Rule of Protein-DNA Recognition: Computational and Experimental Advances

Trevor Siggers (Boston University, Boston, MA),
Marcus Noyes (Langone Medical Center, New York University, New York, NY)

June 3 – 8, 2018

1 Background

Genetic information of the cell is stored in its DNA, providing a stable medium for information storage and retrieval. To make use of this stored information, the DNA template must be transcribed to RNA molecules, which in turn are translated into protein molecules. Gene transcription (i.e., turning a gene 'on') is regulated by DNA-binding proteins called transcription factors, whose function is to bind to specific DNA sites in the vicinity of their target genes. Understanding and representing the biophysical rules of protein-DNA recognition remains a central problem in molecular biology. Since transcription factor binding sites are short (typically 6-12 nucleotides in length) and degenerate (proteins can bind to many different sequences) the problem of identifying them in the genome presents a significant mathematical challenge. This BIRS workshop brought together researchers who use diverse theoretical, computational and experimental approaches to study protein-DNA interactions, from detailed molecular analysis, simulation, and modeling to massively parallel experiments designed to find all instances of proteins bound to their genomic sites in living cells.

This workshop focused on recent progress in measuring and modeling the biophysical mechanisms that enable DNA-binding proteins (such as transcription factors) to bind their cognate sites within the vast genome. The main purpose of the workshop was to bring together researchers with diverse mathematical, computational and experimental approaches to studying protein-DNA recognition. The goal was to obtain new mathematical and computational approaches to transcription factor binding measurement modeling and prediction. This field has developed from pioneering work on sequence analysis by researchers such as Michael Waterman and Temple Smith, and mathematical representations of protein-DNA binding specificity by Peter von Hippel and Gary Stormo (keynote speaker and previous organizer). The BIRS workshop functioned to integrate these issues with the current experimental techniques and data, to directly address possible routes forward. A second goal was to improve methods for detecting the impact of genomic DNA differences between people (e.g., single nucleotide polymorphisms, SNPs) on transcription factor binding and gene regulation. Advances in modeling the impact of genomic variation on gene regulation will have tremendous impact on understanding the link between genomic variation and human health and disease. Recent breakthroughs in developing high-throughput experimental techniques, for both the understanding of natural transcription factor binding activity as well as the influence non-coding SNPs might have on DNA-binding in vivo, make the development of novel mathematical and computational frameworks for analyzing protein-DNA interaction data a high priority. The workshop helped to fulfill this urgent need.

The organizers are both early career scientists in this field, that have made important contributions in both the measurement and modeling of protein-DNA interactions. Dr. Trevor Siggers, is an Assistant Professor in the Department of Biology at Boston University. He has been instrumental in developing and applying

the protein-binding microarray (PBM) technique for the high-throughput characterization of protein-DNA binding specificity. His research integrates computational models of protein-DNA binding with genome-scale experimental datasets to understand specificity of gene-regulatory networks and transcription factor evolution. He was awarded a National Science Foundation (NSF) Postdoctoral Research Fellowship in Biological Informatics, a National Institutes of Health (NIH) K22 Research Scholar Development award, and an NIH R01 that utilizes computational and experimental approaches to study proteins coordinating inflammation. His laboratory is currently funded by the NSF and NIH. Dr. Marcus Noyes is an Assistant Professor in the Department of Biochemistry and Molecular Pharmacology and a core faculty member of the Institute for Systems Genetics at the NYU School of Medicine. Dr. Noyes developed a bacterial one-hybrid assay for the simple characterization of transcription factor specificity, a method that has been employed by a variety of labs to characterize factors from a host of model organisms. In fact, over half of the transcription factors in the fruit fly have been characterized by this method. For this work he received the Harold M. Weintraub Award and was recruited to start his own lab directly out of graduate school as a Lewis-Sigler Fellow at Princeton University. His work at Princeton demonstrated that screens of synthetic transcription factors can be used to accurately predict the function of naturally occurring factors. In his lab NYU he is continuing this work.

Our invitation to participate in the workshop was met with great enthusiasm by the researchers in the field. We took great care to invite participants that include prominent experimental biophysicists and biologist (Joe Thornton, Barak Cohen, Manuel Llinas, Jussi Taipale, Tim Hughes) who are the forefront of the field, as well as mathematicians, physicists and computer scientists (Alexandre Morozov, Mona Singh, Gary Stormo, Yaakov Levy, Harmen Bussemaker) whose main focus is on mathematical modeling of protein-DNA interactions. Attendees included a number of promising young scientists in the list (e.g., Raluca Goran, Matt Weirauch, Polly Fordyce, Jose Rodriquez-Martinez, Jeff Viestra, Yaron Orenstein). We had eight spots filled by graduate students and postdoctoral associates who are relatively new to the field, and we had a number of attendees from Latin American countries (Alejandra Medina Rivera, UNAM, Mexico); Victor Tierrafria (Trainee, UNAM, Mexico); Julio Collado-Vides (UNAM, Mexico); Jose Rodriguez-Martinez (U. Puerto Rico, PR).

One of the main goals for this meeting was to facilitate free exchange of ideas and foster new collaborations, both formal and informal. We believe that BIRS, with its common areas and a beautiful location, provided an ideal setting for these activities. Similar workshops held in 2013 (Banff) and 2015 (Oaxaca) were hugely successful and the overwhelming consensus has been to continue this dialog every 2-3 years. This meeting brought together an unusual group of people who all share a passion for understanding how proteins recognize their DNA binding sites, and fills an important niche that does not overlap with any other conferences. A brief survey of past participants revealed a number of collaborations (approxiamtely 20) that were initiated as a direct result of the previous BIRS meetings. One resulting publication, between four previous attendees, even acknowledges BIRS as making the joint research possible (Yang et al. (2013) NARS 42:D148). Given the success of these two previous meetings the overall agenda and goals for the 2018 meeting followed closely the format of these previous meetings. Rather than presenting experiments and modeling in separate sessions, we mixed experimental and theoretical presentations within a given lecture block whenever possible. Because many invited attendees have both experimental and theoretical/computational components in their labs, they were encouraged to cover recent advances using both approaches in their talks. Furthermore, we stressed the desire for informal presentations and discussions of unpublished data, to maximize the immediate influence that the meeting has on how the field will evolve.

For the structured part of the meeting, we followed the format in which 30-minute talks (+10 min for questions) was scheduled in groups of three or four. Every invited participant gave a talk. Each group of talks was be followed by a tea/coffee break during which the participants had a chance to discuss the latest lectures in more detail. Interspersing short sessions with breaks provided us the flexibility to allow extended time for discussions of specific topics generate interest; we found this to be particularly beneficial for previous meetings, and was so again.

Based on informal and formal feedback from participants, the meeting was a great success. We received numerous emails and informal verbal feedback confirming that participants enjoyed the meeting and found it useful. For example, Dr. Aseem Ansari (U. Wisconsin, a senior investigator in the field) wrote, “Allow me to congratulate you once again on organizing an outstanding meeting”. Another senior scientist, Dr. Manuel Llinas wrote more simply, “fabulous meeting”. Participants strongly agreed that this meeting brought together a unique combination of scientists, allowed for productive exchange of ideas, and should be a continuing

workshop held every few years if possible.

2 Recent Developments and Open Problems

We designed the workshop to bring together researchers at the forefront of diverse experimental and computational aspects of the common field of transcription and protein-DNA binding. Over the past 15 years advances in microarray technologies, next-generation sequencing and computational studies have lead to a renaissance in the study of protein-DNA interactions both in vivo and in vitro, leading to massive data resources to draw upon to understand how protein complexes interact with DNA and lead to transcription. The workshop reiterated the need for continued efforts in cataloging the DNA-binding specificity of factors and for the close examination of binding features of homolgous proteins to understand specificity and evolution of transcription factors (TFs). We continue to see that questions of specificity between TFs require examination of moderate and lower affinity binding sites, further highlighting the need for rich experimental datasets that probe the full range of affinity, and for improve computational/mathematical representations of binding that account for lower affinity binding sites. Finally, it remains clear that there remains a disconnect between protein-DNA binding and function in vivo, and that not all binding is functional, motivating the need to bridge this gap in our understanding. Increasingly we are seeing the integration of large-scale binding data and functional data, such as high-throughput reporter assays, to understand how protein-DNA binding relates to gene regulation. This next step in our general understanding will continue to require new computational and mathematical approaches to model these phenomena. Many of the discussions held at the workshop will likely lead to novel approaches and collaborations towards the goal of a coherent description of how protein-DNA binding leads to gene regulation.

3 Scientific Progress Made

Dr. Gary Stormo gave the keynote talk to open our workshop; he discussed recent works from his lab and described open problems in the field. He described work that adapted the Spec-seq technology to measure cooperativity parameters and how they were able to increase the throughput by over 300-fold. He also described work on adapting the Spec-seq technology to measure the effects of DNA methylation on binding affinity. Dr. Stormo described three computational projects: BEESEM is a maximum likelihood approach to obtain accurate energy values from HT-SELEX experiments; a demonstration of the superiority of energy PWMs over the traditional probability PWMs; a reanalysis of ChIP-seq data using optimized additive PWMs and showing that they are nearly as accurate as more complex models that include shape parameters. Recommendations for the field included using energy PWMs as motifs for TFs, and that optimal motifs should be obtained that depend only on sequence and are agnostic with respect to binding mechanism. Finally he outlined some open problems which primarily focused on using the motifs obtained in vitro to unravel the much more complicated processes that occur in vivo, including cooperativity and competition between factors, the effects of nucleosome modifications and overall chromatin structure.

Dr. Richard Mann discussed three current projects from his lab focus on delineating mechanisms of specificity for the Hox gene family. The first project was an X-ray crystallography study comparing four ternary structures of the Hox protein AbdB bound to four different DNA sequences with Exd. These four DNA sequences differ in affinity for AbdB-Exd by greater than 20-fold; however, the contacts present in the four structures were nearly identical. The second project described experiments to examine the DNA sequence preferences of the protein trimer composed of Hox, Exd, and full-length Hth. Orientation and spacing differences were defined using a variety of SELEX-seq experiments. These differences were due, in part, to the selection of DNA sequences that contained three correctly spaced minor groove width minima, one for each of the three homeodomains. The third project described experiments in which the Hox gene Scr was mutagenized so that it could no longer interact with Exd. ChIP-seq experiments were carried out to demonstrate that only a subset of binding events in vivo is dependent on the interaction with Exd.

Dr. Juan Fuxman Bass presented studies on gene regulatory network (GRN) rewiring by disease-associated genetic variants and by transcription factor (TF) isoforms. GRNs can be perturbed by genetic variants in regulatory regions that affect TF binding or by variants that alter the coding sequence of TFs. Dr. Fuxmann Bass previously developed a pipeline based on enhanced yeast one-hybrid (eY1H) assays to

determine altered TF binding to noncoding variants. By evaluating 109 noncoding variants he detected both loss and gain of TF interactions with mutant loci that are concordant with target gene expression changes. This is also the case even for mutations affecting a single regulatory element such as the ZRS enhancer of SHH that when mutated leads to limb malformations. GRNs can also be rewired by coding variants in TFs. By studying 58 missense variants in 22 TFs, gain and loss of DNA targets were detected mostly in cases of variants affecting the DNA binding domain.

Dr. Raluca Gordan discussed work from her lab focused on how paralogous transcription factors (TFs) distinguish unique targets genes, and how DNA mismatches affect TF-DNA binding. Dr. Gordan's lab has found that even in vitro paralogous TFs interact differently with their genomic targets. The specificity differences between TF paralogs are concentrated in the medium and low affinity binding ranges, which explains why they were missed by previous DNA-binding data and models. These differences identified in vitro help explain the differential in vivo binding profiles of paralogous TFs as captured by in vivo ChIP-seq data. DNA mismatches (or mispairs) occur when two non-complementary bases are aligned on opposite strands of a DNA duplex, forming a 'mispair'. Dr. Gordan described SaMBA (Saturation Mismatch Binding Assay), the first assay to characterize the effects of mismatches on TF-DNA binding in high-throughput. She described how SaMBA quantitatively assesses the effects of the mismatches on the binding specificity of TFs. Her lab found that certain DNA mismatches within TF binding sites can significantly increase TF binding levels compared to the wild-type sequences. The high affinity of TFs for mispaired DNA has the potential to influence gene expression and DNA repair processes, especially in mismatch repair-deficient cells.

Dr. Eric Ortlund discussed work in his lab to understand how paralogous TFs evolve divergent DNA specificities. He described how his lab examined how the glucocorticoid receptor and its paralogs evolved to bind activating response elements [(+)GREs] and negative glucocorticoid response elements (nGREs). He showed that binding to nGREs is a property of the glucocorticoid receptor (GR) DNA-binding domain (DBD) not shared by other members of the steroid receptor family. He showed that the ancestral DBD from which GR and its paralogs evolved was capable of binding both nGRE and (+)GRE sequences because of the ancestral DBD's ability to assume multiple DNA-bound conformations. Subsequent amino acid substitutions in duplicated daughter genes selectively restricted protein conformational space, causing this dual DNA-binding specificity to be selectively enhanced in the GR lineage and lost in all others. These amino acid substitutions subdivided both the conformational and functional space of the ancestral DBD among the present-day receptors, allowing a paralogous family of transcription factors to control disparate transcriptional programs despite high sequence identity.

Dr. Trevor Siggers described a new high-throughput methodological approach from his lab – nuclear extract PBM (nextPBM) – in which protein-DNA binding can be assayed directly using cell nuclear extracts, providing a platform in which to understand the role of cell-specific cofactors and post-translational modifications on protein-DNA binding. Dr. Siggers presented a study of the myeloid lineage factor PU.1 in which he demonstrated how the nextPBM approach could be used to characterize both autonomous PU.1 binding, and cooperative binding with IRF8. Comparison of DNA-binding profiles of purified PU.1 protein and PU.1 from nuclear extracts using the nextPBM approach readily identified the cooeprativley bound DNA sites. He outlined how using a single-nucleotide variant (SNV)-based approach to study protein-DNA binding specificity they could characterize the impact of SNPs on protein-DNA binding and define the impact of DNA bases on protein-DNA binding and cooperativity. He then outlined how this approach can be used to discover new cooperative acting transcriptional complexes.

Dr. Alejandra Medina-Rivera discussed her current research on vascular endothelial cells (EC). EC play an essential role in maintaining homeostasis of arteries and veins, to discover gene regulatory features of aortic endothelial cells Alejandra and her collaborators first ascertained the genomic occupancy of six histone modifications, CTCF, cohesin complex member RAD21, and JUN (c-JUN) transcription factor (TF) in human aortic endothelial cells (HAEC). Comparing this data to equivalent data generated for human umbilical vein endothelial cells (HUVEC) revealed 3000 active chromatin regions specific to each cell type and included known arterial venous marker genes. Furthermore, by performing motif discovery analysis using word counts Alejandra was able to retrieve a combined motif JUN-ETV, which is consistent with the motifs shown by Jose Rodriguez-Martinez (Univ. of Puerto Rico, USA) in his talk. To further identify aortic endothelial enhancers that are likely to play a functional role in EC gene regulation, Alejandra performed inter-species comparisons of JUN and H3K27ac occupancy in primary cow, rat and mouse aortic endothelial cells.

Dr. Yael Mandel-Gutfreund presented combined experimental and computational effort in her lab to

identify proteins with dual DNA binding and RNA binding functions. In the context of tight interactions between transcriptional and post-transcriptional regulation, proteins that bind specifically both DNA and RNA are highly likely to be key players in mediating the cross talk between the different processes of gene expression pathway. Dr. Mandel-Gutfreund presented the technology used to extract RBPs in cells, named RNA Interactome Capture. Employing this technology, they managed to capture over 800 high-confidence RBPs including 175 candidate DRBPs. Yael presented different computational tools they have employed to search for the unique properties of DRBPs that possibly grant their dual binding ability. Finally, Yael presented results from a collaborative work with the group of Martha Bulyk (previous workshop attendee), studying the DNA binding specificities of selected DRBPs.

Dr. Marcus Noyes discussed advanced synthetic screens of Cys2His2 zinc finger (C2H2 ZF) libraries that provide insight into how adjacent zinc fingers influence the DNA-binding preferences of one another. Dr. Noyes and others have taken a reductionist approach in the attempt to understand the rules of how individual zinc finger domains. These attempts have been quite successful at predicting the function of a single zinc finger but fail to capture the rules of how these domains function in the context of a many-fingered transcription factor. To address this issue Dr. Noyes reported his lab's study of how adjacent zinc fingers influence the specificity of one another. He demonstrated how the binding preference of an adjacent finger can influence the target preference of its neighbor. Further, that the influence is controlled by the side chains used by the adjacent finger. Dr. Noyes outlined how these preliminary results are being expanded in his lab to exhaustively screen the influence adjacent fingers have on one another and how these results will be modeled to provide a predictive code to understand the base preference of all human C2H2 ZF transcription factors.

Dr. Miles Pufall described his lab's work on specificity of the CRISPR-Cas9 system. They used SelexGLM to construct a comprehensive model for DNA-binding specificity and observed that 13-bp of complementarity in the PAM-proximal DNA contributes to affinity. Using Spec-seq to measure the effect of mismatches throughout the binding site on affinity, they found dramatic differences in the specificity of binding for Cas9-RNPs with different gRNA sequences. They systematically compared the impact of gRNA:DNA mismatches on affinity and endonuclease activity, using a newly developed technique, SEAM-seq. These simple and accessible experiments identified opposing effects for mismatches on DNA-binding and cutting for complexes with higher specificity. These paired techniques allow development of integrative models to estimate catalytic efficiency and that will lead to a better understanding of how sequence recognition is coupled with cleavage.

Dr. José A. Rodríguez-Martínez presented his worked on combinatorial DNA-binding by human bZIP transcription factors. Using HT-SELEX, the DNA-binding interactomes of 270 bZIP (36 homodimers and 234 homodimers) were examined. The DNA interactomes of 80 heterodimers and 22 homodimers revealed that 72 percent of heterodimer motifs correspond to conjoined half-sites preferred by partnering monomers. Importantly, their analysis identified several bZIP heterodimer-specific DNA binding sites that were strongly bound by the heterodimer but not bound or weakly bound by either homodimer. He also demonstrated that they identified 156 disease and quantitative trait associated single nucleotide polymorphisms predicted to significantly alter binding by bZIP proteins.

Dr. Alexandre Morozov discussed the major role of Mediator complex in establishing and maintaining higher-order chromatin structure in yeast. Morozov and his collaborators, James Broach (Penn State U, USA) and Stefan Bjorklund (Umea U, Sweden), used ChIP-seq to establish that Mediator binding was enriched at the boundaries of chromosomally interacting domains (CIDs; topological domains that are principal units of 3D chromatin organization in yeast). Surprisingly, Mediator binds preferentially to the strongest CID boundaries which delineate the most well-defined chromatin domains corresponding to promoters of highly transcribed genes. Moreover, Mediator co-localizes with the RSC chromatin remodeler complex and the cohesin loading complex, pointing to their yet poorly understood collaborative roles in establishing higher-order chromatin structure. Thus Mediator forms an essential part of "transcription factories" in yeast.

April Mueller discussed specificity of the Cys2His2-type zinc fingers (ZFs). She discussed her work concerning inaccuracies in predictions of ZF specificity and investigations into how the orientation of a ZF domain with respect to its neighbor changes the DNA binding specificity by using a protein-centered bacterial-1-hybrid assay. Her work elucidates variations in ZF binding strategy between ZFs in six different binding modes, revealing an importance to account for these interactions in future prediction methods.

Dr. Tom Tullius discussed the question of whether DNA-binding proteins recognize intrinsic structural

features in DNA (DNA shape), or whether protein binding induces changes in DNA structure. The Tullius lab has developed an experimental approach to this question, which involves the use of the chemical probe hydroxyl radical to make a nucleotide-resolution map of minor groove shape variation in a naked DNA molecule. In collaboration with the lab of Remo Rohs (another workshop attendee), they have used analyzed hydroxyl radical cleavage patterns of 11 protein binding sites. For 7 of the 11 sites, regions of narrow minor groove width found in the naked DNA persisted in the protein-DNA complex, and were found to interact electrostatically with arginine residues from the protein. Dr. Tullius then showed how his lab is now using Illumina DNA sequencing to make this experimental method for mapping DNA structure into a high-throughput technique, with the potential to make structural maps of large stretches of naked genomic DNA in context.

Ignacio Ibarra presented an analysis of publicly available SELEX data to infer features that contribute to the formation of protein-DNA ternary complexes. For handling sparse counts in long k-mer enrichment tables obtained from CAP-SELEX data, his research group developed a method to quantify the contribution of DNA structure features in a reference k-mer using tiled k-mers. Applying this approach, they reported a strong contribution for DNA structure features in a specific combination of transcription factor families, Forkhead+Ets, in which prediction improvements are significantly greater with respect to all tested family pairs. He also discussed how these models can be used to predict ternary complexes binding in QTL/GWAS data which are currently unexplained by single transcription factor binding models.

Jeff Spencer discussed a DNA-cleavage screen that was used to investigate the effects of amino acid substitutions on Cas9 nuclease activity and specificity. A screen of over 8500 mutations revealed how mutations in the Pam-interacting domain of Cas9 affect DNA-target mismatch tolerance. Mutations to this domain decrease the ability of the enzyme to cleave mismatched target sequences as demonstrated by a GFP cleaving assay in mammalian cells. The elucidation of this domain as a contributor to the overall sequence discriminatory potential of the enzyme establishes new directions for research into the mechanism of target selection by Cas9. It also provides new avenues for generating variants with improved fidelity.

Dr. Matt Weirauch discussed his recently published work on viral TFs contributing to autoimmunity. The key finding of this work is that the Epstein-Barr virus (EBV) EBNA2 protein occupies up to half of the genetic risk loci in the human genome for a set of seven autoimmune diseases. These results are important because they might provide a shared molecular mechanism underlying the already established ties between EBV and many of these diseases, offering many therapeutic possibilities. These finding were based on two novel algorithms developed in the lab for this study. The first algorithm is Regulatory Element Locus Intersector (RELI), which systematically estimates the significance of the overlap between genetic loci of a given disease and the genomic binding events of transcription factors (TFs). The second algorithm is Measurement of Allelic Ratios Informatics Operator (MARIO), which identifies genetic risk allele-dependent TF binding events. Collectively, these results nominate mechanisms operating across disease risk loci, suggesting new paradigms of disease origins.

Dr. Manuel Llinas discussed his lab's work on malaria. A unique feature of the malaria parasite is that the genome encodes only a single expanded family of DNA binding proteins called the Apicomplexan AP2 (ApiAP2) proteins. His lab is focused on an in-depth characterization of these transcription factors. Using a combination of in vitro and in vivo approaches such as protein binding microarrays (Campbell et al. 2010) and ChIP-seq the lab has found a wide range of DNA binding specificities and targets for these ApiAP2 proteins. Interestingly, several of these proteins bind highly similar motifs and appear to be bifunctional, acting as both repressors and activators of transcription. Such overlapping specificity has previously been reported for homeobox domains, among others and questions about how specificity is achiever remain open areas of investigation.

Dr. Charles Vinson discussed recent work his lab has been doing in extending protein binding microarray (PBM) technology to explore sequence-specific DNA binding. The first project studied the effect of modified cytosine nucleotides, such as 5-methylcytosine (5mC) and its oxidative byproducts 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxycytosine (5caC), on transcription factor (TF) binding specificity. Dr. Vinson presented results for CEBPB—ATF4 heterodimers binding double-stranded DNA containing cytosine or double-stranded DNA where cytosine in all CG dinucleotides is enzymatically methylated to 5mC. Several binding sites of the form CGAT—G are identified to be well bound only when the CG dinucleotide is methylated. Dr. Vinson described mutant TF constructs designed to bind only 5mC-containing sites and presented data on CEBPB(V285A) and several point mutations in Zta, a bZIP protein encoded in

the Epstein-Barr virus genome.

Dr. Julio Collado-Vides presented on a comprehensive collection of knowledge about regulation of initiation of transcription in E.coli. His analysis showed that proximal sites are required in the sigma 70 type of promoters, that activators are located in general at upstream positions relative to the binding regions of RNA polymerase, whereas repressor sites show a wider distribution. Multiple site analysis offered an initial perspective of cooperative or synergistic interactions, although it is clear that it is hard to generate mechanistic hypothesis bases on relative distances alone. He also discussed the importance of the extension of controlled vocabularies in the organization of gene regulation, showing the differences between the classical definitions and those currently used particularly given the high throughput technologies in studies of gene regulation. Dr. Yaakov (Koby) Levy discussed the complexity of protein-DNA interactions that may exhibit some conflicting thermodynamic and kinetic properties. The fast association between proteins and DNA is governed by nonspecific interactions that allow protein sliding along DNA where the protein binds DNA nonspecifically and performs a helical motion when it is placed in the major groove. His lab has explored using various computational approaches the interplay between the molecular characteristics of the proteins (e.g., DNA recognition motifs, degree of flexibility, and oligomeric states) and the nature of sliding, intersegment transfer events and the overall efficiency of the DNA search. Another important aspect of the search is how the in-vivo conditions (for example, crowding in the cell or coverage of DNA by nucleosomes) affect the efficiency of DNA search. Dr. Levy discussed the molecular features of proteins and of the nucleic acids that allow fast dynamics and high affinity binding on both single- and double-stranded DNA.

Chaitanya Rastogi discussed the NRLB (No Read Left Behind) algorithm that builds high-quality protein-DNA recognition models directly from high-throughput SELEX data. It does so by combining a biophysical model of protein-DNA interaction with a statistical model of sequencing read selection and uses maximum-likelihood methods to infer binding free energy parameters associated with sequence features. This methodology enables NRLB to build models without counting kmers, seeding, filtering reads, and aligning sequences. The resulting sequence-to-affinity models capture DNA binding specificity over the entire affinity range for arbitrarily large footprints. NRLB models predict human Max homodimer binding in near-perfect agreement with existing low-throughput measurements over the entire (¿100-fold) affinity range.

Dr. Remo Rohs discussed his work on high-throughput prediction of DNA shape and its role in protein-DNA binding. High-throughput DNA shape prediction established a cross talk between the fields of structural biology and genomics. The primary goal of DNA shape analysis remains the quest for mechanistic insights into protein-DNA readout modes based on sequencing data without the need of structure determination. Dr. Rohs described studies from his lab that emphasized the importance of interactions between nucleotide positions within a binding site and its flanks, although the definitions of DNA sequence versus shape still differ in structural biology and genomics.

Dr. Harmen Bussemaker discussed two algorithms that his team recently developed for building feature-based models of DNA sequence readout by transcription factors from various types of high-throughput SELEX data. The NRLB algorithm allows them to build models of unprecedented quality which capture DNA binding specificity almost perfectly over a ¿100-fold affinity range and arbitrary footprint size. In another methodological contribution, the Bussemaker lab recently developed a new framework for systematically analyzing the role of DNA shape readout in protein-DNA recognition. The final part of the talk addressed the effect of DNA modifications such as cytosine methylation on transcription factor binding, an emerging topic in the field.

Julia Rogers presented work from her PhD in Martha Bulyk's lab studying mechanisms of DNA binding specificity within the forkhead family of transcription factors. Forkhead factors share a highly conserved DNA binding domain, and typically recognize a canonical forkhead DNA motif (GTAAACA). However, some forkhead subfamilies have evolved the ability to recognize a very different DNA motifs (GACGC). In her talk, Julia presented the crystal structures of the human protein FoxN3 in complex with both the canonical and alternate DNA sequences. Unexpectedly, FoxN3 adopts a remarkably similar conformation to recognize both motifs, making contacts with different DNA bases using the same amino acids. These structures present a new mechanism by which a single DBD can recognize two dramatically different DNA sequences, revealing the importance of DNA structure in transcription factor DNA recognition.

Ashley Penvose discussed her work on the Nuclear Receptors (NRs) proteins. In her research she used custom-designed protein binding microarrays to characterize the binding specificity of the type II NRs. She have found an unexpected amount of shared binding specificity, including high affinity binding to half-sites

and direct repeats of non-canonical spacer lengths. For some NRs, such as PPARy:RXRa, spacing preference is a strong determinant in binding specificity; while for others, such as LXRa:RXRa, spacing does not play a large role in determining binding specificity. Integration of ChIP-Seq, RNA-Seq, and DNase-Seq with her PBM-derived binding models revealed a strong disconnect between binding specificity and regulatory specificity.

Dr. Bart Deplancke discussed his work on KRAB-type Zinc Finger proteins (KZFP). He described how they have recently uncovered one very poorly characterized KZFP, ZFP30, as a top hit in a large-scale TF overexpression screen aimed at identifying novel pro-adipogenic regulators. Experimental follow-up now revealed that ZFP30 promotes adipogenesis by directly targeting and activating a retrotransposon-derived Pparg2 enhancer, suggesting a process of adipogenic exaptation. Their findings provide a new understanding of both adipogenic and KZFP-KAP1 complex-mediated gene regulation. Dr. Deplancke also discussed his lab's SMiLE-seq methodology, which is a new microfluidic tool aimed at determining the DNA binding specificity of single and heterodimeric TFs over a wide affinity range.

Dr. Yaron Orenstein presented his new algorithm DLPRB: deep neural network (DNN) approach for learning protein-RNA binding. The algorithm predicts the complexes formed by binding of proteins to RNAs. The key computational challenge is to efficiently and accurately infer RNA-binding models that will enable prediction of novel protein-RNA interactions to additional transcripts of interest. He presented two different network architectures: a convolutional neural network (CNN), and a recurrent neural network (RNN). The results in inferring accurate RNA-binding models from high-throughput in vitro data exhibit substantial improvements, compared to all previous approaches for protein-RNA binding prediction (both DNN and non-DNN based). By visualizing the binding specificities, novel biological insights underlying the mechanism of protein RNA-binding can be gained.

Dr. Joe Thornton discussed work using ancestral protein reconstruction and biochemical and cell-based assays to examine the evolution of TF specificity. They have shown that a major evolutionary transition occurred about 500 million years ago, from a single ancestral protein that recognized estrogen-receptor response elements (ERE) to a family of extant factors that now recognize glucocorticoid-receptor-like response (GRE) elements. Three historical substitutions in the protein's recognition helix were the primary cause of this shift in specificity, but they were tolerated only because another set of historical permissive substitutions elsewhere in the protein, which do not affect specificity, occurred during the same period. They also found that epistatic interactions among amino acids in the recognition helix are critical to this family of proteins' DNA specificity, undermining the idea of an easily reducible "recognition code" for protein-DNA interactions, and it is only because of this epistasis that it was possible for novel DNA specificity to readily evolve through short mutational paths.

Dr. Tim Hughes described collaborative efforts with the Weirauch and Taipale labs (Dr. Weirauch and Dr. Taipale were both workshop attendees) to complete a collection of motifs for human transcription factors, and steps towards comprehensive analysis of their effector activities, centred on a recently published expert curation of human transcription factors. Several groups have implemented large-scale motif analyses in the last decade, both in vivo and in vitro, using both new and established techniques, but there is ongoing disagreement on the number of transcription factors. The new TF catalog presented contains hundreds of proteins not listed in previous widely used human TF lists, thus enabling a new perspective to be gained on the domain structures, functional and genetic properties, expression patterns, and evolutionary trajectories of human TFs. This new index provides an updated reference set of human TFs and their binding motifs as well as a revised synopsis of their functional properties, underscoring that mapping the mechanisms that regulate and express the human genome remains an ongoing challenge.

Víctor H. Tierrafría talked about the development of the Microbial Conditions Ontology (MCO), which is a collection of terms to describe growth conditions of experiments performed in bacteria in a controlled and structured way. This description is based on a ten-item framework of annotation that provides the minimal information necessary to support experimental reproducibility and comparability. This implementation represents another step towards efficient, accurate, and interoperable retrieval, comparison, and analyses of the biological information in accordance with the purpose of RegulonDB to follow the FAIR data principles.

Dr. Barak Cohen discussed recent work from his lab investigating chromosome position effects. Chromosome position effect refers to the phenomenon where a gene is expressed at very different levels when it is moved to a new location in the genome. Cohen discussed efforts using a new version of Massively Parallel Reporter Gene Assays (MPRAs) a library of reporter genes containing different cis-regulatory elements

was assayed in several different chromosome locations. They key result was that the relative strength of cis-regulatory elements was preserved across chromosomal locations even though the activities of all elements were scaled either up or down at different locations. Overall the results suggest a modular organization of the genome in which different types of cis-regulatory elements contribute independently to gene expression without complex interactions.

4 Extended Out-reach Initiative

A novel part of this year's workshop was an extended 1-day out-reach initiative that had researchers from the workshop meet with high-school students from the Oaxaca area to discuss their work and their careers in research and science. With the help of CMO's Claudia Arias Cao Romero, Dr. Noyes (workshop co-organizer) was put in touch with Dr. Bruno Cisneros, Professor of Mathematics at the UNAM in the Instituto de Matemáticas. They organized an outreach program to take place the Friday that followed the 2018 "Rules of Protein-DNA Recognition" CMO meeting. Several participants from the meeting spent this extra day in Oaxaca and gave short, introductory level talks about their work or their field in a one-day seminar at the IEBO school in Ciénega de Zimatlán, Oaxaca. Critical to the success of the event was Dr. Cisneros' arrangement for translators Dr. Marcelino Ramírez Ibañez and Dr. Beatriz Carely Luna, as several speakers did not speak Spanish. Two schools from the region transported children in the early hours of the morning, one was over 2 hours away, just to be at this event. Roughly 125 students were able to attend the seminars that ranged in topics from "Why are there mountains?" to "Fun with brains" to "You are a bag of 100 trillion cells". By all accounts it was an extremely successful event. Several speakers noted it was the highlight of their trip and how special it was to have this opportunity to give something back when we are so privileged to attend these meetings in the first place! Most importantly, the students were thrilled from beginning to end, asking countless questions after each talk and all wanted to take pictures with the speakers afterwards. One student said to the organizer afterwards, "I'm going to figure out how to join your lab!" It was truly a great success. Dr. Noyes has committed to continue to organize these events for future meetings.

5 Outcome of the Meeting

The primary outcome of the meeting was a continued dialog between researchers with diverse backgrounds and approaches about a common problem. This workshop is highly unique in this respect and many attendees comment on how refreshing it is to hear diverse talks and interact with new researchers thinking about similar issues. Furthermore, bringing together researchers with unique computational and experimental approaches offered a fantastic opportunity to foster discussions that will undoubtedly lead to productive collaborations. A follow-up workshop would certainly multiply the outcome and impact of this very successful workshop.