PROTEIN NUCLEIC ACID INTERACTIONS
GRANT # DE-FG02-96ER62166

FINAL REPORT

Helen M. Berman and Janet Thornton
(1999-2004)

1. Introduction

The overall goal of this collaborative project is to develop methods for analyzing protein-nucleic acid interactions.

Nucleic acid-binding proteins have a central role in all aspects of genetic activity within an organism, such as transcription, replication, and repair. Thus, it is extremely important to examine the nature of complexes that are formed between proteins and nucleic acids, as they form the basis of our understanding of how these processes take place. Over the past decade, the world has witnessed a great expansion in the determination of high-quality structures of nucleic acid-binding proteins. As a result, the number of such structures has seen a constant increase in the Protein Data Bank (PDB) (1) and the Nucleic Acid Database (NDB) (2). These structures, especially those of proteins in complex with DNA, have provided valuable insight into the stereochemical principles of binding, including how particular base sequences are recognized and how the nucleic acid structure is quite often modified on binding.

In this project, we designed several approaches to characterize and classify the properties of both protein-DNA and protein-RNA complexes.

In work done in the previous grant period, we developed methods to use experimental data to evaluate nucleic acid crystal structures in order to ensure that the structures utilized in future studies would be of high quality. The methodology was collated in the standalone software package SFCHECK (3) [A], and an applied survey of structures in the NDB produced very positive results. With this quality control mechanism in place, we then analyzed DNA-binding sites on proteins by studying the distortions observed in DNA structures bound to protein. From our observations, we found that DNA-binding proteins present a very different binding surface to those that bind other proteins and defined three modes of protein binding [B]. Following this survey, we classified DNA-binding proteins into eight different structural/functional groups [C]. This classification highlighted the diversity of protein-DNA complex geometries found in nature and emphasized the importance of interactions between alpha helices and the major groove–the main bind partner with DNA in roughly half of the protein families under study.

These studies gave us the insight to seed our future work in the current project described here, as we observed the repeated presence of the helix-turn-helix (HTH) and zinc-coordinating motifs, and how they present an alpha helix on the surfaces of structurally diverse proteins ready for interaction with DNA [D, E, F, G, H]. Structure-based methods for predicting DNA-binding included scanning of 3D structural templates, use of the electrostatic potential to select generic DNA-binding residue patches, and a statistical model based on geometrical measures such as the recognition helix/second helix hydrophobic interaction area of the HTH motif. A recent study worked to incorporate structural data into a sequence-based method of motif detection. Another, more general, study at Rutgers developed a classification model to annotate DNA-binding proteins based on their three-dimensional (3D) structural features.

As a necessary step to further understanding protein RNA interactions [I] we developed new visualization tools and methods to classify RNA conformation [J,K].

The tools developed are available for use by the scientific community through the NDB.

These projects were a joint effort between the Berman and Thornton groups.

2. Structural Analysis of Protein-DNA interactions

2.1 DNA-Binding Motif Prediction Based on Geometric Statistics

"Statistical models for discerning protein structures containing the DNA-binding helix-turn-helix motif", McLaughlin and Berman, 2003 [D]

In this work, a method for discerning protein structures containing the DNA-binding HTH motif was developed using statistical models based on geometrical measurements of the motif. With a decision tree model, key structural features required for DNA-binding were identified. These features included a high average solvent-accessibility of residues within the recognition helix and a conserved hydrophobic interaction between the recognition helix and the second alpha helix preceding it. The PDB was searched using a more accurate model of the motif to identify structures that have a high probability of containing the motif, including those that had not been reported previously.

A descriptive model and a predictive model of the DNA-binding HTH motif were created. The descriptive model was used to characterize the motif by elucidating its conserved structural features. Knowledge of these conserved features led to further understanding of the structural requirements of DNA-binding and provided an intuitive means of manually classifying the motif.

The predictive model was used to scan through the entire PDB and identify possible candidates for the motif. The survey identified known HTH proteins with high probability, as well as several previously undocumented proteins likely to possess the motif. Because the model was based on geometrical measurements and not on measurements used to compare primary amino acid sequence, it was able to identify candidates for the motif that have no detected sequence homology to the sequences of the structures used to create the model. By doing so, the model acted to supplement sequence-based methods used to identify the motif and identified previously unrecognized candidates for the motif.

2.2 DNA-Binding Protein Prediction Based on Alpha Helical Structure

“A structure-based method for identifying DNA-binding proteins and their sites of DNA-interaction”, McLaughlin et al., 2004 [E]

A general classification model of a DNA-binding protein chain was created based on identification of alpha helices within the chain likely to bind DNA. Using the model, all chains in the PDB were classified. For many of the chains classified with high confidence, previous documentation for DNA-binding was found, yet no sequence homology to the structures used to train the model was detected. The DNA-binding chain classification model is useful for protein chains that bind to DNA via an alpha helix. The utility of the chain model is two-fold: It can be used to speed the process of identifying the function of a new protein as DNA-binding when no sequence homology to documented DNA-binding is detected. Also, it can identify possible sites of DNA-interaction by listing the alpha helices likely to interact with DNA.

The DNA-binding alpha helix classification model was created using a method similar to that described for the recognition helix of a DNA-binding HTH motif (D). The dataset of DNA-binding alpha helices used to create the model included all alpha helices known to interact with DNA based on solved protein-DNA complex structures in the PDB. In addition, a classification model of a DNA-binding protein chain was developed. The chain model considers the number of alpha helices classified as DNA-binding and their associated classification confidence. The chain model was used to search through the entire PDB to identify possible DNA-binding structures.

From this study, four new candidates for DNA-binding were found, including two structures solved through structural genomics efforts. For each of the candidate structures, possible sites of DNA-binding were indicated by listing the residue ranges of alpha helices likely to interact with DNA. This result indicates that the chain model can be used to supplement sequence-based methods for annotating the function of DNA-binding.

This study provided the stimulus to develop a web interface, qprof.rutgers.edu, that allows access to the database of the measurements used in this study and is a tool for making queries for structures having a user-defined set of structural measurements. The tool, Query of PROtein Features (QPROF), allows the user to create a list of geometrical measurement criteria and search for protein structures that satisfy those criteria. The searchable geometrical measurements are based on a set of secondary structural elements (SSEs) that include a reference element (either an alpha helix or beta strand) and topologically adjacent elements, i.e. elements that come before and after the reference element in the protein chain. Future classifications can be done through the web interface dna-binders.rutgers.edu.

2.3 DNA-Binding Prediction Using Structural Motif Templates

“Using structural motif templates to identify proteins with DNA binding function”, Jones et al., 2003 [F]

This work developed a method for predicting DNA binding function from structure using 3D templates. A structural template library of seven HTH motifs was created from non-homologous DNA-binding proteins in the PDB. The templates were used to scan complete protein structures using an algorithm that calculated the root mean squared deviation (rmsd) for the optimal superposition of each template on each structure, based on alpha-carbon backbone coordinates. Distributions of rmsd values for known HTH-containing proteins (true hits) and non-HTH proteins (false hits) were calculated. A threshold value of 1.6 Å rmsd was selected that gave a true hit rate of 88.4% and a false positive rate of 0.7%. The false positive rate was further reduced to 0.5% by introducing an accessible surface area threshold value of 990 Å^2 per HTH motif. The template library and the validated thresholds were used to make predictions for target proteins from a structural genomics project.

The importance of using structural templates lies in their ability to identify HTH motifs in structures from more than one homologous (fold) family. The templates can match HTH motifs from different sequence and different fold families within the designated threshold value. The ability to use a single structural motif to identify proteins across families will be invaluable for structural genomics projects. In these projects the targets are selected to have very low sequence identity to any currently in the PDB, and hence it is likely that they will belong to a new sequence family and have a new protein fold.

The key element in the methodology was the use of extended templates that included two residues before the start and at the end of the HTH motif. In this way, it was possible to reduce the false positives from 368 to 61. The inclusion of the accessible surface area (ASA) threshold value also contributed to the elimination of further false positives.

This simple method of using 3D structural templates to make predictions about the potential DNA binding function of proteins has been validated using scans of complete proteins in the PDB, and then used to make predictions for structural genomics targets.

Using the HTH motif templates as a prototype, a computer server has been constructed (http://www.ebi.ac.uk/thornton-srv/databases/DNA-motifs) that enables users to scan the uploaded coordinates of any 3D protein structure against the current template library. Further libraries will be added to this server as other DNA-binding motifs are extracted and validated.

2.4 DNA-Binding Prediction Using Electrostatic Potentials

"Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins", Jones et al., 2003 [G]

This work analyzed residue patches on the surface of DNA-binding proteins and developed a method for predicting DNA-binding sites using a single feature of these surface patches. Surface patches and the DNA-binding sites were initially analyzed for accessibility, electrostatic potential, residue propensity, hydrophobic, and residue conservation. A reliable computational method to help identify DNA-binding sites on the protein surface is important for functional annotation. The method can also facilitate directed mutagenesis experiments, in which specific residues are mutated and their effect on DNA binding is analyzed.

It was observed that the DNA-binding sites were, in general, amongst the top 10% of patches with the largest positive electrostatic scores. From this, we developed a prediction method in which patches of surface residues were selected that excluded residues with negative electrostatic scores. This method was used to make predictions for a data set of 56 non-homologous DNA-binding proteins. 68% of the data set was correctly predicted.

2.5 A Combined Structure- and Sequence-Based DNA-Binding Prediction Method

"Detecting DNA-binding helix-turn-helix motifs using sequence and structure information", Pellegrini-Calace et al., 2004 [H]

This study analyzed the potential for transferring structural knowledge back into the sequence level for the development of a new sequence-based method for DNA-binding HTH motif prediction. In particular, we aimed to verify whether the information contained in the sequence of the structural motif (here called partial domain, PD) resulted in a more powerful discriminator than the one derived from the full DNA-binding domain (FD).

Previous evolutionary studies proved that, apart from the active domains, DNA-binding HTH protein families exhibit a near-maximal divergence in both amino acid sequence and structural elements outside of the DNA-binding motif. Therefore, we aimed to derive models able to recognize independently evolved HTH motifs in either distantly related sequences or unrelated proteins. Hidden-Markov-Models (HMMs) previously proved to be among the best profile-based methods, were chosen for the pattern homology detection. Two HMM libraries, corresponding respectively to the PD and FD multiple sequence alignments were set up, tested, and compared to the method of 3D structural templates described by Jones, Barker, Nobeli, and Thornton (F).

From this study, we found that HTH sequence information is highly complementary to the corresponding structural information, and the structure/sequence combined method constituted a significant improvement over the two single-feature approaches.

3. Structural Analysis of RNA Structure and Interactions

3.1 "Protein-RNA interactions: a structural analysis", Jones et al., 2001 [I]

A detailed computational analysis of 32 protein-RNA complexes is presented in this work. Data was drawn from the NDB and the PDB. A number of physical and chemical properties of the intermolecular interfaces were calculated and compared with those observed in protein-double-stranded DNA and protein-single-stranded DNA complexes. In addition, the distribution of observed atom–atom contacts in the protein–nucleic acid complexes were calculated and compared to expected values. The interface properties of the protein-RNA complexes revealed the diverse nature of the binding sites.

This analysis presented a similar picture to that observed in DNA binding proteins, in that there is no single archetypal RNA binding site. In this dataset, the largest analyzed in this way, there were 32 proteins representing 14 structural families. When the predominant secondary structure element of each binding site was analyzed, the sites were equally divided between alpha helix and beta-strand, with only one example of an alpha/beta interface. The RNAs bound included elongated single-stranded, looped single-stranded, single-stranded with multiple loops, and double-helix structures. The size and polarity of the RNA binding sites varied widely, as do the modes of recognition used by the protein and the RNA structures recognized. Thus, the picture presented was far more complicated than that of protein–DNA complexes.

A comparison between protein-RNA and protein-DNA complexes showed that while base and backbone contacts (both hydrogen bonding and van der Waals) were observed with equal frequency in the protein-RNA complexes, backbone contacts were more dominant in the protein-DNA complexes. In the protein-RNA complexes, van der Waals contacts played a more prevalent role than hydrogen bond contacts, and preferential binding to guanine and uracil was observed. The positively charged residue arginine, and the single aromatic residues phenylalanine and tyrosine, play key roles in the RNA binding sites.

Similar modes of secondary structure contacts were observed in proteins binding RNA to those that bind DNA. However, when looking more closely at amino acid preferences and base versus backbone contacts, similarities are much harder to find. The unpaired state of many of the bases in RNA structures means that they are more readily available to make contacts with amino acid residues than those in the tightly paired double helices of dsDNA. Hydrogen bond contacts to all parts of the RNA are far less common than in the protein–dsDNA complexes. The ratios of contacts made to the nucleic acid bases and the backbone showed the differences between protein–RNA and protein–dsDNA complexes, and the similarities between the contacts made to RNA and ssDNA.

In terms of size, the protein–RNA complexes were observed as intermediary between the two types of protein–DNA complexes, but they were the least well packed of the three types of complexes. The poor packing of the protein–RNA complexes is a result of the complex tertiary structure that the RNA chains form. The atom contact analysis showed that the purine base guanine is preferentially contacted by proteins in both RNA and dsDNA structures.

The protein–RNA interface parameters calculated here can be calculated for any protein–RNA complex using the protein–nucleic acid server on the World Wide Web (http://www.biochem.ucl.ac.uk/bsm/DNA/server). This tool allows the user to upload the 3D coordinates of any protein–nucleic acid complex and receive back a report of its interface parameters. This server provides a simple means of comparing new complexes with those already known.

3.2 RNA Visualization Tools

“Tools for the automatic identification and classification of RNA base pairs”, Yang et al., 2003 [J]

Three programs were developed to aid in the classification and visualization of RNA structure. Three programs were developed: 1) BPViewer provides a web interface for displaying 3D coordinates of individual base pairs or base pair collections; 2) RNAView automatically identifies and classifies the types of base pairs that are formed in nucleic acid structures using the formalism developed by Leontis and Westhof (4); 3) RNAMLview can be used to rearrange various parts of the RNAView 2D diagram to generate a standard representation (like the cloverleaf structure of tRNAs) or any layout desired by the user.

With the base pair annotation and the 2D graphic display, RNA motifs are rapidly identified and classified. The full diagram convention simplifies and clarifies the annotation, description, and comparison of secondary structure, RNA motifs, and tertiary interactions present in a folded RNA.

A survey was carried out for 41 unique structures selected from the NDB. With these tools, statistics for the occurrence of each edge and of each of the 12 base pair families as proposed by Leontis and Westhof were compiled for the combinations of the four bases: A, G, U, and C. In addition, RNAview 2D projections have been calculated for all structures in the NDB and are available on the NDB Atlas pages (http://ndbserver.rutgers.edu/atlas/).

Web servers for BPViewer and RNAView are provided by the NDB. The application RNAMLview can also be downloaded from the NDB site.

3.3 Method For Defining RNA Conformations

“RNA conformational classes”, Schneider et al., 2004 [K]

3000 nucleotides of 23S and 5S ribosomal RNA from a near-atomic resolution structure of the large ribosomal subunit (NDB code rr0033, PDB code 1jj2) were analyzed in order to classify their conformations. Fourier averaging of the six 3D distributions of torsion angles and analyses of the resulting pseudo-electron maps, followed by clustering of the preferred combinations of torsion angles were performed on this dataset. 18 non-A-type conformations and 14 A-RNA related conformations were discovered and their torsion angles were determined.

In this work, the multidimensional RNA conformational space and the very large number of possible correlations among the individual torsion angles were simplified by focusing on the interrelationships of the conformation angles that define the phosphodiester linkage and the other backbone torsion angles. Using a Fourier averaging method developed earlier for analyzing hydration patterns (5), coupled with a clustering technique, led to identification of the new conformational groups. These dinucleotide conformations are fully described by torsion angles and their Cartesian coordinates are available.

The conformational space of RNA consists primarily of the A-type building block and a minority of diverse other conformations. This work shows that there are distinct classes of these minority conformations. The identified RNA conformations and their idealized coordinates can facilitate analysis of RNA structures and their computer simulations.

Sequence preferences of the conformations were observed in very few cases, often in purine-rich regions.

The study also suggests that the multidimensionality of the RNA conformational space can be approached by analysis of conformations at the phosphodiester link. We deduce this central role of the two torsion angles involved at this link, zeta and alpha+1, from the fact that they exhibit the highest variability, yet are limited into well defined regions, noise notwithstanding. We suggest that the character and importance of the zeta-alpha+1 scattergram can be compared with the cornerstone of protein structural science, the Ramachandran plot of protein backbone torsion angles phi and psi.

4. Conclusion

In our work to study nucleic acid-protein interactions, there have been two central themes of study. First were detailed computational analyses of DNA- and RNA-protein complexes, which have led to a better understanding of the chemical and physical properties that exist between these two types of molecular interface. A classification analysis of DNA-binding proteins was supplement to the DNA-protein study in order to further our understanding of DNA-binding, as well as propel our work into the second theme of our work: DNA-binding prediction. In total, five different methods were developed: four that focused on the HTH motif in particular, and a fifth that allowed for a more general classification of DNA-binding proteins. Finally, we have also made advances in the more clouded field of RNA structure, as we developed software tools as well as a new method for defining RNA conformational classes.

One exciting application of our study is how the classification of protein-nucleic acid complexes will aid our interpretation of genome sequences. For example, although preliminary studies of the available genomes show that many proteins will probably fall into existing DNA-binding families (notably those with HTH, zipper-type, and ßßα zinc-finger motifs), there are exciting possibilities of discovering further modes of DNA-binding. Genome analysis will not only facilitate identification of such proteins, but also allow us to determine functionally important target sites on the DNA and, in combination with structural data, how higher-order oligomers are formed within the cell. Ultimately, this will expand our understanding of the regulation of protein expression and DNA packaging, rearrangement, repair, and replication, which are indispensable to the viability of organisms.

The DNA-binding predictive models should prove to be invaluable tools for automatic motif recognition as the PDB grows with concomitant deposition of protein structures of unknown function. In addition, the methods can be modified to recognize other DNA-binding structural motifs and protein structural domains to further annotate incoming structures. For example, we can expand the general alpha helical DNA-binding model by describing structural characteristics of alpha helices in each DNA-binding class, and either make individual classification models for these different types of DNA-binding alpha helices, or incorporate their characteristic structural measurements into a single model.

The 3D structural template methodology used for finding the HTH motif will be a prototype for functional predictions for proteins that recognize DNA with small contiguous structural motifs. The aim now is to repeat this methodology's success for other sequential DNA-binding motifs, such as the helix–hairpin–helix, helix–loop–helix, and ribbon–helix–helix. To do this, we must calculate and validate new rmsd and ASA thresholds for each motif. Once complete, such a tool will be able to scan a protein structure of unknown function with a library of many different types of motif in one single operation, and thus make predictions about their presence or absence. We have also shown that DNA-binding sites are among the surface patches with the most positive electrostatic potential. An obvious next step is to combine the 3D structural template method with electrostatic potential data to make both methods more robust and increase their specificity to DNA-binding motifs.

Finally, we believe that the combined knowledge of RNA base pair geometries and preferred RNA conformations is a necessary step in understanding protein RNA interactions.

References:

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

2. Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, Demeny T, Hsieh SH, Srinivasan AR, and Schneider B. (1992). The Nucleic Acid Database: a comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.* **63**, 751-759.

3. Vaguine AA, Richelle J, and Wodak SJ. (1999). SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crys.* **55**, 191-205.

4. Leontis NB and Westhof E. (2001). Geometric nomenclature and classification of RNA base pairs. *RNA.* **7**, 499-512.

5. Schneider B, Cohen DM, Schleifer L, Srinivasan R, Olson WK, and Berman HM. (1993). A systematic method to study the spatial distribution of water molecules around nucleic acid bases. *Biophys. J.* **65**, 2291-2303.

Publications from previous grant period:

A. Das U, Chen S, Fuxreiter M, Vaguine AA, Richelle J, Berman HM, and Wodak SJ. (2001). Checking nucleic acid crystal structures. *Acta Cryst.* **57**, 813-828.

B. Jones S, van Heyningen P, Berman HM, and Thornton JM. (1999). Protein-DNA interactions: a structural analysis. *J. Mol. Biol.* **287**, 877-896.

C. Luscombe NM, Austin SE, Berman HM, and Thornton JM. (2000). An overview of the structures of protein-DNA complexes. *Genome Biol.* **1**, 1-37.

Publications from this grant period:

D. McLaughlin WA and Berman HM. (2003). Statistical models for discerning protein structures containing the DNA-binding helix-turn-helix motif. *J. Mol Biol*. **330**, 43-55.

E. McLaughlin WA, Kulp DW, de la Cruz J, Lu XJ, Lawson CL, and Berman HM.(2004). A structure-based method for identifying DNA-binding proteins and their sites of DNA-interaction. *J. Structural and Functional Genomics. **5**, 255-265*

F. Jones S, Barker JA, Nobeli I, and Thornton JM. (2003). Using structural motif templates to identify proteins with DNA binding function. *Nucleic Acids Res*. **31**, 2811-2823.

G. Jones S, Shanahan H, Berman HM, and Thornton JM. (2003). Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res*. **31**, 7189-7198.

H. Pellegrini-Calace M, Berman HM, and Thornton JM. (2004). Detecting DNA-binding helix-turn-helix motifs by structure-based sequence information. *Nucleic Acids Res.* **(in preparation).**

I. Jones S, Daley DTA, Luscombe NM, Berman HM, and Thornton JM. (2001). Protein-RNA interactions: a structural analysis. *Nucleic Acids Res.* **29**, 943-954.

J. Schneider B, Moravek Z, and Berman HM. (2004). RNA conformational classes. *Nucleic Acids Res.* **32**, 1666-1677.

K. Yang H, Jossinet F, Leontis N, Chen L, Westbrook J, Berman HM, and Westhof E. (2003). Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res*. **31**, 3450-3460.

Websites:

Protein-Nucleic Acid Interaction Server
http://www.biochem.ucl.ac.uk/bsm/DNA/server

Taxonomy of Protein-DNA Complex Structures
http://www.biochem.ucl.ac.uk/bsm/prot_dna/prot_dna.html

Function Prediction Using Structural Templates
http://www.ebi.ac.uk/thornton-srv/databases/DNA-motifs

QPROF – Query of Protein Features
http://qprof.rutgers.edu

DNA-Binding Prediction Tool

http://dna-binders.rutgers.edu

BPView
http://ndbserver.rutgers.edu/services/BPviewer

RNAview
http://ndbserver.rutgers.edu/services/rna_viewer

RNAMLView may be downloaded from
http://ndbserver.rutgers.edu/services