in the Chapter 2, the decoys are 100 residue long fragments of larger proteins. The decoys in Figures 3-12A, 3-12C, and 3-12E from parent proteins 1b7u,[138] 1ujn,[139] and 1rt6,[140] have both α-helices and β sheets, while Figure 3-12B (from parent protein 1fhx[141]) represents an all β-sheet protein and Figure 3-12D (from

parent protein 2wrp[142]) contains only α-helices. Most of these decoys are folded rather tightly and resemble small proteins. The decoy from 1rt6 (Figure 3-12E), however, is a fragment of a very large multi-domain protein, HIV-1 reverse transcriptase. This particular decoy contains residues in two domains even though they are connected through the same chain. Decoys like this one can account for some of the poor RMSD values calculated with other references.

For 1rt6-109, the RMSDs are shifted to the right with average values between 15 Å and 25 Å (Figure 3-13), indicating that most decoys are less similar to it than the more compactly folded structures. Our database contains some less compact, semi-folded decoys and a search for such a protein may result in finding a reliable decoy where searches of other databases may not.

3.5 Summary of Methods

A constraint range of +/- 2 Å and +/- 4 Å was found to be too small to obtain good results while a constraint range of +/- 12 Å is far too large. The block method was able to find the lowest RMSD structures, but it required too much computer time and too many constraints to do so. The method of counting the number of constraints that each decoy satisfies has yielded the best results thus far. Target 1b4c was studied previously using a *de novo* protein structure prediction algorithm which employed Rosetta.[143] The 3.6 Å RMSD of our best decoy was slightly better than their best-scoring cluster which had an RMSD of 4.6 Å. We will further discuss the application of this method to other proteins.

Table 3-1. Comparison of input for the four target proteins

Target	Number of decoys in database	Range of distance constraints, Å	Number of residues in sequence
1bba	501	3.8 – 30.2	36
1b0n-b	498	3.8 – 23.8	31
1ctf	498	4.7 – 24.7	78
1dtk	216	5.3 – 27.4	57

Table 3-2. RMSDs for decoys satisfying the most constraints

Decoy	RMSD (Å)	Decoy	RMSD (Å)
1z7q-n	10.7	1lk7-473	14.1
1ivr	11.9	1lk5-473	14.1
1gtm-c	11.9	1lk7-702	14.1
1qmy	11.9	1lk5-15	14.1
1qol-a	12.0	1lk5-244	14.1
1qol-b	12.0	1lk5-702	14.1
1qol-e	12.0	1lk7-15	14.1
1qol-f	12.0	1kv7-1	14.4
1qol-h	12.0	1n68-1	14.4
1t3q-c	13.8	1pf3-1	14.4
1t3q-f	13.9	1khv-1	14.4
1s18	14.0	1khw-1	14.5
1lk7-244	14.1		

A constraint distance acceptance range of +/- 2 Å was used.

Table 3-3. Lowest RMSD decoys in database using 1b4c as a reference

Decoy	RMSD (Å)
1m31-b-2	3.6
1m31-a-2	3.6
1m31-a-1	4.8
1m31-b-1	4.8
1nsh-b-2	4.9
1nsh-a-2	4.9
1wlm-7	5.9
1wlm-8	5.4
1wlm-9	5.1
1wlm-10	5.6
1psr-a-1	5.3
1psr-b-1	5.3

Table 3-4. Decoys remaining after 32 constraints using the block method

Decoy	RMSD (Å)
1m31-a-2	3.6
1m31-b-2	3.6
1mka-49	10.2
1psr-1	5.3
1psr-1	5.3

Table 3-5. Lowest RMSD decoys found in varying the order of constraint application

Trial 1		Trial 2		Trial 3	
Decoy	RMSD (Å)	Decoy	RMSD (Å)	Decoy	RMSD(Å)
1agr-3-8	11.3	1m31-a-2	3.6	1m31-a-2	3.6
1jr4-1-96	15.1	1m31-b-2	3.6	1m31-b-2	3.6
1rif-4	12.9	1nzc-4-95	12.7	1vgw-a-2-1	15.4
1vid-1-95	15.2	1f8x-1-11	13.2	1vgw-d-2-1	15.4
2a72-1-2	11.8	1f8x-2-11	13.2	1vgw-e-2-1	15.5
2a72-2-2	11.8	1f8y-1-11	12.9	1vgz-4-1	15.4
2af0-1-22	12.0	1f8y-2-11	13.1		
2bt2-1-16	11.0				
2bt2-2-16	11.0				
2bt2-3-17	11.0				
2bt2-4-15	11.1				
2bt2-5-17	10.9				
2bv1-1-11	11.2				
2bv1-2-10	11.3				
1ezt-1-8	10.8				
1fqk-2-12	14.5				
1h1d-1-106	15.1				

Table 3-6. Lowest RMSD decoys found using the count method for both trials

Trial 1		Trial 2	
Decoy	RMSD (Å)	Decoy	RMSD (Å)
1m31-a-2	3.6	1m31-a-1	4.8
1m31-b-2	3.6	1m31-a-2	3.6
1hz4-141	11.0	1m31-b-1	4.8
1hz4-142	11.3	1m31-b-2	3.6

Trial 1 uses the original set of constraints. In Trial 2, the constraint involving atoms in the loop region is replaced by one between atoms in defined areas of secondary structure.

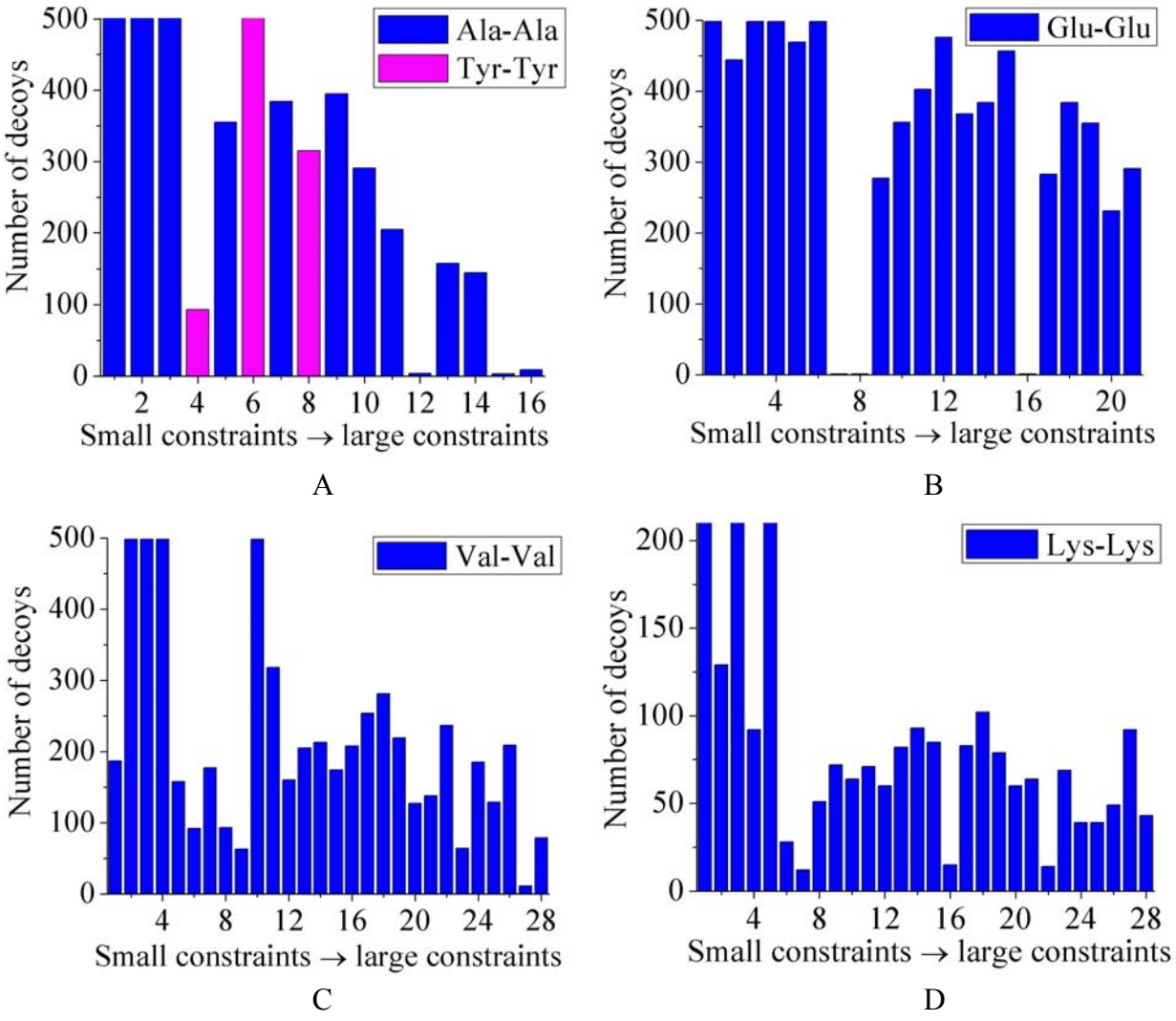


Figure 3-1. The results of counting the number of decoys that satisfy each constraint. Constraints are numbered from shortest to longest distance. A) 1bba, bars in pink correspond to Tyr-Tyr constraints. B) 1b0n-b. C) 1ctf. D) 1dtk. Constraints were selected between the most prevalent residue type for each target.

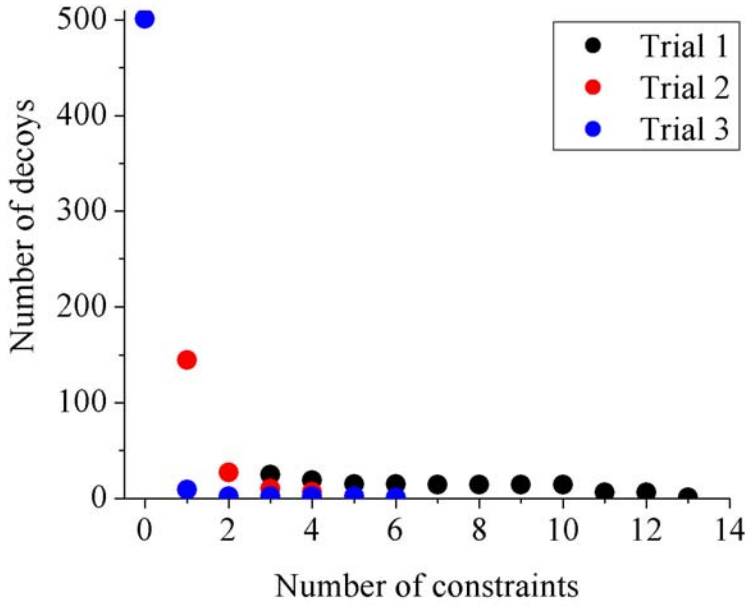
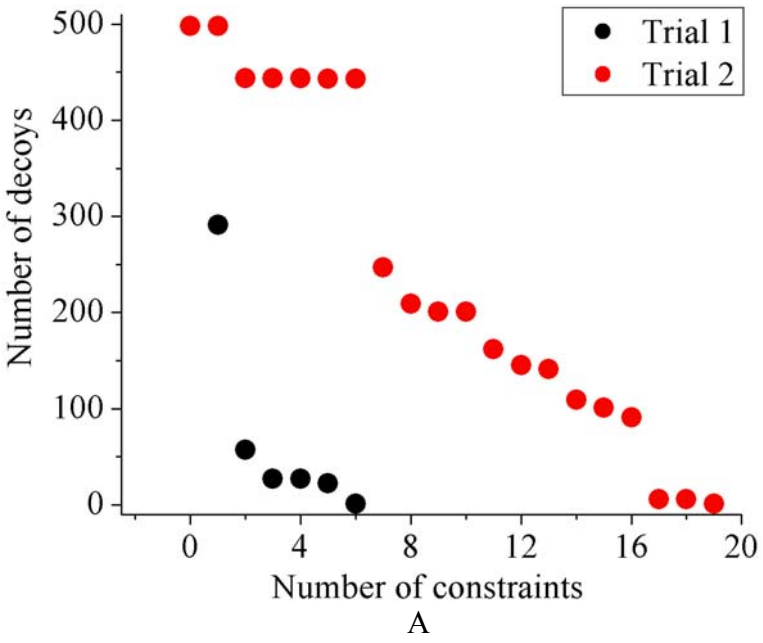
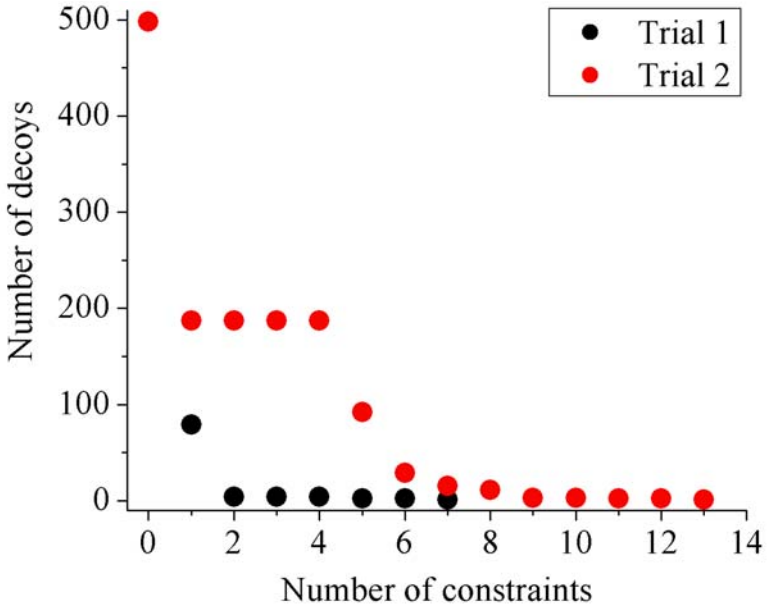


Figure 3-2. Application of randomly ordered constraints for 1bba. The three trials used the same constraints in different, random orders.



A



A

Figure 3-3. Results for using the same set of constraints in different orders. In trial 1, larger constraints were applied first and then smaller constraints until only the target structure remained in the database. In trial 2, constraints were applied from small distances to larger distances. A) 1b0n-b. B) 1ctf. In each case, the final structure remaining in the database was the native structure of the target protein and therefore satisfied all constraints.

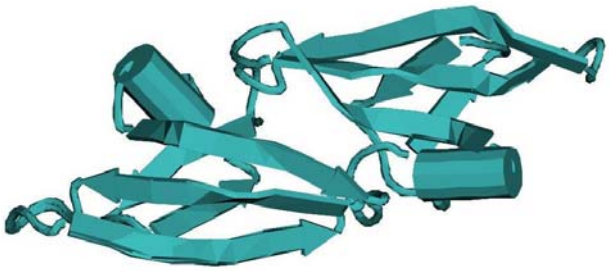
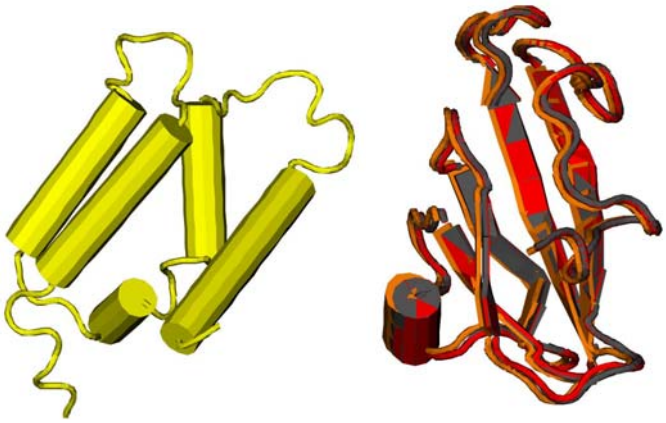
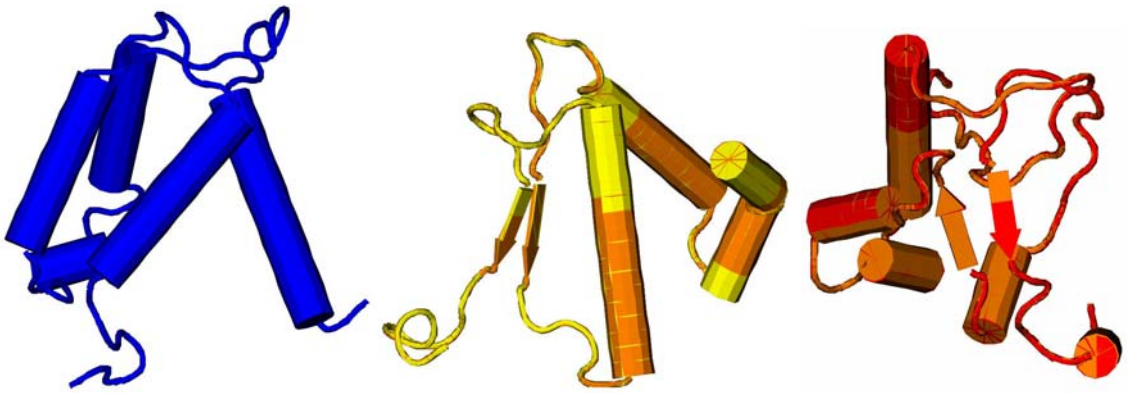


Figure 3-4. The superimposed images of the results of the 2ezm search. The top scoring decoys are 2ezm-1, 2ezn-1, and 1iiy-1. All three PDB codes represent the same protein. 1iiy contains a ligand which was not included in the decoys.



A B

Figure 3-5. Results from Trial 1. A) 1b4c. B) The final three remaining decoys after eight constraints with a +/- 2 Å distance range, 1kv7-1, 1n68-1 and 1pf3-1.



A B C

Figure 3-6. The target protein and the final four remaining decoys after 13 constraints with a +/- 4 Å distance range. A) The native structure of 1b4c. B) 1h0h-a-334 and 1h0h-k-334. Each had an RMSD of 12.2 Å. C) 1nek-a-246 and 1nen-a-246. Each had an RMSD of 13.3 Å. 1b4c is represented using a slightly different orientation than that in Figure 3-5.

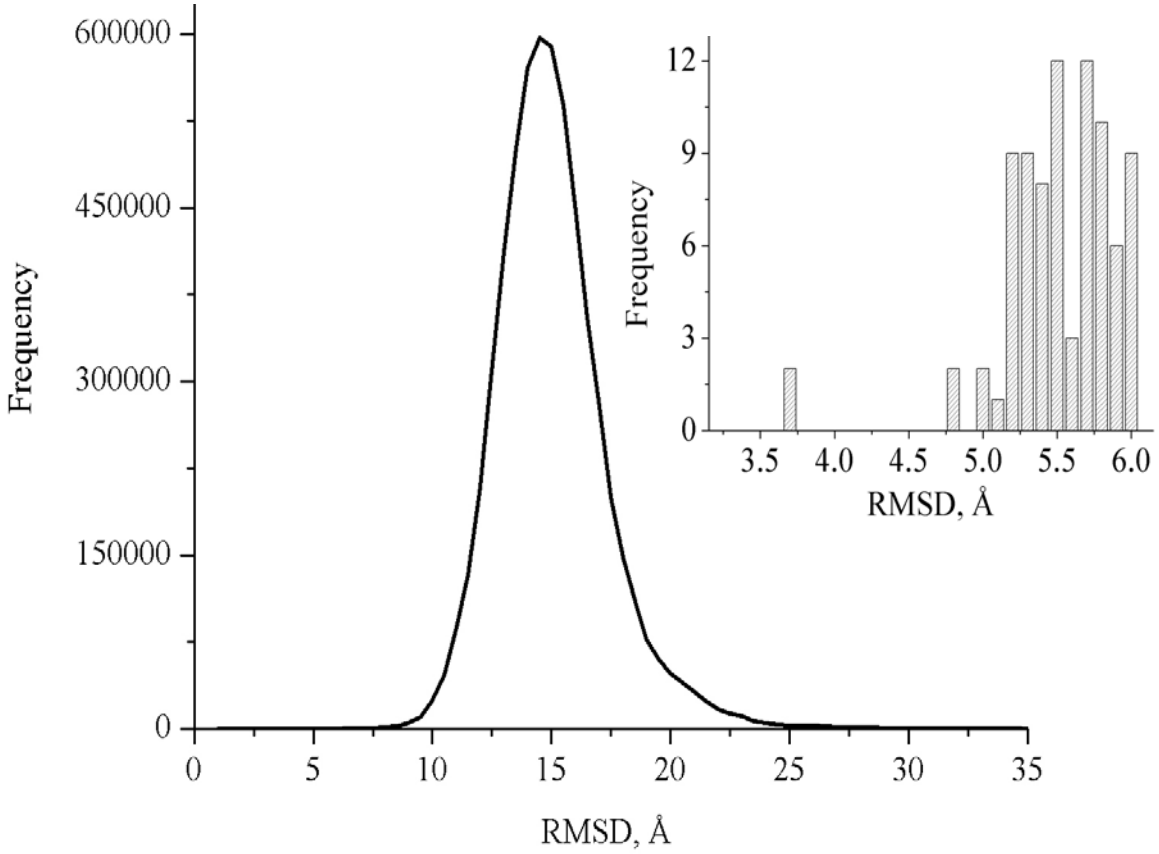


Figure 3-7. Histogram of RMSDs for all decoys in the database using 1b4c as a reference. The histogram of RMSDs for all decoys with RMSDs less than 7.0 Å is also included.

Figure 3-8. Decoys with the lowest RMSDs in database using 1b4c as a reference. A) 1b4c. B) 1m31-a-2. C) 1nsh-a-2. D) 1wlm-9. E) 1psr-a-1

Figure 3-9. Schematic diagram of how an insertion in a loop region can affect the search process. The red structure represents the native structure of our example target protein and the black structure represents a decoy.

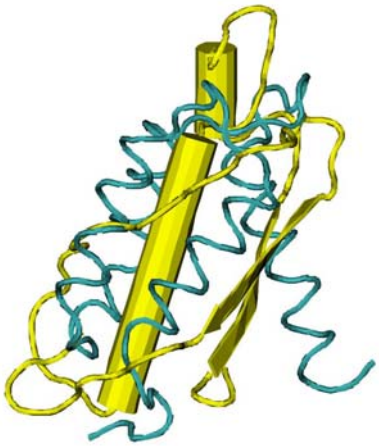


Figure 3-10. Decoy 1mka-49 (shown in yellow) satisfied many constraints for the 1b4c target (shown in blue).

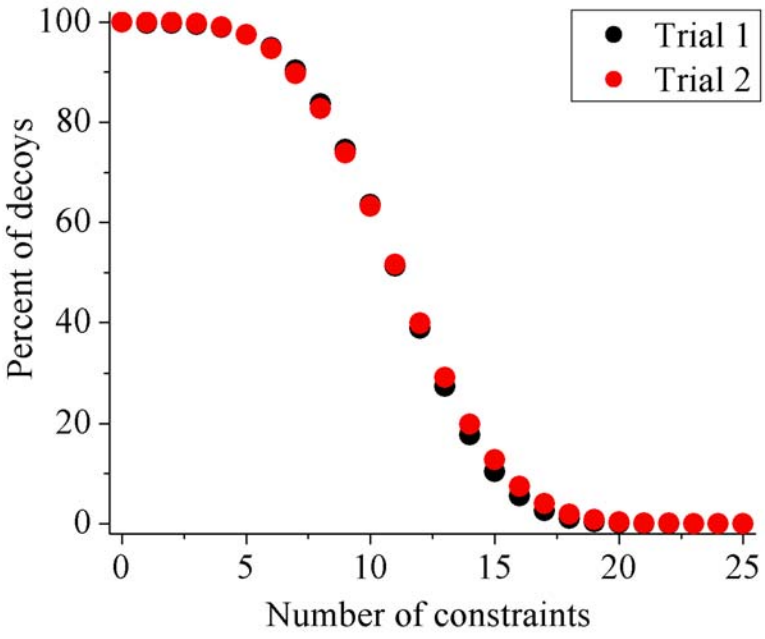


Figure 3-11. Graph of the number of decoys vs. the number of constraints each decoy satisfies for both trials. Fifty percent of the decoys satisfy 11 constraints.

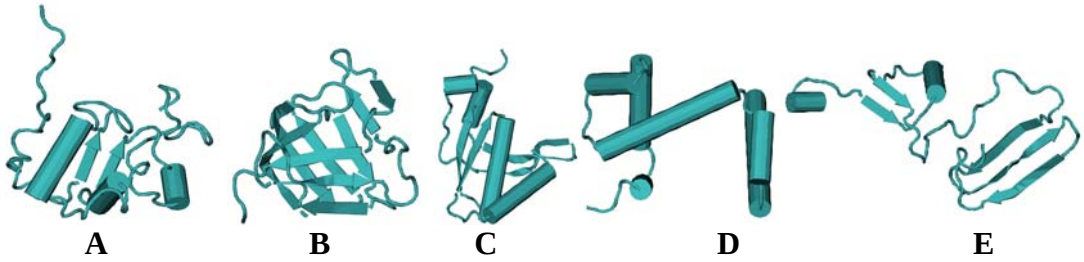


Figure 3-12. Five decoys used to determine a random average RMSD for our decoy database. A) 1b7u-109. B) 1fhx-13. C) 1ujn-75. D) 2wrp-15. E) 1rt6-109.

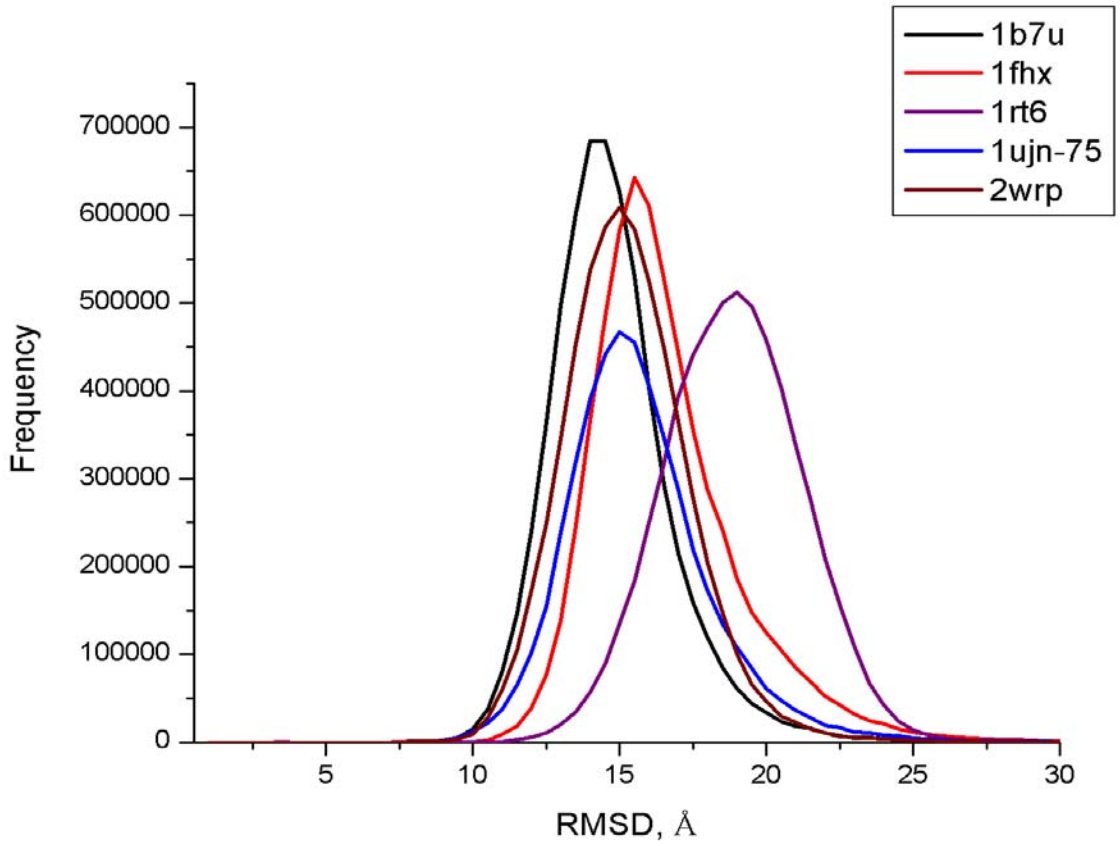


Figure 3-13. Histograms of RMSDs for five randomly chosen decoys, 1b7u, 1fxh, 1rt6, 1ujn, 2wrp.

CHAPTER 4
RESULTS: USING OUR DECOY SET TO FIND FOUR PROTEINS

We attempted to find the structures of four proteins using our database. The target proteins were PDB codes: 1b4c,[131] 1ghh,[144] 1ubi,[145] and 2ezk.[146] We chose these specific proteins because they were previously used to evaluate other methods.[143,147] All target proteins contain fewer than 100 residues and are therefore not explicitly included in the decoy set. Twenty-five distance constraints were chosen for each protein using the secondary structure prediction method JPred[115] (to avoid the loop regions). First we will evaluate the decoy generation method followed by an analysis of the decoy discrimination process.

4.1 Completeness of Decoy Set

Because we know the structures of the target proteins *a priori*, we can evaluate our decoy set by calculating RMSDs for all of the decoys using each target protein as a reference. The RMSD distributions (Figure 4-1) are similar for each of the targets and show that most of the decoys have RMSDs within 12 – 20 Å. We also targeted another five proteins and found a similar distribution. Because it is commonly assumed that a good structure prediction for a small protein is one with an RMSD lower than 6.0 ,[134] it will be difficult to find the few good decoys in the set.

Because the distributions are skewed gaussian, only a few decoys are expected to have RMSDs under 6.0 Å. Assuming a perfect gaussian distribution and using the standard deviation and mean RMSD for each target, we calculated the number of decoys expected to have RMSDs under the following cutoffs: 6 Å, 7 Å, 8 Å, and 9 Å (Table 4-1). Comparing the number extrapolated from a perfect gaussian to the number of decoys found within each RMSD cutoff in our decoy set, we find that the decoy set number is consistently much lower. It is harder to find low RMSD decoys than if the distribution was perfectly Gaussian.

4.2 Evaluation of Decoy Discrimination

4.2.1 Target 1b4c[131], Apo-S100β

Our first target protein is 1b4c (Figure 4-2), a homodimer of S100 beta subunits, each 92 residues in length. It has been classified as a metal-binding protein. Due to the nature of the database, the structure of only one chain was chosen as the target. Using 1b4c as a reference, we found 85 decoys with RMSDs less than 6 Å. Like 1b4c, 1m31 (apo-Mts1)[148] and 1nsh (apo-S100A11)[149] are both are metal binding proteins. PDB code 1psr[150] is the psoriasin protein while is 1wlm[151] is CGI-38 and currently has no known classification.

Four decoys were found to satisfy twenty-five constraints, which ranged in distance from 11.9 Å to 26.9 Å. They have the lowest RMSDs in the database using 1b4c as a reference (3.6 Å and 4.8 Å) and are all from the same parent protein, 1m31 (Figure 4-2). Hypothesis 1 and 2 were satisfied; a low RMSD decoy was in the database and this decoy shared a small set of similar distances with the target. Target 1b4c was studied by Meiler and Baker using a *de novo* protein structure prediction algorithm which employed Rosetta[143]. The 3.6 Å RMSD of our best decoy was slightly better than their best-scoring cluster which had an RMSD of 4.6 Å.

4.2.2 Target 1ghh,[144] DNA-Damage-Inducible protein I (DinI)

Our next target protein, 1ghh, is composed of 81 residues and can be found in Figure 4-3. Of the 8 million decoys in the database, 85 had RMSDs less than 6 Å with 1ghh as a reference. The RMSD distribution was very similar to that seen for 1b4c (Figure 4-1). The distance constraints ranged from 11.4 Å to 21.1 Å. Our method successfully identified the lowest RMSD decoys in the database. Eight decoys satisfied all twenty-five constraints; their structures and RMSDs are shown in Figure 4-3.

Because no attempt has been made to remove redundant structures from the database, some of the top scoring RMSD decoys come from the same parent proteins with different PDB codes.

For example, 1iwg,[152] 1oy6,[153] and 1t9u[154] represent acriflavine resistance protein B. ISHp608 transposase is represented by 2a6m and 2a6o.[155] PDB code 1vh2[156] is the autoinducer-2 synthesis protein.

Sequence homology using BLAST[33] was unable to find any structures similar to 1ghh in the PDB. Our method has an advantage in that we are able to generate low RMSD structures with little sequence homology. Often structural relationships are more conserved than sequence.[1,2,157,158] We found decoys with RMSDs as low as 4.9 Å which was very similar to the 4.8 Å RMSD value found by Meiler and Baker[143] for this protein.

The eight top scoring decoys each have three β-sheets and at least two α-helical regions. Despite the low RMSD values, the target protein has a pair of parallel β–sheets and a pair of anti-parallel β-sheets (see Figure 4-3) while the decoys have only anti-parallel β-sheet orientations. In all of these structures, the β-sheets have, as usual, distances of ~5 Å between α-carbons on adjacent strands. The small distance between the β-strands allows for a low RMSD between the overall structures despite an incorrect topology.

The study of this protein indicates that proteins with β-sheets may have low RMSD (~5 – 6 Å) decoys with incorrect topology. Our preliminary results on other proteins also show low RMSD decoys with various β-sheet orientations. For these types of proteins, RMSD alone may not be a useful indicator of a successful prediction.

4.2.3 Target 1ubi,[145] Ubiquitin

PDB code 1ubi represents the well studied ubiquitin protein (Figure 4-4). It is composed of 76 residues. The RMSDs of all the decoys were calculated using 1ubi as a reference and seven decoys were found to have RMSDs less than 6.0 Å.

The chosen distance constraints ranged from 7.3 Å – 19.5 Å. Two decoys from 1z2m[159] satisfied twenty-five constraints and can be found in Figure 4-4. Parent protein 1z2m is an interferon-induced ubiquitin-like protein and therefore not surprisingly similar to our target. The RMSDs for both decoys were 3.9 Å, the lowest RMSD decoys in the database using 1ubi as a reference. This RMSD value was similar to the top-scoring cluster found using Rosetta,[143] 3.4 Å.

4.2.4 Target 2ezk,[146] Mu End DNA-Binding ibeta Subdomain of Phage Mu Transposase

Target 2ezk has 93 residues. It was selected as our final target protein because Kihara et. al[147] used it to test their method and had some difficulty finding a low RMSD model. A BLAST search of this target found one other protein with sequence homology, 2ezl.[146] With only 93 residues, 2ezl was too small to be included in our database. RMSDs for all the decoys in the database were calculated and 41 decoys had RMSDs between 7.7 Å and 8.0 Å. No decoy in our database had an RMSD less than 6.0 Å; our database does not contain a good decoy. This 93 residue segment is not similar to any piece of a larger protein.

The distance constraints ranged from 12.0 to 18.5 Å. Nine decoys satisfied twenty-five constraints. Six decoys came from parent protein 1ngk[160] while 1v2a[161] was the parent protein for three decoys (Figure 4-5). The parent proteins seem to have functions unrelated to that of the target. Mycobacterium tuberculosis Hemoglobin O (1ngk) has possible functions in oxygen storage and transport, while 1v2a is a glutathione transferase isoenzyme. All of the decoys from 1v2a had RMSDs of 7.7Å using the target as a reference while the decoys from 1ngk had RMSDs of 11.6 – 11.7 Å.

4.2.5 Comparison of Search Process for All Target Proteins

For each target protein, half of the decoys satisfied at least 10 to 12 constraints (Figure 4-6A). All the search proteins show a similar Gaussian distribution of decoy scores (Figure 4-6B).

Most decoys satisfy at least one constraint but very few satisfy all twenty-five. The RMSD distribution for each protein (Figure 4-1) is similar in shape to Figure 4-6B, suggesting a relationship between the score (the number of satisfied constraints) and the RMSD. The low RMSD structures satisfy more constraints than those with high RMSDs. The strong correlation between RMSD and score is seen more clearly in Figure 4-7; low RMSD decoys have high scores, high RMSD decoys have low scores. Also, the average RMSD decreases with an increase in score. As seen in Figure 4-7, three of the target proteins have low RMSD decoys (< 6 Å) that satisfy all constraints. In general, decoys with scores between 10 and 15 have RMSDs between 15 and 20 Å, while decoys with a score of less than 10 have RMSDs greater than 25 Å.

For each target protein, there are a few decoy structures that have high RMSDs and high scores. These decoys generally span more than one domain giving them an unfolded and non-protein-like appearance. Often, one section of the decoy is similar in structure to a target protein thereby satisfying several constraints, while the large RMSD comes from the second section of the protein being so far from the first. In the PDB, occasionally multi-domain proteins are poorly labeled. For example, in 1xi5,[162] residue 838 and 839 are nearly 152 Å apart. Some of the high RMSD, high scoring decoys in this study came from parent proteins 1xi5 and 1xi4.[163]

4.3 Conclusions

We found that it is possible to search our decoy database using distance constraints to find reasonably accurate protein models with RMSDs less than 6 Å. A distance range of +/- 5 Å as the constraint acceptance criterion yields the best results. To avoid dependence on the order of application of constraints, we counted the total number of constraints that each decoy satisfied. Decoys that satisfied the most constraints systematically had the lowest RMSDs.

Our final results showed that 3 of the 4 target proteins had RMSDs less than 5 Å as summarized in Table 4-2. Even low resolution structures have been found to give insight into

the function of proteins.[105] Structures of this resolution can also be used as starting points in density generation for X-ray structures.[111] In each of these trials, twenty-five constraints were needed to eliminate all but a few representative structures. More studies must be performed before we can state with confidence that this number accurately represents the amount of distance information needed to determine structure. We also analyzed the RMSDs for several proteins and found that in general the average RMSD range for decoys in our database is ~15 Å. Like the PDB, our database contains many semi-redundant structures. Removal of such decoys may further decrease the search time of an already fast screening process.

Table 4-1. The number of decoys with RMSDs under each threshold

	Under 6Å		Under 7Å		Under 8Å		Under 9Å	
	real	hypothetical	real	hypothetical	real	hypothetical	real	hypothetical
1b4c	85	896	354	3,958	1,774	14,766	5,983	46,528
1ghh	43	2,885	208	10,082	1,088	30,486	7,308	79,759
1ubi	7	4,540	12	15,001	264	42,939	3,797	106,478
2ezk	0	876	0	3,690	41	13,276	2,182	40,787

There are 8,060,245 decoys in the set.

Table 4-2. Summary table of results

Target	Parent protein of found decoy	RMSD
1b4c	1m31	3.6
1ghh	2a6m	4.9
1ubi	1z2m	3.9
2ezk	1v2a	7.7

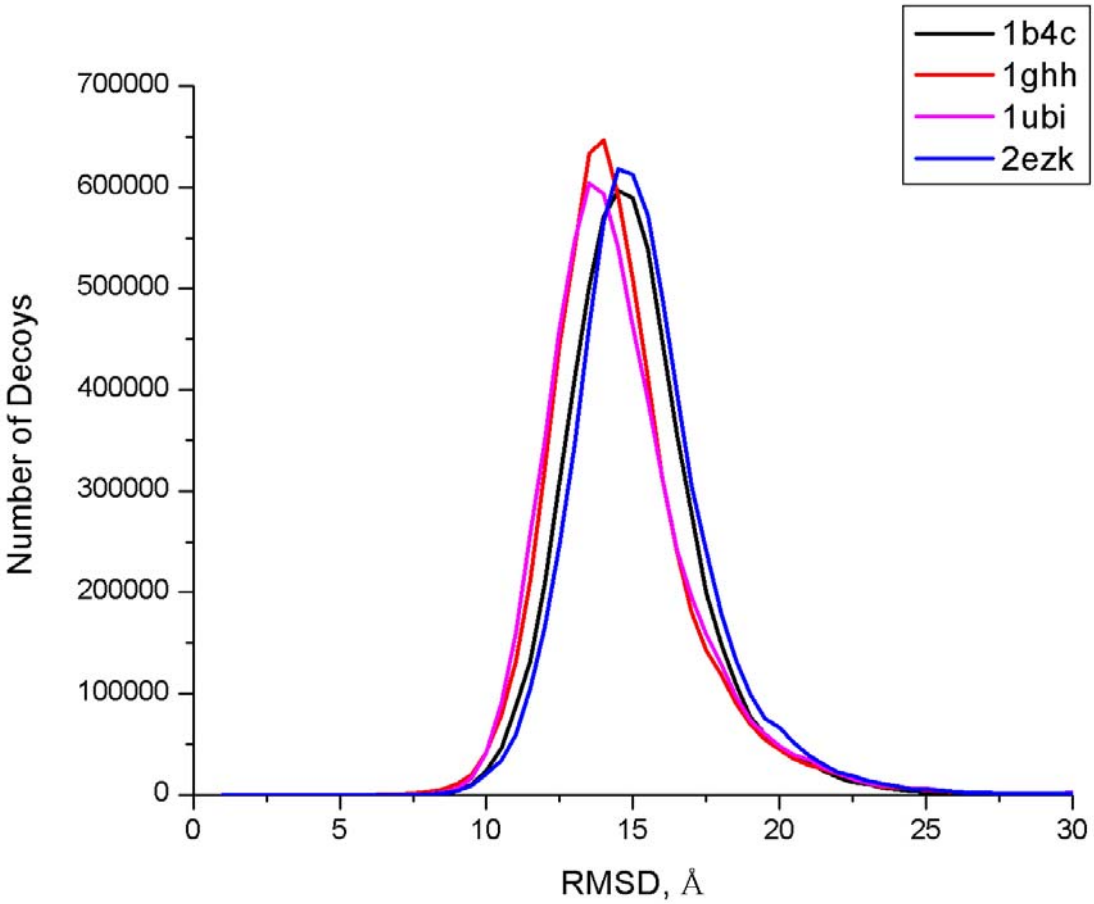


Figure 4-1. RMSD histograms for all studied proteins, 1ghh, 1ubi, 2ezk, and 1b4c

A B C

Figure 4-2. Decoys with the lowest RMSDs in database using 1b4c as a reference. Parent protein 1m31 has two chains, a and b. The first two decoys from each chain are the top scoring decoys with low RMSDs. A) 1b4c. B) 1m31-a-2 and 1m31-b-2. Each decoy had an RMSD of 3.6 Å. C) 1m31-a-1 and 1m31-b-1. Each decoy had an RMSD of 4.8 Å.

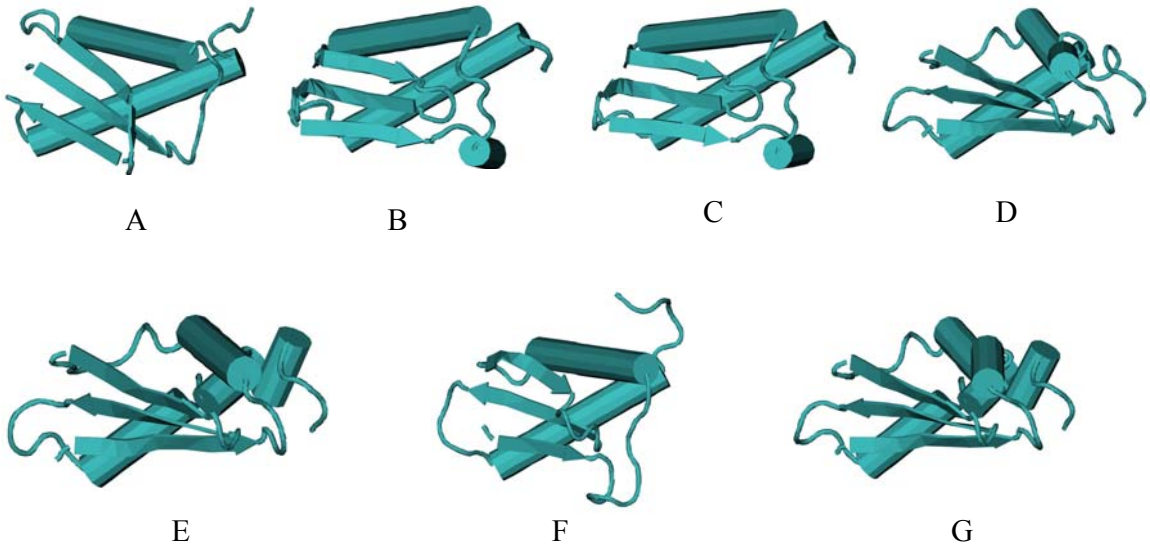


Figure 4-3. Target 1ghh and top scoring decoys. Decoys from the same parent proteins are shown together. A) Target 1ghh. B) 2a6m-1-25 and 2a6m-2-26, each with an RMSD of 4.9 Å. C) 2a6o-1-25 and 2a6o-2-25, each with an RMSD of 4.9 Å. D) 1oy6-1-45 has an RMSD of 5.3 Å. E) 1t9u-1-45 has an RMSD of 5.2 Å. F) 1vh2-1-42 has an RMSD of 5.1 Å. G) 1iwg-1-45 has an RMSD of 5.3 Å.

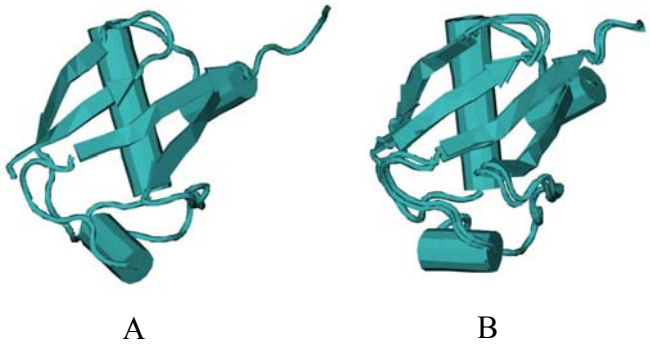


Figure 4-4. Target 1ubi and top scoring decoys. A) 1ubi. B) 1z2m-1 and 1z2m-2 both have an RMSD of 3.9 Å.

A B C

Figure 4-5. Target 2ezk and top scoring decoys. A) Three decoys from 1v2a (1v2a-a-75, 1v2a-b-75, 1v2a-c-75) and B) six decoys from 1ngk (1ngk-e-15, 1ngk-h-15, 1ngk-i-15, 1ngk-j-15, 1ngk-k-15, 1ngk-l-15) satisfied all constraints. The RMSD for decoys from 1v2a was 7.7 Å while those for 1ngk ranged from 11.6 – 11.7 Å.

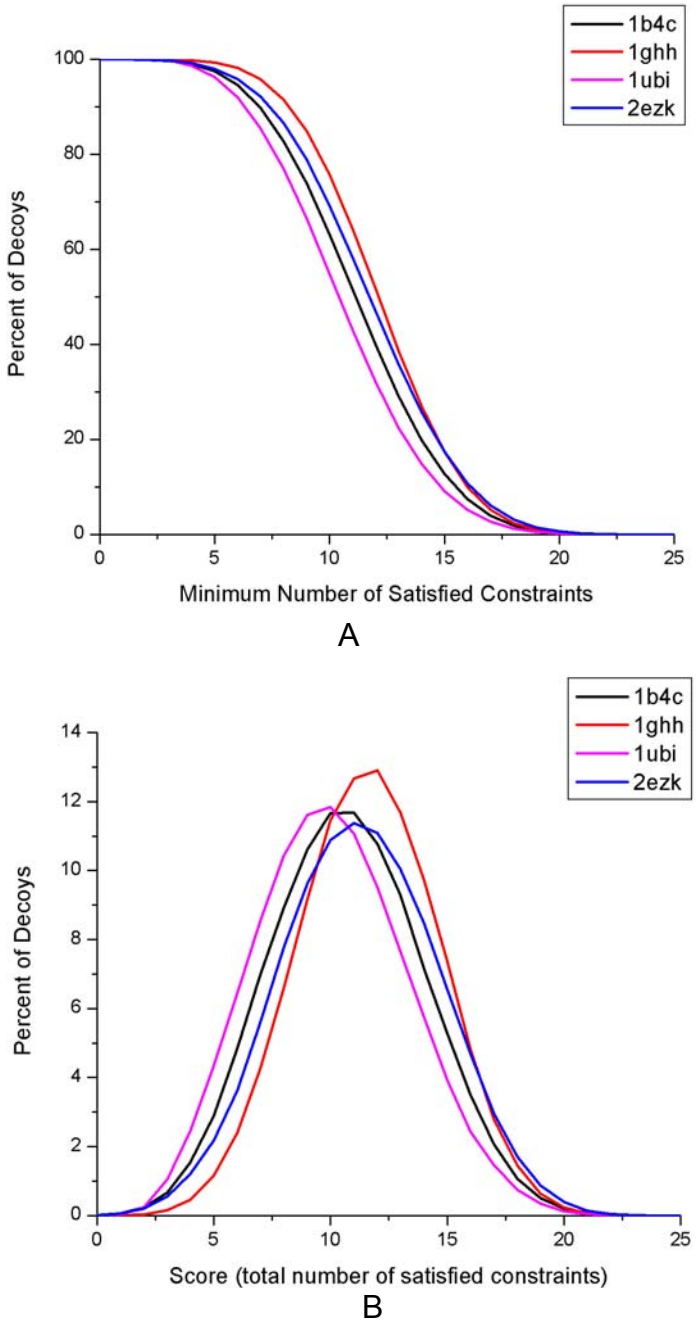


Figure 4-6. Analysis of the scoring procedure: A) the y-axis represents the percent of decoys satisfying at least a certain number of constraints. For example, 100 % of decoys satisfy 0 or more constraints, while fifty percent of the decoys satisfy at least 10 – 12 constraints. B) The y-axis represents the percent of decoys satisfying the exact number of constraints. Very few decoys satisfy exactly 0 or 25 constraints.

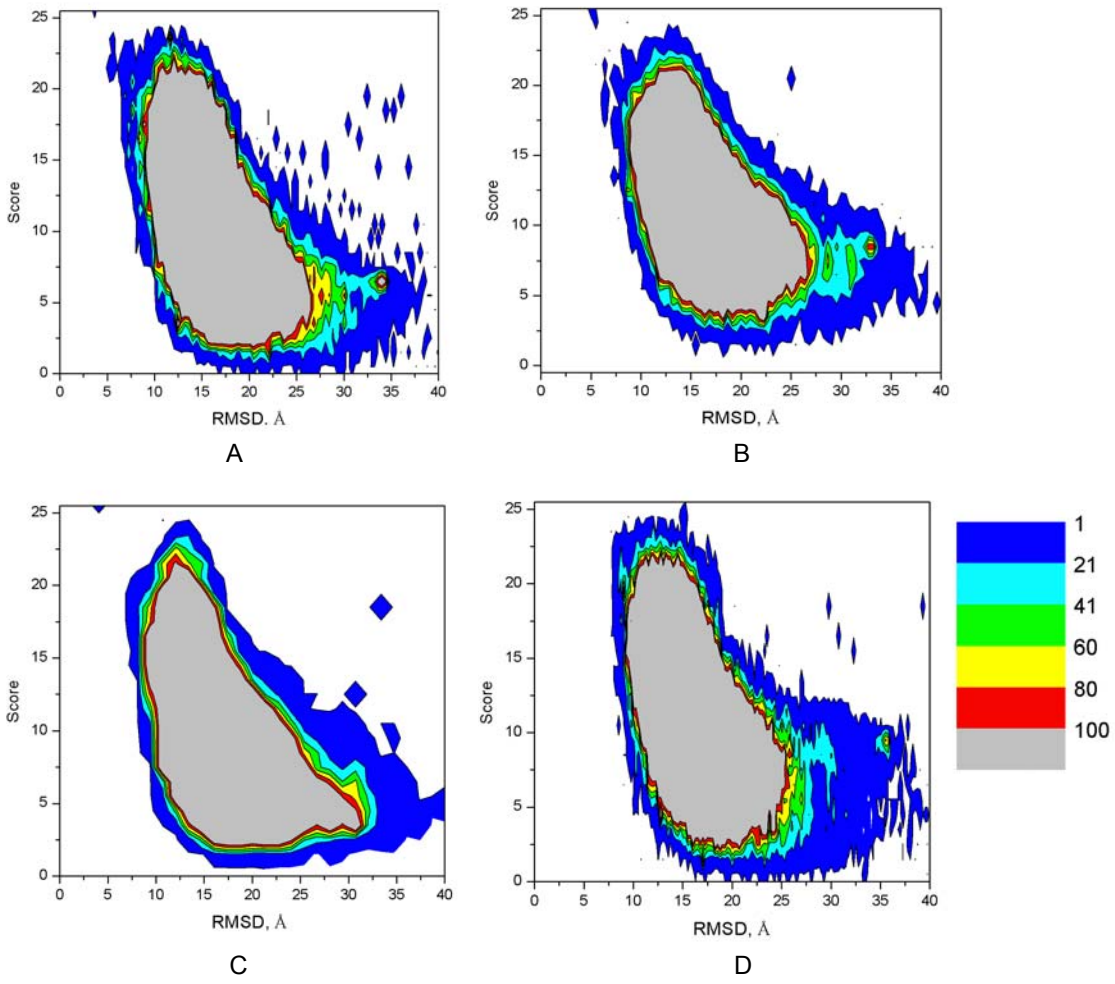


Figure 4-7. Figure showing the relationship between RMSD and score. A) 1b4c. B) 1ghh. C) 1ubi. D) 2ezk. In general, the low RMSD structures have high scores and the high RMSD structures have low scores.

CHAPTER 5
RESULTS: USING SPECIFIC DECOY SETS TO FIND FOUR PROTEINS

To use our decoy discrimination procedure with Rosetta-generated decoy structures, two parameters were optimized: (1) the number of decoys in the data set and (2) the constraint distance acceptance range. Once optimized, these parameters determined the number of distance constraints needed in the search process. We will first explain how the parameters were optimized and then discuss the search results using the optimized parameters for four proteins: 1b4c,[131] 1ghh,[144] 1ubi,[145] and 2ezk.[146]

5.1 Parameter Optimizations

5.1.1 Decoy Set Size

For each target protein, three sets were generated containing 1,000, 10,000, and 50,000 decoys. RMSDs were then calculated using the target protein as a reference. For a given protein, the RMSD distribution is relatively constant despite the number of decoys generated as shown for 1b4c in Figure 5-1. The distribution, however, varies from protein to protein. Analysis of the RMSD ranges in Table 5-1 reveals that slightly better decoys (lower RMSD structures) are generated in the 10,000 decoy set than in the 1,000 decoy set. Increasing the set to include 50,000 decoys, however, does not show a significant improvement in the quality of decoys generated to justify the extra computational cost associated with their generation. In all cases, increasing the set size generates slightly worse structures (ones with higher RMSDs) as well.

For 1b4c more than 40% of the decoys have low RMSDs (less than 6.0 Å). Increasing the set size from 1,000 to 10,000 decoys generates a slightly lower RMSD structure than the best decoy in the 1,000 set. In the largest decoy set, both the lowest and highest RMSD decoys can be found. More than a third of the decoys for 1ubi have low RMSDs in each of the three

different size sets. As seen for 1b4c, an increase in set size for 1ubi generates both lower and higher RMSD structures. The RMSD distributions for 1ghh and 2ezk are broad; only about 10% of the decoys have RMSDs less than 6.0 Å. For 1ghh, there is a slightly better RMSD structure found in the 10,000 decoy set than in the 1,000 while the 10,000 and 50,000 decoy sets have the same RMSD range. For 2ezk, there is no improvement in low RMSD structures by increasing the set size from 10,000 to 50,000. This increase does, however, generate decoys with slightly higher RMSDs.

For all of the target proteins, Rosetta generates structures with RMSDs of 3.6 Å or lower. We choose a decoy set size of 10,000 as a balance in terms of cost/performance ratio because it generates low RMSD structures relatively quickly. The lowest RMSD structure for each protein generated in the 10,000 decoy set is shown in Figure 5-2 superimposed on the native structure of the target protein.

5.1.2 Constraint Distance Acceptance Range

The upper and lower bounds placed on a distance constraint make up the constraint distance acceptance range. Such a range is needed in order to properly simulate experimental conditions in which the measured distances are not exact. A decoy with a calculated distance within the acceptance range, satisfies the constraint. We tested constraint distance acceptance ranges of +/- 1Å, 3Å, and 5Å for sets of twelve and twenty-five constraints.

5.1.2.1 Twelve constraints

Using the present constraint selection procedure, which incorporates information from a secondary structure prediction method, we chose a set of twelve distances from the native structures of each target protein. Each constraint met the following criteria: it involved atoms in defined regions of secondary structure and its length was between 5 and 30 Å.

A constraint distance acceptance range of +/- 1 Å is too tight for most of the target proteins. Fifty percent of decoys for 1b4c satisfy only four constraints (Figure 5-3A). The highest scoring decoy satisfied eleven constraints and had an RMSD of 5.1 Å (Table 5-2). For 1ghh, fifty percent of decoys satisfied six constraints and the highest scoring decoy (score of eleven) had an RMSD of 5.4 Å. The third protein, 1ubi, had two decoys with a score of twelve; their RMSDs were 3.0 Å and 4.1 Å. Although not the lowest in the database, the top-scoring decoys for 1b4c, 1ghh, and 1ubi had RMSDs in the range for good predictions (< 6.0 Å). Those for 2ezk, however, did not. The three top scoring decoys had a score of eleven and RMSDs ranging from 8.1 – 9.8 Å. An acceptance range of +/- 1 Å is far too restrictive for 2ezk, eliminating low RMSD structures. Because the constraint distance acceptance range failed to assign the highest scores to the lowest RMSD decoys, an increase in the acceptance range was necessary.

Increasing the constraint distance acceptance range from +/- 1 Å to +/- 3 Å shows some slight improvement in structure prediction; all four target proteins had a top scoring decoy with a RMSD of 4.6 Å or lower (see Table 5-2). Fifty percent of decoys satisfied nine to eleven constraints. Use of this range, however, results in too many high RMSD decoys satisfying all of the constraints. The top scoring decoys had high RMSDs ranging from 7.9 Å for 1b4c to 13.3 Å for 1ubi (Table 5-2). The total number of top scoring decoys is also higher than when using the lower acceptance range. More constraints must be used in order to employ a constraint distance acceptance range of +/- 3 Å.

Using a constraint distance acceptance range of +/- 5 Å, at least one of the top scoring decoys for 1ghh, 1ubi, and 2ezk also had the lowest RMSD in the set. This constraint range, however, has the same drawbacks as the +/- 3 Å range; 1ghh has nearly one thousand decoys that

satisfy twelve constraints while 1ubi has almost four thousand. Although 1ubi has over three thousand decoys with RMSDs under 6 Å, many of the top scoring decoys have larger RMSDs—as high as 14 Å. Using only twelve constraints in the search procedure does not adequately distinguish the good decoys from the bad. More constraints must be used.

5.1.2.2 Twenty-five constraints

A set of twenty-five constraints was chosen; twelve of which were taken from the previous constraint set. As was seen for twelve constraints, an acceptance range of +/- 1 Å is very tight—fifty percent of decoys satisfy ~6 constraints for 1b4c and ~10 – 12 constraints for the other three target proteins (Figure 5-3B). None of the target proteins had a decoy that satisfied all twenty-five constraints using an acceptance range of +/- 1 Å. Three of the four target proteins had top scoring decoys with RMSDs under 6 Å (Table 5-2) but the lowest RMSD decoy in each set was not assigned the highest score. Only 1ghh had top scoring decoys out of the range for reliable predictions.

Increasing the acceptance range from +/- 1 Å to +/- 3 Å improves predictions for 1ghh and 1ubi, but the RMSDs of the top scoring decoys are higher for 1b4c. Both low (4.3 Å) and high (8.2 Å) RMSD decoys for 2ezk satisfy all twenty-five constraints. For each of the target proteins, fifty percent of decoys satisfy 16 – 18 constraints using an acceptance range of +/- 3 Å and 21 – 23 constraints for an acceptance range of +/- 5 Å. For the latter acceptance range, all target proteins had at least one top scoring decoy with a low RMSD (< 6.0 Å). For 1ghh, 1ubi, and 2ezk, the lowest RMSD decoy in the set had a score of twenty-five. For 1b4c, an acceptance range of +/-5 Å had a top scoring decoy with the lowest RMSD when compared to the other acceptance ranges. In summary, using twenty-five constraints and a constraint distance acceptance range of +/- 5 Å works best for this type of decoy set.

5.2 Search Results

We have found the optimal parameters to be a set size of 10,000 decoys and a constraint distance acceptance range of +/- 5 Å with a set of twenty-five distance constraints. We will present results for the four target proteins using the optimized parameters.

Our scoring procedure is tested by the correlation between the decoy's score and its RMSD (Figure 5-4). A good scoring procedure assigns lower RMSD decoys higher scores. For 1b4c, the expected trend holds true, low RMSD structures have high scores. For 1ghh, 1ubi, and 2ezk, the trend is not detectable. This may be due to the large number of low RMSD structures generated for 1b4c compared to the other target proteins. The lack of a trend is also caused by Rosetta's ability to accurately reproduce local structure in most of its decoys as well as our use of several short distance constraints for the target proteins. Short distances give information about a protein's local structure, while large distances give clues about its overall structure. Rosetta does a great job in predicting secondary structure; most constraints between residues close one another in the chain are, therefore, satisfied by almost every decoy. In summary, the lack of a clear correlation between score and RMSD may be the result of poor constraint choices.

Another way to view our results is to look at the average RMSD for each score (Figure 5-5). For each target protein, the average RMSD decreased with an increase in score indicating the use of constraints to distinguish between good and bad decoys is effective. The decoy set for 1ghh has the highest average RMSD, 9.7 Å. The average RMSD for the 1ghh decoys satisfying twenty-five constraints is only 4.9 Å; many high RMSD decoys were eliminated by applying several distance constraints. The average RMSD for decoys of 1ubi was 7.6 Å and the average of those satisfying twenty-five constraints was lowered to 4.2 Å. The other two target proteins also show a decrease in the average RMSD upon constraint application, albeit less drastic; 2ezk

is lowered from 8.6 Å to 7.7 Å and 1b4c is lowered from 7.1 Å to 6.0 Å. For 2ezk and 1b4c, poor constraint choices may have hindered the decoy discrimination process.

5.2.1 Target 1b4c

All twenty-five constraints were satisfied by eight decoys of 1b4c with RMSDs ranging from 4.6 – 7.7 Å and an average RMSD of 6.0 Å. Some of these structures can be found in Figure 5-6. The top scoring decoys are very similar to each other; the greatest area of variation from the native structure can be found in the loop regions. Also, the first α-helix appears to be somewhat displaced in most of the decoys.

The lowest RMSD structure in the database (Figure 5-2A) has a score of 20. The five constraints this decoy did not satisfy involved residues in the first α-helix (residues 7 – 16). As can be seen in Figure 5-2A, helix 1 of the decoy is slightly displaced from helix 1 of the target protein. All of the constraint distances in the decoy were less than 7Å from the target distances of the native structure. Using a slightly larger constraint distance acceptance range would result in the lowest RMSD decoy satisfying all twenty-five constraints.

5.2.2 Target 1ghh

For 1ghh, four decoys satisfied twenty-five constraints with RMSDs ranging from 3.2 – 6.2 Å. These structures can be found in Figure 5-7. The lowest RMSD decoy in the decoy set satisfied all constraints. Our decoy discrimination procedure successfully identified the low RMSD decoys. The average RMSD in the decoy set was 9.7 Å and dropped to 4.9 Å for decoys satisfying all twenty-five constraints.

Of the top scoring decoys, only the lowest RMSD decoy had the same topology as the native structure of the target protein. There are three β-sheets in the native structure (Figure 5-6); β-1 is located between β-2 and β-3. β-1 is parallel to β-2 and anti-parallel to β-3. In the top

scoring decoys 1059 and 1073, β-2 is the middle β strand and is still parallel to β-1 as in the native structure. In decoy 9935, β-3 is the middle β strand and runs anti-parallel to both of the other strands. Because β-strands are within hydrogen bonding distance to each other, decoys with this secondary structure can have low RMSDs and incorrect topologies, as seen in our previous work (Chapter 4).

5.2.3 Target 1ubi

Eighty-six decoys of 1ubi satisfy twenty-five constraints with RMSDs ranging from 2.4 – 12.6 Å and an average of 4.2 Å. The best structure in the decoy set has an RMSD of 2.4 Å and was found to satisfy all constraints (Figure 5-2C). Only nine of the top scoring decoys had RMSDs greater than 6 Å. Decoy number 3631 (Figure 5-8) had an RMSD of 12.6 Å. This decoy shared similar topology with the target structure for the first fifty residues; the RMSD for this section was only 3.4 Å. Deviation from the target structure appears in a loop region after the decoy's fourth β-sheet.

In this set, 86 decoys were found to satisfy all constraints. As seen in the previous example (1ghh), structures with incorrect topology, inverted β-sheets for example, can sometimes satisfy several constraints. A slightly tighter constraint range of +/- 3 Å had only six decoys that satisfied 24 constraints with RMSDs ranging from 2.4 – 5.8 Å. A tighter constraint range may prevent such incorrectly aligned β-sheets from satisfying so many constraints.

5.2.4 Target 2ezk

For 2ezk, 732 decoys satisfied twenty-five constraints with RMSDs ranging from 2.9 – 14.4 Å with an average of 7.7 Å. Although the lowest RMSD decoy in the set (Figure 5-2D) satisfies all constraints, the RMSD range for the top scoring decoys is similar to that of the whole decoy set. The search method did not adequately distinguish between the good and bad decoys.

This may be because the decoy set contains only a small number of structures with RMSDs lower than 6 Å. Choosing better distance constraints may also improve the discrimination process.

A tighter constraint range of +/- 3 Å led to only eight decoys satisfying all twenty-five constraints (Figure 5-9A). The RMSDs, however, range from 4.3 – 8.2 Å. Excluding residues 1 – 10 lowers the RMSD range of the top scoring decoys to 2.4 – 4.0 Å, indicating this is the greatest region of deviation from native structure. Constraints were not chosen from this region because JPred did not predict any defined secondary structure.

5.3 Conclusions

Using our present method of choosing constraints, twenty-five distances must be measured to distinguish between reliable and unreliable decoys using a constraint distance acceptance range of +/- 5 Å. Decoys with slightly lower RMSDs are generated in the 10,000 decoy set when compared to the 1,000 set. In general, there is no significant difference between the decoys generated in the 10,000 versus the 50,000 decoy set. Rosetta generates low RMSD structures for each of our target proteins and our scoring procedure is effective in assigning these low RMSD decoys high scores. The RMSDs of the best top scoring decoys were: 4.6 Å for 1b4c, 3.2 for 1ghh, 2.4 Å for 1ubi, and 2.9 Å for 2ezk. For 1ubi and 2ezk, several decoys satisfied all twenty-five constraints with a large range of RMSD values. A different set of constraints may be more effective in distinguishing between good and bad decoys. In our next study, we will use a more reliable method for choosing constraints.

Table 5-1. RMSD ranges

	1,000 (% < 6.0 Å)	10,000 (% < 6.0 Å)	50,000(% < 6.0 Å)
1b4c	4.2 – 16.3 (42.7)	3.6 – 17.2 (45.2)	3.4 – 17.7 (44.6)
1ghh	4.2 – 14.6 (9.8)	3.2 – 16.8 (9.5)	3.2 – 16.8 (10.3)
1ubi	2.9 – 14.9 (32.7)	2.4 – 15.0 (34.6)	1.8 – 18.5 (34.9)
2ezk	3.7 – 15.0 (4.9)	2.9 – 15.2 (7.8)	2.9 – 17.9 (7.8)

In parentheses is the percentage of reliable structures generated for each decoy set for each target protein.

Table 5-2. Comparison of scores for each protein with different acceptance ranges

		1b4c		1ghh	
		score	Rmsd range*	score	Rmsd range
12 constraints	+/- 1 Å	11	5.1 (1)	11	5.4 (1)
	+/- 3 Å	12	4.6 – 7.9 (82)	12	4.1 – 12.2 (41)
	+/- 5 Å	12	3.8 - 14.3 (1343)	12	3.2 – 13.1 (998)
25 constraints	+/- 1 Å	13	4.8 – 5.8 (4)	17	6.1 – 12.1 (4)
	+/- 3 Å	22	5.7 (1)	23	4.6 – 5.4 (3)
	+/- 5 Å	25	4.6 – 7.7 (8)	25	3.2 – 6.2 (4)
Lowest RMSD in decoy set		3.6		3.2	
		1ubi		2ezk	
		score	Rmsd range	score	Rmsd range
12 constraints	+/- 1 Å	12	3.0 – 4.1 (2)	11	8.1 – 9.8 (3)
	+/- 3 Å	12	2.4 - 13.3 (518)	12	3.6 – 12.6 (668)
	+/- 5 Å	12	2.4 - 14.0 (3929)	12	2.9 – 14.3 (3207)
25 constraints	+/- 1 Å	19	4.1 (1)	18	5.9 (1)
	+/- 3 Å	24	2.4 – 4.1 (6)	25	4.3 – 8.2 (8)
	+/- 5 Å	25	2.4 – 12.6 (86)	25	2.9-14.3 (732)
Lowest RMSD in decoy set		2.4		2.9	

*the number in parenthesis is the number of decoys with that particular score. All data is for the 10,000 decoy set.

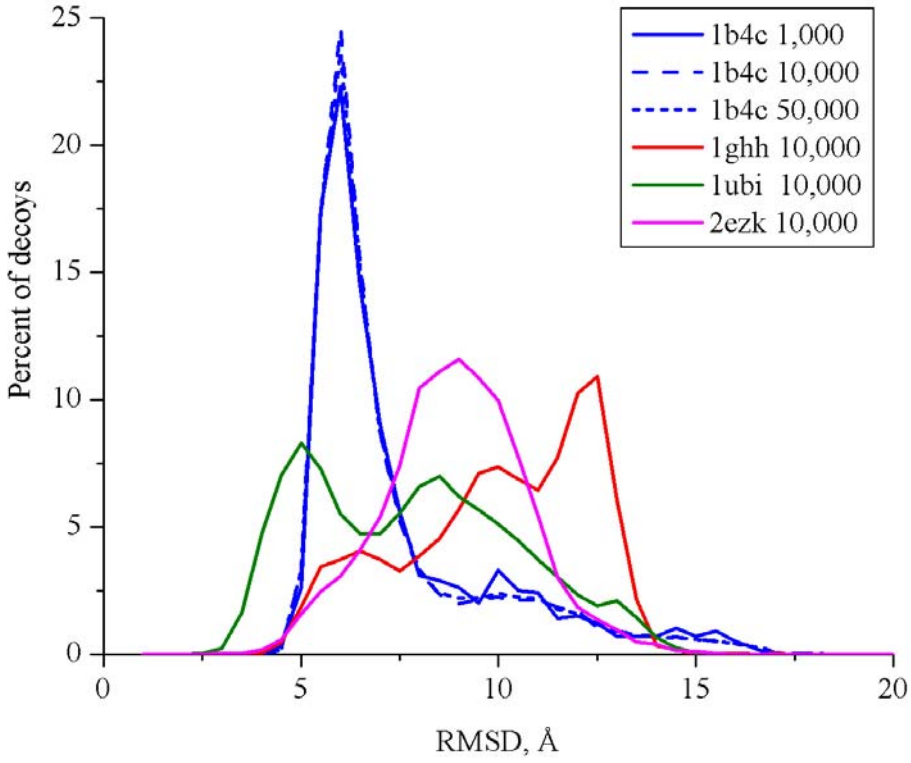


Figure 5-1. RMSD distributions for all four target proteins using the 10,000 decoy sets. For 1b4c, the RMSD distribution for sets of 1,000 and 50,000 decoys are also shown. The bin size was 0.5 Å. The frequency was calculated as a percentage of the total number of decoys in the set.

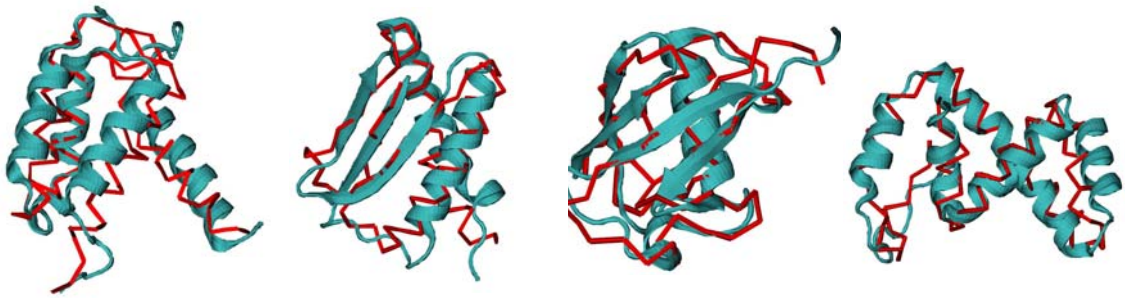
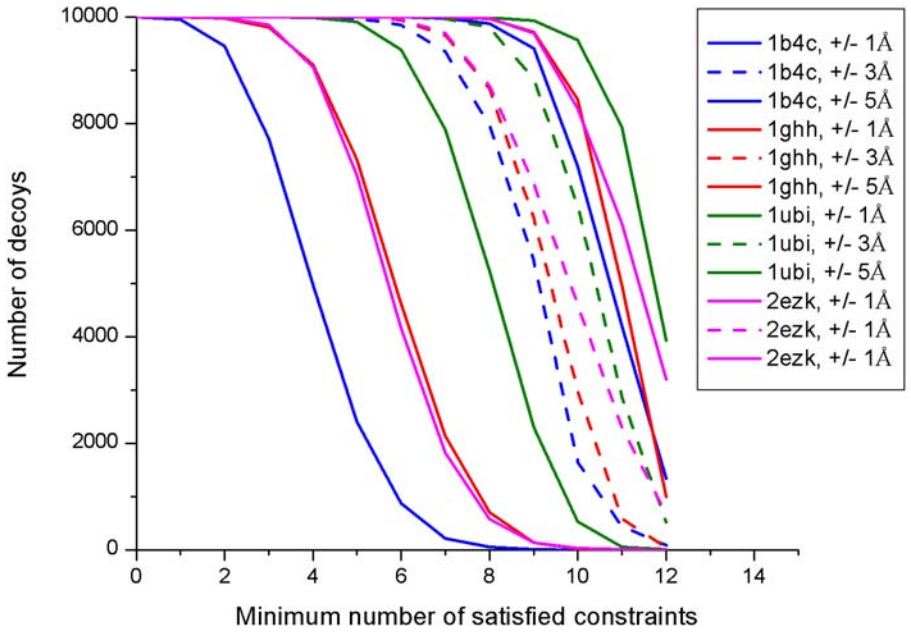
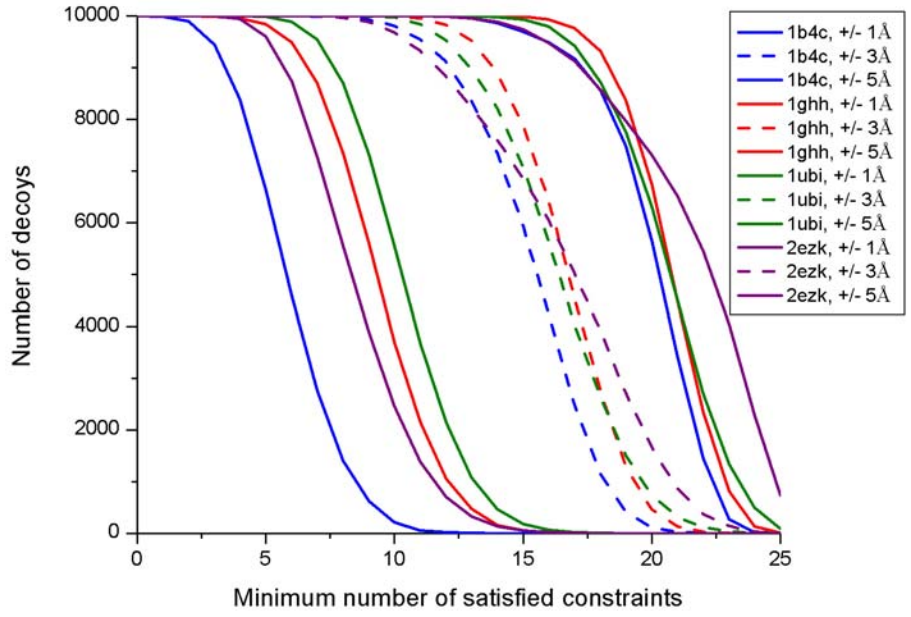


Figure 5-2. Lowest RMSD structures in the 10,000 decoy set. The target protein is shown in blue and the decoy structure is overlapping in red. A) 1b4c with decoy #6426. The RMSD is 3.6 Å. B) 1ghh with decoy # 6104. The RMSD is 3.2 Å. C) 1ubi with decoy # 5423. The RMSD is 2.4 Å. D) 2ezk with decoy # 5532. The RMSD is 2.9 Å.

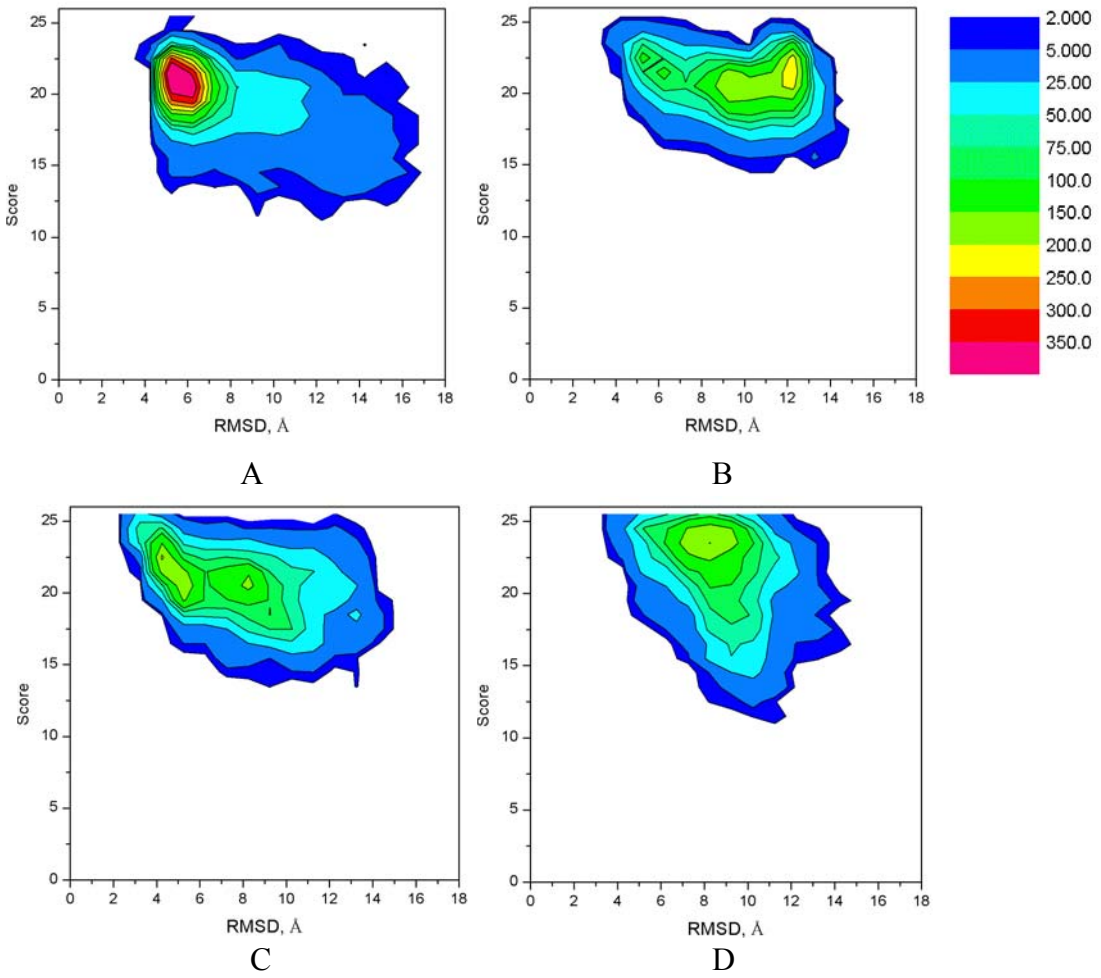


A



B

Figure 5-3. The number of structures remaining vs. score for each protein, for +/-1, 3, and 5 Å using A) twelve constraints and B) twenty-five constraints



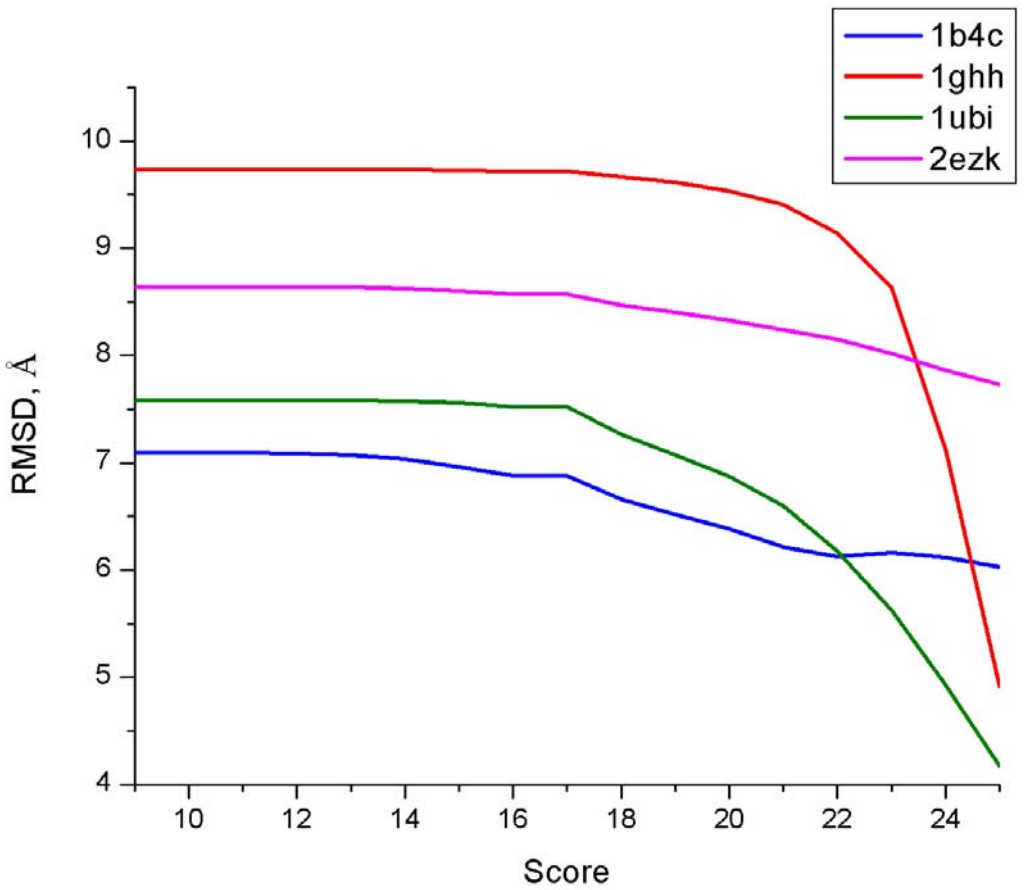
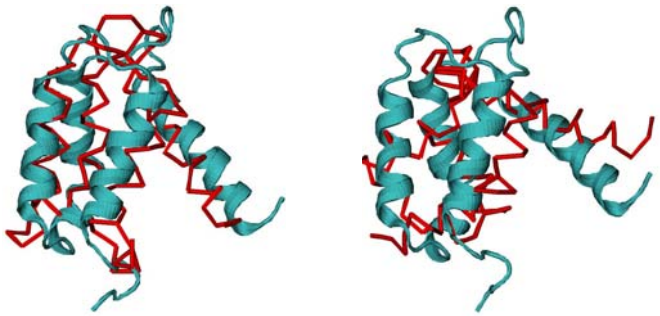


Figure 5-4. Correlation between score and RMSD. A) 1b4c. B) 1ghh. C) 1ubi. D) 2ezk.

Figure 5-5. Average RMSD for each protein at different scores.



A B

Figure 5-6. Top scoring decoys for 1b4c's 10,000 decoy set: A) Decoy #8500 has an RMSD of 4.6 Å. B) Decoy #8827 has an RMSD of 7.7 Å.

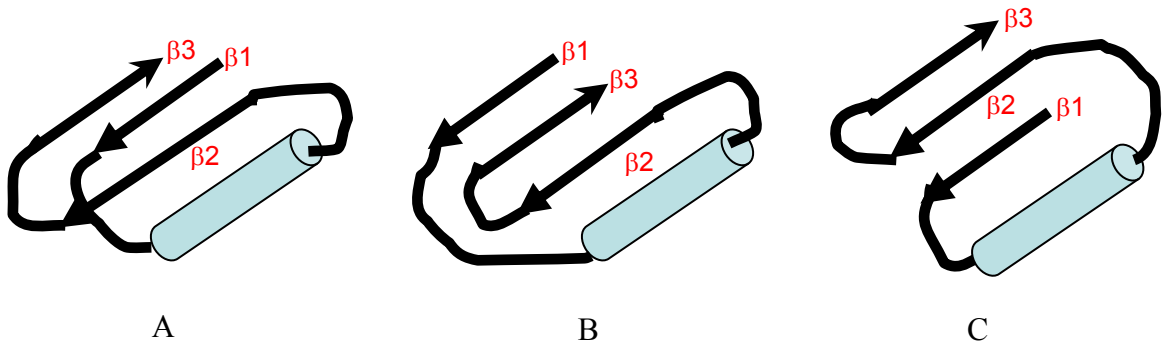


Figure 5-7. Representation of the β-sheet orientation for the native structure of target protein 1ghh and the top scoring decoys. A) Orientation for the native structure of 1ghh and decoy # 6104 with an RMSD of 3.2 Å. B) Orientation for decoy # 1059 and # 1073 with RMSDs of 5.8 Å and 6.2 Å respectively. C) Orientation for decoy # 9935, which has an RMSD of 4.5 Å.

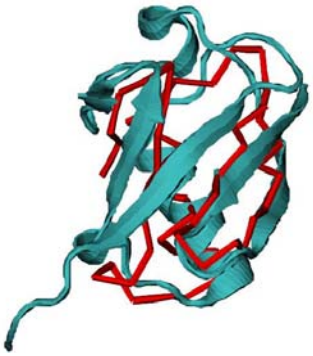


Figure 5-8. Top scoring decoy, for 1ubi, # 3631 with a high RMSD, 12.6 Å. When residues 1 – 50 aligned to the native structure of 1ubi the RMSD is 3.4 Å.

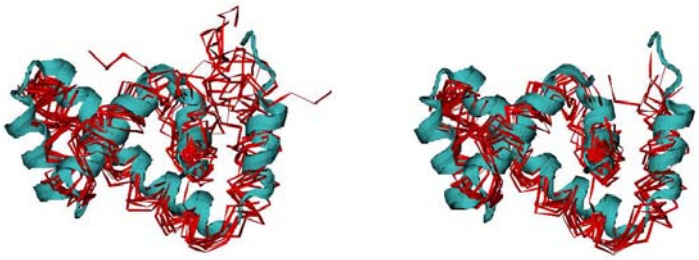


Figure 5-9. Top scoring decoys for 2ezk. A) When all residues are aligned the RMSDs range from 4.3 – 8.2 Å. B) When residues 10 – 93 are aligned, the RMSDs range from 2.4 – 4.0 Å.

CHAPTER 6
RESULTS: USING GENERAL AND SPECIFIC DECOYS SETS TO STUDY TWELVE CASP7 TARGETS

We used our general and specific decoy sets to predict the structures of twelve CASP7 targets. For a given target, the same set of twenty-five constraints was used for both types of decoy sets. Unless otherwise indicated, a constraint distance acceptance range of +/- 5 Å was employed.

6.1 General Decoy Set

For each target, twenty-five constraints were chosen. As seen in Table 6-1, the number of top scoring decoys for each target ranged from two to over ten thousand. The C^{α} RMSDs of the top scoring decoys were calculated and five targets were found to have successful predictions (a decoy with an RMSD under 6.0 Å). To determine whether the lack of reliable predictions for the remaining seven targets was due to a breakdown in decoy generation or decoy discrimination, we calculated the C^{α} RMSDs between each decoy and each target (Figure 6-1).

The RMSD distribution is similar for most of the target proteins; most decoys have RMSDs within 10 – 20 Å of their target with an average RMSD of ~16 Å. Target T335 is the exception. Its RMSD distribution is shifted to the left giving rise to an average RMSD of only 9.9 Å. Over 160 thousand decoys have RMSDs less than 6.0 Å, low enough to be considered a reliable prediction. It is not surprising, therefore, to find over ten thousand decoys satisfying all twenty-five constraints for this small target (42 residues), which also has a very common structural motif.

For target T335 and four others (T288, T309, T340, T359), the best decoy in the set had an RMSD under 6.0 Å and satisfied all constraints. For three targets, T348, T349, and T358, low

RMSD decoys were generated in the set but the discrimination procedure failed to assign them top scores. The remaining four targets (T306, T311, T353, T363), had no low RMSD decoys in the set; the decoy generation method failed to provide accurate structures, indicating no larger proteins contain pieces that look like these four targets.

Comparisons between the JPred predictions and the real structure for each target can be found in Table 6-2. Because constraints are chosen from the JPred prediction, it is important to determine the prediction's accuracy. Poor structure prediction can lead to poor constraint choices.

6.1.1 Targets That Worked

For five CASP7 targets (T288, T309, T335, T340, T359), the lowest RMSD decoy in the set was found to satisfy all twenty-five distance constraints. In each case, the lowest RMSD decoy was also under 6.0 Å.

6.1.1.1 Target T288

Target T288 corresponds to 2gzv, the PDZ domain of human PICK1 (a fragment of a PRKCA-binding protein). The PDB structure is missing two residues (27 and 28) located in a loop region, making the target 91 residues long. Overall JPred does a good job predicting the secondary structure as seen in Table 6-2. Although it does not predict the α-helix between residues 41 – 45 or the small β-strand between residues 60 – 61, its predictions for the rest of the structure are never off by more than two residues.

The α-helix from residue 67 to 76 was selected as a reference and all distance constraints involved an atom from this helix. Eleven decoys satisfied twenty-five constraints with RMSDs ranging from 3.4 – 5.6 Å (Figure 6-2); all of the top scoring decoys had RMSDs within the range of reliable structure predictions.

The six decoys with the lowest RMSDs in the set satisfied all constraints. All of the top scoring decoys are from the PDZ domain of various proteins. PDB codes 1tp3, 1tp5, 1tq3, 1be9, and 1bfe are crystal structures of the same protein, the PDZ3 domain of synaptic PSD-95 protein, complexed with different ligands. Parent protein 1um7 is the PDZ domain of synaptic-assoicated protein 102 while 1b8q is the extended neuronal nitric oxide synthase PDZ domain. The greatest difference between the structures of the decoys and the target occurs at the C-terminus from residues 87 – 91; no constraints involved atoms in this region.

In addition to RMSD, determining the longest continuous segment (LCS) of a decoy structure that has an RMSD under a specific threshold is sometimes used to evaluate the similarity between two structures (see Chapter 2). The longest continuous segment under 5 Å (LCS-5) is 91 residues for the 1tq3 decoy (the entire structure) and 89 residues for the 1b8q decoy. When a lower threshold is used, greater differences appear between the structures. For the decoy from 1tq3, the longest continuous segment under 2 Å (LCS-2) is composed of 15 – 86 (72 residues total), while the LCS-2 for the 1b8q decoy has only 29, from residues 33 – 61.

6.1.1.2 Target T340

Target T340 represents 2HE4, the second PDZ domain of human NHERF-2 (SLC9A3R2) interacting with a mode 1 PDZ binding motif. It is composed of 90 residues. The JPred prediction (Table 6-2) is fairly accurate for this target. It does not, however, predict the α-helix between residues 40 – 44 or the β-strand between residues 58 – 59.

The helix between residues 64 – 73 was selected as the reference structure from which all constraints were chosen. Sixty decoys satisfied twenty-five constraints with RMSDs ranging from 3.0 – 14.0 Å. The top scoring decoys with the highest and lowest RMSDs are shown in Figure 6-3. The best decoy in the database had an RMSD less than 6.0 Å, as did over 80% of the

top scoring decoys. Twenty of the top scoring decoys, with RMSDs ranging from 3.0 – 4.4 Å, came from parent protein 1wf7 (Figure 6-3B) while 1wif (Figure 6-3C) was the parent protein of sixteen top scoring decoys with RMSDs ranging from 4.4 – 4.7 Å. Both parent proteins are PDZ domains of a larger protein. For these top scoring decoys, the longest continuous segments under 5 Å included all residues; the LCS-2 was 63 residues for 1wf7 and 39 for 1wif.

The top scoring decoy with the highest RMSD is from parent protein 1qln (Figure 6-3D), the structure of transcribing T7 RNS polymerase initiation complex. It has no noticeable similarities to the target structure as it is completely α-helical while the target is mostly a β-barrel. The longest continuous segment of this decoy with an RMSD under 5 Å is 25 residues, which included the reference helix. Most of the other top scoring decoys with high RMSDs have a similar α-helical arrangement and the same LCS-5. Tightening the constraint distance acceptance range to +/- 3 Å only lowers the score of 1qln from 25 to 18. It is, therefore, poor constraint choices that must be contributing to the high RMSD structure predictions for this target.

An alternate set of twenty-five constraints was chosen by selecting eight atoms from various regions of secondary structure and calculating the distances between them. Twenty-three decoys satisfied all twenty-five constraints with RMSDs ranging from 3.0 – 4.4 Å. This set of constraints was much better than the previous set, which led to much better results. Among the top scoring structures were twenty decoys from parent protein 1wf7 with RMSDs ranging from 3.0 – 4.4 Å, as well as decoys from parent proteins 1uf1, 1uit, and 1v5l (Figure 6-3E, F, G). The parent proteins of all the top scoring decoys are structures of the PDZ domains of various proteins similar to target T340.

6.1.1.3 Target T359

Target T359 represents 2iwn, the 3rd PDZ domain of multiple PDZ domain protein MPDZ. The gap in the crystal structure between residues 28 – 31 resulted in the protein having a total of 93 residues. JPred does a good job predicting secondary structure for this protein. It does, however, miss an α-helix between residues 45 – 49 and a β-strand between residues 64 – 65 (Table 6-2). Also, it predicts the first, second, and last β-sheets to be slightly longer than they are in the real structure.

The helix between residues 71 – 79 was selected as a reference and all constraints involved an atom from this helix. Fifteen decoys were found to satisfy all twenty-five constraints with RMSDs ranging from 3.6 – 13.1 Å (Figure 6-4). The lowest RMSD decoys in the database satisfied all constraints. The parent proteins of six of the top scoring decoys (PDB codes: 1p1d, 1um1, 1il6, 1ueq, 1wfv) were PDZ domains of various proteins and had RMSDs of 6.8 Å or less. These decoys differed from each other only slightly, mostly in the loop regions. Despite having a lower RMSD, 1p1d (3.6 Å RMSD) had 55 residues in its longest continuous chain with a 2 Å threshold while 1wfv (6.8 Å RMSD) had 60 residues in its LCS-2.

The remaining nine top scoring decoys were redundant structures of parent protein 1w5e and had an RMSD of 13.1 Å. A tighter constraint distance acceptance range of +/- 3 Å was applied and these high RMSD decoys were found to satisfy only 14 constraints. Tightening the constraint distance acceptance range, however, also lowered the score of the lowest RMSD decoy in the decoy set from 25 to 20. So in this case, little is gained by changing the acceptance range; better constraints must be chosen.

Decoys of 1w5e have certain secondary structural elements in just the right places allowing them to satisfy many constraints. For example, twenty-eight residues comprised the LCS-5 for

this decoy, which included the α-helix that was used as a reference in choosing the constraints. Many constraints involved distances between an α-helix and a β-sheet so it is not surprising that all the top scoring decoys had these structural features.

6.1.1.4 Target T309

Target T309 corresponds to 2h4o, a 62 residue hypothetical protein from bacillus subtilis (yonk). JPred predicts an α-helix between residues 27 – 40. In the real structure, however, the α-helix is only between residues 34 – 40. The other regions of secondary structure are β-sheets located between residues 13 – 15, 20 – 24, and 29 – 33, none of which were predicted by JPred. Because of the discrepancies in the secondary structure prediction and the inherent lack of structure in the target, it was difficult to choose good constraints. Surprisingly, we were able to obtain successful predictions for this target.

Because most of the protein is unstructured, constraints were chosen between all regions of predicted secondary structure. Nine decoys satisfied twenty-four constraints. The RMSDs of the top scoring decoys ranged from 5.7 – 14.4 Å and came from four parent proteins: 1esc, 1esd, 1ese, and 1wk1 (Figure 6-5). The first three codes represent esterase and their top scoring decoys had RMSDs ranging from 5.7 – 6.6 Å. The longest continuous segment under 5 Å RMSD is from residues 1 – 58 (almost the whole structure) for 1esc, 1esd, 1ese and between residues 3 – 42 for 1wk1.

For 1wk1, the Lectin C-type domain derived from a hypothetical protein from C. elegans, the top scoring decoys had RMSDs of 14.1 – 14.4 Å. The high RMSD structures have some similarities to the target protein; the decoy's three β-sheets within residues 10 – 34 align closely to those of the target protein. As can be seen in Figure 6-5, the target protein does not have defined secondary structure between residues 1 – 10 or 42 – 62. The RMSD between the two

structures from residues 10 – 34 is ~3.8 Å, and, as mentioned above, its LCS-5 is composed of 40 residues, including this region.

Removing the distance constraint between residues 5 and 35 from the constraint set results in only three decoys (from parent protein 1esc) satisfying all twenty-four constraints. The three decoys each have an RMSD of 5.7 Å. The results are, therefore, significantly improved when bad constraints are not used. This particular constraint is not effective because residue 5 is in a region of undefined secondary structure. Poor secondary structure predictions can lead to poor constraints choices which can significantly hinder the performance of our method.

6.1.1.5 Target T335

Target T335 is 2hep, the UPF0291 protein ynzC from Bacillus subtilis. The target sequence was made of 85 residues but only 42 appear in the NMR structure. JPred correctly predicts the two main α-helices of this protein.

Twenty-five constraints were selected from eight residues, four on each α-helix. Over ten thousand decoys satisfied all twenty-five constraints (Figure 6-6); their RMSDs ranged from 2.1 – 9.5 Å. Over eight thousand of the top scoring decoys had RMSDs less than 6.0 Å. Target T335 is the smallest protein in this study and contains a fairly common structural motif. The lowest RMSD decoy in the set, from parent protein 1qsp, had all but two residues in its longest continuous segment with an RMSD under 2 Å.

Because of their compact structures, some higher RMSD decoys satisfied all twenty-five constraints. For example, a decoy from parent protein 1b6c (Figure 6-6C) had an RMSD of 9.5 Å and a LCS-5 of 30 residues. This decoy just barely satisfied all the constraints; using a constraint distance acceptance range of +/- 3.0 Å, the 1b6c decoy satisfied only eleven of the twenty-five constraints. Using a tighter acceptance range, 142 decoys satisfied all constraints;

their RMSDs ranged from 2.1 – 7.6 Å and 82% of them had RMSDs under 6.0 Å. For this target, tightening the constraint distance acceptance range significantly improved the results.

6.1.1.6 CASP comparisons

We employed the Global Distance Test (GDT) analysis (see Chapter 2 for details) to evaluate the performance of our method compared to other methods used in CASP7. Most methods were successful in predicting low RMSD structures for targets T288, T340, and T359 (Figure 6-7). Our predictions were not among the best for these targets even though their RMSDs were less than 6.0 Å.

Target T309 was a difficult target; it is largely unstructured which is a common pitfall for many methods. Our predictions for this target are much better those of other methods. Even our high RMSD top scoring decoy (cyan line in Figure 6-7) showed better GDT results than the average prediction; over fifty percent of the residues in this decoy satisfied a distance cutoff of 6.0 Å.

CASP results are mixed for target T335; the GDT results are highly scattered. Our top scoring decoy, with an RMSD of 2.1 Å, is one of the best predictions while our other top scoring decoy, with an RMSD of 9.5 Å, is average.

6.1.2 Targets That Could Have Worked But Did Not

For three targets, low RMSD decoys were generated but the discrimination procedure did not assign them top scores. The breakdown in decoy discrimination can be explained by various reasons as discussed in the next section.

6.1.2.1 Target T348

Target T348 represents 2hf1, the putative Tetraacyldisaccharide-1-P 4-kinase from Chromobacterium violaceum. It has 61 residues. The JPred prediction is not very good; it predicts an α-helix between residues 4 – 10 which does not appear in the target structure and it

fails to predict two other α-helices (29 – 31, 46 – 48) and three β-strands (18 – 20, 41 – 42, 50 – 51), as seen in Table 6-2.

The helix between residues 58 – 61 was selected as a reference and half of the constraints involved one of these residues. Eighteen decoys were found to satisfy all twenty-five constraints with RMSDs ranging from 6.9 – 11.0 Å (Figure 6-8 C, D, E, F). The top scoring decoys with the lowest RMSDs of 6.9 Å (2poo, 1poo, 1cvm, 1qlg) represent the enzyme phytase. They all have a β-sheet pattern similar to the target structure, but differ somewhat in the loop regions. The longest continuous segment with an RMSD under 5 Å for these four decoys is between residues 22 – 61, which includes the reference helix.

Different forms of trans-hydrogenase are represented by 1hzz, 1l7d, 1l7e, 1ptj, 1u2d, and 1xlt (Figure 6-8E); decoys from these parent proteins also satisfied all twenty-five constraints. The structures of these decoys have two β-sheets and a terminal α-helix and their RMSDs ranged from 10.6 – 10.7 Å. Although the overall RMSD was quite different, when the structures were aligned in smaller pieces (residues 4 – 22, 18 – 41, and 41 – 61), the RMSD of each section was only 5.7 – 5.8 Å. The LCS-5 for a decoy from 1hzz was composed of residues 23 – 45.

Three other top scoring decoys are depicted in Figure 6-9 to show the wide range of structures that satisfy all constraints. A decoy from parent protein 1s6l had a RMSD of 8.7 Å. As seen previously, the β-sheets of this decoy are inverted compared to the target. A decoy from parent protein 1e88 had an RMSD of 11.0 Å. Similar to the other top scoring decoys with high RMSDs, when the 1e88 and the target were aligned in fragments (residues 20 – 40, 41 – 60), the RMSDs of each section were within the range for good structure predictions (5.7 Å and 5.9 Å respectively). The longest continuous segment with an RMSD less than 5.0 Å is 19 residues. All of these decoys were able to satisfy twenty-five constraints despite their great difference in

structure. To prevent these high RMSD decoys from satisfying all constraints, better constraints must be chosen. Due to the highly inaccurate secondary structure prediction for this target, good constraint choices were quite difficult.

The best decoy in the database, from 1tl2, (Figure 6-8B) was found to have an RMSD of 5.3 Å and a score of 22. Two of the unsatisfied constraints involved residue 10, which was predicted by JPred to be in an α-helix but was not in a defined region of secondary structure in the target. This decoy has a LCS-5 of 56 residues, almost the entire structure.

6.1.2.2 Target T349

Target T349 represents 2hfv, Pseudomonas aeruginosa hypothetical protein RPA1041. It is composed of 75 residues. JPred does not predict the first β-sheet between residues 2 – 7 and predicts residues 58 and 65 to be in β-strands but they are in structurally undefined regions of the target structure. It gives a reliable prediction for the remaining secondary structural elements.

The helix between residues 9 – 23 was selected as a reference and most constraints involved an atom from this region. Sixty-four decoys satisfied all constraints with RMSDs ranging from 10.5 – 13.6 Å (Figure 6-9). A decoy from parent protein 1ta3 (Figure 6-9D) had an RMSD of 13.6 Å, but it was quite similar to the target in two regions: from residues 49 – 71 the RMSD was 3.0 Å and from residues 1 – 18 the RMSD was 5.9 Å. The longest continuous segment with an RMSD less than 5.0 Å included 30 residues. A decoy from parent protein 1t9u had an RMSD of 10.5 Å (Figure 6-9E) and satisfied all twenty-five constraints. Its LCS-5 had 32 residues, which were also located near the C-terminus of the decoy.

In the database, twenty-eight decoys had RMSDs less than 5.6 Å and scores ranging from 14 – 21. A decoy from 1uj5 (Figure 6-9B) had an RMSD of 5.6 Å and a score of 21, while a decoy of 1wel (Figure 6-9C) satisfied only fourteen constraints and had the same RMSD.

Several of the unsatisfied constraints for both of these decoys involved atoms in the final α-helix composed of residues 55 – 64. The two decoys were similar to each other having an RMSD between them of 5.4 Å. Increasing the constraint distance range from 5 Å to 7 Å raises the scores of the lowest RMSD decoys to 23, but also allows other high RMSD decoys to satisfy more constraints.

The decoy from 1uj5 had only 39 residues in its LCS-5, while 1wel had 63. Therefore, 1wel should have satisfied more constraints than 1uj5, even though they had very similar RMSD values, but this did not happen. For the decoy from 1uj5, the placement of secondary structure is closer to the JPred prediction than that of the target. The target and decoy differ greatly in the loop regions; the decoy has two short β-sheets where the target has a long loop. The decoy also ends with a β-sheet unlike the target which has a largely unstructured terminal region.

Another set of constraints was selected in an attempt to better the results. Nine atoms were chosen and distances were calculated between them. Satisfying all the new constraints were 42 decoys with RMSDs ranging from 5.5 Å – 11.1 Å. This was a slight improvement as the lowest RMSD decoy was found to satisfy all constraints.

6.1.2.3 Target T358

Target T358 represents 2hjj, protein ykfF from Escherichia coli. It is 66 residues long. The JPred prediction is pretty good for this protein. The first nine residues are missing in the crystal structure.

The helix between residues 5 – 14 was selected as a reference structure and all constraints involved an atom in this region. Five decoys satisfied all twenty-five constraints with RMSDs ranging from 8.3 – 11.8 Å (Figure 6-10). A decoy from 1p99 had an RMSD of 8.3 Å. Forty-two residues comprised the LCS-5 for this decoy, which included the reference helix. Three top

scoring decoys, 1efd, 1k2v, and 1k7s, had RMSDs of 9.3 Å and LCS-5s of 27 residues, including the reference helix. The longest continuous segment for 1x9d (RMSD = 11.8 Å), also include the reference helix but was composed of only 24 residues.

The best decoys in the database, from parent proteins 1oe9, 1w7i, and 1w7j (Figure 6-10B), had RMSDs of 5.5 Å and scores of 22. Their longest continuous segments with RMSDs under 5.0 Å were 37 residues long. The segments did not include residues in the reference helix which may explain why these decoys did not satisfy all constraints. It is difficult to know *a priori* which regions of the target will be most similar to any particular decoy.

6.1.2.4 CASP comparisons

As done for the targets that worked, we employed the GDT analysis to evaluate our method's performance on the targets that could have worked but did not. Target T348 was difficult for most CASP participants. There were no predicted structures with 100 % of the residues within a 10.0 Å distance cutoff. However, our lowest RMSD decoy in the set (1tl2, pink line in Figure 6-11A) did satisfy this requirement but it was not one of the top scoring decoys. A decoy from 1cvm satisfied all constraints and performed well compared to other CASP predictions (blue line in Figure 6-11A).

The results for target T349 were quite mixed; some groups did well, while others struggled. Our top scoring decoys were not among the best predictions (red and blue blue lines in Figure 6-11B) and the lowest RMSD decoys in the set were average predictions (green and cyan lines in Figure 6-11B).

T358 was a difficult target. Our lowest RMSD decoy in the set would have been one of the best structure predictions had it satisfied all constraints (red line in Figure 6-11C). Two of our top scoring decoys (green and blue lines) were slightly better than average predictions while the other decoy (cyan line Figure 6-11C) was not very good.

6.1.3 Targets That Never Had a Chance

Low RMSD decoys were not generated for the remaining four targets. Lack of structures similar to these targets shows our decoy set to be incomplete. These targets are not represented in the database; either they are not fragments of larger proteins or their parent protein was not included in our database. If they exist, the similar proteins may be less than 100 residues long or they may contain gaps in their PDB structures, excluding them from our decoy set.

6.1.3.1 Target T306

Target T306 (Figure 6-12A) corresponds to 2hd3, a small fragment of Ethanolamine Utilization Protein (EutN) from Escherichia coli. It has 95 residues. The JPred prediction is not very accurate. It does not predict either of the α-helices. It also predicts a long β-sheet composed of residues 75 – 87 which is split into two smaller β-sheets in the real structure (Table 6-1).

Twenty of the twenty-five constraints were chosen from the same reference structure, the β-sheet composed of residues 40 – 45. Two decoys with parent proteins, 1jhw and 1j72, were found to satisfy twenty-four constraints (Figure 6-12C). The RMSDs of the top scoring decoys were 13.4 Å and 13.5 Å respectively. Both parent proteins represent a macrophage capping protein, Cap G, which is composed of four β-sheets and a long α-helix (Figure 6-12C). The longest continuous segment with an RMSD under 5.0 Å is small for these decoys, composed of only 24 residues. The reference β-strand is also in this region, which helps explain why such high RMSD decoys satisfy most of the constraints.

Unlike the top scoring decoys, the target structure has a β-barrel center, as does the lowest RMSD decoy in the database (from parent protein 1fgu, Figure 6-12B). This decoy, however,

was found to satisfy only eighteen constraints. Four of the six unsatisfied constraints involved residue 43, which is part of a β-sheet in the target protein and a small α-helix in the decoy.

Several other slight structural differences exist between the lowest RMSD decoy and the target protein. The decoy has a loop from residues 10 – 22, a small α-helix from 40 – 49, and a β-sheet from residues 60 – 67. The target, however, has an α-helix composed of residues 16 – 20, a loop region from 46 – 52, and an α-helix from residues 61 – 67. In the C-terminus, the target ends with two short β-sheets while the decoy finishes with one short β-sheet followed by an α-helix. All of these differences give rise to an RMSD between the decoy and target of 8.1 Å. The longest continuous segment with an RMSD less than 5.0 Å is composed of 49 residues for this decoy, which included the reference β-strand.

No decoys satisfied all twenty-five constraints. Due to JPred's poor secondary structure predictions for this target, choosing good constraints was quite challenging. The lowest RMSD decoy in the database satisfies a unique set of eighteen constraints. When only those constraints are used, the lowest RMSD decoy is assigned a perfect score. Regardless of constraint choices, however, for this target, no reliable predictions can be made because no decoys with RMSDs less than 6.0 Å exist in the database.

6.1.3.2 Target T311

Target T311 is associated with two parent proteins, 2icp and 2ict, which represent bacterial antitoxin HigA, each crystallized at a different pH. We used 2ict as our reference structure and it was composed of 87 residues. The JPred prediction for this protein was fairly accurate having only a slight discrepancy in the position of the last α-helix (Table 6-2).

The α-helix composed of residues 57 – 74 was selected as the reference helix; all constraints involved an atom in this region. Twenty-three decoys satisfied all twenty-five

constraints and had RMSDs ranging from 10.1 – 10.2 Å (Figure 6-13). The target and the decoys have different local structures for the first and last 15 residues. The remainder of the structure is mostly helical, with both proteins having similar sized helices. The difference lies in the orientation of these helices thereby increasing the overall RMSD between the two structures. Because the decoy's helices are rotated only slightly, the distances between them are similar to those of the target which explains why these high RMSD decoys satisfied so many constraints. For the top scoring decoys, the longest continuous segment with an RMSD under 5.0 Å is 49 residues long, from residue 31 to 79. These decoys came from eleven parent proteins, nine of which were from some form of carbamoyl phosphate synthetase (PDB codes: 1a9x, 1bxr, c30, 1c3o, 1cs0, 1jdb, 1kee, 1mv6, 1t36). Parent protein 1ceb represents recombinant kringle 1 domain of human plasminogen while 1cs8 is procathepsin L.

In the database, 64 slightly different structures of parent protein 1f6g had the lowest RMSD, 6.6 Å. This decoy has a much longer terminal α-helix than the target or the top scoring decoys. It also has an initial α-helix that is quite similar in size to the target. All of the low RMSD decoys satisfied twenty constraints. Three of the unsatisfied constraints involved residue 64 and an atom in a loop region. The other two constraints involved residue 50 which is on a small α-helix in the target structure and part of the large terminal α-helix of the decoy. As seen for T306, no reliable decoys are generated for this target.

6.1.3.3 Target T353

Target T353 represents 2hfq, protein NE1680 from Nitrosomonas europaea. It has 85 residues. JPred accurately predicts the secondary structure for this target. The α-helix composed of residues 29 – 42 was selected as a reference and all constraints involved an atom in this helix. Twenty-six decoys were found to satisfy twenty-four constraints with RMSDs

ranging from 10.8 – 15.2 Å (Figure 6-14). The longest continuous segments under 5.0 Å RMSD for the top scoring decoys were 26 residues for 1ekf (from residue 28 to 53) and 27 residues (from residue 27 to 53) for 1j49. Both LCS-5s included the reference helix.

The best decoy in the database, from parent protein 1jrp, had an RMSD of 6.4 Å (Figure 6-14B) and a score of 17. The decoy matches the target structure exceptionally well from residues 12 – 43, with an RMSD of ~3.1 Å in this region. The longest continuous segment with an RMSD under 5.0 Å is composed of 54 residues (from residue 9 to 62). The lowest RMSD decoy did not satisfy all constraints indicating several poor constraints were chosen. The lowest RMSD decoy, however, was not good enough to be considered a reliable model even if it did satisfy all constraints.

6.1.3.4 Target T363

Target T363 represents 2hj1, the 3D domain-swapped dimer of hypothetical protein from Haemophilus influenzae. The PDB structure contains 77 residues. The JPred prediction for this sequence is fairly accurate. The α-helix between residues 25 – 34 was chosen as a reference and all twenty-five constraints involved an atom in this region. Three decoys of parent protein 1sxg satisfied all twenty-five constraints and had RMSDs of 9.3 – 9.4 Å (Figure 6-15). The longest continuous segment with an RMSD under 5.0 Å was composed of 36 residues and included the reference α-helix.

The lowest RMSD decoy came from parent protein 1hux and is shown in Figure 6-15. It was found to satisfy twenty-four constraints. Comprising the LCS-5 for this decoy were residues 1 – 59. Twenty-seven decoys satisfied the same set of twenty-four constraints with RMSDs ranging from 6.4 – 11.7 Å.

6.1.3.5 CASP comparisons

We performed the GDT analysis on the four proteins for which our database had no low RMSD decoys (Figure 6-16). Target T306 was difficult to predict for most CASP participants; no models have more than 60 % of the residues within a 5.0 Å distance cutoff. Our lowest RMSD decoy (not top scoring), was one of only two models to have 85 % of the residues under a 10.0 Å distance cutoff (blue line in Figure 6-16A). Our top scoring decoys were about average compared to the other CASP models.

Predicting the structure of target T353 was also difficult for most groups. Our lowest RMSD decoys (1jrp, 1jro) were the only models to have 95% of the residues under a distance cutoff of 10.0 Å (pink lines in Figure 6-16C). Our top scoring decoys, however, were again average models.

Our results for T311 and T363 were not very good. Most groups found target T311 pretty easy to predict, while the results for T363 were mixed. In both cases, our top scoring decoys were not among the best predictions.

6.1.4 Summary of Results for General Decoy Set

We studied twelve CASP7 targets. The lowest RMSD decoy in the database for five of the targets was assigned the highest score. Three targets had low RMSD decoys but they were not assigned the highest score while four targets did not have any low RMSD decoys in the database. Because constraints were chosen with the aid of a secondary structure prediction method, predicting incorrect secondary structure can lead to poor constraint choices which results in bad structure predictions. We also found that many high scoring high RMSD decoys have regions of great similarity to the target which usually contain the reference structure.

When comparing our method to the other methods used in CASP7, we find that our method performed quite well for some of the hardest targets and not well for some of the easiest targets.

6.2 Specific Decoy Sets

We studied the same 12 targets as in section 6.1 and used the same sets of constraints. To generate specific decoy sets for each target, we used the Rosetta algorithm. Each set contained exactly 10,000 decoys. The C^{α} RMSDs between each target and each decoy were calculated and the RMSD distribution can be found in Figure 6-17. The RMSD distribution varies greatly from target to target. As seen in the general decoy set, target T335 has a very high number of low RMSD decoys; 95.6 % of decoys are under 6.0 Å. Targets T309 and T306 have RMSD distributions centered around 15.0 Å and 14.0 Å respectively, and have no decoys with RMSDs under 6.0 Å. The RMSD of the best decoy for each target is listed in Table 6-3.

6.2.1 Targets That Worked

Four target proteins, T311, T335, T358, and T349, had the lowest RMSD decoy in their set satisfy all twenty-five constraints. In each case, the lowest RMSD was under 6.0 Å. The number of top scoring decoys, range of RMSD values, as well as the RMSD of the best decoy in the set can be found in Table 6-3.

For T311, 181 decoys had perfect scores and their RMSDs ranged from 4.8 – 14.6 Å. The highest (#61) and lowest (#3545) RMSD decoys satisfying all constraints are shown in Figure 6-18(A, B). The first fifty residues of the high scoring high RMSD decoy (#61) are similar to the target; the RMSD in this region is only 1.7 Å. This decoy has a high RMSD despite its very similar local structure because it also has a few consecutive incorrect dihedral angles. Rotation around this bond makes the whole decoy structure quite different from the target despite their great similarities in local structure. The worst decoy in the set had an RMSD of 19.0 Å and

satisfied only 6 constraints. Our method successfully assigned the worst decoy in the database, a low score. Rosetta generated 158 decoys with RMSDs under 6.0 Å and our discrimination procedure assigned these decoys scores ranging from 13 – 25 with an average of 23.

The 84 top scoring decoys for T349 had RMSDs ranging from 4.0 – 11.0 Å. The highest (#3665) and lowest (#4480) RMSD decoys can be found in Figure 6-18(C, D). Despite the high overall RMSD, the high scoring high RMSD decoy was similar to the target from residues 6 to 24 and 49 to 66 having RMSDs of only 1.8 and 1.5 Å respectively. Both sections correspond to α-helices, with the first being the reference helix. The highest RMSD decoy in the set (15.3 Å) satisfied 14 constraints. Fifty-four decoys were generated with RMSDs less than 6.0 Å and they satisfied between 14 – 25 constraints with an average of 23.

Target T358 had 20 decoys satisfying all constraints with RMSDs ranging from 4.1 – 11.4 Å (Figure 6-18(E, F)). The 11.4 Å RMSD decoy (#3334) was similar to the target from residues 1 to 20, which included the reference helix, and from residues 48 to 60. The RMSDs of these sections were 2.1 Å and 2.0 Å, respectively. The highest RMSD decoy for T358 satisfied 13 constraints and had an RMSD of 15.7 Å. The 122 decoys with RMSDs under 6.0 Å had scores ranging from 10 – 25 with an average of 20, slightly lower than the other targets.

Many decoys satisfied all constraints for T335. The range of RMSDs, however, was small, from 1.4 – 8.2 Å (Figure 6-18(G, H)). The lowest RMSD decoy for this target (#9623) was the lowest RMSD decoy generated for any target. Like the target T335, the high scoring high RMSD decoy, #2312, had an α-helix from residues 4 to 20. The target and decoy had an RMSD of 1.8 Å between residues 1 – 21. Almost all the decoys in this set had RMSDs under 6.0 Å and their scores ranged from 12 – 25 with an average of 23.5. One of the highest RMSD decoys,

with an RMSD of 11.4 Å, also satisfied 24 constraints. In this case, very few decoys had high RMSDs and it was hard to separate the good from the bad.

Four decoys, T288, T348, T359, and T363, had top scoring decoys with low RMSDs but the lowest RMSD decoy in the set did not satisfy all constraints.

For target T288, 86 decoys satisfied all twenty-five constraints and their RMSDs ranged from 3.6 – 13.2 Å (Figure 6-19). The 13.2 Å RMSD decoy (#7663) was similar to the target from residue 46 to 83, which included the reference helix. The RMSD of this region was 1.9 Å. The best decoy in database (#369), with an RMSD of 3.5 Å, satisfied 23 constraints. The two constraints it did not satisfy were between atoms in the reference helix and the second β-sheet. The highest RMSD decoy in the database satisfied 21 constraints and had an RMSD of 17.0 Å. Poor constraint choices may be the result of this high RMSD decoy satisfying so many constraints. Rosetta generated 25 decoys with RMSDs under 6.0 Å and their scores ranged from 19 – 25 with an average score of 24.

No decoys for T348 satisfied all 25 constraints. Two decoys satisfied 23 constraints and had RMSDs of 5.6 Å and 12.8 Å (Figure 6-20). They did not satisfy the same set of 23 constraints. The 12.8 Å RMSD decoy (#7088) was similar to the target from residues 15 to 31 and 49 to 61, which included the reference helix (from residues 58 – 61). The RMSDs in these regions were 2.0 Å and 1.6 Å respectively. The lowest RMSD decoy (#1017) satisfied only 12 constraints and had an RMSD of 4.2 Å. The secondary structure prediction for this target was not very good resulting in poor constraint choices. The highest RMSD decoy in the set satisfied 8 constraints and had an RMSD of 15.3 Å. The 151 low RMSD decoys (RMSD less than 6.0 Å) had scores ranging from 7 – 25 with an average of 16. To improve our results, better constraints must be chosen.

Forty decoys of T359, with RMSDs ranging from 6.0 – 15.2 Å, satisfied all constraints (Figure 6-21). The high scoring, high RMSD decoy (#3012) and the target had an RMSD of 2.1 Å from residues 51 – 89, which included the reference helix from residues 71 to 79. The best decoy in the set, #112, had an RMSD of 4.8 Å and satisfied 23 constraints, while the highest RMSD decoy (17.7 Å) satisfied 17 constraints. Only four decoys were generated that had RMSDs less than 6.0 Å. Their scores ranged from 22 – 25 with an average of 23.

For target T363, 11 decoys satisfied all twenty-five constraints and their RMSDs ranged from 5.7 – 10.5 Å (Figure 6-22). The 10.5 Å decoy (#5181) was similar to the target from residue 1 to 39 having an RMSD of 2.8 Å in this region. In addition to this region containing the reference helix, it is also the most structurally defined area; the target is fairly unstructured from residue 51 to the C-terminus. The two best decoys in the set (#4551, #6388) with RMSDs of 5.1 Å satisfied 19 and 24 constraints, while the highest RMSD decoy (16.8 Å) satisfied 11 constraints. Twenty-five decoys had RMSDs under 6.0 Å and their scores ranged from 16 – 25 with an average of 20.

6.2.2 Targets That Did Not Work

The Rosetta method generated low RMSD decoys for targets T340 and T353, but our method was unable to discriminate the good decoys from the bad. Target T340 had only five low RMSD decoys in the database with scores ranging from 12 – 24 and an average of 21, while the 57 low RMSD decoys for T353 had scored ranging from 15 – 24 with an average score of 20.

Six decoys of T340 satisfied all constraints and had RMSDs ranging from 7.4 – 14.5 Å (Figure 6-23). The 7.4 Å RMSD decoy (#9880) was similar to the target from residues 1 to 11 and 50 to 74, which included the reference helix. Both sections had RMSDs of 1.9 Å. The 14.5 Å RMSD decoy (#9412) was most similar to the target from residue 2 to 17 and 63 to 83. They had RMSDs of 1.9 and 2.9 respectively. The best decoy in the set, #94, had an RMSD of 3.7 Å

but satisfied only 12 constraints, while the next best decoy, with an RMSD of 3.8 Å, satisfied 24 constraints. The decoy with the highest RMSD in the set (17.6 Å) satisfied 20 constraints, while the next highest RMSD decoy (17.2 Å) satisfied 10 constraints.

Twenty decoys of T353 satisfied all twenty-five constraints with RMSDs ranging from 7.0 – 13.5 Å (Figure 6-24). The top scoring decoys were similar to the target from residue 1 to 47 for the 7.0 Å decoy and from residue 1 to 43 for the 13.5 Å decoy. The similar regions included the reference helix from residue 29 to 42. The RMSDs between target and decoy for these sections were 2.8 Å and 3.0 Å respectively. The higher RMSD decoy (#3009) also had a segment of high similarity to the target between residues 59 – 74 (1.7 Å). The best decoy in the set, #5124, had an RMSD of 4.3 Å and satisfied 22 constraints. The three unsatisfied constraints involved residue 29, which is located in an α-helix that is slightly displaced compared to the target. The highest RMSD decoy (16.2 Å) satisfied only 13 constraints.

6.2.3 Targets That Never Had a Chance

For targets T306 and T309, no low RMSD decoys were generated. No decoys for T306 satisfied all constraints. However, two decoys satisfied 24 constraints and had RMSDs of 12.5 Å and 13.5 Å (Figure 6-25). The 12.5 Å RMSD decoy (#1604) was most similar to the target between residues 15 – 32 and 28 – 59 having RMSDs of 3.0 Å and 2.8 Å in these segments. The 13.5 Å decoy (#5643) and the target have similarities between residues 36 – 45, 72 – 82, and 56 – 89 with RMSDs of 2.5 Å, 2.8 Å, and 3.1 Å. The reference β-strand was included in a low RMSD region of each of the top scoring decoys. The lowest RMSD decoy in database (#935) satisfied only16 constraints and had an RMSD of 8.0 Å; five of the unsatisfied constraints involved residue 40, which is located within the β-barrel structure of both the target and the decoy. The highest RMSD decoy (18.0 Å) satisfied only 9 constraints.

Only one decoy of T309 (#210) satisfied all constraints and it had an RMSD of 11.0 Å (Figure 6-26). The target has defined secondary structure only between residues 13 and 40, and the RMSD between the target and decoy in this region (from residues 13 and 39) was 3.7 Å. Therefore, if only the regions of defined secondary structure are considered, the method worked well for this target. The best decoy in database, #5810, satisfied 22 constraints and had an RMSD of 8.1 Å. Two of the unsatisfied constraints involved residue 30, which is located within one of the β-sheets. The highest RMSD decoy (19.8 Å) satisfied 16 constraints.

6.2.4 Summary of Results Using the Specific Decoy Set

Rosetta generated low RMSD decoys for ten of the twelve targets. Even the high RMSD decoys usually had some local structural similarities to the target. Low RMSD decoys were assigned top scores for eight targets. Two targets had low RMSD decoys in the database but they did not satisfy all constraints while low RMSD decoys were not generated for two other targets, T306 and T309. Both T306 and T309 were difficult to predict for most CASP participants.

6.3 Comparisons of Decoy Sets

A comparison of the results using each type of decoy set can be found in Table 6-4. Rosetta generated low RMSD decoys for all but two targets, T309 and T306. The general decoy set generated low RMSD decoys for target T309 (5.7 Å) but the best decoy in the set for target T306 had an RMSD of 8.1 Å. Also, the lowest RMSD decoy for ten of the twelve targets is lower for Rosetta decoys than the general set. The general decoy set generated better decoys for Target T309 and T359.

The discrimination process was equally effective for both types of decoy sets. Three targets (T288, T359, T335) had successful predictions using both decoy sets. Target T306 did not have low RMSD decoys in either set. This is not a very common type of protein. The three

targets for which low RMSDs were generated but were not found in the general decoy set (T348, T349, T358) and two targets for which the general decoy set did not generate low RMSD decoys (T311, T363), had successful predictions using the specific decoys sets. Using the general decoy set, T309 and T340 had low RMSD decoys with top scores, whereas using the specific decoy set, successful predictions were not obtained. Finally, for T353, the specific decoy set generates a low RMSD decoy but it is not assigned a top score, while the general decoy set does not generate a low RMSD decoy. When both methods are considered, ten of the twelve targets had successful predictions.

Table 6-1. Results for 12 targets

Target	Lowest RMSD in decoy set	Range of RMSDs for top scoring decoys	Number of top scoring decoys
288	3.4	3.4 – 5.6	11
340	3.0	3.0 – 14.0	60
359	3.6	3.6 – 13.1	15
309	5.7	5.7 – 14.4	9*
335	2.1	2.1 – 9.5	10582
348	5.3	6.9 – 11.0	18
349	5.5	10.5 – 13.6	64
358	5.5	8.3 – 11.8	5
306	8.1	13.4 – 13.5	2*
311	6.6	10.1 – 10.2	23
353	6.4	10.8 – 15.2	26*
363	6.4	9.3 – 9.4	3

*no decoys satisfied all constraints; the number represents the number of decoys satisfying 24 constraints

Table 6-2. JPred predictions compared to target structures

T288				T306				T309			
Jpred		Real		JPred		Real		JPred		Real	
5 – 11	β	5 – 10	β	4 – 7	β	2 – 10	β	5 – 10	β		
17 – 23	β	19 – 23	β	10 – 15	β					13 – 15	β
30 – 36	β	31 – 36	β			17 – 19	α	18 – 21	α		
		41 – 45	α	24 – 30	β	23 – 30	β			20 – 24	β
53 – 57	β	52 – 57	β	40 – 46	β	36 – 45	β	27 – 40	α	29 – 33	β
		60 – 61	β	54 – 59	β	53 – 59	β			34 – 40	α
66 – 75	α	67 – 76	α			61 – 67	α	46 – 52	β		
80 – 86	β	80 – 86	β	75 – 87	β	76 – 81 84 – 85	β				
				92 – 93	β	93 – 94	β				

T311				T335				T340			
JPred		Real		JPred		Real		JPred		Real	
3 – 12	α	1 –13	α	3 – 18	α	5 – 19	α	6 – 11	β	6 – 11	β
16 – 23	α	17 – 24	α	24 – 46	α	24 – 40	α	18 – 24	β	19 – 23	β
28 – 35	α	28 – 36	α	51 – 55	β			31 – 36	β	30 – 35	β
43 – 52	α	43 – 53	α	64 – 73	α					40 – 44	α
57 – 75	α	57 – 74	α					52 – 56	β	50 – 55	β
		78 – 80	α							58 – 59	β
83 – 85	α							66 – 74	α	65 – 73	α
								80 – 86	β	78 – 84	β

T348				T349				T353			
JPred		Real		JPred		Real		JPred		Real	
4 – 10	α					2 – 7	β	3 – 11	β	3 –11	β
		18 – 20	β	9 – 21	α	9 – 23	α	17 – 24	β	17 – 24	β
26 – 28	β	25 – 28	β	27 – 30	β	26 – 28	β	29 – 43	α	29 – 43	α
		29 – 31	α	36 & 43	β			59 – 61	β	59 – 61	β
35 – 36	β	33 – 38	β	46 – 50	β	45 – 51	β	65 – 74	α	65 – 74	α
		41 – 42	β	54 – 64	α	55 – 64	α	76 – 79	β	76 – 79	β
		46 – 48	α								
		50 – 51	β								
58 – 62	α	54 – 60	α								

T358				T359				T363			
JPred		Real		JPred		Real		JPred		Real	
5 – 14	α	5 – 14	α	5 – 11	β	2 – 11	β	3 – 9	β	1 – 10	β
15 – 21	β	18 – 21	β	17 – 22	β	19 – 22	β	14 – 20	β	13 – 22	β
27 – 33	β	29 – 33	β	35 – 40	β	35 – 40	β	25 – 34	α	27 – 34	α
39 – 42	β	39 – 43	β			45 – 49	α			36 – 39	α
52 – 60	α	50 – 58	α	57 – 61	β	56 – 61	β	51 – 56	β		
						64 – 65	β	69 – 72	β		
				71 – 79	α	71 – 80	α				
				84 – 91	β	84 – 94	β				

Table 6-3. Results for each of the 12 targets

CASP Target	Number of top scoring decoys	RMSD range of top scoring decoys, (Å)	Lowest RMSD in decoy set, (Å)	Number of decoys with RMSDs under 6.0 Å
335	3619	1.4 – 8.22	1.4	9,562
311	181	4.8 – 14.6	4.8	158
358	20	4.1 – 11.4	4.1	122
349	84	4.0 – 11.0	4.0	54
363	11	5.7 – 10.5	5.1	25
288	86	3.6 – 13.2	3.5	25
359	40	6.0 – 15.2	4.8	4
348	2*	5.6, 12.8	4.2	151
353	20	7.0 – 13.5	4.3	57
340	6	7.4 – 14.5	3.7	5
309	1	11.0	8.1	0
306	2*	12.5, 13.5	8.0	0

*top score was less than 25

Table 6-4. Comparison of results for each target using both types of decoy sets

Target	Specific Set	General Set
288	3.6 – 13.2	3.4 – 5.6
335	1.4 – 8.22	2.1 – 9.5
359	6.0 – 15.2	3.6 – 13.1
349	4.0 – 11.0	10.5 – 13.6
358	4.1 – 11.4	8.3 – 11.8
348	5.6, 12.8	6.9 – 11.0
363	5.7 – 10.5	9.3 – 9.4
311	4.8 – 14.6	10.1 – 10.2
353*	7.0 – 13.5	10.8 – 15.2
340	7.4 – 14.5	3.0 – 14.0
309	11.0	5.7 – 14.4
306*	12.5, 13.5	13.4 – 13.5

Targets T353 and T306 were not predicted successfully by either method. The entries in red had successful predictions. Those in green had low RMSD decoys in the set but they did not satisfy all constraints, while those in blue did not have low RMSD decoys.

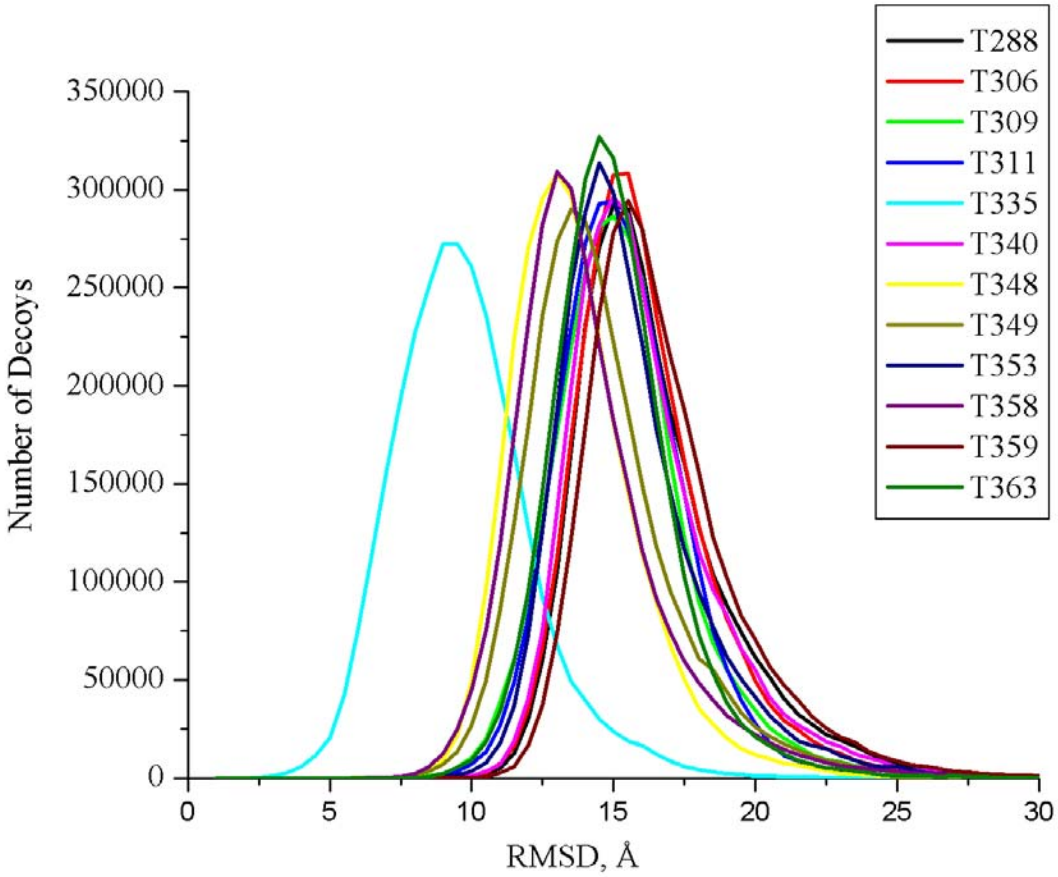


Figure 6-1. RMSD distributions for each target protein

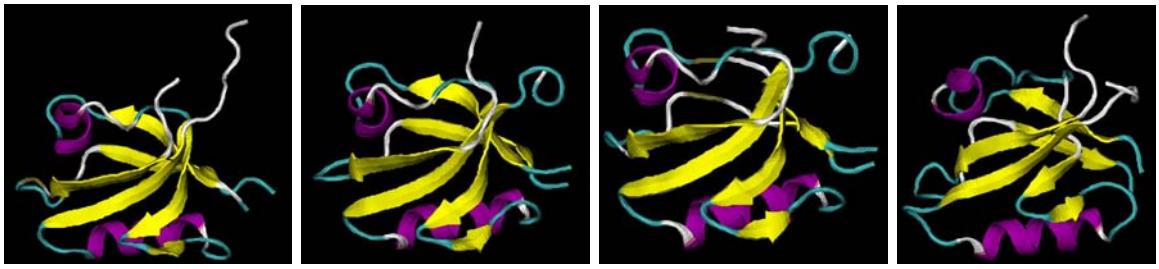


Figure 6-2. Target T288 and the top scoring decoys for T288. A) Target T288. B) 1tq3 had an RMSD of 3.4 Å. C) 1um7 had an RMSD of 3.5 Å. D) 1b8q had an RMSD of 5.6 Å.

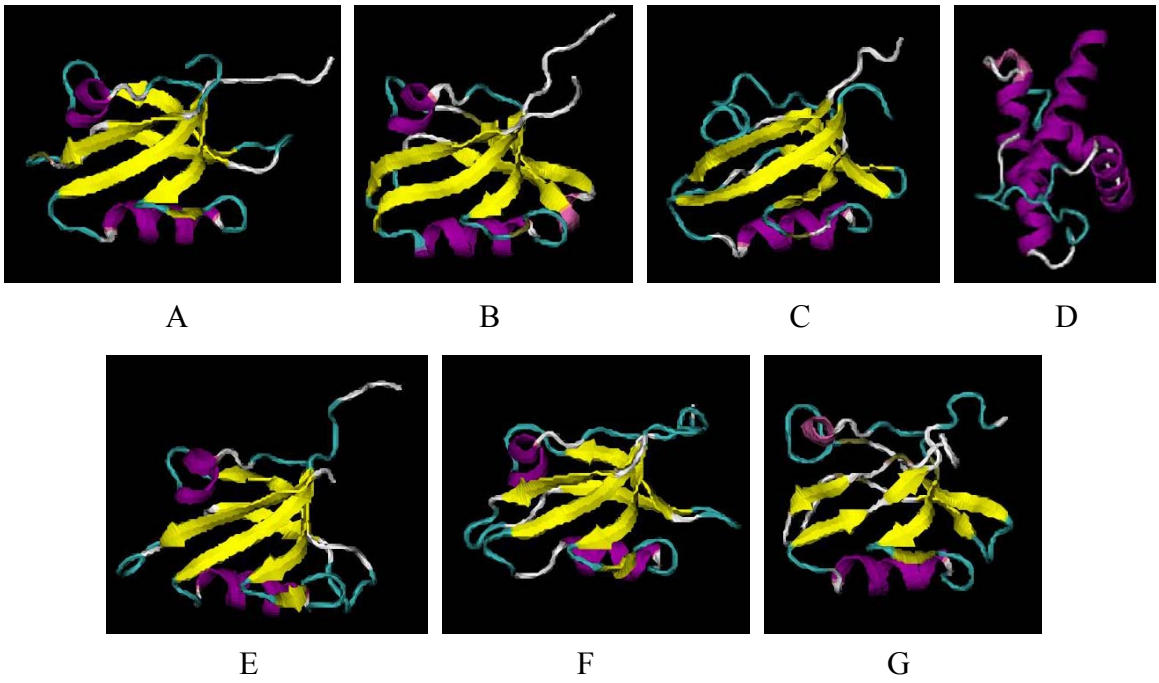


Figure 6-3. Target T340 and some of the top scoring decoys. A) Target T340. B) 1wf7 with an RMSD of 3.0 Å. C) 1wif with an RMSD of 4.4 Å. D) 1qln with an RMSD of 14.0 Å. E) 1uf1 with an RMSD of 4.4 Å. F) 1uit with an RMSD of 3.4 Å. G) 1v5l with an RMSD of 3.3 Å.

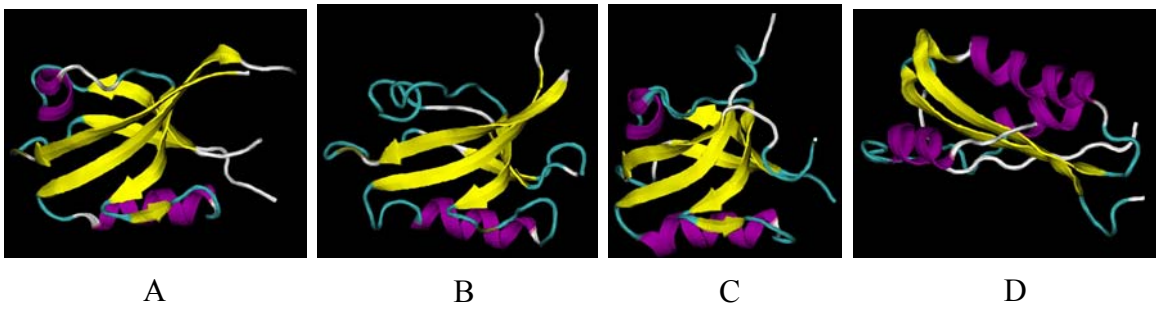
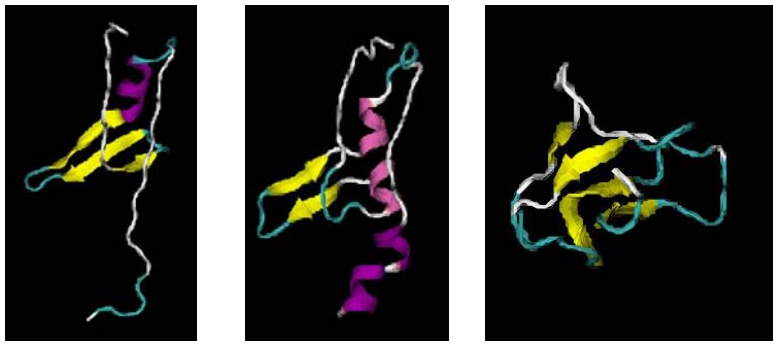
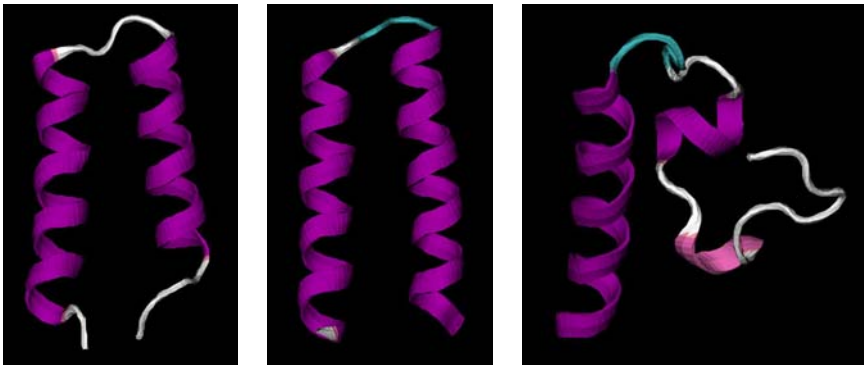


Figure 6-4. Target T359 and its top scoring decoys. A) Target T359. B) 1p1d and 1um1, each had an RMSD of 3.6 Å. C) 1ueq and 1wfv, each with an RMSD of 6.8 Å. D) 1w5e with an RMSD of 13.1 Å.



A B C

Figure 6-5. Target T309 and its top scoring decoys. A) Target T309. B) 1ese, 1esd, and 1esc, each with an RMSD of 5.7 Å. C) 1wk1 with an RMSD of 14.4 Å.



A B C

Figure 6-6. Target T335 and its top scoring decoys. A) Target T335. B) 1qsp, 1wa5, and 1z3h, each with an RMSD of 2.1 Å. C) 1b6c has an RMSD of 9.5 Å.

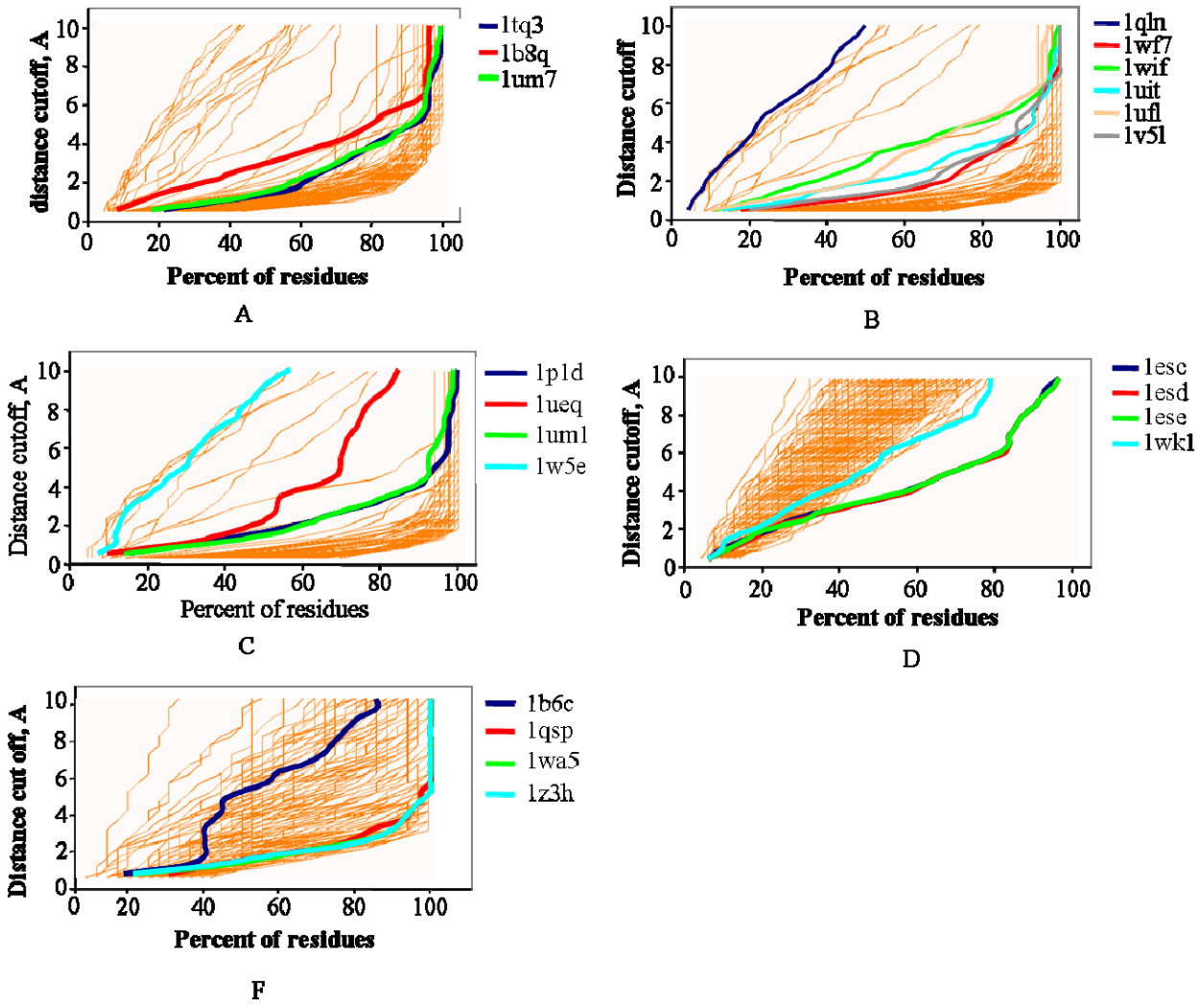
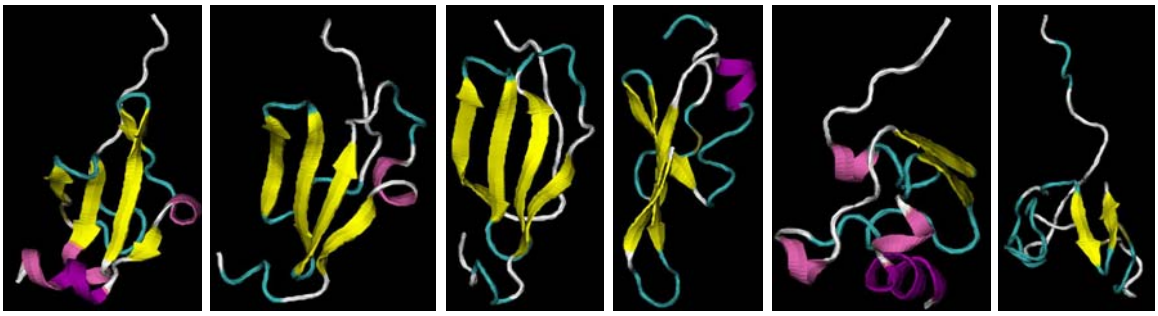
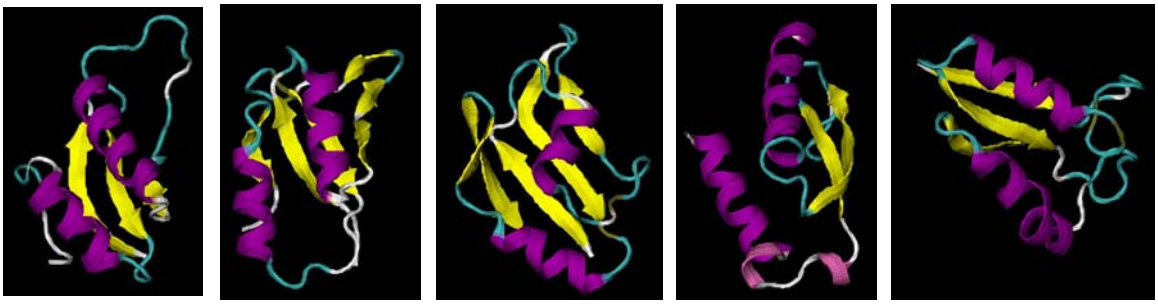


Figure 6-7. Use of Global Distance Test (GDT) analysis to compare our top scoring decoys with the results from other methods used in CASP7. A) Target T288. B) Target T340. C) Target T359. D) Target T309, decoys from 1esc, 1esd, and 1ese have RMSDs of 5.7 Å of, while 1wk1 has an RMSD of 14.4 Å. E) Target T335, the 1b6c decoy has an RMSD of 9.5 Å, while the other three have RMSDs of 2.1 Å



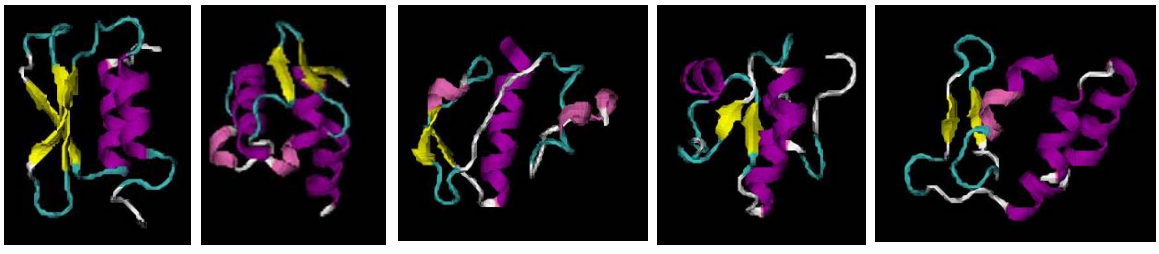
A B C D E F

Figure 6-8. Target T348, the best decoys in the database, and the top scoring decoys. A) Target T348. B) The best decoy in database, 1tl2, had an RMSD of 5.3 Å. Top scoring decoys: C) 1cvm, 1poo, 2poo, and 1qlg, each with an RMSD of 6.9 Å. D) 1s6l had an RMSD of 8.7 Å. E) 1hzz, 1l7d, 1l7e, 1ptj, 1u3d, and 1xlt have RMSDs ranging from 10.6 – 10.7 Å. F) 1e88 had an RMSD of 11.0 Å.



A B C D E

Figure 6-9. Target T349, best decoy in database, and top scoring decoys. A) Target T349. B) 1uj5 had an RMSD of 5.6 Å and a score of 21. C) 1wel had an RMSD of 5.5 Å and a score of 14. D 1t9u had an RMSD of 10.5 Å and a score of 25. E) 1ta3 had an RMSD of 13.6 Å and a score of 25.



A B C D E

Figure 6-10. Target T358, lowest RMSD decoys in database, and top scoring decoys. A) Target T358. B) The best decoys in database, 1oe9, 1w7i, and 1w7j, each have an RMSD of 5.5 Å. C) The top scoring decoy, 1p99, had an RMSD of 8.3 Å. D) The top scoring decoys, 1efd, 1k2v, and 1k7s, each have an RMSD of 9.3 Å. E) The top scoring decoy, 1x9d, had an RMSD of 11.8 Å.

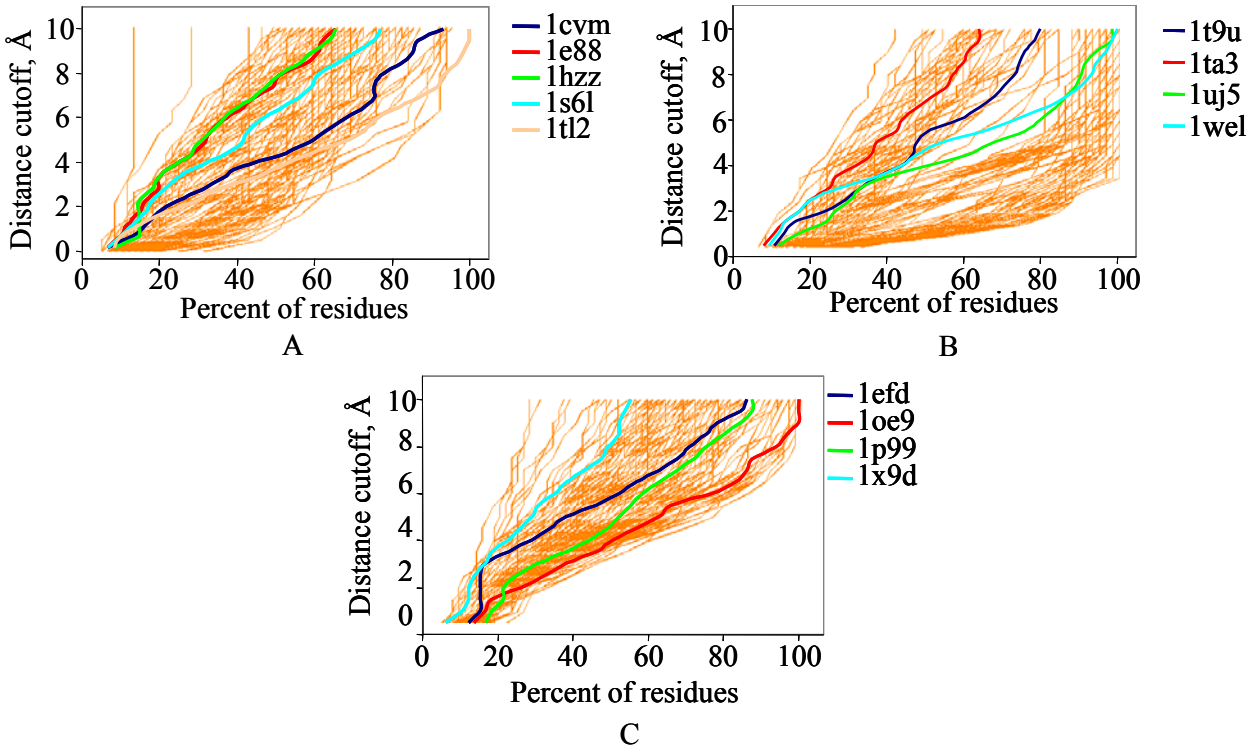
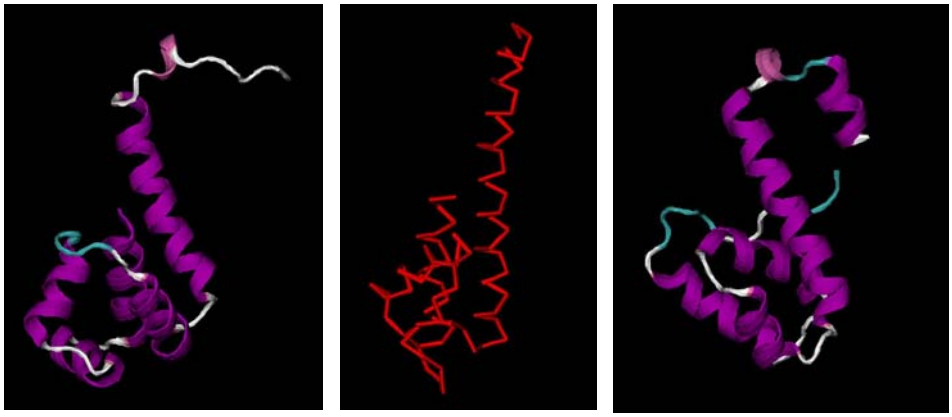


Figure 6-11. Use of Global Distance Test (GDT) analysis to compare our results with those from other methods used in CASP7. A) Target T348. The lowest RMSD decoy in the set (1tl2) has a score of 22. The remaining decoys in the figure satisfied all constraints. B) Target T349. The lowest RMSD decoys are 1wel and 1uj5 which satisfied 14 and 21 constraints respectively, which 1t9u and 1ta3 satisfied all constraints. C) Target T358. The lowest RMSD decoy in the set, 1oe9, satisfied 22 constraints. The remaining decoys satisfied all constraints.

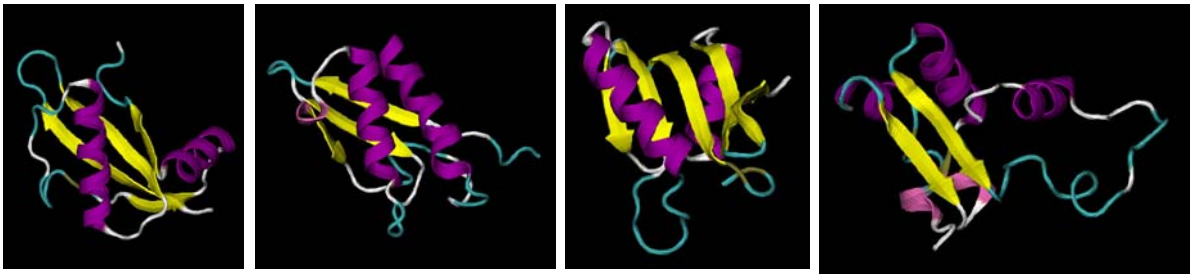
A B C

Figure 6-12. Target T306, lowest RMSD decoys in the database, and the top scoring decoys. A) Target T306. B) Best decoy in database, 1fgu, had an RMSD of 8.1 Å. C) Top scoring decoys, 1jhw and 1j72, had RMSDs ranging from 13.4 – 13.5 Å.



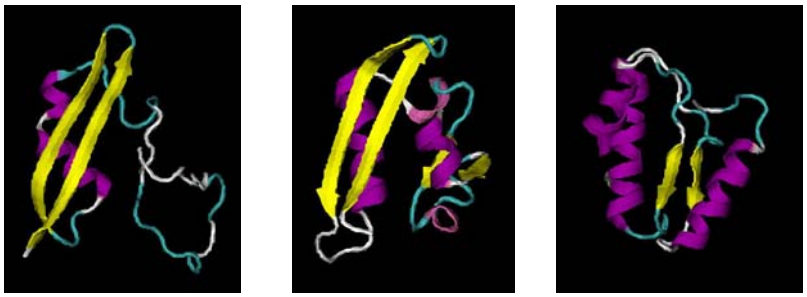
A B C

Figure 6-13. Target T311, best decoy in database, and top scoring decoy. A) Target T311. B) 1f6g has the lowest RMSD in the database, 6.6 Å. The PDB entry for the parent protein contains only the α-carbons. B) The top scoring decoys, 1a9x, 1bxr, c30, 1c3o, 1cs0, 1jdb, 1kee, 1mv6, and 1t36, had RMSDs ranging from 10.1 – 10.2 Å.



A B C D

Figure 6-14. Target T353, best RMSD decoy in database, and top scoring decoys. A) Target T353. B) The lowest RMSD decoys in database, 1jrp and 1jro, each have an RMSD of 6.4 Å. C) The top scoring decoys, 1ekf, had an RMSD of 10.8 Å. D) Both 1j49 and 1j4a, are top scoring decoys and have an RMSD of 11.6 Å.



A B C

Figure 6-15. Target T363, best decoy in database, and a top scoring decoy. A) Target T363. B) The lowest RMSD decoy is 1hux with an RMSD of 6.4 Å. C) The top scoring decoy, 1sxg, had an RMSD of 9.3 Å.

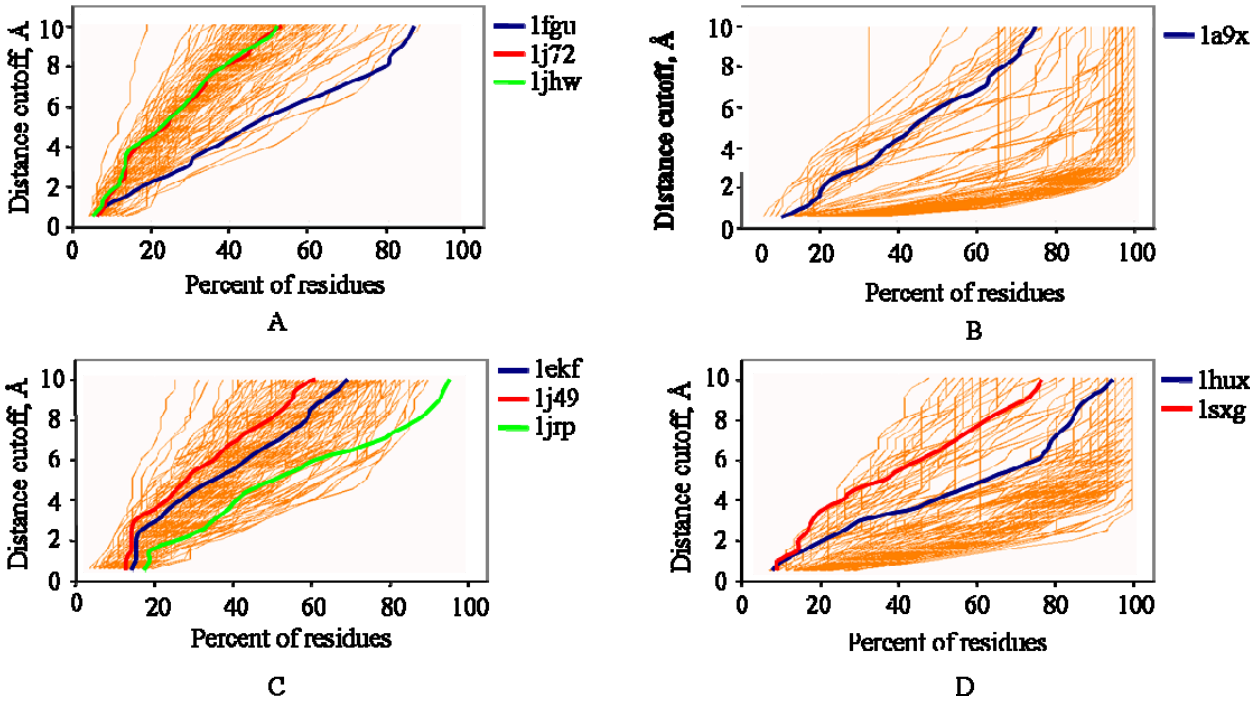


Figure 6-16. Use of Global Distance Test (GDT) analysis to compare our results with those from other methods used in CASP7. A) Target T306. The lowest RMSD decoy in the set, 1fgu, satisfied 18 constraints. The other two decoys satisfied all constraints. B) Target T311. The lowest RMSD decoy in the set, 1f6g, is not shown, while 1a9x is a top scoring decoy. C) Target T353. The lowest RMSD decoy in the set, 1jrp, satisfied 17 constraints. D) Target T363. The lowest RMSD decoy in the set, 1hux, satisfied 24 constraints, while 1sxg satisfied all constraints.

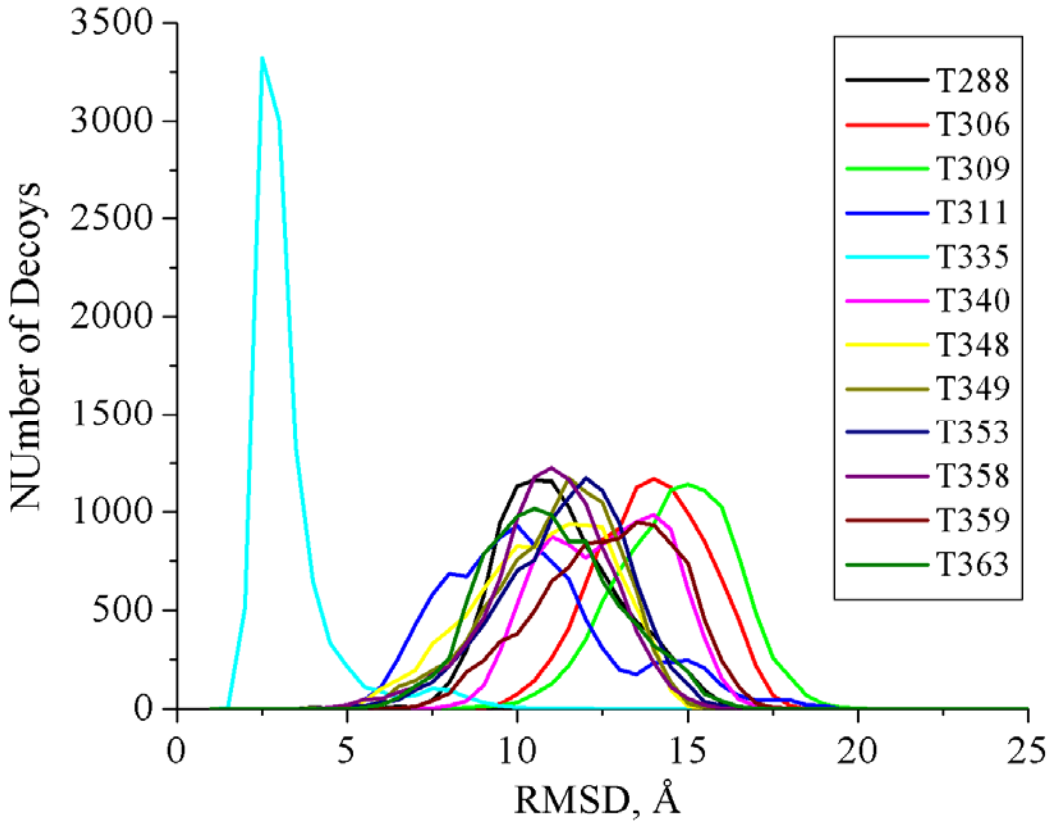


Figure 6-17. Histogram of C^{α} RMSDs for all twelve CASP targets.

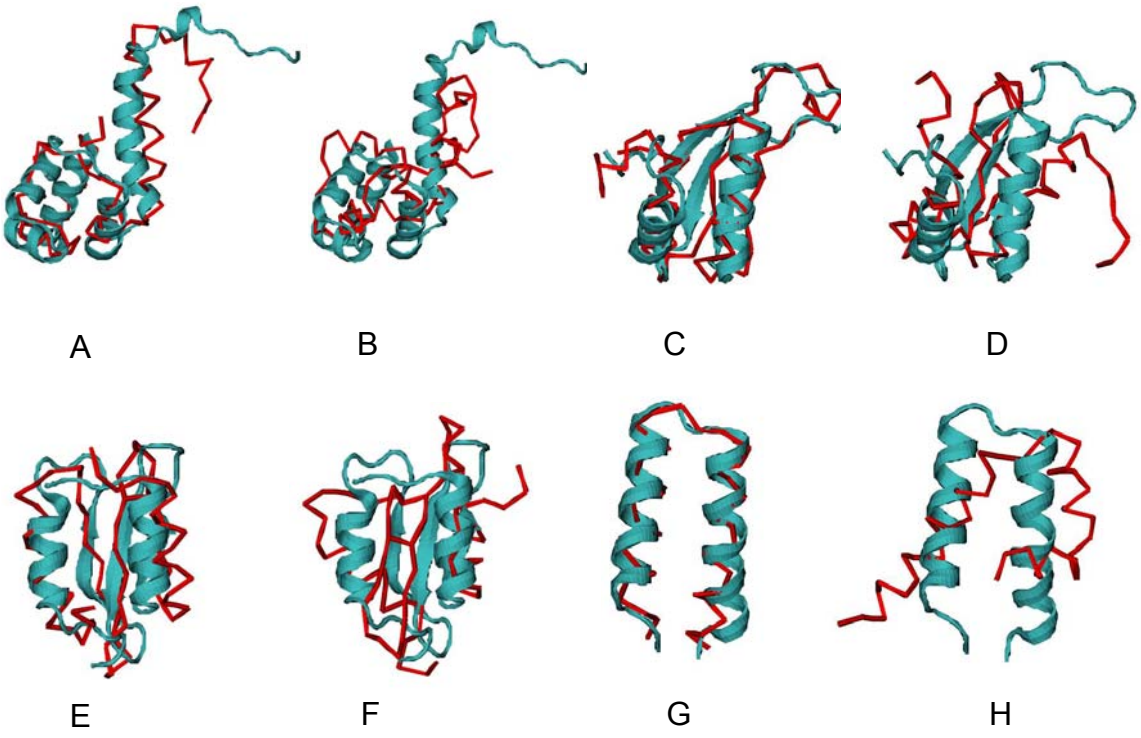
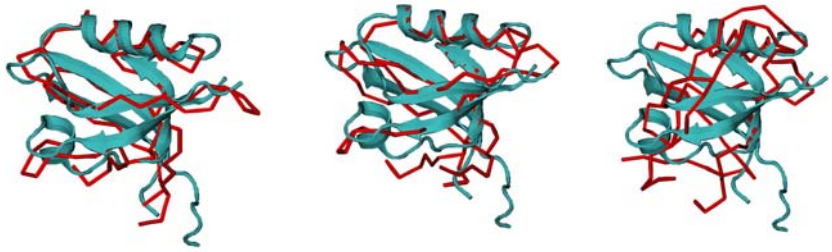


Figure 6-18. Top scoring decoys for target that worked. A) T311, #3545, with an RMSD of 4.8 Å. B) T311, #61, with an RMSD of 14.6 Å. C) T349, #4480, with an RMSD of 4.0 Å. D) T349, #3665, with an RMSD of 11.0 Å. E) T358, #1572, with an RMSD of 4.1 Å. F) T358, #3334, with an RMSD of 11.4 Å. G) T335, #9623, with an RMSD of 1.4 Å. H) T335, #2312, with an RMSD of 8.2 Å. For each target, the top scoring decoys with the lowest and highest RMSDs are shown.

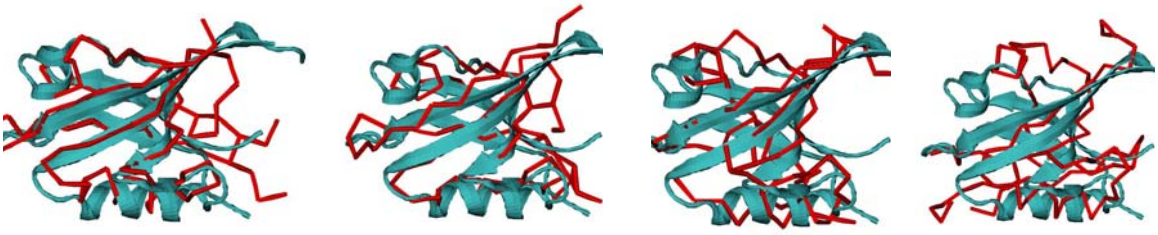


A B C

Figure 6-19. Results for T288. A) The best decoy in data set, #369, had an RMSD of 3.5 Å. B) The top scoring decoy, #5124, had an RMSD of 3.6 Å. C) The top scoring decoy, #7663, had an RMSD of 13.2 Å.

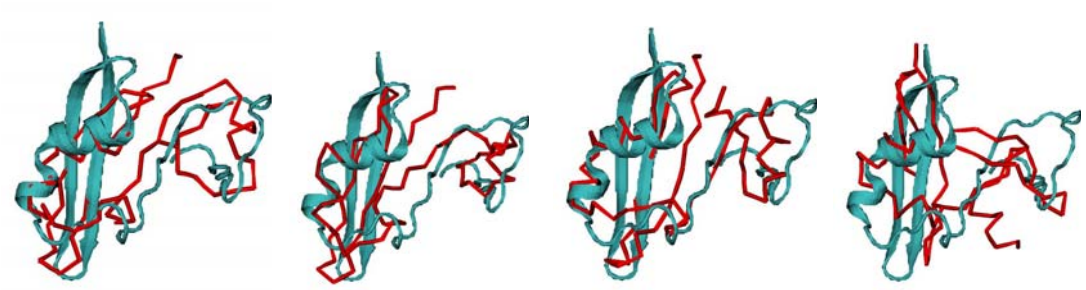
A B C

Figure 6-20. Results for T348. A) The best decoy in data set, #1017, had an RMSD of 4.2 Å. B) The top scoring decoy, #4218, had an RMSD of 5.6 Å. C) The top scoring decoy, #7088, had an RMSD of 12.8 Å.



A B C D

Figure 6-21. Results for T359. A) One of the best decoys in data set, #112, had an RMSD of 4.8 Å. B) The other best decoy in data set, #6536, also had an RMSD of 4.8 Å. C) The top scoring decoy, #5817, had an RMSD of 6.0 Å. D) The top scoring decoy, #3012, had an RMSD of 15.2 Å.



A B C D

Figure 6-22. Results for target T363. A) The best decoy in data set, #4551, had an RMSD of 5.1 Å. B) The best decoy in data set, #6388, had an RMSD of 5.1 Å. C) The top scoring decoy, #6376, had an RMSD of 5.7 Å. D) The top scoring decoy, #5181, had an RMSD of 10.5 Å.

A B C

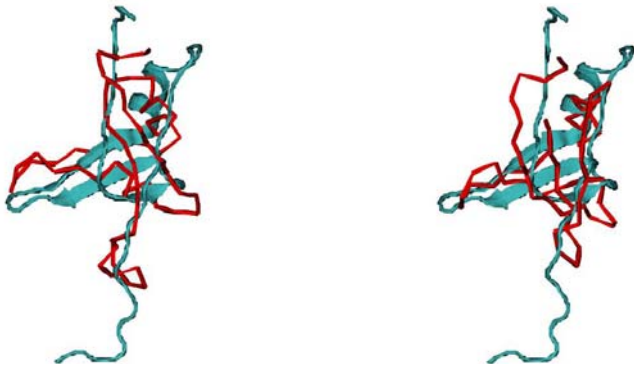
Figure 6-23. Results for target T340. A) The best decoy in data set, #94, had an RMSD of 3.7 Å. B) The top scoring decoy, # 9880, had an RMSD of 7.4 Å. C) The top scoring decoy, # 9412, had an RMSD of 14.5 Å.

A B C

Figure 6-24. Results for target T353. A) The best decoy in data set, #5124, had an RMSD of 4.3 Å. B) The top scoring decoy, # 5488, had an RMSD of 7.0 Å. C) The top scoring decoy, # 3009, had an RMSD of 13.5 Å.

A B C

Figure 6-25. Results for target T306. A) The best decoy in data set, #935, had an RMSD of 8.0 Å. B) The top scoring decoy, #1604, had an RMSD of 12.5 Å. C) The top scoring decoy, #5643, had an RMSD of 13.5 Å.



A B

Figure 6-26. Results for Target T309. A) The best decoy in database, #5810, had an RMSD of 8.1 Å. B) The top scoring decoy, #210, had an RMSD of 11.0 Å.

CHAPTER 7
COMPARISONS OF GENERAL AND SPECIFIC DECOY SETS

7.1 Comparing the performance of the general and specific decoy sets on four target proteins

For most targets, many decoys from both the general and specific sets satisfied at least twelve constraints. We found it necessary to use twenty-five distance constraints to adequately distinguish between reliable and unreliable decoys while employing a constraint distance acceptance range of +/- 5 Å. To avoid dependence on the order of application of constraints, we counted the total number of constraints that each decoy satisfied. Decoys that satisfied the most constraints systematically had the lowest RMSDs.

The general decoy set contains a fixed number of decoys, 8,060,245, while the optimum size for the specific decoy set, in terms of cost and performance, was found to be 10,000 decoys. Decoys with slightly lower RMSDs are generated in the 10,000 set compared to the 1,000 set, but significantly lower RMSD decoys are not usually generated in the 50,000 set.

Rosetta generates low RMSD structures for each of our four target proteins and our scoring procedure effectively assigned these low RMSD decoys high scores. The RMSDs of the best top scoring decoys ranged from 2.4 – 4.6 Å. Several decoys satisfied all twenty-five constraints for 1ubi and 2ezk; the top scoring decoys had a large range of RMSD values. An alternate constraint set may have been more effective in distinguishing between good and bad decoys.

We had similar findings using our general decoy set. The RMSDs of the top scoring decoys ranged from 3.6 – 7.7 Å. Our final results showed that three of the four target proteins had successful structure predictions. Such low resolution structures may be used as starting points in density generation for X-ray structures[111] and may also be useful in determining protein functions.[105] We also analyzed the RMSDs for several proteins and found that in general the

average RMSD range for decoys in our database is ~15 Å. Like the PDB, our database contains many semi-redundant structures. Removal of such decoys may further decrease the search time of an already fast screening process.

7.2 CASP7 results

We studied twelve CASP targets using the general and specific decoy sets. Using the general set, the lowest RMSD decoy for five of the CASP7 targets satisfied all distance constraints. Three targets had low RMSD decoys but were not assigned the highest score while four targets did not have any low RMSD decoys in the database. Because constraints are chosen with the aid of a secondary structure prediction method, predicting incorrect secondary structure can often lead to poor constraint choices which results in bad structure predictions. When comparing our method to the other methods used in CASP7, we find that our method's performance was decent. It does quite well for some of the hardest targets, like T309, and not well for some of the easiest targets, like T288, T340, and T359.

Some targets had no low RMSD decoys in the general decoy set. Two reasons for this occurrence are: (1) the target contained an unusual protein fold not seen in the PDB or (2) the protein fold was excluded from the database. These protein folds may be more common for smaller structures but parent proteins with fewer than 100 residues were not used in decoy generation. Another explanation is that the most similar proteins in the PDB were missing small fragments of structure, thereby excluding them from the decoy set.

For the specific decoy sets, low RMSD decoys were generated for all but two targets, T309 and T306. A low RMSD decoy was generated for target T309 (5.7 Å) in the general decoy set, but for target T306, the best decoy in the general set had an RMSD of 8.1 Å. Also, for ten of the twelve targets, the specific set had lower RMSD decoys compared to those generated in the general set. The general decoy set generated better decoys for Target T309 and T359.

The discrimination process was equally effective for both types of decoy sets. Three targets (T288, T359, T335) had successful predictions using both decoy sets. Most methods used in CASP7 performed well for these targets as well. Target T306 did not have low RMSD decoys in either set and it was very difficult for most other groups to predict. The three targets for which low RMSDs were generated but were not found in the general decoy set (T348, T349, T358) and two targets for which the general decoy set did not have low RMSD decoys (T311, T363), had low RMSD predictions using the specific decoys sets. Using the general decoy set, T309 and T340 had low RMSD decoys with top scores, whereas use of the specific decoy set did not result in successful predictions. Finally, for T353, the specific decoy set generates a low RMSD decoy but it is not assigned a top score, while the general decoy set does not generate a low RMSD decoy. When both methods are considered, ten of the twelve targets had successful predictions.

In both types of decoy sets, whenever high RMSD decoys satisfied all constraints it was because the decoy and the target had regions of great similarity. In those cases where we selected a reference structure and chose constraints involving atoms from that region, such high RMSD high scoring decoys were common, especially when the reference structures were included in the region of similarity. Better results were seen when a set of residues were selected and constraints were chosen between them.

CHAPTER 8
INTRODUCTION: AZOBENZENE ISOMERIZATION[a]

8.1 Isomerization Mechanism

Azobenzene can adopt cis and trans conformations in the electronic ground state with the trans isomer lower in energy by approximately 0.6 eV.[164] The trans to cis energy barrier was found experimentally to be about 1.6 eV.[165] Azobenzene is known to undergo a reversible photoisomerization between these conformations. A trans to cis isomerization occurs upon excitation at 365 nm (3.40 eV) and a cis to trans isomerization takes place at 420 nm (2.95 eV).[166] A thermally induced cis to trans isomerization is also possible in the ground state. Due to their facile inter-conversion at appropriate wavelengths, azobenzenes have the potential to be used in optical switching and image storage devices[167-170] as well as molecular scissors[171] and as targets for coherent control in molecular electronics.[172]

There are two pathways by which isomerization is thought to take place. The rotation pathway occurs by an out of plane torsion of the CNNC dihedral angle labeled ϕ in Figure 8-1. The inversion pathway involves an in-plane inversion of the NNC angle between the azo group and the adjacent carbon of the benzene ring. The inversion angle is labeled φ in Figure 8-1. An interesting and somewhat puzzling aspect of the photochemistry of azobenzenes is the difference in trans to cis quantum yield upon excitation to the dark $S_1(n\pi^*)$ state (Φ= 0.20 [166,173] – 0.36 [174]) and bright $S_2(\pi\pi^*)$ state (Φ=0.09 [166,173] – 0.20 [174]). Even within this large experimental range, $\Phi(S_1)$ is clearly larger than $\Phi(S_2)$. When the rotation pathway is blocked by restricting the NN bond rotation with a crown ether,[175] cyclophane structure,[176] or within a cyclodextrin cavity,[177]

---

[a]Adapted with permission from Crecca, C. R., Roitberg, A. E. Theoretical Study of the Isomerization Mechanism of Azobenzene and Disubstituted Azobenzene Derivatives, J. Phys. Chem. A, 2006;110:8188-8203 Copyright 2006 American Chemical Society

the difference in quantum yield disappears. This observation led to the belief that isomerization occurs by different mechanisms after the $n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ excitations.

Most researchers agree that the inversion mechanism dominates in the ground state,[178-181] but until recently there was much debate over which mechanism dominates after excitation to each excited state. Monti's[178] minimal basis set CI calculations provided the first theoretical explanation: excitation to S1 resulted in isomerization via the inversion pathway while the rotation pathway dominated after S2 excitation. His potential energy curves were adopted by most experimentalists and used to explain their results.

Time-resolved UV-visible absorption spectroscopy of azobenzene by Lednev shows that upon excitation of trans-azobenzene at λexc = 280 to 347 nm, two transients are formed.[182-184] One was determined to be fast decaying, 1 ps, corresponding to the S_2 state and the other was longer-lived, 10 – 16 ps, corresponding to the S_1 state. Lednev used Monti's potential energy curves to explain his results. Therefore, these transients have been assigned assuming the rotational pathway dominates after S_2 excitation.

Fujino[185] performed time-resolved Raman spectroscopy to show that the S_1 state that formed after S_2 excitation had a similar NN stretching frequency as that of the S_0 state. This indicates the NN double bond remains intact after the excitation and therefore provides evidence for the inversion mechanism in the S_1 state. In later work, Fujino[186] presented results from a time-resolved fluorescence experiment that denied the existence of a rotational pathway that starts from the S_2 state, in contrast with Monti's work. They showed that isomerization always occurs in the S_1 state regardless of excitation wavelength. In order to explain the differing quantum yields, he proposed an additional relaxation channel that must be opened upon S_2 excitation and produces mostly trans isomers.

Much theoretical work has been done to investigate the photochemistry of azobenzene. Cattaneo and Persico[179] performed complete active space self-consistent field (CASSCF) and CIPSI calculations to generate potential energy curves of the ground and excited states. Ishikawa et al.[187] obtained three-dimensional potential energy surfaces of S_0, S_1, S_2, and S_3 states using CASSCF and multireference configuration interaction method with singles and doubles (MRCISD). Quennville[188] used CASSCF to generate potential energy curves for the lowest five electronic states. Tiago et al.[189] performed two-dimensional surface scans for S_0, S_1, and S_2 using TDDFT. Ciminelli et al.[190] used a combination of Tully's surface hopping approach with a direct semiempirical calculation to study the dynamics in the excited states. Cembran et al.[191] calculated the lowest singlet and triplet excited state PES along the torsion pathway using complete active space with second-order perturbation theory (CASPT2). Gagliardi et al.[192] also focused on the torsion pathway but used MS-CASPT2 and TDDFT. Diau[193] used CASSCF to look at the inversion, rotation, and concerted-inversion pathways on the S_1 surface.

The most recent theoretical conclusions agree that the $n\rightarrow\pi^*$ state has a slight inversion barrier and a nearly barrierless rotation pathway.[179,180,187,189,190,193] Several researchers have found an S_1-S_0 conical intersection along the rotation pathway with a CNNC dihedral angle of ~90.0°.[187-191,193] It is generally agreed that when excited to the S_1 state, relaxation to the S_0 state occurs through the conical intersection along the midpoint of the rotation pathway.[187,191,193] Recent experimental work has shown support for this mechanism[194]. The comprehensive studies of Fujino and Tahara[185] showed that isomerization does not occur directly on the S_2 state, but that it relaxes to a lower lying excited state, where it then isomerizes. Some calculations point to an S_2-S_1 conical intersection near the trans-azobenzene Franck-Condon region which leads to a direct S_2 to S_1 relaxation.[188,190]

Many models have been unable to explain the difference in quantum yield that is seen upon excitation to the S_2 state. Diau proposed a new isomerization pathway that is open after S2 excitation.[193] This channel produces more trans isomers than cis thereby lowering the trans to cis quantum yield. This mechanism is explored in our studies.

In addition to investigating the preferred isomerization mechanism, we also look at how substituting the phenyl rings of azobenzenes affects the isomerization process. In order to study these effects, we examined the pathways by generating potential energy surfaces of the ground and excited states of azobenzene [Azo] and four of its derivatives, 4,4'-diaminoazobenzene [Azon], 4,4'-nitro-aminoazobenzene [AzoNO_2NH_2], N-[4-(4-(Acetylamino)phenylazo)phenyl]-acetamide [Azonco], and 4,4'-dinitroazobenzene [AzoNO_2NO_2] (Figure 8-2). The azobenzenes will be from now on be referred to by the name in brackets.

Absorption spectroscopy by Blevins and Blanchard on the Azo, Azon, and Azonco systems suggest that the ground state isomerization barrier is reduced when electron-donating substituents are placed on the benzene rings.[180] Our results, however, indicate that electron-donating groups, like NH_2 and $HNCOCH_3$, increase the ground state inversion barrier while electron withdrawing groups, like NO_2, decrease it. Lack of solvent effects in our calculations may be the reason for these discrepancies as will be discussed further in this paper.

8.2 Applications of Azobenzenes in Biomolecules

Recently, a photoswitchable molecular glue for DNA has been developed which can reversibly control the hybridization of mismatch-containing DNAs with the aid of an external light stimulus.[195] These small synthetic molecules bind specifically to mismatch DNA and serve to stabilize the mismatched DNA duplex, thereby acting as a glue holding together two single stranded DNAs. Azobenzene was incorporated into naphthyridine carbamate dimmers, which bind specifically to GG-mismatches in DNA. When the azobenzene undergoes isomerization,

the positions and orientations of the naphthyridines will also change and therefore enable to adherence of two single-stranded DNAs that contain the GG-mismatch. The stabilization of the DNA duplex by the glue was evaluated by melting temperature comparisons. The cis-azobenzene-containing glue stabilized the GG mismatch DNA more strongly than the trans complex. It was also found that the cis complex disassembled upon cis to trans isomerization by 430 nm photoilumination. Thus, this reversible, photoswitchaable molecular glue for DNA has the potential to be used in controlling biological functions triggered by DNA hybridization. It may also be useful in the reversible construction of DNA-based nanoarchitectures.

Azobenzene has also been incoporated into an ionotropic glutamate receptor which acts as a photoswitch and controls an ion channel in cells.[196] The switch covalently modifies target proteins and can reversibly present and withdraw a ligand from its binding site by the photoisomrization of azobenzene. Upon photoswitching to the active state, a tethered glutamate is placed near the binding site. The photostationary state can be altered using different wavelengths of light thereby setting the fraction of active channels in an analog fashion. The switch can be turned on with short pulses at one wavelength, kept on in the dark for a few minutes, and turned off with long pulses at another wavelength. In this way, sustained activation with minimal radiation is achieved. The process provides quick and reversible control of protein function.

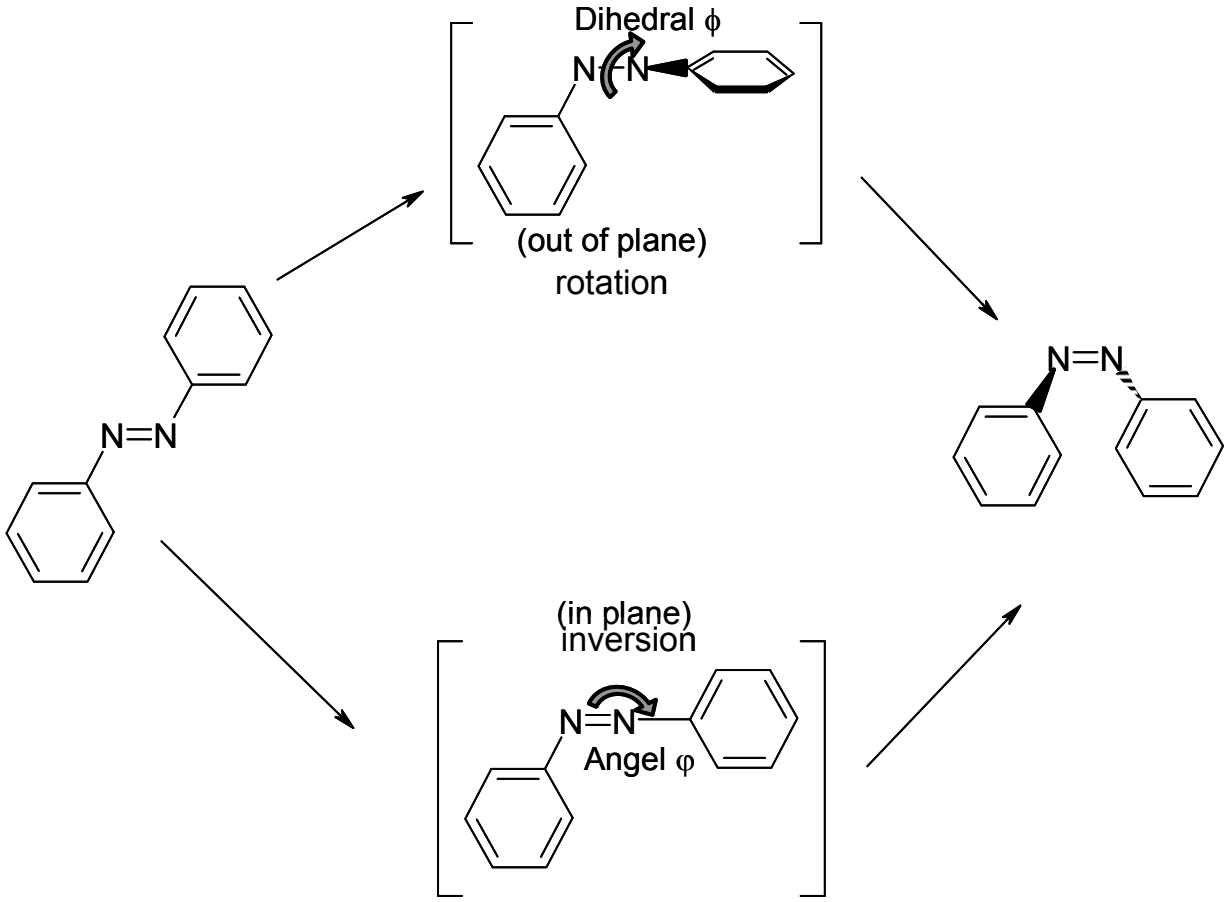


Figure 8-1. Schematic diagram of the rotation and inversion pathways of the trans → cis isomerization of azobenzenes. The rotation pathway is obtained by a torsion of the azo group around the CNNC dihedral angle ϕ. The inversion pathway is obtained by an in-plane inversion of the NNC angle (angle φ) formed between the azo group and the attached carbon of one of the benzene rings.

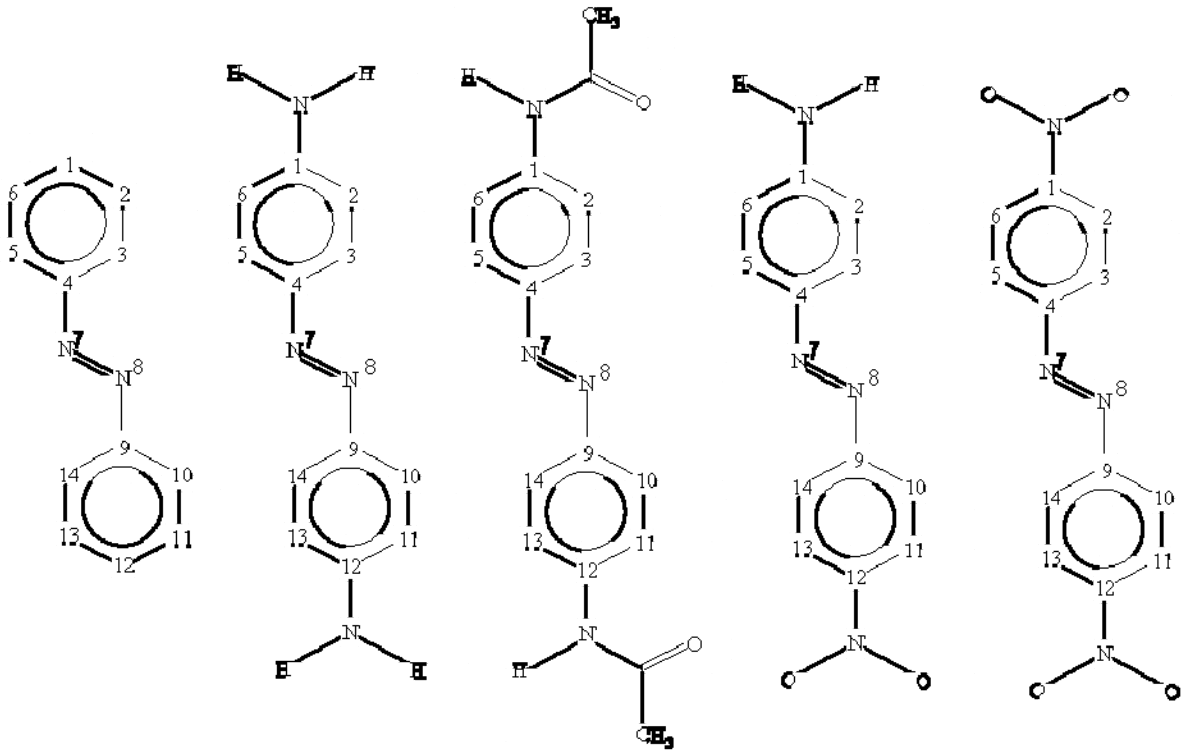


Figure 8-2. Structures of compounds investigated in this work: (a) Azo (b) Azon (c) Azonco (d) $AzoNO_2NH_2$ (e) $AzoNO_2NO_2$. This numbering scheme will be referred to throughout the text.

CHAPTER 9
COMPUTATIONAL DETAILS[b]

9.1 Ground-State Calculations

All calculations were performed using Gaussian 03.[197] All ground-state geometries were computed using *ab initio* density-functional theory with the B3LYP[198] functional and the 6-31G* basis set[199] as this method was previously found to accurately reproduce experimental results.[200]

To investigate the rotation and inversion pathways, the potential energy surface was generated by scanning the NNC angle (angle φ in Figure 8-1, 7-8-9 in Figure 8-2) from 80.0° to 180.0° and the CNNC dihedral angle (dihedral angle ϕ in Figure 8-1, 4-7-8-9 in Figure 8-2) from -40.0° to 220.0° at a 10.0° interval. For each calculation, the NNC angle and the CNNC dihedral angle were fixed at the appropriate values while the rest of the degrees of freedom were optimized. The remaining points in the potential energy surface were found through symmetry.

The potential energy surface for the concerted inversion pathway was generated in the same manner except the NNC and CNN angles were scanned synchronously. Two potential energy surfaces were generated for the 4,4'-nitro-aminoazobenzene due to its asymmetrically substituted benzene rings. Azo(NO_2)NH_2 refers to the surface with NO_2 on the same side as the NNC angle being inverted (9-8-7 in Figure 8-2) while AzoNO_2(NH_2) represents the surface with NH_2 on the same side as the inverted NNC angle (4-7-8 in Figure 8-2).

Charges were calculated using the CHelpG method to determine the electron donating or withdrawing nature of each substituent. Electron donating groups were identified as those that showed a decrease in charge on the ortho and para positions of the substituted azobenzene

[b] Adapted with permission from Crecca, C. R., Roitberg, A. E. Theoretical Study of the Isomerization Mechanism of Azobenzene and Disubstituted Azobenzene Derivatives, J. Phys. Chem. A, 2006;110:8188-8203 Copyright 2006 American Chemical Society

compared to the charge on the unsubstituted azobenzene. While electron withdrawing groups had a negative charge difference on the atoms in the meta positions.

9.2 Excited-State Calculations

All calculations were performed using Gaussian 03. Time dependent density-functional theory (TDDFT) with the B3LYP functional and the 6-31G* basis set were used for the excited-state calculations as they were found to give reliable results.[166] The excited-state potential energy surfaces were generated by calculating single point vertical excitation energies for each of the points in the ground-state potential energy surfaces. Vertical excitations were also calculated from the fully optimized ground state cis and trans minima.

CHAPTER 10
RESULTS: UNSUBSTITUTED AZOBENZENE[c]

10.1 Optimized Ground-State Geometry

The optimized geometries of cis and trans azobenzene were found and the results are shown in Table 10-1. The trans isomer is about 15.2 kcal mol^{-1} or 0.66 eV lower in energy than the cis isomer. This is just slightly higher than the experimental value of 0.6 eV.[164] Different experimental methods suggest different structures for the trans isomer. Electron diffraction[201] results indicate the phenyl rings of the trans isomer are 30° out of plane while the X-ray[202] data show a planar structure. Our results agree with the X-ray data as well as with the results of several theoretical calculations.[179,187,191,203] The structure of the cis isomer is less controversial. Our DFT results are very similar to both X-ray data[204] and other theoretical predictions.[179,187,189,191,200,203]

10.2 Electronic Excitation Energies

For the singlet vertical excitations of the trans isomer of azobenzene, the first transition, $n \rightarrow \pi^*$, is symmetry forbidden and therefore has a very weak oscillator strength, while the second transition, $\pi \rightarrow \pi^*$, is much more intense. The excitation energies for these transitions are shown in Table 10-2. The assignment of symmetry is done by visual inspection. Evaluation of our molecular orbitals (Figure 10-1) reveals that the first transition originates from the lone pair on the central nitrogens and is of 88% $n \rightarrow \pi^*$ character as calculated from the CI coefficients. The second transition is 78% $\pi \rightarrow \pi^*$ and is delocalized throughout the entire molecule. It has been suggested that the second excited state relaxes to the first via a conical intersection above the ground state trans minimum.

---

[c] Adapted with permission from Crecca, C. R., Roitberg, A. E. Theoretical Study of the Isomerization Mechanism of Azobenzene and Disubstituted Azobenzene Derivatives, J. Phys. Chem. A, 2006;110:8188-8203 Copyright 2006 American Chemical Society

The TDDFT calculated energy for the $S_1 \leftarrow S_0$ transition of trans azobenzene, 2.55 eV, is fairly close to that of the known experimental value, 2.79 eV.[166] Although CASSCF[187] and configuration interaction by perturbative iterative selection (CIPSI)[179] calculations have given values that agree slightly better with experiment for this transition, 2.85 eV and 2.81 eV respectively, the $S_2 \leftarrow S_0$ transition is much better described by TDDFT with an energy of 3.77 eV compared to an experimental value of 3.95 eV. CASSCF predicts an energy of 7.62 eV and the CIPSI energy is 4.55 eV. TDDFT consistently predicts slightly lower energies than the experimental values while the CASSCF values are generally much higher. These values are summarized in Table 10-2.

The $S_1 \leftarrow S_0$ transition occurs at about the same energy for both trans and cis. Unlike the trans excitations, however, the $S_1 \leftarrow S_0$ transition from the cis isomer shows slight intensity due to the loss of symmetry making the transition allowed. The $S_2 \leftarrow S_0$ transition from the cis isomer is much less intense and slightly higher in energy than that of the trans isomer.

10.3 Potential Energy Surfaces

10.3.1 Ground State

A ground state three dimensional potential energy surface and a contour map were calculated for azobenzene (Figure 10-2). The surface is very symmetric with two cis and two trans minima. Cis to trans barrier heights were determined from these plots by finding the energy of the highest point on the potential energy surface along the pathway and subtracting from it the energy of the cis minimum. Proper identification of these points as true transition states was done checking for the existence of only one imaginary frequency in normal modes analysis. The peak along the inversion pathway (angle reaction coordinate) was taken to be at an angle of 180.0° and a dihedral angle of 180.0° and is represented in Figure 10-2b by point 1. The peak of the rotational pathway (dihedral angle reaction coordinate) was taken to be at a

dihedral angle of 90.0° while the angle was the same as that of the trans minimum, 110.0°. In the rotation pathway, the peak was a saddle point and is labeled point 2 in Figure 10-2b.

Azobenzene is known to undergo a thermal cis to trans isomerization in the ground state so only the cis barriers will be discussed. The barrier along the inversion pathway, 24.9 kcal mol^{-1}, was lower than that of the rotation pathway, 36.2 kcal mol^{-1}, indicating that in the ground state, the inversion mechanism is favored. This is in agreement with previous reports.[178-181] A cis to trans barrier height for azobenzene was measured experimentally[180] to be 25.8 kcal mol-1, in good agreement with our results.

We can explain the difference in energy barriers between mechanisms by looking at how the NN distance changes along each pathway. Along the inversion pathway, the NN distance decreases (increases in bond order) from the trans isomer to the transition state (point 1 in Figure 10-2B) and then increases in length (decreases in bond order) as it approaches the cis isomer. The inversion transition state shows the strongest NN bond along the pathway. The opposite trend is seen along the rotation pathway. The NN distance increases from the trans isomer to the rotation transition state (point 2 in Figure 10-2B) and then decreases in length as it approaches the cis isomer. The NN distance found in the rotation transition state is approximately that of a single bond. There is a high energy cost involved in a decrease of the NN bond order in the rotation pathway which is seen as an increase in the energy barrier.

10.3.2 Excited State 1 (n → π*)

Potential energy surfaces and contour maps were calculated for the first two excited states (Figure 10-3). Our surfaces are similar to those of previous calculations.[187,189] Vertical excitations from the trans minima reach the points labeled 1 and 4 while excitations from the cis minima arrive at the points labeled 3 and 6. Points 2 and 5 depict the placement of the S_1 minima.

10.3.2.1 Rotation pathway

There is essentially no energy barrier along the rotation pathway of the first excited state as has also been reported in previous calculations.[179,180,187,189,190,193] The potential energy surface along this pathway has only a shallow slope above the area corresponding to the trans minimum (from points 1 to 2 and 4 to 5 in Figure 10-3B), 0.21 kcal (mol*degree)$^{-1}$, and a very steep slope on the cis side (from points 3 to 2 and 6 to 5 in Figure 10-3B), 0.33 kcal (mol*degree)$^{-1}$. These slopes are also shown schematically in Figure 10-4. The figures suggest that when excited from the cis conformation there is a much faster relaxation to the excited state minimum than if excited from the trans conformation. This phenomenon has been shown experimentally by femtosecond transient absorption measurements.[205]

A conical intersection was found between the ground and first excited states. It can be seen when the minimum of the excited state is very close in energy to the maximum barrier height along the rotation pathway in the ground state as can be seen in Figure 10-5. We have located our conical intersection at an NNC angle of 140.0 and a CNNC dihedral angle of 90.0 (point 5 in Figures 10-2b and 10-4a). The location of this conical intersection is in agreement with several other groups.[187-191,193] The splitting between the surfaces is estimated to be 0.65 kcal mol^{-1}.

Stilbene, which can only isomerizes via the rotation mechanism, has been found to have an S_1-S_0 conical intersection along the midpoint of the rotation pathway and is also known to have an isomerization yield of 0.5. It is interesting to find that azobenzene has a conical intersection near the same location yet shows a very different quantum yield. This can be explained by looking at the difference in slope on the S_1 surface on either side of the conical intersection in azobenzene. As mentioned previously, the S_1 slope above the cis minimum (point 6 in Figure 10-

4) is greater than the corresponding slope on the trans side (point 4 in Figure 10-4). The crossing probability close to the conical intersection can be related to the non-adiabatic coupling between S_0 and S_1 [206] written as $d_{S1-S0} = \langle \varphi_{S0} | \nabla \varphi_{S1} \rangle$. A larger slope corresponds to a larger change in wavefunction (right side of formula). There is a greater probability, therefore, of jumping from S_1 to S_0 when starting from the cis side rather than the trans side resulting in more trans isomers in the S_0 state, because the transition carries the momentum from the excited state. In other words, while oscillating on the S_1 surface near the conical intersection, more relaxation occurs when the wave packet moves from point 6 to point 5 than from point 4 to point 5, depositing more population on the trans side than on the cis side of the ground state surface, hence producing a quantum yield lower than 0.5. The slopes (cis and trans sides) on the S_1 surface of stilbene are essentially equal giving rise to more similar S_0 and S_1 wavefunctions than those of Azobenzene. The probability of jumping from S_1 to S_0 is equal when coming from either side of the conical intersection in the case of stilbene. This results in the experimentally seen quantum yield of 0.5.

10.3.2.2 Inversion pathway

There is a slight trans→cis energy barrier along the inversion pathway as can be seen in Figure 10-4. The S_1 trans→cis energy barrier is 9.6 kcal mol^{-1}. There is no conical intersection between the ground and first excited state along this pathway making the inversion mechanism highly improbable. This is in agreement with previous calculations.[188,193]

Our results indicate that the isomerization can easily occur through an excitation to the first excited state, relaxation to the excited state minimum along the rotation pathway, followed by descent to either the cis or trans conformation via the conical intersection, providing for the known cis yield (0.20-0.36) after excitation to the first excited state.

10.3.3 Excited State 2 ($\pi \rightarrow \pi^*$)

The potential energy surfaces of the second excited state are shown in Figure 10-6. As in the ground state surface, cis and trans minima appear on the surface of the S_2 state along the inversion and rotation pathways. The cis minima are extremely shallow. The trans→cis energy barriers were computed in the same manner as the ground state cis→trans barriers. The inversion barrier was found to be 30.1 kcal mol^{-1} while that of the rotation pathway was 29.6 kcal mol^{-1}. Due to these substantial energy barriers, it is unlikely that isomerization occurs on the S_2 surface. Rapid relaxation from the S_2 state to the S_1 state is energetically more favorable. This is in agreement with Kasha's rule.[207] We examined energy gaps between the two states along the inversion, rotation, and concerted inversion pathways in order to investigate this process.

10.3.3.1 Rotation pathway

The possibility of a conical intersection between the S_2 and S_1 states along the rotation pathway with an angle of 117° and a dihedral angle of 180° has been previously suggested.[188] For Azo, the states differ by 23.48 kcal mol^{-1} at the trans minimum as can be seen in Figure 10-7A. We do not find a conical intersection between S_1 and S_2 along the rotation pathway and can therefore rule out this pathway as an isomerization mechanism.

10.3.3.2 Inversion pathway

A conical intersection between the S_2 and S_1 states has been previously located near the ground state trans minima[190]. While we do not find a curve crossing in this exact area, we do see the energy difference between the S_1 and S_2 states become smaller along the inversion pathway as can been seen in Figure 10-7b. This point is a few degrees away from the S_2 minima. At a CNNC dihedral angle of 180.0° and an NNC angle of 100.0°, the energy gap between the S_1 and S_2 surfaces appears to be the smallest, 15.70 kcal mol^{-1}. This energy gap may be small

enough to allow for rapid relaxation to the first excited state. This explains why experimentalists see two transients, a shorter one corresponding to the S_2 state before is relaxes to a longer lived species corresponding to the S_1 state.[182-184]

10.3.3.3 Concerted inversion pathway

The above mechanism does not explain the difference in quantum yield that is seen upon excitation at different wavelengths for unsubstituted azobenzene. To explain this process, we invoke Diau's[193] proposal of an additional isomerization channel (concerted-inversion) that is opened by exciting to the S_2 state. The concert-inversion pathway involves a synchronous inversion of the NNC and CNN angles. In our calculations, the CNNC dihedral angle is fixed at 180.0°. The concerted inversion pathway is plotted in Figure 10-7C.

As in the inversion pathway, the S_1 and S_2 surfaces are close in energy at an NNC angle of 100.0°. This energy gap is significantly smaller than that of the rotation or inversion pathway, 5.17 kcal mol^{-1}. It seems likely that rapid relaxation from the S_2 to S_1 state can occur due to this small energy gap which will again give rise to two transients as seen experimentally. A potential problem of the concerted-inversion mechanism is the existence of an energy barrier on the S_1 surface. The energy barrier (labeled b in Figure 10-7C) is measured by subtracting the energy of the S_1 minimum from the S_1 energy at the S_1-S_0 conical intersection, 31.21 kcal mol^{-1}. The available energy is calculated by subtracting the S_1 minimum energy from the S_1 energy at the S_2-S_1 conical intersection (labeled a in Figure 10-7C), 50.43 kcal mol^{-1}. There is enough energy available to overcome the energy barrier so the channel is open.

10.4 Summary of Unsubstituted Azobenzene

Excitation to the S_1 state leads to isomerization via the rotation mechanism. Our conclusion is based on the finding of a conical intersection between the S_1 and S_0 states near the midpoint of this pathway (NNC=110, CNNC=90.0). The rotation pathway has also been found

to be without a significant barrier, unlike the inversion pathway. Excitation to the S_2 state results in rapid relaxation to the S_1 surface via the conical intersection found at NNC=100 and CNNC=180 along the concerted inversion pathway. The energy gap between these surfaces is significantly smaller than those seen in other pathways. Once on the concerted-inversion S_1 surface there is an energy barrier of ~31.2 kcal mol^{-1}. Only when excitation to the S_2 state occurs is there enough energy to overcome this barrier. The conical intersection between the S_1 and S_0 states is located at NNC=170 and CNNC=180. More trans isomers would be produced because the crossing of these states is on the trans side of the potential energy curve. This is in agreement with the experimental observation of differing quantum yields upon excitations at different wavelengths. The concerted-inversion pathway has a nearly planar transition state in which the NN double bond stays intact. This explains Fujino's observation that the S_1 state formed after S_2 excitation had a similar NN stretching frequency as that of the S_0 state.[185] It should also be noted that because the S_2 state relaxes to the S_1 state at a geometry similar to that of both the electronic ground state as well as the direct S_1 excited state in the Franck-Condon region, the spectra of both S_1 states should be quite similar as seen in Fujino's work.[208] A schematic diagram of these mechanisms is shown in Figure 10-8.

Table 10-1. Optimized Geometries of cis and trans Isomers of Azobenzene

	Angles/deg			Distances/Å		Energy [a]/ kcal mol^{-1}
	∠CNNC	∠NNCC	∠NNC	d_{NN}	d_{CN}	
trans	180.0	0.0	114.8	1.261	1.419	0.0
trans X-ray[b]	180.0	0.0	114.1	1.247	1.428	
trans ED[c]	180.0	30.0	114.5	1.268	1.427	
Cis	9.8	50.3	124.1	1.250	1.436	15.2
Cis X-ray[d]	0.0	53.3	121.9	1.253	1.449	

[a]Energies are relative to the trans isomer. [b]Reference [202], [c]Reference [201]. [d]Reference [204].

Table 10-2. Vertical Excitation Energies (eV) of trans and cis Azobenzene.

		TDDFT[a]	Exp.[b]	CASSCF[c]	CIPSI[d]
trans	S_1	2.55 (0.0)	2.79	2.85	2.81
	S_2	3.77 (0.77)	3.95	7.62	4.55
cis	S_1	2.57 (0.04)	2.82	3.65	2.94
	S_2	4.12 (0.07)	4.77	8.62	4.82

[a] Intensity is in parenthesis. [b] Reference [166], [c] Reference[187], [d] Reference [179].

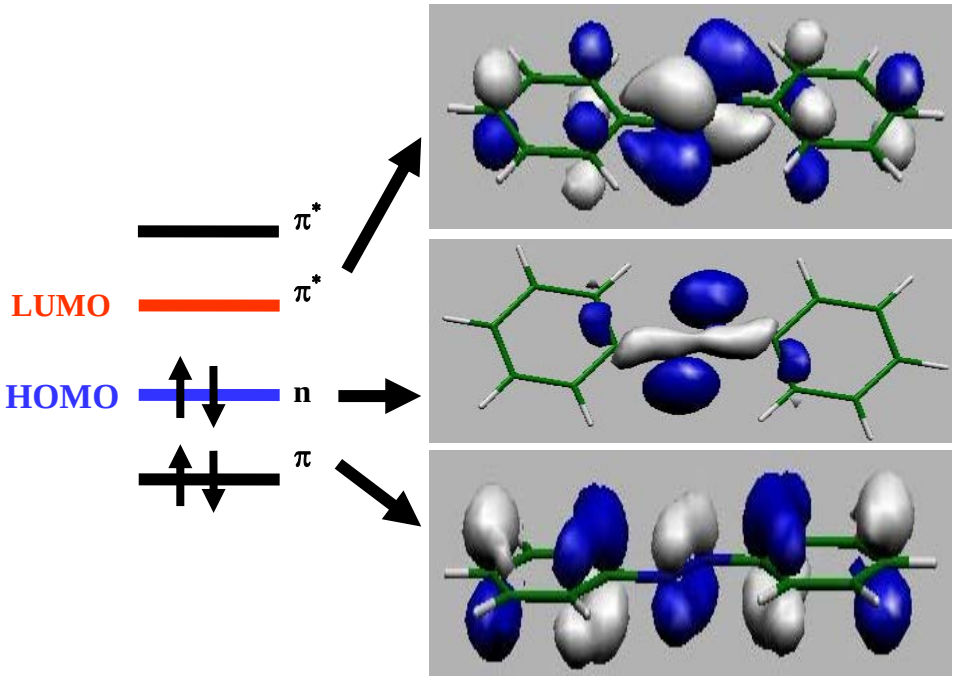


Figure 10-1. Molecular orbitals of Azo involved in the $S_1 \leftarrow S_0$ and $S_2 \leftarrow S_0$ transitions. This figure also represents the molecular orbitals of Azon and Azonco as they are very similar to those of Azo.

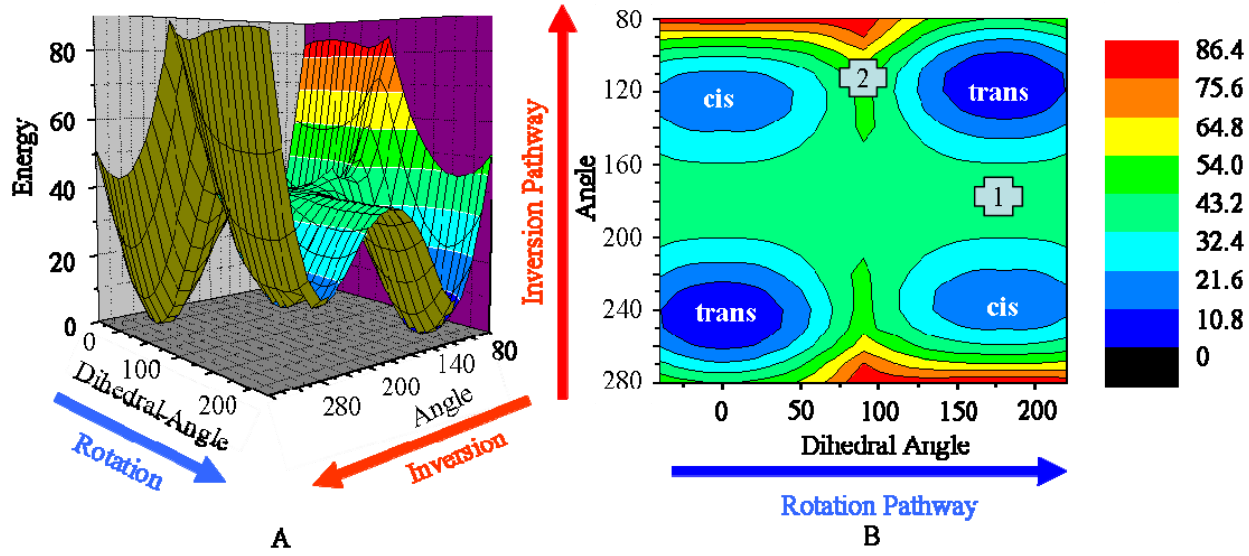


Figure 10-2. Ground state potential energy surface of Azo. A) Potential energy surface. B) Contour map. Angles in degrees, energy in kcal mol^{-1} relative to the energy of the ground state trans isomer. In B, point 1 marks the position of the inversion transition state while point 2 indicates the position of the rotation transition state. The cis and trans minima are also labeled.

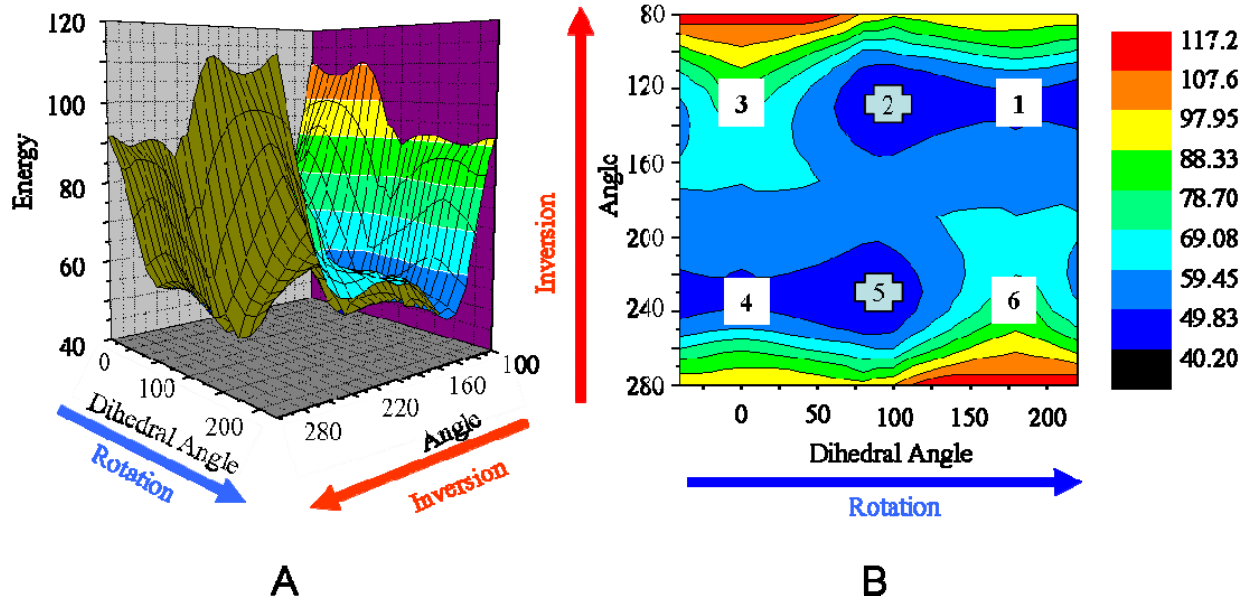


Figure 10-3. First excited state potential energy surface of Azo. A) Potential energy surface. B) Contour map. Points 1 and 4 represent where the molecule is on the S_1 surface after excitation from the ground state trans minima whereas excitation from the ground state cis minima will place the molecule at points 3 and 6. Points 2 and 5 represent the S_1 minima as well as mark the location of the S_1/S_0 conical intersection. Angles in degrees, energy in kcal mol^{-1}, relative to the energy of the ground state trans isomer.

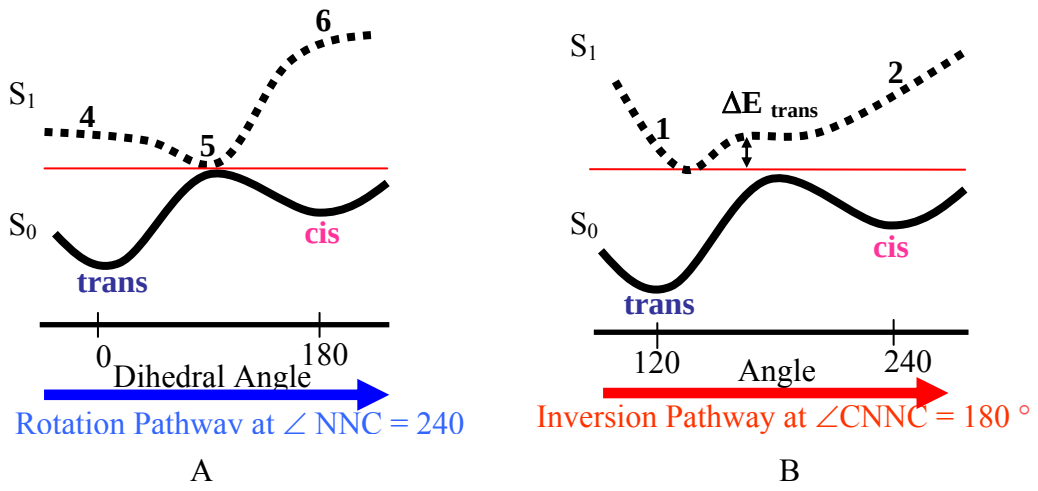


Figure 10-4. Schematic representation of pathways in the first excited state of Azo. A) The rotation pathway. B) The inversion pathway. The curves in A are along the angle of 240° while those in B are along the dihedral of 180°. The labeled points are the same as those in Figure 5b. The arrow in b depicts the inversion barrier in the S_1 state.

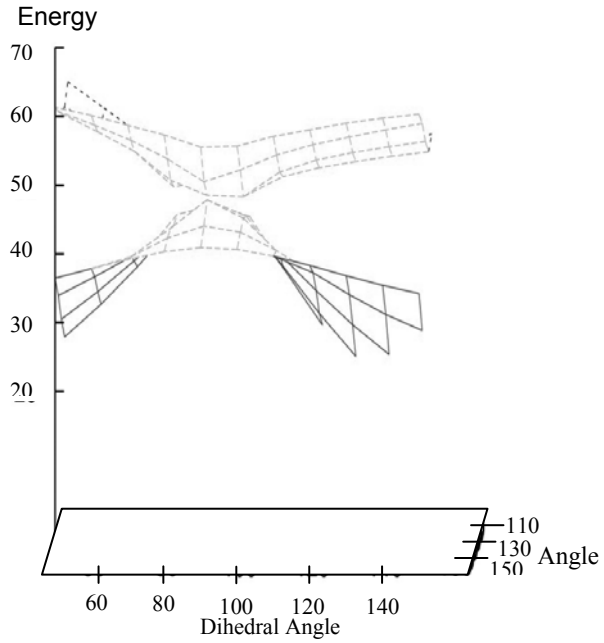


Figure 10-5. Conical Intersection of S_0 and S_1 states of Azo. Angles in degrees, energy in kcal mol^{-1}.

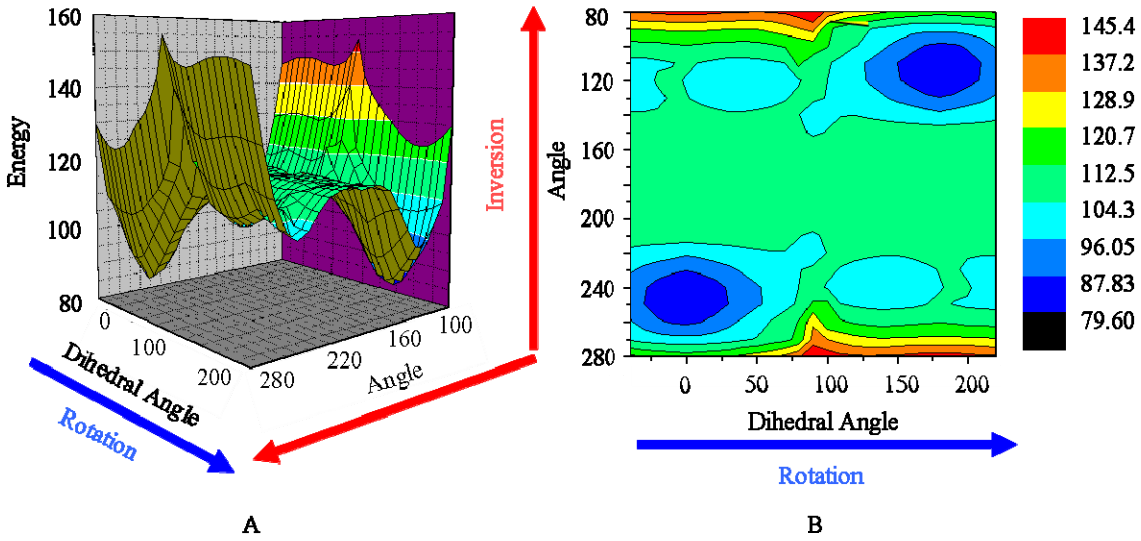


Figure 10-6. Second excited state potential energy surface of Azo. A) Potential energy surface. B) Contour map. Angles in degrees, energy in kcal mol^{-1}, relative to the energy of the trans isomer.

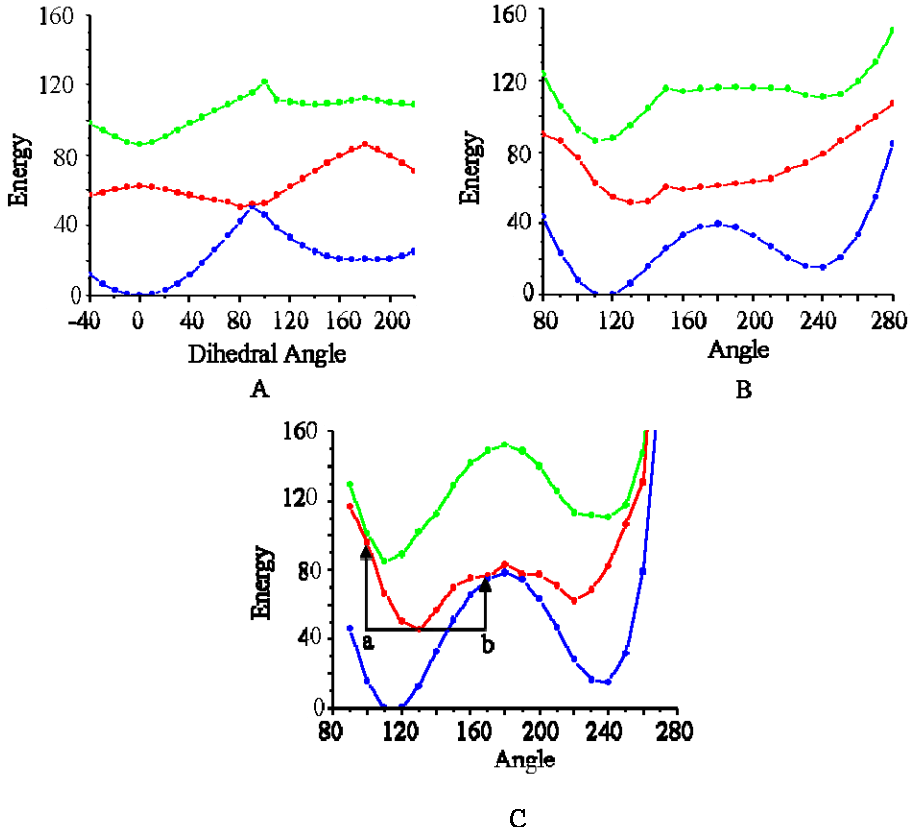


Figure 10-7. A) Rotation Pathway along the angle of the ground state minimum of Azo, ∠NNC=110°. B) Inversion and C) concerted-inversion pathways of Azo along ∠CNNC=180.0 °. S0 in blue, S1 in red, S2 in green. Angles in degrees, energy in kcal mol-1. In C, arrow a represents the available energy while arrow b represents the energy barrier.

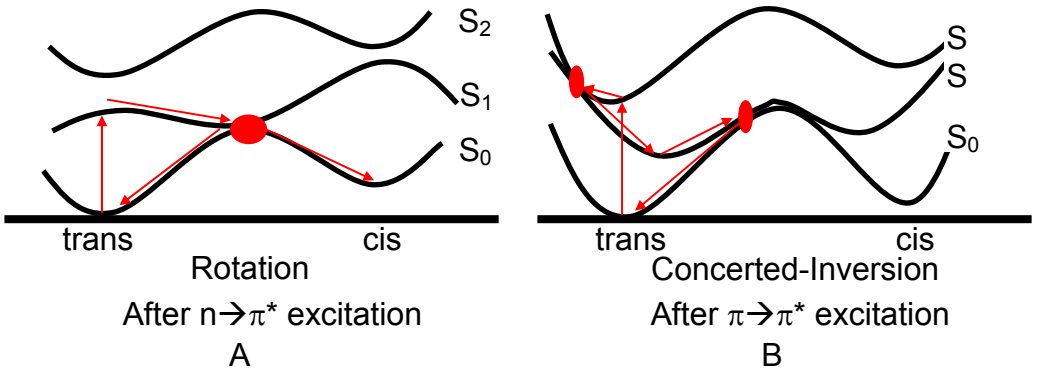


Figure 10-8. Scheme of the trans→cis isomerization process after A) n→π* excitation and B) π→π* excitation. The ovals indicate locations of curve crossings.

CHAPTER 11
RESULTS: SUBSTITUTED AZOENZENES[d]

11.1 Optimized Ground-State Geometry

The optimized geometries of the cis and trans isomers of the azobenzenes were found using the same technique as for the unsubstituted azobenzene. Important bond distances, angles, and dihedrals are summarized in Table 11-1. The values listed for Azo(NO_2)NH_2 are those of the NO_2 substituted ring, while the values for NH_2 substituted ring are represented by AzoNO_2(NH_2).

11.1.1 NN Distance

For each azobenzene studied, the NN bond is shorter for the cis isomer than the trans isomer. The NN distances were quite similar between the azobenzenes ranging from 1.260 Å to 1.267 Å for the trans isomer and 1.247 Å to 1.256 Å for the cis isomer. AzoNO_2NO_2 has the shortest NN distance for both conformations followed by Azo. The substituents appear to contribute only slightly to the NN bond as evidenced by the very small increase in bond length upon substitution of the rings with electron donating groups and a small decrease in length when substituted with electron withdrawing groups.

11.1.2 NNC Angle, CNNC Dihedral Angle, and NNCC Dihedral Angle

Like the NN distances, the NNC angles are very similar. For the trans conformation, the angles range from 114.1° to 115.6°, while the range for the cis isomer was from 124.0° to 125.5°. The CNNC dihedral angle of the trans isomers are all about the same, 180.0° while the NNCC dihedral angle were about 0.0°. The CNNC dihedral angle for the cis isomers is slightly

[d] Adapted with permission from Crecca, C. R., Roitberg, A. E. Theoretical Study of the Isomerization Mechanism of Azobenzene and Disubstituted Azobenzene Derivatives, J. Phys. Chem. A, 2006;110:8188-8203 Copyright 2006 American Chemical Society

larger in substituted azobenzenes ranging from 9.8° for Azo to 11.8° for Azon. The NNCC dihedral angle was smallest for $AzoNO_2(NH_2)$ and largest for $Azo(NO_2)NH_2$.

11.1.3 Relative Energy Differences

The difference between the cis and trans ground state energies was calculated and found to be very similar ranging from 14.8 kcal mol^{-1} for $AzoNO_2NO_2$ to 16.8 kcal mol^{-1} for Azon. Electron donating substituents appeared to increase the relative energy difference while electron withdrawing groups lowered the energy difference. The push-pull system showed a slight increase in relative energy difference when compared to Azo.

11.2 Comparison of Charges

We define electron donating groups as those that activate the ortho and para positions while electron withdrawing groups are those that activate the meta positions. Activation is determined by change in charges relative to the unsubstituted azobenzene. Blevins and Blanchard[180] suggested the CH_3CONH groups of Azonco would act as electron withdrawing substituents. Using the CHarges from ELectrostatic Potentials (CHELPG) method to calculate charges, however, we found that Azonco demonstrates electron donating behavior similar to that of Azon. The charge differences were calculated by subtracting the charge on the unsubstituted azobenzene from that of the substituted azobenzene. As can be seen in Figure 11-1, the carbons that are ortho to the substituent, C2, C6, C11, and C13, (refer to numbering scheme in Figure 10-2) have similar charge differences with an average of -0.226 for Azon and -0.234 for Azonco. The para carbons, C4 and C9, are only slightly activated with average charge differences of -0.126 for Azon and -0.117 for Azonco. The activation of the ortho carbons is enhanced by the electron withdrawing effect of the azo group. The azo group activates the positions meta to itself, which are the same as those ortho to the substituent. In effect, the azo group will act synergistically with the electron donating substituents.

Interesting behavior results when an electron donating substituent, NH_2, is placed on one benzene ring para to the azo group and an electron withdrawing group, NO_2, is placed in the para position of the other benzene ring. This creates a push pull system as in $AzoNO_2NH_2$. As seen in Azon, the NH_2 group activates the positions ortho to itself, C2 and C6, which are the same as those positions meta to the azo group as depicted in Figure 11-1. When an electron withdrawing group like NO_2 is placed on the ring para to the azo group, there is a mixing of charges. Both groups try to activate the positions that are meta with respect to themselves. This will obviously result in a conflict because the meta positions of the azo group are ortho to the NO_2 group. What we see is a difference in charge of -0.125 at C14, which is meta to the NO_2 group and -0.129 at C11, which is meta to the azo group. Therefore C11 and C14 are the activated carbons in $AzoNO_2NH_2$. $AzoNO_2NO_2$ also shows a mixing of charges. Similar results are seen in $AzoNO_2NO_2$, C3, C6, C11, and C14 are activated.

It can now be stated with confidence that Azonco and Azon have electron donating groups, $AzoNO_2NO_2$ had electron withdrawing groups and $Azo(NO_2)NH_2$ is a push pull system with both an electron donating and an electron withdrawing group.

11.3 Electronic Excitation Energies

The singlet vertical excitations of the trans isomers of the substituted azobenzenes are very similar to unsubstituted azobenzene. The first transition, $n \rightarrow \pi^*$, is symmetry forbidden and therefore has a very weak oscillator strength, while the second transition, $\pi \rightarrow \pi^*$, shows some intensity. Visual inspection is used to assign symmetry.

The excitation energies for the $S_0 \leftarrow S_1$ transition for all the azobenzenes were similar, as shown in Table 11-2. The molecular orbitals (Figures 10-1 and 11-2) show again that the first transition originates from the lone pair on the central nitrogens. Figure 10-1 can be used to

represent the molecular orbitals for Azo, Azon, and Azonco. The excitation was of nearly pure $n \rightarrow \pi^*$ character for all but $AzoNO_2NH_2$ and $AzoNO_2NO_2$. These systems show some additional charge transfer to their NO_2 substituents.

The second transition, $\pi \rightarrow \pi^*$, is delocalized throughout the entire molecule for all but the push pull system. $AzoNO_2NH_2$ shows an excitation primarily from the π orbitals of the benzene ring with the NH_2 substituent as well as from the π orbitals of the central nitrogens. As in the $n \rightarrow \pi^*$ transition, $AzoNO_2NH_2$ and $AzoNO_2NO_2$ both show a charge transfer to the NO_2 substituents. The molecular orbitals involved in the second transition are pictured in Figures 10-1 and 11-2.

$AzoNO_2NH_2$ exhibits an intense trans excitation with the smallest energy, 2.99 eV, while the $\pi \rightarrow \pi^*$ excitations of Azonco and Azon are particularly close in energy, 3.25 eV and 3.26 eV respectively. $AzoNO_2NO_2$ has an excitation of 3.48 eV. The $S_2 \leftarrow S_0$ transition of Azo is highest in energy, 3.77 eV, and the least intense of all the azobenzenes. It appears that adding both electron donating and electron withdrawing substituents to Azo decreases the excitation energy and increases the intensity of the $S_2 \leftarrow S_0$ transition.

We have found again that the first and second excited states at the optimized ground state trans geometry are very close in energy. Azo shows the largest energy gap, 1.22 eV, followed by $AzoNO_2NO_2$, 1.17 eV, and Azonco, 0.66 eV. Azon and $AzoNO_2NH_2$ have very similar energy gaps, 0.550 and 0.546 eV respectively. The energy differences between the first two excited states are summarized in Table 11-2.

The energy of the steady-state absorption spectroscopy maximum of Azo was 3.96 eV,[180] slightly higher than the TDDFT maximum of 3.77 eV. Azon showed an experimental excitation of 3.15 eV, while the calculated energy was 3.26 eV. Azonco showed an excitation of 3.41 eV,

slightly higher than the calculated energy of 3.25 eV. Both experimental and theoretical results show Azo to have the highest energy transition. TDDFT predicts the excitation energies of Azon and Azonco to be about the same, while experiment shows these energies to differ by 0.26 eV.

It is also interesting to compare differences between the cis and trans excitations. Unlike the trans excitations, the $S_1 \leftarrow S_0$ transition from the cis isomer shows slight intensity. For Azo and $AzoNO_2NH_2$, the $S_1 \leftarrow S_0$ transition occurs at about the same energy for both trans and cis. Azon and Azonco have trans $S_1 \leftarrow S_0$ excitations slightly higher in energy than cis excitations while $AzoNO_2NO_2$ shows a higher energy cis excitation.

The $S_2 \leftarrow S_0$ transition from the cis isomer is much less intense and slightly higher in energy than that of the trans isomer. There is a greater difference between the cis and trans $S_2 \leftarrow S_0$ transitions than the $S_1 \leftarrow S_0$ transitions. The greatest difference is seen in Azonco with almost 0.5 eV separating the cis and trans excitations.

11.4 Potential Energy Surfaces

11.4.1 Ground State

Ground state three dimensional potential energy surfaces and contour maps were calculated for each azobenzene (Figure 11-3). As mentioned previously, for the push pull system, $Azo(NO_2)NH_2$ represents the surface with NO_2 on the same side as the NNC angle being inverted while $AzoNO_2(NH_2)$ represents the surface with NH_2 on the same side as the inverted NNC angle. As can be seen in these figures, the ground state surfaces of the azobenzenes are very similar. Cis to trans barrier heights were determined as described in Chapter 9. The energy barriers can be found in Table 11-3. For the push pull system, the barriers for both $Azo(NO_2)NH_2$ and $AzoNO_2(NH_2)$ were considered together.

In all five systems, the barrier along the inversion pathway was lower than that of the rotation pathway, indicating that in the ground state, the inversion mechanism is still favored.

The unsubstituted azobenzene, Azo, was found to have an inversion barrier of 24.9 kcal mol^{-1}. $Azo(NO_2)NH_2$ and $AzoNO_2NO_2$ have barriers lower than Azo, 17.2 kcal mol^{-1} and 20.8 kcal mol^{-1} respectively. In both of these systems, the inversion angle is adjacent to a phenyl ring with an electron withdrawing substituent. Azonco, Azon, and $AzoNO_2(NH_2)$, each have an electron donating substituent on the phenyl ring adjacent to the inversion angle and showed higher barriers than Azo, 25.5 kcal mol^{-1}, 26.8 kcal mol^{-1},and 28.5 kcal mol^{-1} respectively. It is clear from our results that substituting the benzene ring attached to the angle being inverted with an electron donating group, raises the inversion barrier height compared to the unsubstituted azobenzene. Substituting the same ring with an electron withdrawing group lowers the barrier height. These observations can be explained upon examination of the molecular orbitals of the inversion transition state (Figure 11-4).

Each of the azobenzenes has an inversion transition state with an angle of 180° and a dihedral angle of 180°. Due to the electron donating substituents on Azon and Azonco, there is more electron density on the phenyl rings than is seen on Azo. There is therefore greater steric hindrance between the lone pairs on the central nitrogens and p orbitals of the phenyl ring adjacent to the 180° NNC angle. The steric effects cause the inversion transition state of Azon and Azonco to be higher in energy than that of Azo. $AzoNO_2NO_2$, on the other hand, has electron withdrawing substituents which accept electron density from the π orbitals of the phenyl rings. $AzoNO_2NO_2$ is slightly stabilized by the ability of the less filled π orbitals of the phenyl ring adjacent to the 180° NNC angle to accept electron density from the lone pair orbitals of the central nitrogens. The lower barrier height of $AzoNO_2NO_2$ compared to Azo is due to this stabilization.

For the push pull system, the smallest barrier appears along the inversion pathway of $Azo(NO_2)NH_2$. This suggests that the preferred mechanism of isomerization in the ground state of the push pull system is the inversion of the NNC angle that is on the same side as the NO_2 substituent. This is in agreement with the results of Kikuchi's[209] studies of a similar push-pull system, 4-dimethylamino-4'-nitroazobenzene. This system has the lowest inversion energy barrier of all the azobenzenes studied. The transition state is stabilized by the vacant orbitals of the nitro substituted phenyl rings accepting electron density from the lone pairs on the central nitrogens. The lone pairs are parallel to the vacant π orbitals on this phenyl ring. The lone pairs are also perpendicular to the occupied orbitals of the amine substituted phenyl ring which has a stabilizing effect as it minimizes the electron-electron repulsion. The combination of these effects results in the $Azo(NO_2)NH_2$ having the lowest inversion energy barrier.

Blevins and Blanchard looked at the ground state cis $\rightarrow$ trans back-conversion for Azo, Azon, and Azonco using theory and experiment. They calculated barrier heights from their experimentally measured isomerization recovery time constants. The experiments did not indicate which pathway the barriers referred to, so we will compare them to both the inversion and rotation cis to trans barriers. A barrier height of 21.2 kcal mol^{-1} was measure for Azon, 23.7 kcal mol^{-1} for Azonco, and 25.8 kcal mol^{-1} for Azo. The experimental data indicates that adding electron donating substituents decreases the energy barrier which conflicts with our results. This may be due to the lack of consideration of solvent effects in our calculations. The dipole moment of the cis isomer and the transition state will be stabilized by the polar solvent. The dipole moments were calculated and can be found in Table 11-4. For Azo, the cis isomer and the inversion transition state have approximately the same dipole moment indicating they will be equally stabilized by a polar solvent. This may explain why our calculated barrier height is

closest to the experimental value for Azo. The inversion transition states of Azon and Azonco are more stabilized by a polar solvent than their corresponding cis isomers due to their greater dipole moment. Stabilization of the transition state will lower the energy barrier as is seen when comparing our calculated results with experiment. Polar solvents will have the greatest effect on the push-pull system due to the large dipole moments that can be found in both the transition state as well as the cis isomer.

The NN distance changes along each pathway in the substituted azobenzenes follow the same trend seen for unsubstituted azobenzene (see Chapter 9). Along the inversion pathway, the NN distance is smallest at the transition state. The values of the NN distances in the transition states can be found in Table 11-5. The inversion transition state of $AzoNO_2NO_2$ has the shortest NN distance, 1.222 Å, followed by Azo, 1.226 Å, and $Azo(NO_2)NH_2$, 1.228 Å. The inversion transition state of Azonco was found to have an NN distance of 1.233 Å while that of Azon was found to be 1.241 Å. $AzoNO_2(NH_2)$ had the longest NN distance, 1.248 Å. The electron donating groups can contribute electron density to the π^* orbitals thereby decreasing the bond order and increasing the length of the NN bond compared to that of the unsubstituted azobenzene. These distances indicate that the central nitrogens of the inversion transition state have a double bond between them.

The opposite trend is seen along the rotation pathway, the NN distance is greatest at the transition state and is approximately that of a single bond. These distances can also be found in Table 11-5. $Azo(NO_2)NH_2$ has the longest NN bond distance, 1.335Å, while $AzoNO_2(NH_2)$ has the shortest NN distance, 1.290Å.

Potential energy surfaces and contour maps were calculated for the first two excited states. Figure 11-5 shows these calculations for the first excited state of all the azobenzenes. The

surface graphs of all substituted azobenzenes appear to be similar to Azo (Figure 10-2a) and are therefore not shown. Slight differences are more visible in the contour plots.

11.4.2 Excited State 1

11.4.2.1 Rotation pathway

As seen in unsubstituted azobenzene, there is essentially no energy barrier along the rotation pathway of the first excited state. There is a shallow slope above the area corresponding to the trans minimum and a very steep slope on the cis side. We can compare the excited state cis and trans energy barriers and slopes (Table 11-6) to approximate relative relaxation times. A steeper slope indicates a quicker relaxation time. We can conclude from this analysis that the lifetime of the first excited state cis isomer is shorter than that of the trans for each of the azobenzenes studied here. Azonco appears to have the steepest trans slope and we predict it will exhibit the shortest S_1 lifetime while $AzoNO_2NO_2$ has the least steep slope and is expected to have the longest S_1 lifetime.

A conical intersection was discovered in each azobenzene between the ground and first excited states. The location of the conical intersection is only slightly different between the azobenzenes. The location as well as the relative energy can be found in Table 11-7. Azonco's conical intersection is located on the trans side of the barrier. This may indicate that Azonco will have a lower cis→trans quantum yield.

11.4.2.2 Inversion pathway

We again see a trans → cis energy barrier along the inversion pathway (Table 11-8). $AzoNO_2(NH_2)$, Azon, $AzoNO_2NO_2$, and Azo have higher barrier heights, 11.5 kcal mol^{-1}, 11.1 kcal mol^{-1} , 10.4 kcal mol^{-1}, and 9.6 kcal mol^{-1} respectively. $Azo(NO_2)NH_2$ and Azonco show very small inversion barriers, 1.3 kcal mol^{-1} and 2.3 kcal mol^{-1}, making it difficult to rule out this pathway as a possible isomerization mechanism for these azobenzenes based on barrier height

alone. Lack of a conical intersection between the ground and first excited state along this pathway makes the inversion mechanism highly improbable. We can conclude that substituting the phenyl rings of azobenzene does not change the isomerization mechanism after S_1 excitation.

11.4.3 Excited State 2

The potential energy surfaces of the second excited state were also generated and can be found in Figure 11-6. Both cis and trans minima appear on this surface along the inversion and rotation pathways of each of the azobenzenes. The trans→cis energy barriers were computed as described previously and can be found in Table 11-9. These barrier heights are too substantial for isomerization to occur on this surface. Rapid relaxation from the S_2 to the S_1 surface is again expected. We will compare the energy gaps between the first and second excited states along the rotation, inversion, and concerted inversion pathways.

11.4.3.1 Rotation pathway

As depicted in Figure 11-7, in general, there is a significant decrease in the energy gap upon substitution of the benzene rings by both electron donating and electron withdrawing groups in agreement with experimental work.[210] These values can be found in Table 11-10. For Azo, the states differ by 23.48 kcal mol^{-1} above the trans minimum. Azo(NO_2)NH_2 shows the smallest energy gap of 8.89 kcal mol^{-1}. These energy gaps are still slightly too high for relaxation to occur along this pathway.

11.4.3.2 Inversion pathway

The energy difference between the states becomes smaller along the inversion pathway near the trans minima as can been seen in Figure 11-8 and Table 11-10. These points are a few degrees away from the minima of the second excited state. In general, at a dihedral angle of 180.0° and angles of 100.0°, the energy gap between the first and second excited state surfaces appears to be the smallest. Azo and AzoNO_2NO_2 have the largest energy gaps, 15.70 kcal mol^{-1}

and 16.01 kcal mol^{-1} respectively. The other azobenzenes show significantly smaller energy gaps, under 4.67 kcal mol^{-1}, making this a very probable pathway.

11.4.3.3 Concerted-inversion pathway

This pathway is depicted in Figure 11-9. Energies of the S_1 and S_2 minima, conical intersections, barrier heights, and available energy can be found in Table 11-11.

Azon, Azonco, and $AzoNO_2NH_2$: For these three azobenzenes, excitation to the S_2 surface in the franck-condon region results in excitation to the S_2 minimum at NNC=110.0 and CNNC=180. This is also the location of the smallest S_2-S_1 energy gap along this pathway, 2.79 kcal mol^{-1} for Azon, 6.24 kcal mol^{-1} for Azonco, and 3.49 kcal mol^{-1} for $AzoNO_2NH_2$. These energy gaps are extremely small and would allow for rapid relaxation from the S_2 surface to the S_1 surface.

As seen in unsubstituted azobenzene, a large energy barrier is seen on the S_1 surface of each of these systems. The energy barriers were measured by subtracting the energy of the S_1 minimum from the S_1 energy at the S_1-S_0 conical intersection (arrow b in Figure 11-9). The available energy is calculated by subtracting the S_1 minimum energy from the S_1 energy at the S_2-S_1 conical intersection (arrow a in Figure 11-9). In each case, the available energy is less than the energy barrier. It is highly improbable that this channel is open for Azon, Azonco, and $AzoNO_2NH_2$. However, highly polar solvents may lower the S_1 energy at the S_1-S_0 conical intersection, which may lower the energy barrier enough to allow for the opening of this channel.

$AzoNO_2NO_2$: $AzoNO_2NO_2$ is quite similar to Azo. The smallest S_1-S_2 energy gap of 6.35 kcal mol^{-1} is found at NNC=100.0° and CNNC=180.0°. This energy gap is smaller than that of the rotation and inversion pathways. The available energy was calculated to be 53.90 kcal mol^{-1} while the barrier was found to be 33.70 kcal mol^{-1}. There appears to be sufficient energy to

overcome the barrier. More trans isomers would be formed as the S_1-S_0 conical intersection appears at NNC=170.0° and CNNC=180.0°.

11.5 Summary of Substituted Azobenzenes

As seen for Azo, the rotation pathway dominates the isomerization process after excitation to the S_1 surface as evidenced by the conical intersection between the S_1 and S_0 states near the midpoint of this pathway (NNC=110, CNNC=90.0) and the lack of a significant barrier. Azon, Azonco, and AzoNH_2NO_2 use the rotation pathway after excitation to the S_2 state as represented schematically in Figure 11-10. There is not enough available energy for these azobenzenes to overcome the concerted-inversion barrier. It may be possible for this channel to open in very polar solvents if the transition state is stabilized. The concerted-inversion channel is open for AzoNO_2NO_2, after excitation to the S_2 surface.

Table 11-1. Optimized Geometries of cis and trans Isomers of Azobenzenes

Structure		Angles/deg			Distances/Å		Energy[a]/ kcal mol^{-1}
		∠CNNC	∠NNCC	∠NNC	R_{NN}	R_{NC}	
trans	Azo	180.0	0.0	114.8	1.261	1.419	0
	Azon	180.0	0.1	115.0	1.267	1.409	0
	Azonco	180.0	0.0	114.9	1.265	1.411	0
	Azo(NO_2)NH_2	179.9	0.2	114.1	1.267	1.415	0
	AzoNO_2(NH_2)	179.9	0.0	115.6	1.267	1.399	0
	AzoNO_2NO_2	179.9	0.2	114.6	1.260	1.427	0
cis	Azo	9.8	50.3	124.1	1.250	1.436	15.2
	Azon	11.8	44.1	124.6	1.256	1.430	16.8
	Azonco	11.1	46.0	124.5	1.253	1.431	16.1
	Azo(NO_2)NH_2	11.5	60.4	125.5	1.254	1.423	15.6
	AzoNO_2(NH_2)	11.5	30.4	125.0	1.254	1.419	15.6
	AzoNO_2NO_2	10.2	52.2	124.0	1.247	1.432	14.8

[a] Energies are relative to their respective trans minima.

Table 11-2. Vertical Excitation Energies in eV of trans and cis Azobenzenes.

Structure		$S_1 \leftarrow S_0$ ($n \rightarrow \pi^*$)			$S_2 \leftarrow S_0$ ($\pi \rightarrow \pi^*$)			$S_2 \leftarrow S_0$ - $S_1 \leftarrow S_0$
		Energy	Intensity	% $n \rightarrow \pi^*$[a]	Energy	Intensity	% $\pi \rightarrow \pi^*$[a]	Energy Diff.
trans	Azo	2.55	0.0	88	3.77 (3.96[b])	0.77	79	1.22
	Azon	2.71	0.0	89	3.26 (3.15[b])	1.03	78	0.55
	Azonco	2.59	0.0	88	3.25 (3.41[b])	1.29	80	0.66
	AzoNO_2NH_2	2.44	0.0	85	2.99	0.86	80	0.55
	AzoNO_2NO_2	2.31	0.0	86	3.48	1.07	80	1.17
cis	Azo	2.57	0.04	78	4.12	0.07	87	1.55
	Azon	2.46	0.09	71	3.70	0.22	77	1.24
	Azonco	2.46	0.10	74	3.72	0.29	73	1.26
	AzoNO_2NH_2	2.46	0.11	60	3.17	0.09	40	0.71
	AzoNO_2NO_2	2.44	0.07	75	3.62	0.03	81	1.18

[a]The % $n \rightarrow \pi^*$ and % $\pi \rightarrow \pi^*$ values are calculated from the CI coefficients. [b]Reference[180], experimental value.

Table 11-3. The cis → trans Energy Barriers Calculated Along the Inversion and Rotation Pathways.

	Azo		Azon		Azonco		$Azo(NO_2)NH_2$		$AzoNO_2(NH_2)$		$AzoNO_2NO_2$	
	Inv.	Rot.	Inv.	Rot.	Inv.	Rot.	Inv.	Rot.	Inv.	Rot.	Inv.	Rot.
∠NNC	180	110	180	120	180	110	180	110	180	120	180	120
∠CNNC	180	90	180	90	180	90	180	90	180	90	180	90
$\Delta E^{\ddagger}_{cis}$	24.9	36.2	26.8	30.5	25.5	34.2	17.2	31.6	28.5	20.8	20.8	29.2
$\Delta\Delta E^{\ddagger a}$	11.3		3.7		8.7		3.6				8.4	

[a] $\Delta\Delta E^{\ddagger}$ is the energy difference in kcal mol^{-1} between the rotation and inversion isomerization barriers. Angles are in degrees.

Table 11-4. Dipole Moments of the inversion transition State and cis Isomer

	Dipole Moment cis	Dipole moment Inversion TS
Azo	3.22	3.22
Azon	2.61	4.44
Azonco	5.35	7.39
$Azo(NO_2)NH_2$	7.53	13.37
$AzoNO_2(NH_2)$	7.53	8.79
$AzoNO_2NO_2$	3.66	5.89

Table 11-5. NN Distances (Å) of Transition States Along the Rotation and Inversion Pathways.

	Inversion	Rotation
Azo	1.226	1.303
Azon	1.241	1.308
Azonco	1.233	1.322
$Azo(NO_2)NH_2$	1.228	1.335
$AzoNO_2(NH_2)$	1.248	1.290
$AzoNO_2NO_2$	1.222	1.297

Table 11-6. Rotational Energy Barriers in the First Excited State [a]

	Trans Barrier	Cis Barrier
Azo	18.5 (0.206)	29.8 (0.331)
Azon	11.6 (0.129)	28.6 (0.318)
Azonco	19.2 (0.213)	29.2 (0.324)
$Azo(NO_2)NH_2$	17.5 (0.194)	31.1 (0.346)
$AzoNO_2(NH_2)$	13.9 (0.154)	32.0 (0.356)
$AzoNO_2NO_2$	11.5 (0.128)	27.3 (0.303)

[a] This barrier is measured as the difference in energy between the excited state minimum and the excited state point corresponding to the ground state trans and cis minima. Energies are in kcal mol^{-1}, slope, in parenthesis, is in units of kcal mol^{-1} $degree^{-1}$.

Table 11-7. Placement and Energy of First Excited State Minimum of the Conical Intersection

	Angles/deg		Energy [a]/ kcal mol^{-1}
	∠NNC	∠CNNC	
Azo	140	90	46.0
Azon	130	90	47.0
Azonco	130	100	43.9
$Azo(NO_2)NH_2$	140	90	42.4
$AzoNO_2(NH_2)$	120	90	38.9
$AzoNO_2NO_2$	150	90	45.4

[a] Energies are relative to their respective trans ground state minimum.

Table 11-8. The trans → cis Inversion Energy Barriers in the First Excited State [a]

	ΔE_{trans}
Azo	9.6
Azon	11.1
Azonco	1.3
$Azo(NO_2)NH_2$	2.3
$AzoNO_2(NH_2)$	11.5
$AzoNO_2NO_2$	10.4

[a] These barriers were found by subtracting the energy of the excited state point above the ground state trans minimum from the energy of the excited state at an angle of 180.0° and a dihedral of 180.0°. Energies are in kcal mol^{-1}.

Table 11-9. The trans → cis Energy Barriers Calculated Along the Inversion and Rotation Pathways on the Second Excited State Surface.

	Azo		Azon		Azonco		$Azo(NO_2)NH_2$		$AzoNO_2(NH_2)$		$AzoNO_2NO_2$	
	Inv.	Rot.	Inv.	Rot.	Inv.	Rot.	Inv.	Rot.	Inv.	Rot.	Inv.	Rot.
∠NNC	180	110	180	110	180	120	180	120	180	110	180	110
∠CNNC	180	90	180	90	180	90	180	90	180	90	180	90
$\Delta E^{\ddagger}_{trans}$	30.1	29.6	40.5	46.2	40.2	34.9	43.1	28.4	27.1	31.1	14.5	27.7
$\Delta\Delta E^{\ddagger a}$	0.5		5.7		5.3		1.3				13.2	

[a] $\Delta\Delta E^{\ddagger}$ is the energy difference between the rotation and inversion isomerization barriers. Angles are in degrees. Energies are in kcal mol^{-1}.

Table 11-10. Energy Differences between S_1 and S_2

	Energy Gap at A=110° D=180° (rotation)	Energy Gap at D=180° A=100° (Inversion)	Energy Gap at D=180° A=100° (Concerted-Inversion)[a]
Azo	26.43	15.70	5.17
Azon	22.06	0.69	2.79
Azonco	17.12	3.56	6.24
$Azo(NO_2)NH_2$	8.89	4.67	3.49
$AzoNO_2(NH_2)$	17.30	2.30	3.49
$AzoNO_2NO_2$	22.83	16.01	6.36

[a]For the concerted-inversion pathway, $Azo(NO_2)NH_2$ and $AzoNO_2(NH_2)$ are the same. Energies are in kcal mol^{-1}.

Table 11-11. Energies of the S_1 and S_2 Minima, Conical Intersections, Barrier Heights, and Available Energy [a].

	S_2 min	S_2 at S_1-S_2 CI	S_1 min	S_1 at S_0-S_1 CI	S_1 barrier[b]	Available Energy[c]
Azo	84.95	100.99 (5.17)	45.39	76.60 (1.64)	31.21	50.43
Azon	73.19	73.19 (2.79)	49.68	77.63 (6.78)	27.95	20.72
Azonco	73.27	73.27 (6.24)	46.11	81.58 (4.56)	35.47	20.92
$Azo(NO_2)NH_2$	67.72	67.72 (3.49)	41.79	71.12 (2.57)	29.33	22.44
$AzoNO_2NO_2$	79.28	98.78 (6.35)	38.53	72.23 (7.12)	33.70	53.90

[a]Energies are in kcal mol-1 and are relative to their respective trans minimum. The numbers in parenthesis refer to the energy gaps between the two states. [b]The S1 barrier is measured as the difference between the S1 minimum energy and the S1 energy at the S_0-S_1 conical intersection. [c]The available energy is the difference between the energy of S_1 at the S_2-S_1 conical intersection and the energy of the S_1 minimum. If the available energy is greater than the S_1 barrier, the concerted-inversion channel can be used.

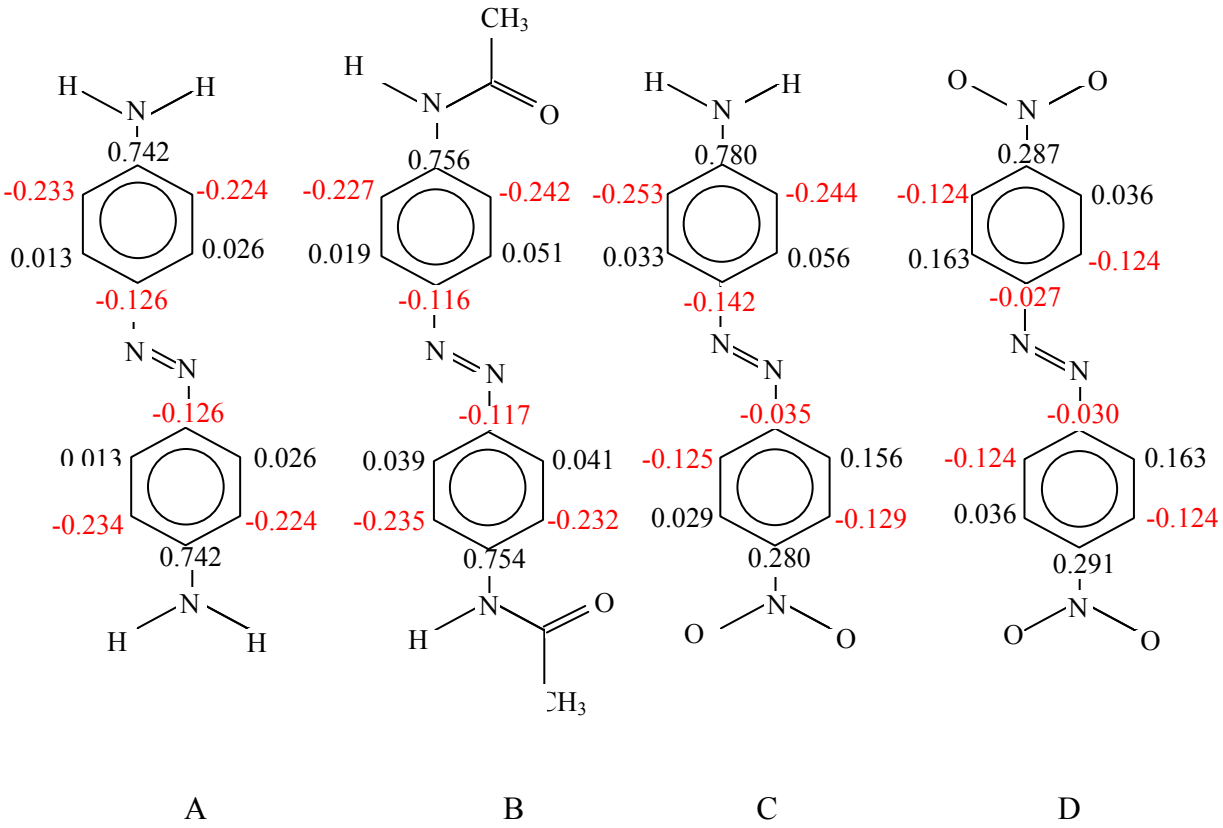


Figure 11-1. Comparison of charge differences in trans isomers. A) Azon. B) Azonco. C) $AzoNO_2NH_2$. D) $AzoNO_2NO_2$. Charge differences were calculated by subtracting the charge on the unsubstituted azobenzene from that of the substituted azobenzene. A negative charge differences (highlighted in bold) indicates that the position has been activated.

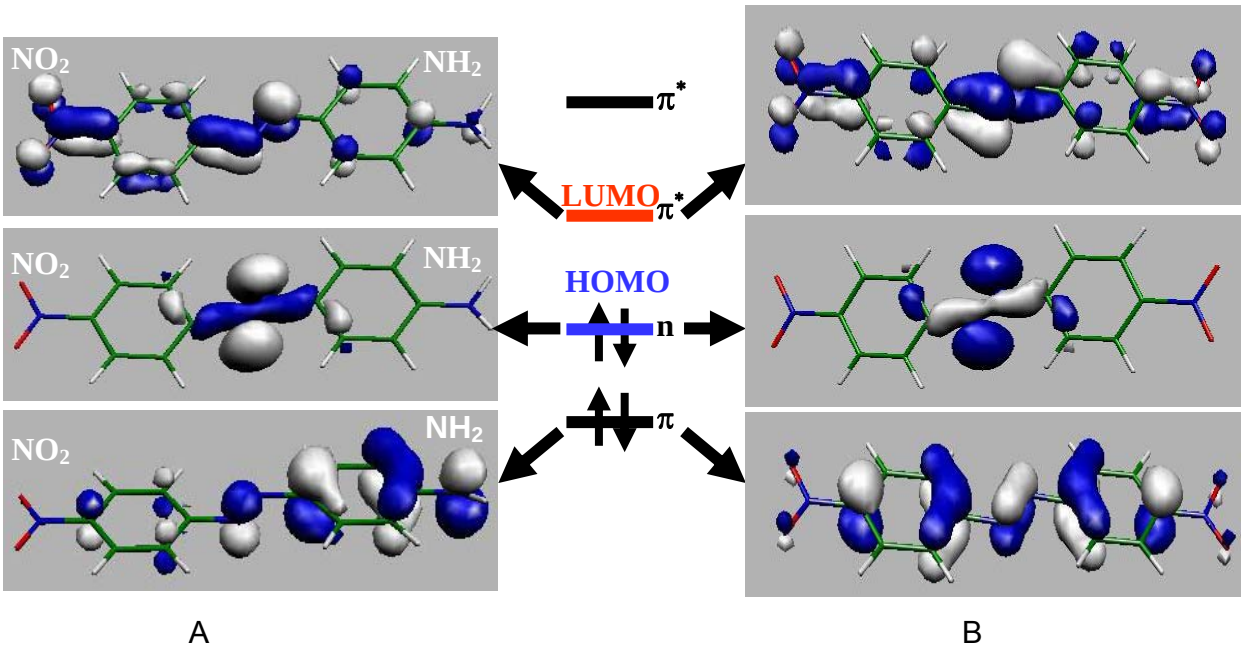


Figure 11-2. Molecular orbitals involved in the $S_1 \leftarrow S_0$ and $S_2 \leftarrow S_0$ transitions. A) AzoNO_2NH_2 B) AzoNO_2NO_2

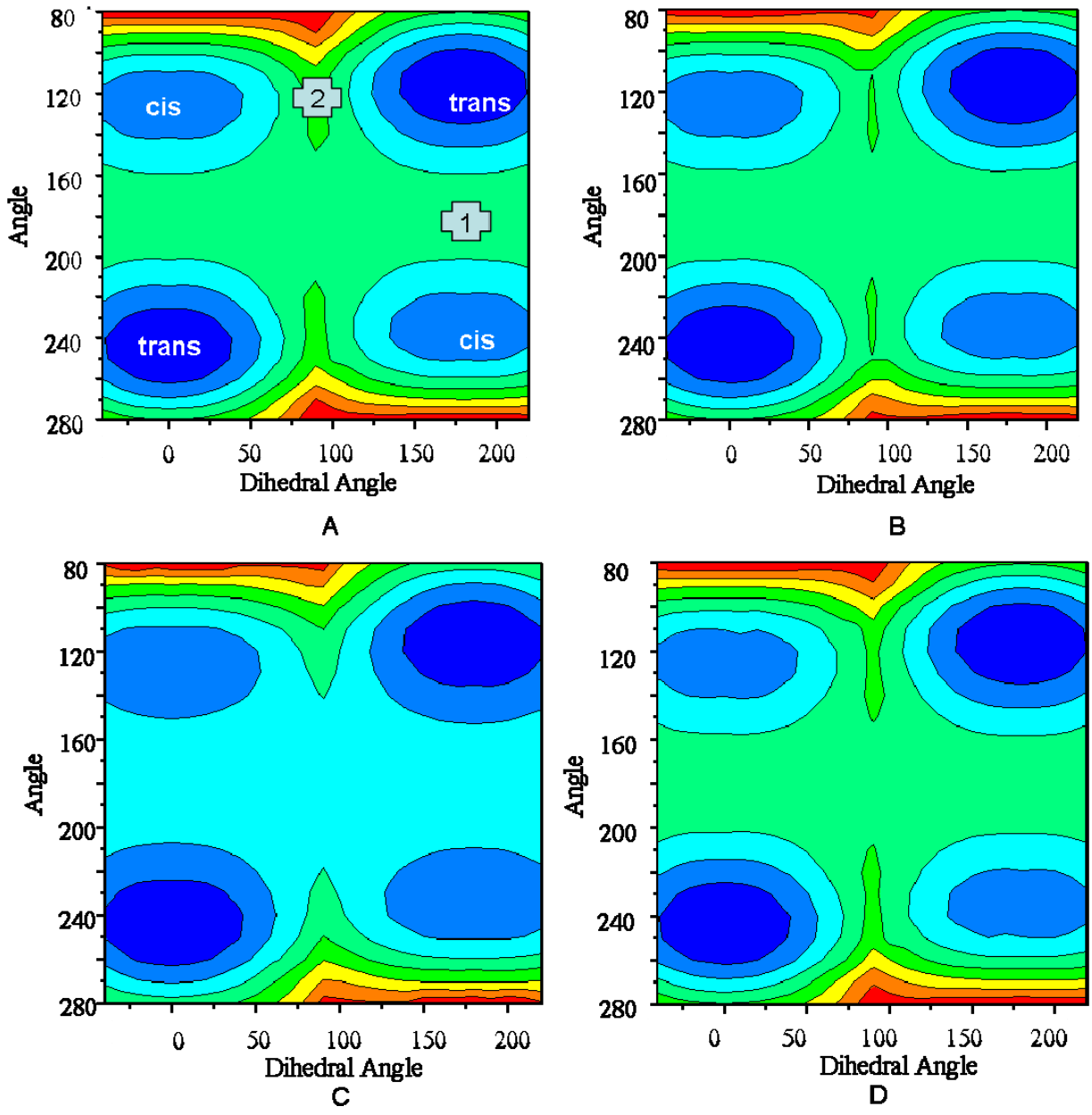


Figure 11-3. Contour maps of the ground state. A) Azo. B) Azon. C) Azonco. D) $Azo(NO_2)NH_2$. E)$AzoNO_2(NH_2)$. F) $AzoNO_2NO_2$. Angles in degrees, energy in kcal mol^{-1}. The energy range for each color id depicted in the legend.

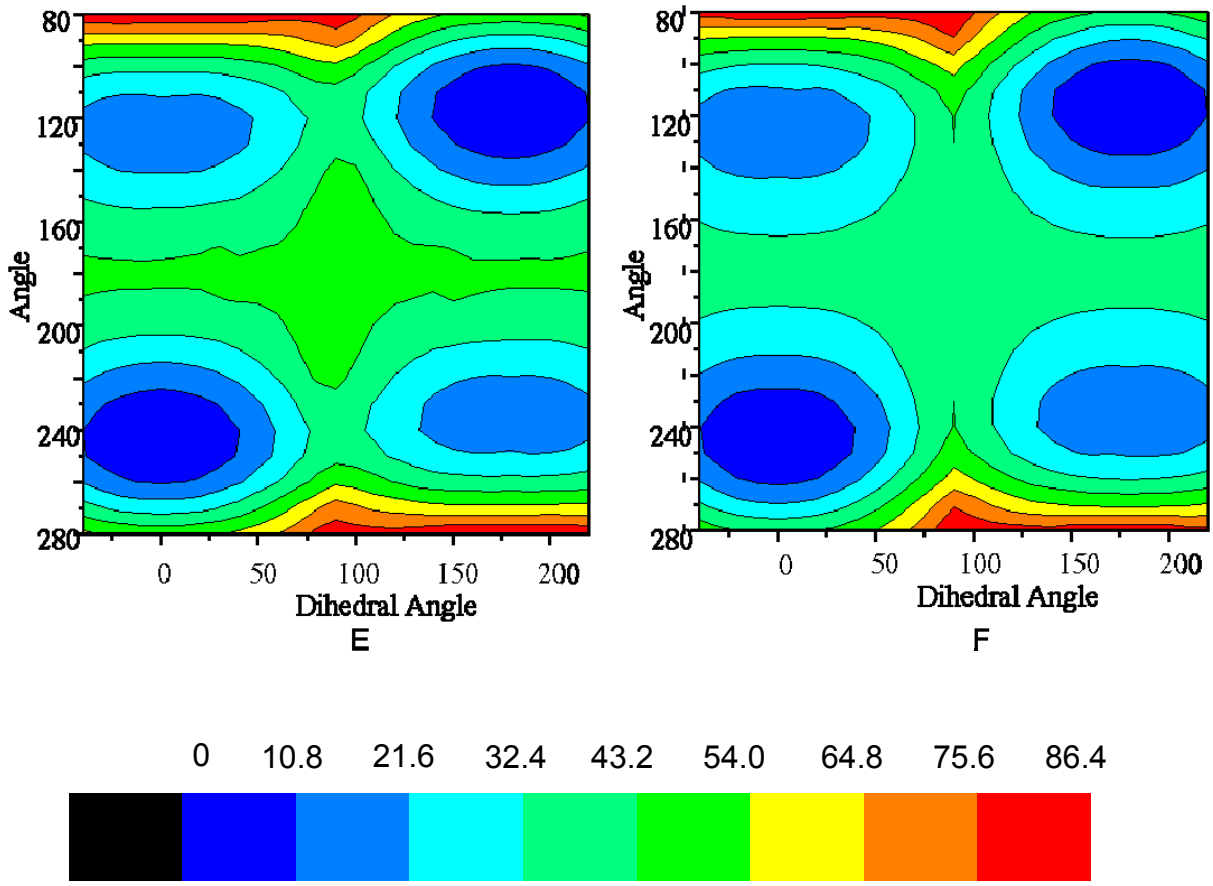


Figure 11-3. Continued

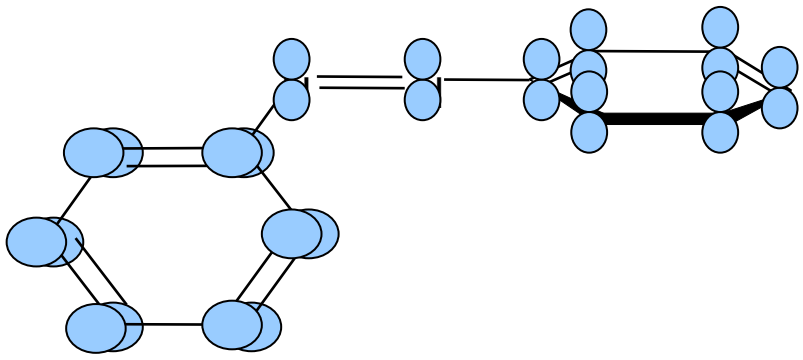


Figure 11-4. Schematic diagram of the molecular orbitals of the inversion transition state.

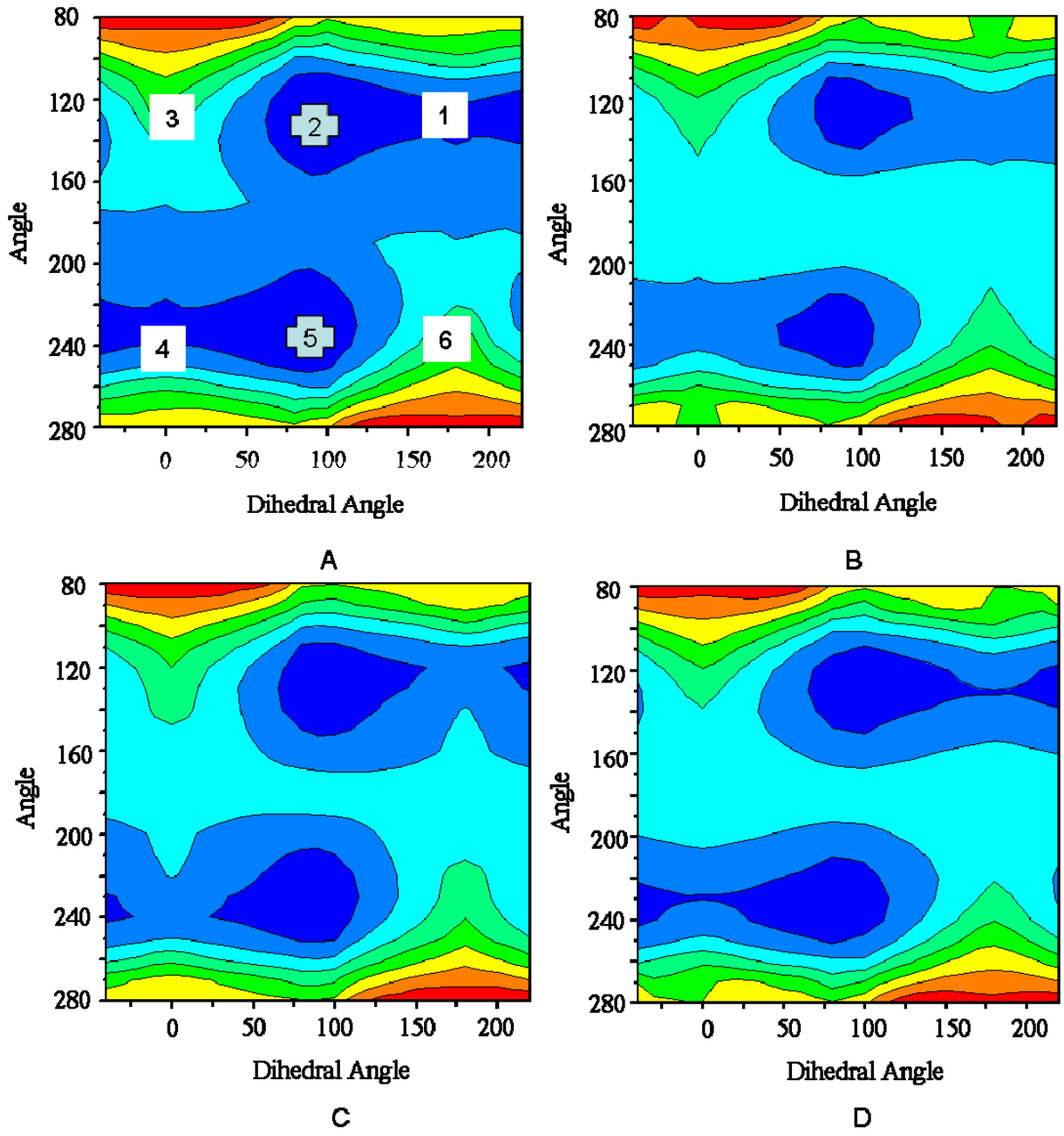


Figure 11-5. Contour maps of the first excited state. A) Azo. B) Azon. C) Azonco. D) $Azo(NO_2)NH_2$. E) $AzoNO_2(NH_2)$. F) $AzoNO_2(NO_2)$. Angles in degrees, energy in kcal mol^{-1}. The energy range for each color is depicted in the legend.

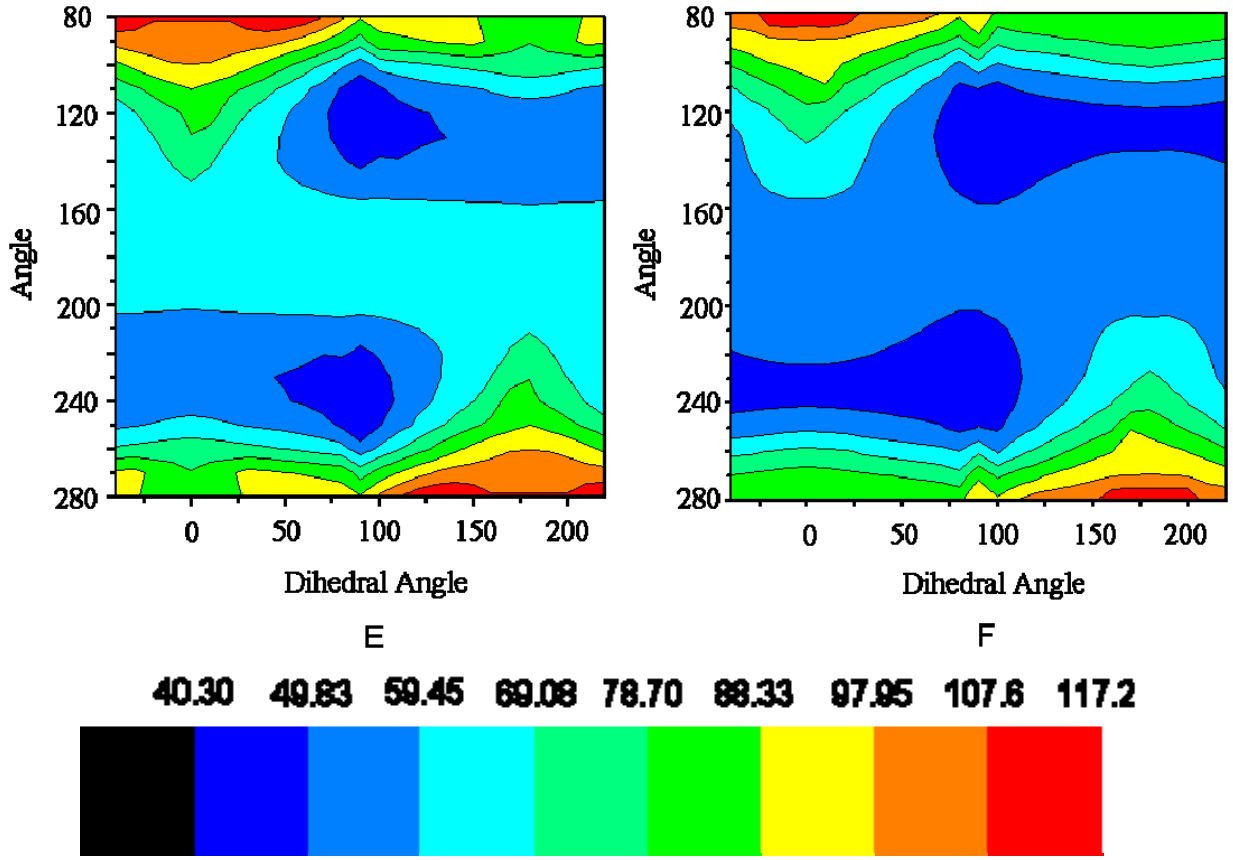


Figure 11-5. Continued

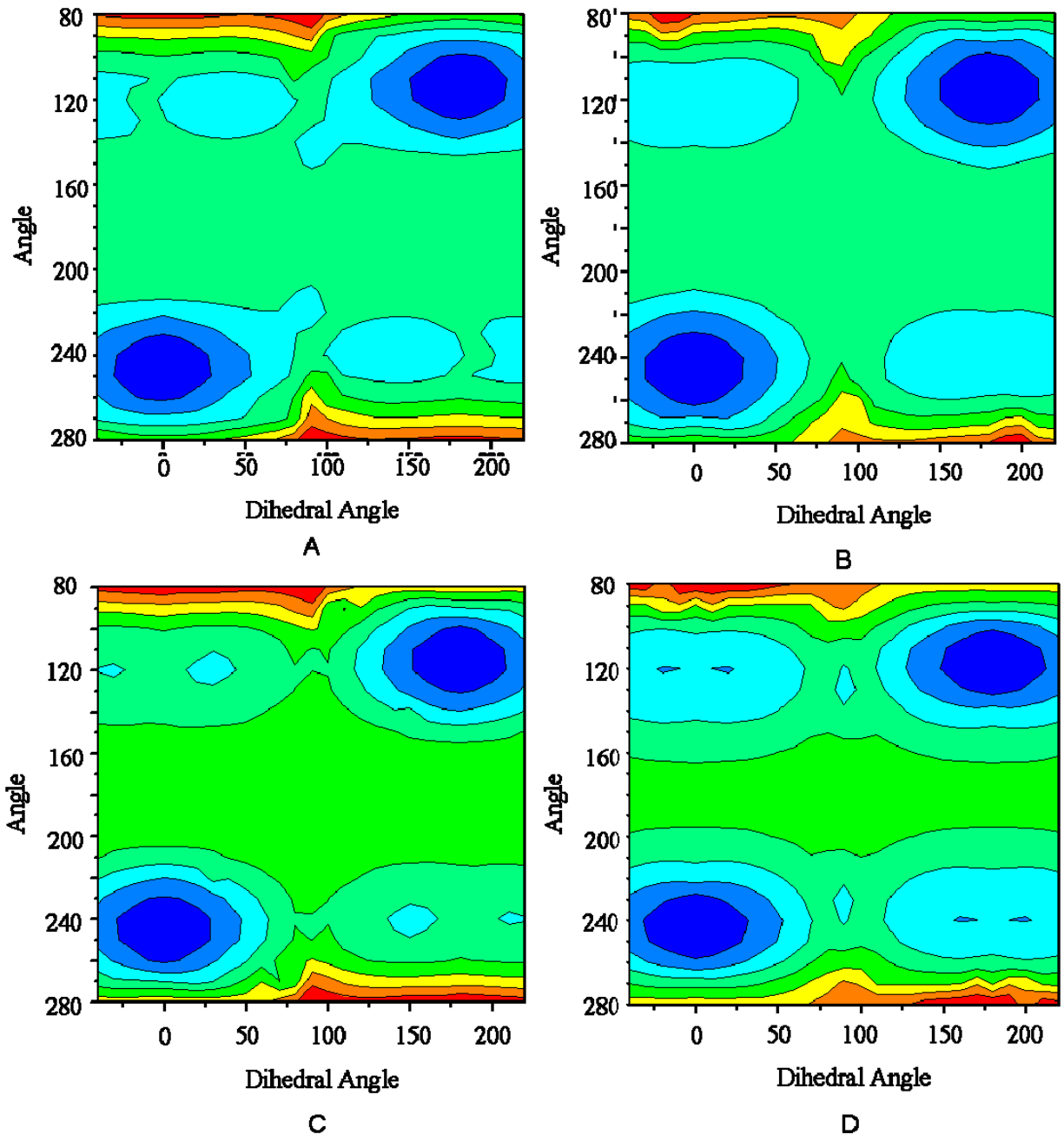


Figure 11-6. Contour maps of the second excited state. A) Azo. B) Azon. C) Azonco. D) Azo(NO_2)NH_2. E) AzoNO_2(NH_2). F) AzoNO_2NO_2. Angles in degrees, energy in kcal mol^{-1}. The energy range is depicted in the legend.

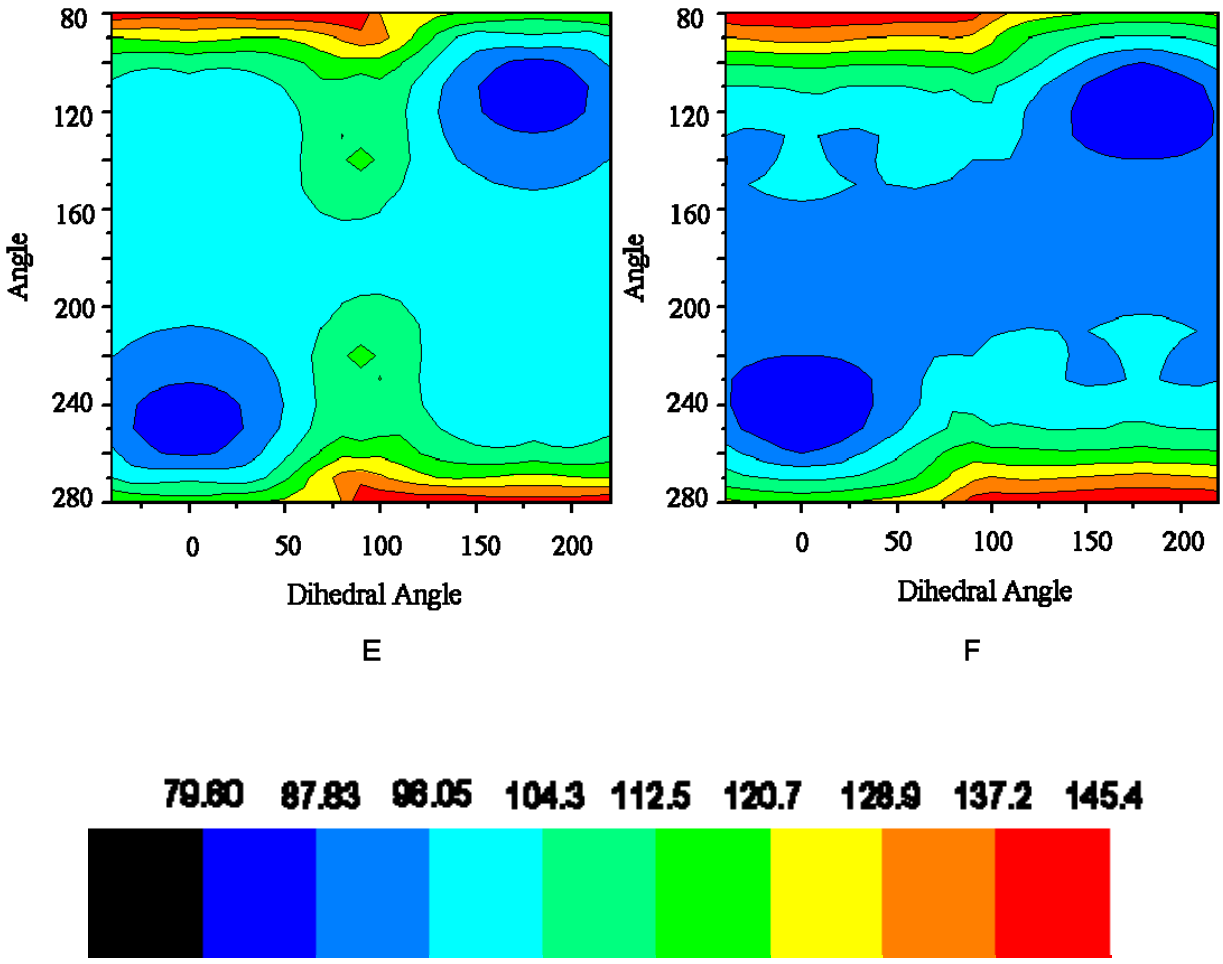


Figure 11-6. Continued

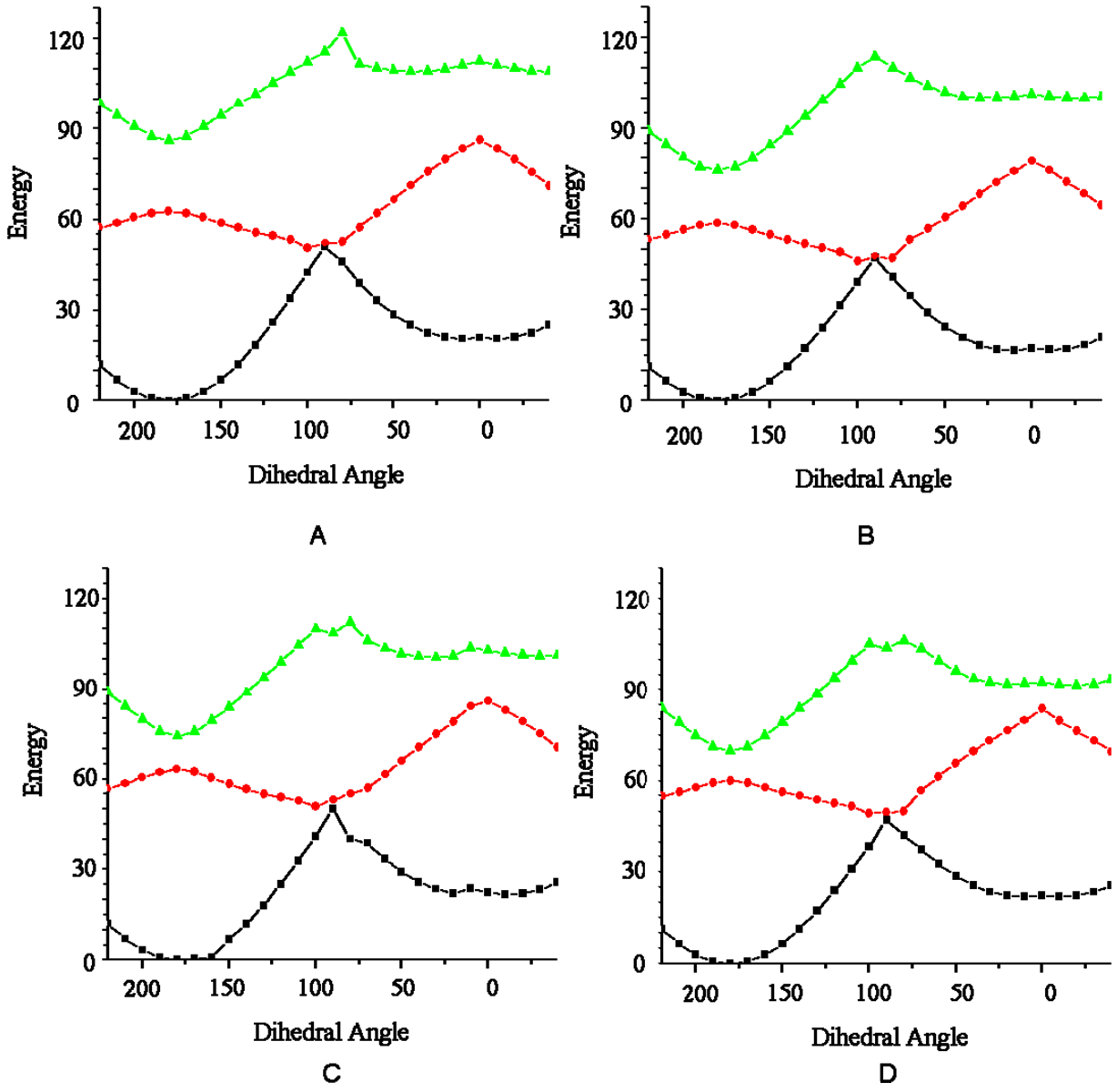


Figure 11-7. Rotation pathway along the angle of the ground state minimum. A) Azo. B) Azon. C) Azonco. D) Azo(NO_2)NH_2. E) AzoNO_2(NH_2). F) AzoNO_2NO_2. Angles in degrees, energy in kcal mol^{-1}.

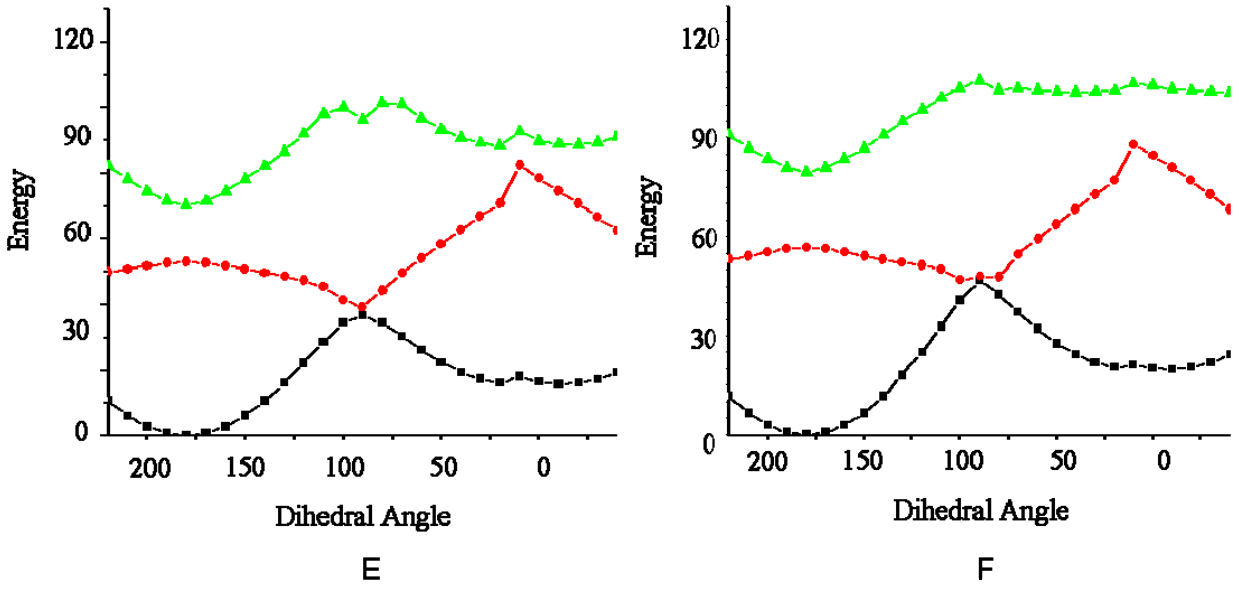


Figure 11-7. Continued

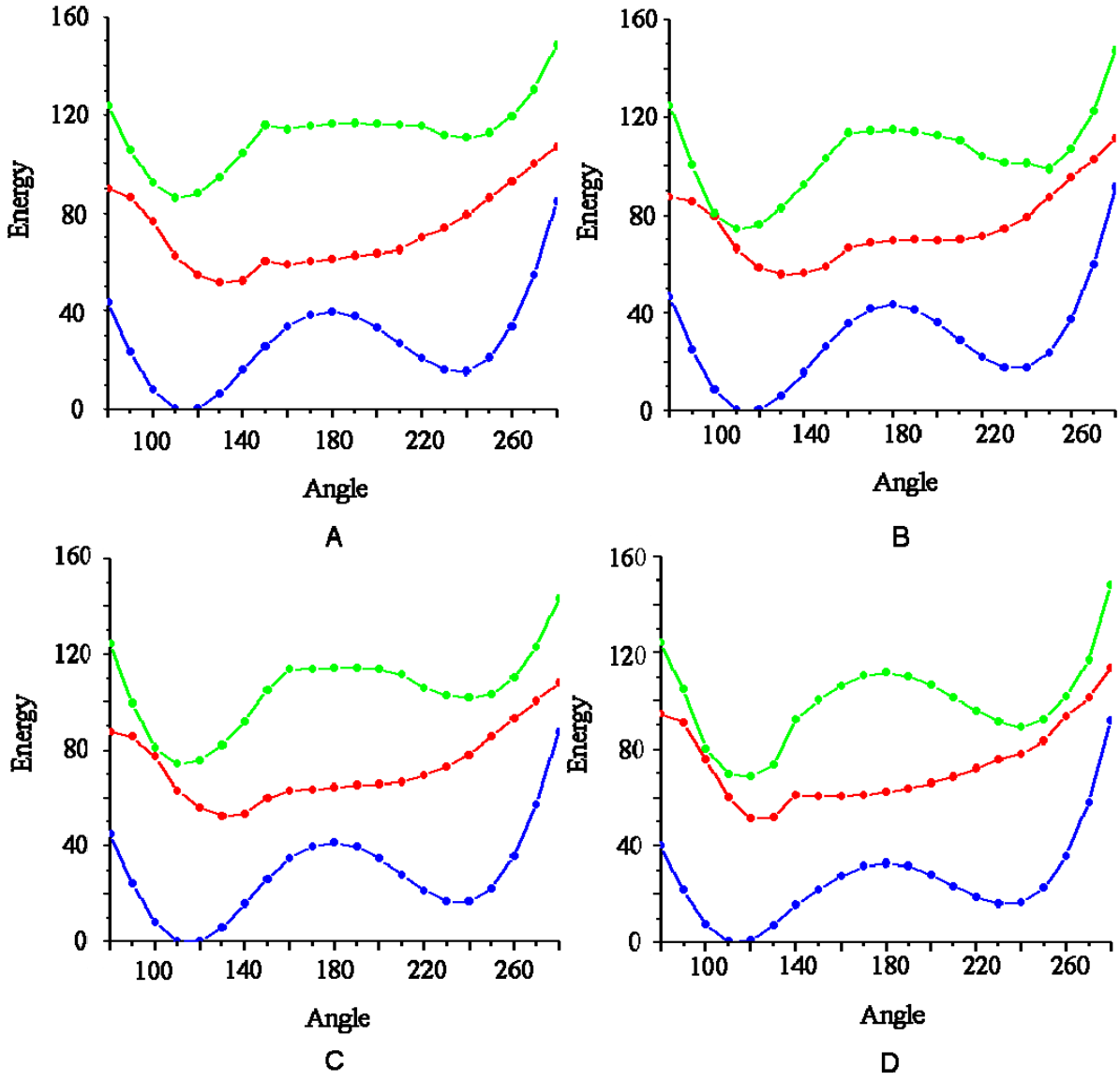


Figure 11-8. Inversion pathway along the dihedral of the ground state minimum. A) Azo. B) Azon. C) Azonco. D) Azo(NO_2)NH_2. E) AzoNO_2(NH_2). F) AzoNO_2NO_2. Angles in degrees, energy in kcal mol^{-1}.

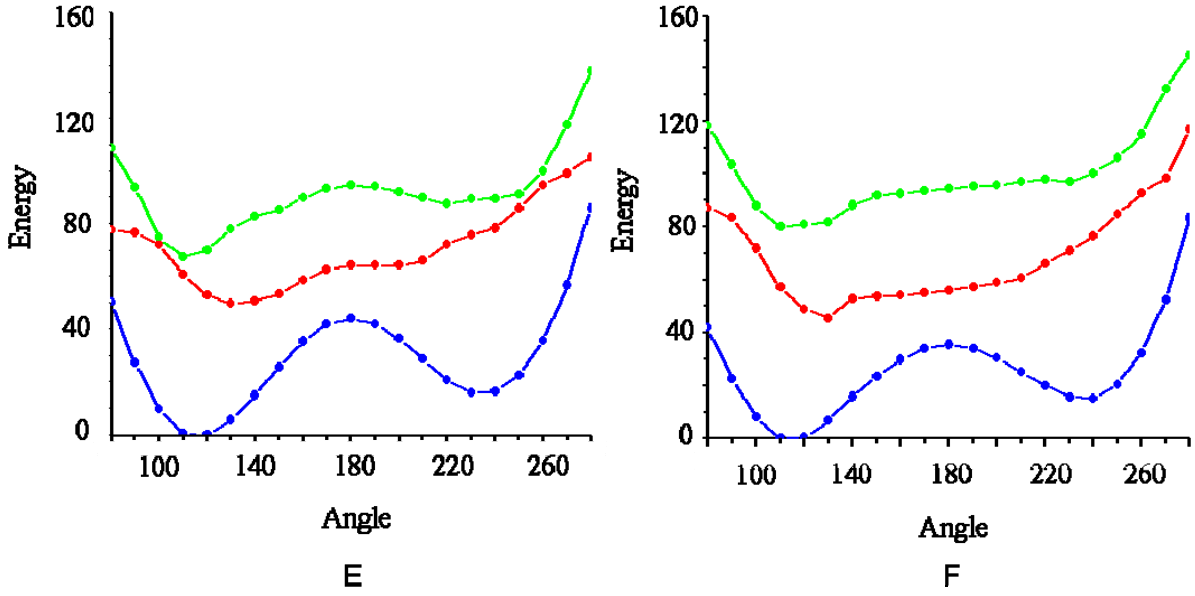


Figure 11-8. Continued

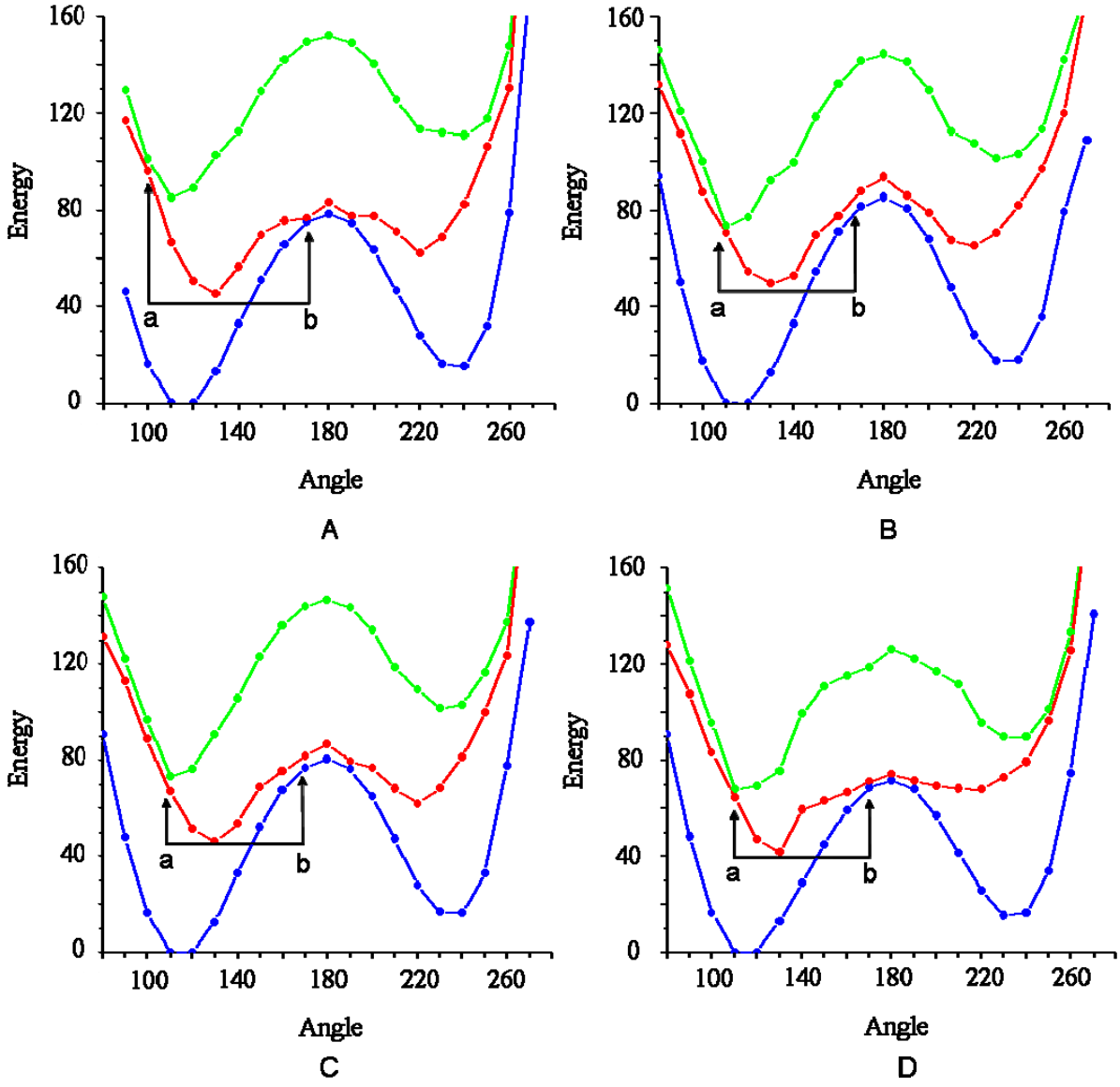


Figure 11-9. Concerted-inversion pathway along the dihedral of the ground state minimum. A) Azo. B) Azon. C) Azonco. D) $AzoNO_2NH_2$. E) $AzoNO_2NO_2$. Angles in degrees, energy in kcal mol^{-1}. Only one graph is necessary for $AzoNH_2NO_2$ because the NNC and CNN angles are being scanned synchronously. Arrow a represents the amount of available energy while arrow b represents the energy barrier. The concerted-inversion pathway is only open when the amount of available energy (arrow a) is greater than the energy barrier (arrow b).

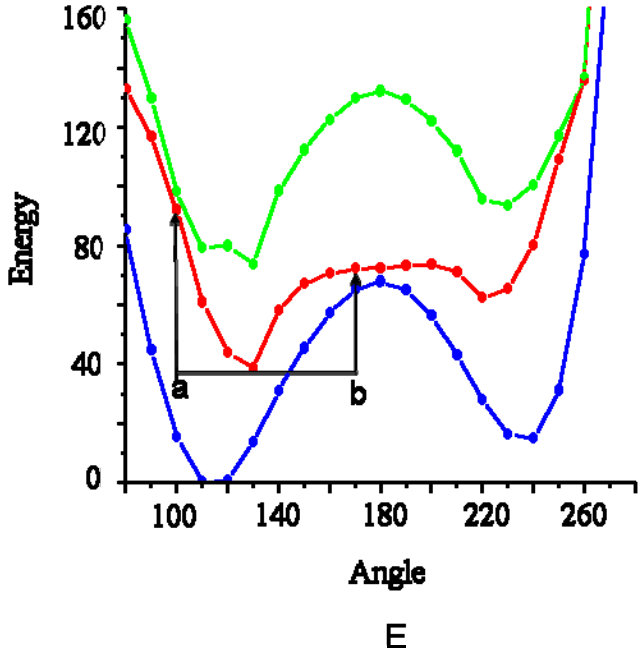


Figure 11-9. Continued

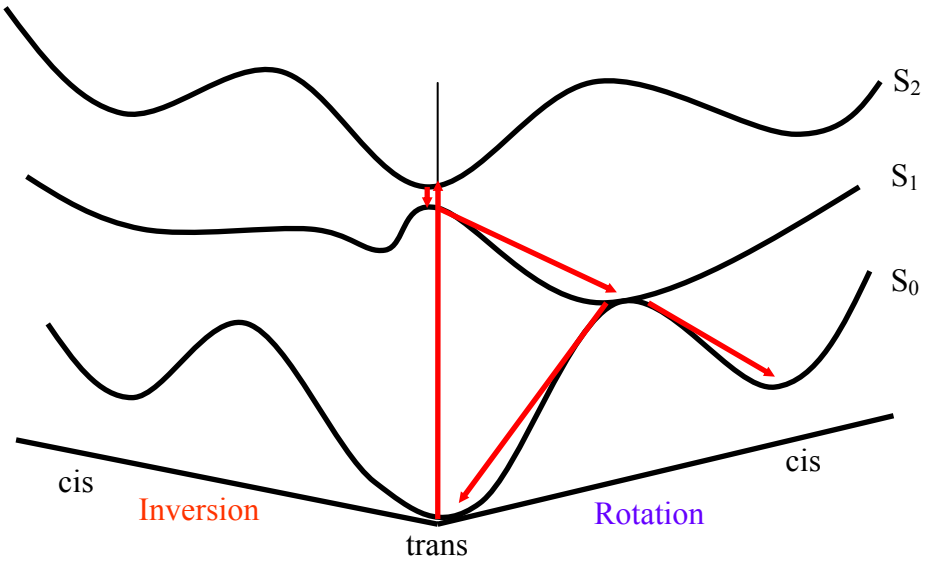


Figure 11-10. Scheme of the trans→cis isomerization process for Azon, Azonco, and $AzoNO_2NH_2$. After both n→π* and π→π* excitation, the rotation pathway dominates the isomerization process.

CHAPTER 12
AZOBENZENE CONCLUSIONS

We have found that adding electron donating substituents to the benzene rings of azobenzene raises the ground state inversion barrier height, making it harder to isomerize. Electron withdrawing groups were found to lower the same barrier. On the potential energy surface of the first excited state, there exists a slight trans → cis barrier along the inversion pathway, while all other pathways are without barriers. A conical intersection between the S_0 and S_1 states was found for each azobenzene along the rotation pathway making this pathway the most likely method of isomerization. The surface of the S_2 state was shown to be extremely close in energy to the S_1 state at specific points, indicating that excitation to the S_2 state leads to rapid relaxation to the S_1 state. Our results indicate this relaxation occurs using the concerted-inversion pathway for Azo and AzoNO_2NO_2. The concerted-inversion energy barriers were too high for the other azobenzenes to overcome. They most likely use the conical intersection found along the rotation pathway as their primary isomerization mechanism, regardless of excitation wavelength.

APPENDIX
LIST OF CONSTRAINTS

Table A-1. List of distances for targets T288 and T306

	T288			T306		
	Atom 1	Atom 2	Distance, Å	Atom 1	Atom 2	Distance, Å
1	5	67	20.1	2	40	12
2	5	71	19.2	2	45	14.7
3	5	76	22	7	40	9.1
4	10	67	21.5	7	45	13.5
5	10	71	16	23	40	16.9
6	10	76	13.1	23	45	5.5
7	19	67	13.3	30	40	9.3
8	19	76	14.5	30	45	17.9
9	23	71	11.6	40	53	14.9
10	23	76	20.2	40	59	13.2
11	31	71	10.1	45	53	12.6
12	31	76	18	45	59	12.7
13	36	67	16.9	40	76	11.5
14	36	71	14.5	40	81	17.7
15	36	76	19	45	76	9.1
16	53	67	12	45	78	5.6
17	53	71	13.7	5	26	8.1
18	57	67	12.5	5	57	9.9
19	57	76	11.6	5	78	10.8
20	67	76	14.4	26	57	11.7
21	67	80	19.9	26	78	11.1
22	67	86	16.5	26	43	6.5
23	71	80	14.4	5	43	10.8
24	71	86	17.4	43	57	7.8
25	76	86	22.7	43	78	6.6

Table A-2. List of distances for targets T309 and T335

	T309			T335		
	Atom 1	Atom 2	Distance, Å	Atom 1	Atom 2	Distance, Å
1	5	35	20.7	5	14	13.7
2	5	40	13.7	5	18	19
3	10	35	12.1	5	24	23.1
4	10	40	13.8	5	29	15.9
5	18	40	17.7	5	35	16.5
6	21	35	9.6	5	40	16.6
7	21	40	15.8	9	18	13.5
8	27	35	24.6	9	24	18
9	27	40	27.9	9	29	11.9
10	34	40	10.1	9	35	14.6
11	35	46	15.6	9	40	17.7
12	35	52	19.1	14	24	12.1
13	40	46	11	14	29	9.7
14	40	52	23	14	35	13.6
15	5	30	25.4	14	40	20
16	10	30	10.9	18	24	11.5
17	18	30	14.9	18	29	13.2
18	30	35	15.3	18	35	17.7
19	30	40	19.6	18	40	24.9
20	30	46	25.2	24	29	8.6
21	30	52	18.7	24	35	15.7
22	5	21	22.9	24	40	24.4
23	21	27	16.8	29	35	9.9
24	21	46	19.5	29	40	17.3
25	21	52	12.9	35	40	8.9

Table A-3. List of distances for target T340

	T340 Set1			T340 Set 2		
	Atom 1	Atom 2	Distance, Å	Atom 1	Atom 2	Distance, Å
1	6	66	22.9	6	22	18.7
2	6	73	19.6	6	31	14
3	11	66	21	6	35	15.9
4	11	69	16	6	52	9.9
5	19	66	15.3	6	55	11.9
6	19	69	11.6	6	68	17.9
7	19	73	12.6	6	81	8.1
8	23	69	11.1	22	31	5.1
9	23	73	17	22	35	11.8
10	30	66	10.4	22	52	10.1
11	30	73	14.8	22	55	14.9
12	35	66	19.2	22	68	10.3
13	35	73	16.7	22	81	14.4
14	52	66	13.8	35	52	13.8
15	52	69	11.2	35	55	17.1
16	52	73	14.4	35	68	17.2
17	55	66	13.3	35	81	13.6
18	55	73	9.8	52	55	11.2
19	66	73	10.6	52	68	9.3
20	66	80	17.4	52	81	6.8
21	66	84	20.5	55	68	7.9
22	69	80	12.9	55	81	6.5
23	69	84	17.8	68	81	10.9
24	73	80	10.7	31	81	11.3
25	73	84	20.2	31	68	10.7

Table A-4. List of distances for target T349

	T349 Set1			T349 Set 2		
	Atom 1	Atom 2	Distance, Å	Atom 1	Atom 2	Distance, Å
1	32	38	10.1	32	38	10.1
2	32	43	15.7	32	43	15.7
3	32	68	9.4	32	58	11.7
4	32	72	18.5	32	78	24.3
5	32	77	22.9	32	81	19.6
6	32	81	19.6	38	43	8.5
7	32	86	21.6	38	52	12.6
8	38	43	8.5	38	58	19.5
9	38	68	11.7	38	68	11.7
10	38	72	11.4	38	78	15.1
11	38	77	14.3	43	52	13.5
12	38	81	10.5	43	58	21
13	38	86	12.6	43	72	10.9
14	43	68	16.3	43	78	15.6
15	43	72	10.9	43	81	13.3
16	43	77	14	52	68	6.4
17	43	81	13.3	52	72	11
18	43	86	16	52	78	20
19	68	77	18.2	52	81	17.6
20	68	81	17	58	68	14.5
21	68	86	22.4	58	72	24.2
22	72	81	8.9	58	81	28.7
23	72	86	16.5	68	81	17
24	77	86	14.1	72	78	9.4
25	81	86	8.4	68	72	13.2

Table A-5. List of distances for targets T348 and T353

	T348			T353		
	Atom 1	Atom 2	Distance, Å	Atom 1	Atom 2	Distance, Å
1	26	58	14	3	29	13.7
2	26	61	13.5	3	35	14.7
3	28	58	10.3	3	42	22.3
4	28	61	9.6	11	29	18.2
5	35	58	10.9	11	35	16.1
6	35	61	12.1	11	42	16.4
7	36	58	14.6	17	29	20
8	36	61	15.3	17	35	17.8
9	4	58	34.8	17	42	17.5
10	4	61	33.1	29	35	9.8
11	8	58	24.3	29	42	19.8
12	8	61	22.2	35	42	10.3
13	10	58	19	29	65	24
14	10	61	17.3	29	73	20.1
15	4	26	22.1	35	65	21.3
16	4	28	24.5	35	72	14.4
17	4	35	24.9	42	65	25
18	8	26	15.2	42	72	19.3
19	8	28	14.2	29	77	14.9
20	8	35	16	42	77	13.1
21	10	26	12.7	29	79	17.1
22	10	28	9.6	35	79	12.6
23	10	35	11.8	42	79	14.5
24	26	35	6	8	29	10.5
25	4	10	18.8	8	42	16

Table A-6. List of distances for targets T358 and T363

	T358			T363		
	Atom 1	Atom 2	Distance, Å	Atom 1	Atom 2	Distance, Å
1	14	23	14.3	4	36	11.3
2	14	27	13.2	14	45	12.7
3	14	30	11.4	20	45	15.5
4	14	38	9.9	25	38	19.6
5	14	42	18.4	25	45	15.8
6	14	48	20.3	31	36	10.6
7	14	51	15.2	31	45	10.1
8	14	61	16.9	36	64	15.6
9	14	67	18.4	36	67	18.8
10	18	27	9.3	36	80	23.9
11	18	30	11.7	36	83	27.9
12	18	38	11.7	45	62	16.1
13	18	42	15	45	66	21.1
14	18	48	18.2	45	80	27.1
15	18	51	15	36	45	12.1
16	18	61	17.3	27	38	15.6
17	18	67	15.1	27	45	11.1
18	23	27	8.7	39	45	9.8
19	23	30	16.7	39	65	16.3
	23	38	16.6	39	80	23.5
21	23	42	13.2	39	83	28.6
22	23	48	17.7	14	41	11.1
23	23	51	16.8	31	41	9.6
24	23	61	17.5	14	67	13.5
25	23	67	10.9	20	40	19.7

Table A-7. List of distances for targets T359 and T311

	T359			T311		
	Atom 1	Atom 2	Distance, Å	Atom 1	Atom 2	Distance, Å
1	5	71	19.1	10	64	14.1
2	5	79	21.9	10	70	9.3
3	11	71	23.6	10	82	22.8
4	11	79	14.4	19	64	19.8
5	19	71	16.1	19	70	18.9
6	19	79	11.2	19	77	27.3
7	22	79	11.8	23	64	21.7
8	35	79	12.6	23	70	21.8
9	40	71	19.4	23	77	29.7
10	40	79	15.8	30	64	16.3
11	57	71	11.2	30	70	21.2
12	57	79	15.8	35	64	20.3
13	61	71	13.5	35	70	22.4
14	71	79	12.4	35	77	28.6
15	71	84	22.0	42	64	16.4
16	71	87	15.9	42	70	12.9
17	71	91	17.2	42	82	24.9
18	79	84	12.7	50	64	7.4
19	79	87	12.9	50	70	12.2
20	79	91	21.5	50	82	24.5
21	5	75	19.2	59	64	10.9
22	11	75	18.3	59	70	16.9
23	40	75	16.0	59	77	27.3
24	57	75	11.8	64	82	27.4
25	75	84	16.7	64	70	10.3

LIST OF REFERENCES

1. Todd AE, Orengo CA, Thornton JM. Evolution of Function in Protein Superfamilies, from a Structural Perspective. Journal of Molecular Biology 2001;307(4):1113-1143.

2. Orengo CA, Todd AE, Thornton JM. From protein structure to function. Current Opinion in Structural Biology 1999;9(3):374-382.

3. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature (London, United Kingdom) 1958;181:662-666.

4. Perutz MF, Muirhead H, Cox JM, Goaman LCG, Mathews FS, McGandy EL, Webb LE. Three-dimensional Fourier synthesis of horse oxyhemoglobin at 2.8 A resolution. I. X-ray analysis. Nature (London, United Kingdom) 1968;219(5149):29-32.

5. Geerlof A, Brown J, Coutard B, Egloff MP, Enguita FJ, Fogg MJ, Gilbert RJC, Groves MR, Haouz A, Nettleship JE, Nordlund P, Owens RJ, Ruff M, Sainsbury S, Svergun DI, Wilmanns M. The impact of protein characterization in structural proteomics. Acta Crystallographica, Section D: Biological Crystallography 2006;D62(10):1125-1136.

6. Powell HR. The Rossmann Fourier autoindexing algorithm in MOSFLM. Acta Crystallographica, Section D: Biological Crystallography 1999;D55(10):1690-1695.

7. Hauptman H. Phasing methods for protein crystallography. Current Opinion in Structural Biology 1997;7(5):672-680.

8. Uson I, Sheldrick GM. Advances in direct methods for protein crystallography. Current Opinion in Structural Biology 1999;9(5):643-648.

9. Taylor G. The phase problem. Acta Crystallographica, Section D: Biological Crystallography 2003;D59(11):1881-1890.

10. Ealick SE. Advances in multiple wavelength anomalous diffraction crystallography. Current Opinion in Chemical Biology 2000;4(5):495-499.

11. Wider G. Structure determination of biological macromolecules in solution using nuclear magnetic resonance spectroscopy. BioTechniques 2000;29(6):1278-1280, 1282, 1284-1290, 1292, 1294.

12. Hore PJ. Nuclear Magnetic Resonance. Compton RG, editor. New York: Oxford University Press; 2001. 90 p.

13. Guntert P. Structure calculation of biological macromolecules from NMR data. Quarterly Reviews of Biophysics 1998;31(2):145-237.

14. Nilges M, Clore GM, Gronenborn AM. Determination of three-dimensional structures of proteins from interproton distance data by hybrid distance geometry-dynamical stimulated annealing calculations. FEBS Letters 1988;229(2):317-324.

15. Havel TF. An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. Progress in Biophysics & Molecular Biology 1991;56(1):43-78.

16. Wagner G, Braun W, Havel TF, Schaumann T, Go N, Wuethrich K. Protein structures in solution by nuclear magnetic resonance and distance geometry. The polypeptide fold of the basic pancreatic trypsin inhibitor determined using two different algorithms, DISGEO and DISMAN. Journal of Molecular Biology 1987;196(3):611-639.

17. Guntert P, Mumenthaler C, Wuthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. Journal of Molecular Biology 1997;273(1):283-298.

18. Guntert P, Qian YQ, Otting G, Muller M, Gehring W, Wuthrich K. Structure determination of the Antp (C39----S) homeodomain from nuclear magnetic resonance data in solution using a novel strategy for the structure calculation with the programs DIANA, CALIBA, HABAS and GLOMSA. J Mol Biol FIELD Full Journal Title:Journal of molecular biology 1991;217(3):531-540.

19. Braun W. Distance geometry and related methods for protein structure determination from NMR data. Quarterly Reviews of Biophysics 1987;19(3-4):115-157.

20. Demeter A, Fodor T, Fischer J. Stereochemical investigations on the diketopiperazine derivatives of enalapril and lisinopril by NMR spectroscopy. Journal of Molecular Structure 1998;471(1-3):161-174.

21. Hubbell WL, Altenbach C. Investigation of structure and dynamics in membrane proteins using site-directed spin labeling. Current Opinion in Structural Biology 1994;4(4):566-573.

22. Jeschke G. Determination of the nanostructure of polymer materials by electron paramagnetic resonance spectroscopy. Macromolecular Rapid Communications 2002;23(4):227-246.

23. Schweiger A, Jeschke G. Principles of Pulse Electron Paramagnetic Resonance Spectroscopy; 2001. 572 pp p.

24. Berliner LJ, Eaton GR, Eaton SS, Editors. Distance Measurements in Biological Systems by EPR. [In: Biol. Magn. Reson., 2000; 19]; 2000. 614 pp p.

25. Rabenstein MD, Shin Y-K. Determination of the distance between two spin labels attached to a macromolecule. Proceedings of the National Academy of Sciences of the United States of America 1995;92(18):8239-8243.

26. Stryer L. Fluorescence energy transfer as a spectroscopic ruler. Annual Review of Biochemistry 1978;47:819-846.

27. Dodson MS. Dimethyl suberimidate cross-linking of oligo(dT) to DNA-binding proteins. Bioconjug Chem FIELD Full Journal Title:Bioconjugate chemistry 2000;11(6):876-879.

28. MacPhee CE, Howlett GJ, Sawyer WH. Mass Spectrometry to Characterize the Binding of a Peptide to a Lipid Surface. Analytical Biochemistry 1999;275(1):22-29.

29. Young MM, Tang N, Hempel JC, Oshiro CM, Taylor EW, Kuntz ID, Gibson BW, Dollinger G. High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. Proceedings of the National Academy of Sciences of the United States of America 2000;97(11):5802-5806.

30. Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BFF, Rapp BA, Wheeler DL. GenBank. Nucleic Acids Research 1999;27(1):12-17.

31. Stoesser G, Tuli MA, Lopez R, Sterk P. The EMBL Nucleotide sequence database. Nucleic Acids Research 1999;27(1):18-24.

32. Zhu H, Bilgin M, Snyder M. Proteomics. Annual Review of Biochemistry 2003;72:783-812.

33. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 1997;25(17):3389-3402.

34. Sanchez R, Sali A. Evaluation of comparative protein structure modeling by MODELLER-3. Proteins 1997;Suppl 1:50-58.

35. Epstein CJ, Goldberger RF, Anfinsen CB. The genetic control of tertiary protein structure. Model systems. Cold Spring Harbor Symposia on Quantitative Biology 1963;28:439-449.

36. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO Journal 1986;5(4):823-826.

37. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins: Structure, Function, and Genetics 1991;9(1):56-68.

38. Rost B. Twilight zone of protein sequence alignments. Protein Engineering 1999;12(2):85-94.

39. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology 1990;215(3):403-410.

40. Pearson WR. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods in Enzymology 1990;183(Mol. Evol.: Comput. Analy. Protein Nucleic Acid Sequences):63-98.

41. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 1994;22(22):4673-4680.

42. Peitsch MC, Schwede T, Guex N. Automated protein modeling - the proteome in 3D. Pharmacogenomics 2000;1(3):257-266.

43. Bates PA, Sternberg MJE. Model building by comparison at CASP3: using expert knowledge and computer automation. Proteins: Structure, Function, and Genetics 1999(Suppl. 3):47-54.

44. Dayringer HE, Tramontano A, Sprang SR, Fletterick RJ. Interactive program for visualization and modeling of proteins, nucleic acids and small molecules. Journal of Molecular Graphics 1986;4(2):82-87.

45. Sali A, Blundell TL. Comparative protein modeling by satisfaction of spatial restraints. Journal of Molecular Biology 1993;234(3):779-815.

46. Vriend G. WHAT IF: a molecular modeling and drug design program. Journal of Molecular Graphics 1990;8(1):52-56, 29.

47. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins: Structure, Function, and Genetics 1999(Suppl. 3):171-176.

48. Fiser A, Do RKG, Sali A. Modeling of loops in protein structures. Protein Science 2000;9(9):1753-1773.

49. De Filippis V, Sander C, Vriend G. Predicting local structural changes that result from point mutations. Protein Engineering 1994;7(10):1203-1208.

50. Stites WE, Meeker AK, Shortle D. Evidence for strained interactions between side-chains and the polypeptide backbone. Journal of Molecular Biology 1994;235(1):27-32.

51. Dunbrack RL, Jr., Karplus M. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nature Structural Biology 1994;1(5):334-340.

52. Novotny J, Rashin AA, Bruccoleri RE. Criteria that discriminate between native proteins and incorrectly folded models. Proteins: Structure, Function, and Genetics 1988;4(1):19-30.

53. Brenner SE, Chothia C, Hubbard TJP, Murzin AG. Understanding protein structure: using scope for fold interpretation. Methods in Enzymology 1996;266(Computer Methods for Macromolecular Sequence Analysis):635-643.

54. Holm L, Sander C. Mapping the protein universe. Science (Washington, D C) 1996;273(5275):595-602.

55. Hubbard TJP, Murzin AG, Brenner SE, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Research 1997;25(1):236-239.

56. Valencia A, Kjeldgaard M, Pai EF, Sander C. GTPase domains of ras p21 oncogene protein and elongation factor Tu: analysis of three-dimensional structures, sequence families, and functional sites. Proceedings of the National Academy of Sciences of the United States of America 1991;88(12):5443-5447.

57. Bourne PE, Weissig H, Editors. Structural Bioinformatics; 2003. 649 pp

58. Godzik A, Skolnick J. Sequence-structure matching in globular proteins: Application to supersecondary and tertiary structure determination. Proceedings of the National Academy of Sciences of the United States of America 1992;89(24):12098-12102.

59. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through the folding motif. Proteins: Structure, Function, and Genetics 1993;16(1):92-112.

60. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature (London, United Kingdom) 1992;358(6381):86-89.

61. Anfinsen CB, Haber E, Sela M, White FH, Jr. Kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proceedings of the National Academy of Sciences of the United States of America 1961;47:1309-1314.

62. Anfinsen CB. Principles that govern the folding of protein chains. Science (Washington, DC, United States) 1973;181(4096):223-230.

63. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. Journal of Molecular Biology 1997;268(1):209-225.

64. Samudrala R, Xia Y, Huang E, Levitt M. Ab initio protein structure prediction using a combined hierarchical approach. Proteins: Structure, Function, and Genetics 1999(Suppl. 3):194-198.

65. Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. Ab initio folding of proteins using restraints derived from evolutionary information. Proteins: Structure, Function, and Genetics 1999(Suppl. 3):177-185.

66. Pillardy J, Czaplewski C, Liwo A, Lee J, Ripoll DR, Kazmierkiewicz R, Oldziej S, Wedemeyer WJ, Gibson KD, Arnautova YA, Saunders J, Ye Y-J, Scheraga HA. Recent improvements in prediction of protein structure by global optimization of a potential energy function. Proceedings of the National Academy of Sciences of the United States of America 2001;98(5):2329-2333.

67. Marqusee S, Robbins VH, Baldwin RL. Unusually stable helix formation in short alanine-based peptides. Proceedings of the National Academy of Sciences of the United States of America 1989;86(14):5286-5290.

68. Blanco FJ, Rivas G, Serrano L. A short linear peptide that folds into a native stable b-hairpin in aqueous solution. Nature Structural Biology 1994;1(9):584-590.

69. Callihan DE, Logan TM. Conformations of Peptide Fragments from the FK506 Binding Protein: Comparison with the Native and Urea-unfolded States. Journal of Molecular Biology 1999;285(5):2161-2175.

70. Park BH, Levitt M. The complexity and accuracy of discrete state models of protein structure. Journal of Molecular Biology 1995;249(2):493-507.

71. Sippl MJ, Hendlich M, Lackner P. Assembly of polypeptide and protein backbone conformations from low energy ensembles of short fragments: Development of strategies and construction of models for myoglobin, lysozyme, and thymosin b4. Protein Science 1992;1(5):625-640.

72. Bowie JU, Eisenberg D. An evolutionary approach to folding small a-helical proteins that uses sequence information and an empirical guiding fitness function. Proceedings of the National Academy of Sciences of the United States of America 1994;91(10):4436-4440.

73. Jones DT. Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. Proteins:Structure, Function, and Genetics 1997;Suppl 1:185-191.

74. Sippl MJ. Knowledge-based potentials for proteins. Current Opinion in Structural Biology 1995;5(2):229-235.

75. Koppensteiner WA, Sippl MJ. Knowledge-based potentials-back to the roots. Biochemistry (Moscow)(Translation of Biokhimiya (Moscow)) 1998;63(3):247-252.

76. Simmerling C, Strockbine B, Roitberg AE. All-Atom Structure Prediction and Folding Simulations of a Stable Protein. Journal of the American Chemical Society 2002;124(38):11258-11259.

77. Qiu L, Pabit SA, Roitberg AE, Hagen SJ. Smaller and faster: the 20-residue Trp-cage protein folds in 4 ms. Journal of the American Chemical Society 2002;124(44):12952-12953.

78. Hansmann UHE, Okamoto Y. Numerical comparisons of three recently proposed algorithms in the protein folding problem. Journal of Computational Chemistry 1997;18(7):920-933.

79. Pedersen JT, Moult J. Protein folding simulations with genetic algorithms and a detailed molecular description. Journal of Molecular Biology 1997;269(2):240-259.

80. Park B, Levitt M. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. Journal of Molecular Biology 1996;258(2):367-392.

81. Huang ES, Subbiah S, Tsai J, Levitt M. Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. Journal of Molecular Biology 1996;257(3):716-725.

82. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. Journal of Molecular Biology 1998;275(5):895-916.

83. Bonneau R, Strauss CEM, Rohl CA, Chivian D, Bradley P, Malmstrom L, Robertson T, Baker D. De novo prediction of three-dimensional structures for major protein families. Journal of Molecular Biology 2002;322(1):65-78.

84. Fetrow JS, Skolnick J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. Journal of Molecular Biology 1998;281(5):949-968.

85. Zhang Y, DeVries ME, Skolnick J. Structure modeling of all identified G protein-coupled receptors in the human genome. [Erratum to document cited in CA144:305279]. PLoS Computational Biology 2006;2(3):200.

86. Bradley P, Chivian D, Meiler J, Misura KMS, Rohl CA, Schief WR, Wedemeyer WJ, Schueler-furman O, Murphy P, Schonbrun J, Strauss CEM, Baker D. Rosetta predictions in CASP5: Successes, failures, and prospects for complete automation. Proteins: Structure, Function, and Genetics 2003;53(Suppl. 6):457-468.

87. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CE, Baker D. Rosetta in CASP4: progress in ab initio protein structure prediction. Proteins: Structure, FUnction, and Genetics 2001;(Suppl 5):119-126.

88. Bowers PM, Strauss CEM, Baker D. De novo protein structure determination using sparse NMR data. Journal of Biomolecular NMR 2000;18(4):311-318.

89. Rohl CA, Baker D. De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. Journal of the American Chemical Society 2002;124(11):2723-2729.

90. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. Science (Washington, DC, United States) 2003;302(5649):1364-1368.

91. Kuhlman B, O'Neill JW, Kim DE, Zhang KYJ, Baker D. Accurate Computer-based Design of a New Backbone Conformation in the Second Turn of Protein L. Journal of Molecular Biology 2002;315(3):471-477.

92. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. Journal of Molecular Biology 2003;331(1):281-299.

93. Rohl CA, Strauss CEM, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with Rosetta. Proteins: Structure, Function, and Bioinformatics 2004;55(3):656-677.

94. Samudrala R, Levitt M. Decoys \"R\" Us: a database of incorrect conformations to improve protein structure prediction. Protein Science 2000;9(7):1399-1401.

95. Wang Y, Zhang H, Li W, Scott RA. Discriminating compact nonnative structures from the native structure of globular proteins. Proceedings of the National Academy of Sciences of the United States of America 1995;92(3):709-713.

96. Subramaniam S, Tcheng DK, Fenton JM. A knowledge-based method for protein structure refinement and prediction. Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology 1996;4:218-229.

97. Holm L, Sander C. Evaluation of protein models by atomic solvation preference. Journal of Molecular Biology 1992;225(1):93-105.

98. Crippen GM. A novel approach to calculation of conformation: distance geometry. Journal of Computational Physics 1977;24(1):96-107.

99. Havel TF, Kuntz ID, Crippen GM. The combinatorial distance geometry method for the calculation of molecular conformation. I. A new approach to an old problem. J Theor Biol FIELD Full Journal Title:Journal of theoretical biology 1983;104(3):359-381.

100. Kuntz ID, Crippen GM, Kollman PA. Application of distance geometry to protein tertiary structure calculations. Biopolymers 1979;18(4):939-957.

101. Collins CJ, Schilling B, Young M, Dollinger G, Guy RK. Isotopically labeled crosslinking reagents: resolution of mass degeneracy in the identification of crosslinked peptides. Bioorganic & Medicinal Chemistry Letters 2003;13(22):4023-4026.

102. Schilling B, Row RH, Gibson BW, Guo X, Young MM. MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides. Journal of the American Society for Mass Spectrometry 2003;14(8):834-850.

103. Kruppa GH, Schoeniger J, Young MM. A top down approach to protein structural studies using chemical cross-linking and fourier transform mass spectrometry. Rapid Communications in Mass Spectrometry 2003;17(2):155-162.

104. Alexandrov NN, Nussinov R, Zimmer RM. Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing 1996:53-72.

105. Bonneau R, Tsai J, Ruczinski I, Baker D. Functional Inferences from Blind ab Initio Protein Structure Predictions. Journal of Structural Biology 2001;134(2 & 3):186-190.

106. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J. On the origin and highly likely completeness of single-domain protein structures. Proceedings of the National Academy of Sciences of the United States of America 2006;103(8):2605-2610.

107. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Research 2000;28(1):235-242.

108. Levitt M. Growth of novel protein structural data. Proceedings of the National Academy of Sciences of the United States of America 2007;104(9):3183-3188.

109. Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A. Critical Assessment of methods of protein Structure Prediction (CASP)-round 6. Proteins: Structure, Function, and Bioinformatics 2005;61(Suppl. 7):3-7.

110. Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using Rosetta. Methods in Enzymology 2004;383(Numerical Computer Methods, Part D):66-93.

111. Rohl CA. Protein structure estimation from minimal restraints using Rosetta. Methods in Enzymology 2005;394(Nuclear Magnetic Resonance of Biological Macromolecules, Part C):244-260.

112. Chivian D, Kim DE, Malmstrom L, Schonbrun J, Rohl CA, Baker D. Prediction of CASP6 structures using automated Robetta protocols. Proteins: Structure, Function, and Bioinformatics 2005;61(Suppl. 7):157-166.

113. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Research 2004;32(Web Server):W526-W531.

114. Chivian D, Kim DE, Malmstroem L, Bradley P, Robertson T, Murphy P, Strauss CEM, Bonneau R, Rohl CA, Baker D. Automated prediction of CASP-5 structures using the Robetta server. Proteins: Structure, Function, and Genetics 2003;53(Suppl. 6):524-533.

115. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. JPred: a consensus secondary structure prediction server. Bioinformatics 1998;14(10):892-893.

116. King RD, Sternberg MJE. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Protein Science 1996;5(11):2298-2310.

117. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. Journal of Molecular Biology 1993;232(2):584-599.

118. Salamov AA, Solovyev VV. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. Journal of Molecular Biology 1995;247(1):11-15.

119. Frishman D, Argos P. Seventy-five percent accuracy in protein secondary structure prediction. Proteins: Structure, Function, and Genetics 1997;27(3):329-335.

120. Li W, Zhang Y, Skolnick J. Application of sparse NMR restraints to large-scale protein structure prediction. Biophysical Journal 2004;87(2):1241-1248.

121. Chen Y, Ding, F., Dokholyan, N. V. Fidelity of the protein structure reconstruction from inter-residue proximity constraints. J Phys Chem B 2007;111:7432-7438.

122. Faulon J-L, Sale K, Young M. Exploring the conformational space of membrane protein folds matching distance constraints. Protein Science 2003;12(8):1750-1761.

123. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Research 2003;31(13):3370 - 3374.

124. Li X, Sutcliffe MJ, Schwartz TW, Dobson CM. Sequence-specific proton NMR assignments and solution structure of bovine pancreatic polypeptide. Biochemistry 1992;31(4):1245-1253.

125. Lewis RJ, Brannigan JA, Offen WA, Smith I, Wilkinson AJ. An evolutionary link between sporulation and prophage induction in the structure of a repressor: anti-repressor complex. Journal of Molecular Biology 1998;283(5):907-912.

126. Leijonmarck M, Liljas A. Structure of the C-terminal domain of the ribosomal protein L7/L12 from Escherichia coli at 1.7 .ANG. Journal of Molecular Biology 1987;195(3):555-580.

127. Berndt K, Guentert P, Wuethrich K. Nuclear magnetic resonance solution structure of dendrotoxin K from the venom of Dendroaspis polylepis polylepis. Journal of Molecular Biology 1993;234(3):735-750.

128. keassar Chen lM. a novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. journal of Molecular Biology 2003;329:159 - 151 174.

129. Bewley CA, Gustafson KR, Boyd MR, Covell DG, Bax A, Clore GM, Gronenborn AM. Solution structure of cyanovirin-N, a potent HIV-inactivating protein. Nature Structural Biology 1998;5(7):571-578.

130. Bewley CA. Solution structure of a cyanovirin-N:Mana1-2Mana complex structural basis for high-affinity carbohydrate-mediated binding to gp120. Structure (Cambridge, MA, United States) 2001;9(10):931-940.

131. Drohat AC, Tjandra N, Baldisseri DM, Weber DJ. The use of dipolar couplings for determining the solution structure of rat apo-S100B(bb). Protein Science 1999;8(4):800-809.

132. Roberts SA, Weichsel A, Grass G, Thakali K, Hazzard JT, Tollin G, Rensing C, Montfort WR. Crystal structure and electron transfer kinetics of CueO, a multicopper oxidase required for copper homeostasis in Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America 2002;99(5):2766-2771.

133. Roberts SA, Wildner GF, Grass G, Weichsel A, Ambrus A, Rensing C, Montfort WR. A Labile Regulatory Copper Ion Lies Near the T1 Copper Site in the Multicopper Oxidase CueO. Journal of Biological Chemistry 2003;278(34):31958-31963.

134. Reva BA, Finkelstein AV, Skolnick J. What is the probability of a chance prediction of a protein structure with an rmsd of 6 .ANG.? Folding & Design 1998;3(2):141-147.

135. Raaijmakers H, Macieira S, Dias JM, Teixeira S, Bursakov S, Huber R, Moura JJG, Moura I, Romao MJ. Gene Sequence and the 1.8 .ANG. Crystal Structure of the Tungsten-Containing Formate Dehydrogenase from Desulfovibrio gigas. Structure (Cambridge, MA, United States) 2002;10(9):1261-1272.

136. Yankovskaya V, Horsefield R, Toernroth S, Luna-Chavez C, Miyoshi H, Leger C, Byrne B, Cecchini G, Iwata S. Architecture of succinate dehydrogenase and reactive oxygen species generation. Science (Washington, DC, United States) 2003;299(5607):700-704.

137. Leesong M, Henderson BS, Gillig JR, Schwab JM, Smith JL. Structure of a dehydratase-isomerase from the bacterial pathway for biosynthesis of unsaturated fatty acids: two catalytic activities in one active site. Structure (London) 1996;4(3):253-264.

138. Sharma AK, Rajashankar KR, Yadav MP, Singh TP. Structure of mare apolactoferrin: the N and C lobes are in the closed form. Acta Crystallographica, Section D: Biological Crystallography 1999;D55(6):1152-1157.

139. Sugahara M, Nodake Y, Sugahara M, Kunishima N. Crystal structure of dehydroquinate synthase from Thermus thermophilus HB8 showing functional importance of the dimeric state. Proteins FIELD Full Journal Title:Proteins 2005;58(1):249-252.

140. Ren J, Esnouf RM, Hopkins AL, Warren J, Balzarini J, Stuart DI, Stammers DK. Crystal Structures of HIV-1 Reverse Transcriptase in Complex with Carboxanilide Derivatives. Biochemistry 1998;37(41):14394-14403.

141. Ferguson KM, Kavran JM, Sankaran VG, Fournier E, Isakoff SJ, Skolnik EY, Lemmon MA. Structural basis for discrimination of 3-phosphoinositides by pleckstrin homology domains. Molecular Cell 2000;6(2):373-384.

142. Lawson CL, Zhang R, Schevitz RW, Otwinowski Z, Joachimiak A, Sigler PB. Flexibility of the DNA-binding domains of trp repressor. Proteins: Structure, Function, and Genetics 1988;3(1):18-31.

143. Meiler J, Baker D. Rapid protein fold determination using unassigned NMR data. Proceedings of the National Academy of Sciences of the United States of America 2003;100(26):15404-15409.

144. Ramirez BE, Voloshin ON, Camerini-Otero RD, Bax A. Solution structure of DinI provides insight into its mode of RecA inactivation. Protein Science 2000;9(11):2161-2169.

145. Alexeev D, Bury SM, Turner MA, Ogunjobi OM, Muir TW, Ramage R, Sawyer L. Synthetic, structural and biological studies of the ubiquitin system: chemically synthesized and native ubiquitin fold into identical three-dimensional structures. Biochemical Journal 1994;299(1):159-163.

146. Schumacher S, Clubb RT, Cai M, Mizuuchi K, Clore GM, Gronenborn AM. Solution structure of the Mu end DNA-binding Ib subdomain of phage Mu transposase: modular DNA recognition by two tethered domains. EMBO Journal 1997;16(24):7532-7541.

147. Kihara D, Lu H, Kolinski A, Skolnick J. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. Proceedings of the National Academy of Sciences of the United States of America 2001;98(18):10125-10130.

148. Vallely KM, Rustandi RR, Ellis KC, Varlamova O, Bresnick AR, Weber DJ. Solution Structure of Human Mts1 (S100A4) As Determined by NMR Spectroscopy. Biochemistry 2002;41(42):12670-12680.

149. Dempsey AC, Walsh MP, Shaw GS. Unmasking the Annexin I Interaction from the Structure of Apo-S100A11. Structure (Cambridge, MA, United States) 2003;11(7):887-897.

150. Brodersen DE, Etzerodt M, Madsen P, Celis JE, Thogersen HC, Nyborg J, Kjeldgaard M. EF-hands at atomic resolution: the structure of human psoriasin (S100A7) solved by MAD phasing. Structure (London) 1998;6(4):477-489.

151. Kobayashi N, Koshiba, S., Inoue, M., Kigawa, T., Yokoyama, S. RIKEN Structural Genomics/Proteomics Initiative (RSGI), Solution structure of mouse CGI-38 protein. To be Published

152. Murakami S, Nakashima R, Yamashita E, Yamaguchi A. Crystal structure of bacterial multidrug efflux transporter AcrB. Nature (London, United Kingdom) 2002;419(6907):587-593.

153. Yu EW, McDermott G, Zgurskaya HI, Nikaido H, Koshland DE, Jr. Structural Basis of Multiple Drug-Binding Capacity of the AcrB Multidrug Efflux Pump. Science (Washington, DC, United States) 2003;300(5621):976-980.

154. Yu EW, Aires JR, McDermott G, Nikaido H. A periplasmic drug-binding site of the AcrB multidrug efflux pump: A crystallographic and site-directed mutagenesis study. Journal of Bacteriology 2005;187(19):6804-6815.

155. Ronning DR, Guynet C, Ton-Hoang B, Perez ZN, Ghirlando R, Chandler M, Dyda F. Active site sharing and subterminal hairpin recognition in a new class of DNA transposases. Molecular Cell 2005;20(1):143-154.

156. Badger J, Sauder JM, Adams JM, Antonysamy S, Bain K, Bergseid MG, Buchanan SG, Buchanan MD, Batiyenko Y, Christopher JA, Emtage S, Eroshkina A, Feil I, Furlong EB, Gajiwala KS, Gao X, He D, Hendle J, Huber A, Hoda K, Kearins P, Kissinger C, Laubert B, Lewis HA, Lin J, Loomis K, Lorimer D, Louie G, Maletic M, Marsh CD, Miller I, Molinari J, Muller-Dieckmann HJ, Newman JM, Noland BW, Pagarigan B, Park F, Peat TS, Post KW, Radojicic S, Ramos A, Romero R, Rutter ME, Sanderson WE, Schwinn KD, Tresser J, Winhoven J, Wright TA, Wu L, Xu J, Harris TJR. Structural analysis of a set of proteins resulting from a bacterial genomics project. Proteins: Structure, Function, and Bioinformatics 2005;60(4):787-796.

157. Martin ACR, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, Laskowski RA, Mitchell JBO, Taroni C, Thornton JM. Protein folds and functions. Structure (London) 1998;6(7):875-884.

158. Russell RB, Ponting CP. Protein fold irregularities that hinder sequence analysis. Current Opinion in Structural Biology 1998;8(3):364-371.

159. Narasimhan J, Wang M, Fu Z, Klein JM, Haas AL, Kim J-JP. Crystal Structure of the Interferon-induced Ubiquitin-like Protein ISG15. Journal of Biological Chemistry 2005;280(29):27356-27365.

160. Milani M, Savard, P.-Y., Oullet, H., Ascenzi, P., Guertin, M., Bolognesi, M. . A TyrCD1/TrpG8 hydrogen bond network and a TyrB10-TyrCD1 covalent link shape the heme distal site of Mycobacterium tuberculosis hemoglobin O. PNAS 2003 v100:5766-5771.

161. Udomsinprasert R, Pongjaroenkit, S., Wongsantichon, J., Oakley, A.J., Prapanthadara, L.A., Wilce, M.C., Ketterman, A.J. . Identification, characterization and structure of a new Delta class glutathione transferase isoenzyme. . BiochemJ 2005 388 763-771.

162. Fotin A, Cheng, Y., Grigorieff, N., Walz, T., Harrison, S.C., Kirchhausen, T. . Structure of an auxilin-bound clathrin coat and its implications for the mechanism of uncoating Nature 2004 432 649-653.

163. Fotin A, Cheng, Y., Sliz, P., Grigorieff, N., Harrison, S.C., Kirchhausen, T., Walz, T. . Molecular model for a complete clathrin lattice from electron cryomicroscopy. Nature 2004 432 573-579.

164. Schulze FW, Petrick HJ, Cammenga HK, Klinge H. Thermodynamic properties of the structural analogs benzo[c]cinnoline, trans-azobenzene, and cis-azobenzene. Zeitschrift fuer Physikalische Chemie (Muenchen, Germany) 1977;107(1):1-19.

165. Talaty ER, Fargo JC. Thermal cis-trans isomerization of substituted azobenzenes. Correction of the literature. Chemical Communications (London) 1967(2):65-66.

166. Rau H. Azo compounds [Photochromium based on E-Z isomerization of double bonds]. Studies in Organic Chemistry (Amsterdam) 1990;40(Photochromism: Mol. Syst.):165-192.

167. Liu ZF, Hashimoto K, Fujishima A. Photoelectrochemical information storage using an azobenzene derivative. Nature (London, United Kingdom) 1990;347(6294):658-660.

168. Hugel T, Holland Nolan B, Cattani A, Moroder L, Seitz M, Gaub Hermann E. Single-molecule optomechanical cycle. Science (New York, NY) 2002;296(5570):1103-1106.

169. Ikeda T, Tsutsumi O. Optical switching and image storage by means of azobenzene liquid-crystal films. Science (Washington, D C) 1995;268(5219):1873-1875.

170. Sekkat Z, Dumont M. Photoassisted poling of azo dye doped polymeric films at room temperature. Applied Physics B: Photophysics and Laser Chemistry 1992;B54(5):486-489.

171. Muraoka T, Kinbara K, Kobayashi Y, Aida T. Light-Driven Open-Close Motion of Chiral Molecular Scissors. Journal of the American Chemical Society 2003;125(19):5612-5613.

172. Zhang C, Du MH, Cheng HP, Zhang XG, Roitberg AE, Krause JL. Coherent Electron Transport through an Azobenzene Molecule: A Light-Driven Molecular Switch. Physical Review Letters 2004;92(15):158301/158301-158301/158304.

173. Bortolus P, Monti S. Cis-trans photoisomerization of azobenzene. Solvent and triplet donors effects. Journal of Physical Chemistry 1979;83(6):648-652.

174. Zimmerman G, Chow L-Y, Paik U-J. The photochemical isomerization of azobenzene. Journal of the American Chemical Society 1958;80:3528-3531.

175. Rau H. Further evidence for rotation in the p,p* and inversion in the n,p* photoisomerization of azobenzenes. Journal of Photochemistry 1984;26(2-3):221-225.

176. Rau H, Lueddecke E. On the rotation-inversion controversy on photoisomerization of azobenzenes. Experimental proof of inversion. Journal of the American Chemical Society 1982;104(6):1616-1620.

177. Bortolus P, Monti S. cis .dblharw. trans Photoisomerization of azobenzene-cyclodextrin inclusion complexes. Journal of Physical Chemistry 1987;91(19):5046-5050.

178. Monti S, Orlandi G, Palmieri P. Features of the photochemically active state surfaces of azobenzene. Chemical Physics 1982;71(1):87-99.

179. Cattaneo P, Persico M. An ab initio study of the photochemistry of azobenzene. Physical Chemistry Chemical Physics 1999;1(20):4739-4743.

180. Blevins AA, Blanchard GJ. Effect of Positional Substitution on the Optical Response of Symmetrically Disubstituted Azobenzene Derivatives. Journal of Physical Chemistry B 2004;108(16):4962-4968.

181. Andersson JA, Petterson R, Tegner L. Flash photolysis experiments in the vapor phase at elevated temperatures. I: Spectra of azobenzene and the kinetics of its thermal cis-trans isomerization. Journal of Photochemistry 1982;20(1):17-32.

182. Lednev IK, Ye TQ, Matousek P, Towrie M, Foggi P, Neuwahl FVR, Umapathy S, Hester RE, Moore JN. Femtosecond time-resolved UV-visible absorption spectroscopy of trans-azobenzene: dependence on excitation wavelength. Chemical Physics Letters 1998;290(1,2,3):68-74.

183. Lednev IK, Ye T-Q, Abbott LC, Hester RE, Moore JN. Photoisomerization of a Capped Azobenzene in Solution Probed by Ultrafast Time-Resolved Electronic Absorption Spectroscopy. Journal of Physical Chemistry A 1998;102(46):9161-9166.

184. Lednev IK, Ye T-Q, Hester RE, Moore JN. Femtosecond Time-Resolved UV-Visible Absorption Spectroscopy of trans-Azobenzene in Solution. Journal of Physical Chemistry 1996;100(32):13338-13341.

185. Fujino T, Tahara T. Picosecond Time-Resolved Raman Study of trans-Azobenzene. Journal of Physical Chemistry A 2000;104(18):4203-4210.

186. Fujino T, Arzhantsev SY, Tahara T. Femtosecond Time-Resolved Fluorescence Study of Photoisomerization of trans-Azobenzene. Journal of Physical Chemistry A 2001;105(35):8123-8129.

187. Ishikawa T, Noro T, Shoda T. Theoretical study on the photoisomerization of azobenzene. Journal of Chemical Physics 2001;115(16):7503-7512.

188. Quennville J. First principles strudies of cis-trans photoisomeriation dynamics and excited states in ethylene, stilbene, azobenzene, and TATB. Urbana: University of Illinois at Urbana-Champaign; 2003.

189. Tiago ML, Ismail-Beigi S, Louie SG. Photoisomerization of azobenzene from first-principles constrained density-functional calculations. Journal of Chemical Physics 2005;122(9):094311/094311-094311/094317.

190. Ciminelli C, Granucci G, Persico M. The photoisomerization mechanism of azobenzene: A semiclassical simulation of nonadiabatic dynamics. Chemistry--A European Journal 2004;10(9):2327-2341.

191. Cembran A, Bernardi F, Garavelli M, Gagliardi L, Orlandi G. On the Mechanism of the cis-trans Isomerization in the Lowest Electronic States of Azobenzene: S0, S1, and T1. Journal of the American Chemical Society 2004;126(10):3234-3243.

192. Gagliardi L, Orlandi G, Bernardi F, Cembran A, Garavelli M. A theoretical study of the lowest electronic states of azobenzene: The role of torsion coordinate in the cis-trans photoisomerization. Theoretical Chemistry Accounts 2004;111(2-6):363-372.

193. Diau EW-G. A New Trans-to-Cis Photoisomerization Mechanism of Azobenzene on the S1(n,p) Surface. Journal of Physical Chemistry A 2004;108(6):950-956.

194. Chang C-W, Lu Y-C, Wang T-T, Diau EW-G. Photoisomerization Dynamics of Azobenzene in Solution with S1 Excitation: A Femtosecond Fluorescence Anisotropy Study. Journal of the American Chemical Society 2004;126(32):10109-10118.

195. Dohno C, Uno S-n, Nakatani K. Photoswitchable Molecular Glue for DNA. Journal of the American Chemical Society 2007;129(39):11898-11899.

196. Gorostiza P, Volgraf M, Numano R, Szobota S, Trauner D, Isacoff EY. Mechanisms of photoswitch conjugation and light activation of an ionotropic glutamate receptor. Proceedings of the National Academy of Sciences of the United States of America 2007;104(26):10865-10870.

197. Gaussian I, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; , Scuseria GER, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, , T.; Kudin KNB, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, , V.; Mennucci BC, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, , H.; Hada ME, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; , Nakajima TH, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; , Hratchian HPC, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, , R.; Stratmann REY, O.; Austin, A. J.; Cammi, R.; Pomelli, C.;, Ochterski JWA, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; , Dannenberg JJZ, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; , Farkas OM, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; , Ortiz JVC, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; , Liu GL, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; , Keith TA-L, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; , Gill PMWJ, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; and Pople, J. , A. Gaussian 03, Revision C.02. Wallingford CT; 2004.

198. Becke AD. Density-functional thermochemistry. III. The role of exact exchange. Journal of Chemical Physics 1993;98(7):5648-5652.

199. Hariharan PC, Pople JA. Influence of polarization functions on MO hydrogenation energies. Theoretica Chimica Acta 1973;28(3):213-222.

200. Biswas N, Umapathy S. Density Functional Calculations of Structures, Vibrational Frequencies, and Normal Modes of trans- and cis-Azobenzene. Journal of Physical Chemistry A 1997;101(30):5555-5566.

201. Traetteberg M, Hilmo I, Hagen K. A gas electron diffraction study of the molecular structure of trans-azobenzene. Journal of Molecular Structure 1977;39(2):231-239.

202. Bouwstra JA, Schouten A, Kroon J. Structural studies of the system trans-azobenzene/trans-stilbene. I. A reinvestigation of the disorder in the crystal structure of trans-azobenzene, C12H10N2. Acta Crystallographica, Section C: Crystal Structure Communications 1983;C39(8):1121-1123.

203. Fliegl H, Koehn A, Haettig C, Ahlrichs R. Ab Initio Calculation of the Vibrational and Electronic Spectra of trans- and cis-Azobenzene. Journal of the American Chemical Society 2003;125(32):9821-9827.

204. Mostad A, Roemming C. Refinement of the crystal structure of cis-azobenzene. Acta Chemica Scandinavica (1947-1973) 1971;25(10):3561-3568.

205. Naegele T, Hoche R, Zinth W, Wachtveitl J. Femtosecond photoisomerization of cis-azobenzene. Chemical Physics Letters 1997;272(5,6):489-495.

206. Tully JC. Molecular dynamics with electronic transitions. Journal of Chemical Physics 1990;93(2):1061-1071.

207. Kasha M. Characterization of electronic transitions in complex molecules. Discussions of the Faraday Society 1950;No. 9:14-19.

208. Fujino T, Arzhantsev SY, Tahara T. Femtosecond/picosecond time-resolved spectroscopy of trans-azobenzene: isomerization mechanism following S2(pp*) <- S0 photoexcitation. Bulletin of the Chemical Society of Japan 2002;75(5):1031-1040.

209. Kikuchi O, Azuki M, Inadomi Y, Morihashi K. Ab initio GB study of solvent effect on the cis-trans isomerization of 4-dimethylamino-4'-nitroazobenzene. Theochem 1999;468(1-2):95-104.

210. Hirose Y, Yui H, Sawada T. Effect of Potential Energy Gap between the n-p and the p-p State on Ultrafast Photoisomerization Dynamics of an Azobenzene Derivative. Journal of Physical Chemistry A 2002;106(13):3067-3071.

BIOGRAPHICAL SKETCH

Christina was born in a small town in Northeastern Pennsylvania. In 1998, she entered Bloomsburg University of Pennsylvania with aspirations of becoming a nurse. Because of the excellent tutelage she received from Dr. Wayne P. Anderson in an introductory organic/biochemistry course, she decided to change her major to chemistry. A year later, she joined Dr. Anderson in his research efforts. Together, they studied the geometric effects on the spectra of Vanadyl complexes as well as potential aluminum catalysts to be used in olefin polymerization. She also had the privilege of working with Dr. Anna Krylov in the summer of 2002 as a participant in the Research Experience for Undergraduates program at the University of Southern California.

In 2003, Christina entered the graduate program at the University of Florida and immediately began working for Dr. Adrian Roitberg. Her initial studies were focused on determining the isomerization pathways of Azobenzene but she is now more interested in bio-systems. During the summer of 2005, she had the opportunity to participate in the National Science Foundation's East Asia and Pacific Summer research Institutes program working for Dr. Jill Gready, at the Australian National University.