Rules of protein-DNA recognition: computational and experimental advances

Alexandre Morozov (Rutgers University),
Gary Stormo (Washington University in St. Louis School of Medicine)

June 16 – June 21, 2013

1 Overview of the Field

Protein-DNA interactions are of fundamental importance in many areas of biology. This workshop's focus was primarily, but not exclusively, on sequence-specific transcription factors (TFs) that bind to particular segments of DNA to control the expression of subsets of genes. The proper functioning of TF-DNA binding is crucial to the development of organisms and their responses to environmental conditions. For example, cells respond to environmental stresses by upregulating some genes and downregulating others in a coordinated and robust manner. Disruptions of the normal TF-DNA interactions are often associated with diseases of various kinds. Understanding which interactions occur, how they are themselves regulated, and the consequences of genetic variations on the normal functioning of cells are critical issues in modern biomedical research.

In most species TFs represent about 5-10% of all genes but the exact repertoire is generally unknown. Individual TFs may be highly specific and control the expression of only a few genes. The classic example of the Lac Repressor is only known to control the expression from a single promoter in the *E. coli* genome. Other TFs control the expression of many genes throughout the genome, often in specific combinations. Natural variations in TF binding site sequences contribute to the normal phenotypic variation among individuals. Mutations in binding sites within a species can lead to disruptions in regulation of gene expression, which is often associated with specific diseases. But variation between TFs genes and their binding sites is also a primary driving force in evolution, where changes in gene regulation show up as changes in phenotypes between species.

Technological advances in recent years have provided enormous databases of biological information, including genome sequences of many species. This allows for the identification of many genes in each species, but the complete elucidation of gene functions requires further work. Most TFs can be identified from their gene sequences, but their target DNA binding sites are not easily predictable by computational methods due to the complexity of the relationship between protein-DNA binding affinity (which determines which sites will be bound), and DNA sequence. Even when binding sites for specific TFs are determined experimentally, the effects of sequence variations are generally not predictable. For humans there are now thousands of individual genome sequences available, including both natural variation and that associated with particular diseases. There are now many examples of genetic variation that affect gene expression, and it is anticipated that such variations play a significant role in phenotypic diversity. Better predictive models, as well as further experimental data, are needed to understand more fully the causes and consequences of gene expression variation. Progress in that area would help accomplish many of the goals of current biomedical research.

The physical basis of specificity in protein-DNA interactions has long been investigated by a variety of approaches. Structural studies, primarily using X-ray crystallography or nuclear magnetic resonance (NMR) techniques, show that sequence-specific binding can be accomplished by a combination of direct interactions between the protein and the bases of DNA and by variation in the structure of DNA-binding motifs (such as bending of the DNA helix or narrowing of the minor groove). Biochemical and biophysical approaches can determine the fundamental forces involved in protein-DNA interactions, and have shown that different proteins can use alternative binding mechanisms. Electrostatic contributions are common to all TF-DNA interactions, but specificity can be obtained through either enthalpic or entropic contributions. New high-throughput technologies allow for quantitative measurements of binding affinity simultaneously for thousands to millions of different DNA sequences *in vitro*, and are also capable of identifying binding site locations *in vivo* genome-wide. The combination of all experimental approaches has elucidated the details of many specific examples of TF-DNA interactions, and provided some general principles of how TF-DNA binding specificity is established. Computational approaches build on those general principles to make predictive models that provide insights into biological processes and drive further experiments.

2 Recent Developments and Open Problems

While technological advances in recent years have generated extensive data about protein-DNA interactions, there are many unanswered questions and open problems. This is a partial list that is relevant to the workshop:

- **Identification of DNA-binding proteins**

  The genome sequence of an organism allows for the identification of most TF coding sequences, but the exact set of TFs is unknown in any organism. Identification is primarily based on the fact that TFs belong to a small set of protein families and thus can be identified by sequence similarity. For example, most bacterial TFs are members of the helix-turn-helix class of proteins, whereas in eukaryotes TFs are usually members of zinc finger, homeodomain and several other well-known protein families. However, predictions based solely on sequence similarity can lead to some false positives. For example, while most zinc finger proteins bind to DNA, in some cases they bind RNA instead (and in some they bind to both), and in other cases they participate in protein-protein interactions. Furthermore, relying solely on protein sequence similarity will also miss some DNA binding proteins that do not belong to any of the standard TF families. Experimental methods can resolve whether or not a protein binds to DNA, but those experiments can be costly and laborious if performed on all possible DNA-binding proteins. Computational approaches that predict DNA-binding proteins based on sequence and structure can aid in the identification of the complete repertoire of TFs.

- **Contributions of structure to sequence-specific binding**

  Sequence-specific binding affinity can be due to at least two mechanisms. One is by direct interaction between protein amino acids and DNA base-pairs in the binding site. Hydrogen bonds, as well as other types of interactions can provide different energies to different sequences. In addition, the structure of DNA is also sequence-dependent. The classic B-form DNA is really an average structure, and different sequences diverge from that average in various ways. For instance, the major and minor groove widths are different for different sequences, and some sequences have intrinsic bends or increased flexibility relative to others. Proteins can use those differences to obtain different binding energies to different sequences, thus generating a sequence-specific binding mechanism. That structural variation can contribute to differential binding affinity has been known for a long time, but there are open questions about how specific proteins utilize the structural information and how to incorporate such information in predictive models, especially on a genomic scale.

- **Predicting binding sites for all TFs and determining the complexity of models required for useful predictions**

  Recent technological advances have greatly increased the knowledge of protein-DNA binding specificities. However, less than half of the TFs in any species have well characterized specificities, so there remains a large amount of data to be collected. Some new experimental methods can help to fill in the

details, including approaches that examine the binding of multiple TFs simultaneously. But in addition to generating the relevant datasets, computational methods for analyzing the data and developing predictive models also need further improvements. Some TFs can be well represented by simple models, such as position weight matrices (PWMs), and good methods exist for estimating the parameters of those models. But for TFs that require more complex models, especially for those with multiple modes of binding, improved computational approaches are still needed.

- **Improved structural modeling of protein-DNA interactions**

  Biophysical modeling, using molecular dynamics simulations and Monte Carlo approaches, can be very useful in understanding the mechanism of protein-DNA interactions, including dynamics and energetic landscapes. There are several challenges to these modeling approaches. Accurate force field parameters are critical to get good models, and the use of implicit versus explicit solvent can also influence the accuracy of the results. The repertoire of allowed moves governs the sampling of conformations and determines the computational time required for suitable sampling runs. Recent improvements in algorithms, as well as continued increases in computer speed, have enabled much more comprehensive simulations to be performed than ever before. New insights have been gained, but it is also clear that further improvements in all aspects, from force field parameters to sampling strategies, are needed to provide highly accurate predictions of protein-DNA energetics and specificty. One important goal is to make predictions from limited data, such as just protein sequence, that would accurately match experimental measurements. That would have many applications in genome analysis and also aid in the design of proteins with desired specificity, facilitating synthetic biology research.

- **Interactions between TFs and coordinated regulation; cooperation and competition between TFs and nucleosomes**

  Even if we knew all of the TFs and the specificity for each one, that would not be sufficient information for accurate modeling of regulatory networks *in vivo*. Within cells the interactions of TFs with each other, both positive and negative, need to be taken into account. TFs can compete for binding to the same, or overlapping, sites, and they can bind cooperatively so that they bind at locations that would not be occupied with any one TF alone. Furthermore, in eukaryotic cells the binding of TFs is in competition with nucleosomes (fundamental units of chromatin in which 147 base-pairs of genomic DNA is tightly wrapped around a histone octamer; typically, nucleosomes cover 75-90% of all genomic DNA), and their positions are governed by many factors, many of them not understood. Technologies are being adapted to assay binding of multiple TFs at once, both *in vitro* and *in vivo*, that can help address some of the important issues about protein-protein competition and cooperativity that are critical for the accurate modeling of gene regulatory networks.

3 Presentation Highlights

A poll of the participants following the workshop found that every talk was considered a highlight by at least some of the audience. The workshop accomplished its primary goal (as stated in the workshop proposal) of bringing together researchers who work on a common topic, the interactions of proteins with DNA, but from a variety of approaches and perspectives. At most scientific meetings, this topic occupies a small portion of a meeting on a broader subject. For example, at meetings focused on genome analysis there might be one or two sessions devoted to protein-DNA interactions, or likewise at meetings focused on bioinformatics/computational biology only a fraction of the sessions would be on this topic. The diverse experimental and computational approaches represented in this workshop means that people tend to go to diverse sets of meetings; indeed most of the participants had not met one another before, even though they work on a common topic. This fact was perceived as a strength of the workshop by everyone polled. In the following we describe specific highlights of the meeting, organized around the specific open questions listed in the previous section.

- **Identifying DNA-binding proteins**

  It is not yet possible to predict DNA-binding proteins by their sequence alone unless they belong to one of the well known TF families, and even then false positive predictions are possible. But given both

the sequence and the structure of the protein, it is possible to predict DNA-binding proteins and also to determine the DNA-binding surface of the protein. The structure may be determined experimentally, or inferred from homology with other proteins of known structure. Talks by Julie Mitchell and Yael Mandel-Gutfreund both addressed the issue of predicting DNA-binding surfaces given a protein sequence and structure. Using different methods, they took advantage of extensive sets of protein-DNA complexes to identify features within the protein structures, such as electrostatic potentials on the surface, that can be used to accurately distinguish DNA-binding proteins and predict DNA-binding domains within the protein.

- **Structural analysis of protein-DNA specificity**

  A particular emphasis of the workshop was on structural analysis of protein-DNA interactions and how they contribute to DNA-binding specificity. Phoebe Rice described her X-ray crystallography studies of transposase proteins that bind to specific sites on DNA and perform enzymatic reactions, cleaving and religating DNA to cause structural rearrangements of the chromosome. Robert Kaptein described his NMR studies to determine the structures of protein-DNA complexes in solution. His studies of the classic DNA binding protein, the Lac repressor, bound to several different DNA sequences, have elucidated the rearrangement of contacts that occur when the protein binds to different DNA sequences. Of particular interest is the altered geometry of the protein-DNA interface when the protein is bound non-specifically.

  It has long been appreciated that specificity of protein-DNA interactions can be obtained with at least two complementary mechanisms. One is referred to as direct readout, where the amino acids of the protein surface make contacts with base-pairs in the DNA site, allowing for energetic differences in binding to different sequences. In addition, the structure of DNA is sequence-dependent, so that proteins that prefer specific structures, such as an overall bend to the DNA or especially narrow or wide grooves, can obtain different binding energies to different sequences even without making direct contacts with the base-pairs. This mechanism is referred to indirect readout. Akinoro Sarai, who has pioneered the analysis of indirect readout contributions to DNA-binding specificity, described the database of protein-DNA complexes that he has managed for many years, as well as approaches for modeling and quantifying the contributions of structure to protein-DNA binding specificity.

  Remo Rohs has also been developing models that account for structural contributions to DNA-binding specificity. To apply such models to genome-wide analyses, he has determined important structural parameters, such as minor groove width, roll, and propeller and helix twist, over all possible pentanucleotide sequences, which can then be combined for a whole genome structural model. The structural parameters can not only contribute to understanding of specific protein-DNA interactions, such as for several homeodomain proteins, but can also be used in the prediction of altered binding specificities in nucleotide variants on a genome-wide scale. In a related talk, Iris Dror described the use of such parameters in the modeling of protein-DNA interactions and how they can complement sequence based parameters, such as PWMs, to provide more accurate models of protein-DNA specificity. Structural analyses also played a large role in the description of specific protein-DNA interaction studies, such as those of Richard Mann on the Hox and Ext homeodomain proteins and by Tali Haran on the association of the tumor suppressor protein P53 and its differential affinity for different sequences. In particular, Tali considered the cost of DNA deformation that contributes to binding affinity differences in sites with different spacings between half-sites.

  One of the important themes of the meeting was the contribution of water to protein-DNA specificity. Gerald Koudelka described the influence of water, and specifically formation of a hydration spine along the DNA surface, on binding of bacterial repressor proteins to DNA. He included models and extensive data on modified DNA sequences, including not only base-pair substitutions but also using nucleotide analogs to determine contributions of specific atoms in the complex. Water contributions to binding was also emphasized by Carlos Comancho, who described how such energetic contributions in zinc finger proteins could be used to improve modeling and how, in general, waters in and near the protein-DNA interface can influence the magnitude of hydrogen bonds between amino acids and base-pairs.

  Richard Lavery described advances in structural modeling of DNA, including explicit solvent calculations with K+ ions in both grooves of the DNA. He also described calculations on the energetics of

structural distortions. These calculations were possible because of improved software and increased computer speed. Phil Bradley has developed biophysical models of protein DNA complexes that use Monte Carlo techniques to predict specificity. In some cases explicit solvent contributions, in particular water interactons at the interface, lead to significant improvements and can still be sampled efficiently by employing rotamers of water configurations. Anjum Ansari described how bending and kinking of DNA can influence protein binding and so contribute to specificity even when the protein does not interact directly with the base-pairs. Models for bending propensity can be used to help predict binding specificity. In some proteins, such as the rad4 DNA repair enzyme, bases can be "flipped out" of the helix and the energetics of that flip can be modeled computationally.

Harmen Bussemaker described experimental data on rates of cleavage by DNAse I, an endonuclease that is generally considered to be sequence non-specific. However, the data show that rates of cleaving different hexamers can vary by up to 1000-fold. Sequence-dependent structural variation lead to those differences and can be modeled by taking into account structural parameters of the DNA.

- **Determining TF motifs from high-throughput technologies**

  Technological advances over the last decade have greatly enhanced our ability to experimentally determine the specificity of protein-DNA interactions. At the same time, computational algorithms have been developed to maximize the information obtained from those experiments. Several of the developers of the experimental methods and computational approaches attended the workshop and described some of the latest enhancements. Martha Bulyk, who invented the protein binding microarray technology, described its new uses with customized arrays to ask directed biological questions. In her talk she highlighted recent work of determining evolutionary changes in specificity and examples of multiple modes of binding for one family of transcription factors. Aseem Ansari, who invented a related technology, called cognate site identifier, focused on developing small molecule mimics of transcription factors that can be designed to binding to specific DNA sequences via minor groove interactions. He also described some visualization tools to represent the protein binding data. Jussi Tapale described large-scale SELEX-seq experiments on a wide variety of transcription factors. His group has pushed the technology to be massively parallel, obtaining specificity information for a large number of factors simultaneously. Scot Wolfe and Marcus Noyes have each contributed to the development and enhancement of bacterial one-hybrid methods for determining the specificity of transcription factors. Scot described recent experiments on homeodomain proteins and models for their specificity and on making homeodomain-zinc finger hybrid proteins to obtain proteins with unusual and desired speficities. Marcus described his work on zinc finger proteins, including randomizations of the protein to select ones with desired specificity. These can then be induced *in vivo* using a hormone induction system to induce their expression and follow the consequences of their subsequent regulatory interactions. Tim Hughes has utilized several of the technologies mentioned above, and gave a talk about a large-scale analysis of zinc finger proteins. He described a library of over 45,000 zinc fingers, and an analysis of over 5,000 zinc fingers for which the binding specifities were determined over a library of 135 different 4 base-pair long binding site variants. From that data one can build predictive models for a comprehensive set of zinc finger proteins.

  In addition to those experimental talks, there were several that developed computational approaches to analyzing the data from high-throughput technologies to obtain improved models of specificity. Gary Stormo provided some history of approaches at modeling specificity, including models that incorporate higher-order interactions, and then described their more recent work utilizing non-linear regression methods and biophysical models to infer specificities. Todd Riley also described computational approaches based on biophysical modeling and non-linear regression to determine the parameters of specificity. This approach can easily incorporate additional features of the DNA sequences, including higher-order interactions, and has been demonstrated to give accurate predictions for PBM and SELEX-seq data. Alex Morozov described some of his work using biophysical modeling to represent TF specificity, including an analysis of MITOMI data. The models incorporate steric exclusion between neighboring DNA-bound TFs and thus allowi for correct treatment of TF binding when overlapping sites are present on the DNA, including the case of cooperative binding by multiple types of TFs. Raluca Gordan described her work on analyzing PBM data, including non-linear regression methods with feature selection. She also described data collected for two bHLH TFs from yeast, Tye7 and

Cbf1, that bind to very similar motifs yet have distinct binding sites *in vivo*. Using PBMs that are derived from genomic sequences, she was assessing the contributions to differential binding specificities of the two proteins, both within the binding sites and in the flanking regions.

Wyeth Wasserman described a new approach for modeling DNA-binding specificity that allows for higher-order interactions as well as variable length binding sites. It includes nearest neighbor contributions to specificity as well as an HMM model for recognition that allows for deletions/insertions between positions of the binding sites. He also described specificity modeling based on *in vivo* location data, such as ChIP-seq and new enhancements to the JASPAR database of TF recognition motifs. He also described a compendium of examples of non-coding sequence variants that effect the expression of neighboring genes.

- **Recognition models and binding motif predictions**

  A long-standing goal in computational biology is the ability to predict the specificity of a DNA-binding protein based only on its sequence. Early attempts at a universal, deterministic recognition code proved impossible, but efforts in the succeeding years have focused on family-specific recognition models that are probabilistic. Two general approaches have been employed. One is data-driven statistical models using machine learning methods. The other is a biophysical approach where Monte Carlo and molecular dynamics simulations predict changes in binding affinity for specific factors to all possible (or at least a large compendium) of potential binding sites. Gary Stormo described the use of a Random Forest regression approach on large quantitative binding sets for both homeodomain and zinc finger proteins to develop predictive models of specificity. Mona Singh described work using support vector machines trained on zinc finger interaction data to develop predictive models of specificity. On the biophysical side, Phil Bradley described his approach, an adaptation of the Rosetta software for protein modeling and prediction, for use on predicting protein-DNA binding specificity. An advantage of this approach is that it can be applied to any transcription with a starting structure for the protein-DNA complex and doesn't require an extensive training set. In all cases progress has been made and predictions of specificity based on protein sequences have improved considerably, but there is still ample room for further improvements.

- **Interactions between TFs and the effect of chromatin structure**

  Most of the workshop focused on the issues of intrinsic specificity of proteins for DNA sequences, usually determined *in vitro*. But *in vivo* there are many complicating factors that can affect protein-DNA interactions and some of the talks included such analyses. For example, Richard Mann described how interacting factors can alter the specificity of each other, as in the example of Hox-Ext binding sites. Tali Haran described work on the tumor suppressor protein P53 where the interactions between dimers alter the binding specificity. Because of cooperativity between dimers, the protein concentration plays a pivotal role in site selection. Matt Slattery described work on analysis of genome-wide binding site analyses. ChIP-seq data for 80 TFs in various tissues was compared to obtain binding location analysis and infer interactions between TFs. Some genome segments are "hot spots" for binding, being associated with a large fraction of TFs, others with a more moderate number and still others associate very specifically with only one or a few. Mona Singh also described work on the analysis of genome location analysis data and associations with cross-species and conservation and chromatin accessibility which together provide for more accurate predictions of regulatory interactions. Alex Morozov also described computational models that include competition between transcription factor binding and nucleosome positioning. Both sequence and structural contributions to binding specificity play a role in determining the overall occupancy of different DNA segments. David Lukatsky focused on how non-specific DNA binding can affect TF occupancies genome-wide. He also discussed the interplay between nucleosome positioning and specific and non-specific TF binding. Sarah Bondos studies the Ubx (Ultrabithorax) protein, a homeodomain protein that is critical for proper development in *Drosophila*. It has the same DNA binding domain as another homeodomain protein, AbdA, but its *in vivo* binding is largely to distinct regions. Her work focuses primarily on the interactions between Ubx and other factors and how that contributes to the overall regulatory program. Ubx protein also has interesting material properties, including homopolymerization and interactions with other proteins and DNA, that may be useful for the design of interesting bioactive sensors. . Norbert Reich described a new technology that combines

protein binding microarrays with TIRF microscopy. This has several advantages, including the ability to measure kinetic binding parameters and also to monitor and measure binding of several different proteins to segments of DNA by means of using alternative fluorescent dyes to mark each protein. It opens up a new approach to modeling more complex associations of multiple proteins and DNA *in vitro*.

- **RNA-binding proteins and recognition motifs**

  One additional topic, protein-RNA interactions and motif identification, was also discussed in a talk by Quaid Morris. Much of gene regulation occurs post-transciptionally and involves protein-RNA interactions. Many of the issues described above for protein-DNA interactions are also relevant to protein-RNA interactions. The field is less mature, which fewer large datasets, and this talk provided an introduction to some of the methods being employed to determine binding site motifs for RNA-binding proteins. Using 35 nucleotide RNA sequences, Quaid Morris and colleagues assayed binding specificities of 209 distinct RNA-binding proteins, and identified binding site motifs for each. Some of the motifs appear to be sequence-only, as for DNA binding proteins, but other require structural modeling that involves base-pairing between segments of the RNA.

4 Scientific Progress Made

The nature of the protein-DNA field is that paradigms emerge over time based on combinations of experimental data and computational modeling. The main progress, as reported by several participants in post-workshop polls, was a better appreciation of alternative approaches to studying protein-DNA interactions, and some new ideas or insights that will be taken back to their groups for further development and testing. Interactions between participants also resulted in several new collaborations which are now underway. These accomplishments are in agreement with our original goals for the workshop, which focused on bringing experimental and computational researchers together in order to foster informal exchange of ideas.

5 Outcome of the Meeting

There was a strong general consensus that this was a very worthwhile meeting; more than one participant said it was the best meeting they had ever attended in this field. As mentioned above, several participants have agreed to share data and have begun discussing new collaborations. Several said they have a better understanding and appreciation of alternative methods of analyzing protein-DNA interactions and will expand the repertoire of methods they use, both experimental and computational. Several also stated that an important outcome of the meeting is a new perspective on their own data, which will result in new experimental directions and new insights into mechanisms of specificity. There was a strong sentiment that BIRS was an outstanding venue for this type of meeting, providing a small group of researchers with overlapping interests but diverse approaches with the means of presenting their own work and initiating many informal discussions. A proposal to organize another similar meeting in a few years was supported overwhelmingly.