Banff International Research Station

for Mathematical Innovation and Discovery

Rules of protein-DNA recognition: computational and experimental advances (13w5042)

June 16-21, 2013

MEALS

*Breakfast (Buffet): 7:00 – 9:30 am, Sally Borden Building, Monday – Friday
*Lunch (Buffet): 11:30 am – 1:30 pm, Sally Borden Building, Monday – Friday
*Dinner (Buffet): 5:30 – 7:30 pm, Sally Borden Building, Sunday – Thursday
Coffee Breaks: As per daily schedule, in the foyer of the TransCanada Pipeline Pavilion (TCPL)
***Please remember to scan your meal card at the host/hostess station in the dining room for each meal.**

MEETING ROOMS

All lectures will be held in the lecture theater in the TransCanada Pipelines Pavilion (TCPL). An LCD projector, a laptop, a document camera, and blackboards are available for presentations.

SCHEDULE

Sunday

16:00 Check-in begins (Front Desk – Professional Development Centre - open 24 hours)
17:30-19:30 Buffet Dinner
20:00 Informal gathering in 2nd floor lounge, Corbett Hall (if desired)
Beverages and small assortment of snacks are available on a cash honor system.

Monday

7:00-8:45 Breakfast
8:45-9:00 Introduction and Welcome by BIRS Station Manager, TCPL
9:00 Lectures

1. 9:00 – 9:30 Richard S. Mann "Deconvoluting the recognition of DNA shape from DNA sequence"
2. 9:30 – 10:00 Phil Bradley "Structure-based specificity prediction"
3. 10:00 – 10:30 Coffee Break
4. 10:30 – 11:00 Gerald B. Koudelka "A crucial role for specific minor groove hydration in DNA binding site recognition by proteins"
5. 11:00 – 11:30 Carlos J. Camacho "Specificity of Transcription Factors: Where are the Waters?"

11:30-13:00 Lunch
13:00-14:00 Guided Tour of The Banff Centre; meet in the 2nd floor lounge, Corbett Hall
14:00 Group Photo; meet in foyer of TCPL (photograph will be taken outdoors so a jacket might be required).
Lectures

1. 14:30 – 15:00 Richard Lavery "Using molecular simulation to probe DNA recognition mechanisms"
2. 15:00 – 15:30 Coffee Break
3. 15:30 – 16:00 Akinori Sarai "Quantifying the Direct and Indirect Readout Specificities in Protein-DNA Recognition"
4. 16:00 – 16:30 Remo Rohs "Genome-wide DNA shape analysis refines rules of protein-DNA

recognition”

5. 16:30 – 17:00 Yael Mandel-Gutfreund “Combining electrostatic and geometric approaches to uniquely characterize DNA binding interfaces”

17:30-19:30 Dinner
19:30 Discussion time

Tuesday

7:00-9:00 Breakfast
9:00 Lectures

1. 9:00 – 9:30 Gary D. Stormo “Measuring, modeling and predicting specificity of protein-DNA interactions”
2. 9:30 – 10:00 Mona Singh “Predicting transcription factor specificities and interactions”
3. 10:00 – 10:30 Coffee Break
4. 10:30 – 11:00 Raluca Gordan “Learning regression models of protein-DNA binding specificity from high-throughput in vitro data”
5. 11:00 – 11:30 Jussi Taipale “Genome-wide analysis of protein-DNA interactions”

11:30-13:30 Lunch
13:30 Lectures

1. 13:30 – 14:00 Todd Riley “Building accurate sequence-to-affinity models from Protein Binding Microarray Data using FeatureREDUCE-PBM”
2. 14:00 – 14:30 Alexandre Morozov “A unified thermodynamic approach to predicting DNA-binding affinities from high-throughput in vitro data”
3. 14:30-15:00 Coffee Break
4. 15:00 – 15:30 Tim Hughes “Decoding C2H2 Zinc Fingers”
5. 15:30 – 16:00 Scot Wolfe “Bacterial one-hybrid analysis of the recognition potential of Cys2His2 Zinc Fingers and Homeodomains”
6. 16:00 – 16:30 Break
7. 16:30 – 17:00 Marcus Noyes “Cross your fingers, hope to bind”
8. 17:00 – 17:30 Julie Mitchell “DBSI: DNA Binding Site Identifier”

17:30-19:30 Dinner
19:30 Discussion time

Wednesday

7:00-9:00 Breakfast
9:00 Lectures

1. 9:00 – 9:30 Robert Kaptein “Lac Repressor-DNA Interactions: Structure, Dynamics, and Allostery”
2. 9:30 – 10:00 Phoebe Rice “Serine resolvases: how the geometry of DNA half-sites dictates oligomerization”
3. 10:00 – 10:30 Coffee Break
4. 10:30 – 11:00 Martha L. Bulyk “DNA Binding Specificity Changes in the Evolution of Forkhead Transcription Factors”
5. 11:00 – 11:30 Tali E. Haran “The relationship between p53 REs structural properties and p53-dependent gene expression”

Coffee Break, TCPL – available from 10:00 am onwards but must finish by 11:00 am

11:30-13:30 Lunch
13:30-17:30 Free Afternoon: a hike, weather permitting
17:30-19:30 Dinner
19:30 Discussion time

Thursday

7:00-9:00 Breakfast
9:00 Lectures

1. 9:00 – 9:30 Wyeth Wasserman “Identification of disrupting cis-regulatory mutations in transcription factor binding sites”
2. 9:30 – 10:00 Matthew Slattery “Divergent transcriptional regulatory logic at the intersection of tissue growth and developmental patterning”
3. 10:00 – 10:30 Coffee Break
4. 10:30 – 11:00 Sarah Bondos “Generating context-specific functions with intrinsically disordered domains”
5. 11:00 – 11:30 David B. Lukatsky “Design principles of non-consensus protein-DNA binding and its effect on eukaryotic genomes”

11:30-13:30 Lunch
Lectures

1. 13:30 – 14:00 Harmen Bussemaker “New structural insights into protein-DNA recognition derived from high-resolution analysis of the intrinsic sequence preferences of DNase I”
2. 14:00 – 14:30 Norbert Reich “TIRF-PBM: a unique approach to investigate how individual transcription factors function in complex”
3. 14:30 – 15:00 Coffee Break
4. 15:00 – 15:30 Iris Dror “Covariation between homeodomains and the DNA shape of their binding sites provides new insights into protein-DNA recognition”
5. 15:30 – 16:00 Anjum Ansari “Dynamics and Mechanism of DNA-Bending Proteins in Binding Site Recognition”
6. 16:00 – 16:30 Break
7. 16:30 – 17:00 Aseem Ansari “Comprehensive Specificity and Energy Landscapes of DNA and RNA interactions with their ligands”
8. 17:00 – 17:30 Quaid Morris “Motif models for RNA-binding proteins”

17:30-19:30 Dinner
19:30 Discussion time

Friday

7:00-9:00 Breakfast
9:00 Discussion time
11:30-13:30 Lunch

Checkout by 12 noon.

** 5-day workshop participants are welcome to use BIRS facilities (BIRS Coffee Lounge, TCPL and Reading Room) until 3 pm on Friday, although participants are still required to checkout of the guest rooms by 12 noon. **

Abstracts to follow in alphabetical order by last name of speaker.

Banff International Research Station
for Mathematical Innovation and Discovery

Rules of protein-DNA recognition: computational and experimental advances (13w5042)

June 16-21, 2013

ABSTRACTS

(in alphabetical order by speaker surname)

Speaker: Anjum Ansari (University of Illinois at Chicago)
Title: **Dynamics and Mechanism of DNA-Bending Proteins in Binding Site Recognition**
Abstract: Proteins that recognize and bind to specific sites on DNA often distort the DNA at these sites. The rates at which these DNA distortions occur is considered to be important in the ability of these proteins to discriminate between specific and nonspecific sites. These rates have proven difficult to measure for most protein-DNA complexes in part because of the difficulty in separating the bimolecular association and dissociation of the protein from the unimolecular conformational rearrangements (DNA bending and kinking). A notable exception is the Integration Host Factor (IHF), a eubacterial architectural protein involved in chromosomal compaction and DNA recombination, which binds with subnanomolar affinity to specific DNA sites and bends them into sharp U-turns. The unimolecular DNA bending step in the IHF-DNA complex has been resolved using both stopped-flow and laser temperature-jump perturbation. I will present recent T-jump measurements on this complex, which reveal two distinct steps in the DNA bending dynamics. The fast phase, on time scales of ~100 s, appears to be nonspecific DNA bending and may correspond to the wrapping/unwrapping dynamics of the arms of the bound protein. I will also present new results on nucleotide-flipping and DNA unwinding dynamics in the context of a 3-nucleotide mismatched bubble that mimics a UV-induced lesion recognized by XPC/Rad4, a DNA damage recognition protein that plays a critical role in nucleotide excision repair machinery.

Speaker: Aseem Z. Ansari (University of Wisconsin – Madison)
Title: **Comprehensive Specificity and Energy Landscapes of DNA and RNA interactions with their ligands**
Abstract: Recent high throughput methods have yielded rich datasets that capture textured molecular recognition events that govern specific interactions between nucleic acids and their protein and/or their small molecule ligands. We have developed a novel data visualization and analysis approach to simultaneously evaluate the multidimensional relationships between various positions of a binding site. These Sequence Specificity and Energy Landscapes (SSL/SEL) provide a rapid and intuitive understanding of the various direct and indirect (including allosteric) forces that govern the specific interactions between natural and designed DNA and RNA binding molecules (proteins or small molecules). We will present our current insights from evaluation of multi-protein complexes that cooperatively bind RNA or DNA sequences (1, 2). We will also describe the fabrication of DNA microarrays with epigenetic marks for evaluation of the role of such modifications on the specificity

of nucleic acid-ligand interactions (3).

(1) Tietjen, et al., (2011) Methods in Enzymol. 497, 3-30.
(2) Campbell, et al. (2012) Cell Rep. 1, 570-581
(3) Warren, et al. (2012) Lab Chip, 12, 376-380

Speaker: Phil Bradley (Fred Hutchinson Cancer Research Center)
Title: **Structure-based specificity prediction**
Abstract: I will describe our work using protein-folding simulation techniques to predict DNA binding specificity. Time permitting, I will discuss the structure of a TAL effector-DNA complex solved by molecular replacement with de novo models.

Speaker: Sarah Bondos (Texas A&M Health Science Center)
Title: **Generating context-specific functions with intrinsically disordered domains**
Abstract: Animal survival relies on the ability of individual Hox transcription factors infer their locations within an animal and respond by regulating the appropriate subset of target genes. Although protein interactions and chromatin remodeling undoubtedly regulate Hox function, these processes cannot fully account for tissue-specific differences in Hox activity. Conversely, genetic studies with Hox chimeras and truncations demonstrate that sequences outside the DNA-binding homeodomain of Hox proteins are critical for tissue-specific functions. However, >30 years of research has failed to identify any molecular mechanism that can account for context-specific function of a Hox protein *in vivo*. Using the Drosophila Hox transcription factor Ultrabithorax as a model system, we find that most of the protein sequence alters DNA binding specificity or affinity by the homeodomain. These effects appear to be mediate by interactions between the structured DNA binding homeodomain, which both provide potential mechanisms to modulate this specificity in response to tissue-specific cues, and coordinate DNA binding with other protein functions.

Speaker: Martha L. Bulyk (Harvard Medical School)
Title: **DNA Binding Specificity Changes in the Evolution of Forkhead Transcription Factors**
Abstract: The evolution of transcriptional regulatory networks entails the expansion and diversification of transcription factor (TF) families. The forkhead family of TFs, defined by a highly conserved winged helix DNA-binding domain (DBD), has diverged into dozens of subfamilies in animals, fungi, and related protists. We have used a combination of maximum likelihood phylogenetic inference and independent, unbiased functional assays of DNA binding capacity to explore the evolution of DNA binding specificity within the forkhead family. We present converging evidence that similar alternative sequence preferences have arisen repeatedly and independently in the course of forkhead evolution. The vast majority of DNA binding specificity changes we observed are not explained by alterations in the known DNA-contacting amino acid residues conferring specificity for canonical forkhead binding sites. Intriguingly, we have found forkhead DBDs that retain the ability to bind very specifically to two completely distinct DNA sequence motifs. We propose an alternate specificity-determining mechanism whereby conformational rearrangements of the DBD broaden the spectrum of sequence motifs that a TF can recognize. DNA binding bispecificity suggests a new source of modularity and flexibility in gene regulation and may play an important role in the evolution of transcriptional regulatory networks.

Speaker: Harmen Bussemaker (Columbia University)

Title: **New structural insights into protein-DNA recognition derived from high-resolution analysis of the intrinsic sequence preferences of DNase I**

Abstract: DNA binding proteins find their cognate sequences within genomic DNA through recognition of specific chemical and structural features. Here we demonstrate that high-resolution DNase I cleavage profiles can provide detailed information about the shape and chemical modification status of genomic DNA. Analyzing millions of DNA backbone hydrolysis events on naked genomic DNA, we show that the intrinsic rate of cleavage by DNase I closely tracks the width of the minor groove. Integration of these DNase I cleavage data with bisulfite sequencing data for the same cell type's genome reveals that cleavage directly adjacent to CpG dinucleotides is enhanced at least eight-fold by cytosine methylation. This phenomenon we show to be attributable to methylation-induced narrowing of the minor groove. Furthermore, we demonstrate that it enables simultaneous mapping of DNase I hypersensitivity and regional DNA methylation levels using dense in vivo cleavage data. Taken together, our results suggest a general mechanism by which CpG methylation can modulate protein-DNA interaction strength via the remodeling of DNA shape. [Joint work with the labs of Remo Rohs and John Stamatoyannopoulos; see Lazarovici et al., PNAS, 2013.]

Speaker: Carlos J. Camacho (University of Pittsburgh)

Title: **Specificity of Transcription Factors: Where are the Waters?**

Abstract: Despite the increasing number of high-quality protein-DNA crystal structures and corresponding binding data, the origin of the specificity of transcription factors is still poorly understood. This is not surprising since, as for protein-protein interactions, implicit solvent models ignore crystallographic water molecules even though they provide snapshots of optimal solutions for the role of solvent in protein interactions. Here, we will present detailed examples of how molecular waters distort the binding free energy in ways that are not possible to be addressed by a protein-DNA code alone. Motivated by high-resolution crystal structures, we describe a simple quantitative approach to explicitly incorporate the role of molecular water in protein interactions. Our empirical energies are fully consistent with mobile water molecules having a strong polarization effect in direct intermolecular interactions. We relate the relatively easy tuning of protein-DNA interactions by a few waters (or kcal/mol) to the low affinity high complementarity required by transcription factors binding DNA.

Speaker: **Iris Dror**[1,2], Tianyin Zhou[1], Yael Mandel-Gutfreund[2], and Remo Rohs[1]
[1] Molecular and Computational Biology Program,
University of Southern California, 1050 Childs Way,
Los Angeles, CA 90089, USA
[2] Faculty of Biology, Technion – Israel Institute of Technology, Haifa, Israel

Title: **Covariation between homeodomains and the DNA shape of their binding sites provides new insights into protein-DNA recognition**

Abstract: It is well established that transcription factors (TFs) identify their binding sites through direct contacts with unique chemical groups of the base pairs mainly in the major groove. A second type of mechanism, which has been relatively less studied, is the readout of DNA shape, in which the protein recognizes the three-dimensional DNA structure (Rohs et al., 2010). Here we focused on the homeodomain family of TFs and analyzed the DNA shape of thousands of sequences in order to study the correlation between the amino acid sequence of homeodomains and the nucleotide sequence and shape of their DNA binding sites. We have found regions in the homeodomains that are significantly correlated with the sequence or with the shape of their preferred binding sites, demonstrating the role of the different homeodomain regions in attaining binding specificity through the different modes of recognition. Next, we predicted specific residues in homeodomains which likely play an important role in DNA

recognition through DNA shape attributes. Furthermore, we show that adding DNA shape information to the characterization of TF binding sites can improve predictions of homeodomain binding specificity. Finally, our work indicates that DNA shape information can provide new mechanistic insights into TF binding.

References

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*, vol. 79, pp. 233–269, 2010.

Speaker: Raluca Gordan (Duke University)

Title: **Learning regression models of protein-DNA binding specificity from high-throughput in vitro data**

Abstract: The DNA binding specificity of a transcription factor (TF) is typically represented using a position weight matrix (PWM) model, which implicitly assumes that individual bases in a TF binding site contribute independently to the binding affinity. This assumption that does not always hold, and thus more complex models of binding specificity have been developed. However, complex models have their own caveats: they typically have a large number of parameters, which makes them hard to learn and interpret.

I will present regression-based models of TF-DNA binding specificity trained using high resolution in vitro data from custom protein binding microarray (PBM) experiments. Our PBMs are specifically designed to cover a large number of putative DNA binding sites for the TFs of interest (i.e., yeast and human TFs from different protein families) in their native genomic context. These high-throughput, quantitative data are well suited for training complex models that take into account not only independent contributions from individual bases, but also contributions from di- and trinucleotides at various positions within or near the binding sites. To ensure that our models remain interpretable, we develop and use various feature selection methods, with the goal of identifying a small number of sequence features that accurately predict TF-DNA binding specificity. We apply our new regression models to the difficult problem of distinguishing between the binding specificities of paralogous TF with highly similar DNA binding motifs.

Speaker: Tali E. Haran (Department of Biology, Technion–Israel Institute of Technology, Israel)

Title: **The relationship between p53 REs structural properties and p53-dependent gene expression**

Abstract: The tumor suppressor p53 is one of the most central hubs in human cells, connected to a complex network in the living cell. Mutations in the p53 tumor suppressor gene are the most frequent genetic alterations in human cancer. How p53 decides between its various cellular functions and what determines the strength of binding of p53 to its response elements (REs) at various levels of protein is currently unknown. The level of p53 in human cells can change many folds in response to a variety of stress and cellular environments. p53 levels could determine relative expression and thus biological consequences, such as cell cycle arrest and apoptosis. Despite extensive studies addressing sequence-specific transactivation by p53 there remains the issue of what determines the differential transactivation of p53 in vivo in response to variations in levels of p53 protein. I will show that differential structural properties of p53 REs affect both p53 binding mechanism as well as the transactivation (TA) level. Moreover, I will show that the relationship is protein-level dependent, and discuss our current understanding of the rules of engagement of p53 to its REs.

Speaker: Tim Hughes (University of Toronto)
Title: **Decoding C2H2 Zinc Fingers**
Abstract: Deriving a C2H2-DNA "recognition code" has been a long-standing challenge in computational biology and gene regulation. Approaches to solving the code began with manually-identified rules based on protein-DNA structures and mutagenesis experiments, and over the last decade have been dominated by machine learning and molecular modelling. Given the large number of potential features, the amount of training data remains a limitation in these analyses. Using the bacterial 1-hybrid system (B1H), we have created a data matrix representing the relative preference of 5,504 individual natural C2H2 zinc fingers to 135 of the 256 possible 4-mers, each in the F3 position of Zif268/Egr1, thus providing ~740,000 distinct measurements. As a first step in using these data to refine the recognition code, we used linear regression to assign weights to each combination of protein/base residues. The linear code highlights known "specificity residues" but suggests that many other residues contribute to sequence specificity. Predicted PWMs inferred by either the linear code or by deriving C2H2-specific PWMs from the data itself compare favourably with an independent gold-standard set, and are consistent with results of de novo ChIP-seq experiments. Ongoing efforts are aimed at direct tests of the new code, asking why it sometimes fails (and whether we can "fix" the failures), and using the predicted motifs for vertebrate C2H2 arrays to predict their biological functions.

Speaker: Robert Kaptein (Utrecht University, The Netherlands)
Title: **Lac Repressor-DNA Interactions: Structure, Dynamics, and Allostery**

Abstract: The *E.coli* lac operon is the classical model for gene regulation in bacteria. An overview will be given of our work on the lac repressor-operator system. An early result was the 3D structure of lac headpiece in 1985, one of the first protein structures determined by NMR. Our studies of the structure and dynamics of complexes of a dimeric headpiece construct with lac operator DNA have provided a detailed picture of how the various lac operator sequences are recognized. Furthermore, we have determined the complex with non-specific DNA and investigated how the repressor searches for its target site by sliding along DNA and binding to the operator through a folding-coupled-to-binding transition.

Recently we have addressed the mechanism of allosteric coupling of the lac repressor. As all allosteric changes occur in the dimer we use a dimeric form of lac repressor (70 kD), which lacks the tetramerization domain. From ^{15}N chemical shifts of the inducer (IPTG) bound and operator bound complexes we could deduce the allosteric mechanism. Furthermore, the ternary complex with both inducer and DNA bound could be characterized. The results are surprising and different from what the crystal structures suggest.

Speaker: Gerald B. Koudelka (SUNY Buffalo)
Title: **A crucial role for specific minor groove hydration in DNA binding site recognition by proteins**
Abstract: The ability of a protein to recognize bases that it does not directly contact is termed indirect read-out. Indirect read-out commonly involves the protein-mediated induction of a B' DNA conformation in the non-contacted bases of the DNA binding site. A stabilizing feature of B'-DNA is a multi-layered spine of solvent, with each layer interlinked by hydrogen bonds. Our model of the 3-dimensional arrangement of these waters envisions that water molecules in the primary solvent layer are arranged to interact only with H-bond accepting groups on the floor of the minor groove. Therefore, solvent molecules in the spine are restricted in their positions, orientations and mobility.
We have been probing the role of the B' spine of hydration in modulating the stability and

sequence specific binding of a DNA binding protein, specifically the bacteriophage P22 repressor (P22R). We find that that proper positioning of a water molecule bound to minor groove functional groups on the noncontacted bases in the P22R binding site is critical to DNA binding by P22R. The presence and proper positioning of this and related solvent molecules apparently are essential for formation of B' conformation and formation of this DNA structure is required for complex formation by this protein. We conjecture similar mechanisms regulate DNA binding by other proteins.
In addition we find that disrupting this hydration spine in the B' region of the complex strongly influences repressor's ability to directly 'read'/recognize specific base sequences at positions 5-6 bases away in the binding site. Hence, the spine of hydration may run along the entire binding site and thereby influence the stability and specificity of the entire P22R-operator complex. This suggestion indicates the presence of a structural connection, mediated by solvent, between direct and indirect readout of DNA sequences by the P22 repressor.

Speaker: Richard Lavery (University of Lyon, France)
Title: **Using molecular simulation to probe DNA recognition mechanisms**
Abstract: In recent years, we have used a variety of simulation techniques including molecular dynamics, free energy calculations and sequence threading to probe DNA interactions with drugs and proteins at the molecular level. One of our main aims has been to analyze the sequence-dependent properties of DNA and to understand their role in recognition processes. I will present our results so far and discuss their limitations.

References

A free energy pathway for the interaction of the SRY protein with its binding site on DNA from atomistic simulations. B. Bouvier, R. Lavery *J. Amer. Chem. Soc.* 131 (2009) 9864-9865.

Protein-DNA binding specificity: a grid-enabled computational approach applied to single and multiple protein assemblies. K. Zakrzewska, B. Bouvier, A. Michon, C. Blanchet, R. Lavery *Phys. Chem. Chem. Phys.* 11 (2009) 10712-10721.

A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. R. Lavery, K. Zakrzewska, D. Beveridge, T.C. Bishop, D.A. Case, T. Cheatham III, S. Dixit, B. Jayaram, F. Lankas, C. Laughton, J.H. Maddocks, A. Michon, R. Osman, M. Orozco, A. Perez, T. Singh, N. Spackova, J. Sponer *Nucleic Acids Res.* 38 (2010) 299-313.

Protein-DNA recognition triggered by a DNA conformational switch. B. Bouvier, K. Zakrzewska, R. Lavery (2011) *Angew. Chem. Int. Ed.* 50, 6516-6518.

Multistep drug intercalation: molecular dynamics and free energy studies of daunomycin binding to DNA. M. Wilhelm, A. Mukherjee, B. Bouvier, K. Zakrzewska, J.T. Hynes, R. Lavery *J. Amer. Chem. Soc.* 134 (2012) 8588-8596.

Speaker: David B. Lukatsky (Ben-Gurion University of the Negev, Israel)
Title: **Design principles of non-consensus protein-DNA binding and its effect on eukaryotic genomes**
Abstract: High-throughput ChIP-seq, HT-ChIP, and ChIP-exo measurements of protein-DNA binding preferences have challenged the understanding of transcriptional regulation in eukaryotic genomes. These measurements have demonstrated quite generally that

transcription regulators bind thousands of active and inactive regions across the genome, and strikingly, in many cases few specific transcription factor binding sites can be identified in the Highly Occupied Target (HOT) regions. In my talk I will propose design principles of non-consensus protein-DNA binding. I will suggest that DNA exerts an effective protein localization potential that acts statistically on all DNA-binding proteins. This effective potential varies in each genomic location along the genome. I will also suggest that the predicted non-consensus protein-DNA binding mechanism provides a genome-wide background for specific promoter elements, such as transcription factor binding sites and TATA-like elements. I will discuss a number of examples, such as the genome-wide yeast transcription regulator and nucleosomal binding preferences, the yeast pre-initiation complex (PIC) binding preferences, and the human CTCF protein-DNA binding preferences.

Speaker: Yael Mandel-Gutfreund (Faculty of Biology Technion, Israel)
Title: **Combining electrostatic and geometric approaches to uniquely characterize DNA binding interfaces**
Abstract: DNA binding proteins interact with DNA via distinct regions on their surface that are characterized by an ensemble of chemical, physical and geometrical properties. We have previously developed different approaches to characterize DNA binding proteins by combining geometric and electrostatic features. Recently we have developed a novel approach to characterize protein structures by the distribution of their overlapping local surface patches. In this approach the protein surface is represented by a bag of overlapping surface patches, which are defined by a central surface residue and its nearest surface neighbors. We characterize each protein by a 'bag-of-surface patches'- a vector representing the distribution of different patches which appear on the protein surface. The similarity between two proteins is finally measured by the distance between their corresponding vectors of surface patches. In the talk I will present the new approaches and their applicability to uniquely identify DNA binding interfaces accurately and efficiently.

Speaker: Richard S. Mann (Columbia University)
Title: **Deconvoluting the recognition of DNA shape from DNA sequence**
Abstract: In previous work (Slattery et al; Joshi et al) we described the importance of the three dimensional structure of the DNA double helix – DNA shape – in the recognition of DNA binding sites by the Hox family of transcription factors and their cofactors. In particular, we found that local minima in the width of the DNA minor groove create electronegative pockets that are binding sites for amino acids with positively charged side chains such as arginine. Moreover, our data argued that DNA shape is a consequence of DNA sequence, and thus lead to the idea that the recognition of specific DNA sequences by DNA binding proteins is mediated by both base readout, typically in the major groove, and shape readout (Rohs, et al.). The relationship between DNA shape and DNA sequence leads to a logical loop that is difficult to tease apart: if shape is a consequence of sequence, then is the recognition of a binding site by a transcription factor mediated by the sequence of base pairs or by the resulting shape of the DNA molecule? One prediction of the shape recognition model is that if the shape-detecting amino acids are mutated the shape of the preferred binding sites would become less important. We tested this prediction by mutating the basic residues of Hox proteins known to insert into narrow regions of the DNA minor groove and carrying out in vitro SELEX-seq experiments. The results from these and other 'shape mutants' will be presented, and argue that the recognition of DNA shape is a key aspect of binding site selection by this family of DNA binding proteins.

References

Joshi R, et al. (2007) Functional specificity of a Hox protein mediated by the recognition of minor groove structure. Cell 131, pp. 530-543.

Slattery M, et al. (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell 147, pp. 1270-1282.

Rohs R, et al. (2010). Origins of specificity in protein-DNA recognition. Annu Rev Biochem 79, pp. 233–269

Speaker: Julie Mitchell (University of Wisconsin – Madison)
Title: **DBSI: DNA Binding Site Identifier**
Abstract: The DNA Binding Site Identifier (DBSI) is a new structure-based method for predicting protein interaction sites for DNA binding. We computed 480 candidate features for identifying protein residues that bind DNA, including new features that capture the electrostatic microenvironment within shells near the protein surface. Our iterative feature selection process identified features important in other models, as well as features unique to the DBSI model, such as a strong banded electrostatic feature with spatial separation that matches the canonical width of the DNA minor groove. DBSI was trained and validated on a data set of 263 protein-DNA complexes, tested on an independent set of protein-DNA complexes. In addition, we assessed the performance on DBSI on data sets of apo and holo protein structures distinct from the training data.

Speaker: Alexandre V. Morozov (Rutgers University)
Title: **A unified thermodynamic approach to predicting protein-DNA binding affinities from high-throughput *in vitro* data**
Abstract: DNA-binding proteins such as transcription factors carry out numerous functions in a cell by binding to their cognate sites on genomic DNA. The primary determinants of factor binding specificity and the minimal number of parameters required to describe them have been a subject of considerable debate in the literature. Studies of protein-DNA interactions typically involve high-throughput *in vitro* measurements of protein-DNA binding affinity, in which the binding process is not affected by chromatin structure and other complicating *in vivo* factors. Here we present a thermodynamic framework for predicting protein-DNA binding energetics from mechanically induced trapping of molecular interactions (MITOMI) and protein-binding microarray (PBM) assays. Our approach accommodates protein-DNA binding models of varying complexity, allowing us to assess the role of higher-order contributions in a systematic way. It rigorously treats steric exclusion and accounts for non-specific binding, secondary binding motifs, and platform-dependent biases.

Speaker: Quaid Morris (University of Toronto)
Title: **Motif models for RNA-binding proteins**
Abstract: Our goal is to build networks describing post-transcriptional gene regulation in metazoa. A large part of this effort is computationally defining RNA-binding protein (RBP) target sites in mRNAs. Our strategy is this: first, we use a microarray based in-vitro assay called RNAcompete to estimate the RBP binding affinity to hundreds of thousands of short RNA sequences, to date in collaboration with Dr. Timothy Hughes we have profiled the sequence binding preferences of more than 200 RBPs; then we summarize RBP sequence-binding preference using motif models (RNAcontext and Malarkey) specifically designed for RBPs; finally we scan mRNAs for potential binding sites. I will discuss our recent efforts to develop

algorithms that find succinct representations of the RNA sequence and structure preferences based on RNAcompete and in vivo binding data, as well as, some results applying those algorithms to three different types of RBPs.

Speaker: Marcus Noyes (Princeton University)
Title: **Cross your fingers, hope to bind**
Abstract: The *Cys_2His_2* zinc finger DNA-binding domain (ZF) is a powerful domain utilized by transcription factors and engineered by researchers to target desired sequences. As transcription factors, the ZF is the most common DNA-binding domain utilized by eukaryotic transcription factors, representing nearly 50% of the human factors. However, attempts to characterize the DNA-binding specificity for many of these factors have failed. The ZF domain has also been engineered to take on novel specificities, typically by selection of proteins from ZF libraries that offer the desired base preference. These engineered domains have been used to target auxiliary domains to a genomic target and edit the genome or modulate transcription. These artificial tools have proven so powerful they have been the focus of a great deal of research and as a result, much of our understanding of this domain has come from engineered rather than endogenous ZFs. Still, most of these engineered domains have their biases. Regardless of the engineering approach, most ZFs have been selected in similar contexts from libraries with limited coding schemes, offer similar biases to G-rich targets, and mostly fail when assembled as modules.
To address some limitations of prior engineering approaches, we optimized a PCR-based method for the construction of large ZF libraries that offer fully randomized codons at 5 or 6 positions. Using a bacterial one-hybrid assay we have selected ZFs from these libraries that bind a wide range of targets in multiple contexts. In collaboration with Mona Singh's lab, we have created a computational pipeline for the confirmation of library diversity and recovery of enriched ZFs from a sometimes high background. Many ZFs selected by this method have been tested with multiple neighboring fingers to provide insight into how adjacent ZF's might influence specificity and assembly. This helps us answer what types of ZFs are the most functional partners. Finally, information gleaned from this partner screen can be used to design functional ZF arrays. Testing these assembled ZFs as artificial transcription factors and zinc finger nucleases *in vivo* is ongoing.

Speaker: Norbert Reich (University of California Santa Barbara)
Title: **TIRF-PBM: a unique approach to investigate how individual transcription factors function in complex**
Abstract: Our interest is to reconcile the limitations of single protein binding studies (PBMs) with binding studies involving cells; the latter show dramatically different patterns for individual TFs from those obtained by classical PBM studies. Our hypothesis is that the binding preferences of individual TFs are altered when they assemble as a complex. TIRF-PBM provides a means to directly investigate such questions. We have developed a high-throughput protein binding microarray (PBM) assay to systematically investigate transcription regulatory protein complexes binding to DNA with varied specificity and affinity. This approach provides unprecedented insights into how the specificity of one protein is altered by auxiliary proteins. Our approach is based on the novel coupling of total internal reflectance fluorescence (TIRF) spectroscopy, swellable hydrogel double-stranded DNA microarrays, and dye-labeled regulatory proteins, making it possible to determine both equilibrium binding specificities and kinetic rates for multiple protein:DNA interactions in a single experiment. Initial applications to general transcription factors TBP, TFIIA, and TFIIB showed that kinetic and thermodynamic measurements by TIRF-PBM are similar to those determined by traditional methods, while simultaneous measurement of the factors in binary

and ternary protein complexes reveals preferred binding combinations. We then extended this to the study of STAT proteins and screening for small molecules that disrupt multi-transcription factor complexes. TIRF-PBM provides a novel and extendible platform for multi-protein transcription factor investigation.

Speaker: Sherwin Montaño[1], Sally Rowland[2], Martin Boocock[2], Marshall Stark[2], and **Phoebe Rice**[1].
[1] The University of Chicago
[2] University of Glasgow
Title: **Serine resolvases: how the geometry of DNA half-sites dictates oligomerization**
Abstract: Serine resolvases are site-specific DNA recombinases that resolve replicon dimers into monomers. Chemistry and strand exchange are carried out within a tetramer that synapses two DNA segments. However, the active tetramer does not form until those four subunits are incorporated within a larger complex that traps three supercoiling nodes. This "synaptosome" acts as a topological filter and ensures that only intramolecular resolution reactions occur.
Formation of the synaptosome requires additional recombinase dimers and/or DNA bending proteins. We study two related examples: Sin and Tn3 resolvases. Their catalytic domains mediate dimerization (and tetramerization when activated) and are connected by a flexible linker to a HTH DNA binding domain. In both cases, the crossover site is a simple inverted repeat but the accessory sites vary. For Sin, each partner duplex includes binding sites for HU (or IHF) and for a second dimer of Sin. The half-sites for the 2nd dimer are in direct repeat, and their geometry triggers formation of a DNA binding-domain-mediated tetramer that mediates synapsis. For Tn3, the DNA topology of the synaptosome is the same, but the protein-protein contacts that mediate its formation are different. Two additional Tn3 recombinase dimers are bound per recombination partner. These binding sites are inverted repeats, but the central spacers are 3bp shorter and 6bp longer than that of the crossover site. Complexes formed on these sites place the catalytic domain dimers in very different orientations relative to the DNA. These differences are not random: they arrange the dimers to form additional protein-protein contacts that mediate synaptosome formation.

Speaker: Todd Riley (Columbia Univeristy)
Title: **Building accurate sequence-to-affinity models from Protein Binding Microarray Data using FeatureREDUCE-PBM**
Abstract: We have recently developed FeatureREDUCE-PBM (Riley et al., submitted), an algorithm that can be used to infer sequence-to-affinity models from protein binding microarray (PBM) data. PBM technology has been used to probe the DNA binding specificity of hundreds of transcription factors from a variety of organisms, but methods for analyzing such data are still in development. FeatureREDUCE-PBM parameterizes the relative affinity of all possible DNA sequences in terms of a small set of free energy parameters associated with base pair substitutions or insertions/deletions and their possible dependencies. The algorithm accounts for PBM-specific biases and by using robust regression techniques achieves quantification of relative binding affinities at an unprecedented level of accuracy. FeatureREDUCE-PBM can also model multiple binding modes, whereby the method can correctly build a separate, independent affinity model for each contributing binding mode and determine the relative affinity of each of these binding modes from the confounded PBM data. In a recent study comparing 26 PBM analysis algorithms, FeatureREDUCE-PBM emerged as the top-performing algorithm (Weirauch et al., Nature Biotechnol., 2013).

Speaker: Remo Rohs (University of Southern California)

Title: **Genome-wide DNA shape analysis refines rules of protein-DNA recognition**

Abstract: Experimentally solved structures of protein-DNA complexes have revealed key mechanisms of DNA readout. Whereas protein-DNA readout is well understood on a transcription factor-family basis, it is still not really known how closely related transcription factors from the same family select their distinct genomic targets. To refine our knowledge on mechanisms employed by transcription factors in achieving DNA binding specificity, we must take advantage of insights gained from both genomics and structural biology. To be able to derive DNA structural features from the wealth of DNA sequence information generated by high-throughput experiments is important for capturing the sequence-structure degeneracy of DNA. Such degeneracy enables dissimilar sequences to adopt similar DNA shapes, or in turn, single nucleotide polymorphisms to disrupt DNA shape in an extended region. To fully embrace the genomic era, we developed a method for high-throughput DNA shape analysis on a genomic scale. Using this approach we can predict DNA structural features at single-nucleotide resolution for any number or length of DNA sequences in an instant manner. Based on an integration of the predicted DNA shape features with sequence, we have developed statistical machine learning approaches to model the DNA binding specificity of transcription factors and other DNA binding proteins, and we achieved a significant improvement in prediction accuracy compared to using sequence alone. Moreover, we found that nuances in DNA shape features can actually explain different binding specificities of closely related transcription factors. Studying DNA shape features of transcription factor binding sites in *cis* regulatory modules of *Drosophila melanogaster* genomes, we also discovered that single nucleotide variants in non-coding regions that change DNA shape more drastically are more likely to be eliminated through purifying selection. We identified regulatory regions that appear non-conserved in terms of nucleotide sequence but are actually conserved in terms of DNA shape. Taken together, these findings provide new mechanistic insights into rules of protein-DNA recognition and regulatory functions of the genome.

Speaker: Akinori Sarai (Kyushu Institute of Technology, Fukuoka, Japan)

Title: **Quantifying the Direct and Indirect Readout Specificities in Protein-DNA Recognition**

Abstract: Sequence-specific recognition of DNA by proteins plays a critical role in regulating gene expression. The recognition of target sequences in the genome can be achieved by a combination of two different mechanisms: the direct readout through the direct interactions between protein and DNA bases; and the indirect readout through the sequence-dependent conformation and/or deformability of DNA structure. In order to understand the recognition mechanism, we need to quantify their specificities. We have used the knowledge-based approach based on the statistical analysis of protein-DNA complex structures to calculate the potential of mean force (statistical potential) and the specificities of the direct and indirect readouts (1, 2). By using this method, it has been shown that both the direct and indirect readouts make important contributions to the specificity of protein-DNA recognition (2, 3). The quantification of specificity has also enabled us to analyze the structure-specificity relationship in protein-DNA recognition. For example, we could show that the cooperative protein-DNA recognition increases the specificity (1).

The major problem of the knowledge-base approach is the limited amount of available structural data. In the case of the indirect readout, we could only consider the dimer sequences and had to use a harmonic approximation to calculate the potential of mean force. In order to complement the knowledge-based approach, we have performed various kinds of computer simulations to derive the energy potentials, which are equivalent to the statistical potentials, for the direct and indirect readouts. In the case of the direct readout, we calculated the potential of

mean force for base-amino acid interactions by considering simply a fragment of DNA that included a base pair and an amino acid side chain (4). In the case of the indirect readout, we used conformational ensembles obtained by molecular dynamics (MD) simulations of DNAs containing all 136 unique tetramer sequences, to calculate the probability distribution functions of the base-pair step parameters (5). Then, we derived the conformational energy of the central base-pair step within each tetramer sequence directly from the probability distribution functions with and without the harmonic approximation (5, 6). We have also developed a method, which uses Bayesian statistics to derive the probability of a particular sequence for a given DNA structure directly from the trajectories of MD simulations (7). Then, we used the information entropy to quantify the specificity of the indirect readout.

These analyses provided insight into the molecular mechanism of protein-DNA recognition. By combining these methods, we have made some applications to drug-DNA interactions (8), nucleosome positioning (9), as well as the genome-scale target prediction of transcription factors (1).

1. H. Kono and A. Sarai "Structure-based prediction of DNA target sites by regulatory proteins" Proteins 35, 114 (1999).
2. M.M. Gromiha, J.G. Siebers, S. Selvaraj, H. Kono and A. Sarai "Intermolecular and Intramolecular Readout Mechanisms in Protein-DNA Recognition" J. Mol. Biol. 337, 285 (2004).
3. A. Sarai and H. Kono "Protein-DNA Recognition Patterns and Predictions" Ann. Rev. Biophys. Biomol. Struct. 34, 379 (2005).
4. F. Pichierri, M. Aida, M. Gromiha and A. Sarai "Free energy maps of base-amino acid interaction for protein-DNA recognition" J. Am. Chem. Soc. 121, 6152 (1999).
5. M.J. Arauzo-Bravo, S. Fujii, H. Kono, S. Ahmad and A. Sarai "Sequence-Dependent Conformational Energy of DNA Derived from Molecular Dynamics Simulations: Toward Understanding the Indirect Readout Mechanism in Protein-DNA Recognition" J. Am. Chem. Soc. 127, 16074 (2005).
6. S. Yamasaki, T. Terada, K. Shimizu, H. Kono, and A. Sarai "A Generalized Conformational Energy Function of DNA Derived from Molecular Dynamics Simulations" Nucleic Acids Res. 37, e135 (2009).
7. S. Yamasaki, T. Terada, H. Kono, K. Shimizu and A. Sarai "A New Method for Evaluating the Specificity of Indirect Readout in Protein-DNA Recognition" Nucleic Acids Res. 40, e129 (2012).
8. M. J. Araúzo-Bravo and A. Sarai "Indirect readout in Drug-DNA Recognition: Role of DNA Sequence-dependent DNA Conformation" Nucleic Acids Res. 36, 376 (2008).
9. M. Fernandez, S. Fujii, H. Kono and A. Sarai "Evaluation of DNA intramolecular interactions for nucleosome positioning in yeast" Genome Inform. 23, 13 (2009).

Speaker: Mona Singh (Princeton University)
Title: **Predicting transcription factor specificities and interactions**

Abstract: I will discuss two projects: (1) a cross-genomic pipeline we have been developing for uncovering DNA binding sites for transcription factors with known specificities, along with transcription factor-transcription interactions and (2) our attempts at predicting de novo Cys2His2 zinc finger binding specificities.

Speaker: **Matthew Slattery**[1,2], Roumen Voutev[1], Lijia Ma[2], Nicolas Nègre[2,3], Kevin P. White[2], Richard S. Mann[1]

1. Department of Biochemistry and Molecular Biophysics, Columbia University, 701 West 168th Street, HHSC 1104, New York, NY 10032

2. Institute for Genomics and Systems Biology and Department of Human Genetics, University of Chicago, 900 East 57th Street, KCBD 10115, Chicago, IL 60637

3. Université de Montpellier 2 and INRA, UMR1333 DGIMI, F-34095 Montpellier, France

Title: **Divergent transcriptional regulatory logic at the intersection of tissue growth and developmental patterning**

Abstract: The Hippo tumor suppressor pathway controls proliferation in a tissue-nonspecific fashion in *Drosophila* epithelial progenitor tissues via the transcriptional coactivator Yorkie (Yki). However, despite the tissue-nonspecific role that Yki plays in tissue growth, the transcription factors that recruit Yki to the DNA, most notably Scalloped (Sd) and Homothorax (Hth), are important regulators of developmental patterning with many tissue-specific functions. Thus, these three transcriptional regulators – Yki, Sd, and Hth – provide a model for exploring the properties of protein-DNA interactions that regulate both tissue-shared and tissue-specific functions. With this goal in mind, we identified the positions in the fly genome that are bound by Yki, Sd, and Hth in the progenitors of the wing and eye-antenna structures of the fly. These data not only provide a global view of the Yki gene regulatory network, they reveal an unusual amount of tissue specificity in the genomic regions targeted by Sd and Hth, but not Yki. The data also reveal that tissue-specific binding is very likely to overlap regulatory DNA regions, provide important clues for how tissue-specific Sd and Hth binding occurs, and support the idea that gene regulatory networks are plastic, with spatial differences in DNA binding significantly impacting network structures.

Speaker: Gary D. Stormo (Washington University, St. Louis)

Title: **Measuring, modeling and predicting specificity of protein-DNA interactions**

Abstract: This talk will cover our approaches for modeling the specificity of transcription factors (TFs) with an emphasis on the use of high-throughput data. This will include some new experimental approaches that we think will improve the accuracy of determining differences in binding affinity while maintaining high-throughput data generation. Finally it will include our approach to predict the specificity of TFs based only on their protein sequence with results for zinc finger and homeodomain TF families.

Speaker: Arttu Jolma[1,2], Jian Yan[1], Thomas Whitington[1], Martin Enge[1], Kazuhiro Nitta[1,3], Teemu Kivioja[1,2], Ekaterina Morgunova[1], Mikko Taipale[1], Juan M. Vaquerizas[4], Nicholas M. Luscombe[4], Minna Taipale[1,2], Esko Ukkonen[2], Patrick Lemaire[3,5] and **Jussi Taipale[1,2]**

[1]*Karolinska Institutet, Department of Biosciences and Nutrition*

[2]*University of Helsinki, Finland*

[3]*IBDML, Marseilles, France*

[4]*EMBL - European Bioinformatics Institute, Cambridge, UK*

[5]*CRBM, MONTPELLIER, France*

Title: **Genome-wide analysis of protein-DNA interactions**

Abstract: Understanding the information encoded in the human genome requires two genetic codes, the first code specifies how mRNA sequence is converted to protein sequence, and the second code determines where and when the mRNAs are expressed. Although the proteins that read the second, regulatory code – transcription factors (TFs) – have been largely identified, the code is poorly understood as it is not known which sequences TFs can bind in the genome. To understand the regulatory code, we have analyzed the occupancy of the majority of all expressed

TFs in human colorectal cancer cells, and analyzed the sequence-specific binding of human, mouse and *Drosophila* TFs using high-throughput SELEX. Our results reveal additional specificity determinants for a large number of factors for which a partial specificity was known, including a commonly observed A- or T-rich stretch that flanks the core motifs. Global analysis of the data revealed that homodimer orientation and spacing preferences, and base-stacking interactions, have a larger role in TF-DNA binding than previously appreciated. Comparison of the human binding profiles with those of house mouse and *Drosophila*, revealed that the monomer binding specificity of TFs evolves very slowly, and has been almost completely fixed between vertebrates and invertebrates despite complete lack of regulatory element conservation. However, TF flanking sequences, and dimer spacing and orientation preferences appear to evolve much faster than monomer binding preferences, indicating that such changes are a potential source of evolutionary novelty. A binding model that is required to understand binding of TFs to the genome, which incorporates information about protein-protein interactions induced by DNA, and inheritance of epigenetic states across cell division will be discussed.

Speaker: Wyeth W. Wasserman (University of British Columbia)
Title: **Identification of disrupting cis-regulatory mutations in transcription factor binding sites**
Abstract: Whole genome sequencing has arrived in applied human clinical research. While much progress has been made in the study of protein altering mutations, the identification of mutations disrupting cis-regulatory elements remains an unresolved challenge. The presentation will focus on three key aspects of the challenge: data resources, binding site discrimination, and scoring the strength of disruption. Applied analysis of cancer genome data will be presented.

Speaker: Scot Wolfe (UMass Medical School)
Title: **Bacterial one-hybrid analysis of the recognition potential of Cys_2His_2 Zinc Fingers and Homeodomains**
Abstract: Cys_2His_2 zinc fingers and homeodomains represent the two largest classes of DNA-binding domains found in the majority of metazoan genomes. This prevalence suggests that these scaffolds provide a robust framework with diverse sequence recognition potential for the evolution of orthogonal/complex regulatory systems. Despite numerous biochemical and structural studies on members of both of these families, important questions remain about the breadth of their sequence recognition potential from both a functional and engineering perspective. We have utilized a bacterial one-hybrid system to probe specific aspects of recognition for both of these families through the selection of preferred target sites for hundreds of natural-occurring and artificially-generated family members, or through the selection of specificity determinants within these frameworks to recognize novel target sequences. These data provide a more nuanced perspective on recognition by these two families that will facilitate the construction of improved recognition models for estimating the recognition preference of family members found in eukaryotic species. These data will also facilitate the construction of artificial DNA-binding domains with improved recognition properties for focused biological or genomic questions. Toward this latter goal we have begun engineering zinc finger-homeodomain chimeras to recognize composite binding sites for targeted genome engineering.