# Rules of Protein-DNA Recognition: Computational and Experimental Advances

Harmen J. Bussemaker (Columbia University, New York, NY, USA),
Remo Rohs (University of Southern California, Los Angeles, CA, USA)

June 21- 26, 2015

## 1 Background

All genetic information in a cell is stored in its DNA, which provides a highly stable medium for information storage and retrieval. To make use of this stored information, the DNA template must be transcribed to messenger RNA, which in turn is translated to proteins. Gene transcription is regulated by proteins called transcription factors, whose function is to bind to specific DNA sites in the vicinity of their target genes. Understanding the physico-chemical rules of protein-DNA recognition is a central problem in molecular biology. Since binding sites are short (typically 6-12 nucleotides in length) and degenerate (multiple DNA sequences have similar binding affinities to a given protein), the problem of identifying them in the genome presents a significant mathematical challenge. This BIRS workshop brought together researchers who use multiple experimental and theoretical approaches to studying protein-DNA interactions, from detailed molecular analysis and simulations to massively parallel experiments designed to find all instances of proteins bound to their genomic sites in living cells.

The workshop focused on recent progress in understanding structural and energetic mechanisms that enable DNA-binding proteins (such as transcription factors) to bind their cognate genomic sites with high affinity and specificity. The main purpose of the workshop was to bring together researchers with diverse approaches and perspectives to studying protein-DNA recognition, not just experimental versus computational but also different approaches within each of those areas. The goal was to obtain a better understanding of how transcription factors achieve specificity and how specificity can be modeled and predicted. A second goal was to improve the methods for the design of proteins with novel DNA-binding interfaces and specificities. Advances in understanding the rules of protein-DNA recognition will lead to better understanding of such fundamental biological processes such as DNA replication and transcription, with numerous future applications in biotechnology and medicine. Recent breakthroughs in developing high-throughput experimental techniques make the development of novel mathematical and computational frameworks for analyzing protein-DNA interaction data a high priority. The proposed workshop helped fulfill this urgent need.

The organizers are both pioneers in this field. Dr. Harmen Bussemaker, a Professor in the Department of Biological Sciences and the Department of Systems Biology at Columbia University, is widely known for his pioneering computational work aimed at understanding gene regulatory networks based on the integration of genome sequence, transcription factor binding, and gene expression data. His credentials include a Lenfest Distinguished Columbia Faculty Award and a John Simon Guggenheim Foundation Fellowship. Dr. Remo Rohs, an Associate Professor in Departments of Biological Sciences, Chemistry, Physics, and Computer Science at the University of Southern California, is known for his pioneering work on integrating DNA shape into studies of protein-DNA binding. Dr. Rohs is also an Alfred P. Sloan Research Fellow and recipient of

the OpenEye Award from the American Chemical Society. The laboratories of both organizers are funded by the National Institutes of Health.

Our invitation to participate in the workshop was met with great enthusiasm by the researchers in the field. The participants included prominent experimental biophysicists and biologists who are the forefront of the field, as well as mathematicians and physicists whose main focus is on mathematical modeling of protein-DNA interactions. We took special care to include a significant number of promising young scientists in the list (e.g. Phil Bradley, Raluca Gordan, Trevor Siggers, Matthew Slattery, Matthew Weirauch). Eight of the confirmed PI participants were female scientists (Martha Bulyk, Polly Fordyce, Raluca Gordan, Christina Leslie, Karolin Luger, Wilma Olson, Mona Singh, Zhiping Weng). We also reserved spots for graduate students and postdoctoral associates who are relatively new to the field, as well as for researchers who have relevant late-breaking findings. We were also careful to include participants from groups traditionally underrepresented in the sciences.

One of the main goals for this meeting was to facilitate free exchange of ideas and foster new collaborations, both formal and informal. A previous workshop in the summer of 2013 was hugely successful and the overwhelming consensus was to continue this dialog every two years. The 2015 organizer, who was an attendant at the 2013 meeting, is aware of at least six new collaborations that were initiated as a direct result of the 2013 BIRS meeting. Given its success, the overall agenda and goals for the 2015 meeting closely followed that of the 2013 meeting. 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 had both experimental and theoretical/computational components in their labs, they were encouraged to cover recent advances using both approaches in their talks. 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 adopted a format in which 30-minute talks (+10 min for questions) will be scheduled in groups of three or four. Every invited participant was encouraged to give a talk, though it will not be mandatory. Each group of talks was followed by a tea/coffee break during which the participants had a chance to discuss the latest lectures in more detail. We also scheduled time for round-table discussions and thematic focus groups that could be devoted to some of the more technical aspects of experiments and mathematical models. We also included social activities such as a hiking trip to promote scientific discussions and forge new relationships.

Judging from the informal and formal feedback of the attendees, the meeting was again a great success. Dr. Gary Stormo, who co-organized the first edition of this meeting in 2013, wrote ”The workshop was great. Nearly every talk was related to things we do and contained relevant new information. The schedule had enough free time to allow for extended discussions which often continued over meals. I got to meet many people in the field for the first, and reacquaint with those I’ve known before. Overall excellent.” Another senior attendee, Dr. Tom Tullius, wrote ”This workshop was one of the best scientific meetings I have ever attended. The participants were chosen very carefully to represent a wide range of experience in the field (postdocs to senior scientists), but with a clear focus on the scientific area of the workshop (protein-DNA recognition). This made for easy and effective communication among all the participants. I learned a lot from everyone. The scientific focus of the workshop almost perfectly fit my own research interests - I realized when I looked at the program that the workshop participants were all scientists whose papers I invariably read when I come across them in the literature. I especially appreciated the opportunity to meet scientists from Central and Latin America, and this led to at least one future collaboration between my lab and a scientist from South America.” Everybody agreed that this meeting brought together a unique combination of researchers, and there was strong support to get this series going in future years.

## 2 Recent Developments and Open Problems

The organizers designed the workshop in a way that brings together researchers at the forefront of experimental and/or computational studies of protein-DNA binding. Work in this field has a long history in structural biology, a field that answered how a large number of proteins bind DNA. Three-dimensional structures of protein-DNA complexes revealed how different families of proteins (e.g., homeodomains or zing finger proteins) bind their respective DNA target sites. However, structural biology did not answer how members of

the same protein family (i.e., paralogous transcription factors) bind to different, albeit very similar, DNA targets in the genome, or why only a very small subset of putative binding sites in the genome is functional in vivo. In recent years, the advent of deep sequencing has generated a vast amount of data that describes DNA binding specificities of transcription factors in much detail. The availability of this data led to the development of many computational approaches for analyzing DNA binding specificity data efficiently. Now is the time when prior knowledge gained from individual structures can be integrated with high-throughput sequencing data. This workshop brought researchers from these different fields together, and many of the discussions during this workshop will likely lead to different approaches, new collaborations, and ultimately new answers to the key questions in the field.

# 3 Scientific Progress Made

Christina Leslie presented a powerful new statistical approach for learning the recognition code of a family of transcription factors or RNA-binding proteins (RBPs) from high-throughput binding assays. Their method, called affinity regression, trains on protein binding microarray (PBM) or RNAcompete experiments to learn an interaction model between two kinds of inputs, protein domain sequences and DNA or RNA probe sequences. Trained on a large mouse homeodomain PBM data set, the model correctly identifies residues that confer DNA-binding specificity and accurately predicts binding motifs for a new independent test set of homeodomains from many divergent species. Similarly, learning from RNAcompete data for diverse RBPs, the model can predict the binding affinities of held-out proteins and identify key RNA-binding residues, despite the high level of sequence divergence. [1]

Zhiping Weng spoke about a method that combined the various features of DNase cleavage and footprint with sequence motif to predict transcription factor binding sites. [2, 3]

Sebastiaan H. Meijsing addressed how transcription factors know where to go in the genome. The glucocorticoid receptor (GR) was one of the first transcriptional regulatory factors for which a DNA recognition sequence was identified now more than three decades ago. However, known recognition sequences only partially explain where GR binds to the genome. For example, genome-wide analysis of GR-bound loci showed that a large subset op peaks contains neither a canonical nor other known motifs that can tether GR to the DNA. Furthermore, although millions of sequences matching recognition sequences are found in the genome, only 1 out of approximately 1000 potential sites actually recruits GR. Here we set out to identify genomic sequences that specify where in the genome GR binds. The analysis of high-resolution ChIP data (ChIP-exo) uncovered that GR can be recruited to the DNA by a broader spectrum of sequences than previously known. Furthermore, another way to specify where GR binds might be through sequence signals that prevent GR recruitment to certain loci. If such signals exist, they should be depleted in genomic regions where GR binds. Computational analysis of ChIP-seq data for GR identified several candidate negative regulatory sequences that interfere with genomic binding of GR. Together, these studies highlighted several known and unknown mechanisms that provide a better understanding of how GR targets the genome. [4]

Martha Bulyk showed how Transcription Activator-Like Effector (TALE) proteins recognize DNA using a seemingly simple DNA-binding code, which make them attractive for use in genome engineering technologies that require precise targeting. While this code is used successfully to design TALEs to target specific sequences, off-target binding has been observed and is difficult to predict. They explored TALE-DNA interactions comprehensively by quantitatively assaying the DNA binding specificities of 21 representative TALEs to 5,000-20,000 unique DNA sequences per protein using custom-designed protein binding microarrays (PBMs). They found that protein context features exert significant influences on binding. Thus, the canonical recognition code does not fully capture the complexity of TALE-DNA binding. They used the PBM data to develop a computational model, Specificity Inference for TAL-Effector Design (SIFTED), to predict the DNA-binding specificity of any TALE. SIFTED is offered as a publicly available web tool that predicts potential genomic off-target sites for improved TALE design. [5]

Gary Stormo presented new methods for determining specificity and cooperativity, including *Spec-seq*, a method for determining the specificity of protein-DNA interactions by sequencing the bound and unbound fractions from a binding reaction. Relative affinity is directly measured by the ratios of those ratios, making the analyses particularly easy and not dependent on fitting to any model. Thousands of alternative binding sites can be assayed in parallel with very high precision. Cooperativity, and how it depends on the sequence,

can also be assayed by using Spec-seq and separating both individually bound and co-bound fractions. The talk also described an EM-based approach to determining protein-DNA specificity from HT-SELEX (or SELEX-seq) experiments.

Wilma Olson discussed how DNA topology confers sequence specificity to nonspecific architectural proteins. The organization of long genomes in the confined spaces of a cell entails special facilitating mechanisms. A variety of architectural proteins play key roles in these processes. One such protein is the bacterial protein HU, which helps to condense DNA by introducing sharp bends in the double helix. The protein binds in a sequence-neutral fashion and randomly distorts linear DNA when introduced in computer-simulated structures at levels comparable to those found in the cell. The intrinsic tendency for DNA to retain a naturally straight structure, however, restricts the nonspecific protein to specific loci when the molecule is covalently closed or looped by a protein. Moreover, the rotational settings imposed on DNA by loop-mediating proteins, such as the E. coli Lac repressor assembly, introduce sequential specificity in HU placement. Thus, an architectural protein with no discernable DNA sequence-recognizing features becomes site-specific and potentially assumes a functional role upon loop formation. [7]

Yaron Orenstein discussed how high-throughput SELEX (HT-SELEX) allows a high-resolution measurement of transcription factor (TF) binding preferences. First, he compared the models learned by two high-throughput in vitro technologies: protein binding microarrays (PBMs) and HT-SELEX, spanning 162 different TFs. While by and large, the technologies agree, there are notable exceptions. For some TFs, the HT-SELEX-derived models are longer versions of the PBM-derived models, whereas for other TFs, the HT-SELEX models match the secondary PBM-derived models. Remarkably, PBM-based 8-mer ranking is more accurate than that of HT-SELEX, but models derived from HT-SELEX predict in vivo binding better. In addition, his work revealed several biases in HT-SELEX data including nucleotide frequency bias, enrichment of C-rich k-mers and oligos and underrepresentation of palindromes. It produced a better understanding of the pros and cons of each technology and can lead to the development of improved binding models.

Alexandre V. Morozov spoke about genome-wide profiling of chromatin accessibility and nucleosome positioning in *Drosophila melanogaster*. Chromatin structure and dynamics play pivotal roles in gene regulation, DNA repair and other processes essential to eukaryotic cells. Up to 90% of genomic DNA is occluded by nucleosomes, the fundamental units of chromatin in which DNA is wrapped around the histone octamer surface[6]. Nucleosomal DNA is thought to be uniformly inaccessible to DNA binding and processing factors, such as micrococcal nuclease (MNase). However, he has found that digestion of Drosophila chromatin with high and low concentrations of MNase reveals two distinct nucleosome types: MNase-sensitive and MNase-resistant. MNase-resistant nucleosomes assemble on sequences depleted of A/T and enriched in G/C-containing dinucleotides, whereas MNase-sensitive nucleosomes form on A/T-rich sequences found at transcription start and termination sites, enhancers, and DNase I hypersensitive sites. Estimates of nucleosome formation energies indicate that MNase-sensitive nucleosomes tend to be less stable than MNase-resistant ones. Strikingly, a decrease in cell growth temperature of about 10C makes MNase-sensitive nucleosomes less accessible, suggesting that observed variations in MNase sensitivity are related to either thermal fluctuations of chromatin fibers or the activity of enzymatic machinery. In the vicinity of active genes and DNase I hypersensitive sites nucleosomes are organized into periodic arrays, likely due to 'phasing' off potential barriers formed by DNA-bound factors or by nucleosomes anchored to their positions through external interactions. The latter idea is substantiated by a biophysical model of nucleosome positioning and energetics, which predicts that the +1 nucleosomes immediately downstream of transcription start sites of active genes are anchored through external interactions.

Mark D. Biggin talked about protein/DNA interactions in vivo, and how to predicting DNA occupancy and function. He discussed how *in vivo* animal transcription factors show a quantitative continuum of DNA binding to highly overlapping sets of genomic regions that are located close to most genes. These continua span functional, quasi-functional, and nonfunctional DNA binding events, with transcription factor regulatory specificities being distinguished by quantitative differences in DNA occupancy patterns. Currently, he is using computational models to describe the biochemical mechanisms that produce these patterns of DNA binding and how combinations of transcription factors cooperate to generate spatial and temporal gene expression.

Chaitanya Rastogi presented computational methods for inferring transcription factor specificity from SELEX-seq data. SELEX-seq is an experimental and computational platform that combines biophysical modeling and deep sequencing in order to determine the DNA binding specificity of a transcription factor complexes [8]. Recent work has demonstrated the protocol's ability to elucidate novel recognition properties

of the eight Drosophila Hox proteins [9]. SELEX-seq analyses require detailed oligomer count information to infer affinities, a challenging computational task given the size of the data. Efficient implementations of the computational pipeline are required as the adoption of SELEX-seq increases. Following the methodology set out in [8, 9], he developed a suite of R/Bioconductor functions, named "SELEX," to facilitate the analysis of SELEX-seq data. Thanks to efficient algorithms, this software can run on a standard laptop computer. The package includes functionality for kmer counting, Markov model construction, and information gain (Kullback-Leibler divergence) calculations, along with integrated solutions for painless annotation and management of SELEX-seq experiments. It will form the foundation for future feature-based models of SELEX-seq data.

Barak Cohen pointed out that only 0.1% of consensus binding sites in mammalian genomes are actually occupied by transcription factors in vivo. And, only a fraction of these occupied sites influence gene regulation. How this specificity is achieved is unknown, but is critical to the cell in executing specific programs of gene expression during development, and in response to cellular and environmental perturbations. He hypothesized that sequence information outside what is traditionally considered canonical binding sites must contribute to specificity in large genomes. His work focuses on identifying these 'extra' sources of sequence information which may include flanking nucleosome positioning signals, determinants of DNA shape, and binding sites for cooperatively interacting transcription factors. Ultimately he hopes to gain the ability to predict the location and specificity of mammalian enhancers from their DNA sequence features. [10, 11, 12, 13].

Ana Carolina Dantas Machado discussed how Protein-DNA interactions orchestrate multiple layers of regulation across a vast array of biological processes. To date, studies have shed light into the mechanistic details of how some of these interactions occur, although many modes of regulation largely remain elusive. There is a widespread role for DNA shape readout in regulating proteins in organisms ranging from viruses to mammals. She discussed how hape recognition readout is not exclusively achieved by arginines, but also by lysines and histidines, as in the case of simian virus 40 large T antigen and the ferric uptake regulator, respectively. Modifications of a DNA base pair, in this case CpG methylation, can affect protein-DNA recognition. For CpG methylation. while base readout can be affected due to the insertion of a methyl group on the major groove, the 3-dimensional structure of the DNA can also be affected and therefore play a role on DNA shape recognition, as seen for the enzyme DNase I. She identified the human myocyte enhancer factor 2 as a potential shape reader and discussed current efforts on unraveling how protein-DNA interactions are governed in this case. Through these studies, her lab has been able to expand our knowledge of readout modes achieved during protein-DNA recognition, and they aim to elucidate how specific DNA binding sites are selected and how they can be affected due to variations at the protein-DNA interface. [24, 25, 26, 27, 28].

Michal Levo talked about the unraveling determinants of transcription factor binding outside the core binding site. Binding of transcription factors (TFs) to regulatory sequences is a pivotal step in the control of gene expression. Despite many advances in the characterization of sequence motifs recognized by TFs, our ability to quantitatively predict TF binding to different regulatory sequences within cells, and the resulting expression level of regulated genes, is still limited [14, 15]. The projects presented in her talk aim to characterize determinants of regulatory protein binding, going beyond the investigation of core TF binding sites (TFBSs). They start by exploring the transcriptional effects of nucleosome-disfavoring sequences, namely poly(dA:dT) tracts, that are highly prevalent in eukaryotic promoters, in the vicinity of TFBSs. By measuring promoter activity and nucleosome occupancy for a large-scale promoter library, designed with systematic manipulations to the properties and spatial arrangement of these tracts, they show that they significantly and causally affect transcription (with changes to nucleosome occupancy over the nearby site negatively correlated with the measured transcriptional effect). They demonstrate that manipulating these elements offers a general genetic mechanism, for tuning expression in a predictable manner, with resolution that can be even finer than that attained by altering transcription factor sites [16]. She further present a novel high-throughput in vitro assay termed BunDLE-seq that provides quantitative measurements of TF binding to thousands of fully designed sequences of 200bp in length, within a single experiment. Applying this binding assay to two yeast TFs she demonstrated that sequences outside the core TF binding site profoundly affect TF binding. TF-specific models based on the sequence or DNA shape of the regions flanking the core binding site are highly predictive of the measured differential TF binding. They further characterized the dependence of both single and co-occurring TF binding events, on the number and location of binding sites and on the TF concentration. Coupling in vitro TF binding measurements, and another application of a method probing nucleosome formation, to in vivo expression measurements carried out with the same template sequences

serving as promoters, offers insights into mechanisms that may determine the different expression outcomes observed [17]. Finally, she briefly presented a newly established in vivo binding assay, aiming to uncover the distribution of single-cell binding configurations formed on a large scale library of synthetically designed regulatory sequences. By revealing co-occurring binding events, of both nucleosomes and TFs, this assays offers means to advance our understanding of coorperative and competitive dynamics governing regulatory binding within cells.

Tom Tullius talked about nucleotide-resolution structural maps of DNA and DNA-protein complexes, in vitro and in vivo. His laboratory has developed the hydroxyl radical as a high-resolution chemical probe of DNA structure [18]. He described two unpublished experiments that take advantage of this chemistry to illuminate aspects of how proteins recognize DNA. The first experiment involves subjecting naked DNA to cleavage by the hydroxyl radical. His lab previously showed that the hydroxyl radical cleavage pattern represents a nucleotide-resolution map of the width of the DNA minor groove [18]. For this experiment they synthesized a 300 bp DNA molecule that contained a dozen transcription factor binding sites. The aim was to determine whether pre-existing DNA structural features are recognized by a protein, or whether protein binding induces changes in DNA structure. They found three distinct behaviors: (1) some protein binding sites have nearly the same structure as naked DNA as they do in the protein-DNA complex; (2) in some binding sites, part of the site is the same structure in naked DNA as in the complex, and part of the site changes in structure upon protein binding; (3) some binding sites change substantially in structure when the protein binds. Even in sites that change structure when protein binds, there is usually a hint of the bound structure in the naked DNA. Work from the Tullius laboratory first demonstrated that bound proteins protect DNA from hydroxyl radical-induced cleavage, thereby producing a ”footprint” showing the precise nature of the protein-DNA interface. In the second experiment, he described how to extend the hydroxyl radical footprinting experiment to the entire human genome. He called this experiment OH-seq (hydroxyl radical-seq). The lab generates hydroxyl radicals in living human cells by brief irradiation with gamma ray photons, a simple and non-invasive procedure. Ligation of sequencing adaptors marks the sites of radical-induced strand breaks. High-throughput sequencing on the Illumina platform provides tens of millions of sequence tags, allowing them to map the frequency of strand breaks at single-nucleotide resolution throughout the human genome. They find that radical-induced damage is greater in regions of open chromatin that are depleted in nucleosomes. Nucleosome-free regions often occur in functional non-coding regions of the genome, for example the promoters of active genes. The OH-seq experiment thus allows them to visualize what parts of a genome are susceptible to attack by the hydroxyl radical, and what parts are resistant, providing new information on genome topography. Tullius suggested that this new method has promise for mapping all protein-DNA interactions (nucleosomes, transcription factors) throughout an entire genome.

Robert Kaptein talked about DNA recognition and target location by the E. coli Lac Repressor. The E.coli lac repressor is a text-book example of a bacterial gene regulatory protein. In his lecture he gave an overview of his group's NMR work on the specific and non-specific interactions of lac repressor with DNA. The structure and dynamics of complexes of a dimeric lac headpiece (DNA-binding domain) with lac operators have provided a detailed picture of how various lac operator sequences are recognized [19]. Operator binding is accompanied by a large conformational change and DNA bending. Furthermore, the interaction with natural operators is asymmetric in contrast to what has been observed in the X-ray structure of the lac repressor-O1 complex. To address the problem of non-specific DNA interaction the NMR structure of the dimeric headpiece with non-operator DNA has also been solved [20]. Generally, these non-specific interactions are assumed to be crucial for rapid target-site location by DNA-binding proteins. In the NMR structure of the non-specific lac headpiece-DNA complex the DNA is not bent and the hinge region of the headpiece remain unfolded. Broadening of NMR lines observed in complexes with non-operator DNA could be shown to reflect the sliding of lac headpiece along the DNA and the rate of sliding could be determined [21]. Surprisingly, however, the 1D diffusion constant for sliding obtained from NMR line-broadening is much smaller than that determined by single-molecule fluorescence methods and cannot account for an enhanced target location by lac repressor. I will discuss possible reasons for this discrepancy.

Cliff Meyer discussed the systematic analysis of H3K27ac ChIP-seq for identification of transcriptional regulators and their target Genes. A deeper understanding of the cis-regulatory control of gene transcription is needed to understand the etiology of cancer and other common diseases. ChIP-seq has been used to generate hundreds of genome-wide maps of the histone modification H3K27ac in a variety of human cell types. This histone modification is associated with active enhancers and has been used to describe super-enhancers, cis-

regulatory regions of pre-eminent importance in the regulation of tissue specific and oncogenic genes [22]. Here we adopt a gene centric approach to define a regulatory potential that summarizes the aggregate activity of multiple cis-regulatory elements on each gene. This model is effective in describing specificity in cis-regulatory activity and is highly predictive of gene expression changes in response to the BET-bromodomain inhibitor JQ1 [2]. Using an extensive database of published H3K27ac profiles from a broad variety of human cell types we show how H3K27ac defined regulatory potentials can accurately model diverse gene sets derived from differential gene expression analyses. In addition we demonstrate a semi-supervised learning approach for identifying cis-regulatory elements associated with a set of differentially expressed genes. His method leverages published H3K27ac data to aid the interpretation of newly generated human or mouse H3K27ac ChIP-seq profiles. In addition the method can be used to interpret gene expression studies, without the production of matched H3K27ac ChIP-seq data.

Matthew Weirauch argued that computational prediction of functional transcription factor (TF) binding sites in cis-regulatory regions is of critical importance for the development of comprehensive gene regulatory models. Position weight matrices (PWMs) have proven to be a robust method for encapsulating TF binding specificities, yet most PWM-derived motifs possess a low information content leading to an overabundance of predicted binding sites. Composite motifs (CMs) representing multimeric cooperative element (CE) binding sites for two or more TFs provide a larger, more informative footprint. We hypothesize that CEs favoring discreet stereospecific TF binding configurations with respect to their relative order, orientation, and spacing serve as markers for cis-regulatory regions. We therefore developed the Combinatorics of Stereospecific Motif Orientation (COSMO) algorithm to identify enriched CMs present in input DNA sequences. When we applied COSMOSto an H3K27ac time-course ChIP-seq dataset from LPS-stimulated B cells, the top two CM predictions matched the ETS-FOX CE (EFCE, p¡1e-44) and the ETS-IRF CE (EICE, p¡1e-42). The EFCE has been identified as an enhancer-specific fate driver for endothelial cell types and possibly others, while the EICE has been previously shown to play an important role during B cell differentiation. We also found differential enrichment for many CMs at varying timepoints, implying functionally distinct roles for these binding sites. By applying COSMO to existing ChIP-seq datasets for additional cell types and antibody targets, his work will extend the number of known CEs as well as improve understanding of their role in driving gene network reprogramming and CE-dependent cell fate decisions.

Aseem Ansari talked about ”emergent” cognate sites and ”specificity locks” revealed by Differential Specificity and Binding Energy Landscapes (DiSELs). His lab developed Sequence and Energy Landscapes (SELs) to view the comprehensive specificity profiles of DNA binding molecules. SELs are organized with respect to a seed motif -derived from motif searching algorithms. As such SELs clearly reveal the contribution of each nucleotide with the binding site as well as the impact of flanking sequences on binding affinity for a given DNA binding molecule. These landscapes also reveal ”emergent” sites that differ from the consensus in non-obvious ways and yet display high affinity for a given DNA binder. Many such ”emergent” sites were tested for binding and found to be bona fide targets of the transcription factors or engineered DNA binding ligands. [29] To differentiate between two highly conserved paralogs with identical amino acid side chains that interact with DNA, we developed DiSEL. This approach identified subtle differences in dependencies on specific bases within a binding site. Individually these subtle preferences do rise significantly above noise in the data but collectively when clustered by DiSEL a clear pattern of differential specificity between two nearly identical TFs emerges. DiSEL captures all previously defined differences and reveals several more that might guide factors to different sets of genes (Bhimsaria, Ahmed, et al. in preparation). Evaluating DiSEL-based differences in specificity led to the realization that residues on the Lhx2 and Lhx4 that don’t directly contact DNA contribute to the differences in the DNA sequence preferences of each protein. Extending this observation to other protein-DNA complexes led to the realization that rather than searching for amino acid - base pair recognition codes it was far more predictive to examine the full set of interactions over two base pairs. The ”recognition envelope” can be grafted onto another protein to impart specific recognition of a given two base pair stack. This approach departs from the traditional approaches of looking at single side chains interacting with single nucleobases -which have often failed to confer specificity when grafted on to different proteins, even if they have the same fold. (Sukumar, et al. in preparation). Finally, the use of programmable DNA minor groove binding small molecules to recruit or displace transcription factors from their target sites was also presented. The application toward dissecting the contribution of sequence versus shape to the overall binding of proteins to their cognate sites was discussed and found to be widely interesting to the audience. A new synthetic route to generating libraries of small molecules that target different sequences makes it

possible to test several different molecules on any given site of interest. Members of such a library would permit researchers to perturb minor groove shapes and examine the contributions to overall binding by a transcription factor of interest. [29]

Philip Bradley describe his method for prediction and design of protein:DNA interactions. His lab's research is aimed at building predictive, atomistic models of protein:DNA interactions which can be used to understand and engineer the DNA binding specificity of proteins. In this talk, he described an approach to structural modeling of protein-DNA interactions, benchmark calculations on C2H2 zinc fingers, predictions of the TAL effector:DNA complex structure, and protein design simulations aimed at modifying the DNA-binding specificity of homing endonucleases. He also described work designing novel tandem repeat proteins with structures unlike those seen in nature, with potential applications as new DNA binding platforms.

Judith Kribelbauer presented her work on characterizing orientation and spacer preferences of Hox transcription factor complexes using SELEX-seq. To investigate how Hox proteins achieve target specificity through complex formation with co-factors, the labs of her co-mentors Harmen Bussemaker and Richard Mann recently developed a high-throughput in vitro methodology, SELEX-seq, and applied it to study the binding specificity of the heterodimeric complex between each of the eight D. melanogaster Hox proteins and their common co-factor Extradenticle (Exd) [33, 34]. However, it is known that a second co-factor, Homothorax (Hth), is also important for Hox function. We therefore applied SELEX-seq to study higher-order complexes between the Hox protein Ultrabithorax (Ubx) or Deformed (Dfd) and the co-factors Exd and Hth. Using novel computational methodology required to make accurate inferences from these complex data, we observe great variation in binding mode, in which the relative orientation of the three proteins as well as distance between their respective binding interfaces varies but is dictated by the DNA sequence. The overall stability of the Hth-Exd-Hox-DNA complex varies with the base sequence of the spacer between the Hth and Exd-Hox half-sites, even though this stretch of DNA presumably is not contacted directly. We therefore investigated to what extent the structural properties of the DNA spacer determine its contribution to the overall binding free energy, with the goal of inferring a 'DNA spacer code' for Hox complexes.

Charles Vinson described his efforts to achieve proper spatiotemporal control of gene expression, transcription factors cooperatively assemble onto specific DNA sequences. The ETS domain protein monomer of GABPa and the B-ZIP domain protein dimer of CREB1 cooperatively bind DNA only when the ETS (C/GCGGAAGT) and CRE (GTGACGTCAC) motifs overlap precisely, producing the ETS-CRE motif (C/G CGGAAGTGACGTCAC). We designed a Protein Binding Microarray (PBM) with 60-bp DNAs containing four identical sectors, each with 177,440 features that explore the cooperative interactions between GABPa and CREB1 upon binding the ETSCRE motif. The DNA sequences include all 15-mers of the form C/GCGGA——CG—, the ETS-CRE motif and all single nucleotide polymorphisms (SNPs), and occurrences in the human and mouse genomes. CREB1 enhanced GABPa binding to the canonical ETSCRE motif CCGGAAGT 2-fold, and up to 23-fold for several SNPs at the beginning and end of the ETS motif, which is suggestive of two separate and distinct allosteric mechanisms of cooperative binding. We show that the ETS-CRE array data can be used to identify regions likely cooperatively bound by GABPa and CREB1 in vivo, and demonstrate their ability to identify human genetic variants that might inhibit cooperative binding.

Timothy Hughes argued that the rapid expansion and diversification of C2H2 zinc finger proteins has made this simple domain the most numerous in many metazoans, including human, where it is found in nearly half of all transcription factors. The C2H2 domain is found across all Eukarya, however, and in most lineages it has not amplified. Here, we show that virtually all metazoans possess multiple C2H2 domains that preferentially bind each 3-base DNA sequence. The C2H2 expansion in metazoans, and particularly chordates, is facilitated by widespread contribution of protein 'backbone' residues to binding energy, allowing the base-contacting 'specificity' residues to mutate without catastrophic loss of affinity for DNA In contrast, the restricted C2H2 binding vocabulary found in plants, fungi, and other lineages is explained by a reliance on DNA-contacting residues for affinity. Thus, simple and fundamental properties of a single small domain backbone have contributed to pervasive differences between major eukaryotic lineages, including striking differences in evolutionary mechanisms of gene regulation.

Lin Yang talked about dissecting the role of DNA shape readout for different transcription factor families - Transcription factor binding sites (TFBSs) are most commonly characterized by the nucleotide preferences at each position of the DNA target. Whereas these sequence motifs are quite accurate descriptions of the DNA binding specificity of transcription factors (TFs), proteins recognize DNA as a three-dimensional object. Therefore, DNA structural features refine the description of TF binding specificities and provide mechanis-

tic insights into protein-DNA recognition. Motif databases contain large numbers of nucleotide sequences identified in binding experiments based on their selection by a TF. To utilize DNA shape information when analyzing the DNA binding specificities of TFs, we developed a new tool for calculating DNA structural features from nucleotide sequences provided by motif databases. The resulting TFBSshape database generates heat maps and quantitative data for the DNA structural features minor groove width, propeller twist, roll, and helix twist for 729 TF datasets from 23 different species derived from the motif databases JASPAR and UniPROBE. As demonstrated for the basic helix-loop-helix and Hox TFs, the TFBSshape database can be used to uncover differential DNA binding preferences of closely related TFs. This approach can also be used to quantify the structural similarity between distinct sequence motifs. The TFBSshape database is freely available at http://rohslab.cmb.usc.edu/TFBSshape/. With the availability of DNA structural data, machine learning methods such as multiple linear regression can be used to construct models of TF-DNA binding specificity that incorporate both DNA sequence and shape information, which can introduce performance increase to sequence-only models and help gain new insights into TF-DNA recognition mechanisms.

Trevor Siggers touched upon adaptation and allostery in his talk. Technologies that allow the characterization of protein-DNA binding to thousands of DNA sequences are providing insights into mechanisms of transcription factor (TF) evolution and function. He discussed his efforts to use protein-binding microarrays (PBMs) to study the evolution of Cys2His2 (C2H2) zinc finger (ZF) proteins, and the role of allostery in NF-kB-dependent gene regulation. Focusing on a model system of C2H2 ZF proteins in S. cerevisiae, we analyzed the how ZF proteins with identical canonical DNA-recognition residues had evolved to bind both common and TF-specific binding sites. We found two distinct mechanisms by which ZFs have evolved to enable the binding to new sequences in a modular fashion. Studying the Bcl3-family of IkB proteins that are recruited to DNA by NF-kB dimers, we have used PBMs to analyze the DNA sequence-determinants of Bcl3 cofactor recruitment. We demonstrate that DNA sequence features can have a strong effect on the recruitment of Bcl3 cofactors to DNA. Further, we show that DNA features of allosteric recruitment can be studied in a high-throughput fashion using the PBMs.

Matthew Slattery talked about the Cap-n-Collar (CNC) transcription factors, which are master regulators of transcriptional responses to cellular stress. His lab has used genome-wide chromatin immunoprecipitation (ChIP-seq) and gene expression data to characterize an ancient CNC regulatory network, conserved from Drosophila and humans. A comparative approach demonstrates that this stress responsive regulatory axis, which includes multiple important feedback loops, is strongly conserved even at the level of network structure. Importantly, distinct gene batteries within this network are induced via disparate regulatory strategies, with a select subset regulated in a switch-like manner while many others respond to stress in a graded fashion. These divergent regulatory strategies, which are mediated by two classes of CNC DNA binding elements, are essential for both initiating and tempering the transcriptional response to cell stress.

# 4 Outcome of the Meeting

The main outcome of the meeting is that a dialogue has begun between researchers with different backgrounds and training who are interested in similar fundamental questions on mechanisms of gene regulation. Researchers who otherwise would not meet at specialized conferences had intense discussions and might start integrating knowledge and methodologies. A follow-up workshop would certainly multiply the outcome and impact of this very successful workshop.

# References

[1] R. Pelossof, I. Singh, J.Y. Li, M.T. Weirauch, T.R. Hughes, C.S. Leslie. Learning the recognition code for transcription factor and RNA-binding protein families from high-throughput binding assays. *Nature Biotech.* **33** (2015), 1242-49.

[2] Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, Rando OJ, Birney E, Myers RM, Noble WS, Snyder M, Weng Z., Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22**(9) (2012), 1798-812.

[3] Wang J, Zhuang J, Iyer S, Lin XY, Greven MC, Kim BH, Moore J, Pierce BG, Dong X, Virgil D, Birney E, Hung JH, Weng Z., Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res.* **41** (2013), D171-6.

[4] Starick SR, Ibn-Salem J, Jurk, M, Hernandez C, Love MI, Chung HR, Vingron M, Thomas-Chollier M, Meijsing S.H. (2015) ChIP-exo signal associated with DNA-binding motifs provide insights into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. Genome Research 25(6): 825-835.

[5] Rogers JM, Barrera LA, Reyon D, Sander JD, Kellis M, Joung JK, Bulyk ML, Context influences on TALE-DNA binding revealed by quantitative profiling. *Nat Commun.* **6** (2015), 7440.

[6] Felsenfeld G, Groudine M. Controlling the double helix. Nature, 2003, 421:448-53

[7] J. Wei, L. Czapla, M.A. Grosner, D. Swigon, and W.K. Olson, *Proc. Natl. Acad. Sci., USA* (2014), **111**(47) 16742-16747.

[8] T.R. Riley, M. Slattery, N. Abe, C. Rastogi, D. Liu, R.S. Mann, and H.J. Bussemaker, *SELEX-seq, a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes* Methods Mol. Biol. **1196** (2014), 255-78.

[9] M. Slattery, T.R. Riley, P. Liu, N. Abe, P. Gomez-Alcala, R. Rohs*, B. Honig*, H.J. Bussemaker, R.S. Mann, *Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox proteins* Cell **147**(6) (2011), 1270-82.

[10] Kwasnieski, J.C., Fiore, C., Chaudhari, H.G., and Cohen B.A., High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* **10** (2014), 1595-602.

[11] Mogno, I., Kwasnieski, J.C., and Cohen, B.A., Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants, *Genome Res.* **23** (2013), 1908-1915.

[12] White, M.A., Myers, C.A., Corbo, J.C., and Cohen, B.A., Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc. Natl. Acad. Sci. USA*. **110** (2013), 11952-7.

[13] Kwasnieski, J.C., Mogno, I., Myers, C.A., Corbo, J.C., and Cohen, B.A., Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc. Natl. Acad. Sci. USA.* **109** (2012), 19498-19503.

[14] Levo, M. and Segal, E. In pursuit of design principles of regulatory sequences. Nat Rev Genet 15, 453-68 (2014).

[15] Slattery, M. et al. Absence of a simple code: how transcription factors read the genome. Trends Biochem Sci 39, 381-399 (2014).

[16] Raveh-Sadka, T. et al. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. Nat Genet 44, 743-50 (2012).

[17] Levo, M. et al. Unraveling determinants of transcription factor binding outside the core binding site. Genome Res 25, 1018-29 (2015).

[18] E. P. Bishop, R. Rohs, S. C. J. Parker, S. M. West, P. Liu, R. S. Mann, B. Honig, and T. D. Tullius. 2011. A map of minor groove shape and electrostatic potential from hydroxyl radical cleavage patterns of DNA. ACS Chemical Biology 6, 1314-1320.

[19] C.G. Kalodimos, A.M.J.J. Bonvin, R. Kopke Salinas, R. Wechselberger, R. Boelens and R. Kaptein, Plasticity in protein-DNA recognition: lac repressor interacts with its natural operator O1 through alternative conformations of its DNA-binding domain , EMBO J. 21 (2002), 2866-2876.

[20] C.G.Kalodimos, et al., Structure and Flexibility Adaptation in Nonspecific and Specific Protein-DNA Complexes, Science 305 (2004), 386-389.

[21] K. Loth et al., Sliding and target location of DNA-binding proteins: an NMR view of the lac repressor system, J.Biomol. NMR 56 (2013) 41-49.

[22] Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes, Whyte WA et al, Cell, 2013, 153, 307-319

[23] Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers, Loven et al, Cell, 2013, 153, 320-334

[24] R. Rohs, X. Jin, S.M. West, R. Joshi, B. Honig, and R.S. Mann, Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.* **79** (2010), 233-269.

[25] Z. Deng, Q. Wang, Z. Liu, M. Zhang, A.C. Dantas Machado, T.P. Chiu, C. Feng, Q. Zhang, L. Yu, L. Qi, J. Zheng, X. Wang, X.M. Huo, X. Qi, X. Li, W. Wu, R. Rohs, Y. Li, and Z. Chen, Mechanistic insights into metal ion activation and operator recognition by the ferric uptake regulator. *Nat. Commun.* 6 (2015), 7642.

[26] Y.P. Chang, M. Xu, A.C. Dantas Machado, X.J. Yu, R. Rohs, and X.S. Chen, Mechanism of origin DNA recognition and assembly of an initiator-helicase complex by SV40 large tumor antigen. *Cell Rep.* **3** (2013), 1117-1127.

[27] A. Lazarovici, T. Zhou, A. Shafer, A.C. Dantas Machado, T. Riley, R. Sandstrom, P.J. Sabo, Y. Lu, R. Rohs, J.A. Stamatoyannopoulos, and H.J. Bussemaker, Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc. Natl. Acad. Sci. USA* **110**, (2013) 6376-6381.

[28] A.C. Dantas Machado, T. Zhou, S. Rao, P. Goel, C. Rastogi, A. Lazarovici, H.J. Bussemaker, and R. Rohs, Evolving insights on how cytosine methylation affects protein-DNA binding. *Brief. Funct. Genomics* **14**(1), (2014) 61-73.

[29] Carlson et al. PNAS 2010, Tietjen et al. Methods in Enzymol. 2011, Campbell et al. Cell Rep 2012.

[30] Moretti et al. ACS Chemical Biology 2008, Erwin, et al. Angewandte Chemie Int. Ed., 2014, Eguchi et al. Biochem J. 2014.

[31] Yanover C, Bradley P., Extensive protein and DNA backbone sampling improves structure-based specificity prediction for C2H2 zinc fingers. Nucleic Acids Res. 2011 Jun;39(11):4564-76.

[32] Mak AN, Bradley P, Cernadas RA, Bogdanove AJ, Stoddard BL, The crystal structure of TAL effector PthXo1 bound to its DNA target. Science. 2012 Feb 10;335(6069):716-9.

[33] Slattery, M., T. Riley, P. Liu, N. Abe, P. Gomez-Alcala, I. Dror, T. Zhou, R. Rohs, B. Honig, H. J. Bussemaker and R. S. Mann (2011). "Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins." Cell 147(6):1270-1282.

[34] Riley ,T.R., Slattery,M., Abe, N.,Rastogi C., Liu, D., Mann, R.S., Bussemaker, H.J. (2014). SELEX-seq: A Method for Characterizing the Complete Repertoire of Binding Site Preferences for Transcription Factor Complexes, Methods in Mol. Biol(1196) 255-278