8-15-2006

Structural analysis of the EGR family of transcription factors: Templates for predicitng protein - DNA internations

Jamie Duke

Structural Analysis of the EGR Family of Transcription Factors:

Templates for Predicting Protein – DNA Interactions

Approved: _____

Gary R. Skuse, Ph.D.
Thesis Advisor

_____

Richard L Doolittle, Ph.D
Head, Department of Biological Sciences

Submitted in partial fulfillment of the requirements for the Master of Science degree in Bioinformatics at the Rochester Institute of Technology.

Jamie L. Duke
May 2006

Thesis/Dissertation Author Permission Statement

Title of thesis or dissertation: Structural Analysis of the EGR Family of Transcription Factors: Templates for Predicting Protein – DNA Interactions.

Name of Author: Jamie L. Duke
Degree: Master's of Science
Program: Bioinformatics
College: Science

I understand that I must submit a print copy of my thesis or dissertation to the RIT Archives, per current RIT guidelines for the completion of my degree. I hereby grant to the Rochester Institute of Technology and its agents the non-exclusive license to archive and make accessible my thesis or dissertation in whole or in part in all forms of media in perpetuity. I retain all other ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all parts of this thesis or dissertation.

Print Reproduction Permission Granted:

I, Jamie L. Duke, hereby grant permission to the Rochester Institute of Technology to reproduce my print thesis or dissertation in whole or in part. Any reproduction will not be for commercial use or profit.

Signature of Author: Jamie L Duke

Date: 20 June 2006

Inclusion in the RIT Digital Media Library Electronic Thesis & Dissertation (ETD) Archive

I, Jamie L. Duke, additionally grant to the Rochester Institute of Technology Digital Media Library (RIT DML) the non-exclusive license to archive and provide electronic access to my thesis or dissertation in whole or in part in all forms of media in perpetuity.

I understand that my work, in addition to its bibliographic record and abstract, will be available to the world-wide community of scholars and researchers through the RIT DML. I retain all other ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation. I am aware that the Rochester Institute of Technology does not require registration of copyright for ETDs.

I hereby certify that, if appropriate, I have obtained and attached written permission statements from the owners of each third party copyrighted matter to be included in my thesis or dissertation. I certify that the version I submitted is the same as that approved by my committee.

Signature of Author: Jamie L Duke

Date: 20 June 2006

Thesis Advisor

Dr. Carlos J. Camacho
Department of Computational Biology
University of Pittsburgh

Thesis Committee

Dr. Gary Skuse
Department of Biological Sciences
Rochester Institute of Technology

Dr. James Halavin
Department of Mathematics & Statistics
Rochester Institute of Technology

Dr. Paul Craig
Department of Chemistry
Rochester Institute of Technology

Abstract

The EGR family of transcription factors is known to be activated in cells exposed to growth factors in a variety of tissues. The overall structure of the family is highly conserved while the amino acid sequence can be quite diverse allowing for a wide array of DNA recognition sequences. Through homology modeling it is possible to reproduce the structure of the DNA binding domain of EGR proteins, which consists of three zinc fingers. It has also been determined through molecular dynamic simulations that most side chains within the domain reach an equilibrium state. However, residues that are essential for DNA binding are seen throughout the simulation as not reaching an equilibrium state, but constantly sampling available conformational space. Furthermore, through cluster analysis the three recognition residues in each zinc finger are found to have side chain conformations that are optimal for DNA recognition. These studies help to show a possible mechanism for zinc finger recognition of DNA and create homology modeled proteins that are able to be used in protein – DNA interaction prediction.

List of Figures

List of Tables

List of Equations

List of Appendices

Acknowledgements

I would like to thank Dr. Carlos Camacho for allowing me to start this project during the summer of 2005 with Bioengineering and Bioinformatics Summer Institute (BBSI) hosted by the Department of Computational Biology at University of Pittsburgh, and for allowing me to continue this project at RIT. Without the NIH and NSF grant through the BBSI at the University of Pittsburgh, I would not have been able to start doing the research that ultimately led to the completion of my thesis. I would also like to thank Dr. Gary Skuse, Dr. Paul Craig and Dr. James Halavin for serving on my thesis committee and helping me through the thesis process. Finally, I would like to thank my parents, Candy and Russ, as well as my sister, Danielle, for all of their encouragement and support while continuing my education.

Table of Contents

Introduction

Understanding interactions between proteins and DNA has become a major area of research in biology within the past few years. Furthermore, the challenges of predicting these interactions have lured both computational biologists and bioinformaticists into joining the expedition. By understanding the general mechanism for protein – DNA binding, it is possible to further understand cell regulation and develop more specific pharmaceuticals for existing diseases.

Focusing on specific families of proteins has allowed for advancements in the field, specifically with the early growth response factor (EGR) family. This family of transcription factors has proved useful due to the fact that it is highly conserved and extensively studied. It has been implicated in a diverse set of cellular processes including cell growth, differentiation, and apoptosis (Gashler and Sukhatme 1995). The EGR family of transcription factors is within the class of immediate early genes (IEGs) which are the group of genes directly activated by extracellular signals. Immediate early genes have been referred to as the "gateway to the genomic response" as they control the regulation of downstream genes, with two subclasses: effector and regulatory genes (Davis et al. 2003). EGR proteins respond to mitogens, differentiation stimuli, tissue injury and signals from neuron excitation (Gashler and Sukhatme 1995). Mitogens are the dominant form of extracellular signals for EGR protein induction and also cause the quickest reaction in EGR production.

All members of the EGR family contain two or three zinc finger domains which are used in conjunction with one another to recognize specific DNA sequences (Benos et al. 2002). Each zinc finger recognizes approximately 5 bases within the DNA strand, and has an overlap in the recognition of bases due to coordination between the other fingers in the protein (Palliard et al. 2004). Christy and Nathans first described EGR proteins in 1989 as containing three repeated zinc finger domains, and they hypothesized that collectively these domains worked as transcription factors to recognize a specific DNA sequence. The gene they describe, *zif268*, is

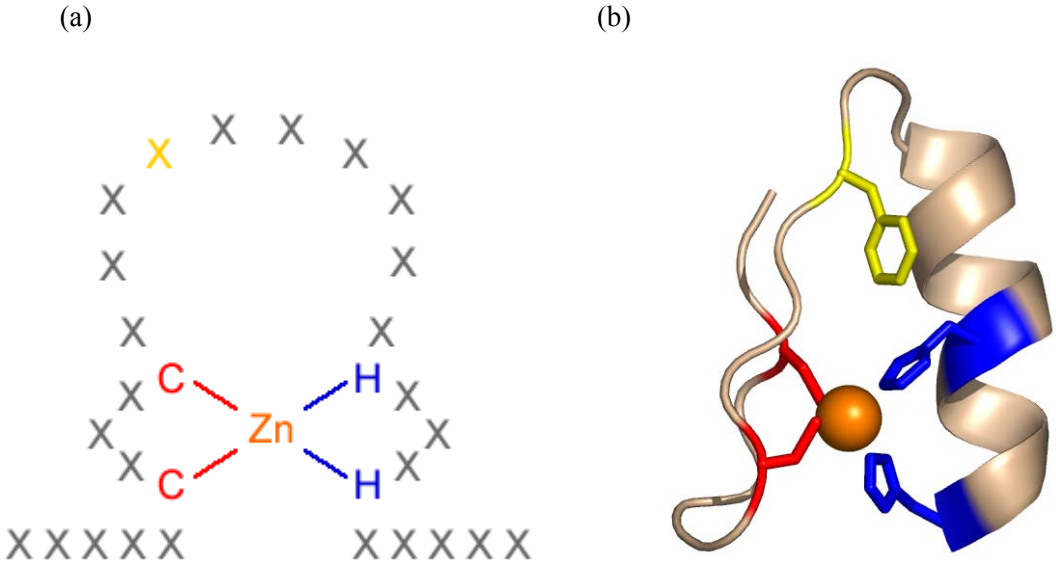


Figure 1. Patterns of zf-C2H2 domain. (a) The two dimensional representation of the zf-C2H2 domain. The four conserved residues that coordinate the zinc ion are shown in the center of the domain. The yellow 'X' four residues away from the second cysteine residue is also a conserved aromatic ring essential to the stability of the domain. Adapted from Prosite figure. **(b)** Three dimensional representation of the zinc finger C2H2 domain. The amino acid and zinc coloration is the same as in (a).

widely expressed in mouse tissues and is also known by the names *NGF-1A*, *egr-1*, and *Krox-24*. The consensus DNA sequence for zinc finger binding of *zif268* is $GCG\frac{G}{T}GGGCG$. A high affinity for guanine in DNA recognition is specific to this family.

Each zinc finger in the EGR family of transcription factors is of the type zf-C2H2, where 'zf' stands for zinc finger and C2H2 represents the 2 conserved cysteine and 2 conserved histidine residues. The zf-C2H2 domain has a consensus pattern of: 'x-C-x(1-5)-C-x(12)-H-x(3-6)-H' (Hulo et al. 2004). The 'X' in the pattern stands for any amino acid with the amount of repeating units in parenthesis. Structural stability in the domain is provided by the four conserved residues working together to coordinate a zinc ion along with a conserved aromatic ring located 4 amino acids away from the second cysteine residue. Figure 1 shows the zf-C2H2 domain in both a two- and three-dimensional manner, as well as the details of how the four conserved residues coordinate the zinc ion. Each zinc finger domain includes one α-helix and two β-strands. The α-helix is inserted into the major groove of the DNA double helix and is responsible for recognizing

at most 5 nucleic acids. The two conserved histidine residues responsible for coordinating the zinc ion are located towards the C-terminal of the α-helix pointing away from the DNA, as can be seen in Figure 2.

The EGR family of proteins has a set method of DNA recognition that is coordinated between the three zinc fingers, with each finger specifically recognizing three nucleotides. Each domain has three key residues for interaction with DNA, the negative one position, the third position and the sixth position in the α-helix, and can be seen in Figure 3. DNA recognition is specific to the residues that are located within the three key residues, and the given recognition sequence can be changed by mutating the protein binding region (Palliard et al. 2004). Secondary interactions also occur within the binding region through water molecules. These secondary interactions help to overlap the recognition of DNA between the zinc finger domains and involve additional residues located within the α-helix. When secondary interactions are included, each domain can recognize between four and seven nucleotides.

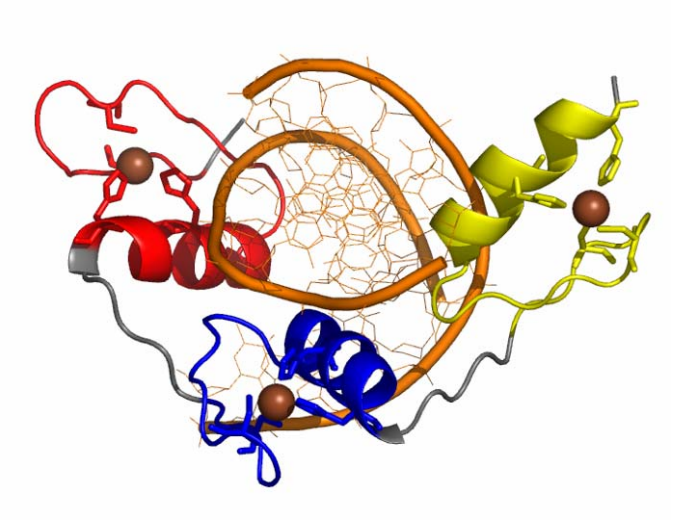


Figure 2. Representative Structure of the EGR Family of Transcription Factors. Seen here is the EGR protein from *Mus musculus* (PDB identifier: 1A1F). The DNA structure is shown in orange, with the backbone outlined. Each of the zinc finger domains is highlighted, starting with the N-terminus domain in yellow, followed by the second domain in blue, and the C-terminus domain in red. The four conserved amino acids, two cysteine and two histidine, can be seen coordinating the space-filled zinc atom. The conserved histidine residues are located within the α-helix, pointing away from the DNA chain. As can be seen from the figure, each of the zinc finger domains is inserted into the major groove of the DNA. Each domain overlaps the recognized bases with the subsequent domain.

The proteins that include the classic zinc finger domain are quite diverse. According to Pfam version 17.0 (Bateman et al. 2004), there are a total of 32,784 different protein domains that have been identified with

a zf-C2H2 domain; 1,390 of those sequences are from humans and 1,085 are from *Mus musculus*. The various zf-C2H2 protein sequences are broken up into 235 different architectures, including the 'zf-C2H2, zf-C2H2, zf-C2H2' architecture that includes the EGR family of transcription factors. This architecture has 723 members of which 128 are from humans. There are 5 recognized human EGR proteins; however there may potentially be an additional 26 human proteins that could belong to the EGR family, determined by sequence length and proximity of zf-C2H2 domains to one another. The remaining human sequences do not exhibit similarity to the EGR family of proteins.

Although proteins with zinc finger domains are highly studied, there are few structures available relative to the number of sequences that contain the domain. There are a total of 43 structures in the Protein Data Bank (PDB) with resolved zinc finger domains as of October 2005. A preliminary BLASTp search of the PDB yielded a total of 9 to 11 structures that would be eligible for use as a template for structure prediction for the EGR family. Of these possible templates, one is of human origin and the remaining are from *Mus*

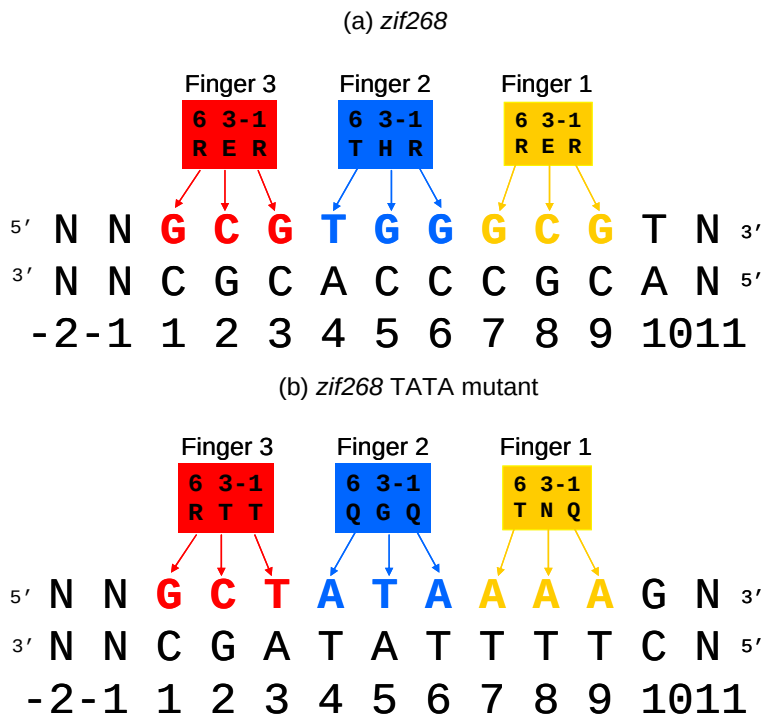


Figure 3. EGR Protein binding motif. (a) DNA binding motif for *zif268*, the representative for the EGR family. (b) DNA binding motif for a *zif268* variant recognizing the TATA box. The key interactions are shown for each domain with the solid arrows representing hydrogen bonds. Adapted from a figure by Palliard et al. 2004.

musculus. Four of the mouse structures are variants of the wild-type EGR family proteins. With the lack of structure data for this large family of proteins, structure prediction is necessary for being able to predict the potential interactions with DNA.

The main goals of this project consist of investigating the diversity of the EGR family in humans including basic bioinformatics analysis, carrying out homology modeling between resolved structures and known human EGR proteins and analyzing the structures to determine if protein – DNA interactions are possible.

Methods

This project has been working towards creating a pseudo-pipeline of programs that is able to take the DNA binding region of an EGR protein and be able to accurately determine a template for homology modeling in hopes of finding the most likely amino acid conformations to determine a structure for use in protein – DNA interaction prediction. While this process is not completely automated, significant progress has been made towards reducing the amount of human intervention in the process. There are numerous points in the following methods which need to be examined for quality purposes. Figure 4 shows the general flow of the methods from an amino acid sequence to a clustered protein structure.

EGR Family Analysis

Before beginning with homology modeling of the EGR family of transcription factors, it was first necessary to gain an understanding of the diversity its members. Members of the family were gathered through the Pfam database, starting with all proteins that have a zinc finger domain of the C2H2 type (Pfam ID: PF00096). It was determined through conversing with Dr. Takis Benos of the University of Pittsburgh that the proteins of interest for the purpose of the project were only the human zinc finger proteins with either two or three zf-C2H2 domains. To gain an understanding of the diversity of the individual domains, the domains were split apart from the

protein and a multiple sequence alignment was performed using ClustalW (Thompson et al. 1994). The alignment was performed using an opening gap penalty of 4.0, a gap extension penalty of 10.0 and the PAM matrix series. The goal of the alignment was to align the conserved cysteine and histidine residues in the zinc finger.

Template Determination

The first step involved in this project was to determine the proper structure(s) to use as a homology modeling template(s) for a given query sequence. With so few predetermined structures for the EGR family, a combination of multiple template structures may be needed to determine an accurate predicted structure. Basic analysis of EGR sequences were done using the results from BLASTp against the PDB database, and performing multiple sequence alignments to determine the best possible alignment and template (Altschul et al. 1997).

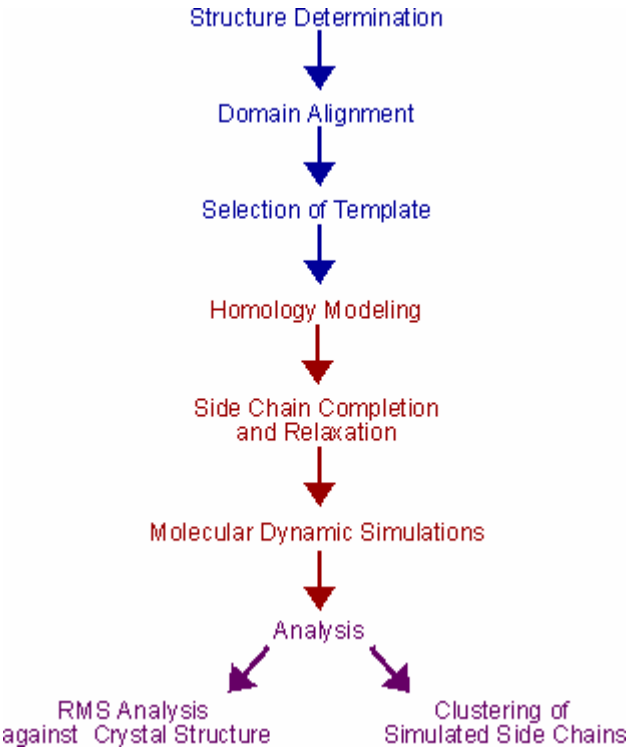


Figure 4. Flowchart of Methods. There are three groups of methods used: template determination, structure modeling, and structural analysis. Template determination, shown in blue, is the first three steps and is vital to having an accurate prediction. Structure modeling, shown in red, is the next three steps and is where all of the primary data are generated. Finally, the analysis is shown in purple and is broken into two different parts: analysis against the crystal structure and clustering of the simulated structure.

Structure Modeling

After the proper template for the query sequence was determined, the program Consensus was used to create

the homology model (Prasad et al. 2003). Although multiple templates may be beneficial during homology modeling, Consensus can only accommodate one structural template. Instead, Consensus uses five different methods to create an alignment between the template structure and the sequence to be modeled with the structure being determined through threading algorithms. Not only does Consensus perform its own sequence alignment between the template and query, it also predicts the accuracy of the secondary structure assignment for the alignment allowing the user to easily assess the prediction. Although Consensus selectively removes regions of dissimilarity and splits loosely connected domains to minimize potential misalignments, the zinc finger domains of the EGR proteins were not broken apart during the homology modeling process due to the high similarity between the models and the templates.

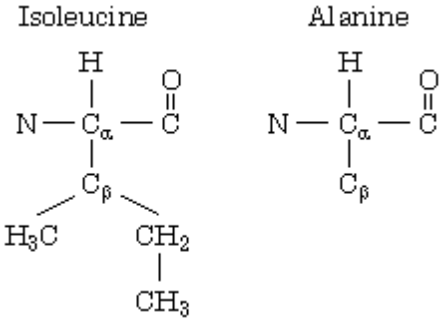


Figure 5. Figures of Isoleucine and Alanine. The structures of these amino acids are only similar through the C-β atom, and thus Isoleucine can be modeled only through its C-β leaving C-γ1, C-γ2, and C-δ not in the model.

The structure that was returned from Consensus was incomplete in respect to the side chains of the amino acids. Since threading algorithms are used, side chains for the query sequence can only be predicted in the model as much as they are in agreement with the template. For example, if an isoluecine from the query sequence is aligned with an alanine in the template, the predicted structure for isoleucine would contain the side chain through the C_β atom, as that is where the structural similarity between the two concludes. Figure 5 further demonstrates this point. Thus, completion of the side chains must be performed to be able to use the structure in further analysis. For this application, side chains were completed and minimized through the use of CHARMM (Brooks et al. 1983). Most importantly, CHARMM added the necessary hydrogen atoms to the amino acids, which are required to be able to do molecular dynamic simulations. The three domains of each EGR protein were manually broken apart for the side chain

completion and molecular dynamic simulations. The independent domains overlap one another by three residues.

The final modeling of the EGR protein involved molecular dynamic (MD) simulations to allow for fluctuation of the individual amino acids. According to Camacho, molecular dynamic simulations can help to predict native-like rotamer conformations of the side chains (2005). Proper prediction of side chain conformation is essential to predict interactions, whether they are between proteins or proteins and DNA. If one side chain is modeled improperly the resulting prediction of interactions could be incorrect.

Simulations are to be performed using GROMACS, in which the protein is simulated using the default settings (Lindahl et al., 2001). The protein is solvated in a dodecahedral water layer where there is a minimum distance from the protein to the edge of the water layer. The size of the water layer was varied, and the optimal water layer was found to be 1.75 angstroms when the simulation commences. Throughout the simulation, the carbonyl carbon, carbonyl oxygen, and nitrogen atoms of the backbone are constrained to keep the protein in the general state that was predicted through homology modeling. The simulation for each zinc finger domain of the protein lasted a total of 4.0 ns with each time step lasting 2 femtoseconds. The first 0.2 ns are removed from the simulation data for equilibrium purposes. Following the simulation, snapshots were taken every 2.5 ps recording structural information to be used for analysis of the simulation.

Structure Analysis

There were two different classes of structural analysis that have taken place: analysis against the crystal structure and analysis of the most highly sampled rotamer conformation. These two different forms of analysis are vital as this method is being verified using known structures for the query protein to have a measurement of the predictive abilities of the homology modeling and the molecular dynamic simulations.

The first form of analysis compared each amino acid for each snapshot of the MD simulation against the crystal structure. Each amino acid was individually analyzed by computing the distance between each heavy atom in the predicted structure and the crystal

$$d = \left(\left|x_1 - x_1\right| + \left|y_1 - y_2\right| + \left|z_1 - z_2\right|\right) \quad \text{(Eq. 1)}$$

structure, defined as the sum of the absolute value of the differences of each of the three dimensions. Before the distance was calculated, the amino acids were aligned to minimize the fluctuations the protein may have encountered during the MD simulation. To align the amino acids, the distance was calculated for each axis in space between the alpha carbons (C_α) and was subtracted from each atom in the predicted structure to align the two amino acids. Figure 6

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d^2} \quad \text{(Eq. 2)}$$

further demonstrates this point visually. The root mean square (RMS) was then calculated, with *d* being the distance between *N* pairs of structurally equivalent atoms. Additionally, the RMS was computed a second time for arginine residues to account for the symmetry in the terminal atoms. For example, atoms that are labeled the same were compared during the RMS calculation,

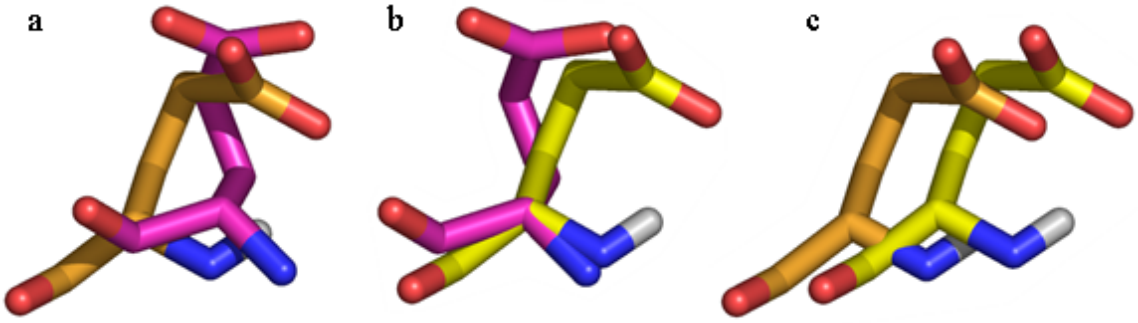


Figure 6. Alpha Carbon Centering. (a) Pictured are two unaligned glutamate residues, one from the crystal structure shown in magenta, and the other from the MD simulation shown in orange. Computing the RMS between these two residues is going to return a value that is misleading. (b) The two residues are aligned, with the crystal structure still shown in magenta, and the modeled structure is shown in yellow. Here the RMS is going to give much more insight into the difference between the two structures. (c) Shown is the difference between the unaligned and the aligned structure.

but in the case of arginine the two terminal atoms are both nitrogen, and are randomly assigned the designation of 'NH1' and 'NH2'. Thus when the terminal atoms were compared to the crystal structure it was also necessary to compare 'NH1' to 'NH2' and vice versa.

The second form of analysis aimed to complete the structural prediction process by clustering the individual side chains of the modeled protein. A neighbor clustering method was employed to gain general insight into the structure of a given amino acid. Kozakov *et* al. have described a clustering method for amino acid side chains that is applicable to this project (2005). Through clustering, we were looking to determine the most frequently sampled conformation throughout the MD simulation. This procedure relied on the calculation of the RMS for the heavy atoms in a side chain between pairs of time points. The procedure for calculating the RMS in this step of the analysis was the same as comparing the simulated structures to the crystal structure. The first 0.2 ns were dropped from the simulated data before clustering commenced to allow for equilibrium purposes. Approximately $1.16 * 10^6$ RMS calculations were required to compare each snapshot to every other snapshot for one residue. Each modeled protein has approximately 90 residues, equating to $1.04 * 10^8$ RMS calculations. Due to the magnitude of the RMS calculations required, three different clustering techniques were attempted throughout the course of this research, one technique used PERL, another used Java, and a third also used PERL but utilizing a different method. Between all three different techniques, conformations were clustered together if the RMS was found to be within a clustering radius of 0.75 angstroms.

The first two methods of clustering are very similar, but were instantiated in different programming languages. The clustering method was first developed using PERL, and is suitable for smaller data sets of approximately 400 models. This program holds every coordinate of every atom for each model throughout the simulation in memory, and then stores the RMS between each residue. Although it was memory intensive, the program was successful in clustering simulations of 1.0 ns in length, however when applied to a 4.0 ns simulation, the program quickly maximized memory usage, and took upwards of 24 to 36 hours to cluster one residue, which

would have equated to approximately three months to cluster one simulated protein. This result was highly unsatisfactory.

Next, a similar clustering algorithm was constructed in Java using the class structure and taking advantage of the hierarchical nature of object oriented programming. The system developed had classes for models within a simulation, residues within a model, and atoms within a residue. There was also a clustering class that applied the previously developed classes. This system was slightly different in the fact that only a subset of residues were accessed and held in memory at one point in time to keep memory usage low. The class method was also not ideal due to the fact that clustering still took approximately a week for one protein.

The third method of analysis developed utilized a program that computes the RMS between two PDB files. In the two previous clustering methods, a majority of the time was spent calculating the RMS, which was not done in an efficient method. By computing the RMS externally from the clustering algorithm and storing the results in a file, the time it takes for computing the RMS and clustering is significantly reduced. Computing the RMS was reduced to approximately one day per domain, and multiple domains were run simultaneously. For each comparison, a RMS file was created, which was accessed in the clustering algorithm. Following calculation of the RMS, the clustering program read in the RMS calculations, and clustered residues that were within 0.75Å radius of one another.

Results

EGR Family Analysis

A multiple sequence alignment was performed between all of the human EGR proteins listed in the Pfam database. A total of 719 domains were aligned encompassing a total of 287 EGR proteins. The alignment is a total length of 32 residues, with each domain being 23 residues in length. By having a low gap opening penalty of 4 and a larger gap extension penalty of 10,

CLUSTAL W is able to align three of the four conserved residues: the two cysteine residues and the first histidine. The fourth residue is the final histidine residue in the α-helix which is noted in the profile of the family as sometimes being replaced by a cysteine, which occurs in 2.2% of the zf-C2H2 domains. These two residues are not seen as easily interchangeable by CLUSTAL W in the alignment procedure as cysteine is polar neutral and histidine is polar basic (Figure 7a). Additionally, the distance between the histidine residues can vary by a difference of four locations at the tail end of the domain. Since there is variability in the location of the histidine, CLUSTAL W prefers to keep the end of the domain intact unless inserting a gap is more favorable than misaligning a residue (Figure 7b). The full alignment can be seen in Appendix A.

Homology Modeling

The three structures under investigation are a C2H2 zinc finger protein, a C2H2 zinc finger variant, and a C2H2 designed zinc finger. The C2H2 zinc finger protein (PDB 1AAY), was crystallized in 1997 with a 1.60 Å resolution and was chosen because of its sequence complementarities to the EGR family consensus (Elrod-Erickson et al., 1996). The second

```
(a)
Z297B_HUMAN_428-450          YECN--ICAKRFMWRDS--FHRHVTS-C----
Q6ZN18_HUMAN_45-70           YNCCWDQCQACFNSSPD--LADHIRS-IH---
Q6ZN29_HUMAN_3-26            YQCKK--CNVVFPRIFD--LITHQKKQ-C---
                               *    *              :

(b)
HELI_HUMAN_140-162           FHCNQ--CGASFTQKGN--LLRH--IKLH---
Q6ZMZ8_HUMAN_69-92           FHCDQ--CSYSCKRKDN--LNLH--KKLKH—
FOG2_HUMAN_363-385           FRCN--HCHFGFQTQRE--LLQHQ--ELH---
Q9C0D4_HUMAN_982-1004        FKCW--FCGRLYEDQEE--WMSHGQR--H---
                               *    *              :
```

Figure 7. Sample Alignments from the Multiple Sequence Alignment of EGR Protein Domains. (a) The misalignment between the final histidine and cysteine residues. Although gaps are inserted into the sequence, if the gaps were removed, the domain would be aligned correctly in this instance. (b) CLUSTAL W prefers to keep the carboxyl terminal of the domain together and places a priority in aligning the residues occurring before the final histidine.

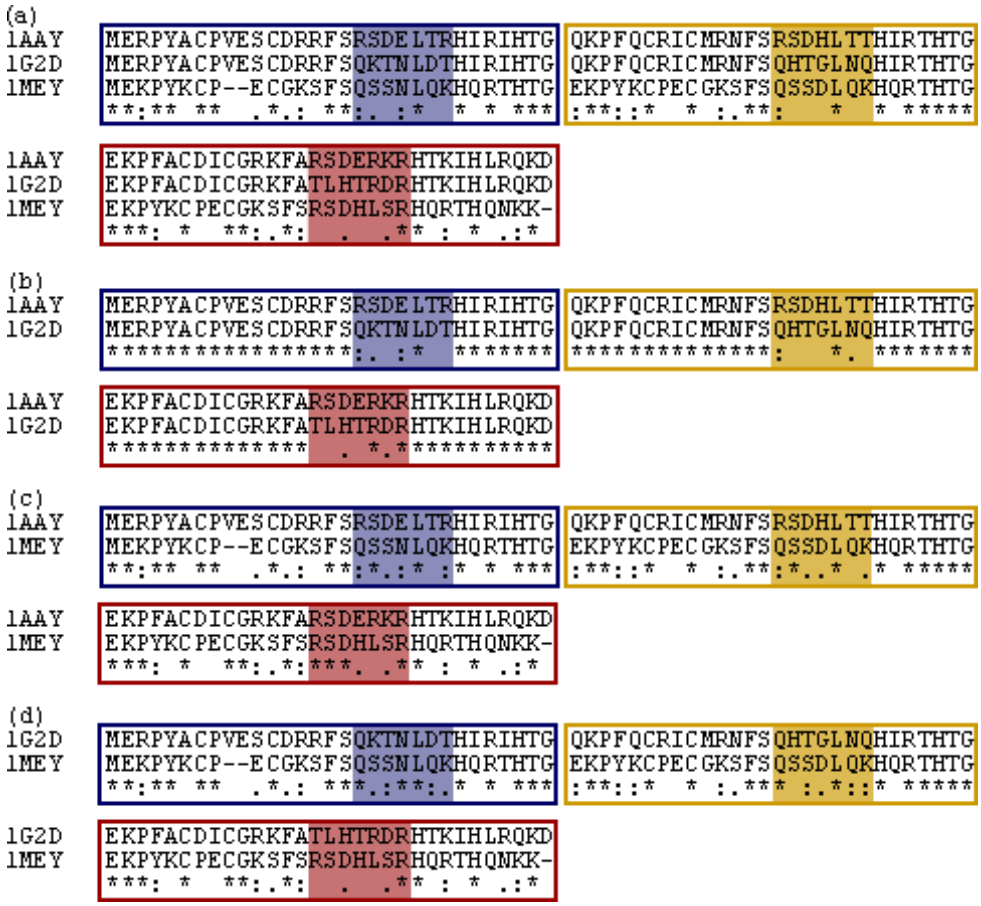


Figure 8. Full Alignments of EGR proteins of interest. (a) Multiple sequence alignment of all EGR Proteins in use. Domain 1 is outlined in blue, domain 2 in yellow, and domain 3 in red. The DNA binding region is also highlighted for each domain. (b) Pairwise alignment between 1AAY and 1G2D. The only difference between these two proteins lies in the DNA binding region. (c) Pairwise alignment between 1AAY and 1MEY. While there is more similarity in the DNA binding regions than between 1AAY and 1G2D at 86% there is only 38% identity. The overall identity and similarity is significantly lower at 49% and 77% respectively. (d) Pairwise alignment between 1G2D and 1MEY. Overall similarity between 1G2D and 1MEY is 76%, while the identity is 42%, and the DNA binding domain is 76% similar and 29% identical.

structure, the C2H2 zinc finger variant (PDB 1G2D), was crystallized in 2000 with a 2.20 Å resolution and was chosen for its relatively high sequence identity with 1AAY and its divergence of the binding domain (Wolf et al., 2001). These proteins have 79% sequence identity, but are only 14% identical in the DNA binding domains. The final structure that was chosen for analysis, the C2H2 designed zinc finger (PDB 1MEY), was crystallized in 1996 at 2.20 Å resolution (Kim and Berg, 1996). This protein has 49% identity with 1AAY, but unlike 1G2D,

1MEY has 38% identity to 1AAY in the binding domains. 1MEY was chosen to be modeled off of 1AAY due to the fact that similarity was slightly higher than that of 1MEY with 1G2D. The pairwise and multiple sequence alignments of these domains can be found in Figure 8. The three homology models that were constructed are the prediction of 1G2D with 1AAY as the template (GD-AA), 1AAY with 1G2D as the template (AA-GD), and 1MEY with 1AAY as the template (ME-AA).

Each model utilizes the sequence from an EGR protein with a known structure and goes through the process of homology modeling, side chain completion, molecular dynamic simulations and clustering. Upon completion of homology modeling, Consensus reports the strength of the alignment in terms of the compatibility of the five alignment processes, the secondary structure prediction, as well as two different structure predictions: the full structure prediction and the structure of the reliable regions. For the purposes of this project the full structure prediction was used. Output from Consensus for all three EGR modeled proteins can be found in Appendix B.

For GD-AA, Consensus had a strong alignment between the target and template, but was only confident about the structure prediction for the first domain and the beginning half of the second domain. This result is unexpected as the only regions of nonidentity are in the first seven amino acids of the α-helix for each domain. It is possible that there is a shift in the secondary structure prediction, which could have caused the lower confidence. However, the RMS between the backbones of GD-AA and 1G2D is 1.29 Å, which confers confidence to the structure prediction.

The structure prediction of AA-GD is more favorable than the prediction of GD-AA. Although the alignments between 1G2D and 1AAY are the same in both homology models, Consensus was more confident in predicting AA-GD and this was seen in the selection status of the output. The only area of difficulty for the prediction lies in the C terminus with the final

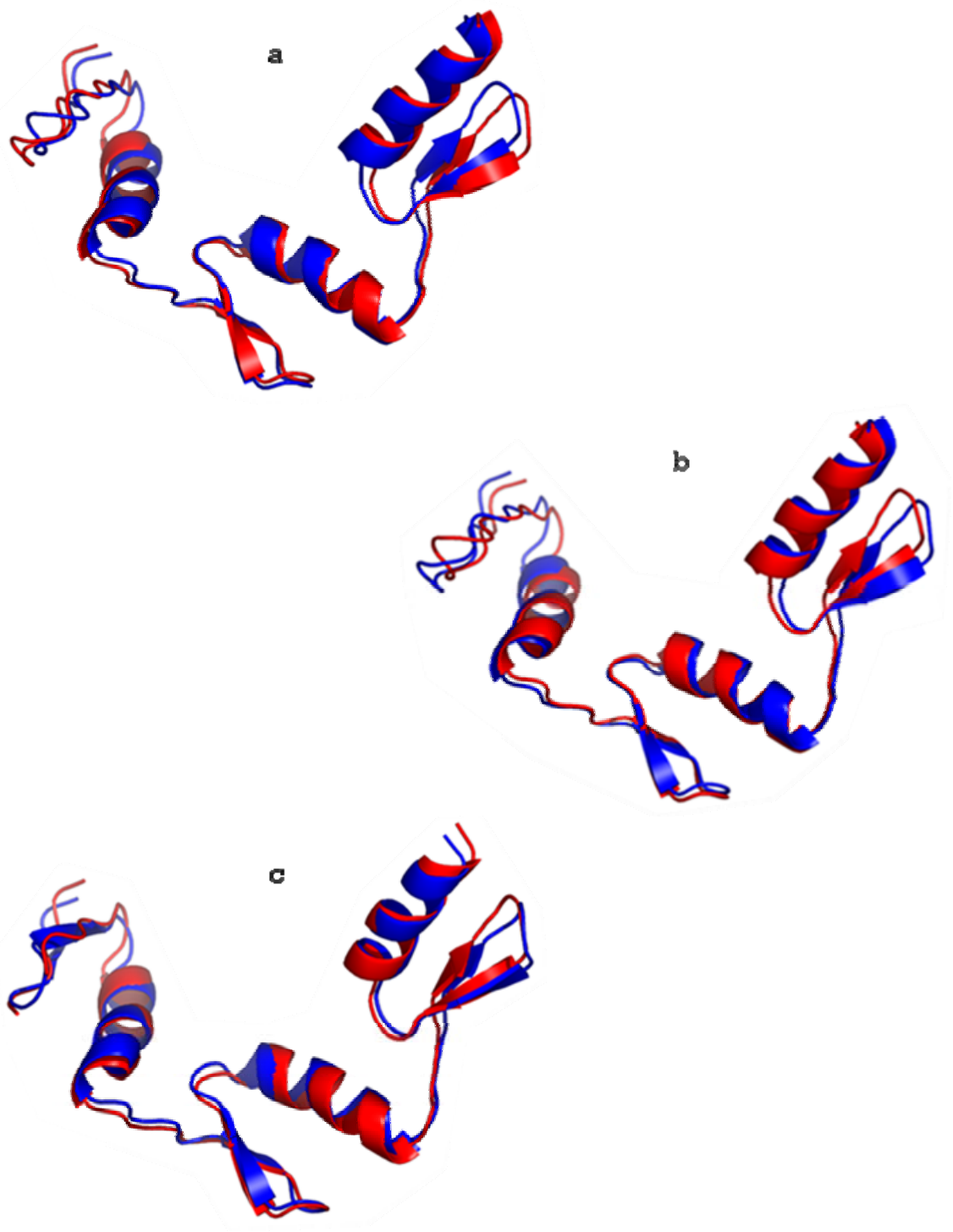


Figure 9. Differences between Homology Models and Crystal Structures. (a) GD-AA (red) modeled against 1G2D. (b) AA-GD (red) modeled against 1AAY (blue). (c) ME-AA (red) modeled against 1MEY (blue). Major differences in the homology models against the crystal structure is seen more prominanlty in the first and third domains, leftmost and rightmost domains respectively, particularly in the β-strands. The second domains appear to have the best fit between all three represented figures.

eleven amino acids of the protein. The RMS is also between the backbone of AA-GD and 1AAY is also very favorable at 1.29 Å.

Although there is less similarity between 1AAY and 1MEY than between 1G2D and 1AAY, Consensus has a very strong structural prediction for ME-AA. Much like AA-GD, ME-AA has a strong alignment throughout the whole protein, and Consensus is confident about the structural prediction through most of the model. Unlike AA-GD, ME-AA does not have a confident secondary structure prediction at either the N or C terminals of the protein, lacking confidence for 9 and 10 amino acids respectively. As the confidence is slightly lower for this model than the previous two, the RMS deviation is therefore slightly higher between ME-AA and 1MEY at 1.35 Å. The predicted models and crystal structures can be seen in Figure 9.

The molecular dynamic simulations for each EGR protein zinc finger domain yields structure and energy information for each time step throughout the simulation. To make the data more manageable, snapshots were taken every 2.5 ps creating 1600 total structures for the 4.0 ns simulation and were used for both types of analysis. Initially, the simulations were lacking a layer of water large enough to allow for proper side chain fluctuations, which was made evident through examining the clustering results. The initial water layer was 0.75 Å, which is insufficient for amino acids with larger side chains, for example arginine and lysine. Figure 10 shows potential hydrogen bonding that was occurring in the second zinc finger domain in the GD-AA simulated structure. These results showed neighboring side chains hydrogen bonding to one another, which was unsubstantiated in the literature for the crystal structure. By expanding the water layer the side chains are given more flexibility and less chance to hydrogen bond with neighboring residues.

Structure Analysis

The structures resulting from the MD simulations were first analyzed against the corresponding crystal structure. For each model there are nine different residues of interest, negative one position, third position and sixth position of the α-helix.

The first model analyzed was GD-AA against the crystal structure for 1G2D. The three positions in first domain exhibit three different characteristic states as can be seen in Figure 11. The first binding position is the negative one position, or the residue before the α-helix begins. The glutamine residue has very high fluctuation in the RMS throughout the course of the simulation. There is no pattern within the fluctuations, but the residue does seem to be going between two different states, one state that is close to the crystal structure, approximately 1.0Å difference, and another that is much further removed at approximately 2.5Å. The third position is occupied with an asparagine which resonates around and equilibrium point of 1.2Å difference from the crystal structure. Most residues in the domain are expected to reach some sort of equilibrium state, which residue 21 exemplifies. The sixth position of the helix is inhabited by a threonine. It is interesting that this residue clearly

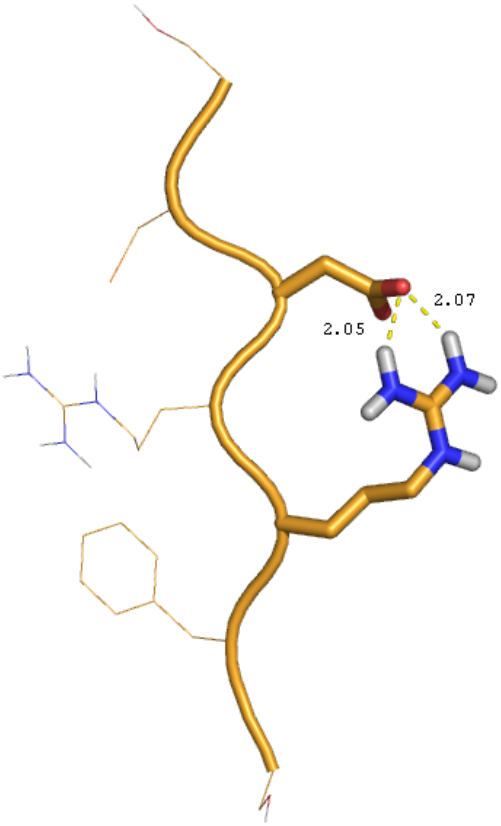


Figure 10. Hydrogen Bonding in Initial Molecular Dynamic Simulation Data Clustering. The aspartate and arginine residues from the second β-strand, shown in stick representation, have potential hydrogen bonds as they are located within 2.1Å. The oxygen atom from the aspartate could hydrogen bond with either NH_2 group in the arginine.

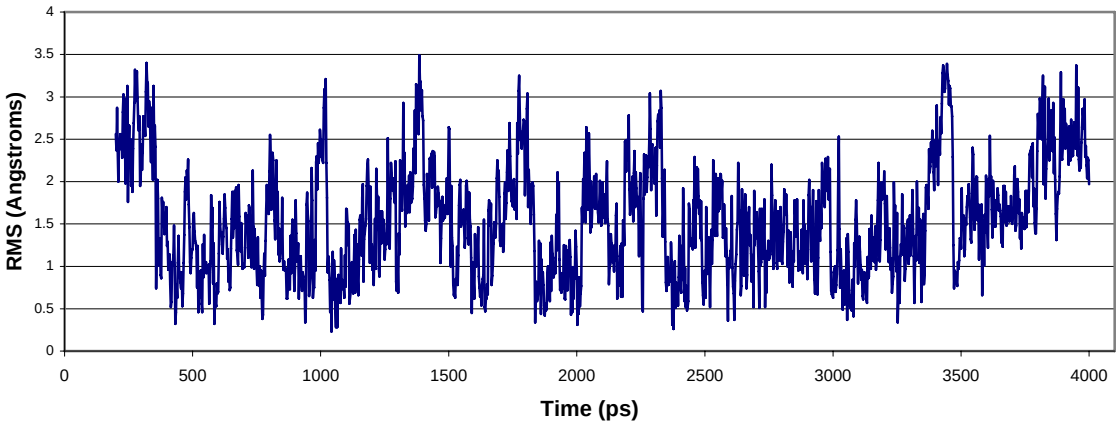
fluctuates between two apparent states, one state existing around 0.3Å difference from the crystal structure and the other state having a 1.1Å difference.

The residues in the second domain are not as interesting as the first domain, seen in Figure 12. Residue 46, a glutamine, in the negative one position in the α-helix and is very similar to the glutamine in the -1 position of the first domain. There seems to be fluctuation between two states, but it is not as distinct as other residues previously seen. The first state is approximately 1.0Å away from the crystal and the second state is approximately 2.25Å away, with the residue spending a significant amount of time in between the two states. The third position in the helix does not yield much information as the residue is a glycine; the fluctuation seen in Figure 12b is between atoms located in the backbone of the structure. The sixth position of the α-helix is also a glutamine residue, which has a similar RMS profile as residue 46. The lower state fluctuates just below 1.0Å difference from the crystal structure and the higher state is situated around 2.25Å. Unlike residue 46, residue 52 does not spend much time in an intermediary state.

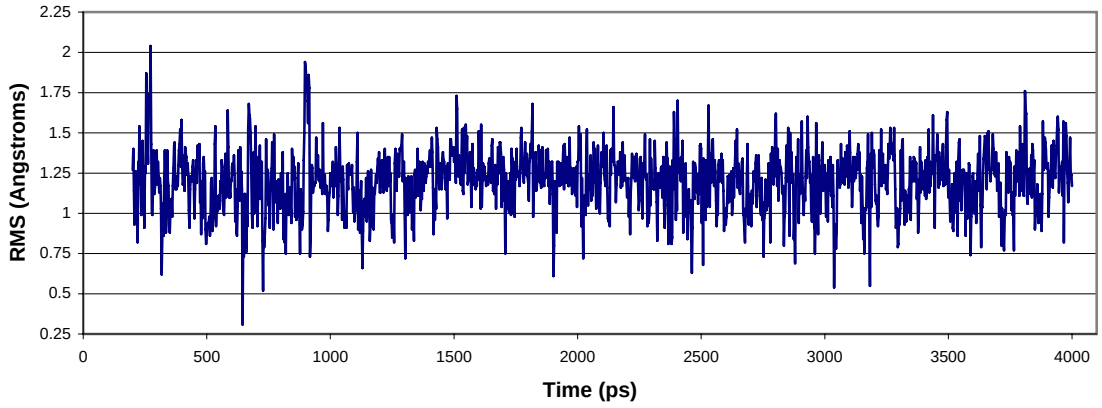
The third domain is comprised of two threonine residues and an arginine residue, seen in Figure 13. Residue 74 in negative one position of the α-helix has two very distinct states, the first fluctuating around 0.45Å and the second state existing at approximately 1.1Å difference from the crystal structure. Residue 77, the threonine in the third position of the helix is in an equilibrium state for most of the simulation around 0.4Å difference from the crystal structure, but in the beginning and the end of the simulation it jumps to the higher state at 1.3Å difference. The final residue of interest for GD-AA is residue 80, an arginine. For the majority of the simulation, the residue is located between 3.0Å and 4.5Å away from the crystal structure. While this residue has a more bulky side chain, it is not expected that it would be so far removed from the crystal structure.

The residues of interest for the first domain of AA-GD are fairly far removed from the crystal structure as can be seen in Figure 14. The negative one position fluctuates around 1.5Å away from the crystal structure with the higher state above 2.5Å difference. Although the lowest

(a) Residue 18 Glutamine (Domain 1 Position -1)



(b) Residue 21 Asparagine (Domain 1 Position 3)



(c) Residue 24 Threonine (Domain 1 Position 6)

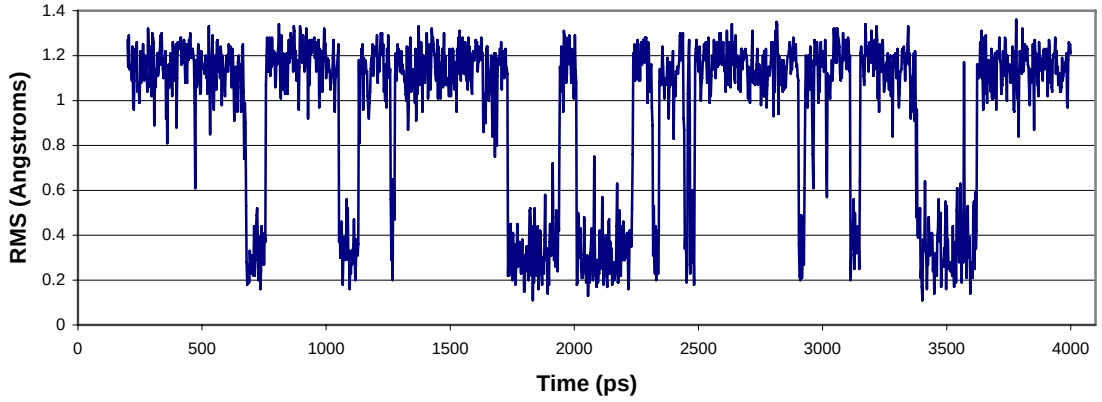
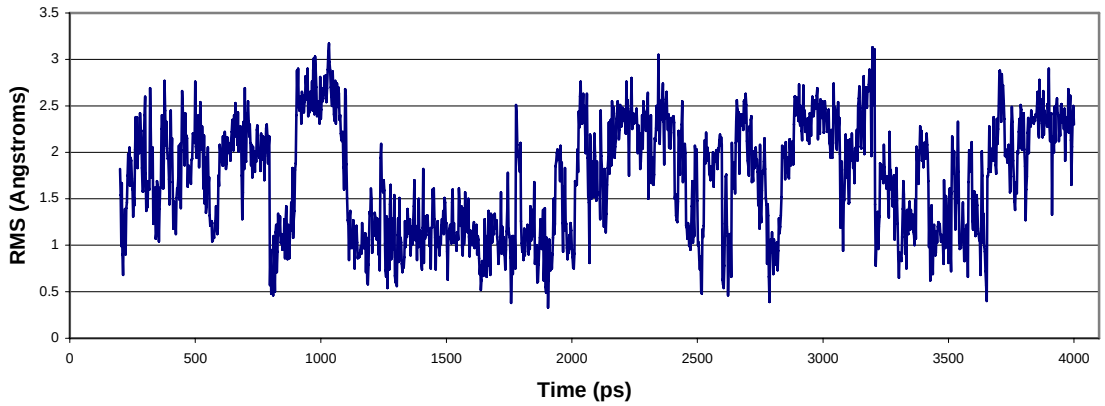
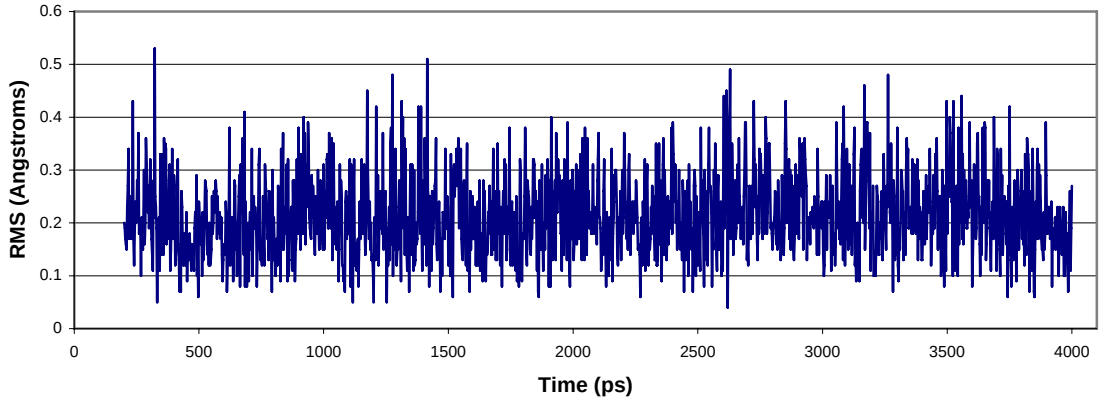


Figure 11. Graphs of Domain 1 of GD-AA. All graphs are the RMS between the crystal structure of 1G2D against the MD simulation data of GD-AA against time in picoseconds.

(a) Residue 46 Glutamine (Domain 2 Position -1)



(b) Residue 49 Glycine (Domain 2 Position 3)



(c) Residue 52 Glutamine (Domain 2 Position 6)

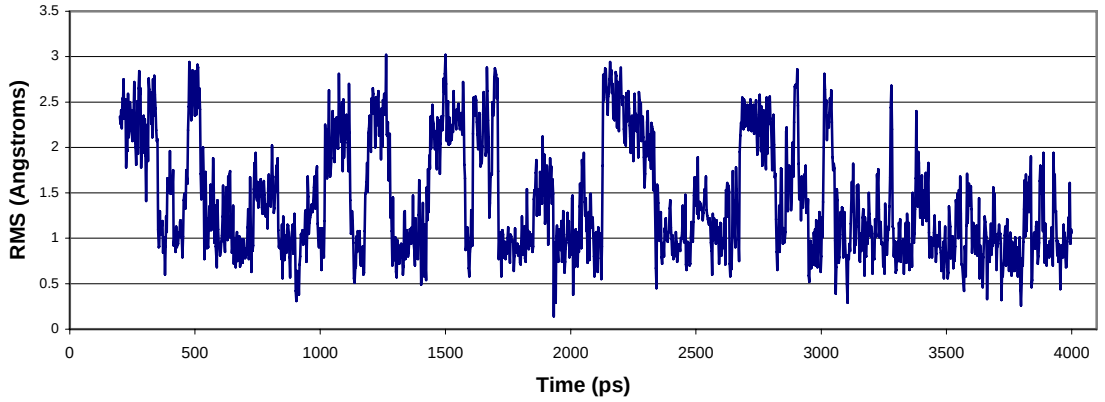
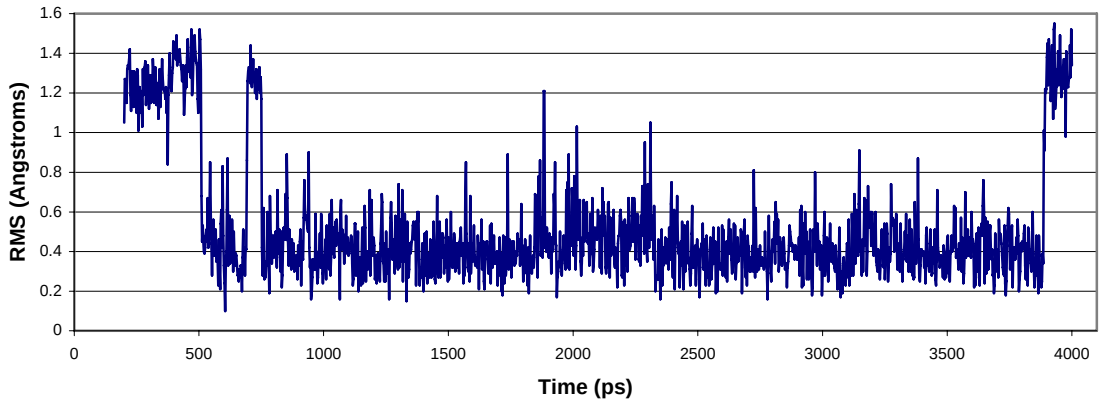
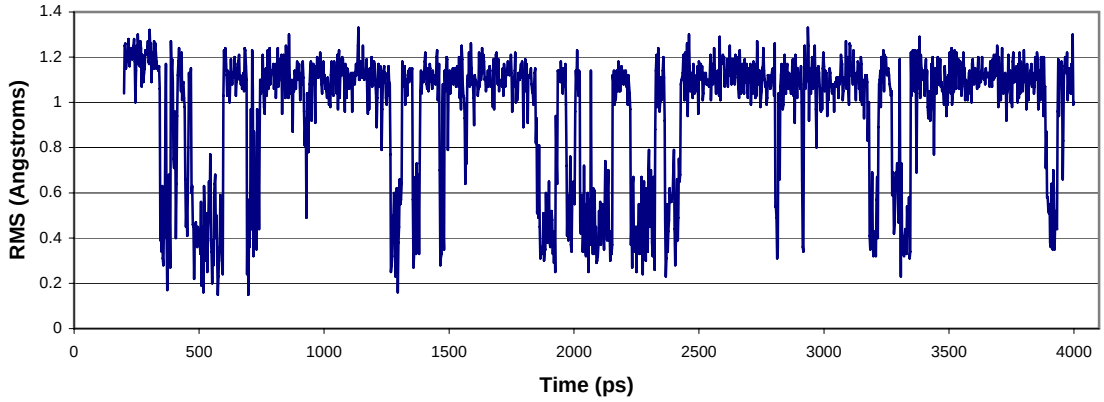


Figure 12. Graphs of Domain 2 of GD-AA. All graphs are the RMS between the crystal structure of 1G2D against the MD simulation data of GD-AA against time in picoseconds.

(a) Residue 74 Threonine (Domain 3 Position -1)



(b) Residue 77 Threonine (Domain 3 Position 3)



(c) Residue 80 Arginine (Domain 3 Position 6)

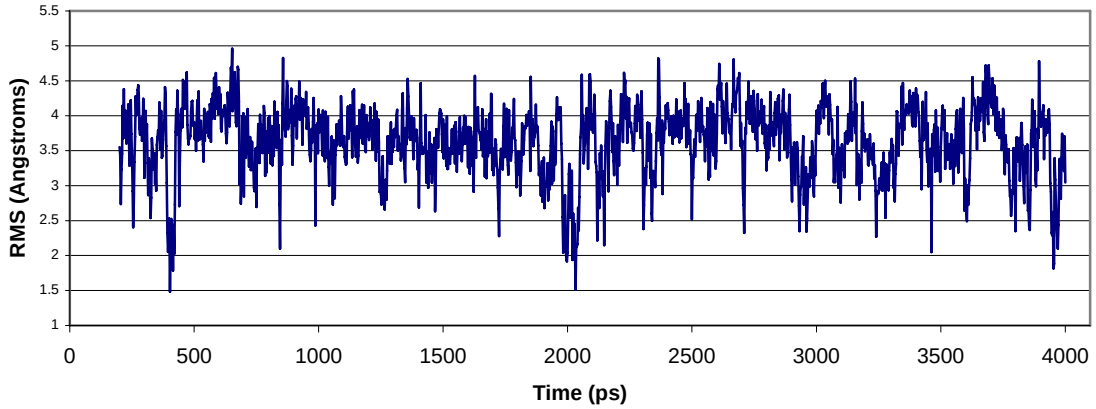


Figure 13. Graphs of Domain 3 of GD-AA. All graphs are the RMS between the crystal structure of 1G2D against the MD simulation data of GD-AA against time in picoseconds.

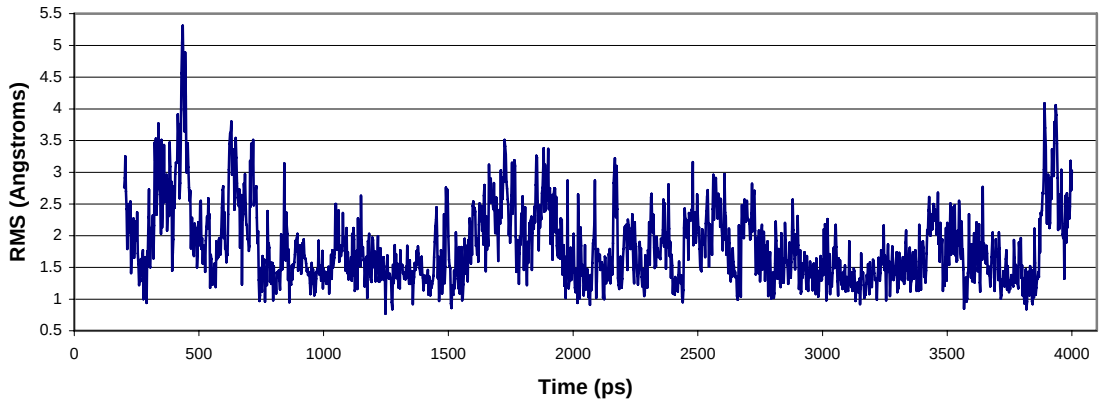
state is removed from the crystal structure, it is more desirable than the results for the previous arginine residue, GD-AA domain 3 position 6. Residue 21 vacillates between two states, and also spends time in an intermediary state. The low state is approximately 0.5Å difference from the crystal structure, where the higher state is 1.1Å away. The sixth position of the helix is occupied by another arginine residue, but unlike the arginine residue in the negative one position, this residue prefers to exist in a higher state around 2.5Å away from the crystal structure.

The residues of the second domain seem to exist in states of equilibrium, except for the first residue as can be seen in Figure 15. The arginine residue at the negative one position does not prefer to exist in any given state as it continually changes states. Towards the end of the simulation, the residue goes through a series of gradual increases followed by a sharp decline in RMS. This trend is interesting as the residue can quickly jump 3.0Å from a high state to a low state. Residue 49, a histidine, highly prefers a state that is 1.25Å away from the crystal structure. This residue, like the threonine residues seen previously also jumps sharply into a higher state of 2.5Å difference for brief periods of time. The sixth position of the α-helix is a threonine and prefers a state that is 0.4Å away from the crystal structure. The higher state resonates around 1.2Å away from the crystal structure.

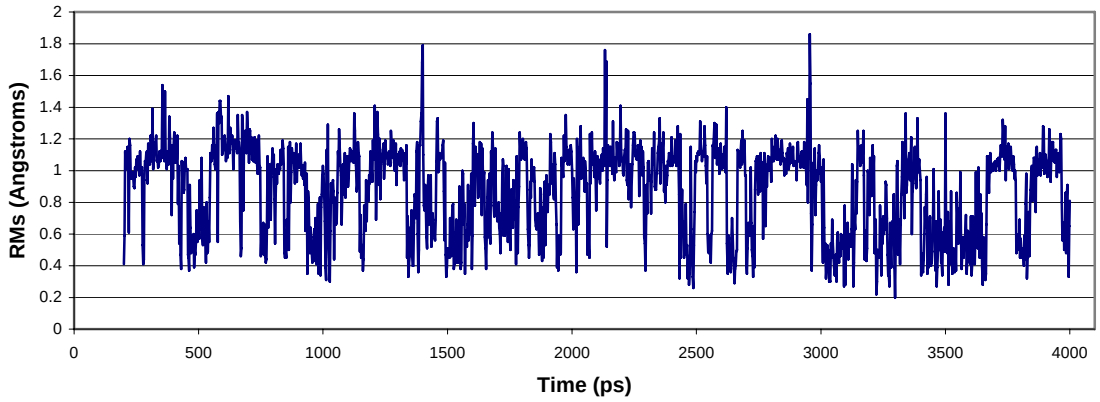
Much like the first domain, the third domain has the same sequence of residues, and has very similar RMS profiles as seen in Figure 16. The negative one position is an arginine residue and exists around 1.5Å away from the crystal structure. The third position of the helix fluctuates between 0.5Å and 1.25Å away from the crystal structure. The sixth position of the helix, an arginine residue, starts at a lower state around 1.5Å, and increases to a higher state around 2.75Å away from the crystal structure.

The model of ME-AA introduces new residues of aspartate and lysine into the DNA binding positions. The negative one position of first domain is a glutamine and is in a state of flux throughout the simulation not preferring one particular state, but the average RMS value

(a) Residue 18 Arginine (Domain 1 Position -1)



(b) Residue 21 Glutamate (Domain 1 Position 3)



(c) Residue 24 Arginine (Domain 1 Position 6)

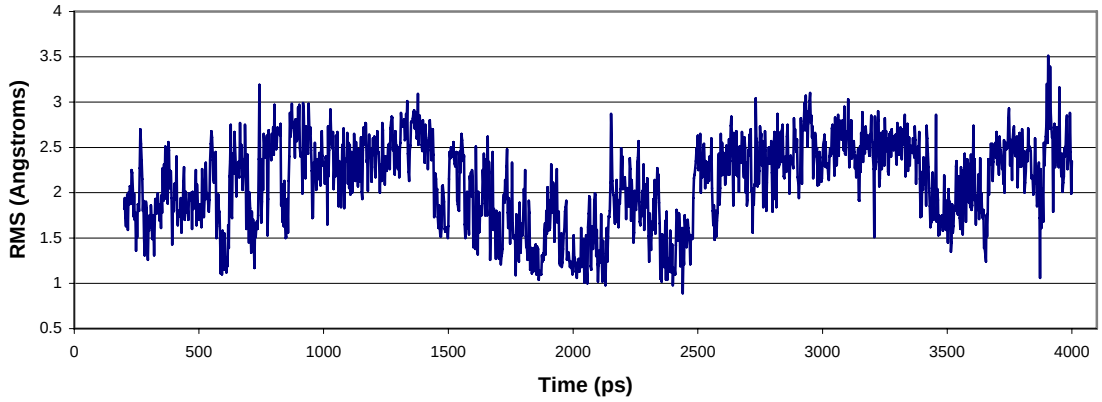
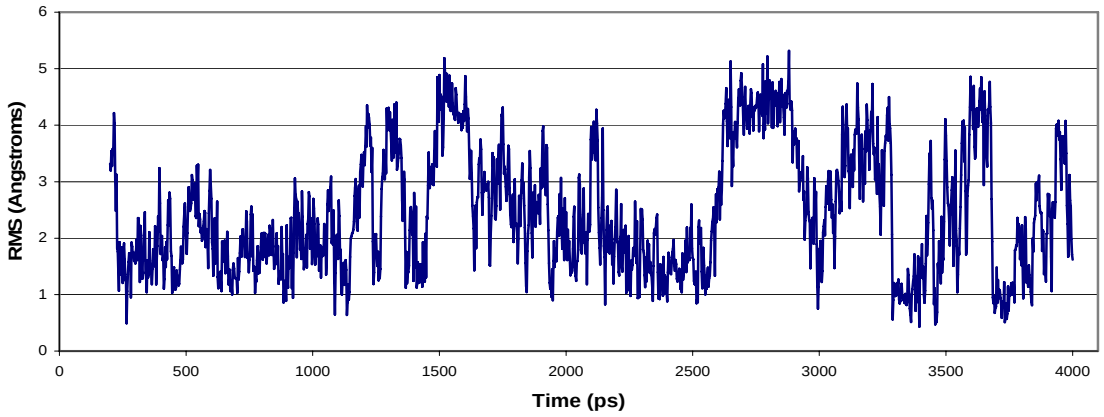
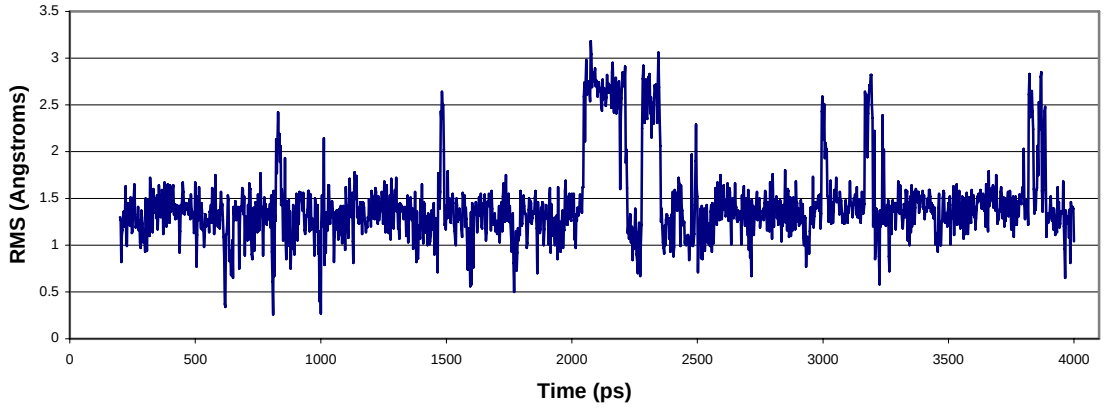


Figure 14. Graphs of Domain 1 of AA-GD. All graphs are the RMS between the crystal structure of 1AAY against the MD simulation data of AA-GD against time in picoseconds.

(a) Residue 46 Arginine (Domain 2 Position -1)



(b) Residue 49 Histidine (Domain 2 Position 3)



(c) Residue 52 Threonine (Domain 2 Position 6)

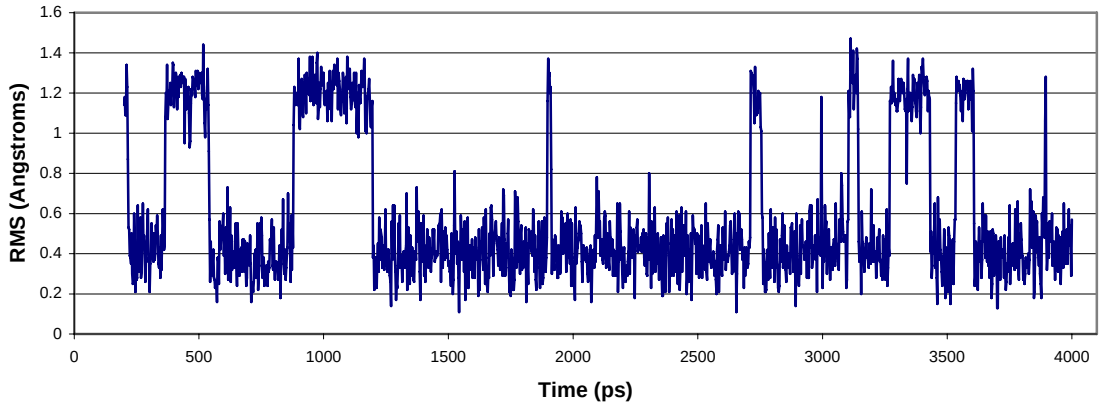
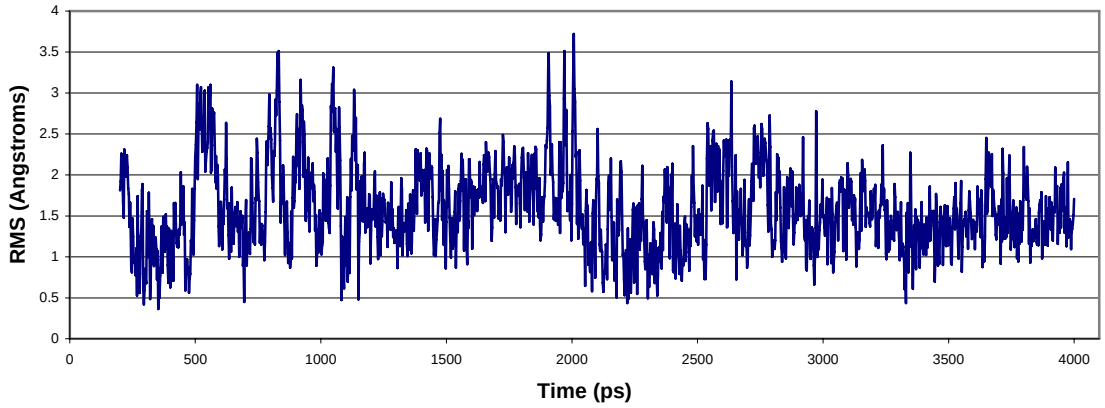
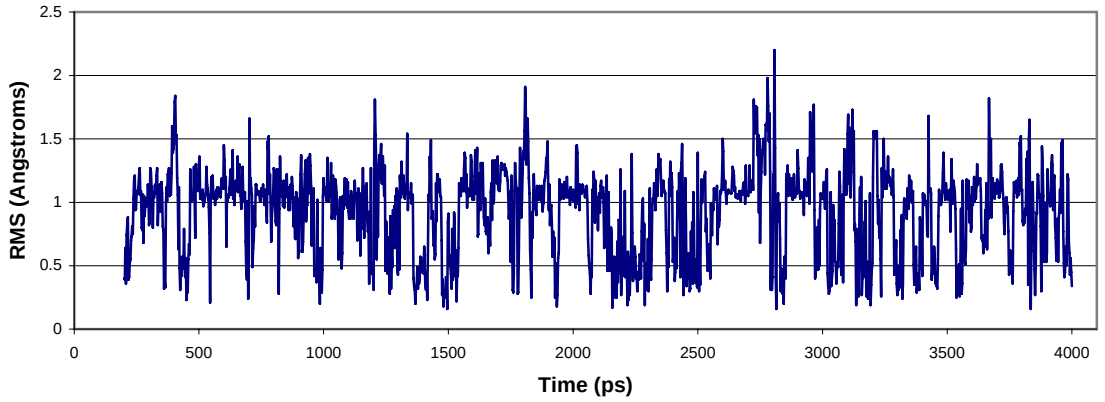


Figure 15. Graphs of Domain 2 of AA-GD. All graphs are the RMS between the crystal structure of 1AAY against the MD simulation data of AA-GD against time in picoseconds.

(a) Residue 74 Arginine (Domain 3 Position -1)



(b) Residue 77 Glutamate (Domain 3 Position 3)



(c) Residue 80 Arginine (Domain 3 Positon 6)

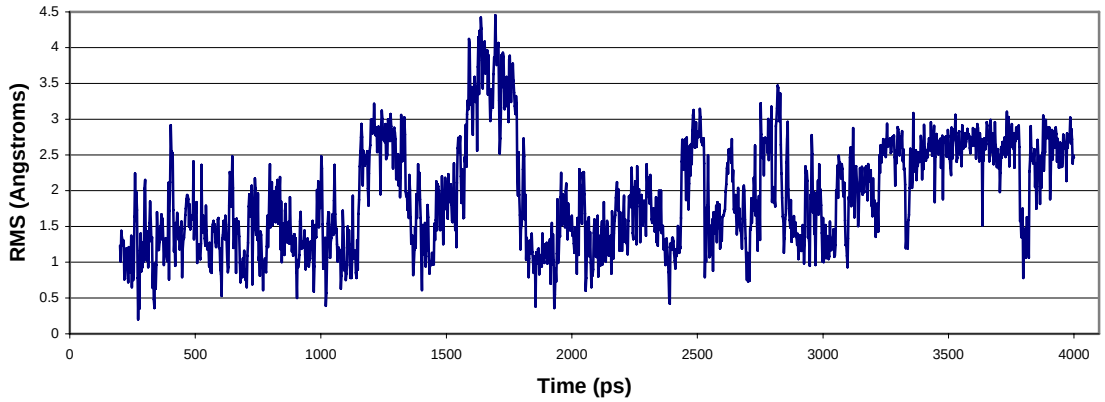


Figure 16. Graphs of Domain 3 of AA-GD. All graphs are the RMS between the crystal structure of 1AAY against the MD simulation data of AA-GD against time in picoseconds.

throughout the simulation is 1.83Å difference from the crystal structure as can be seen in Figure 17. Residue 17, an asparagine, fluctuates around an RMS of 1.2Å. The sixth position of the α-helix is occupied by a lysine residue, which is functionally similar to an arginine and a strong residue for binding to molecules with negative charges. Unlike other arginine residues that have been analyzed, the lysine here fluctuates between 1.0Å and 3.0Å difference from the crystal structure.

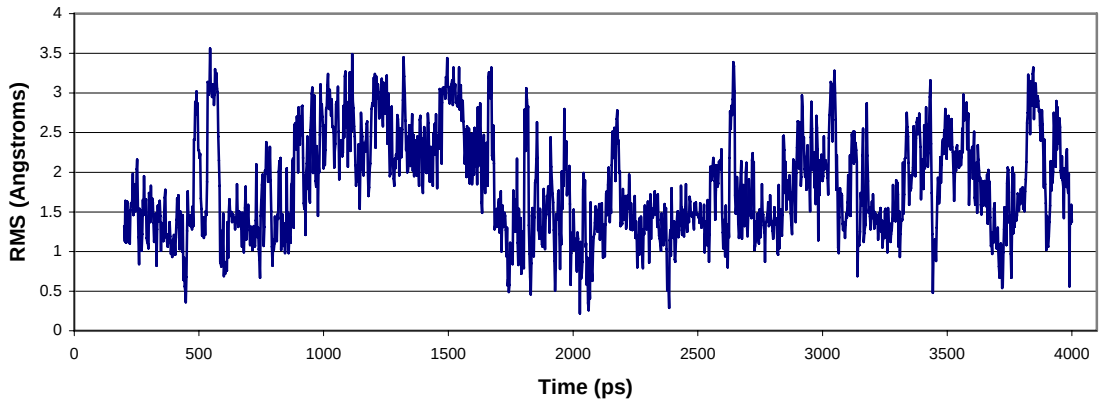
The second domain is interesting in the way it is constructed and thus how it behaves within the simulation. The negative one position is a neutrally charged glutamine which has a good deal of fluctuation, but is not uncommon when compared to other glutamine residues modeled and is seen in Figure 18. The third position of the α-helix is an aspartate, which is negatively charged. It fluctuates around the 1.0Å difference from the crystal structure. What is interesting is that the negative residue is generally positioned between two positively charged residues, as domains 1 and 3 of AA-GD exhibit, not between a neutral and positively charged residue. The lysine in position 6 of the α-helix resonates around 2.1Å difference from the crystal structure.

The third domain of ME-AA has a unique RMS profile and can be seen in Figure 19. The negative one position arginine exists primarily around 1.5Å away from the crystal structure, but also fluctuates to a much higher state above 3.0Å. The histidine residue in position 3 of the α-helix fluctuates randomly between two states, one approximately at 1.0Å difference from the crystal structure, and the second state around 2.25Å difference. The final residue of interest, an arginine in position 6 of the α-helix prefers a state that is over 4.0Å away from the crystal structure, but the RMS will periodically fall towards the observed state in the crystal.

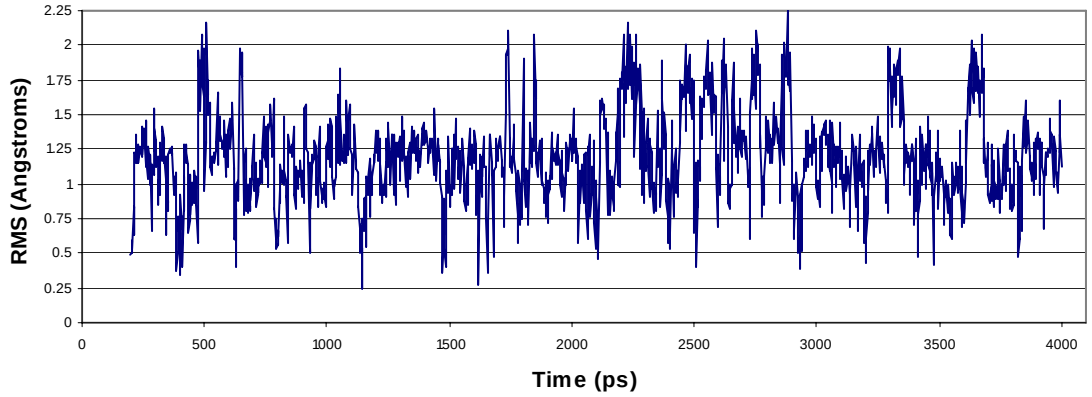
Clustering Results

Clustering the side chains is the final step in determining the overall structure of the homology modeled protein. Since we are trying to determine the structure that is closest to the

(a) Residue 16 Glutamine (Domain 1 Position -1)



(b) Residue 19 Asparagine (Domain 1 Position 3)



(c) Residue 22 Lysine (Domain 1 Position 6)

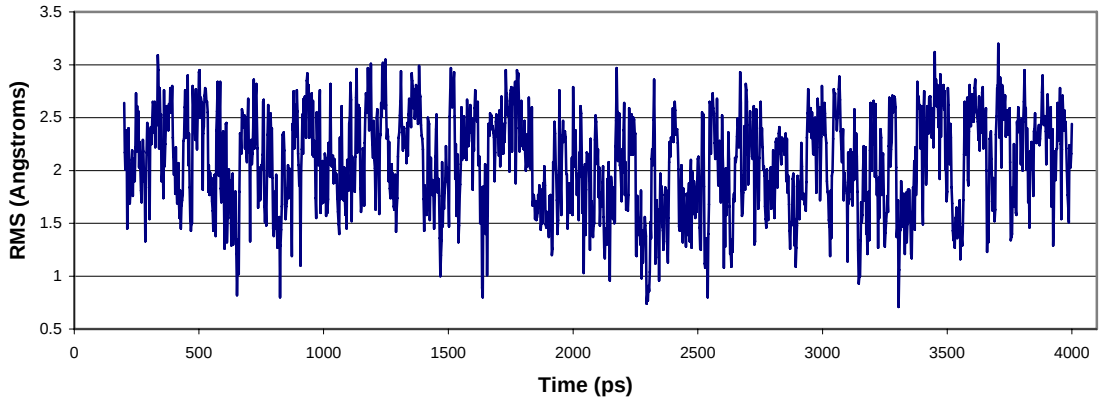
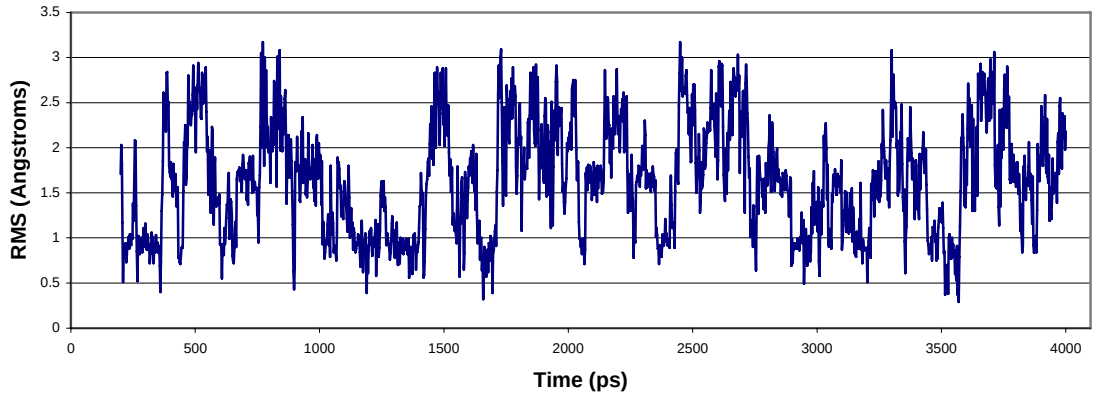
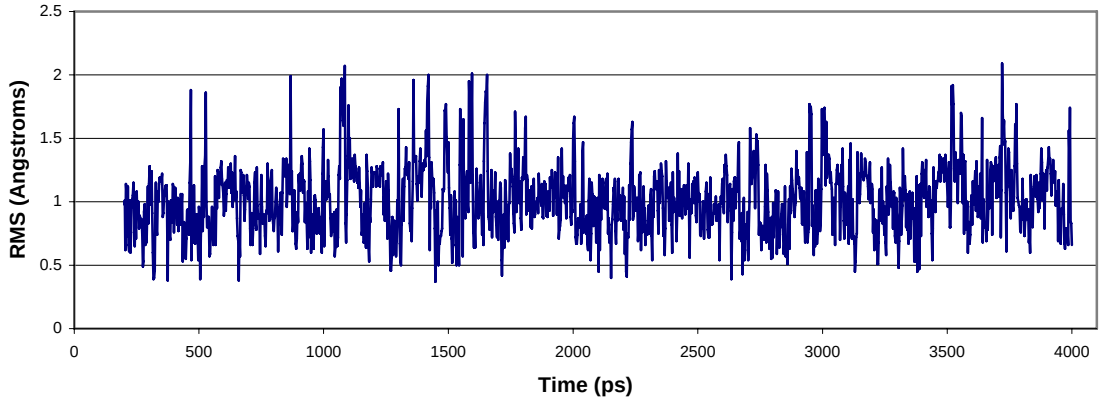


Figure 17. Graphs of Domain 1 of ME-AA. All graphs are the RMS between the crystal structure of 1MEY against the MD simulation data of ME-AA against time in picoseconds.

(a) Residue 44 Glutamine (Domain 2 Position -1)



(b) Residue 47 Aspartate (Domain 2 Position 3)



(c) Residue 50 Lysine (Domain 2 Position 6)

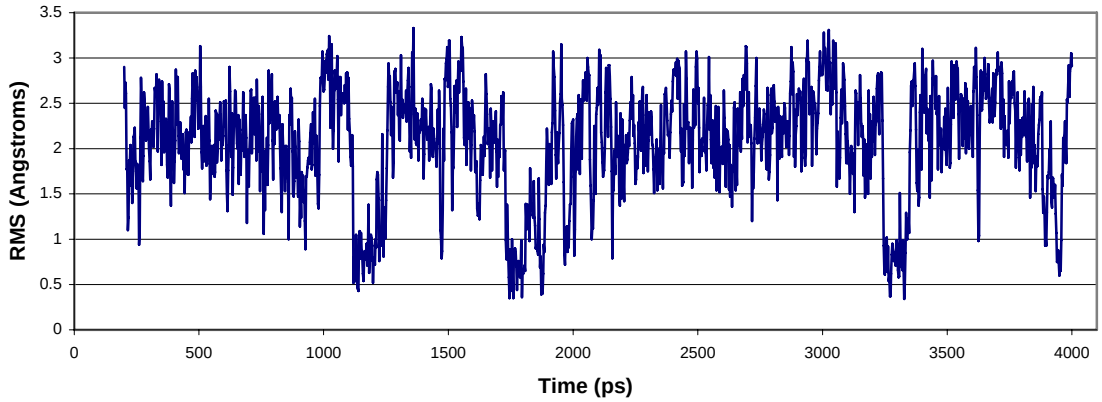
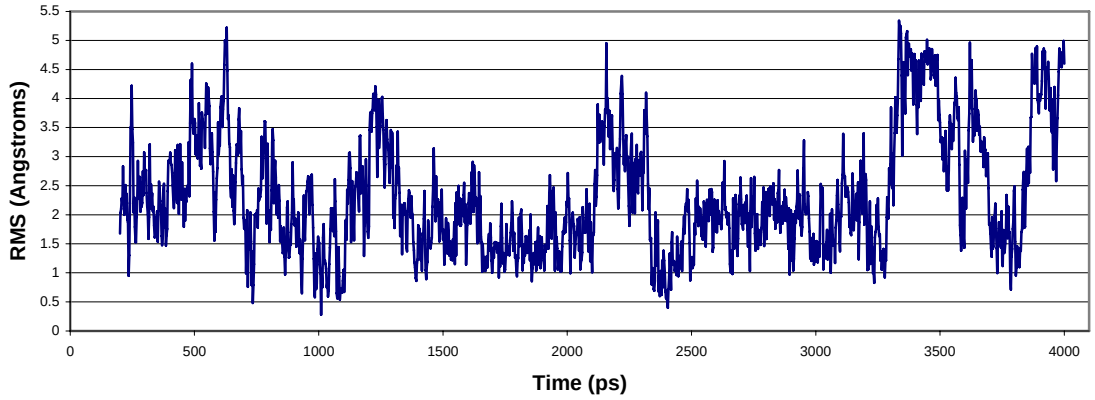
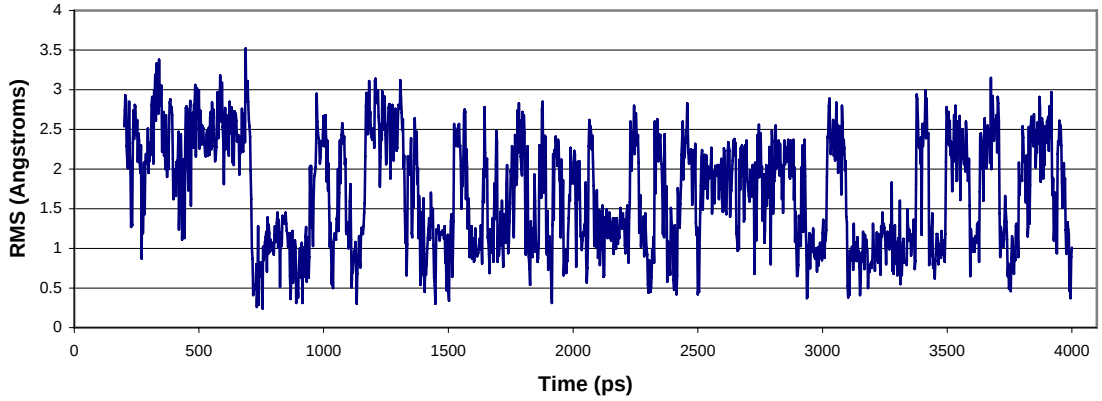


Figure 18. Graphs of Domain 2 of ME-AA. All graphs are the RMS between the crystal structure of 1MEY against the MD simulation data of ME-AA against time in picoseconds.

(a) Residue 72 Arginine (Domain 3 Position -1)



(b) Residue 75 Histidine (Domain 3 Position 3)



(c) Residue 78 Arginine (Domain 3 Position 6)

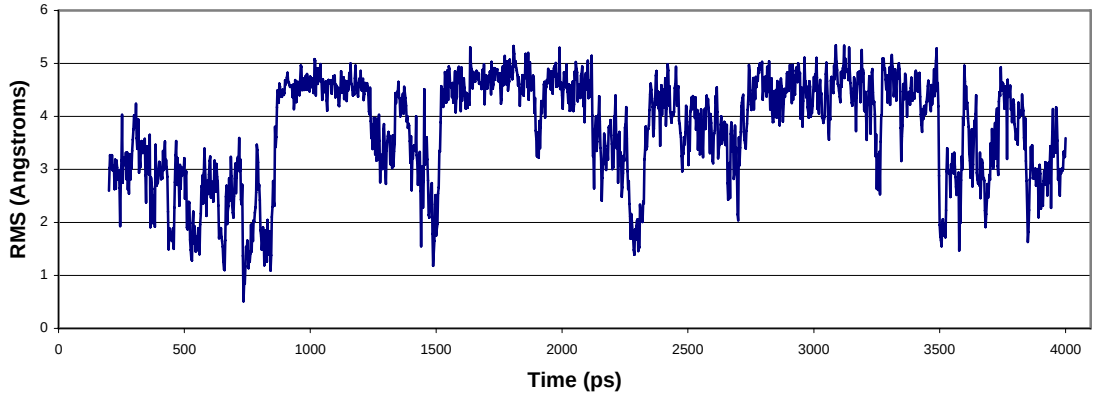


Figure 19. Graphs of Domain 3 for ME-AA. All graphs are the RMS between the crystal structure of 1MEY against the MD simulation data of ME-AA against time in picoseconds.

crystal structure, the model for a given amino acid with the highest amount of neighbors through RMS analysis within a 0.75Å radius is selected. The residues from the PDB files of various models are concatenated together. The reason this is possible is due to the fact that the backbone is held constant throughout the simulation so the residues, when concatenated, form one connected protein.

In creating one protein of three domains, the loop regions between domains are more difficult to reconstruct as there is duplication of residues between the different domains. Information on the individual domains for each homology model and the overlapping residues for each domain can be found in Appendix C. For purpose of simplicity, the first residue of overlap between the two domains is captured from the N-terminal domain and the second residue of overlap is captured from the C-terminal domain. For example, if the sequence for domain one is **...GQKP** and the sequence for domain two is **QKPF...** the glycine and glutamine from domain one and the lysine and phenylalanine from domain two are kept. By selecting residues in this manner any terminal residue from each domain is discarded. Additionally, the terminal residues generally have a high difference in the RMS from the crystal structure conferring additional support towards selecting non-terminal residues. Examining the completed structures uses seven residues of interest per domain: the three residues used for DNA binding, the two conserved cysteine residues and two conserved histidine residues.

The completed structure for GD-AA can be seen in Figure 20 as having fairly good conservation of the side chains from the crystal structure. As can be seen in Figure 20a, the residues of interest are almost identical in structure to the crystal. There is a rotation of the threonine in position 6 of the α-helix and the foreground cysteine (C1) is slightly rotated from the crystal, but still conforming to an architecture that is appropriate for binding the zinc ion. Most of the residues that are seen in a conformation differing from the crystal structure lie within the β-strands, which is expected as they are exposed to solvent. The second domain, Figure 20b has a

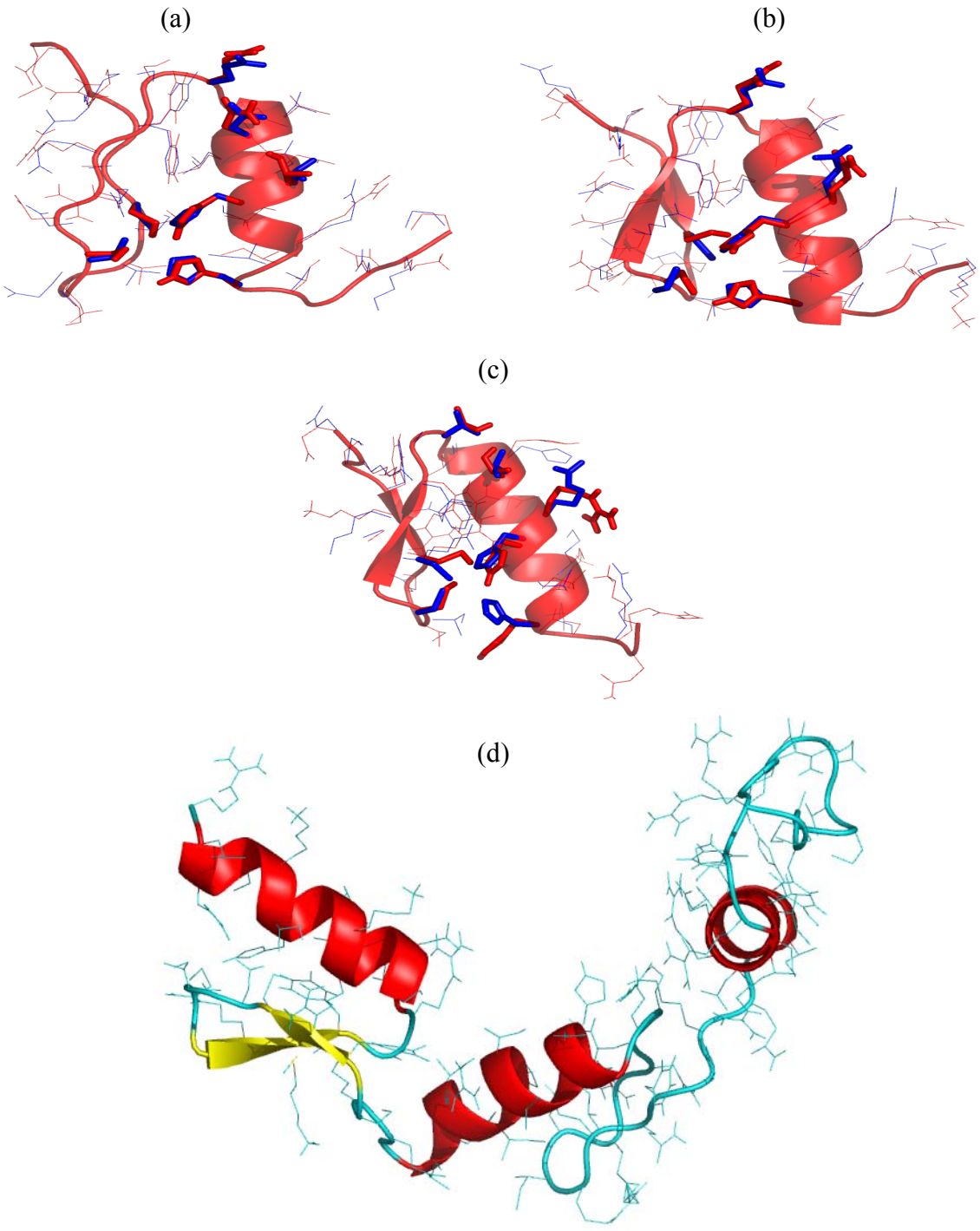


Figure 20. GD-AA Clustered Domain Results. The homology model is shown in red with the crystal structure superimposed in blue with the conserved residues and DNA binding residues in stick form for (a), (b), and (c). (a) Domain 1. (b) Domain 2. (c) Domain 3. (d) All domains for GD-AA concatenated together and colored by secondary structure: yellow is for a β-sheet, red for an α-helix, and cyan for the remaining residues. Domain 1 is on the left.

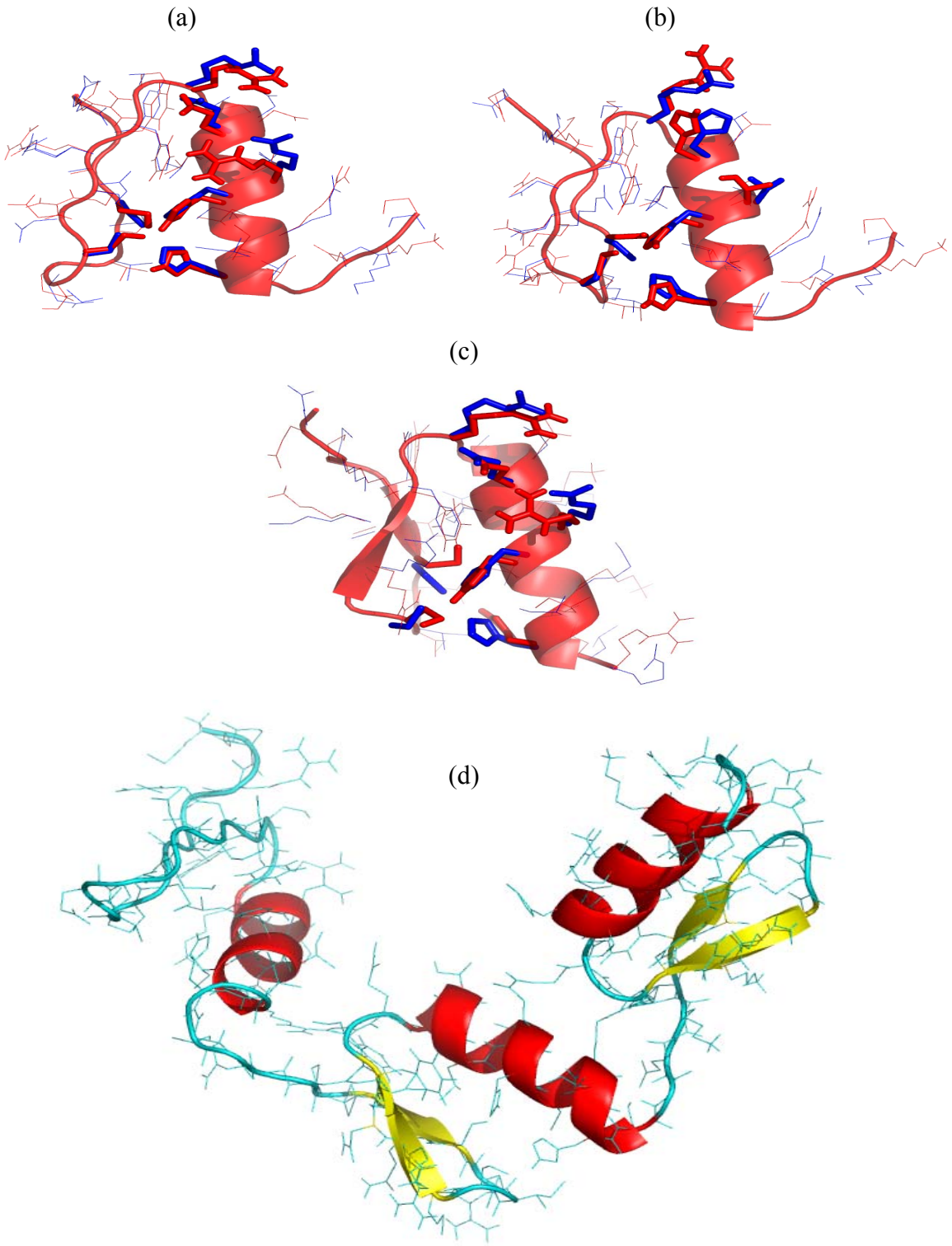


Figure 21. AA-GD Clustered Domain Results. The homology model is shown in red with the crystal structure superimposed in blue with the conserved residues and DNA binding residues in stick form for (a), (b), and (c). (a) Domain 1. (b) Domain 2. (c) Domain 3. (d) All domains for AA-GD concatenated together and colored by secondary structure: yellow is for a β-sheet, red for an α-helix, and cyan for the remaining residues. Domain 1 is on the left.

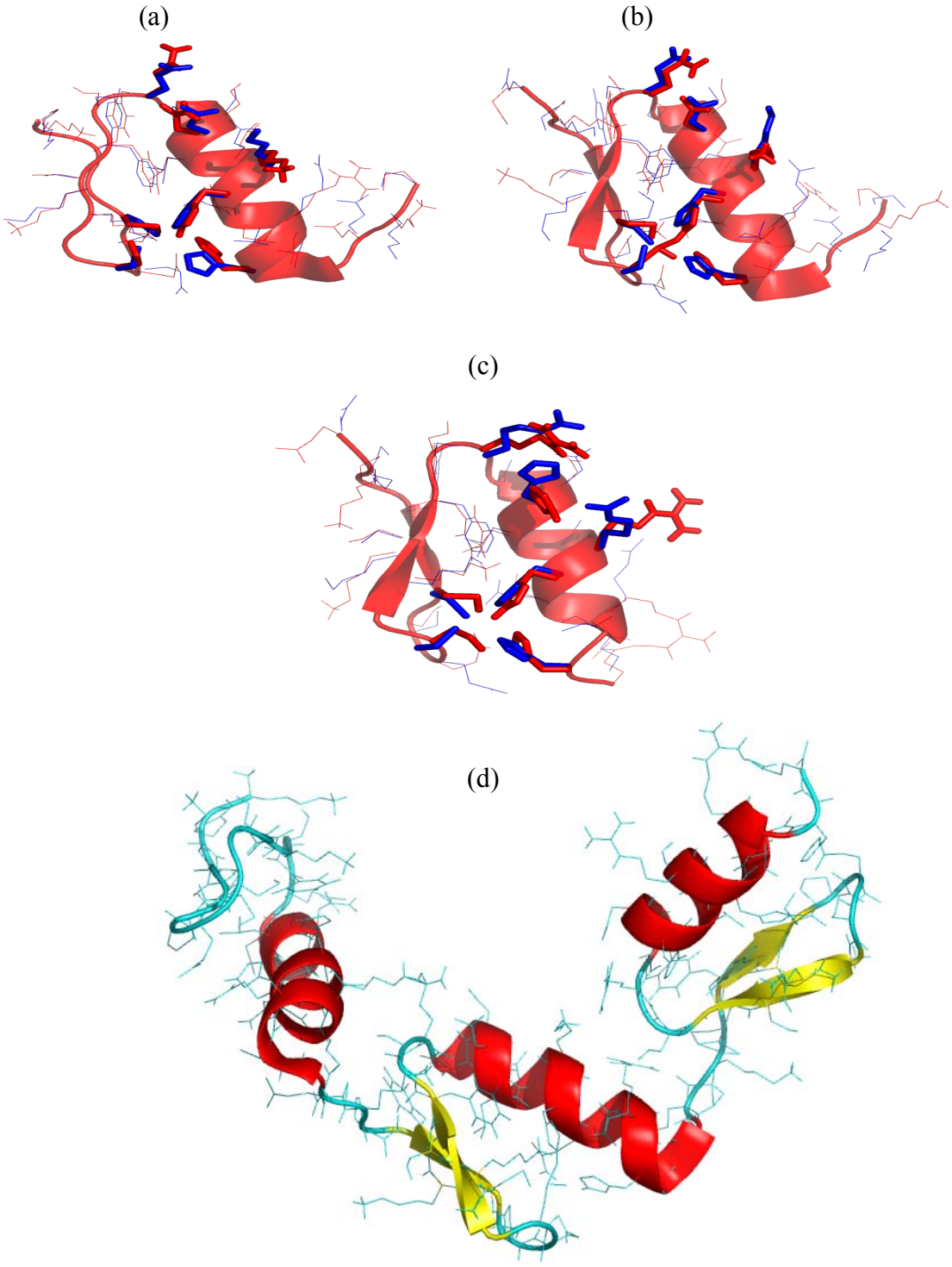


Figure 22. ME-AA Clustered Domain Results. The homology model is shown in red with the crystal structure superimposed in blue with the conserved residues and DNA binding residues in stick form for (a), (b), and (c). (a) Domain 1. (b) Domain 2. (c) Domain 3. (d) All domains for GD-AA concatenated together and colored by secondary structure: yellow is for a β-sheet, red for an α-helix, and cyan for the remaining residues. Domain 1 is on the left.

Table 1. Residues of Interest for GD-AA. Listed for each domain is the time in which the residue was found and the RMS difference from the crystal structure.

Domain 1			Domain 2			Domain 3		
Residue	Time (ps)	RMS	Residue	Time (ps)	RMS	Residue	Time (ps)	RMS
C1	1820	1.08	C1	2432.5	0.45	C1	485	0.95
C2	815	0.91	C2	325	1.26	C2	2010	0.49
H1	1320	1.23	H1	2105	0.64	H1	1072.5	0.99
H2	3302.5	1.18	H2	2885	1.24	H2	1665	2.80
Helix -1	1120	0.97	Helix -1	377.5	0.86	Helix -1	2927.5	0.39
Helix 3	2622.5	0.25	Helix 3	3382.5	1.14	Helix 3	2820	0.71
Helix 6	1352.5	0.84	Helix 6	1605	1.25	Helix 6	515	3.62

less conserved structure than the first domain. The DNA binding residues are highly similar to those in the crystal. The third position of the α-helix cannot be seen due to the fact that the residue is glycine and the side chain is only a hydrogen atom. The glutamine in the sixth position of the α-helix is rotated about the C_δ which puts the residue in a less desirable position to make contact with the DNA. The third domain, Figure 20c, is the least desirable clustering formation for the protein. Positions negative one and three of the α-helix are identical to the crystal, but the sixth position, an arginine, is rotated in the exact opposite direction of the crystal structure. The RMS between these two residues is 3.62Å, and can be seen at time point 1287.5ps in Figure 13c. This is also the only of the domains in this structure to such a conflict between the conserved residues and the crystal structure. The RMS and the time point for the residues of interest can be found in Table 1. Both the cysteine (C1) in the background and the histidine (H2) in the foreground are rotated out of position and thus would not bind to the zinc atom if present. The residues in the β-strands have much higher similarity to the crystal structure than the residues in either domain one or two. A full table of RMS versus the crystal structure and time point for each residue within each domain can be found in Appendix D. The complete model for GD-AA can be seen in Figure 20d having the β-strands only recognized in the first domain, which is the only domain that did not have the β-strands recognized after homology modeling (Figure 9a).

Table 2. Residues of Interest for AA-GD. Listed for each domain is the time in which the residue was found and the RMS difference from the crystal structure.

Domain 1			Domain 2			Domain 3		
Residue	Time (ps)	RMS	Residue	Time (ps)	RMS	Residue	Time (ps)	RMS
C1	1095	0.94	C1	3792.5	1.07	C1	2592.5	0.87
C2	280	2.05	C2	2437.5	0.49	C2	1410	0.84
H1	882.5	1.77	H1	3847.5	1.52	H1	1117.5	1.25
H2	2850	1.27	H2	2490	1.44	H2	655	2.84
Helix -1	3102.5	1.13	Helix -1	2970	2.3	Helix -1	3402.5	1.39
Helix 3	237.5	0.97	Helix 3	1087.5	1.27	Helix 3	2592.5	1.07
Helix 6	1585	2.18	Helix 6	837.5	0.67	Helix 6	1415	1.55

The completed structure for AA-GD is slightly more disorganized than the completed structure for GD-AA. The first domain, Figure 21a, does not show much similarity between the residues of interest and the crystal structure. Although the arginine (position negative one) and the glutamate (position three) in the α-helix are quite similar to the crystal structure, the arginine in the sixth position is removed from crystal and bent inward towards the middle of the domain. The RMS between this residue and the crystal is 2.18Å. Additionally, the conserved residues are not in favorable positions to coordinate a zinc ion. The cysteine residues are angled away from one another, and the second histidine is turned down away from the crystal. The RMS and time positions for the AA-GD can be found in Table 2. The second domain, Figure 21b, is more ordered than the first domain. The DNA binding residues in this domain oriented in a similar manner to the crystal structure with a slight difference in the positioning of the alpha carbon. The only residues with positions that could be of concern are the C1 and H2 residues, which are both angled in a different manner than the crystal structure but in relative position to continue to coordinate a zinc ion. The quality of the third domain lied between the first and second domains. The arginine (position negative one) and glutamate (position three) are in line with the crystal structure. Once again, the arginine in the sixth position is turned inward towards the middle of the domain, leaving it in a difficult, but not impossible position for binding to DNA. The RMS between the crystal and the clustered arginine residue is 1.55Å, which is low for this residue. In

Table 3. Residues of Interest for ME-AA. Listed for each domain is the time in which the residue was found and the RMS difference from the crystal structure.

Domain 1			Domain 2			Domain 3		
Residue	Time (ps)	RMS	Residue	Time (ps)	RMS	Residue	Time (ps)	RMS
C1	2385	0.5	C1	3605	0.64	C1	1707.5	0.6
C2	3412.5	1.25	C2	1627.5	0.62	C2	2287.5	0.98
H1	2847.5	0.49	H1	1412.5	0.67	H1	1265	0.93
H2	1170	2.3	H2	1665	2.22	H2	660	1.88
Helix -1	735	1.18	Helix -1	1320	0.71	Helix -1	2680	1.98
Helix 3	2420	1.11	Helix 3	3735	1.07	Helix 3	2262.5	2.08
Helix 6	3075	2.21	Helix 6	957.5	2.4	Helix 6	2005	4.51

addition, both cysteine residues are angled away from one another, and the H2 residue is once again angled downwards.

The completed structure for ME-AA has good structural preservation of the conserved residues as can be seen in Figure 22. The first domain, Figure 22a, is conserved in the residues of interest except for position six of the α-helix, a lysine residue. The lysine has a kink in the side chain near the beta carbon and is directed away from the supposed location of the DNA and has an RMS of 2.21Å against the crystal structure. The second domain is similar to the first domain, except that the lysine in the sixth position of the α-helix is rotated about the beta-carbon to produce a side chain that is directed away from the location of the DNA with an RMS difference of 1.34Å difference from the crystal. The third domain is slightly more chaotic in terms of the side chains of interest on the α-helix. All three residues on the α-helix are not aligned with the crystal structure. Detailed RMS descriptions can be found in Table 3. All three of the residues are directed away from where the DNA would be located within the structure. The conserved residues also have more fluctuation from the crystal structure that the other two domains, but are still oriented properly for zinc coordination.

Discussion

Protein Structure Quality

The structures of the final state of the clustered proteins are favorably aligned with the respective crystal structure. When looking at the whole DNA binding region of the protein, the final structures exhibit RMS values that are desirable for the purpose of protein – DNA interaction prediction (Table 4). The structures have much better RMS scores in between the backbone than the initial structures from homology modeling through Consensus. Given that the RMS scores between all atoms within the models is around 1.7Å for each completed structure, there is confidence that these structures are of sound quality, but that there is some difference between the structures. There may possibly be regions within the protein that are more dissimilar than other regions that can contribute to the increased RMS score. Suspect areas are the loop regions between the domains, the loop region between the β-strands as well as the N and C terminals, as the homology modeling process ranked those regions as not having confidence in the prediction. Differences can be visualized by referring back to Figure 9.

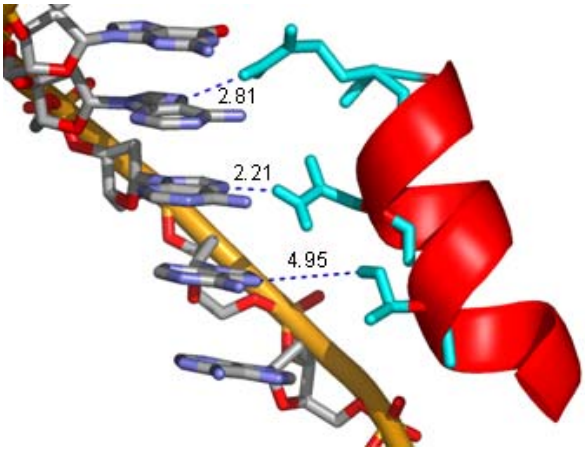
Protein – DNA Interaction

Although the completed proteins were not used in protein – DNA interaction prediction programs, it is still possible to observe potential interactions between the protein domains and the DNA present in the crystal structures. By superimposing the DNA from the crystal structure onto

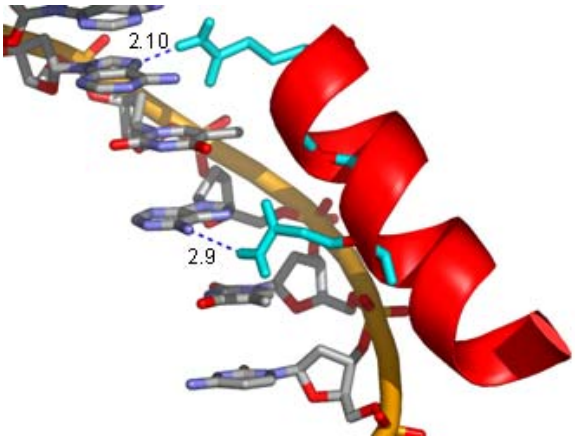
Table 4. RMS values between completed models and the crystal structure. Three different RMS calculations were made between a given crystal structure and completed model: between all atoms (amino acids), between C-α atoms, and between atoms in the backbone.

RMS	1G2D vs. GD-AA	1AAY vs. AA-GD	1MEY vs. ME-AA
Amino Acids	1.68 Å	1.75 Å	1.70 Å
C-α	0.52 Å	0.59 Å	0.60 Å
Backbone	0.57Å	0.63Å	0.66Å

(a)



(b)



(c)

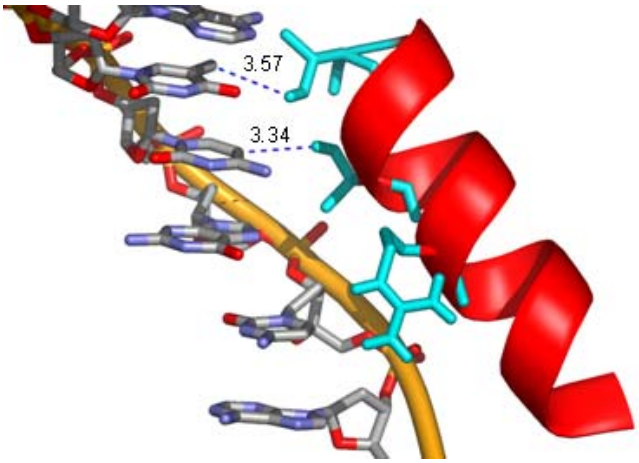
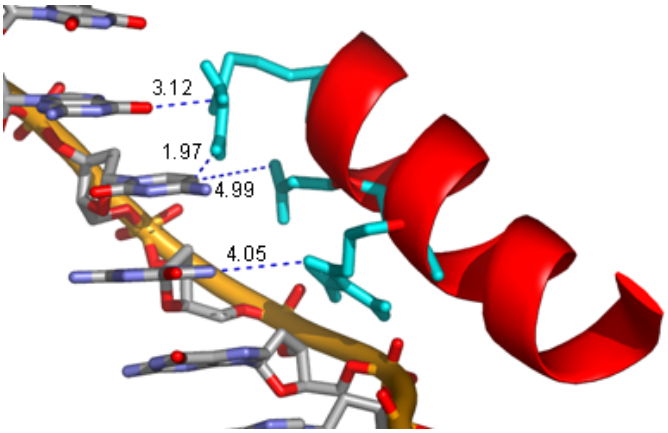


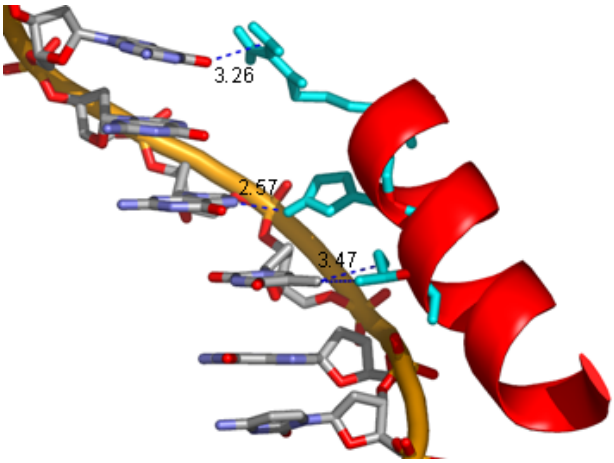
Figure 23. Potential Protein - DNA Interactions for GD-AA. The DNA helix from the crystal structure for 1G2D was superimposed upon the completed structure for GD-AA. For each figure the three residues in the α-helix are the -1, 3 and 6 positions from top to bottom. (a) Domain 1: Residues Q, N, and T. (b) Domain 2: Residues Q, G (not pictured), Q**.** (c) Domain 3: Residues T, T, R.

the completed structure of the homology model insight can be gained as to the probability of the protein making the necessary DNA interactions for binding. The completed structure for GD-AA is the best in terms of potential protein – DNA interactions (Figure 23). As a generalization, the negative one residue for all three domains has very strong interaction potential. Although the exact orientation is not ideal, the distance for interaction is in the scope needed for short range forces to take effect, less than 3 Å, in two of the three domains. The first domain was the only one of the three to have the key residues oriented in a

(a)



(b)



(c)

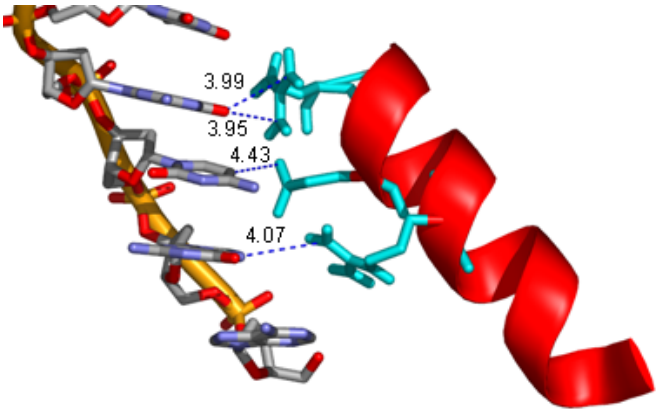


Figure 24. Potential Protein DNA Interactions for AA-GD. The DNA helix from the crystal structure for 1AAY was superimposed upon the completed structure for AA-GD. For each figure the three residues in the α-helix are the -1, 3 and 6 positions from top to bottom. (a) Domain 1: Residues R, E and R. (b) Domain 2: Residues R, H and T. (c) Domain 3: Residues R, E and R.

manner to create a possible interaction between the protein and the DNA. The second and third domains are each missing an interaction. Domain two has the second residue as a glycine, which will not have a side chain to make a viable interaction, and the arginine in the third domain is oriented away from the DNA and thus no interaction is possible in this state.

The structure for AA-GD has more possible interactions between the key residues and DNA than the previous structure of GD-AA. In this model, each domain has at least three possible interactions between the protein and DNA due to the orientation of the side chains (Figure

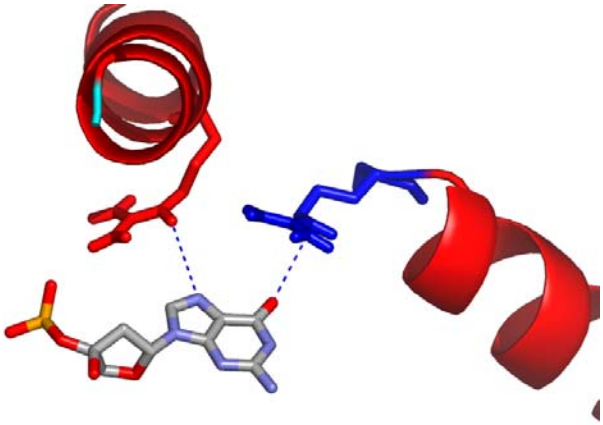
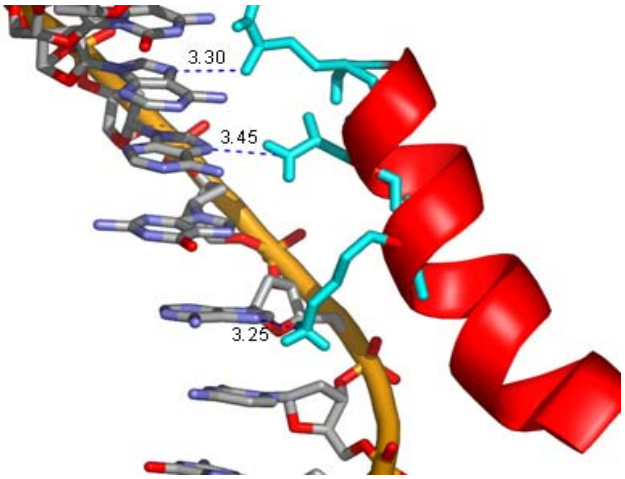


Figure 25. Favorable Coordination of a Guanine by Residues in Two Domains. The arginine from domain one is shown in red, and the arginine from domain two is shown in blue. Although both residues are coordinating the same nucleotide, a guanine, they are interacting with different areas of the nucleotide.

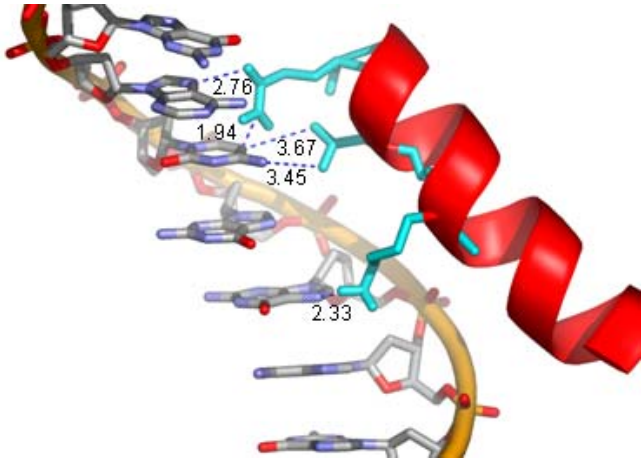
24). The first domain has four potential interactions, with the first residue, an arginine, potentially interacting with two different bases in the DNA. Each NH_2 group of the arginine could make contact with the DNA. Both of these interactions are more favorable than either of the two interactions for the third and sixth residues in the α-helix, which are no closer than 4 Å to the DNA. The second domain has each residue making contact with only one base, but the arginine in the negative one position is not interacting with the expected nucleotide. The arginine is actually interacting with the guanine that the arginine in the sixth position of the α-helix in domain one is interacting, but on different atoms. Interestingly, the two residues do not conflict with one another as would be expected since the two domains were modeled independent of one another (Figure 25). Both the histidine and threonine are in favorable orientations to make the proper interactions. While there is potential for making correct contact with the DNA for domain three, the distance of the side chains is not close enough to create hydrogen bonds or short range forces.

The model of ME-AA has some interesting side chain conformations that complicate the binding of DNA. ME-AA is the only structure that contains lysine residues in the key residues for binding to DNA. All three domains for position six of the α-helix do not have a desirable conformation for binding to DNA in the proper and experimentally defined manner (Figure 26). For positions negative one and three in the α-helix in domain one, the orientation of the residues allows for potential interaction between the protein and DNA. The lysine in position six is

(a)



(b)



(c)

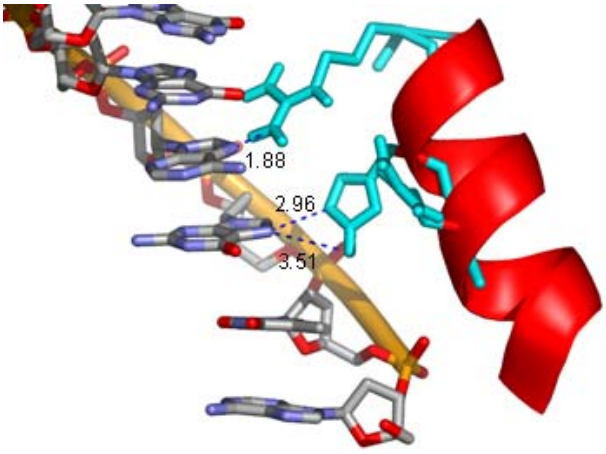


Figure 26. Potential Protein DNA Interactions for ME-AA. The DNA helix from the crystal structure for 1MEY was superimposed upon the completed structure for ME-AA. For each figure the three residues in the α-helix are the -1, 3 and 6 positions from top to bottom. (a) Domain 1: Residues Q, N and K. (b) Domain 2: Residues Q, D and K. (c) Domain 3: Residues R, H and R.

skipping a nucleotide and interacting with the next nucleotide, an adenine, which is also being coordinated by the glutamine in position negative one in domain two. The interactions on the adenine nucleotide are unfavorable due to the fact that the two arginine residues are attempting to occupy the same space (Figure 27a). The aspartate residue in domain two can have two possible interactions with same nucleotide, both in an advantageous orientation. Once again, the lysine residue in domain two is positioned so that it is skipping a nucleotide, however the way that domain three is

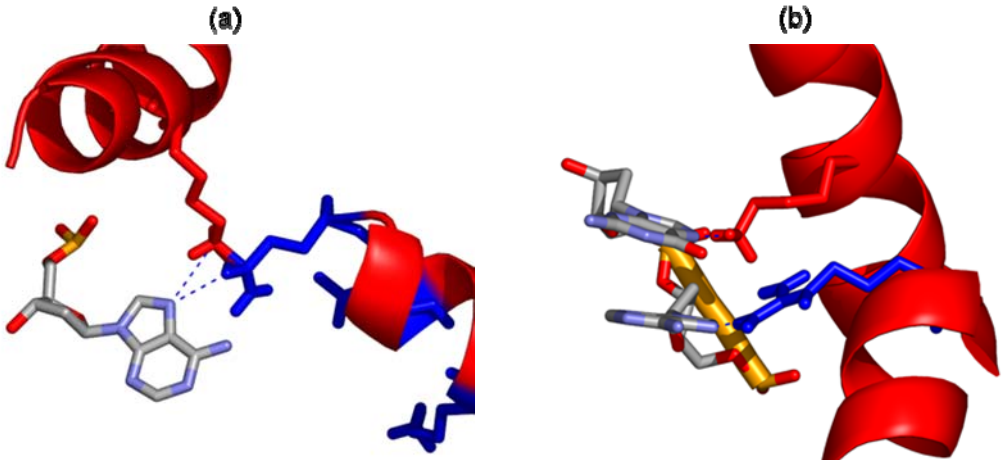


Figure 27. Favorable and Unfavorable DNA Interactions within ME-AA. (a) Unfavorable DNA interactions between a lysine (red) in domain one and a glutamine (blue) in domain two. Both residues are attempting to interact with the same atom in the adenine and are occupying the same space. (b) Favorable interactions with a lysine (red) in domain two and an arginine (blue) in domain three.

positioned allows for there to be no conflict in the interactions between the arginine in the negative one position and the DNA (Figure 27b). The histidine in position three of domain three is not properly oriented to make a strong interaction with the DNA, but is within proximity to have both short and long range interactions. The arginine in position six is not in a desirable conformation for binding to the DNA.

Through close examination of the individual amino acids with respect to the nucleotides available, one can deduce that these structures are of sufficient quality to be able to perform protein – DNA interaction prediction algorithms. Some of the residues discussed will need to be further examined in terms of the clustering results to find a better orientation, but to do so in an automated fashion would require the need for a robust algorithm for scoring clusters. Additionally, the DNA shown in these examples has been slightly distorted due to the fact that it originated from a crystal structure with a bound EGR protein, thus precise measurements are biased towards interactions occurring after the protein has bound to the DNA.

Protein – DNA Proposed Binding Mechanism

The residues of interest from analysis against the crystal structure shed light into a possible mechanism for EGR proteins binding to DNA. An EGR protein will not bind to DNA in the absence of zinc ions (Huang and Adamson 1993). Until the point of binding, the structure of the protein is able to fluctuate and is seen throughout the MD simulation. Through analysis of the simulations, certain residues have sparked heightened interest, specifically position six in the alpha helix. Generally occupied by an arginine or lysine, position six is the only residue in the helix that consistently has a highly sampled state that is removed from the state of the crystal structure. Additionally, the negative one position of the α-helix appears to fluctuate between two states throughout the simulation while taking time to sample the available space in between. The third position of the α-helix is the one residue that tends to reach an equilibrium state and is generally an amino acid with a smaller side chain.

It is possible that the information gathered on the states of the positions of the individual side chains could be related to the binding mechanism of EGR proteins, which is still yet to be empirically determined. One possible mechanism that comes to light from this research is that the third position of the α-helix remains steady, while the negative one position of the α-helix samples the conformational space. When the negative one position residue finds a desirable ligand it would rotate to a position to bring the third position of the α-helix closer to the DNA and increase the binding potential. After that point, the sixth position of the helix would come into play and reach out to make the final interaction between the protein and the ligand, in this case double stranded DNA.

There are many possible sources that could undermine the binding mechanism hypothesis. The current method for analyzing the predicted structure of homology models is through RMS analysis, which has been criticized as being a measure not specific enough to yield in-depth information about the differences in structure. Referring back to the graphs in Figures 11 – 19, calculating the RMS is sufficient to tell how much difference there is between two

different states, but is not specific enough to be able to cluster the data on this information alone. Because the comparison in these graphs is against a steady state can there be information gained. RMS analysis is also used in the cluster determination for the completed structures of the proteins. Here each time step is compared to every other time step for a given residue, and the time step with the highest amount of neighbors is reported and used in the structure of the completed protein. Since RMS is not being used in a conventional hierarchical or non-hierarchical clustering method, some of the reservations in using RMS as a means of determining structural information can be circumvented.

Another source of concern stems from the neighbor clustering method instantiated in finding the best structure to make the completed structure. Although the neighbor clustering algorithm reports all possible clusters, there is no straightforward way to identify markedly different clusters of a given residue. It is possible to manually search through the high ranking clusters to find disparate clusters, but that is neither time efficient or able to be automated. It is not guaranteed that the second or third highest ranking cluster is any different than the highest ranking cluster, since clusters are not distinct, non-overlapping sets. This problem presents itself when looking at residues with a high amount of clusters accommodating less than a majority of the time steps but receiving similar amounts of neighbors.

There are different approaches that could have been taken for clustering the side chains of the amino acids. One way would have been to analyze the side chains against a rotamer library using RMS, and then using the rotamer with the highest amount of similar structures in the final structure reported (Kimura et al. 2001). Another method would be to find the centroid for each side chain in a given residue and use hierarchical clustering methods in three-dimensional space to find the largest cluster within a given radius. The disadvantage of this clustering method would be that valuable information about the side chain and the direction that it is facing would be lost due to reduction of dimensionality. To alleviate the hindrance in the previous method, one could find the centroid of the last carbon in the side chain and any atoms extending beyond that

point and perform hierarchical clustering. The advantage to this method would be that the subject that is of highest concern, the location of the tip of the side chain, would be forefront consideration of this method. Ultimately, the atoms of the side chain would be neglected, but are of little concern since they are not interacting with the DNA.

Future Research

There are many areas of this project that can be further expanded in addition to ideas previously mentioned. There are two other areas for improvement within the current method for predicting the structure of EGR proteins. The first improvement would be to make the method completely automated so that any user interested in EGR proteins would be able to create their own homology model and perform MD simulations. Automating the procedure would be labor intensive as there are steps in the method that at this point in time require human interaction and require making knowledge-based decisions. The second major area for improvement would be to develop a scoring function for addressing the quality of a homology model, and the completed structure after the simulation and clustering is complete. A scoring algorithm would also be beneficial in the clustering process, potentially ranking rotamers seen throughout the simulation. Residues with a more energetically favorable state would appear higher in the list of clustered residues, thus not relying solely on the RMS calculation between structures.

The next step for the completed predicted structure for zinc finger proteins would involve the use of a derivative of the ClusPro program, initially developed by Comeau *et al*. for predicting interactions between proteins (2004). ClusPro uses two unbound structures and performs automated rigid body docking, and ranks the results based upon clustering of the resolved conformations on the evaluation of the properties of free energy, electrostatic and desolvation calculations. Since this method uses two unbound structures it can also be applied to predict protein – DNA interactions, which is under development in the Camacho laboratory. The initial clustered result for GD-AA with a 0.75Å water layer was tested to see if protein – DNA

interactions could be predicted. The structures for the protein domains were predicted to a degree that they were able to pick up the experimentally proven interactions between 1G2D and the given DNA sequence.

Aside from predicting protein – DNA interactions, there are many more experiments that can be performed using MD simulations. Since we have MD simulation results for the homology modeled proteins without the presence of ions or DNA, the next step would be to add these incrementally to see how the protein reacts to the new conditions. One starting point would be to add a zinc ion into each of the zinc finger domains. It would also be advantageous to simulate the protein in the presence of DNA or a negatively charged object representing DNA both with and without zinc ion to see how the protein thus reacts. Another interesting simulation would be with all three domains together, and allow for the backbone to be unconstrained in the loop regions between the zinc finger domains.

Conclusion

By exploring the EGR family of transcription factors there have been many insights gained into structure prediction for this specific family. Since the EGR family is one of the largest families of transcription factors, the method developed can easily be used for any of the proteins within the family. Using molecular dynamic simulations on singular domains yields a final clustered protein with side chains that have near-native conformations. By analyzing the results and superimposing a DNA helix one can infer information about potential protein – DNA interactions before performing actual interaction algorithms. Analyzing the completed domains before using interaction algorithms can help determine if there are residues that should be further explored. By looking at the results, and the analysis against the crystal structure, a hypothesis for EGR proteins binding to DNA can be surmised: the negative one position of the α-helix is responsible for initial contact with the DNA, and the other key residues have interactions that are initiated after initial binding.

References

Altschul, S.F.; Madden, T.L.; Schäffer A.A.; Zhang J; Zhang Z.; Miller W; Lipman D.J., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.* 1997, **25**, 3389-3402.

Bateman, A.; Coin, L.; Durbin, R.; Finn, R.D.; Hollich, V.; Griffiths-Jones, S.; Khanna, A.; Marshall, M.; Moxon, S.; Sonnhammer, E.L.L.; Studholme, D.J.; Yeats, C.; Eddy, S.R., "The Pfam Protein Families Database", *Nucleic Acids Res. Database Issue* 2004, **32**, D138-D141.

Benos, P.V.; Lapedes, A.S.; Stormo, G.D., "Probabilistic Code for DNA Recognition by Proteins of the EGR Family", *J. Mol. Biol.* 2002, **323**, 701-727.

Brooks, B.R.; Bruccoleri, R.E.; Olafson, B.D.; States, D.J.; Swaminathan, S.; Karplus, M., "CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamic Calculations", *J. Comput. Chem.* 1983, **4**, 187-217.

Camacho, C.J., "Modeling Side-Chains Using Molecular Dynamics Improve Recognition of Binding Region in CAPRI Targets", *Proteins* 2005, **60**, 245-251.

Christy, B.; Nathans, D., "DNA Binding Site of the Growth Factor-Inducible Protein Zif268", *Proc. Natl. Acad. Sci. USA.* 1989, **86**, 8737-8741.

Comeau, S.R.; Gatchell, D.W.; Vajda, S.; Camacho, C.J., "ClusPro: an automated docking and discrimination method for the prediction of protein complexes", *Bioinformatics* 2004, **20**, 45-50.

Davis, S.; Bozon, B.; Laroche, S., "How Necessary is the Activiation of the Immediate Early Gene *zif268* in Synaptic Plasticity and Learning?", *Behav. Brain Res.*2003, **142**, 17-30.

Elrod-Erickson, M.; Rould, M.A.; Nekludova, L.; Pabo, C.O., "Zif268 Protein-DNA Complex Refined at 1.6 A: A Model System for Understanding Zinc Finger – DNA Interactions", *Structure*, 1996, **4**, 1171-1180.

Gashler, A.; Sukhatme, V.P., "Early Growth Response Protein 1 (Egr-1): Prototype of a Zinc-finger Family of Transcription Factors", *Prog. Nucleic Acid Res. Mol.Biol.*, 1995, **50**, 191-224.

Huang, R.P.; Adamson, E.D., "Characterization of the DNA-binding Properties of the Early Growth Response-1 (Egr-1) Transcription Factor: Evidence for Modulation by a Redox Mechanism", *DNA Cell Biol.* 1993, **12(3)**, 265-273.

Hulo, N.; Sigrist, C.J.A.; Le Saux, V.; Langendijk-Genevaux, P.S.; Bordoli, L.; Gattiker, A.; De Castro, E.; Bucher, P.; Bairoch, A., "Recent improvements to the PROSITE database", *Nucleic Acids Res.* 2004, **32**, 134-137.

Kim, C.A.; Berg, J.M., "A 2.2 A Resolution Crystal Structure of a Designed Zinc Finger Bound to DNA", *Nat. Struct. Biol.*, 1996, **3**, 940-945.

Kimura, S.R; Brower, R.C; Vajda, S; Camacho, C.J., "Dynamical view of the Positions of Key Side Chains in Protein – Protein Recognition", *Biophys J.* 2001, **80**, 635-642.

Klug, A; Rhodes, D., " 'Zinc Fingers': A Novel Protein Motif for Nucleic Acid Recognition", *Trends Biochem. Sci.*1987, **12**, 464 – 469.

Kozakov, D.; Clodfelter, K.H.; Vajda, S.; Camacho, C.J., "Optimal Clustering for Detecting Near-Native Conformation in Protein Docking", *Biophys. J.* 2005, **89**, 867-875.

Lindahl, E.; Hess, B.; Van Der Spoel, D., "GROMACS 3.0: A Package for Molecular Simulation and Trajectory Analysis", *J. Mol. Model.* 2001, **7**, 306-317.

Paillard, G.; Deremble, C.; Lavery, R., "Looking into DNA recognition: zinc finger binding specificity", *Nucleic Acids Res.* 2004, **32**, 6673-6682.

Prasad, J.C.; Comeau, S.R.; Vajda, S.; and Camacho, C.J., "Consensus alignment for reliable framework prediction in homology modeling", *Bioinformatics*, 2003, **19,** 1682-1691.

Sukhatme, V.P.; Cao, X.; Chang, L.C.; Tsai-Morris, C.; Stamenkovich, D.; Ferreira, P.C.P; Cohen, D.R.; Edwards, S.A.; Shows, T.B.; Curran, T.; Le Beau, M.M.; Adamson, E.D., "A Zinc Finger-Encoding Gene Coregulated with *c-fos* during Growth and Differentiation, and after Cellular Depolarization", *Cell* 1988, **53**, 37-43.

Thompson, J.D.; Higgins, D.G.; Gibson, T.J., "CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-specific Gap Penalties and Weight Matrix Choice", *Nucleic Acids Res.* 1994, **22**, 4673 – 80.

Wolf, S.A.; Grant, R.A., Elrod-Erickson, M.; Pabo, C.O., "Beyond the 'Recognition Code': Structures of Two Cys2Hys2 Zinc Finger / TATA Box Complexes", *Structure* 2001, **9**, 717-723.

Appendix

Appendix A – Multiple Sequence Alignment of EGR Protein Domains

```
Q69BL7_HUMAN_146-168        FQCNQ--CGASFTQKGN--LLRH--IKLH---
Q69BM3_HUMAN_112-134        FQCNQ--CGASFTQKGN--LLRH--IKLH---
Q69BM4_HUMAN_145-167        FQCNQ--CGASFTQKGN--LLRH--IKLH---
AIOL_HUMAN_146-168          FQCNQ--CGASFTQKGN--LLRH--IKLH---
IKAR_HUMAN_145-167          FQCNQ--CGASFTQKGN--LLRH--IKLH---
Q6PQC6_HUMAN_114-136        FHCNQ--CGASFTQKGN--LLRH--IKLH---
Q6PQD0_HUMAN_114-136        FHCNQ--CGASFTQKGN--LLRH--IKLH---
Q6PQC8_HUMAN_140-162        FHCNQ--CGASFTQKGN--LLRH--IKLH---
Q96JP3_HUMAN_147-169        FHCNQ--CGASFTQKGN--LLRH--IKLH---
Q9H2S9_HUMAN_85-107         FHCNQ--CGASFTQKGN--LLRH--IKLH---
HELI_HUMAN_140-162          FHCNQ--CGASFTQKGN--LLRH--IKLH---
Q6ZMZ8_HUMAN_69-92          FHCDQ--CSYSCKRKDN--LNLH--KKLKH--
Q5VWB8_HUMAN_446-469        FKCPS--CPLLFAQKRT--MLEH--LKNTH--
Q9H937_HUMAN_446-469        FKCPS--CPLLFAQKRT--MLEH--LKNTH--
Q68DQ8_HUMAN_858-881        FKCPS--CPLLFAQKRT--MLEH--LKNTH--
ZN592_HUMAN_892-915         FKCPE--CPLLFVQKPE--LMQH--VKSTH--
Q6MZM6_HUMAN_434-457        TICKI--CELSFETEHV--LLQH--MKDNH--
Q6N043_HUMAN_388-411        TICKI--CELSFETEHV--LLQH--MKDNH--
Q9H740_HUMAN_224-247        TICKI--CELSFETEHV--LLQH--MKDNH--
Q9HCI8_HUMAN_404-427        TICKI--CELSFETEHV--LLQH--MKDNH--
Q6N085_HUMAN_388-411        TICKI--CELSFETEHV--LLQH--MKDNH--
Q9NXS0_HUMAN_92-115         TICKI--CELSFETEHV--LLQH--MKDNH--
PRDM6_HUMAN_570-595         FKCER--CERSFTQATQ--LSRH--QRMPNEC
Q8NEI5_HUMAN_266-288        FKCHL--CDRCFGQQTN--LDRH--LKKH---
Q5T967_HUMAN_201-223        FSCSL--CSYTFAYRTQ--LERH--MTSH---
ZN469_HUMAN_237-259         FACSS--CNYTFAKKEQ--FDRH--MNKH---
Q86X64_HUMAN_2-24           FKCTV--CDYTAAQKPQ--LLRH--MEQH---
Q6ZN18_HUMAN_112-136        FKCVVGGCNASFASQGG--LARH--VPTH---
Q96BG3_HUMAN_120-144        FKCVVGGCNASFASQGG--LARH--VPTH---
Q6Y8G5_HUMAN_121-143        FVC--NICGRAFTTKGN--LKVH--YMTH---
Q6Y8G6_HUMAN_461-483        FVC--NICGRAFTTKGN--LKVH--YMTH---
FOG1_HUMAN_320-342          FVC--LICLSAFTTKAN--CERH--LKVH---
O95276_HUMAN_92-114         FKC--LICNGAFTTGGS--LRRH--MGIH---
Q9NV14_HUMAN_160-185        FKCPALGCVRIFKRIGF--LNKHA-MTVH---
FOG2_HUMAN_363-385          FRCN--HCHFGFQTQRE--LLQHQ--ELH---
Q96IQ9_HUMAN_145-169        FRCSALSCTETFPSMQE--LVAHS--KLH---
Q9H0G6_HUMAN_31-55          FRCK--DCSAAYDTLVE--LTVHMNETGH---
Q9P254_HUMAN_212-236        FRCK--DCSAAYDTLVE--LTVHMNETGH---
Q63HK5_HUMAN_214-238        FRCK--DCSAAYDTLVE--LTVHMNETGH---
Q6N003_HUMAN_215-239        FRCR--QCSATYDTLVE--LTVHMNETGH---
Q8N260_HUMAN_215-239        FRCR--QCSAAYDTLVE--LTVHMNETGH---
Q8NDK8_HUMAN_284-306        FVCK--NCSQMFYTEKG--LSSHM--CFH---
Q9H0D2_HUMAN_284-306        FVCK--NCSQMFYTEKG--LSSHM--CFH---
Q7M4M3_HUMAN_77-99          FRCR--NCDEGFSYNYQ--LHTHM--SIH---
Q5R2W1_HUMAN_736-758        FICH--VCNKAFRTNFR--LWSHF--QSH---
Q5R2W1_HUMAN_842-864        FMCK--LCHRTFKTAFS--LWSHE--QTH---
Q6N003_HUMAN_926-948        FYCS--DCASQFRTPST--YISHL--ESH---
Q8N260_HUMAN_926-948        FYCS--DCASQFRTPST--YISHL--ESH---
Q69BM1_HUMAN_337-361        FECN--MCGYRSHDRYE--FSSHIARGEH---
Q69BM3_HUMAN_407-431        FECN--MCGYRSHDRYE--FSSHIARGEH---
Q9C0D4_HUMAN_982-1004       FKCW--FCGRLYEDQEE--WMSHGQR--H---
Q86X64_HUMAN_77-100         LKCP--VCSFVYGTKWE--FNRHLKN-KH---
Q9C0D4_HUMAN_108-130        FICS--HCSYISYTKGE--FQRHLV--KH---
                              *    *             :
```

```
Q9NWV2_HUMAN_81-103          FKCS--ICQRHFKNLKT--FVKHQQ--LH---
AIOL_HUMAN_174-196           FKC--HLCNYACQRRDA--LTGH--LRTH---
Q69BL7_HUMAN_174-196         FKC--HLCNYACQRRDA--LTGH--LRTH---
Q69BM3_HUMAN_140-162         FKC--HLCNYACQRRDA--LTGH--LRTH---
IKAR_HUMAN_173-195           FKC--HLCNYACRRRDA--LTGH--LRTH---
Q69BM4_HUMAN_173-195         FKC--HLCNYACRRRDA--LTGH--LRTH---
HELI_HUMAN_168-190           FKC--PFCSYACRRRDA--LTGH--LRTH---
Q6PQC6_HUMAN_142-164         FKC--PFCSYACRRRDA--LTGH--LRTH---
Q6PQC8_HUMAN_168-190         FKC--PFCSYACRRRDA--LTGH--LRTH---
Q6PQD0_HUMAN_142-164         FKC--PFCSYACRRRDA--LTGH--LRTH---
Q96JP3_HUMAN_175-197         FKC--PFCNYACRRRDA--LTGH--LRTH---
Q9H2S9_HUMAN_113-135         FKC--PFCNYACRRRDA--LTGH--LRTH---
Q71UL5_HUMAN_4-26            FEC--KLCHQRSRDYSA--MIKH--LRTH---
Q9Y2Y4_HUMAN_401-423         FSC--SLCPQRSRDFSA--MTKH--LRTH---
Q6ZMZ9_HUMAN_134-156         FEC--KVCGQAFRQRSA--LTVH--KQCH---
Q6ZTI2_HUMAN_93-117          FRCEFEGCERRFANSSD--RKKH--SHVH---
ZIC4_HUMAN_220-244           FRCEFEGCERRFANSSD--RKKH--SHVH---
ZIC1_HUMAN_332-356           FKCEFEGCDRRFANSSD--RKKH--MHVH---
Q96IQ9_HUMAN_176-198         FKC--ENCLLRFRTHRS--LFKH--LHVC---
Q6PIF0_HUMAN_243-265         FEC--PNCHERFARNST--LKCH--LTAC---
Q6DKI8_HUMAN_56-80           FECPVEGCCARFSARSS--LYIH--SKKH---
O95276_HUMAN_64-86           FKCSQ--CGRGFVSAGV--LKAH--IRTH---
PRDM6_HUMAN_542-564          FKCGY--CGRAFAGATT--LNNH--IRTH---
Q8WXE2_HUMAN_275-299         FVCPFQGCNRRFIQSNN--LKAH--ILTH---
Q96MM3_HUMAN_275-299         FVCPFQGCNRRFIQSNN--LKAH--ILTH---
PRD12_HUMAN_271-293          FVCRF--CNRRFSQSST--LRNH--VRLH---
Q96AM6_HUMAN_172-194         FICKF--CGRHFTKSYN--LLIH--ERTH---
Q96LB6_HUMAN_172-194         FICKF--CGRHFTKSYN--LLIH--ERTH---
Q8TAX0_HUMAN_175-197         FVCKF--CGRHFTKSYN--LLIH--ERTH---
ZIC1_HUMAN_302-326           FPCPFPGCGKVFARSEN--LKIH--KRTH---
ZIC4_HUMAN_190-214           FPCPFPGCGKVFARSEN--LKIH--KRTH---
Q6PJS0_HUMAN_212-234         FECKE--CGRSFRNSSC--LNDH--IQIH---
Q96MU6_HUMAN_365-387         FACVV--CGKYFRNSSC--LNNH--VRIH---
Q6ZS14_HUMAN_108-131         FQCKF--CVRYFRSKNL--LIEHT-RKVH---
Q7Z7H7_HUMAN_473-495         FECKD--CGETFNKSAA--LAEH--RKIH---
Q9BU03_HUMAN_71-93           FICFT--CARSFLSSKA--LITH--QRSH---
Q9H609_HUMAN_71-93           FICFT--CARSFPSSKA--LITH--QRSH---
Q6ZN24_HUMAN_186-208         IRCEF--CGEFFENRKG--LSSH--ARSH---
Q6ZN24_HUMAN_370-392         IRCEF--CGEFFENRKG--LSSH--ARSH---
Q7LDZ1_HUMAN_178-200         IRCEF--CGEFFENRKG--LSSH--ARSH---
Q7LDZ1_HUMAN_362-384         IRCEF--CGEFFENRKG--LSSH--ARSH---
Q7LDZ1_HUMAN_5-27            TTCEV--CGACFETRKG--LSSH--ARSH---
Q6ZN24_HUMAN_80-102          MRCDF--CGAGFDTRAG--LSSH--ARAH---
Q7Z6T2_HUMAN_769-793         FICEMPNCGAVFSSRQA--LNGH--ARIH---
Q9NUN9_HUMAN_99-123          FICEMPNCGAVFSSRQA--LNGH--ARIH---
Q9NQ73_HUMAN_769-793         FICEMPNCGAVFSSRQA--LNGH--ARIH---
Q96PN7_HUMAN_1013-1037       FICEMPNCGAVFSSRQA--LNGH--ARIH---
OVOL1_HUMAN_118-140          FTCRV--CQKAFTYQRM--LNRH-MKC-H---
Q86XL8_HUMAN_95-117          FTCRV--CQKAFTYQRM--LNRH-MKC-H---
Q5R2W1_HUMAN_708-731         WPCEK--CGKMFTVHKQ--LERH-QEL-LC--
Q6ZMW8_HUMAN_890-913         FICRK--CQMMFTDEDA--AVNH-QKS-FC--
Q5T149_HUMAN_335-357         FQCAL--CQKSFTQLAH--LQKH-HLV-H---
Q8IZ20_HUMAN_335-357         FQCAL--CQKSFTQLAH--LQKH-HLV-H---
Q5T149_HUMAN_411-433         FQCSV--CRSRFTQHIH--LKLH-HRL-H---
Q8IZ20_HUMAN_411-433         FQCSV--CRSRFTQHIH--LKLH-HRL-H---
Q9H9H3_HUMAN_125-147         FSCDI--CGKLFTRREH--VKRH-SLV-H---
                               *    *              :
```

```
ZBT10_HUMAN_726-748          FSCDI--CGKLFTRREH--VKRH-SLV-H---
Q6DKI8_HUMAN_26-50           FTCPVEGCGKSFTRAEH--LKGH-SIT-H---
Q6IPT8_HUMAN_399-421         FQCPT--CGVRFTRIQN--LKQH-MLI-H---
Q86VJ7_HUMAN_399-421         FQCPT--CGVRFTRIQN--LKQH-MLI-H---
Q86XX5_HUMAN_399-421         FQCPT--CGVRFTRIQN--LKQH-MLI-H---
Q86VJ7_HUMAN_427-449         FQCDR--CGKKFTRAYS--LKMH-RLK-H---
Q86XX5_HUMAN_427-449         FQCDR--CGKKFTRAYS--LKMH-RLK-H---
Q6IPT8_HUMAN_427-449         FQCDR--CGKKFTRAYS--LKMH-RLK-H---
ZSCA1_HUMAN_292-314          FQCAD--CGMVFTWVTH--FIEH-QKT-H---
ZSCA1_HUMAN_380-402          FQCSV--CGKAFPWMVH--LIDH-QKL-H---
Q8NDX6_HUMAN_129-151         FECDI--CDMRFIQKYH--LERH-KRV-H---
Q9NWV2_HUMAN_55-77           FECDV--CHKTFRQLVH--LERH-QLT-H---
Q9HD72_HUMAN_355-377         VACEI--CGKIFRDVYH--LNRH-KLS-H---
Q5TC79_HUMAN_429-452         FHCHV--CGKSFPFQAI--LNQH-FRKNH---
Q8WYT0_HUMAN_338-361         FHCHV--CGKSFPFQAI--LNQH-FRKNH---
Q5VYE9_HUMAN_432-455         FRCEI--CGKCFPFQGT--LNQH-LRKNH---
ZBT34_HUMAN_428-451          FRCEI--CGKCFPFQGT--LNQH-LRKNH---
Q8N8K7_HUMAN_426-448         FVCPY--CHKKFRRQAY--LRKH-LST-H---
Q96Q84_HUMAN_426-448         FVCPY--CHKKFRRQAY--LRKH-LST-H---
Q96T92_HUMAN_427-449         FVCPY--CHKKFRRQAY--LRKH-LST-H---
Q7M4M3_HUMAN_21-43           FSCEI--CEKKFYAMAH--VRKH-MVA-H---
Q9NZF0_HUMAN_20-42           YACRE--CGRTFRDRNS--LLKH--MIIH---
ZN644_HUMAN_448-470          YACRE--CGRTFRDRNS--LLKH--MIIH---
HKR2_HUMAN_31-53             YACQE--CGCTFSNNSS--LVKH--WHVH---
Q9BU03_HUMAN_143-165         FACTE--CGQDFAQEAG--LHQH--YIRH---
Q9H609_HUMAN_143-165         FACTE--CGQDFAQEAG--LHQH--YIRH---
O14860_HUMAN_13-35           FVCNE--CGKAFRDRPG--FIRH--YIIH---
Q7Z6T2_HUMAN_842-864         FPCKE--CGKVFFKIKS--RNAH--MKTH---
Q7Z6T3_HUMAN_854-876         FPCKE--CGKVFFKIKS--RNAH--MKTH---
Q9NUN9_HUMAN_172-194         FPCKE--CGKVFFKIKS--RNAH--MKTH---
Q96PN7_HUMAN_1086-1108       FPCKE--CGKVFFKIKS--RNAH--MKTH---
Q9NQ72_HUMAN_854-876         FPCKE--CGKVFFKIKS--RNAH--MKTH---
Q9NQ73_HUMAN_842-864         FPCKE--CGKVFFKIKS--RNAH--MKTH---
Q8NDK8_HUMAN_477-499         FPCRE--CERVFDKIKS--RNAH--MKRH---
Q9H0D2_HUMAN_735-757         FPCRE--CERVFDKIKS--RNAH--MKRH---
ZSCA1_HUMAN_320-342          FPCPE--CGKVFLHNSV--LTEH--GKIH---
Q8WV14_HUMAN_211-233         FPCPE--CGMRFKRKFA--VEAH--QWIH---
Q9BU03_HUMAN_112-134         FPCPD--CGKTFGQAVS--LRRH--RQMH---
Q9H609_HUMAN_112-134         FPCPD--CGKTFGQAVS--LRRH--RQMH---
Q9HD72_HUMAN_292-314         LPCGL--CGKVFTDANR--LRQH--EAQH---
O14857_HUMAN_2-24            YVCT--ECERAFIRKSN--FITHQRIH-----
Q7Z3Q9_HUMAN_133-155         HVCS--ECGKAFCYKSE--FIRHQRSH-----
Q7Z7H7_HUMAN_420-442         YVCD--ECGRSFSVISE--FVEHQIMH-----
PRDM7_HUMAN_346-368          YVCR--ECGEGFSRKSD--LLSHQRTH-----
PRDM7_HUMAN_374-396          YVCR--ECERGFSRKSV--LLIHQRTH-----
Q96AF5_HUMAN_332-354         YTCE--QCGRGFDWKSV--FVIHHRTH-----
Q9UFF2_HUMAN_119-141         YTCE--QCGRGFDWKSV--FVIHHRTH-----
ZN446_HUMAN_332-354          YTCE--QCGRGFDWKSV--FVIHHRTH-----
Q9H686_HUMAN_229-251         YTCE--QCGRGFDWKSV--FVIHHRTH-----
Q96AF5_HUMAN_395-417         YPCE--ECGCSFSWKSQ--LVIHRKSH-----
Q9H686_HUMAN_292-314         YPCE--ECGCSFSWKSQ--LVIHRKSH-----
Q9UFF2_HUMAN_182-204         YPCE--ECGCSFSWKSQ--LVIHRKSH-----
ZN446_HUMAN_395-417          YPCE--ECGCSFSWKSQ--LVIHRKSH-----
Q9UFF2_HUMAN_210-232         HFCS--DCGRAFDWKSQ--LVIHRKGH-----
ZN446_HUMAN_423-445          HFCS--DCGRAFDWKSQ--LVIHRKGH-----
Q9H686_HUMAN_320-342         HFCS--DCGRAFDWKSQ--LVIHRKGH-----
                               *    *             :
```

```
Q8N2J5_HUMAN_379-401        YKCP--ECDSSFSHKSS--LTKHQITH-----
Q8N782_HUMAN_73-95          YECE--ECDKAFSFKSN--LESHRITH-----
Q8N246_HUMAN_335-357        HECG--ECRKTFSYKSN--LIRHRRVH-----
O14859_HUMAN_43-65          YKCK--QCGKGFSRRSA--VNVHCKVH-----
ZNF27_HUMAN_1-23            FKCV--ECGKGFSRRSA--LNVHHKLH-----
Q8N4W9_HUMAN_53-75          CKCH--QCGKVFSPRSL--LAEHEKIH-----
ZNF18_HUMAN_29-51           CKCD--YCGKGFSDFSG--LRHHEKIH-----
Q13580_HUMAN_30-52          YKCN--ECGKTFGQNSD--LLIHKSIH-----
Q8N782_HUMAN_45-67          HKCN--ECGKTFSQKSY--LACHRSIH-----
Q9H963_HUMAN_62-84          YKCN--QCGKTFSYKSS--LVIHKAIH-----
ZNF79_HUMAN_70-92           YKCN--ECGKFFSESSA--LIRHHIIH-----
O14889_HUMAN_12-34          YRCN--ECGKAFSVRSS--LTTHQAIH-----
O43694_HUMAN_47-69          YQCH--ECRKPLVSVSS--LTTHQTIH-----
Q15923_HUMAN_1-21           --CN--ECGKAFNQSAC--LMQHQRIH-----
ZN253_HUMAN_200-222         YRCE--ECGKAFNQSAN--LTTHKRIH-----
O14891_HUMAN_44-66          YRCE--ECGKAFGQSSS--LIHHQRIH-----
Q96MU6_HUMAN_279-303        KPCELEECGKASPVSSS--LTQHVRIH-----
Q15929_HUMAN_9-31           FKCI--VCGKAFNSSSN--LTTHKKIH-----
ZN253_HUMAN_172-194         FKCI--ICGKAFKRSST--LTTHKKIH-----
O43698_HUMAN_35-57          FKCE--ECGKAFNHPSA--LTTHKFIH-----
O43698_HUMAN_63-85          YKCE--ECEKAFNRFSY--LTKHKIIH-----
Q9H3U2_HUMAN_32-54          YKCE--ECGKAFNRFST--LTKHKRIH-----
Q8N229_HUMAN_6-28           YKCE--ECGKAFSVFST--LTKHKIIH-----
Q8N229_HUMAN_34-56          YKCE--ECGKAFNQSSI--FTKHKIIH-----
ZNF15_HUMAN_29-51           YKCK--ECGKAFNQSST--LMKHKIIH-----
ZN253_HUMAN_228-250         YKCE--ECGKAFNWSSD--LNKHKKIH-----
Q9H3U2_HUMAN_4-26           YKCE--ECDRAFSQSSN--LTEHKKIH-----
Q15929_HUMAN_37-59          YRCE--ECGKAFKRSSH--LTVHKIVH-----
Q9H3U2_HUMAN_60-82          YKCE--ECGKAFNQSYQ--LTRHKIVH-----
ZNF15_HUMAN_1-23            YKCE--ECGKSFILSSH--LTTHKIIH-----
Q13580_HUMAN_2-24           YKCK--VCDKAFAWNSH--LVRHTRIH-----
Q13580_HUMAN_86-108         YKCK--VCDKAFAWNSH--LVSHTSIH-----
Q8N782_HUMAN_17-39          YKCE--ECDKAFRHNSA--LQRHRRIH-----
ZNF27_HUMAN_29-51           YNCE--ECGKAFIHDSQ--LQEHQRIH-----
Q7Z7K7_HUMAN_195-217        YKCK--ECGKAFHHFSY--LVKHQRIH-----
Q8TBC5_HUMAN_413-435        YACG--ECGEAFAWLSH--LMEHHSSH-----
Q9H9A0_HUMAN_413-435        YACG--ECGEAFAWLSH--LMEHHSSH-----
Q9BRK7_HUMAN_310-332        YACG--ECGEAFAWLSH--LMEHHSSH-----
O14860_HUMAN_44-66          IECI--ECGKAFNRRSY--RTWHQQRH-----
Q8N2C6_HUMAN_45-67          YECV--ECGKAFNRRSP--LTRHQRIH-----
O14852_HUMAN_2-24           HECN--QCGKAFNRSSN--HIHHQKVH-----
Q16524_HUMAN_11-33          YECN--ACGEAFIRSKS--LARHQVLH-----
Q16524_HUMAN_67-89          YECS--ECGKAFSRSKC--LIRHQSLH-----
Q16524_HUMAN_39-61          YKCN--ACGRAFCSNRN--LIDHQRIH-----
Q8N2C6_HUMAN_17-39          YVCI--QCGKAFCRTTN--LIRHFSIH-----
Q15933_HUMAN_33-55          YECS--QCGKAFRQSTH--LTQHQRIH-----
Q15936_HUMAN_39-61          YECS--QCGKAFRQSTH--LTQHQRIH-----
O14885_HUMAN_33-55          YECH--QCGKAFSQRAH--LTIHQRIH-----
Q8N8C0_HUMAN_33-55          YECQ--ICGKPFRKRAH--LTQHNRIH-----
Q6ZMZ9_HUMAN_106-128        YECS--ECGKVIRRKAW--FDQHQRIH-----
Q8N8C0_HUMAN_89-111         YECL--ECRKTFRRSAH--LIRHQRIH-----
Q8N246_HUMAN_279-301        FECS--ECEESFSKKCH--LILHKIIH-----
ZN212_HUMAN_455-477         YSCT--ECEKSFVQKQH--LLQHQKIH-----
Q8N2J5_HUMAN_317-339        YECA--ECEISFRHKQQ--LTLHQRIH-----
ZN212_HUMAN_316-338         YECS--ECEITFRYKQQ--LATHLRSH-----
ZNF32_HUMAN_1-23            YECQ--ECGKSFRQKGS--LTLHERIH-----
                              *    *             :
```

```
Q7Z7K7_HUMAN_279-301        YECK--ECGKSFTSGST--LNQHQQIH-----
Q7Z7K7_HUMAN_307-329        YHCK--QCGKSFTVGST--LIRHQQIH-----
Q8N246_HUMAN_251-273        YECS--KCEKAFTCKNT--LVQHQQIH-----
O14861_HUMAN_41-63          YECT--ECGKSFSVKGK--LIQHQRIH-----
ZN154_HUMAN_63-85           YECT--ECGKSFSHNSS--LIKHQRIH-----
Q7Z3Q9_HUMAN_161-183        YGCT--DCGKAFSHKST--LIKHQRIH-----
Q15933_HUMAN_61-83          YECN--DCGKAFSHSSS--LTKHQRIH-----
Q15936_HUMAN_67-89          YECN--DCGNPFSHSSS--LTKHQRIH-----
Q7KZ25_HUMAN_73-95          YVCN--DCGKAFSQSSS--LIYHQRIH-----
ZK23_HUMAN_1-23             YECS--ECGKSFRQRSG--LIQHRRLH-----
ZN154_HUMAN_35-57           YECS--ECGKSFTQNSG--LIKHRRVH-----
ZK23_HUMAN_29-51            YECS--ECGKSFSQSAS--LIQHQRVH-----
O14885_HUMAN_5-27           YECE--ECGKEFRHISS--LIAHQRMH-----
ZN154_HUMAN_7-29            YECS--ECGKFFPYSSS--LRKHQRVH-----
ZNF17_HUMAN_29-51           YECS--ECGKFFVDSCT--LKSHQRVH-----
ZNF17_HUMAN_1-23            YECN--KCGKFFRYCFT--LNRHQRVH-----
Q5T149_HUMAN_307-329        YECN--ICGKSFGQLSN--LKVHLRVH-----
Q8IZ20_HUMAN_307-329        YECN--ICGKSFGQLSN--LKVHLRVH-----
ZNF12_HUMAN_29-51           YECN--ECGKFFSRLSY--LTVHYRTH-----
ZN126_HUMAN_29-51           YECN--QCGKAFSKHSH--LQCHKRTH-----
ZN126_HUMAN_57-79           YECN--QCGKAFSQHGL--LQRHKRTH-----
ZN126_HUMAN_1-23            YECN--QCGKAFAQHSS--LKCHYRTH-----
O14892_HUMAN_1-21           --CP--ECGKSFCQKVT--STQHQRTH-----
ZNF21_HUMAN_29-51           YACT--ECGKAFREKST--FTVHQRTH-----
ZNF12_HUMAN_1-23            YKCS--ECGKCFCRKST--LTTHLRTH-----
O14892_HUMAN_27-49          YECN--ACGKTFYHKSV--LTRHQIIH-----
O14892_HUMAN_55-77          YECY--ECGKTFCLKSD--LTIHQRSH-----
ZN174_HUMAN_354-376         YTCG--ECGNCFGRQST--LKLHQRIH-----
Q5MPB1_HUMAN_54-76          YICE--ECGIRCKKPSM--LKKHIRTH-----
Q82133_HUMAN_41-63          YICE--ECGIRCKKPSM--LKKHIRTH-----
ZEP2_HUMAN_1186-1208        YICE--ECGIRCKKPSM--LKKHIRTH-----
Q99302_HUMAN_3-25           YVCE--ECGIRCKKPSM--LKKHIRTH-----
Q8N8C0_HUMAN_61-83          YECK--ECGKVFICCST--LIQHKRTH-----
HKR2_HUMAN_59-81            YMCG--HCGKCFRESSS--LAKHQRVH-----
ZN174_HUMAN_382-404         YQCG--QCGKSFRQSSN--LHQHHRLH-----
ZBT12_HUMAN_387-409         HSCG--ICGKCFTQKST--LHDHLNLH-----
Q8TAX0_HUMAN_231-253        FKCQ--ECGKGFCQSRT--LAVHKTLH-----
Q96AM6_HUMAN_228-250        FKCQ--ECGKGFCQSRT--LAVHKTLH-----
Q96LB6_HUMAN_228-250        FKCQ--ECGKGFCQSRT--LAVHKTLH-----
Q9H963_HUMAN_34-56          YKCD--ICGKVFNQKRY--LAYHHRCH-----
Q9H963_HUMAN_90-112         HKCN--ECGKVFNQKAY--FASHHRLH-----
Q15921_HUMAN_36-58          YECE--ECGKSFSRSSH--LAQHQRTH-----
Q96K08_HUMAN_183-205        YTCE--ECGKAFSRSSF--LVQHQRIH-----
O14852_HUMAN_30-53          YTCV--ECGKAFSQSSH--LIQHQ-IHPH---
Q7KZ25_HUMAN_45-67          FKCD--ECGKGFVQGSH--LIQHQRIH-----
ZN396_HUMAN_251-273         QKCD--ECGKIFSQSSA--LILHQRIH-----
Q5U5Z4_HUMAN_242-264        FKCK--ECLKAFSQSSA--LIQHQRTH-----
O14887_HUMAN_11-33          HTCD--ECGKSFCYISA--LHIHQRVH-----
O14888_HUMAN_11-33          HTCD--ECGKSFCYISA--LHIHQRVH-----
O14890_HUMAN_12-34          HTCD--ECGKSFCSISA--QHIHQRVH-----
Q15921_HUMAN_8-30           HKCN--ECGKSFCRLSH--LIQHQRTH-----
O14891_HUMAN_16-38          FKCI--ECGKAFRLSSK--LIQHQRIH-----
Q6RFR8_HUMAN_138-160        HECM--ICGKAFRLHSH--LIQHQRIH-----
Q8N4W9_HUMAN_25-47          HKCD--DCGKAFTSHSH--LVGHQRIH-----
Q15923_HUMAN_27-49          YTCT--ECGKAFTQNSS--LVEHERTH-----
ZNF26_HUMAN_29-51           FKCS--ECGKAFTQKSS--LSEHQRVH-----
                              *    *             :
```

```
Q15933_HUMAN_5-27          YGCN--ECGKTFSHSSS--LSQHERTH-----
Q15936_HUMAN_11-33         YGCN--ECGKTFSHSSS--LSQHERTH-----
Q96K08_HUMAN_155-177       WKCN--ECGKTFTQSSS--LTQHQRTH-----
O14889_HUMAN_40-62         YKCN--ECGKVFTQNAH--LANHRRIH-----
O43694_HUMAN_19-41         LLCH--ECGKVFTQNSH--LVRHRGIH-----
Q92669_HUMAN_44-66         FTCH--ECGKKFSQNSH--LIKHRRTH-----
O14857_HUMAN_30-52         YECS--DCGKSFTSKSQ--LLVHQPVH-----
Z297B_HUMAN_373-394        YPC---QCGKSFTHKSQ--RDRHMSMH-----
ZN174_HUMAN_326-348        YKCD--DCGKSFTWNSE--LKRHKRVH-----
O14887_HUMAN_39-61         IKCD--VCGKEFSQSSH--LQTHQRVH-----
O14890_HUMAN_40-62         YKCD--VCGKEFSQSSR--LQTHQRVH-----
O14888_HUMAN_39-61         YKCY--VCGKEFSQSSH--LQTHQRVH-----
O14859_HUMAN_15-37         LKCD--ECGKEFSQGAH--LQTHQKVH-----
O14899_HUMAN_19-41         FKCE--ECGKEFSWSAG--LSAHQRVH-----
Q9UFH1_HUMAN_26-48         YKCE--VCGKGFSQRSN--LQAHQRVH-----
ZNF13_HUMAN_29-51          YKCE--ECGKGFSRASN--LLAHQRGH-----
ZN229_HUMAN_377-399        YKCE--ECGKAFGRSSN--LLVHQRVH-----
O14899_HUMAN_47-69         YTCQ--QCGKGFSQASH--FHTHQRVH-----
O14899_HUMAN_75-97         YICD--VCCKGFSQRSH--LIYHQRVH-----
Q9UFH1_HUMAN_54-76         YKCD--ACGKGFRWSSG--LLIHQRVH-----
ZN229_HUMAN_349-371        YRCD--VCGKGFRYKSV--LLIHQGVH-----
ZNF13_HUMAN_1-23           YQCD--ACGKGFSRSSD--FNIHFRVH-----
GLIS1_HUMAN_322-346        YACQIPGCSKRYTDPSS--LRKHVKAH-----
Q5VTL4_HUMAN_322-346       YACQIPGCSKRYTDPSS--LRKHVKAH-----
GLIS3_HUMAN_471-495        YACQIPGCTKRYTDPSS--LRKHVKAH-----
Q5VZV9_HUMAN_472-496       YACQIPGCTKRYTDPSS--LRKHVKAH-----
Q9BZE0_HUMAN_293-317       YYCKMPGCHKRYTDPSS--LRKHIKAH-----
Q6ZTI2_HUMAN_123-147       YTCKVRGCDKCYTHPSS--LRKHMKVH-----
ZIC4_HUMAN_250-274         YTCKVRGCDKCYTHPSS--LRKHMKVH-----
ZIC1_HUMAN_362-384         YLCKM--CDKSYTHPSS--VRKHMKVH-----
Q5U5Z4_HUMAN_214-236       YKCST--CEKAFIHNSS--LRRHQKNH-----
Q6ZN79_HUMAN_172-194       YQCNL--CEKAYTNCFR--LRRHKMTH-----
Q6ZN79_HUMAN_200-222       YACHL--CGKAFTQCSH--LRRHEKTH-----
Q8TBC5_HUMAN_441-463       YACQG--CWKTFHFSLA--LAEHQKTH-----
Q9H9A0_HUMAN_441-463       YACQG--CWKTFHFSLA--LAEHQKTH-----
Q9BRK7_HUMAN_338-360       YACQG--CWKTFHFSLA--LAEHQKTH-----
Q96MU6_HUMAN_337-359       YECKD--CGKACGGFYL--LNEHGKTH-----
Y0352_HUMAN_633-655        YQCKV--CHKFFRGRST--IKCHLKTH-----
Q5U5Z4_HUMAN_270-292       YIC--KECGKAFSHSAS--LCKHLRTH-----
ZN396_HUMAN_279-301        YAC--DECAKAFSRSAI--LIQHRRTH-----
Q6PJS0_HUMAN_240-262       HKC--TYCGKAFTRSTQ--LTEHVRTH-----
Q8NEI5_HUMAN_209-231       YTC--RYCGKIFPRSAN--LTRHLRTH-----
Q6Q7C8_HUMAN_195-219       YLCDYPDCGKAFVQSGQ--LKTHQRLH-----
Q7RTV3_HUMAN_195-219       YLCDYPDCGKAFVQSGQ--LKTHQRLH-----
ZNF21_HUMAN_1-23           YEC--PVCWKAFSQKSQ--LIIHQRTH-----
Q6ZN79_HUMAN_228-250       YKC--HQCGKAFIQSFN--LRRHERTH-----
ZN396_HUMAN_307-329        YKC--HDCGKAFSQSSN--LFRHRKRH-----
Q92669_HUMAN_16-38         YKC--SWCGKSFSQNTN--LHTHQRTH-----
ZNF79_HUMAN_42-64          YKC--SECGKAFSQSTN--LIIHQKTH-----
Q7KZ25_HUMAN_17-39         YKC--NKCTKAFGCSSR--LIRHQRTH-----
Q8TB80_HUMAN_18-40         YSC--SACGKCFGGSGD--LRRHVRTH-----
O14858_HUMAN_11-33         YECKQ--CSKAFPVYSS--YLRHEKIH-----
O14858_HUMAN_39-61         YECKQ--CSKAFPDYSS--YLRHERTH-----
Q15930_HUMAN_48-70         YECKE--CGKAFSSFKY--FCRHERTH-----
Q8NAJ8_HUMAN_36-58         YECKQ--CGKAFVSFTS--FRYHERTH-----
Q6RFS2_HUMAN_30-52         FECKR--CGKAFRSSSS--FRLHERTH-----
                             *    *             :
```

```
Q8NAJ8_HUMAN_92-114         CECKK--CGKAFHNFSS--LQIHERMH-----
Q15930_HUMAN_21-42          YKCK---CGKAFSDLFS--FQSHETTH-----
Q6RFS2_HUMAN_58-80          YHCKE--CGKAYSCRAS--FQRHMLTH-----
Q6LDH8_HUMAN_11-33          YECKQ--CGKLLSHRSS--FRRHMMAH-----
Q6RFS2_HUMAN_2-24           YECKQ--CGKAFSYLNS--FRTHEMIH-----
Q6RFR8_HUMAN_166-188        YECKE--CGKAFSYHSS--FSHHQRIH-----
Q7M4M1_HUMAN_8-30           YECKE--CGKPFSFLTG--FRVHMRMH-----
Q6Y8G5_HUMAN_93-115         HGCTR--CGKNFSSASA--LQIHERTH-----
Q6Y8G6_HUMAN_433-455        HGCTR--CGKNXSXXSA--LQIHERTH-----
Q6LDH8_HUMAN_39-61          HKCTV--CGKAFDSPSV--FQRHERTH-----
GLIS1_HUMAN_262-286         NKCMFEGCSKAFSRLEN--LKIHLRSH-----
Q5VTL4_HUMAN_262-286        NKCMFEGCSKAFSRLEN--LKIHLRSH-----
GLIS3_HUMAN_411-435         NKCTFEGCEKAFSRLEN--LKIHLRSH-----
Q5VZV9_HUMAN_412-436        NKCTFEGCEKAFSRLEN--LKIHLRSH-----
Q9BZE0_HUMAN_235-257        HRCP--TCSKSFSRLEN--LKIHNRSH-----
Q8N2J5_HUMAN_289-311        FTCM--ECGKSFRLKIN--LIIHQRNH-----
ZNF32_HUMAN_29-51           FECT--HCGKSFRAKAN--LVTHQRIH-----
Q8WXE2_HUMAN_245-269        FRCTFEGCGKRFSLDFN--LRTHVRIH-----
Q96MM3_HUMAN_245-269        FRCTFEGCGKRFSLDFN--LRTHVRIH-----
OVOL1_HUMAN_146-168         HLCTY--CGKGFNDTFD--LKRHVRTH-----
Q86XL8_HUMAN_123-145        HLCTY--CEKGFNDTFD--LKRHVRTH-----
O00110_HUMAN_125-147        HLCRC--CGKGFHDAFD--LKRHMRTH-----
Q7Z5E7_HUMAN_74-96          HLCLY--CGKLYSRKYG--LKIHMRTH-----
Q5XG86_HUMAN_438-460        YLCQQ--CGAAFAHNYD--LKNHMRVH-----
ZBTB7_HUMAN_438-460         YLCQQ--CGAAFAHNYD--LKNHMRVH-----
O73453_HUMAN_420-442        YLCIH--CNAKFVHNYD--LKNHMRIH-----
Q5TC79_HUMAN_373-395        LTCIY--CAKSFNQKGS--LDRHMRLH-----
Q8WYT0_HUMAN_282-304        LTCIY--CAKSFNQKGS--LDRHMRLH-----
Q5VYE9_HUMAN_376-398        LICIY--CGKSFNQKGS--LDRHMRLH-----
ZBT34_HUMAN_372-394         LICIY--CGKSFNQKGS--LDRHMRLH-----
ZN212_HUMAN_427-449         LICGY--CGKSFSHPSD--LVRHQRIH-----
BTEB1_HUMAN_143-167         HKCPYSGCGKVYGKSSH--LKAHYRV-H----
KLF13_HUMAN_167-191         HKCHYAGCEKVYGKSSH--LKAHLRT-H----
Q86TP0_HUMAN_199-223        HICHIPGCGKVYGKTSH--LRAHLRW-H----
Q8WWU2_HUMAN_319-343        HICHIPGCGKVYGKTSH--LRAHLRW-H----
Q8WWU4_HUMAN_553-577        HICHIPGCGKVYGKTSH--LRAHLRW-H----
Q8WWU3_HUMAN_336-360        HICHIPGCGKVYGKTSH--LRAHLRW-H----
SP3_HUMAN_621-645           HICHIPGCGKVYGKTSH--LRAHLRW-H----
Q68DP2_HUMAN_11-35          HICHIPGCGKVYGKTSH--LRAHLRW-H----
Q69B26_HUMAN_568-592        HICHIPGCGKVYGKTSH--LRAHLRW-H----
Q69B27_HUMAN_621-645        HICHIPGCGKVYGKTSH--LRAHLRW-H----
SP1_HUMAN_626-650           HICHIQGCGKVYGKTSH--LRAHLRW-H----
SP4_HUMAN_647-671           HICHIEGCGKVYGKTSH--LRAHLRW-H----
Q6BEB4_HUMAN_296-320        HVCHVPGCGKVYGKTSH--LKAHLRW-H----
Q6NTF6_HUMAN_294-318        HVCHVPGCGKVYGKTSH--LKAHLRW-H----
Q7Z615_HUMAN_356-380        HSCHIPGCGKVYGKTSH--LKAHLRW-H----
SP8_HUMAN_374-398           HSCHIPGCGKVYGKTSH--LKAHLRW-H----
Q7Z718_HUMAN_276-300        HSCHIPGCGKVYGKASH--LKAHLRW-H----
SP7_HUMAN_294-318           HSCHIPGCGKVYGKASH--LKAHLRW-H----
KLF10_HUMAN_369-393         HICSHPGCGKTYFKSSH--LKAHTRT-H----
O75411_HUMAN_358-382        HICSHPGCGKTYFKSSH--LKAHTRT-H----
KLF11_HUMAN_394-418         YVCSFPGCRKTYFKSSH--LKAHLRT-H----
KLF1_HUMAN_279-303          HTCAHPGCGKSYTKSSH--LKAHLRT-H----
Q6PIJ5_HUMAN_279-303        HTCAHPGCGKSYTKSSH--LKAHLRT-H----
KLF4_HUMAN_387-411          HTCDYAGCGKTYTKSSH--LKAHLRT-H----
Q8N717_HUMAN_421-445        HTCDYAGCGKTYTKSSH--LKAHLRT-H----
                              *    *              :
```

```
Q5T3J9_HUMAN_346-370        HTCDYAGCGKTYTKSSH--LKAHLRT-H----
KLF2_HUMAN_272-296          HTCSYAGCGKTYTKSSH--LKAHLRT-H----
KLF12_HUMAN_317-341         HRCDFEGCNKVYTKSSH--LKAHRRT-H----
Q8WWI3_HUMAN_317-341        HRCDFEGCNKVYTKSSH--LKAHRRT-H----
KLF3_HUMAN_260-284          HRCDYDGCNKVYTKSSH--LKAHRRT-H----
Q6PIR1_HUMAN_260-284        HRCDYDGCNKVYTKSSH--LKAHRRT-H----
KLF5_HUMAN_373-397          HYCDYPGCTKVYTKSSH--LKAHLRT-H----
KLF8_HUMAN_274-298          HQCDFAGCSKVYTKSSH--LKAHRRI-H----
KLF6_HUMAN_200-224          HRCHFNGCRKVYTKSSH--LKAHQRT-H----
Q9BT79_HUMAN_200-224        HRCHFNGCRKVYTKSSH--LKAHQRT-H----
KLF7_HUMAN_219-243          HRCQFNGCRKVYTKSSH--LKAHQRT-H----
Q7Z3H8_HUMAN_186-210        HRCQFNGCRKVYTKSSH--LKAHQRT-H----
KLF15_HUMAN_321-345         HKCTFPGCSKMYTKSSH--LKAHLRR-H----
BTEB4_HUMAN_127-151         HRCPFPDCAKAYYKSSH--LKSHLRT-H----
KLF14_HUMAN_195-219         HQCPFPGCTKAYYKSSH--LKSHQRT-H----
O95276_HUMAN_36-58          YKCFY--CHRAYKKSCH--LKQHIRS-H----
Q86VQ7_HUMAN_283-307        YCCNYENCGKAYTKRSH--LVSHQRK-H----
Q8N805_HUMAN_283-307        YCCNYENCGKAYTKRSH--LVSHQRK-H----
Q5TC79_HUMAN_401-423        FVCRM--CGKKYTRKDQ--LEYHIRK-H----
Q8WYT0_HUMAN_310-332        FVCRM--CGKKYTRKDQ--LEYHIRK-H----
Q5VYE9_HUMAN_404-426        FVCKF--CGKKYTRKDQ--LEYHIRG-H----
ZBT34_HUMAN_400-422         FVCKF--CGKKYTRKDQ--LEYHIRG-H----
BTEB4_HUMAN_157-181         FACDWQGCDKKFARSDE--LARHHRT-H----
KLF13_HUMAN_197-221         FACSWQDCNKKFARSDE--LARHYRT-H----
KLF14_HUMAN_225-249         FSCDWLDCDKKFTRSDE--LARHYRT-H----
BTEB1_HUMAN_173-197         FPCTWPDCLKKFSRSDE--LTRHYRT-H----
KLF10_HUMAN_399-423         FSCSWKGCERRFARSDE--LSRHRRT-H----
O75411_HUMAN_388-412        FSCSWKGCERRFARSDE--LSRHRRT-H----
KLF11_HUMAN_424-448         FNCSWDGCDKKFARSDE--LSRHRRT-H----
KLF15_HUMAN_351-375         FACTWPGCGWRFSRSDE--LSRHRRS-H----
KLF12_HUMAN_347-371         YKCTWEGCTWKFARSDE--LTRHYRK-H----
Q8WWI3_HUMAN_347-371        YKCTWEGCTWKFARSDE--LTRHYRK-H----
KLF3_HUMAN_290-314          YKCTWEGCTWKFARSDE--LTRHFRK-H----
Q6PIR1_HUMAN_290-314        YKCTWEGCTWKFARSGE--LTRHFRK-H----
KLF8_HUMAN_304-328          YKCTWDGCSWKFARSDE--LTRHFRK-H----
KLF1_HUMAN_309-333          YACTWEGCGWRFARSDE--LTRHYRK-H----
Q6PIJ5_HUMAN_309-333        YACTWEGCGWRFARSDE--LTRHYRK-H----
KLF5_HUMAN_403-427          YKCTWEGCDWRFARSDE--LTRHYRK-H----
Q7Z3W8_HUMAN_5-29           YRCSWEGCEWRFARSDE--LTRHFRK-H----
Q9BT79_HUMAN_230-254        YRCSWEGCEWRFARSDE--LTRHFRK-H----
KLF6_HUMAN_230-254          YRCSWEGCEWRFARSDE--LTRHFRK-H----
KLF7_HUMAN_249-273          YKCSWEGCEWRFARSDE--LTRHYRK-H----
Q7Z3H8_HUMAN_216-240        YKCSWEGCEWRFARSDE--LTRHYRK-H----
Q5T3J9_HUMAN_376-400        YHCDWDGCGWKFARSDE--LTRHYRK-H----
Q8N717_HUMAN_451-475        YHCDWDGCGWKFARSDE--LTRHYRK-H----
KLF4_HUMAN_417-441          YHCDWDGCGWKFARSDE--LTRHYRK-H----
KLF2_HUMAN_302-326          YHCNWDGCGWKFARSDE--LTRHYRK-H----
Q68DP2_HUMAN_41-65          FVCNWMYCGKRFTRSDE--LQRHRRT-H----
Q69B26_HUMAN_598-622        FVCNWMYCGKRFTRSDE--LQRHRRT-H----
Q69B27_HUMAN_651-675        FVCNWMYCGKRFTRSDE--LQRHRRT-H----
SP3_HUMAN_651-675           FVCNWMYCGKRFTRSDE--LQRHRRT-H----
Q86TP0_HUMAN_229-253        FVCNWMYCGKRFTRSDE--LQRHRRT-H----
Q8WWU2_HUMAN_349-373        FVCNWMYCGKRFTRSDE--LQRHRRT-H----
Q8WWU3_HUMAN_366-390        FVCNWMYCGKRFTRSDE--LQRHRRT-H----
Q8WWU4_HUMAN_583-607        FVCNWMYCGKRFTRSDE--LQRHRRT-H----
SP4_HUMAN_677-701           FICNWMFCGKRFTRSDE--LQRHRRT-H----
                              *    *              :
```

```
Q6BEB4_HUMAN_326-350        FVCNWLFCGKSFTRSDE--LQRHLRT-H----
Q6NTF6_HUMAN_324-348        FVCNWLFCGKSFTRSDE--LQRHLRT-H----
Q7Z615_HUMAN_386-410        FVCNWLFCGKRFTRSDE--LQRHLRT-H----
SP8_HUMAN_404-428           FVCNWLFCGKRFTRSDE--LQRHLRT-H----
SP2_HUMAN_548-572           FVCNWFFCGKRFTRSDE--LQRHART-H----
Q7Z718_HUMAN_306-330        FVCNWLFCGKRFTRSDE--LERHVRT-H----
SP7_HUMAN_324-348           FVCNWLFCGKRFTRSDE--LERHVRT-H----
SP1_HUMAN_656-680           FMCTWSYCGKRFTRSDE--LQRHKRT-H----
EGR2_HUMAN_340-364          YPCPAEGCDRRFSRSDE--LTRHIRI-H----
Q68CZ5_HUMAN_340-364        YPCPAEGCDRRFSRSDE--LTRHIRI-H----
EGR3_HUMAN_275-299          HACPAEGCDRRFSRSDE--LTRHLRI-H----
EGR1_HUMAN_338-362          YACPVESCDRRFSRSDE--LTRHIRI-H----
EGR4_HUMAN_380-404          FACPVESCVRSFARSDE--LNRHLRI-H----
EGR2_HUMAN_398-420          FAC--DYCGRKFARSDE--RKRHTKI-H----
Q68CZ5_HUMAN_398-420        FAC--DYCGRKFARSDE--RKRHTKI-H----
EGR1_HUMAN_396-418          FAC--DICGRKFARSDE--RKRHTKI-H----
EGR3_HUMAN_333-355          FAC--EFCGRKFARSDE--RKRHAKI-H----
EGR4_HUMAN_438-460          FAC--DVCGRRFARSDE--KKRHSKV-H----
Q86VQ7_HUMAN_313-337        YSCNWESCSWSFFRSDE--LRRHMRV-H----
Q8N805_HUMAN_313-337        YSCNWESCSWSFFRSDE--LRRHMRV-H----
BTEB1_HUMAN_203-225         FRCP--LCEKRFMRSDH--LTKHARR-H----
KLF13_HUMAN_227-249         FSCP--ICEKRFMRSDH--LTKHARR-H----
KLF10_HUMAN_429-451         FACP--MCDRRFMRSDH--LTKHARR-H----
O75411_HUMAN_418-440        FACP--MCDRRFMRSDH--LTKHARR-H----
KLF11_HUMAN_454-476         FVCP--VCDRRFMRSDH--LTKHARR-H----
BTEB4_HUMAN_187-209         FSCP--LCSKRFTRSDH--LAKHARR-H----
KLF14_HUMAN_255-277         FSCP--LCPKQFSRSDH--LTKHARR-H----
Q7Z718_HUMAN_336-358        FTCL--LCSKRFTRSDH--LSKHQRT-H----
SP7_HUMAN_354-376           FTCL--LCSKRFTRSDH--LSKHQRT-H----
Q68DP2_HUMAN_71-93          FVCP--ECSKRFMRSDH--LAKHIKT-H----
Q69B27_HUMAN_681-703        FVCP--ECSKRFMRSDH--LAKHIKT-H----
Q8WWU2_HUMAN_379-401        FVCP--ECSKRFMRSDH--LAKHIKT-H----
Q8WWU3_HUMAN_396-418        FVCP--ECSKRFMRSDH--LAKHIKT-H----
Q8WWU4_HUMAN_613-635        FVCP--ECSKRFMRSDH--LAKHIKT-H----
SP3_HUMAN_681-703           FVCP--ECSKRFMRSDH--LAKHIKT-H----
Q69B26_HUMAN_628-650        FVCP--ECSKRFMRSDH--LAKHIKT-H----
Q86TP0_HUMAN_259-281        FVCP--ECSKRFMRSDH--LAKHIKT-H----
Q6BEB4_HUMAN_356-378        FACP--ECGKRFMRSDH--LAKHVKT-H----
Q6NTF6_HUMAN_354-376        FACP--ECGKRFMRSDH--LAKHVKT-H----
Q7Z615_HUMAN_416-438        FACP--VCNKRFMRSDH--LSKHVKT-H----
SP8_HUMAN_434-456           FACP--VCNKRFMRSDH--LSKHVKT-H----
SP1_HUMAN_686-708           FACP--ECPKRFMRSDH--LSKHIKT-H----
SP4_HUMAN_707-729           FECP--ECSKRFMRSDH--LSKHVKT-H----
SP2_HUMAN_578-600           FECA--QCQKRFMRSDH--LTKHYKT-H----
Q86VQ7_HUMAN_343-365        YKCD--QCSREFMRSDH--LKQHQKT-H----
Q8N805_HUMAN_343-365        YKCD--QCSREFMRSDH--LKQHQKT-H----
EGR2_HUMAN_370-392          FQCR--ICMRNFSRSDH--LTTHIRT-H----
Q68CZ5_HUMAN_370-392        FQCR--ICMRNFSRSDH--LTTHIRT-H----
EGR1_HUMAN_368-390          FQCR--ICMRNFSRSDH--LTTHIRT-H----
EGR3_HUMAN_305-327          FQCR--ICMRSFSRSDH--LTTHIRT-H----
EGR4_HUMAN_410-432          FQCR--ICLRNFSRSDH--LTSHVRT-H----
KLF12_HUMAN_377-399         FKCA--DCDRSFSRSDH--LALHRRR-H----
Q8WWI3_HUMAN_377-399        FKCA--DCDRSFSRSDH--LALHRRR-H----
KLF3_HUMAN_320-342          FQCP--DCDRSFSRSDH--LALHRKR-H----
Q6PIR1_HUMAN_320-342        FQCP--DCDRSFSRSDH--LALHRKR-H----
KLF8_HUMAN_334-356          FRCT--DCNRSFSRSDH--LSLHRRR-H----
                              *    *              :
```

```
KLF1_HUMAN_339-361          FRCQ--LCPRAFSRSDH--LALHMKR-H----
Q6PIJ5_HUMAN_339-361        FRCQ--LCPRAFSRSDH--LALHMKR-H----
Q5T3J9_HUMAN_406-428        FQCQ--KCDRAFSRSDH--LALHMKR-H----
Q8N717_HUMAN_481-503        FQCQ--KCDRAFSRSDH--LALHMKR-H----
KLF4_HUMAN_447-469          FQCQ--KCDRAFSRSDH--LALHMKR-H----
KLF2_HUMAN_332-354          FQCH--LCDRAFSRSDH--LALHMKR-H----
KLF5_HUMAN_433-455          FQCG--VCNRSFSRSDH--LALHMKR-H----
KLF6_HUMAN_260-282          FKCS--HCDRCFSRSDH--LALHMKR-H----
Q7Z3W8_HUMAN_35-57          FKCS--HCDRCFSRSDH--LALHMKR-H----
KLF7_HUMAN_279-301          FKCN--HCDRCFSRSDH--LALHMKR-H----
Q7Z3H8_HUMAN_246-268        FKCN--HCDRCFSRSDH--LALHMKR-H----
Q5XG86_HUMAN_466-490        YQCD--SCCKTFVRSDH--LHRHLKK-DGC--
ZBTB7_HUMAN_466-490         YQCD--SCCKTFVRSDH--LHRHLKK-DGC--
O73453_HUMAN_448-472        YQCE--FCYKSFTRSDH--LHRHIKR-QSC--
KLF15_HUMAN_381-403         YQCP--VCEKKFARSDH--LSKHIKV-H----
Q8NDX6_HUMAN_157-179        YQCE--RCHQCFSRTDR--LLRHKRM-C----
Q8TB80_HUMAN_46-68          YTCE--ICNKCFTRSAV--LRRHKKM-H----
Q5XG86_HUMAN_410-432        YECN--ICKVRFTRQDK--LKVHMRK-H----
ZBTB7_HUMAN_410-432         YECN--ICKVRFTRQDK--LKVHMRK-H----
O73453_HUMAN_392-414        YMCT--ICEVRFTRQDK--LKIHMRK-H----
Z297B_HUMAN_428-450         YECN--ICAKRFMWRDS--FHRHVTS-C----
Q6ZN18_HUMAN_45-70          YNCCWDQCQACFNSSPD--LADHIRS-IH---
Q96BG3_HUMAN_53-78          YNCCWDQCQACFNSSPD--LADHIRS-IH---
Q8TAX0_HUMAN_203-225        YTC--DICHKAFRRQDH--LRDH-RY-IH---
Q96AM6_HUMAN_200-222        YTC--DICHKAFRRQDH--LRDH-RY-IH---
Q96LB6_HUMAN_200-222        YTC--DICHKAFRRQDH--LRDH-RY-IH---
Q8IUL5_HUMAN_310-333        YPC--QACGKRFSRLDH--LSSHFRT-IH---
Q8N2Y5_HUMAN_120-143        YPC--QACGKRFSRLDH--LSSHFRT-IH---
Q96BR9_HUMAN_310-333        YPC--QACGKRFSRLDH--LSSHFRT-IH---
Q96BX0_HUMAN_105-128        YPC--QACGKRFSRLDH--LSSHFRT-IH---
ZBTB8_HUMAN_369-392         YPC--ETCGKRFTRQEH--LRSHALS-VH---
GLIS1_HUMAN_292-316         YLCQHPGCQKAFSNSSD--RAKHQRT-H----
GLIS3_HUMAN_441-465         YLCQHPGCQKAFSNSSD--RAKHQRT-H----
Q5VTL4_HUMAN_292-316        YLCQHPGCQKAFSNSSD--RAKHQRT-H----
Q5VZV9_HUMAN_442-466        YLCQHPGCQKAFSNSSD--RAKHQRT-H----
Q9BZE0_HUMAN_263-287        YVCPYEGCNKRYSNSSD--RFKHTRT-H----
Q6N043_HUMAN_418-441        YVCQV--CNYRSSSFSD--VETHFRTSH----
Q9H740_HUMAN_254-277        YVCQV--CNYRSSSFSD--VETHFRTSH----
Q6MZM6_HUMAN_464-487        YVCQV--CNYRSSSFSD--VETHFRTSH----
Q6N085_HUMAN_418-441        YVCQV--CNYRSSSFSD--VETHFRTSH----
Q9HCI8_HUMAN_434-457        YVCQV--CNYRSSSFSD--VETHFRTSH----
Q9NXS0_HUMAN_122-145        YVCQV--CNYRSSSFSD--VETHFRTSH----
ZNF18_HUMAN_1-23            FQCTI--CKKAFLRSSD--FVKHQRT-H----
Q6PIF0_HUMAN_215-237        HVCQY--CEKQFDHFGH--FKEHLRK-H----
Q6PIF0_HUMAN_279-301        YECQV--CNSVFNSWDQ--FKDHLVI-H----
Q9HD72_HUMAN_383-405        YSCPV--CGLRFKRKDR--MSYHVRS-H----
Q9Y2Y4_HUMAN_373-395        YACSV--CGKRFSLKHQ--METHYRV-H----
Y0352_HUMAN_605-627         YSCKV--CGKRFAHTSE--FNYHRRI-H----
O14861_HUMAN_13-35          YQCK--ECGKGFNNNTK--LIQHQRI-H----
Q6PJS0_HUMAN_128-150        YKCK--ECGKGFKYFAS--LDNHMGI-H----
Q5T967_HUMAN_241-263        FKCT--ECGKAFKYKHH--LKEHLRI-H----
Q9BYX9_HUMAN_4-26           YPCM--ICGKKFKSRGF--LKRHMKN-H----
Z297B_HUMAN_400-422         YGCG--VCGKKFKMKHH--LVGHMKI-H----
Q5T7W1_HUMAN_115-137        YECG--ICGKKYKYYNC--FQTHVRA-H----
Q5T7W2_HUMAN_115-137        YECG--ICGKKYKYYNC--FQTHVRA-H----
Q6ZT53_HUMAN_115-137        YECG--ICGKKYKYYNC--FQTHVRA-H----
                              *    *             :
```

```
Q5T7W1_HUMAN_156-178        YTCD--ICGKKYKYYSC--FQEHRDL-H----
Q5T7W2_HUMAN_156-178        YTCD--ICGKKYKYYSC--FQEHRDL-H----
Q6ZT53_HUMAN_156-178        YTCD--ICGKKYKYYSC--FQEHRDL-H----
Q5T7W1_HUMAN_244-266        YTCE--FCGKQYKYYTP--YQEHVAL-H----
Q5T7W2_HUMAN_232-254        YTCE--FCGKQYKYYTP--YQEHVAL-H----
Q6ZT53_HUMAN_224-246        YTCE--FCGKQYKYYTP--YQEHVAL-H----
Q6ZMZ8_HUMAN_41-63          YKCD--QCGYLSKTANK--LIEHVRV-H----
Q8WXE2_HUMAN_217-239        HVCA--ECGKAFVESSK--LKRHFLVH-----
Q96MM3_HUMAN_217-239        HVCA--ECGKAFVESSK--LKRHFLVH-----
Q9NWV2_HUMAN_27-49          YVCN--ICFKHFETPSK--LARHYLIH-----
ZN469_HUMAN_156-178         RDCH--HCGKRFPKPFK--LQRHLAVH-----
ZBT12_HUMAN_359-381         FMCP--RCGKQFNHSSN--LNRHMNVH-----
ZN580_HUMAN_120-142         FTCG--ACGKAFKRSSH--LSRHRATH-----
Q658W5_HUMAN_197-219        YACH--LCGRRFTLRSS--LREHLQIH-----
ZBTB2_HUMAN_254-276         YACH--LCGRRFTLRSS--LREHLQIH-----
ZBT25_HUMAN_236-258         HLCH--YCGERFDSRSN--LRQHLHTH-----
PRD12_HUMAN_243-265         MRCV--ICHRGFNSRSN--LRSHMRIH-----
Q658W5_HUMAN_306-328        YECT--ICGRKFIQKSH--WREHMYIH-----
ZBTB2_HUMAN_363-385         YECT--ICGRKFIQKSH--WREHMYIH-----
ZBT25_HUMAN_347-369         MSCT--ICGHKFPRKSQ--LLEHMYTH-----
Q8WV14_HUMAN_183-205        HVC--TDCGRRFTYPSL--LVSHRR-MH----
SP2_HUMAN_518-542           HVCHIPDCGKTFRKTSL--LRAHVR-LH----
O00110_HUMAN_192-214        HVC--EDCGFTSSRPDT--YAQHRA-LH----
ZFHX2_HUMAN_624-646         HTC--DQCAISFSSQDL--LTSHRR-LH----
ZN580_HUMAN_150-172         HTC--PLCPRRFQDAAE--LAQHVR-LH----
Q68DQ8_HUMAN_993-1016       FPC--RLCERSFCSAPS--LRRHVRVNH----
Q9H937_HUMAN_581-604        FPC--RLCERSFCSAPS--LRRHVRVNH----
Q5VWB8_HUMAN_581-604        FPC--RLCERSFCSAPS--LRRHVRVNH----
ZN592_HUMAN_1013-1036       YPC--RQCEQSFHTPNS--LRKHIRNNH----
Q9UPR8_HUMAN_711-734        YQC--KQCEESFHYKSQ--LRNHEREQH----
Q8NEI5_HUMAN_237-260        YRC--KYCDRSFSISSN--LQRHVRNIH----
PRDM6_HUMAN_514-536         YQC--GHCSQSFSQPSE--LRNHV-VTH----
Q8TBE5_HUMAN_138-161        YKC--ELCSFRCSDRSN--LSHHRRRKH----
Q9H5V7_HUMAN_138-161        YKC--ELCSFRCSDRSN--LSHHRRRKH----
Q9H2T0_HUMAN_138-161        YKC--ELCSFRCSDRSN--LSHHRRRKH----
Q9NZF0_HUMAN_158-181        YIC--KMCPFTTSAKSV--LKKHTEYLH----
ZN644_HUMAN_586-609         YIC--KMCPFTTSAKSV--LKKHTEYLH----
Q7Z5E7_HUMAN_102-124        LKC--KVCLRPFGDPSN--LNKHIR-LH----
Q5MPB1_HUMAN_82-106         YHCTY--CNFSFKTKGN--LTKHMKSKAH---
Q82133_HUMAN_69-93          YHCTY--CNFSFKTKGN--LTKHMKSKAH---
Q99302_HUMAN_31-55          YVCKH--CHFAFKTKGN--LTKHMKSKAH---
ZEP2_HUMAN_1214-1238        YVCKL--CNFAFKTKGN--LTKHMKSKAH---
Q9NV14_HUMAN_229-253        YFCLHFNCNESFKLPFQ--LAQHTKS--H---
ZN644_HUMAN_410-432         YPCTK--CNVNFREKKH--LHRHMMY--H---
Q63HK5_HUMAN_1041-1064      YQCKL--CNRTFASKHA--VKLHLSKT-H---
Q9H0G6_HUMAN_858-881        YQCKL--CNRTFASKHA--VKLHLSKT-H---
Q9P254_HUMAN_1039-1062      YQCKL--CNRTFASKHA--VKLHLSKT-H---
Q6DKI8_HUMAN_89-114         SRCPVSTCNRLFTSKHS--MKAHMVRQ-H---
ZBT12_HUMAN_415-438         YRCSY--CDVRFAHKPA--IRRHLKEQ-H---
Q6ZN29_HUMAN_3-26           YQCKK--CNVVFPRIFD--LITHQKKQ-C---
Q9UBK3_HUMAN_85-108         YRCED--CDQLFESKAE--LADHQKFP-C---
Q5VWB8_HUMAN_415-437        YKCAM--CDTVFTHKPL--LSSHFDQH-----
Q9H937_HUMAN_415-437        YKCAM--CDTVFTHKPL--LSSHFDQH-----
Q68DQ8_HUMAN_827-849        YKCAM--CDTVFTHKPL--LSSHFDQH-----
Q8N8K7_HUMAN_291-313        YRCPE--CDKVFSCPAN--LASHRRWH-----
Q96T92_HUMAN_292-314        YRCPE--CDKVFSCPAN--LASHRRWH-----
                              *    *              :
```

```
Q96Q84_HUMAN_291-313        YRCPE--CDKVFSCPAN--LASHRRWH-----
ZN580_HUMAN_92-114          YSCPE--CARVFASPLR--LQSHRVSH-----
Q96JV0_HUMAN_737-759        YQCKH--CDSKLQSTAE--LTSHLNIH-----
Q96JV0_HUMAN_896-918        YHCSQ--CDRVLMSMQG--LRSHERSH-----
Q6Q7C8_HUMAN_167-189        IRCNI--CNRVFPREKS--LQAHKRTH-----
Q7RTV3_HUMAN_167-189        IRCNI--CNRVFPREKS--LQAHKRTH-----
ZNF26_HUMAN_1-23            YECNE--CEKAYPRKAS--LQIHQKTH-----
Q71UL5_HUMAN_32-54          YQCTI--CTEYCPSLSS--MQKHMKGH-----
Q9Y2Y4_HUMAN_428-450        YRCSL--CGAGCPSLAS--MQAHMRGH-----
Q6ZN29_HUMAN_184-206        YQCDQ--CTVAFPTLEL--WQEHQHVH-----
FOG1_HUMAN_290-314          RVCPFPQCRKSCPSASS--LEIHMRS--H---
FOG2_HUMAN_296-320          SLCPFPQCTKSFSNARA--LEMHLNS--H---
Q6ZMW8_HUMAN_162-185        KRCPF--CRALFKAKSA--LESHIRSR-H---
Q6ZN29_HUMAN_366-389        KRCPF--CRALFKAKSA--LESHIRSR-H---
Q8TBE5_HUMAN_110-132        HRCHL--CPFASAYERH--LEAHMRS--H---
Q9H2T0_HUMAN_110-132        HRCHL--CPFASAYERH--LEAHMRS--H---
Q9H5V7_HUMAN_110-132        HRCHL--CPFASAYERH--LEAHMRS--H---
POGZ_HUMAN_494-516          FRCPH--CTKRLKNNIR--FMNHMKH--H---
Q5SZS4_HUMAN_494-516        FRCPH--CTKRLKNNIR--FMNHMKH--H---
Q5SZS3_HUMAN_439-461        FRCPH--CTKRLKNNIR--FMNHMKH--H---
Q5SZS1_HUMAN_399-421        FRCPH--CTKRLKNNIR--FMNHMKH--H---
Q5SZS2_HUMAN_441-463        FRCPH--CTKRLKNNIR--FMNHMKH--H---
Q9UPR8_HUMAN_683-705        YQCPI--CEHIADNSKD--LESHMIH--H---
POGZ_HUMAN_619-641          LLCPY--CLKVFKNGNA--FQQHYMR--H---
Q5SZS1_HUMAN_524-546        LLCPY--CLKVFKNGNA--FQQHYMR--H---
Q5SZS2_HUMAN_575-597        LLCPY--CLKVFKNGNA--FQQHYMR--H---
Q5SZS3_HUMAN_564-586        LLCPY--CLKVFKNGNA--FQQHYMR--H---
Q5SZS4_HUMAN_619-641        LLCPY--CLKVFKNGNA--FQQHYMR--H---
Q7Z6T2_HUMAN_351-373        LTCSI--CLKEFKNLPA--LNGHMRS--H---
Q7Z6T3_HUMAN_351-373        LTCSI--CLKEFKNLPA--LNGHMRS--H---
Q9NQ73_HUMAN_351-373        LTCSI--CLKEFKNLPA--LNGHMRS--H---
Q9NQ72_HUMAN_351-373        LTCSI--CLKEFKNLPA--LNGHMRS--H---
Q96PN7_HUMAN_512-534        LTCSI--CLKEFKNLPA--LNGHMRS--H---
Q9H0G6_HUMAN_92-116         LKCMY--CGHSFESLQD--LSVHMIKTKH---
Q9P254_HUMAN_273-297        LKCMY--CGHSFESLQD--LSVHMIKTKH---
Q63HK5_HUMAN_275-299        LKCMY--CGHSFESLQD--LSVHMIKTKH---
Q96MY0_HUMAN_186-208        LKCSI--CGHLFSSCSD--LEKHAES--H---
Q9UBK3_HUMAN_139-161        QECKE--CDQVFPDLQS--LEKHMLS--H---
Y0352_HUMAN_508-530         FSCSV--CANSFVDWHL--LEKHMAV--H---
AIOL_HUMAN_202-224          YKCE--FCGRSYKQRSS--LEEHKER-C----
Q69BM1_HUMAN_59-81          YKCE--FCGRSYKQRSS--LEEHKER-C----
Q69BL7_HUMAN_202-224        YKCE--FCGRSYKQRSS--LEEHKER-C----
HELI_HUMAN_196-219          HKCN--YCGRSYKQRSS--LEEHKER-CH---
Q6PQC8_HUMAN_196-219        HKCN--YCGRSYKQRSS--LEEHKER-CH---
Q6PQD0_HUMAN_170-193        HKCN--YCGRSYKQRSS--LEEHKER-CH---
IKAR_HUMAN_201-224          HKCG--YCGRSYKQRSS--LEEHKER-CH---
Q96JP3_HUMAN_208-231        YKCN--YCGRSYKQQST--LEEHKER-CH---
Q9H2S9_HUMAN_146-169        YKCN--YCGRSYKQQST--LEEHKER-CH---
Q6ZN20_HUMAN_244-267        LKCP--KCNWHYKYQQT--LEAHMKE-KH---
ZN409_HUMAN_453-476         LKCP--KCNWHYKYQQT--LDVHMQE-KH---
Q5T967_HUMAN_171-194        LTCP--YCDRGYKRFTS--LKEHIKY-RH---
Q8IUL5_HUMAN_282-304        FKCP--YCTHVVKRKAD--LKRHLR--CH---
Q8N2Y5_HUMAN_92-114         FKCP--YCTHVVKRKAD--LKRHLR--CH---
Q96BR9_HUMAN_282-304        FKCP--YCTHVVKRKAD--LKRHLR--CH---
Q96BX0_HUMAN_77-99          FKCP--YCTHVVKRKAD--LKRHLR--CH---
Q9H9H3_HUMAN_97-119         LKCP--HCSYVAKYRRT--LKRHLL--IH---
                              *    *             :
```

```
ZBT10_HUMAN_698-720          LKCP--HCSYVAKYRRT--LKRHLL--IH---
ZBTB8_HUMAN_341-363          HKCP--FCPYTAKQKGI--LKRHIR--SH---
ZFHX2_HUMAN_103-127          FKCT--VCRVSYNQSST--LEIHMRSVLH---
ZFHX2_HUMAN_46-70            YKCT--VCKESFTQKNI--LLVHYNSVSH---
PRD12_HUMAN_299-323          YKCQ--VCQSAYSQLAG--LRAHQKSARH---
Q6ZN20_HUMAN_299-323         FRCE--VCNYSTTTKGN--LSIHMQSDKH---
ZN409_HUMAN_508-532          YRCD--VCNYSTTTKGN--LSIHMQSDKH---
Q6ZN20_HUMAN_367-391         WRCE--VCDYETNVARN--LRIHMTSEKH---
Q9BYX9_HUMAN_35-57           YRCT--DCDYTTNKKIS--LHNHLES--H---
Q6ZS14_HUMAN_600-623         YKCT--MCNYSTTTLKG--LRVHQQH-KH---
Q9UPR8_HUMAN_655-677         YRCR--LCHYTSGN-KG--YIKQHLR-VH---
Q5T1J7_HUMAN_442-465         VFCN--ACGK-TFYDKG-TLKIHYNA-VH---
Q8NAR2_HUMAN_13-36           VFCN--ACGK-TFYDKG-TLKIHYNA-VH---
Q6T3A3_HUMAN_363-386         VFCN--ACGK-TFYDKG-TLKIHYNA-VH---
Q6ZN30_HUMAN_441-464         VFCN--ACGK-TFYDKG-TLKIHYNA-VH---
BNC1_HUMAN_357-380           VFCT--ACEK-TFYDKG-TLKIHYNA-VH---
Q6T3A3_HUMAN_957-980         IMCN--ICHK-MYSNKG-TLRVHYKT-VH---
Q9NXV0_HUMAN_325-348         IMCN--ICHK-MYSNKG-TLRVHYKT-VH---
Q9H6J0_HUMAN_251-274         IMCN--ICHK-MYSNKG-TLRVHYKT-VH---
Q6ZN30_HUMAN_1035-1058       IMCN--ICHK-MYSNKG-TLRVHYKT-VH---
BNC1_HUMAN_928-951           ITCH--LCQK-TYSNKG-TFRAHYKT-VH---
Q5T1J7_HUMAN_834-857         KICY--VCKK-SFKSSY-SVKLHYRN-VH---
Q9NXV0_HUMAN_123-146         KICY--VCKK-SFKSSY-SVKLHYRN-VH---
Q6T3A3_HUMAN_755-778         KICY--VCKK-SFKSSY-SVKLHYRN-VH---
Q6ZN30_HUMAN_833-856         KICY--VCKK-SFKSSY-SVKLHYRN-VH---
Q8NAR2_HUMAN_405-428         KICY--VCKK-SFKSSY-SVKLHYRN-VH---
Q9H6J0_HUMAN_49-72           KICY--VCKK-SFKSSY-SVKLHYRN-VH---
BNC1_HUMAN_720-743           FQCD--ICKK-TFKNAC-SVKIHHKN-MH---
Q9C0D4_HUMAN_136-159         YQCE--YCDYGAIRNDY-IVK-HTKR-VH---
OVOL1_HUMAN_174-197          YKCS--LCDK-AFTQRC-SLESHLKK-IH---
Q86XL8_HUMAN_151-174         YKCS--LCDK-AFTQRC-SLESHLKK-IH---
O00110_HUMAN_153-176         FRCS--ACGK-AFTQRC-SLEAHLAK-VH---
Q7M4M1_HUMAN_36-59           YKCK--DCGN-AFIWRA-SLQYHVKK-VH---
Q7Z5E7_HUMAN_131-154         YRCE--FCGK-VLVRRR-DLERHVKS-RH---
Q96MY0_HUMAN_528-551         CACT--DCGQ-VATNRT-DLEIHVKR-CH---
FOG1_HUMAN_241-264           FPCK--DCGIWYRSERN--LQAHLLY-YC---
ZN592_HUMAN_1153-1176        AQCL--LCGLCYTSASS--LSRHLFI-VH---
Q86XK6_HUMAN_120-143         FTCH--HCGKQLRSLAG--MKYHVMA-NH---
ZN512_HUMAN_197-220          FTCH--HCGKQLRSLAG--MKYHVMA-NH---
Q86XK6_HUMAN_210-233         LKCH--HCGKPYRSKAG--LAYHLRS-EH---
ZN512_HUMAN_287-310          LKCH--HCGKPYRSKAG--LAYHLRS-EH---
Q86XK6_HUMAN_363-386         YKCL--LCQKEFVSESG--VKYHINS-VH---
ZN512_HUMAN_440-463          YKCL--LCQKEFVSESG--VKYHINS-VH---
Q8N8K7_HUMAN_525-548         FSCK--HCPSTFFSSPG--LTRHINK-CH---
Q96Q84_HUMAN_525-548         FSCK--HCPSTFFSSPG--LTRHINK-CH---
Q96T92_HUMAN_526-549         FSCK--HCPSTFFSSPG--LTRHINK-CH---
Q96MY0_HUMAN_705-728         FQCK--KCFYKTRSSTV--LTRHIKL-RH---
Q8NDX6_HUMAN_101-123         FVCE--HCFGAFRSSYH--LKRHIL--IH---
Q96JV0_HUMAN_959-982         FRCK--LC--SFKSSYNSRLKTHILK-AH---
                               *    *              :
```

Appendix B – Consensus Output

Target GD-AA : Template specified/identified is 1AAY Chain A

Columns 1-2 correspond to the target.
Columns 3-4 correspond to the template.
Column 5 is the confidence (0-9) for that pair of aligned residues according to sequence alignments
Column 6 is the selection status.

An 'S' indicates that the corresponding pair of residues has been confidently aligned and may be used in homology modeling.

A '.' indicates otherwise.

Please disregard column 7. This is only for diagnostic purposes and will soon be removed.

--------- GD-AA.consensus ---------

R 3	R 1	9	.	H
P 4	P 2	9	.	
Y 5	Y 3	9	.	
A 6	A 4	9	.	
C 7	C 5	9	.	
P 8	P 6	9	.	
V 9	V 7	9	.	
E 10	E 8	6	.	
S 11	S 9	7	S	D
C 12	C 10	9	S	D
D 13	D 11	9	S	D
R 14	R 12	9	S	D
R 15	R 13	9	S	D
F 16	F 14	9	S	D
S 17	S 15	9	S	D
Q 18	R 16	9	S	D
K 19	S 17	9	S	D
T 20	D 18	9	S	D
N 21	E 19	9	S	A
L 22	L 20	9	S	A
D 23	T 21	9	S	E
T 24	R 22	9	S	E
H 25	H 23	9	S	AE
I 26	I 24	9	S	AE
R 27	R 25	9	S	AE
I 28	I 26	9	S	E
H 29	H 27	9	S	E
T 30	T 28	9	S	E
G 31	G 29	9	S	E
Q 32	Q 30	9	S	E
K 33	K 31	9	S	E
P 34	P 32	9	S	E
F 35	F 33	9	S	E
Q 36	Q 34	9	S	E
C 37	C 35	9	S	E
R 38	R 36	9	S	E
I 39	I 37	9	S	E
C 40	C 38	9	S	E

M 41	M 39	9	S	E
R 42	R 40	9	S	E
N 43	N 41	9	S	E
F 44	F 42	9	S	AE
S 45	S 43	9	S	AE
Q 46	R 44	9	S	E
H 47	S 45	9	S	E
T 48	D 46	9	S	E
G 49	H 47	9	.	G
L 50	L 48	9	.	DG
N 51	T 49	9	.	DG
Q 52	T 50	9	.	AG
H 53	H 51	9	.	AG
I 54	I 52	9	.	EG
R 55	R 53	9	.	EG
T 56	T 54	9	.	EG
H 57	H 55	9	.	EG
T 58	T 56	9	.	EG
G 59	G 57	9	.	EG
E 60	E 58	9	.	EG
K 61	K 59	9	.	EG
P 62	P 60	9	.	EG
F 63	F 61	9	.	EG
A 64	A 62	9	.	EG
C 65	C 63	9	.	EG
D 66	D 64	9	.	EG
I 67	I 65	9	.	EG
C 68	C 66	9	.	EG
G 69	G 67	9	.	EG
R 70	R 68	9	.	EG
K 71	K 69	9	.	EG
F 72	F 70	9	.	AEG
A 73	A 71	9	.	AEG
T 74	R 72	9	.	EG
L 75	S 73	9	.	EG
H 76	D 74	9	.	EG
T 77	E 75	9	.	EG
R 78	R 76	9	.	EGI
D 79	K 77	9	.	EGI
R 80	R 78	9	.	EGI
H 81	H 79	9	.	EGI
T 82	T 80	9	.	EGI
K 83	K 81	9	.	EGI
I 84	I 82	9	.	EGI
H 85	H 83	9	.	EGI
L 86	L 84	9	.	EGI
R 87	R 85	9	.	EGI

Target 1AAY : Template specified/identified is 1G2D Chain C

Columns 1-2 correspond to the target.
Columns 3-4 correspond to the template.
Column 5 is the confidence (0-9) for that pair of aligned residues according to sequence alignments
Column 6 is the selection status.

An 'S' indicates that the corresponding pair of residues has been confidently aligned and may be used in homology modeling.

A '.' indicates otherwise.

Please disregard column 7. This is only for diagnostic purposes and will soon be removed.

--------- AA-GD.consensus ---------

M 1	M 1	5	.	H
E 2	E 2	7	.	DH
R 3	R 3	9	.	DH
P 4	P 4	9	S	D
Y 5	Y 5	9	S	D
A 6	A 6	9	S	D
C 7	C 7	9	S	D
P 8	P 8	9	S	D
V 9	V 9	9	S	D
E 10	E 10	9	S	D
S 11	S 11	9	S	D
C 12	C 12	9	S	D
D 13	D 13	9	S	D
R 14	R 14	9	S	D
R 15	R 15	9	S	D
F 16	F 16	9	S	D
S 17	S 17	9	S	D
R 18	Q 18	9	S	D
S 19	K 19	9	S	D
D 20	T 20	9	S	D
E 21	N 21	9	S	D
L 22	L 22	9	S	D
T 23	D 23	9	S	D
R 24	T 24	9	S	A
H 25	H 25	9	S	A
I 26	I 26	9	S	A
R 27	R 27	9	S	A
I 28	I 28	9	S	E
H 29	H 29	9	S	E
T 30	T 30	9	S	E
G 31	G 31	9	S	E
Q 32	Q 32	9	S	E
K 33	K 33	9	S	E
P 34	P 34	9	S	E
F 35	F 35	9	S	E
Q 36	Q 36	9	S	E
C 37	C 37	9	S	E
R 38	R 38	9	S	E
I 39	I 39	9	S	E
C 40	C 40	9	S	E
M 41	M 41	9	S	E

R 42	R 42	9	S	E
N 43	N 43	9	S	E
F 44	F 44	9	S	AE
S 45	S 45	9	S	AE
R 46	Q 46	9	S	E
S 47	H 47	9	S	E
D 48	T 48	9	S	E
H 49	G 49	9	S	A
L 50	L 50	9	S	A
T 51	N 51	9	S	E
T 52	Q 52	9	S	E
H 53	H 53	9	S	E
I 54	I 54	9	S	E
R 55	R 55	9	S	E
T 56	T 56	9	S	E
H 57	H 57	9	S	E
T 58	T 58	9	S	E
G 59	G 59	9	S	E
E 60	E 60	9	S	E
K 61	K 61	9	S	E
P 62	P 62	9	S	E
F 63	F 63	9	S	E
A 64	A 64	9	S	E
C 65	C 65	9	S	E
D 66	D 66	9	S	E
I 67	I 67	9	S	E
C 68	C 68	9	S	E
G 69	G 69	9	S	E
R 70	R 70	9	S	E
K 71	K 71	9	S	E
F 72	F 72	9	S	AE
A 73	A 73	9	S	AE
R 74	T 74	9	S	AE
S 75	L 75	9	S	E
D 76	H 76	9	S	E
E 77	T 77	9	S	E
R 78	R 78	9	S	E
K 79	D 79	9	.	EI
R 80	R 80	9	.	EI
H 81	H 81	9	.	AEI
T 82	T 82	9	.	AEI
K 83	K 83	9	.	EI
I 84	I 84	9	.	EI
H 85	H 85	9	.	EI
L 86	L 86	9	.	EI
R 87	R 87	9	.	EI
Q 88	Q 88	9	.	EI
K 89	K 89	9	.	EI

Target ME-AA : Template specified/identified is 1AAY Chain A

Columns 1-2 correspond to the target.
Columns 3-4 correspond to the template.
Column 5 is the confidence (0-9) for that pair of aligned residues according to sequence alignments
Column 6 is the selection status.

An 'S' indicates that the corresponding pair of residues has been confidently aligned and may be used in homology modeling.

A '.' indicates otherwise.

Please disregard column 7. This is only for diagnostic purposes and will soon be removed.

--------- ME-AA.consensus ---------

K 3	R 1	9	.	F
P 4	P 2	9	.	F
Y 5	Y 3	9	.	F
K 6	A 4	9	.	F
C 7	C 5	9	.	F
P 8	P 6	9	.	F
E 9	V 7	7	.	F
C 10	C 10	9	.	F
G 11	D 11	9	.	F
K 12	R 12	9	S	D
S 13	R 13	9	S	D
F 14	F 14	9	S	D
S 15	S 15	9	S	D
Q 16	R 16	9	S	D
S 17	S 17	9	S	D
S 18	D 18	9	S	D
N 19	E 19	9	S	A
L 20	L 20	9	S	A
Q 21	T 21	9	S	E
K 22	R 22	9	S	E
H 23	H 23	9	S	AE
Q 24	I 24	9	S	AE
R 25	R 25	9	S	AE
T 26	I 26	9	S	E
H 27	H 27	9	S	E
T 28	T 28	9	S	E
G 29	G 29	9	S	E
E 30	Q 30	9	S	E
K 31	K 31	9	S	E
P 32	P 32	9	S	E
Y 33	F 33	9	S	E
K 34	Q 34	9	S	E
C 35	C 35	9	S	E
P 36	R 36	9	S	E
E 37	I 37	9	S	E
C 38	C 38	9	S	E
G 39	M 39	9	.	
K 40	R 40	9	S	D
S 41	N 41	9	S	D
F 42	F 42	9	S	A

S 43	S 43	9	S	A
Q 44	R 44	9	S	E
S 45	S 45	9	S	E
S 46	D 46	9	S	E
D 47	H 47	9	S	E
L 48	L 48	9	S	E
Q 49	T 49	9	S	E
K 50	T 50	9	S	AE
H 51	H 51	9	S	AE
Q 52	I 52	9	S	E
R 53	R 53	9	S	E
T 54	T 54	9	S	E
H 55	H 55	9	S	E
T 56	T 56	9	S	E
G 57	G 57	9	S	E
E 58	E 58	9	S	E
K 59	K 59	9	S	E
P 60	P 60	9	S	E
Y 61	F 61	9	S	E
K 62	A 62	9	S	E
C 63	C 63	9	S	E
P 64	D 64	9	S	E
E 65	I 65	9	S	E
C 66	C 66	9	S	E
G 67	G 67	9	S	E
K 68	R 68	9	S	E
S 69	K 69	9	S	E
F 70	F 70	9	S	AE
S 71	A 71	9	S	AE
R 72	R 72	9	S	E
S 73	S 73	9	S	E
D 74	D 74	9	S	E
H 75	E 75	9	S	E
L 76	R 76	9	.	EI
S 77	K 77	9	.	EI
R 78	R 78	9	.	EI
H 79	H 79	9	.	EI
Q 80	T 80	9	.	EI
R 81	K 81	9	.	EI
T 82	I 82	9	.	EI
H 83	H 83	9	.	EI
Q 84	L 84	9	.	EI
N 85	R 85	9	.	EI

Appendix C – Domain Information for Each Homology Modeled EGR Protein

Areas of overlap are underlined.

AA-GD
RPYACPVESCDRRFSRSDELTRHIRIHTGQKPFQCRICMRNFSRSDHLTTHIRTHTGEKPFA
CDICGRKFARSDERKRHTKIHLR

AA-GD Domain 1 (Residues 1 – 32)
RPYACPVESCDRRFSRSDELTRHIRIHTGQKP

AA-GD Domain 2 (Residues 30 – 60)
QKPFQCRICMRNFSRSDHLTTHIRTHTGEKP

AA-GD Domain 3 (Residues 57 – 85)
EKPFACDICGRKFARSDERKRHTKIHLR

GD-AA
MERPYACPVESCDRRFSQKTNLDTHIRIHTGQKPFQCRICMRNFSQHTGLNQHIRTHTGE
KPFACDICGRKFATLHTRDRHTKIHLRQK

GD-AA Domain 1 (Residues 1 – 34)
MERPYACPVESCDRRFSQKTNLDTHIRIHTGQKP

GD-AA Domain 2 (Residues 32 – 61)
QKPFQCRICMRNFSQHTGLNQHIRTHTGEK

GD-AA Domain 3 (Residues 60 – 89)
EKPFACDICGRKFATLHTRDRHTKIHLRQK

ME-AA
KPYKCPECGKSFSQSSNLQKHQRTHTGEKPYKCPECGKSFSQSSDLQKHQRTHTGEKPY
KCPECGKSFSRSDHLSRHQRTHQ

GD-AA Domain 1 (Residues 1 – 30)
KPYKCPECGKSFSQSSNLQKHQRTHTGEKP

ME-AA Domain 2 (Residues 28 – 58)
EKPYKCPECGKSFSQSSDLQKHQRTHTGEKP

ME-AA Domain 3 (Residues 56 – 82)
EKPYKCPECGKSFSRSDHLSRHQRTHQ

Appendix D – RMS versus Crystal Structure for Clustered Models

AA - GD

Residue Number	Residue Type	Time (ps)	RMS	Residue Number	Residue Type	Time (ps)	RMS
1	R	870	0.77	44	R	2970	2.3
2	P	485	0.82	45	S	1560	2.05
3	Y	2752.5	1.45	46	D	1557.5	1.27
4	A	3470	0.63	47	H	1087.5	0.82
5	C	1095	0.94	48	L	1037.5	0.47
6	P	1767.5	1.67	49	T	385	0.88
7	V	1555	1.01	50	T	837.5	0.67
8	E	1625	2.61	51	H	3847.5	1.52
9	S	1542.5	2.95	52	I	2415	0.57
10	C	280	2.05	53	R	1245	0.66
11	D	3592.5	2.48	54	T	1817.5	1.04
12	R	3400	3.04	55	H	2265	1.7
13	R	1600	0.66	56	T	1237.5	0.39
14	F	3425	1.72	57	G	317.5	0.29
15	S	1517.5	0.95	58	E	2300	1.74
16	R	3102.5	1.13	59	K	3112.5	0.65
17	S	2672.5	0.64	60	P	3182.5	0.41
18	D	1442.5	2.06	61	F	795	1.46
19	E	237.5	0.97	62	A	317.5	0.14
20	L	1115	0.57	63	C	2592.5	0.87
21	T	1090	1.18	64	D	1055	2.44
22	R	910	2.43	65	I	2675	1.09
23	H	882.5	1.77	66	C	1410	0.84
24	I	3047.5	1.85	67	G	3040	0.25
25	R	2610	1.76	68	R	830	3.68
26	I	1897.5	2.13	69	K	1265	1.23
27	H	2850	1.27	70	F	3562.5	1.5
28	T	837.5	335	71	A	2205	0.31
29	G	3010	0.87	72	R	3402.5	1.39
30	Q	3132.5	2.65	73	S	3367.5	0.33
31	K	3177.5	3.53	74	D	2490	1.91
32	P	2622.5	0.58	75	E	2592.5	1.07
33	F	3887.5	1.61	76	R	3342.5	2.14
34	Q	607.5	0.77	77	K	860	2.07
35	C	3927.5	0.9	78	R	1415	1.55
36	R	2102.5	4.02	79	H	1117.5	1.25
37	I	2365	0.94	80	T	1515	1.0
38	C	2437.5	0.49	81	K	3647.5	2.27
39	M	2717.5	0.51	82	I	3400	0.51
40	R	547.5	4.9	83	H	655	2.84
41	N	2235	0.49	84	L	272.5	1.17
42	F	200	0.46	85	R	1062.5	2.55
43	S	1745	0.78				

GD-AA

Residue Number	Residue Type	Time (ps)	RMS	Residue Number	Residue Type	Time (ps)	RMS
1	M	2410	1.3	45	S	525	0.84
2	E	460	2.74	46	Q	1120	0.97
3	R	3970	2.27	47	H	2235	0.87
4	P	2622.5	0.61	48	T	1080	1.16
5	Y	3580	1.49	49	G	2622.5	0.25
6	A	317.5	0.25	50	L	3827.5	0.6
7	C	2432.5	0.45	51	N	305	1.02
8	P	3597.5	1439.5	52	Q	1352.5	0.84
9	V	2622.5	0.41	53	H	1320	1.23
10	E	3105	2.73	54	I	1082.5	1.59
11	S	3760	0.98	55	R	3207.5	2.21
12	C	325	1.26	56	T	3087.5	2.07
13	D	890	1.47	57	H	3302.5	1.18
14	R	2645	1.92	58	T	1805	0.32
15	R	1107.5	1.63	59	G	1465	0.16
16	F	525	1.47	60	E	1200	1.52
17	S	1030	0.68	61	K	640	0.61
18	Q	377.5	0.86	62	P	3467.5	0.48
19	K	1097.5	1.34	63	F	2010	1.57
20	T	317.5	1.22	64	A	2622.5	0.49
21	N	3382.5	1.14	65	C	485	0.95
22	L	940	0.28	66	D	2325	0.92
23	D	3225	1.91	67	I	992.5	1.12
24	T	1680	0.75	68	C	2010	0.49
25	H	2105	0.64	69	G	2683.5	0.54
26	I	3972.5	1.54	70	R	2620	1.04
27	R	1837.5	1.13	71	K	1622.5	1.71
28	I	1350	0.68	72	F	1222.5	1.56
29	H	2285	1.24	73	A	2622.5	0.15
30	T	3507.5	0.29	74	T	2927.5	0.39
31	G	3207.5	0.15	75	L	3757.5	0.7
32	Q	1377.5	0.83	76	H	3087.5	1.36
33	K	2982.5	0.32	77	T	2820	0.71
34	P	317.5	0.32	78	R	1817.5	1.61
35	F	3660	1.56	79	D	862.5	1.96
36	Q	2680	0.38	80	R	1287.5	3.62
37	C	1820	1.08	81	H	1072.5	0.99
38	R	3945	2.19	82	T	3252.5	0.91
39	I	2137.5	0.87	83	K	2417.5	1.72
40	C	815	0.91	84	I	1045	0.65
41	M	2342.5	0.79	85	H	1665	2.8
42	R	2140	5.41	86	L	2012.5	1.07
43	N	2625	1.12	87	R	2017.5	3.46
44	F	3955	0.7	88	Q	2225	0.93

ME – AA

Residue Number	Residue Type	Time (ps)	RMS	Residue Number	Residue Type	Time (ps)	RMS
1	K	3617.5	1.78	42	Q	1320	0.71
2	P	3702.5	0.55	43	S	1707.5	0.33
3	Y	1537.5	1.71	44	S	417.5	0.51
4	K	2650	0.94	45	D	3735	1.07
5	C	2385	0.5	46	L	2090	0.5
6	P	392.5	0.57	47	Q	1307.5	1.97
7	E	1435	2.12	48	K	957.5	2.4
8	C	3412.5	1.25	49	H	1412.5	0.79
9	G	532.5	0.67	50	Q	1592.5	0.76
10	K	1640	0.59	51	R	1710	2.28
11	S	992.5	0.75	52	T	2560	0.46
12	F	1135	1.45	53	H	1665	2.22
13	S	322.5	0.82	54	T	2622.5	0.56
14	Q	735	1.18	55	G	860	0.39
15	S	600	0.48	56	E	1752.5	2.14
16	S	1665	0.66	57	K	2640	2.56
17	N	2420	1.11	58	P	1732.5	0.35
18	L	3112.5	0.46	59	Y	3780	1.49
19	Q	2910	0.76	60	K	1090	1.86
20	K	3075	2.21	61	C	1707.5	0.6
21	H	2597.5	0.66	62	P	2187.5	0.61
22	Q	2297.5	1.04	63	E	3995	2.49
23	R	3812.5	2.01	64	C	2287.5	0.98
24	T	3247.5	0.46	65	G	3007.5	0.25
25	H	1170	2.3	66	K	817.5	0.71
26	T	1995	0.37	67	S	620	0.47
27	G	482.5	0.21	68	F	3277.5	1.46
28	E	3640	1.98	69	S	2882.5	0.45
29	K	2115	1.1	70	R	2680	1.98
30	P	1400	0.47	71	S	1120	0.44
31	Y	1105	0.66	72	D	2455	0.55
32	K	402.5	3.06	73	H	2262.5	2.08
33	C	3605	0.64	74	L	3957.5	0.51
34	P	1960	0.42	75	S	2935	0.84
35	E	2465	2.86	76	R	2005	4.51
36	C	1627.5	0.62	77	H	1265	0.93
37	G	2260	0.22	78	Q	2327.5	0.69
38	K	2097.5	1.81	79	R	2650	4.92
39	S	2397.5	0.9	80	T	1457.5	1.52
40	F	2985	0.95	81	H	660	1.88
41	S	922.5	0.79				