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, particularly research being done 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. All members of the EGR family contain three zinc finger domains which are used in conjunction with one another to recognize specific DNA sequences (Benos et al). 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 (Paillard et al).

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 pattern of: x-C-x(1-5)-C-x(12)-H-x(3-6)-H (Hulo et al). 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 through the four conserved residues working together to coordinate a zinc atom along with a conserved aromatic ring located 4 amino acids away from the second cysteine residue. Figure 1 shows the zf-C2H2 domain in a two-dimensional manner, as well as the details into how the four conserved residues coordinate the zinc atom. 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 about 5 nucleic acids. The two conserved histidine residues responsible for coordinating the zinc atom are located within the α-helix pointing away from the DNA, as can be seen in Figure 2.

The proteins that include the classic zinc finger domain are quite diverse. According to Pfam, version 17.0, there are a total of 32,784 different protein sequences that have 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 five recognized human EGR proteins; however there may be an additional 26 human proteins that could potentially belong to the EGR family. The potential proteins were determined through comparison against known EGR family members on the basis of sequence length and proximity of the zf-C2H2 domains. The 97 remaining human sequences do not exhibit significant similarity to the EGR family of proteins, and are excluded from our inquiry.

**Figure 1. 2-D Pattern of zf-C2H2 domain.** The four conserved residues that coordinate the zinc ion are shown in the center of the domain. The green ‘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 (PDOC00028).
Although proteins with zinc finger domains are highly studied, there are relatively few structures available for the amount of sequences that contain the domain. There are a total of 43 structures in the PDB with resolved zinc finger domains. 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 structure is from humans and the remaining are from *Mus musculus*. Four of the mouse structures are either variants of the wild-type EGR family proteins. With the lack of structure data that exists 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 testing the structures with ClusPro to determine the specific protein – DNA interactions.

**Methodology**

The first steps involved in this project are to determine the proper structure(s) to use as a template to perform the homology modeling for the sequence being determined. With so few predetermined structures, a combination of multiple structures may be needed to determine an accurate structure. Bioinformatics analysis of human EGR will be performed, including BLASTp analysis of query sequences against the PDB database, and using those results to perform multiple sequence alignments to determine the type of homology modeling to be preformed. An overview of the methodology for the project can be seen in Figure 3.

Predicting the structure for human EGR proteins will be done using one of two different methods depending on the percentage of sequence identity with known structure. The first method is for proteins that have a sequence identity greater than 30%, indicating homology with known structures. Such sequences will be run through the program Consensus to create a

---

**Figure 3. Flow Diagram of the Methodology.** After analysis of the human EGR proteins through bioinformatics, a suitable structure for the template strand needs to be determined. The percentage of identity between the template structure and the query strand determines the type of modeling to be performed, through either homology modeling or secondary structure prediction. The determined structures will then be analyzed with ClusPro to predict the determined protein – DNA interactions.
homology model of the query sequence based off of the predetermined template structure. 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. Furthermore, Consensus selectively removes regions of dissimilarity, and splits loosely connected domains to minimize potential misalignments (Prasad et al). The second method is for proteins that are less than 30% similar to known structures. Since we are mainly interested in the protein – DNA interactions, secondary structure prediction will be performed with the program SABLE (Adamczak et al). We are looking to properly identify the α-helical regions of the protein, since they directly involved in determining the protein – DNA interactions.

The final step involves the use of the program ClusPro, initially developed by Comeau, Gatchell, Vajda and Camacho for predicting interactions between proteins. 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 (Comeau et al). Since this method uses two unbound structures it can also be applied to predict protein – DNA interactions. We are specifically looking for recognition of DNA sequences that are specific for the EGR family.

**Potential Results**

The anticipated results for the homology modeling portion of this project are structures that align well with other known structures of the family, but are still diverse enough to potentially recognize different binding sites on the DNA. Theoretically, the protein – DNA interactions produced should be similar to the sequence produced from the SAMIE algorithm developed by Benos et al, which uses a probabilistic recognition code to determine the DNA recognition sequence from the amino acids in the protein.
References


