NATIONAL CENTER FOR THE MULTISCALE ANALYSIS OF GENOMIC AND CELLULAR NETWORKS

WINTER 2009

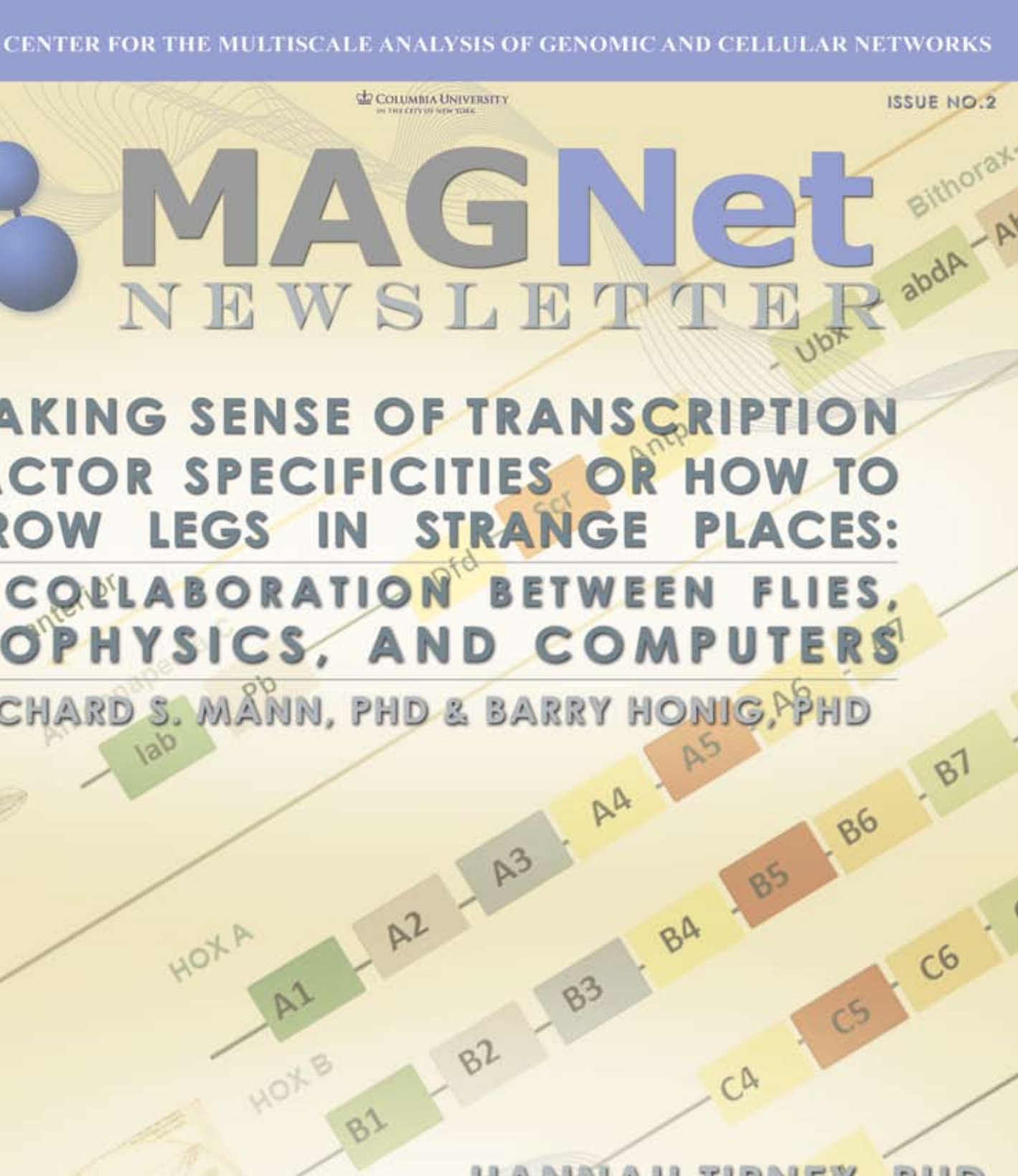
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

ISSUE NO.2

MAGNet NEWSLETTER

MAKING SENSE OF TRANSCRIPTION FACTOR SPECIFICITIES OR HOW TO GROW LEGS IN STRANGE PLACES: A COLLABORATION BETWEEN FLIES, BIOPHYSICS, AND COMPUTERS

RICHARD S. MANN, PHD & BARRY HONIG, PHD



HANNAH TIPNEY, PHD
LAWRENCE HUNTER, PHD

THREE R'S OF COMPUTER ASSISTED BIOMEDICAL DISCOVERY: READING, REASONING AND REPORTING.

HARMEN J. BUSSEMAKER, PHD

DISSECTING TRANSCRIPTION FACTOR FUNCTION ON MULTIPLE SCALES

WINTER 2009
ISSUE NO.2

MAGNet Newsletter
http://magnet.c2b2.columbia.edu

MAGNet NEWSLETTER

FEATURES

SECTIONS

INTRODUCTION

Welcome to the second issue of the MAGNet Center Newsletter. The last year has been a period of intense effort not just at Columbia but, jointly, across the entire community of NCBCs as we prepare to compete for the renewal of our Centers in 2010. As each center is gearing up to demonstrate its scientific accomplishments, software tools, and impact on the research community, we are individually reflecting on our core mission and on its implications for the biomedical sciences. Roughly speaking, MAGNet's mission has been the creation of integrative tools for the assembly and analysis of molecular interaction networks, within specific cellular contexts. Integration is a much-hyped term to describe that the set is better than the sum of its parts. In biology, this concept has been much utilized in fusing multiple clues supporting specific hypotheses: for instance, the hypothesis that protein A regulates the expression of protein B. Yet that is but one of the ways in which knowledge can be integrated. For instance, as we have discovered, integration of multiple computational inferences, discrete layers of representation, and even diverse methodological approaches can be equally valuable if done right. Using molecular interactions as the basis to integrate and analyze biological data is a leitmotif that infuses this issue's articles by several MAGNet investigators, including Drs. Bussemaker, Honig and Mann from Columbia University, and guest writers Drs. Hannah Tipney and Larry Hunter from the University of Colorado at Denver.

Drs. Honig and Mann tackle the issue of combinatorial regulation by multiple transcription factors during early *Drosophila* embryo development. It is clear that the complexity of multicellular organisms could not possibly arise from transcriptional programs driven by individual transcription factor proteins. We now know that transcription factors interact with regulatory regions of the chromatin in the context of transcriptional regulation complexes. These help both stabilize the binding by increasing affinity and also provide context-specific regulation of genetic programs, driven by the presence or absence of specific co-factors. By moving from single to multi-transcription factor interactions with the DNA molecule, for instance from individual Hox proteins to PBC-Hox complexes, researchers thought they could solve the transcription factor binding-specificity problem. However, they soon realized that combining two proteins into a complex did not necessarily address the specificity of the individual interactions and that some other process contributing to specificity would have to be revealed by complex formation. Indeed, this simple observation may have lead to the discovery of the role of co-factors in changing transcription factor's conformation to expose "hidden" features that contribute to the specificity of DNA binding, specifically in relationship to the shape of DNA's minor groove. This progress, which is creating an entire new field of "DNA shape" analysis, would not have been possible without the integration of structural, functional, and sequence information to understand Hox factor binding-specificity using techniques developed within the MAGNet center.

In a corresponding article by Dr. Harmen Bussemaker, the same issue of complex-derived specificity in *Drosophila* regulation is explored from a completely different perspective. Indeed, Bussemaker and colleagues at the Netherlands Cancer Institute and University of Chicago observed that contrary to the *in vitro* model of sequence-based DNA binding specificity of individual transcription factors, large scale binding assays showed that a large fraction of the transcriptionally active proteins are binding to hotspots (2kb-3kb DNA regions that together account for about 5% of the total chromatin). Surprisingly, these do not contain the classical DNA-binding motifs for these proteins. Indeed, even in the presence of single point mutations in the transcription factors' DNA binding domains, researchers found hotspot binding virtually unaffected, showing that the process is not DNA-binding-domain mediated but rather effected by additional molecular interactions with nuclesome proteins. This again suggests that transcriptional processes should be studied in the context of multi-protein complexes rather than one transcription factor at the time. Additionally, it suggests that in order to understand these multi-protein binding processes, one may have to abandon a purely functional or sequence-based view of the protein-DNA interactions and start integrating information from 3D structural models.

Finally, Drs. Tipney and Hunter reflect on the fact that an interaction-centric view of biology is not only useful in the context of studying protein-protein or protein-DNA interactions but may be extended to encompass virtually any aspect of gene and cellular function. Starting from a model that explicitly represents interactions between ontological terms as a graph, they show that it is possible to integrate biological knowledge, producing a signature that can reveal genes critically involved in the presentation of specific phenotypes. They call this a "3R system," based on the fact that the model is built by "Reading" the literature using NLP approaches, and that it must then "Reason" about specific facts mapped to this graphical models, and then "Report" findings in a succinct, hypothesis-centric fashion, i.e., expressed as simple, testable predicates. This approach allowed the identification of four novel gene candidates for tongue formation, which were experimentally validated through whole mount *in situ* hybridizations to E11.5 and E12.5 mouse embryos. Surprisingly, one of them is the Hoxa2, an important transcription factor in early development, whose expression had been previously studied in unrelated contexts.

- Andrea Califano, Ph.D.

THREE R'S OF COMPUTER ASSISTED BIOMEDICAL DISCOVERY: READING, REASONING AND REPORTING.

HANNAH TIPNEY, PHD[1] & LAWRENCE HUNTER, PHD[1]

[1]School of Medicine, University of Colorado, Denver

Microarray experiments, genome-wide association studies, and a plethora of new methods exploiting low-cost sequencing technology now routinely produce data at genomic scale. Researchers have long known that most biological phenomena, especially those relevant to human health, involve complex interactions among dozens, hundreds or even thousands of gene products. The technologies that make possible simultaneous observations regarding the presence or activity of all of the gene products in a biological sample have already yielded a bonanza of biomedical insights.

In large part due to this genome-scale technology, human knowledge relevant to biomedical research is exploding. The PubMed bibliographic database contains more than 17 million publications, adding nearly 800,000 in 2008 alone. In addition to this traditional scientific literature, the latest Nucleic Acids Research database issue (2009) lists 1170 more structured collections of molecular biology information, including critical resources such as GenBank for macromolecular sequences and the PDB for structures, dozens of model organism databases with curated function information, and growing collections of microarray data, genotype repositories, and more.

However, the combination of genome-scale assays and huge increases in human knowledge of molecular biology poses its own challenges. The groups of genes identified in a particular experiment—and the many interactions among them—need to be understood in the context of all that is already known about them. For a typical experiment, that can mean hundreds of genes, tens of thousands of interactions and at least as many publications and database entries that have to be digested to fully exploit one's own results. Exploring genome-scale results in light of everything else that has ever been published is a huge challenge. Genome-scale data rarely respects disciplinary boundaries, so papers and results from many fields, likely some unfamiliar to the experimentalist, have to be appreciated. No wonder this task can seem overwhelming for bench scientists! Unfortunately, failure to take full advantage of this wealth of prior knowledge can cause important results to be overlooked or misinterpreted, wasting time, effort and money.

For many years, efforts have been made to centralize all of the information relevant to the interpretation of genome-scale data into an integrated, easy to use form. The National Library of Medicine's NCBI, the European Bioinformatics Institute, and various model organism databases all have made extensive and valuable efforts in this regard. Yet these efforts have not been entirely successful, for several reasons. First, much of the necessary information is expressed in unstructured form, written in the natural language of journal articles and the like. Valiant (and valuable) efforts by biocurators to manually process the entire literature and represent its content formally appear unable to keep up with the rate of publication and the many potentially important facts expressed in each article. Second, much human thought about biomolecular function is not explicitly stated in any database or publication, but is instead the result of inferences regarding possible functions of a molecule, made by considering factors such as homology, location, interaction partners, expression patterns, knockout phenotypes and so on. A third problem involves the best way to present this enormous amount of information to a bench scientist trying to interpret a large dataset. A stack of hundreds of gene summaries is not much easier to digest than hundreds of journal publications, nor necessarily an easy path to understanding one's data in context.

Our laboratory has been developing computational approaches that address each of those problems. We call them "3R systems," since they have to read the literature, reason about implicit

information, and report the aspects that are relevant to the interpretation of a dataset. We recently published an article in PLoS Computational Biology describing the Hanalyzer, a 3R system that helped interpret a complex craniofacial development expression array dataset, leading to the discovery of four novel genes involved in the development of the mammalian tongue.

Our systems are built on a foundation of community-curated ontologies such as the Gene Ontology, and on public database identifiers for genes and gene products. We construct a knowledge network in which the nodes are ontology terms or gene identifiers, the nodes are linked together by edges that represent various types of relationships, and where each edge is quantified with a reliability score. The knowledge network is initially built by reading (using text mining programs and semantic database integration techniques), and then expanded by various kinds of reasoning. We create a data network that describes the results of a particular genome-scale experiment. In this network, the nodes are identifiers specifying significant genes or gene products, and links are drawn between genes that interacted in the experiment, each quantified with the degree of interaction (e.g. by correlation coefficient over a time course). Finally, we then construct visualizations that make it easy for a scientist to tell where the networks align (meaning that there was existing knowledge about a set of genes and relationships observed in the data) or don't (possible new discoveries), and to explore all of the knowledge relevant to those relationships in a uniform system, based on the popular Cytoscape platform.

The initial population of a knowledge network begins with the extraction of gene product interaction information from existing databases of protein-protein interactions and protein-DNA interactions (transcription factors). We then add the results of our highly effective concept recognition system, OpenDMAP, as well as other text mining approaches (such as gene name co-occurrence over all PubMed abstracts) to cast a net over as much of the literature as we can.

We augment this reading-based network by reasoning about relationships. Many relationships between gene products can be inferred on the basis of shared characteristics. Inferences can be made on the basis of participation in a particular metabolic or signaling pathway, biological process, shared molecular function or functional domain, co-localization to a particular subcellular compartment, related phenotype on knockout, and various other shared characteristics. More complex inferences, involving reasoning over ontology term cross-products are also possible, for example linking a calcium transport gene to a calcium signaling gene. While each of these inferences is potentially erroneous, all reliabilities are quantified, and multiple independent lines of reasoning often strengthen the belief in a linkage.

The data network from an experiment is combined with the knowledge network, visualized with a set of Cytoscape plugins that filter and color-code the relationships based on the strength of the data and knowledge underlying each. By clicking on a relationship, a user can see all of the sources of knowledge that support it, drilling down to each for more detail as needed. In the craniofacial example described in the PLoS Computational Biology paper, the Hanalyzer was first used to try to explain a group of genes with a particular tissue- and time-specific expression profile. By exploring the relationships that were well supported in the knowledge network, the analyst decided that these genes were likely involved in tongue development. She then added in the relationships that were strong in the data, but not reflected in the knowledge network, identifying four genes that had no published association with craniofacial muscle development, but she hypothesized were also involved in tongue development. Remarkably, all four hypotheses were biologically validated through whole mount in situ hybridizations to E11.5 and E12.5 mouse embryos.

We construct a knowledge network in which the nodes are ontology terms or gene identifiers

The Hanalyzer demonstrates the potential of 3R systems, but is just the first step in the development of more powerful systems with improvements in each "R". Text mining is a rapidly evolving field, and the growth of repositories like PubMedCentral is opening many opportunities for full text natural language processing. Many existing methods in computational reasoning and visualization can likely be applied productively in the future. We hope to enhance the power and utility of 3R systems to the point where they are routinely used by bench scientists to interpret genome-scale experiments and to help in the generation of novel and significant hypotheses.

The Hanalyzer is available for download at http://hanalyzer.sourceforge.net/

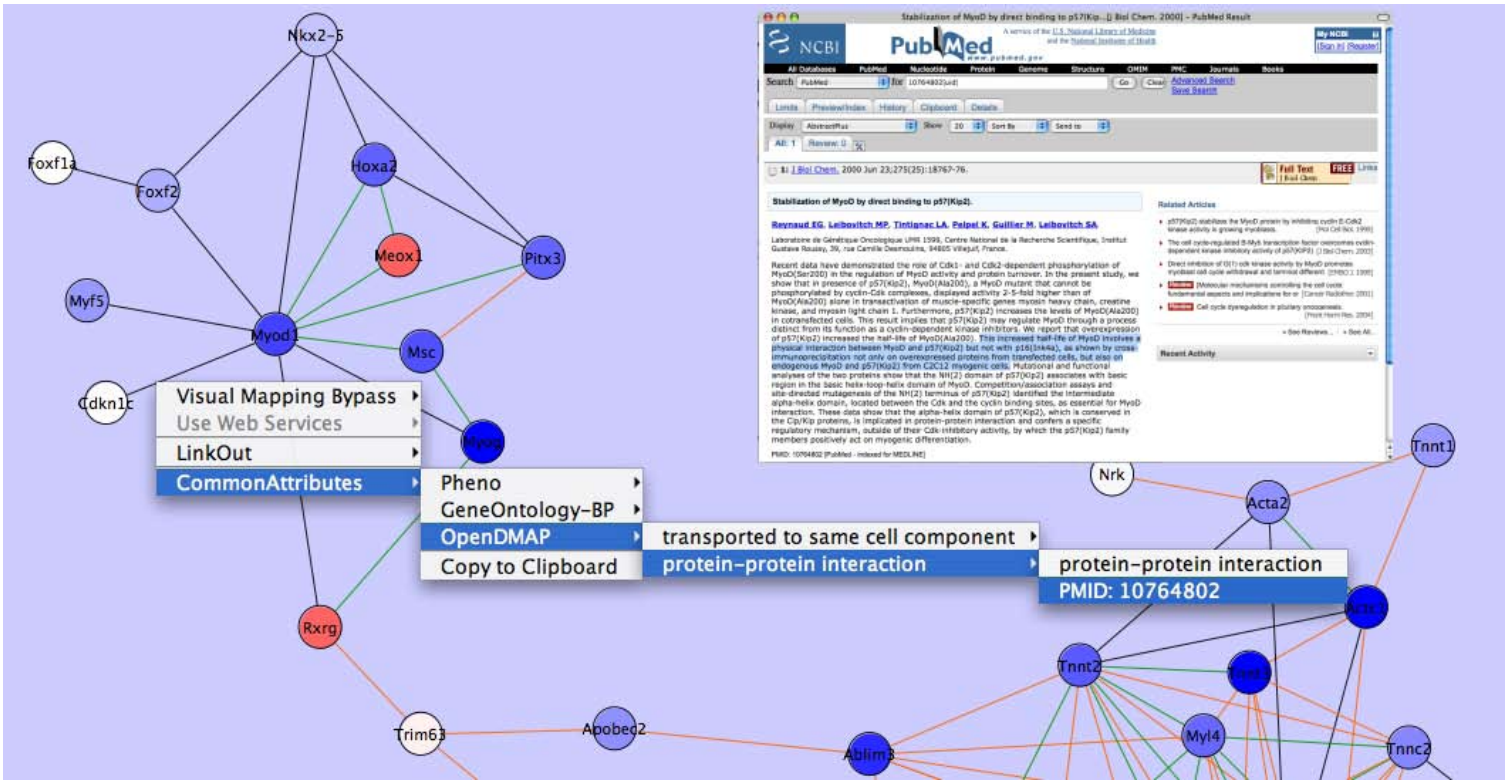


Figure 1: Text mining in Hanalyzer

A SAMPLE OF RELEVANT PAPERS FROM OUR LAB

Bada M, Hunter L (2007) "Enrichment of OBO ontologies." J Biomed Inform 40: 300-315.

Baumgartner WA Jr, Lu Z, Johnson HL, Caporaso JG, Paquette J, Lindemann A, White EK, Medvedeva O, Cohen KB, Hunter L. Concept recognition for extracting protein interaction relations from biomedical text. Genome Biol. 2008;9 Suppl 2:S9.

Baumgartner WA, Jr., Cohen KB, Fox LM, Acquaah-Mensah G, Hunter L (2007) "Manual curation is not sufficient for annotation of genomic databases." Bioinformatics 23: i41-48.

Gabow A, Leach S, Baumgartner WA, Jr., Hunter L, Goldberg D (2008) "Improving protein function prediction methods with integrated literature data." BMC Bioinformatics 9.

Hunter L, Cohen K (2006) "Biomedical language processing: what's beyond PubMed?" Molecular Cell 21: 589-594.

Hunter L, Lu Z, Firby J, Baumgartner WA, Jr., Johnson HL, et al. (2008) "OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and celltype-specific gene expression." BMC Bioinformatics 9: 78.

Leach S, Gabow A, Hunter L, Goldberg DS (2007) "Assessing and combining reliability of protein interaction sources." Pac Symp Biocomput: 433-444.

Leach S, Tipney H, Feng W, Baumgartner, WA, Kasliwal, P, Schuyler, R, Williams T, Spritz R, Hunter L (2009) "3R Systems for biomedical discovery acceleration, with applications to craniofacial development." PLoS Computational Biology. In press.

Tipney H, Leach S, Feng W, Spritz R, Williams T, et al. (2009) Leveraging existing biological knowledge in the identification of candidate genes for facial dysmorphology. BMC Bioinformatics. 10(Suppl 2):S12

DISSECTING TRANSCRIPTION FACTOR FUNCTION ON MULTIPLE SCALES

HARMEN J. BUSSEMAKER, PHD[1,2]

[1]Department of Biological Sciences, Columbia University

[2]Center for Computational Biology and Bioinformatics, Columbia University

Introduction

Transcription factors (TFs) play a central role in the regulation of genome expression. By interacting with DNA in a highly sequence-specific manner, these proteins coordinate interaction with the polymerase complexes that transcribe DNA to messenger RNA. Whenever cells respond to internal or external signals, relayed by signal transduction pathways, it is the transcription factors that are charged with the ultimate task of increasing or decreasing the transcription rate of specific genes. The number of different TFs ranges from hundreds in simple organisms such as yeast to thousands in mammalian cells. However, in spite of many years of detailed research by many groups, a general mechanistic and quantitative framework for understanding how they function is still largely lacking. MAGNet investigator Dr. Harmen Bussemaker and his collaborators are taking a transcription-factor-centric computational approach to deciphering gene regulatory networks. Their research has yielded some surprising results.

Transcription binding *in vitro*: from sequence to affinity

Before there can be any hope of understanding how the transcriptional machinery interacts with the genome in a living cell, we need good quantitative models of how individual TF proteins interact with "naked" DNA in a test tube. Until recently, researchers mostly thought in terms of discrete cis-regulatory elements in DNA, and related bioinformatics tools were based on the binary classification between sequence "bound" and "unbound" by the TF. It has become increasingly clear, however, that quantification of TF-DNA binding affinity is essential for understanding TF function. From the point of view of a TF, the double-stranded DNA molecule at the core of each chromosome looks like an affinity landscape, where each position on the chromosome has its own dissociation constant Kd (equal to the concentration at which the site is 50% occupied), which in turn depends on the local DNA sequence. High-throughput chromatin-

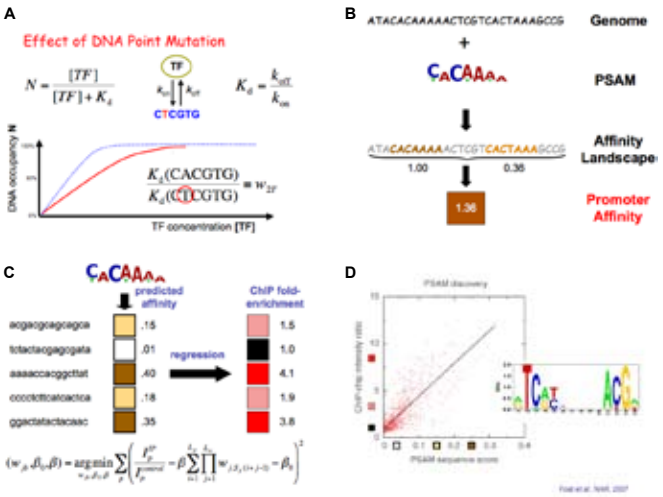


Figure 1: Inferring binding free energy parameters directly from high-throughput protein-DNA interaction data. **(A)** The biophysical model underlying the MatrixREDUCE software. **(B)** Position-specific affinity matrices (PSAM) allow one to see DNA sequence as an affinity landscape from the point of view of a specific TF. **(C)** A genome-wide fit to ChIP-chip serves to determine the relative affinity parameters that constitute the PSAM. (D) Plot of predicted promoter affinity versus ChIP fold-enrichment.

immunoprecipitation (ChIP) [1,2] and protein binding microarray (PBM) [3] experiments have generated data about TF-DNA interaction on an unprecedented scale. However, sophisticated computational methods are required to distill accurate sequence-to-affinity models from these data. The Bussemaker lab has pioneered methods that directly estimate the binding free energy parameters (ΔΔG) that define the sequence specificity of TFs by fitting biophysical models to high-throughput data (Figure 1) [4-7]. Ongoing work in the Bussemaker Lab aims to incorporate structural information about the TF-DNA interface as part of

these models, towards the ambitious goal of inferring a universal protein-DNA recognition code (Figure 2).

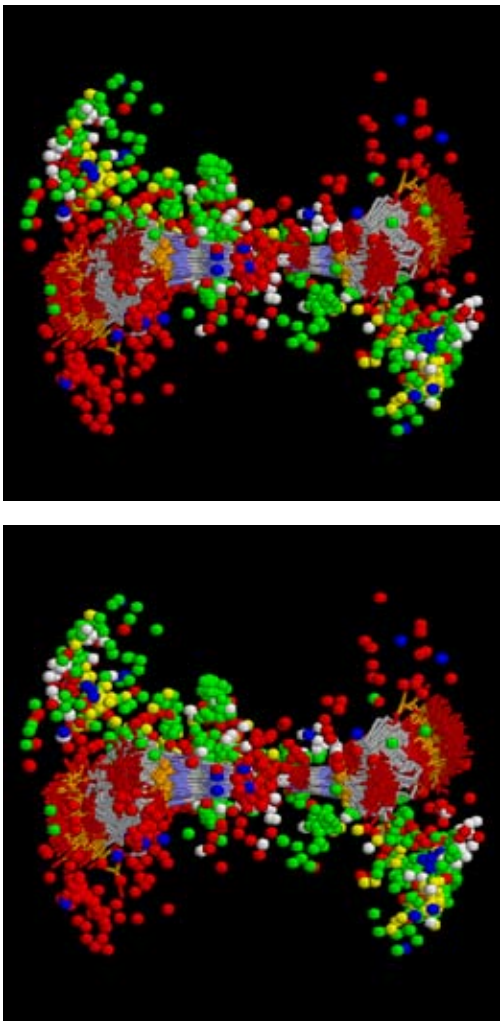


Figure 2: Decoding protein-DNA recognition.

The Bussemaker lab is developing methods that integrate structural information with high-throughput genomics data to better predict protein-DNA binding affinities. Shown here are all instances in the Protein Data Bank of an interaction between an A-T base pair in double-stranded DNA and a lysine side-chain in the DNA-binding domain of a TF (side (left) and top (right) views relative to the base pair coordinate frame). The spheres denote the position of each C_α backbone carbon; the side-chains themselves are not shown. Colors correspond to different TF structural families.

Images created by Dr. Xiang-Jun Lu in the Bussemaker Lab.

Hotspots of trancription factor binding *in vivo*.

While the relationship between DNA sequence and TF binding is still relatively straightforward in the test tube, this is not at all the case in the living cell, where interactions with nucleosomes and other chromatin-associated proteins all contribute to the binding profile of the TF along the chromosome. The recruitment of a TF to specific locations on the chromosome can in fact be completely independent of its DNA-binding domain. That this is true for about 5% of the fruit fly genome is the surprising discovery that was made in collaboration with Dr. Bas van Steensel at the Netherlands Cancer Institute and Dr. Kevin White, currently at the University of Chicago. Using DamID technology (which works differently from ChIP but provides similar information) [8] to map the binding of a large number of chromatin-associated proteins, it was found that most TFs in *Drosophila* are specifically recruited to 2-3kb large regions that together constitute about 5% of the genome [9]. Strikingly, these "hotspots" did not contain any predicted binding sites for the TFs. To further investigate this phenomenon, the binding of two variants of the well-known Bicoid (Bcd) protein was analyzed. The first variant consisted of only the DNA-binding domain of Bcd. As expected, it bound only to regions with high *in vitro* binding affinity. The second variant consisted of the entire Bcd protein but carried a point mutation that inactivates the DNA-binding domain. This protein was no longer bound the *in vitro* binding sites, but was still recruited very specifically to the hotspots. The Bussemaker lab is currently investigating to what extent interactions with nucleosomes and other TFs can account for this phenomenon.

Multi-gene domain organization of chromatin.

Proteins often function as part of a protein complex or biochemical pathway. The cell has therefore developed mechanisms for coordinately regulating the expression of multiple genes, such as those that involve transcription factors. Recently, however, it has become increasingly clear that physical proximity of genes along the genome can also serve to coordinate the regulation of functionally related genes. The Bussemaker and Van Steensel labs recently discovered that at least 50% of all fruit fly genes are organized into multi-gene chromatin domains bound by specific combinations of proteins (Figure 3). These domains are functionally coherent both in terms of gene expression and in terms of functional annotation, and evolutionary selection acts against chromosomal rearrangements that break them up.

Conclusion.

Taken together, these results underscore the complexity of *in vivo* transcription factor function. Dissecting the various mechanisms that contribute to their target specificity is likely to keep researchers busy in coming years, and additional surprises are undoubtedly still in store.

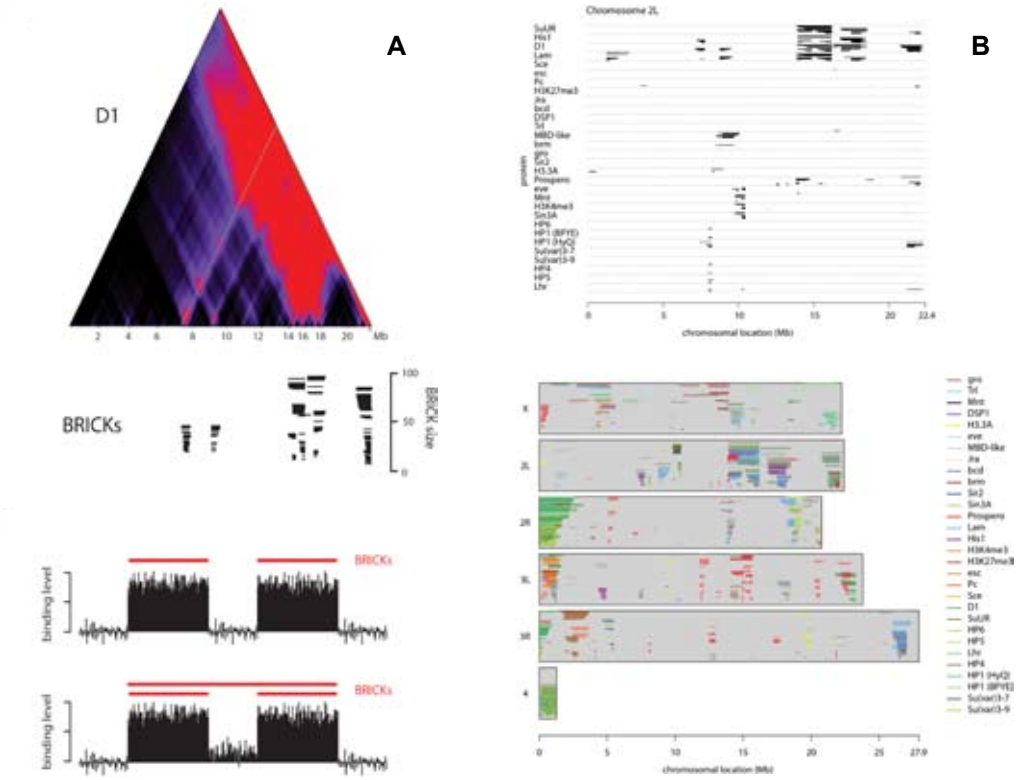


Figure 3: Global Chromatin Domain Organization of the *Drosophila* Genome.

(A) "Domainograms" allow visualization of chromosomal clustering of protein binding on all length scales simultaneously. A dynamic programming algorithm was developed that parses the binding profile along the chromomosome in terms of discrete domains or "BRICKs".

(B) Overview of the BRICKSs detected for all proteins that were mapped, revealing the large degree of chromatin domain organization in *Drosophila*.

Figures reproduced from De Wit et al. [10]

LITERATURE CITED

1. Ren, B. et al., Genome-wide location and function of DNA binding proteins. Science 290 (5500), 2306 (2000).

2. Iyer, V. R. et al., Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 409 (6819), 533 (2001).

3. Berger, M. F. et al., Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol 24 (11), 1429 (2006).

4. Foat, B. C., Houshmandi, S. S., Olivas, W. M., and Bussemaker, H. J., Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. Proc Natl Acad Sci U S A 102 (49), 17675 (2005).

5. Foat, B. C., Morozov, A. V., and Bussemaker, H. J., Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. Bioinformatics 22 (14), e141 (2006).

6. Foat, B. C., Tepper, R. G., and Bussemaker, H. J., TransfactomeDB: a resource for exploring the nucleotide sequence specificity and condition-specific regulatory activity of trans-acting factors. Nucleic Acids Res 36 (Database issue), D125 (2008).

7. Bussemaker, H. J., Foat, B. C., and Ward, L. D., Predictive modeling of genome-wide mRNA expression: from modules to molecules. Annu Rev Biophys Biomol Struct 36, 329 (2007).

8. van Steensel, B., Delrow, J., and Henikoff, S., Chromatin profiling using targeted DNA adenine methyltransferase. Nat Genet 27 (3), 304 (2001).

9. Moorman, C. et al., Hotspots of transcription factor colocalization in the genome of *Drosophila melanogaster*. Proc Natl Acad Sci U S A 103 (32), 12027 (2006).

10. de Wit, E. et al., Global chromatin domain organization of the *Drosophila* genome. PLoS Genet 4 (3), e1000045 (2008).

MAKING SENSE OF TRANSCRIPTION FACTOR SPECIFICITIES OR HOW TO GROW LEGS IN STRANGE PLACES: A COLLABORATION BETWEEN FLIES, BIOPHYSICS, AND COMPUTERS

BARRY HONIG, PHD [1,2] **RICHARD S. MANN, PHD** [1]

[1]DEPARTMENT OF BIOCHEMISTRY AND BIOPHYSICS, COLUMBIA UNIVERSITY
[2]CENTER FOR COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, COLUMBIA UNIVERSITY

Introduction

"In evolution, nature built these regulatory circuits; now the world is run by these switches" - Dr. Mark Ptashne, quoted in a NY Times article by PHILIP J. HILTS published February 24, 1998.

As most molecular biologists know, the "switches" Ptashne is referring to in the above quote are controlled by transcription factors (TFs), proteins that bind DNA and cause genes to be expressed or not. In a relatively simple case, such as the switch that governs phage lambda's choice between lytic and lysogenic states (a problem that Ptashne spent many years studying) the transcription factor is lambda repressor (Ptashne, 1992). Lambda repressor has two important functional domains: one that recognizes DNA and one that binds to other lambda repressor molecules. These two domains, in combination with precisely spaced binding sites in lambda DNA that help to promote repressor-repressor interactions, are sufficient to control nearby genes and thereby decide whether this switch will be flipped to the lytic or lysogenic state.

Remarkably, the principles discovered in lambda and other simple transcriptional switches are also used in more complex biological settings, such as in eukaryotic gene regulation. However, life is much more complex as a eukaryote. For one, genomes are generally much larger, making it more difficult for transcription factors to find the right binding sites. Second, higher eukaryotes tend to have many cell types, where it is not uncommon for the same transcription factor to regulate two very different sets of genes. Third, transcription factor binding sites can often be highly degenerate, which for most transcription factors results in the presence of more than one potential binding site in every gene. Finally, eukaryotic DNA exists as chromatin, where it is wrapped around large protein particles called nucleosomes. Thus, eukaryotic transcription factors not only have to find binding sites in DNA, they have to navigate through chromatin. Nevertheless, somehow eukaryotic cells "know" how to read and accurately interpret regulatory DNA, allowing them to create complex biological structures such as eyes, hearts, and limbs. As molecular biologists who study transcription factors in eukaryotes, our goal is to do what the cell can do, namely read the DNA and "know" what it means.

Some of the problems faced by eukaryotes are solved by the use of combinations of transcription factors to regulate genes. Such a combinatorial mechanism is especially useful for allowing the same transcription factor to execute different functions in different cell types: in cell type A, to regulate gene X, transcription factors a, b, and c might be required, while in other cell types, and at other genes, other unique combinations of transcription factors may be used.

A special, yet widely used type of combinatorial control is when transcription factors bind to DNA cooperatively. The definition of cooperative DNA binding is when two or more TFs bind to DNA with a much higher affinity together than the sum of their individual affinities. Most typically, and exemplified by lambda repressor, this occurs when the proteins can interact with each other, as well as with DNA. If the binding sites are arranged appropriately, the protein-protein interaction can significantly stabilize the final protein-DNA complex, thus making its formation 'cooperative'.

Yet, despite these insights, we are still far from interpreting the sequences of eukakryotic regulatory DNA with any accuracy. Certainly, the principles learned from simpler systems must apply in eukaryotes, but our general inability to predict the function of regulatory DNAs suggests that there may be more to this problem than a simple direct readout of the DNA sequence by DNA binding domains. By trying to solve the problem of how a unique subset

of eukaryotic transcription factors, encoded by the Hox genes, function, we may have discovered such a new twist on DNA recognition by DNA binding proteins.

Hox transcription factors

The homeotic or 'Hox' genes were first discovered by genetic experiments carried out by Ed Lewis in the 1950s and 1960s in the fruit fly, *Drosophila melanogaster* (Lewis, 1978). These genes first caught the attention of geneticists and developmental biologists because mutations in them caused body parts, such as the legs, antennae, and wings, to switch identities. For example, the bithorax (bx) mutation, one of the first to be discovered, results in a small structure known as the haltere, which helps the fly balance during flight, to develop as a wing instead of a haltere (Figure 1). Analogously, the original mutation in the Antennapedia (Antp) gene causes antennae to develop as legs. These swaps in developmental fates were termed homeotic transformations, following the concept of homeosis as first defined by William Bateson who described, in a now classic book, similar types of aberrations in wild populations of animals (Bateson, 1894). In fact, we now know that aberrations such as extra digits on hands or feet, a phenotype that Bateson included in his book, are due to changes in Hox gene expression (Goodman, 2002).

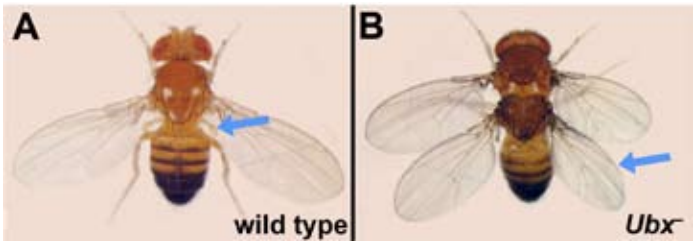


Figure 1: Hox mutant phenotypes.

(A) Wild type fruit fly with one pair of wings and one pair of halteres (blue arrow).

(B) Ubx mutant fruit fly, showing a transformation of haltere to wing (blue arrows). After Lewis, 1978.

Moving ahead several decades, we now know that probably all multicellular animals have a series of Hox genes (8 in the fruit fly; 39 in humans) (McGinnis and Krumlauf, 1992). Each Hox gene is expressed in a specific subset of cells in developing embryos, most typically in different regions along the anterior-posterior (AP) axis (Figure 2). In other words, most cells of a developing embryo express some combination of Hox genes that depends on their AP position within the body plan.

Hox genes all encode transcription factors that bind DNA using a DNA binding domain known as the homeodomain. Thus, the answer to homeosis -- why cells make a leg instead of an antenna – is somehow embedded into the function of these transcription factors. It is worth here emphasizing what the genetic results tell us: namely, that the presence or absence of single Hox transcription factors determines the developmental outcomes of entire body parts. Thus, while it is almost certainly the case that these transcription factors never work alone, which one is present, and how they bind DNA and regulate their target genes, is the key to which structure an animal will build.

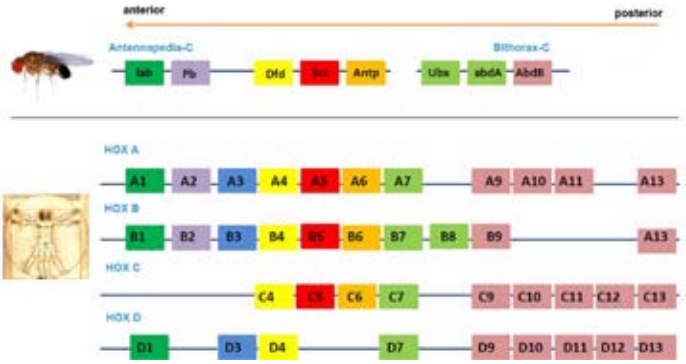


Figure 1: Hox complexes.

Fruit flies have a single set of eight Hox genes, split into two complexes, the Antennapedia Complex (Antennapedia-C) and the Bithorax complex (Bithorax-C). Humans have four sets of Hox genes (39 in total), Hox A, Hox B, Hox C, and Hox D, likely resulting from duplications that have occurred during evolution. Within each Hox cluster, and in both fruit flies and vertebrates, the genes are ordered along the chromosome in the same order they are expressed along the embryonic anterior to posterior axis. In vertebrates, there has also been additional expansion of the posterior AbdB-like genes.

The homeodomain enigma

In principle, if we knew how Hox proteins bind and regulate the correct target genes, we'd be a long way towards understanding how these big developmental decisions, like whether to make an antenna or a leg, get made. The same is true for many transcription factors that sit atop developmental hierarchies. Another good example is Eyeless, which, like the Hox factors, is also a homeodomain protein. Also first discovered in fruit flies, this highly conserved transcription factor is known to be important for making eyes in many animal species, whether the eye is a compound eye as in the fruit fly or a human eye (Gehring, 1996). The problem, however, is that

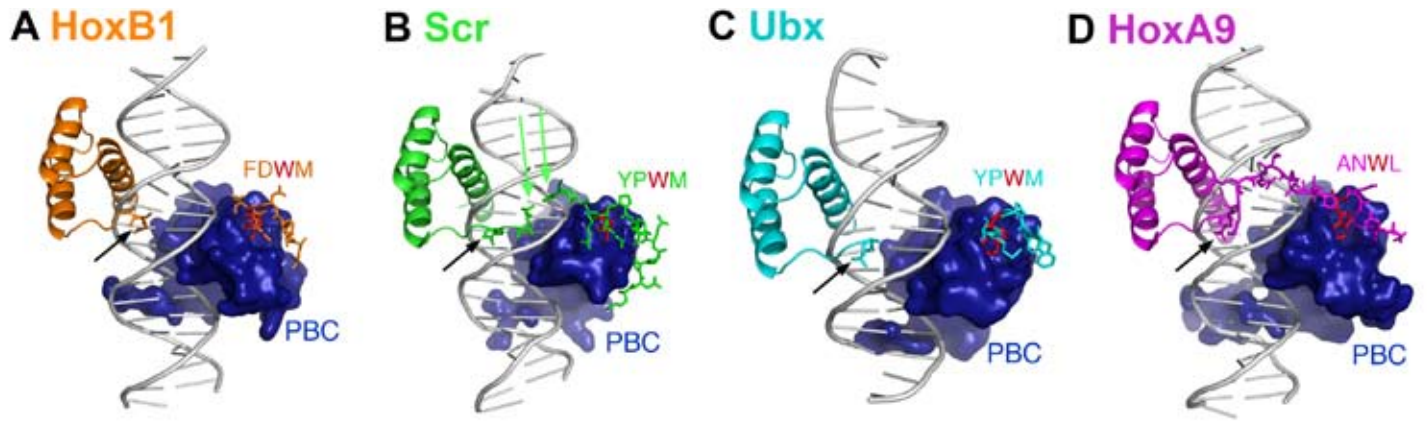


Figure 3. Hox-PBC complexes.

Crystal structures of four Hox-PBC-DNA complexes, HoxB1-Pbx (A), Scr-Exd (B), Ubx-Exd (C), and HoxA9-Pbx (D) are shown. All four complexes show a very similar arrangement of the Hox and PBC homeodomains, and a nearly identical protein-protein interaction mediated by the Hox 'YPWM' motif (or close variants) and the PBC homeodomain. In addition to showing a similar overall arrangement of the homeodomains, all four structures show Arg5 of the Hox homeodomain interacting with the minor groove. However, of these four, only the Scr-Exd structure was solved with a Hox-specific binding site (fkh250) and only this structure shows additional basic side chains inserting into the DNA minor groove (green arrows, B). Structures were described in (Joshi et al., 2007; LaRonde-LeBlanc and Wolberger, 2003; Passner et al., 1999; Piper et al., 1999).

homeodomain proteins have notoriously low DNA binding specificities. In fact, nearly all Hox proteins bind to very similar DNA sequences *in vitro* (Affolter et al., 2008; Berger et al., 2008; Noyes et al., 2008). What's worse is that Hox proteins even have similar DNA binding specificities to a wealth of other homeodomain proteins present in eukaryotic cells, including Eyeless! Yet, which genes a Hox protein regulates to make a leg must be dramatically different from those that Eyeless regulates to make an eye. Clearly, something must be missing in these *in vitro* DNA binding experiments that cells use to allow them to carry out their specific functions *in vivo*.

As described above, transcription factors often use the trick of binding DNA cooperatively with other factors to increase their specificity. About 12 years ago, a protein that binds cooperatively with Hox proteins was described (Mann and Chan, 1996). In flies, this protein is called Extradenticle (Exd), while its three mammalian orthologs are called Pbx1, Pbx2 and Pbx3 (Moens and Selleri, 2006). Over the past decade, many studies have shown that Exd/Pbx (which are collectively referred to as PBC proteins) are critical for Hox proteins to carry out their specific functions *in vivo*. Interestingly, PBC proteins are also homeodomain proteins, and the cooperative complex formed between Hox and PBC factors is a head-to-tail dimer for which several X-ray crystal structures now exist (Figure 3) (Joshi et al., 2007; LaRonde-LeBlanc and Wolberger, 2003; Passner et al., 1999; Piper et al., 1999).

To some researchers, the finding of PBC-Hox cooperative binding "solved" the specificity problem, at least for Hox proteins. However, in reality this finding only raised more questions than it answered. In particular, PBC proteins have the capacity to form cooperative heterodimers with nearly all of the Hox factors. Thus, to a first approximation, PBC-Hox cooperativity only provides an additional set of protein-DNA contacts (via the PBC homeodomain), but these new contacts are the same for all of the Hox factors. In other words, for PBC proteins to actually help distinguish between the specificities of different Hox proteins, they would have to do something more than just provide an additional set of protein-DNA contacts. Instead, they might uncover hidden specificity information that was built into the Hox proteins. One way to think about this is that PBC factors might change the conformation of a Hox protein, such that it could now "read" the DNA in ways that it could not in the absence of PBC proteins.

PBC proteins reveal latent specificity information built into Hox proteins

Although there was plenty of circumstantial evidence to support such a model (Mann and Chan, 1996), recently, in part due to funding from a MAGNet grant, we have obtained direct evidence supporting this idea (Joshi et al.,

AcademyeBriefings

Crossing Paths

The RECOMB Regulatory Genomics / Systems Biology / DREAM Conference

Conference Chairs:
Manolis Kellis, Massachusetts Institue of Technology
Andrea Califano, Columbia University
Gustavo Stolovitzky, IBM Computational Biology Center

Sponsored by the Merck Research Laboratories, IBM, and the NIH Roadmap Magnet Center

Held at the Broad Institue, Cambridge, MA | October 29 - November 2, 2008

Powerful new experimental tools have created opportunities for diverse groups of researchers to contribute to studies of biology - not just biologists and chemists but physicists, engineers, mathematicians, and computer scientists. Although each group brings unique strengths to bear, they also speak different dialects and are excited by different challenges. One consequence is a proliferation of small and specialized conferences.

On October 29 - November 2, 2008, in Cambridge, Massachusetts, the RECOMB Regulatory Genomics/ Systems Biology/ DREAM conference bucked this trend by bringing three separate conferences together in a single venue. Over four tightly scheduled days, the meeting combined the 5th RECOMB Satellite Conference on Regulatory Genomics, chaired by Manolis Kellis, the 4th RECOMB Satellite Conference on Systems Biology, chaired by Andrea Califano, and the 3rd DREAM Conference, chaired by Gustavo Stolovitzky.

The first two conferences had previously spun off from the RECOMB conference on Research in Computational Molecular Biology. RECOMB was founded in 1997 as a forum for computer science issues in biology, and was last held in March-April 2008 in Singapore. The DREAM conference (Dialog for Reverse Engineering Assessments and Methods) began in 2006 with a more focused goal of evaluating systems biology tools for building biological networks.

New York Academy of Sciences
Building communities, advancing science since 1817

How good network inference aglorithms really are?

At a microscopic level, organisms are ruled by interacting systems of biomolecules. Historically, scientists painstakingly elucidated chains of molecular events using experiments that reveal individual interactions. In recent years, researchers have built richer, interconnected networks to mathematically summarize their knowledge of these interactions. This systems biology enterprise, largely stimulated by high-throughput tools like microarrays that measure mRNA levels as an indicator of gene expression, is a vital and increasingly important activity in both basic biology and in medicine.

A nagging concern, however, is how accurately these networks represent the biology. For complex systems like biological networks, there are practical limits on how well even massive amounts of data can uniquely define the underlying structure and yield useful predictions of measurable events. Indeed, although its advocates call this process "reverse engineering," the topology and the detailed molecular interactions of the "inferred" networks will likely never be known with precision.

To address such concerns it is important to define a formal framework for assessing the quality of biological network prediction algorithms. For an activity of this type to be successful it is important that the research community participate and offer critical input. The DREAM conference and competition were established to enable and foster this participation. The main objective of the DREAM initiative is to catalyze the interaction between experiment and theory in the area of cellular network inference. The fundamental question for DREAM is simple: How can researchers assess how well they are describing the networks of interacting molecules that underlie biological systems? From the earliest planning stages of the DREAM project, a key component of the initiative was the development of a competition in which different teams competed in using the same, blinded data to infer the networks that had generated it. Perhaps only in this way can the community know whether the networks that their methods produce can be trusted. The idea was inspired in part by other competitions, notably the CASP assessment of algorithms for protein-structure prediction.

The protein-folding challenge, however, begins with a precise amino-acid sequence and ends with a three-dimensional structure that is experimentally well defined. For reverse engineering of networks, both the specification of the data and the evaluation of the results are much harder.

A major problem for DREAM is identifying gold-standard networks whose structure can be taken as known. The best-understood networks are those created by people, but many researchers have expressed concerns that these networks would hold little interest for the larger biological community. What has emerged as an acceptable compromise is to select a range of "challenges" that span both large and small biological networks as well as mathematical networks, and, in between, a synthetic network implemented in yeast.

The purpose of DREAM is not to produce the best possible network, but to evaluate the best tools for producing networks. What is still needed, and what DREAM aims to achieve, is a fair comparison of the strengths and weaknesses of the methods and a clear sense of the reliability of the network models they produce. To achieve the goal of meaningful comparisons, the DREAM project provides a gold standard against which the competitors' results are scored. Ultimately, the first and most important step in seeking to understand data is human insight and combining intuition with computational tools to reveal new and powerful strategies.

Information about upcoming conferences as well as challenge data from previous competitions are available at the DREAM project web site, http://wiki.c2b2.columbia.edu/dream/.

2007). Acquiring this evidence required the combination of methods from biochemistry, genetics, biophysics, and structural biology, providing a compelling argument for why multidisciplinary approaches in biology can be extremely powerful and need to be supported in the future.

In the center of this work lie two X-ray crystal structures in which the same two proteins, the Hox protein Sex combs reduced (Scr) and the PBC factor Extradenticle (Exd), were crystallized on two different binding sites. The first binding site, called fkh250 is a native binding site from the forkhead (fkh) gene, a natural target of Scr in *Drosophila* (Andrew et al., 2000; Ryoo and Mann, 1999). The activation of fkh by Scr is critical for making salivary glands during embryogenesis, and no other Hox factor has this ability. Thus, the regulation of fkh by Scr is what we refer to as a Hox-specific function, and contrasts with other Hox functions that may not require such exquisite specificity. Importantly, and consistent with this idea, Scr-Exd heterodimers bind to fkh250 10 to 20 times better than other Hox-Exd heterodimers (Ryoo and Mann, 1999). Thus, unlike most other well-studied Hox binding sites (including those used in previous structures), fkh250 exhibits the type of specificity *in vitro* which would be expected of an Hox-specific binding site. The second binding site, called fkh250con, differs from fkh250 in only three positions. In contrast to fkh250, most Hox-Exd heterodimers bind fkh250con with high affinity, approaching the affinity that Scr-Exd has for fkh250. Thus, unlike fkh250, fkh250con does not exhibit specificity for a particular Hox factor. The Scr specificity exhibited by fkh250, and the lack of specificity exhibited by fkh250con, is also observed when these binding sites are used to drive artificial reporter genes *in vivo* in *Drosophila* embryos, strengthening the argument that these binding sites contain all of the information required to produce Hox-specific (as in the case of fkh250) or Hox-non-specific (as in the case of fkh250con) readouts *in vivo* (Ryoo and Mann, 1999).

The two crystal structures described in Joshi et al. (2007) thus provide a unique and direct comparison between the same protein complex, binding its 'specific' binding site (fkh250) and a 'Hox-non-specific' binding site (fkh250con). This comparison revealed several unexpected features, some of which are likely to be critical for understanding DNA binding specificity by these proteins. For one, the DNA minor groove had a different shape in the two structures. In fkh250 the minor groove is narrower in the center of the Hox-Exd binding site compared to the equivalent region of the fkh250con binding site. Theoretical and computational work showed that this narrower minor groove leads to a more negative electrostatic potential (Figure 4). In other words, the fkh250 binding site has a negatively charged pocket right in the center of the Hox-Exd binding site. In contrast, the equivalent position in the fkh250con binding site, although still negative, is not as negative (Figure 4). Using a separate set of computational methods, we also found that the shapes of the minor grooves seen in both complexes are likely to be a direct result of the DNA sequence, and is not induced by protein binding. In other words, the shape differences present in these two DNA sequences are intrinsic to these DNAs.

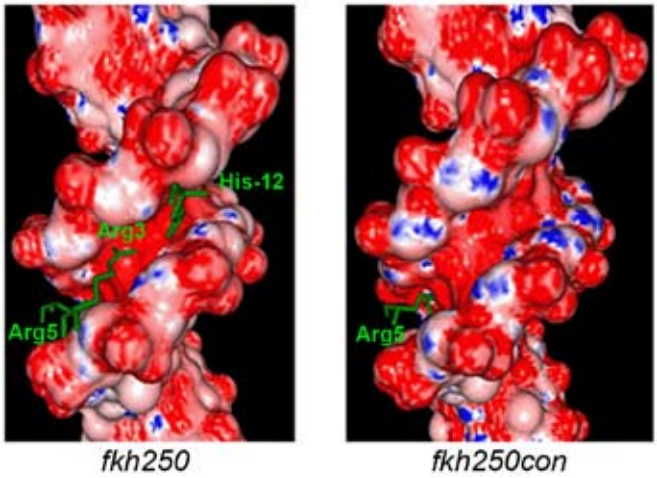


Figure 4. Shape and charge differences between fkh250 and fkh250con

Shown are DELPHI images of fkh250 (left) and fkh250con (right), illustrating the shape and charge differences between these two DNAs. Red: negative; blue: positive. Arg5 of Scr is present in both minor grooves, but Arg3 and His-12 are only observed inserting into the minor groove of the fkh250 binding site. The shape of the minor groove is narrower where these side chains are inserting in the fkh250 binding site compared to the equivalent region of the fkh250con binding site. See Joshi et al (2007) for details.

The negative pocket present in the fkh250 binding site is, we believe, critical for its Hox-selectivity, and a model summarizing this is shown in Figure 5. In particular, these findings suggest that the localized negative electrostatic potential in fkh250 results in it being a poor (i.e. low-affinity) binding site for most Hox-Exd heterodimers. If so, then why can Scr-Exd bind to this site with such a high affinity (~10 nM Kd)? The answer is that Scr has basic residues (an Arginine and a Histidine) in its N-terminal arm (a part of the homeodomain) and nearby linker that insert into this negative pocket, thus counteracting its repulsion. Most other Hox proteins do not have these basic residues

in the equivalent position, and thus cannot overcome the negative repulsion of the fkh250 binding site.

How well do these results support the model that Exd reveals latent specificity information built into Hox proteins? The answer is: remarkably well. Scr needs Exd to position the Arg and His so that they can insert into the negative pocket (Figure 5). Without Exd, the peptide (Scr N-terminal arm and linker) that these residues come from is not structured and is more likely to be interacting with water rather than DNA. Thus, although the Arg and His residues are present in Scr, they are not able to "read" the

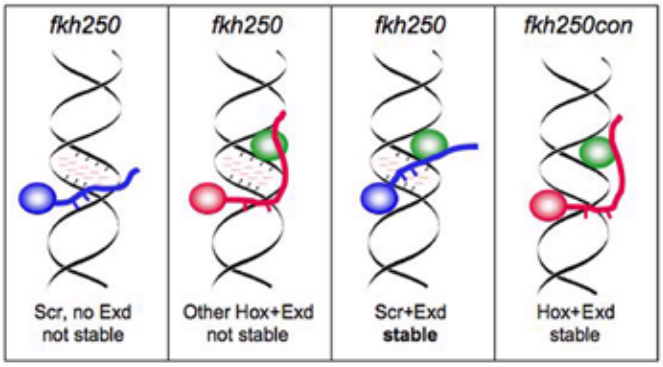


Figure 5. Hox specificity depends on local shape recognition: a model.

The first three panels show Hox-Exd dimers in the presence of the fkh250 binding site, which has a narrow minor groove (small arrows) and negative electrostatic potential (pink dashes) in the center of the binding site. The fourth panel shows the fkh250con binding site, which does not have these characteristics. Scr-Exd, but not other Hox-Exd dimers, can effectively bind to the fkh250 binding site because Exd positions a normally unstructured peptide so that the basic side chains (short blue lines) can insert into the negative pocket formed by the narrow minor groove. In contrast, because fkh250con does not have this negative pocket, it is less selective and can bind multiple Hox-Exd dimers.

DNA unless Exd is there to help force these residues into the appropriate conformation (Figure 5).

Local shape recognition: a common mode of DNA recognition?

The results with Scr-Exd-fkh250 raise another novel, and potentially general, feature of DNA recognition by transcription factors. Specifically, because the basic side chains present in Scr are inserting into a negative pocket formed by the unusually narrow minor groove in fkh250, we suggest that they are reading a DNA shape, not a specific DNA sequence. This mode of DNA recognition contrasts with the classical way that proteins are thought to bind specific DNA sequences, which depend on hydrogen bonds formed between amino acid side chains and DNA base pairs.

The idea that proteins read a DNA shape – in this case, the shape of the minor groove – appears to be a previously unknown mode of protein-DNA recognition, one that we call local shape recognition. Indirectly, of course, the shape of the minor groove is a consequence of the DNA sequence. Our results, however, suggest that different DNA sequences can generate similar shapes. Conversely, distinct DNA sequences that fit the same "consensus" binding site for a particular factor (such as the AT-rich DNA sequences that homeodomains like to bind to) can have different shapes. Thus, if local shape recognition is generally used by DNA binding proteins, it may be a mechanism to distinguish between DNA sequences that all conform to the same consensus sequence.

Our recent results, in fact, strongly suggest that local shape recognition of DNA by transcription factors may be widely used in biology. Systematic analyses of all available protein-DNA structures present in the Protein Data Bank (PDB) reveal that the pattern of minor groove width varies tremendously in DNA sequences recognized by a large number of DNA binding domains. Moreover, in most of the minor groove width minima present in these structures, there is an Arginine side chain, suggesting, as is the case for Scr-Exd, that the insertion of basic amino acid side chains into narrow minor grooves may be commonly used by DNA binding proteins.

Implications and future prospects

As alluded to at the start of this article, molecular biologists are still unable to do what living cells can do, namely, read a DNA sequence and interpret its regulatory properties. The results summarized here, which came from a unique convergence of structural biology, biochemistry, developmental biology, and biophysics, suggest that we may also have to take local DNA structure into consideration to fully decode regulatory DNA. Although almost always a double helix, small deviations from the canonical B-DNA structure, such as short stretches in which the minor groove is narrower than usual, can have profound effects on DNA recognition by transcription factors and perhaps other factors that interact with DNA. Work in the future must first continue to test the generality of these findings. If, however, these findings are as general as they currently seem to be, we must next devise new tools to decipher this previously unappreciated mode of protein-DNA recognition through which nature achieves in some cases exquisite specificity. No doubt that the deepest insights will continue to come when a multidisciplinary approach is applied, as exemplified by the work described here.

LITERATURE CITED

Affolter, M., Slattery, M., and Mann, R. S. (2008). A lexicon for homeodomain-DNA recognition. Cell 133, 1133-1135.

Andrew, D. J., Henderson, K. D., and Seshaiah, P. (2000). Salivary gland development in *Drosophila melanogaster.* Mech Dev 92, 5-17.

Bateson, W. (1894). Materials for the study of variation: treated with special regard to discontinuity in the origin of species (London: Macmillan and Co.).

Berger, M. F., Badis, G., Gehrke, A. R., Talukder, S., Philippakis, A. A., Peña-Castillo, L., Alleyne, T. M., Mnaimneh, S., Botvinnik, O. B., Chan, E. T., et al. (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell 133, 1266-1276.

Gehring, W. J. (1996). The master control gene for morphogenesis and evolution of the eye. Genes Cells 1, 11-15.

Goodman, F. R. (2002). Limb malformations and the human HOX genes. Am J Med Genet 112, 256-265.

Joshi, R., Passner, J. M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M. A., Jacob, V., Aggarwal, A. K., Honig, B., and Mann, R. S. (2007). Functional specificity of a Hox protein mediated by the recognition of minor groove structure. Cell 131, 530-543.

LaRonde-LeBlanc, N. A., and Wolberger, C. (2003). Structure of HoxA9 and Pbx1 bound to DNA: Hox hexapeptide and DNA recognition anterior to posterior. Genes Dev 17, 2060-2072.

Lewis, E. B. (1978). A gene complex controlling segmentation in *Drosophila*. Nature 276, 565-570.

Mann, R. S., and Chan, S. K. (1996). Extra specificity from extradenticle: the partnership between HOX and PBX/EXD homeodomain proteins. Trends Genet 12, 258-262.

McGinnis, W., and Krumlauf, R. (1992). Homeobox genes and axial patterning. Cell 68, 283-302.

Moens, C. B., and Selleri, L. (2006). Hox cofactors in vertebrate development. Dev Biol 291, 193-206.

Noyes, M. B., Christensen, R. G., Wakabayashi, A., Stormo, G. D., Brodsky, M. H., and Wolfe, S. A. (2008). Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell 133, 1277-1289.

Passner, J. M., Ryoo, H. D., Shen, L., Mann, R. S., and Aggarwal, A. K. (1999). Structure of a DNA-bound Ultrabithorax-Extradenticle homeodomain complex. Nature 397, 714-719.

Piper, D. E., Batchelor, A. H., Chang, C. P., Cleary, M. L., and Wolberger, C. (1999). Structure of a HoxB1-Pbx1 heterodimer bound to DNA: role of the hexapeptide and a fourth homeodomain helix in complex formation. Cell 96, 587-597.

Ptashne, M. (1992). A Genetic Switch: Phage λ and Higher Organisms, 2nd edition edn (Cambridge, MA: Blackwell Scientific Pulbications Inc.).

Ryoo, H. D., and Mann, R. S. (1999). The control of trunk Hox specificity and activity by Extradenticle. Genes Dev 13, 1704-1716.

RNA VIRUSES AS PROBES OF EVOLUTION

RAUL RABADAN LAB

Viruses are obligate intracellular parasites and the evolution of a virus is inexorably linked to the nature and fate of its host. One therefore expects that virus and host genomes should have common features.

The innate immune response provides a first line of defense against pathogens by targeting generic differential features that are present in foreign organisms but not in the host. These mechanisms generate selection forces acting both on pathogens and hosts that further determine their co-evolution. Our lab studies the fingerprints of these selection forces acting in parallel on both host innate immune genes and ssRNA viral genomes. Biases are identified in the coding regions of innate immune response genes in plasmacytoid dendritic cells and then used to predict significant host innate immune genes. We then compare the significant motifs in highly expressed innate genes, to those in ssRNA viruses and study the evolution of these motifs in the H1N1 influenza genome. The deeply under-represented motif pattern of CpG in an AU context - which is found in the both ssRNA viruses and innate genes, and has decreased throughout the history of H1N1 influenza - is immunostimulatory and has been selected against during the co-evolution of viruses and host innate immune genes. This shows how differences in host immune biology can drive the evolution of viruses that jump to a species with different immune priorities than the original host.

Since the spring of 1977, two subtypes of influenza A virus (H3N2 and H1N1) have been seasonally infecting the human population; this pattern is very different from what was observed after previous influenza pandemics. In 1918, 1957, and 1968, new viral strains completely supplanted the prior ones. The reappearance, in May of 1977 in Northern China, of the H1N1 virus, a virus that had been considered extinct in the human population since 1957, had serious consequences. Children became especially sick because they had never been exposed to the H1N1 virus. Since 1978, the influenza vaccine has contained H1N1 viral antigens, in addition to the previously circulating H3N2. In the last few years, there has been an international effort to sequence influenza isolates, and to make publicly available the extensive information that has been gathered on them. We have studied the distribution of patient ages within the populations that exhibit the symptomatic disease caused by each of the different subtypes of influenza virus; when information is pooled across multiple geographical locations and seasons, striking differences emerge between these subtypes. The symptomatic flu due to H3N2 is distributed across all age groups, whereas H1N1 causes symptomatic disease mainly in a younger population. This trend is probably a remnant of the effect that was observed in 1977, i.e. young persons were more affected by the H1N1 virus than were older ones. The above findings suggest that a previous exposure to an influenza subtype confers a long-lasting protection, even more than 30 years after the differential event (1977). Each subtype affects its own characteristic spectrum of age groups. This "signature" is relevant to age-related risk assessments, modeling of epidemiological networks for specific age groups, and age-specific vaccine design.

Viruses present such a diversity and fast evolution that standard techniques of sequence alignment fail to provide significant similarity between emerging viruses and the ones already known. We are developing and implementing algorithms to analyze High Throughput Sequencing data that allow the identification of emerging viruses without relying on sequence alignment.

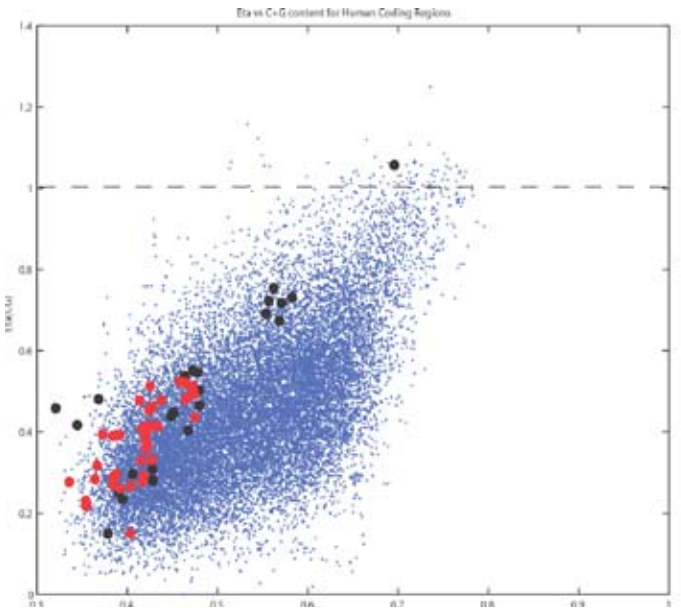


Figure 1: Comparison between human genes and human ssRNA viruses. CpG odds ratio versus C+G content for human genes in blue. Superimposed on top of these genes are the human single stranded RNA viruses (human ssRNA+ (black) and Human ssRNA−(red)). We can appreciate how human ssRNA viruses follow a similar distribution as human genes.

HELPING RESEARCHERS MANAGE AND ANALYZE GENOMIC DATA

ARIS FLORATOS AND ANDREA CALIFANO LABS

High throughput genomic technologies are increasingly becoming key drivers of significant discoveries in biomedical research. The massive nature of the data produced by such technologies, however, means that in order to utilize them effectively it is necessary to develop and deploy non-trivial management and analysis infrastructure. NIH programs like the National Centers for Biomedical Computing (which the MAGNet Center is a member of) and the cancer Biomedical Informatics Grid (caBIG®) fund the development of tools whose aim is to help researchers meet these challenges.

In an effort to further encourage and support the adoption of these sophisticated tools, caBIG® has recently launched the Enterprise Support Network (ESN, https://cabig.nci.nih.gov/esn), a collection of resources and organizations providing services, mentoring and expertise to members of the biomedical research community who are interested in exploring the caBIG® software and technology offerings. A component of the ESN, the Molecular Analysis Tools Knowledge Center (MATKC, https://cabig-kc.nci.nih.gov/Molecular/KC/) is one of six such subject-specific centers recently established. It is jointly managed by our labs at Columbia University and by the Broad Institute of MIT and Harvard University. geWorkbench (www.geworkbench.org), the bioinformatics platform of the MAGNet Center, is one of the 4 applications that the Center has been tasked to support; the other three are (1) caArray (https://array.nci.nih.gov/), a microarray gene expression data repository developed by NCI, (2) GenePattern (www.genepattern.org), a data analysis platform, and (3) caIntegrator (http://caintegrator-info.nci.nih.gov/), a framework for enabling the correlated ("translational") analysis of clinical and laboratory data. The Center maintains and monitors user and developer forums, a Wiki site offering documentation and support for the four applications, a community bug tracking and feature request system, and a knowledge base comprising articles that describe in detail how to address common technical and analysis issues. The Center's mission, as with each of the Knowledge Centers, is to serve as an authoritative repository of knowledge and information about the supported tools and technologies.

COMMUNITY-DRIVEN KNOWLEDGE SHARING FOR THE DISCOVERY AND VISUALIZATION OF WORKFLOWS IN GEWORKBENCH

GAIL KAISER LAB

We are investigating knowledge sharing for computational scientists, demonstrated in a prototype called **genSpace** that is implemented as an add-on to MAGNet's geWorkbench (http://www.geworkbench.org). It is expected that scientists collaborating in the same lab on the same project share: Data (specimens, samples, materials, analyses), Tools (instruments, software, hardware), and most significantly Knowledge (open discussion, whiteboard). Our primary motivation is to address the temporal (time) and physical (space) constraints preventing this model from scaling to communities of scientists working on different projects but who could potentially learn from each other's expertise, experience, etc. and thus produce better results for humanity.

Most current generation Computer-Supported Cooperative Work systems enable data sharing and/or tool sharing (e.g., PNNL Collaboratories, UIUC BioCoRE). But knowledge sharing (how/when/where/why to use tools and data) has previously been limited to labor intensive approaches such as publications, email mailing lists, wikis, spontaneous on-line or real-world chats, etc. Our scientists already have too many demands on their time. We instead seek to enable automatic knowledge sharing that requires zero "extra work".

Our approach leverages the now very popular, and intuitively easy to use, social network concepts such as collaborative filtering ("people like you …") to disseminate knowledge on how best to use geWorkbench and its numerous integrated analysis and visualization tools.

genSpace logs, aggregates, and data mines geWorkbench users' activities to recommend what have proven through frequent use to be the most useful tools and tool sequences (workflows). Individual users can opt-in or opt-out to activity logging as desired, e.g., due to privacy or confidentiality concerns, but still obtain recommendations based on activities by other users. genSpace can answer the following questions that a novice, or even intermediate to expert, geWorkbench (or analogous analysis tool integration system) user might ask:

- What do I do first?
- Which tools work well together?

- Where does this tool fit in a typical workflow?
- Who do I know who also uses this tool?
- How can I get help (from an expert who is online right now)?

Further information is available at http://www.psl.cs.columbia.edu/genspace.

USING GEWORKBENCH TO ACCESS THE TERAGRID INFRASTRUCTURE

ARIS FLORATOS LAB

The TeraGrid (http://www.teragrid.org/) is operated by a consortium of National Laboratories and universities, and uses a high-speed network to share major computing and data resources (with access to more than 750 Teraflops of computing power and 30 Petabytes of online and archival storage). It is currently the largest such "cyber-infrastructure" for open scientific research. caGrid (www.cagrid.org), is the grid middleware layer of the caBIG® initiative. caGrid, built on some of the same underlying technologies as the TeraGrid, provides a number of enhancements to promote the exposure and reuse of clinical and laboratory data through its emphasis on shared data models and adherence to data interchange standards. As such, the caGrid infrastructure currently focuses more on data and security than on computing power. Christine Hung from our lab worked closely with Ravi Maduri of the University of Chicago/Argonne National Labs and others to develop and demonstrate a gateway that provided a secure mechanism to transfer caGrid computing jobs to the TeraGrid. This facility can be used to give researchers access to the raw computing power of the TeraGrid when required. In the project, geWorkbench was utilized as the front end user interface tool, and a caGrid service was developed that provided access to a TeraGrid job queue. A geWorkbench hierarchical clustering algorithm was placed on the TeraGrid host. The entire process from launching a caGrid job using geWorkbench to running the job on the TeraGrid and retrieving the results for display in geWorkbench was successfully demonstrated at the caBIG® 2008 Annual Meeting.

This project is described in full detail in a paper presented at Teragrid'08http://www.teragrid.org/events/teragrid08/Papers/papers/100.pdf.

MODELING NOISE IN TRANSCRIPTIONAL REGULATION: INFORMATION FLOW IN REGULATORY CASCADES

CHRIS WIGGINS LAB

The past decade has seen great advances in our understanding of the role of noise in gene regulation and the physical limits to signaling in biological networks. In recent work (Aleksandra M. Walczak, Andrew Mugler, Chris H. Wiggins, "A stochastic spectral analysis of transcriptional regulatory cascades", PNAS 2009, to appear) we introduced a spectral method for the computation of the joint probability distribution over all species in a biological network. The spectral method exploits the natural eigenfunctions of the master equation of birth-death processes to solve for the joint distribution of modules within the network, which then inform each other and facilitate calculation of the entire joint distribution. We illustrate the method on a ubiquitous case in nature: linear regulatory cascades. The efficiency of the method makes possible numerical optimization of the input and regulatory parameters, revealing design properties of, e.g., the most informative cascades. We find, for threshold regulation, that a cascade of strong regulatory events converts a unimodal input to a bimodal output, that multimodal inputs are no more informative than bimodal inputs, and that a chain of up-regulations outperforms a chain of down-regulations. We anticipate that this numerical approach may be useful for modeling noise in a variety of small network topologies in biology.

We are collaborating with the laboratory of Prof. David Miller at Vanderbilt University, who uses pioneering cell-sorting and microarray-based technologies to profile mRNA isolated from individual neurons, gradually expanding our knowledge. Using the limited existing knowledge, we already have some preliminary results, and we are currently using novel computational techniques that we developed to identify sets of genes that are synergistically interacting with respect to synapse formation.

GRID-ENABLEMENT OF BIOINFORMATICS WORKFLOWS

ARIS FLORATOS LAB

The caBIG® Integrative Cancer Research Workflow Working Group (led by Kiran Keshav, a member of our lab) is chartered with developing useful workflows from available caGrid-enabled data and analytical services. The group's first proof of concept project was to implement a microarray-based workflow drawing gene expression data from caArray (https://array.nci.nih.gov/), performing preprocessing using a GenePattern (http://www.genepattern.org/) grid service, and then running hierarchical clustering using a grid-enabled geWorkbench component. More recently, the group demonstrated a proteomics workflow using the Computational Proteomics Analysis System (CPAS, https://proteomics.fhcrc.org/CPAS/), the Protein Information Resource (PIR), and caBIO (https://cabig.nci.nih.gov/tools/cabio), a caBIG® programmatic interface for accessing biological information.

Additionally, the Workflow Working Group investigated workflow authoring and invocation using the Taverna Workbench (http://taverna.sourceforge.net/). The group has worked with the caGrid development team to prototype a workflow authoring tool to enable the orchestration, discovery and invocation of caBIG® grid services and has provided the caBIG® community with guidelines for creating workflows as well as feedback on tool and process enhancements. All artifacts from the Workflow Working Group can be found on the Gforge website at http://gforge.nci.nih.gov/projects/workflow/.

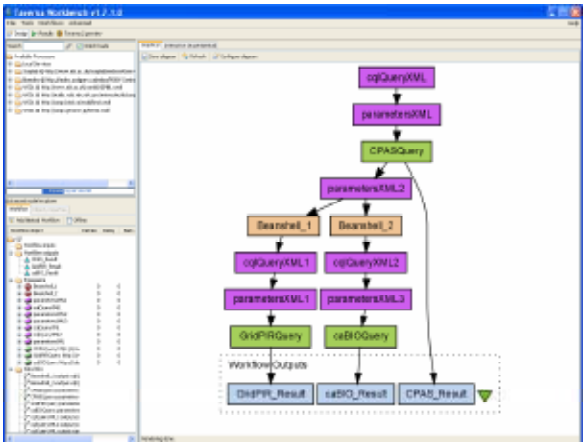


Figure 1: A caGrid workflow defined in TAVERNA.

THE GERMLINE ALGORITHM DISSECTS RECENT POPULATION STRUCTURE BY HIDDEN RELATEDNESS

ITSIK PE'ER LAB

Itsik Pe'er's Lab recently developed GERMLINE, a robust algorithm for identifying segmental sharing indicative of recent common ancestry between pairs of individuals. GERMLINE efficiency, orders of magnitude better than previous methods, facilitates analysis of high throughput data: thousands of samples genomewide. Application of this method to current datasets facilitates novel insights on recent effects on population structure, based on the fact that pairs of individuals from closely inbred populations are more likely to share significant chunks of their genomes due to recent common ancestry. This is clearly demonstrated by the analysis of 1000 samples from the New York Health Study, a public-access dataset hosted by C2B2 within the Intragen database (IntragenDB, http://intragen.c2b2.columbia.edu/). A connected component emerges that essentially identifies the individuals self-reported as Ashkenazi Jewish (blue), separating them from other New Yorkers with European ancestry (green). This is expected from random graph theory given a difference in average degrees between nodes representing individuals of different ethnic groups. Inclusion of Ashkenazi samples collected by the Hebrew University Genetic Resource (cyan) confirms this analysis. PhD student Sasha Gusev, the developer of GERMLINE and MSc student Pier Palamara were able to use the population-specific chance of hidden relatedness to observe geographic separators between clusters of different populations. Furthermore, the locations of these segments that are shared between individuals without recombination are focused at specific loci, including the HLA and large CNV regions, suggesting a biological mechanism for conservation of intact haplotypes.

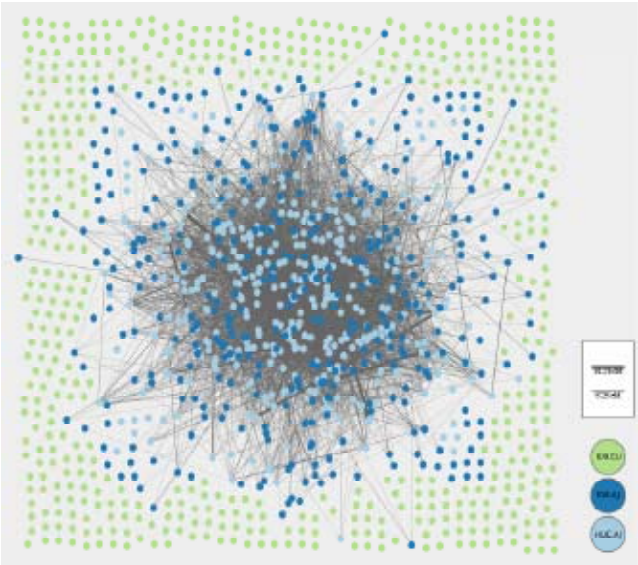


Figure 1: GERMLINE analysis of IntragenDB samples clearly separates Ashkenazi Jews (blue) from other New Yorkers with European ancestry (green).

ADDING SEMANTIC DIMENSION TO RANKING OF PUBMED SEARCH RESULTS

KENNETH ROSS LAB

An ever-increasing amount of data and semantic knowledge in the domain of life sciences is bringing about new data management challenges. Many life sciences researches search PubMed as part of their daily activities. With the number of articles in PubMed growing from year to year, and with many queries returning thousands of high-quality matches, there is a clear need for relevance ranking of results. Such ranking is not currently available in PubMed.

Kenneth Ross and his students Julia Stoyanovich and William Mee are developing a system that will add the semantic dimension to literature search. Their system incorporates

several families of novel ranking functions that use MeSH annotations to determine the relevance of an article to a user's query. The system implements novel adaptive algorithms that compute the ranking efficiently on the scale of PubMed.

When complete, the system will allow ranked browsing, and will also provide a two-dimensional visualization of results that plots article relevance against publication date.

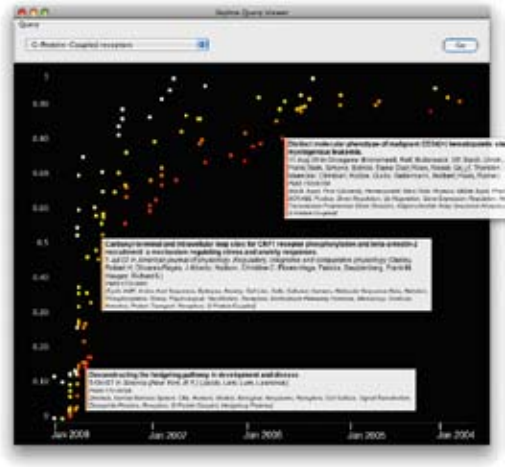
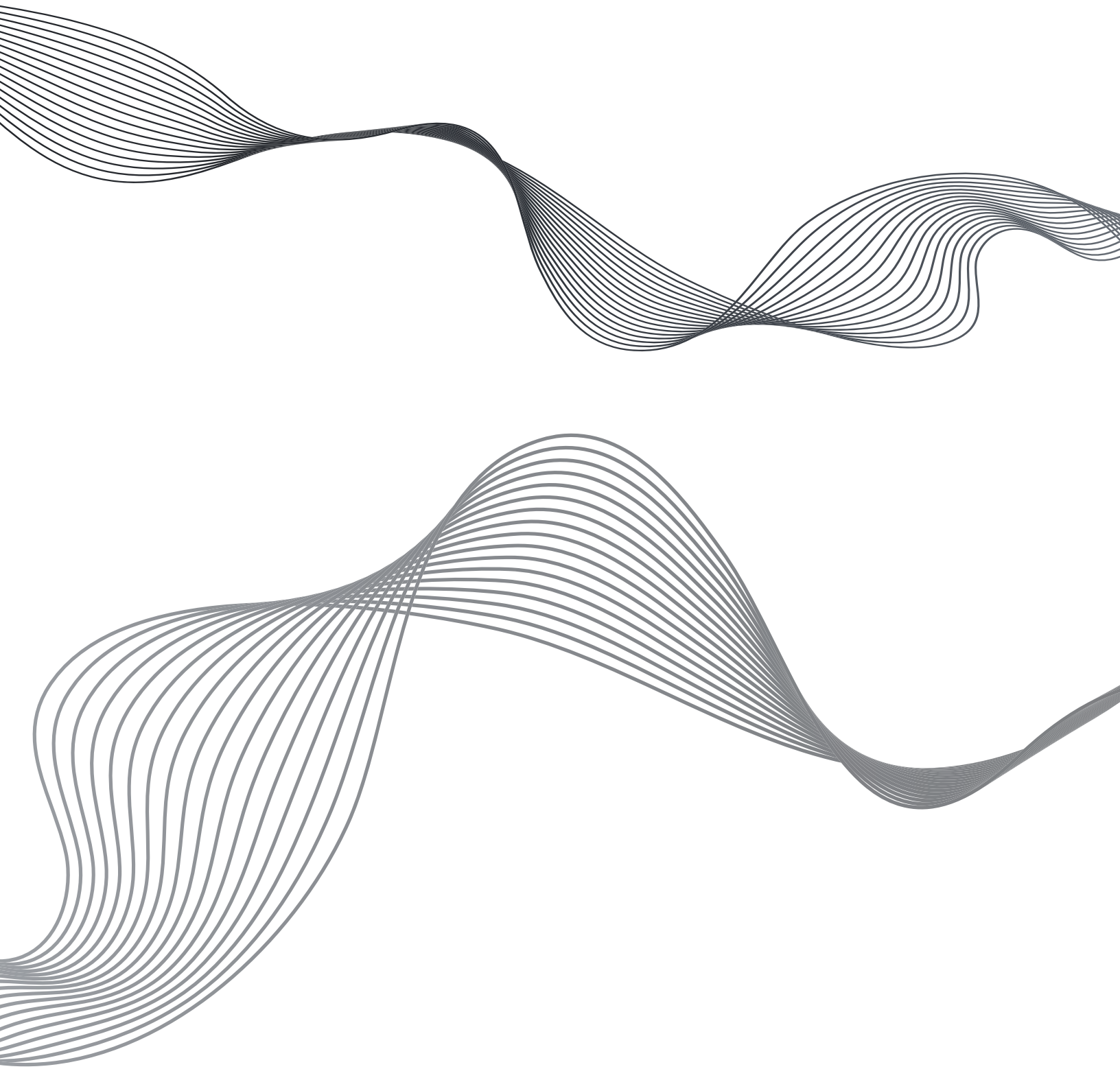
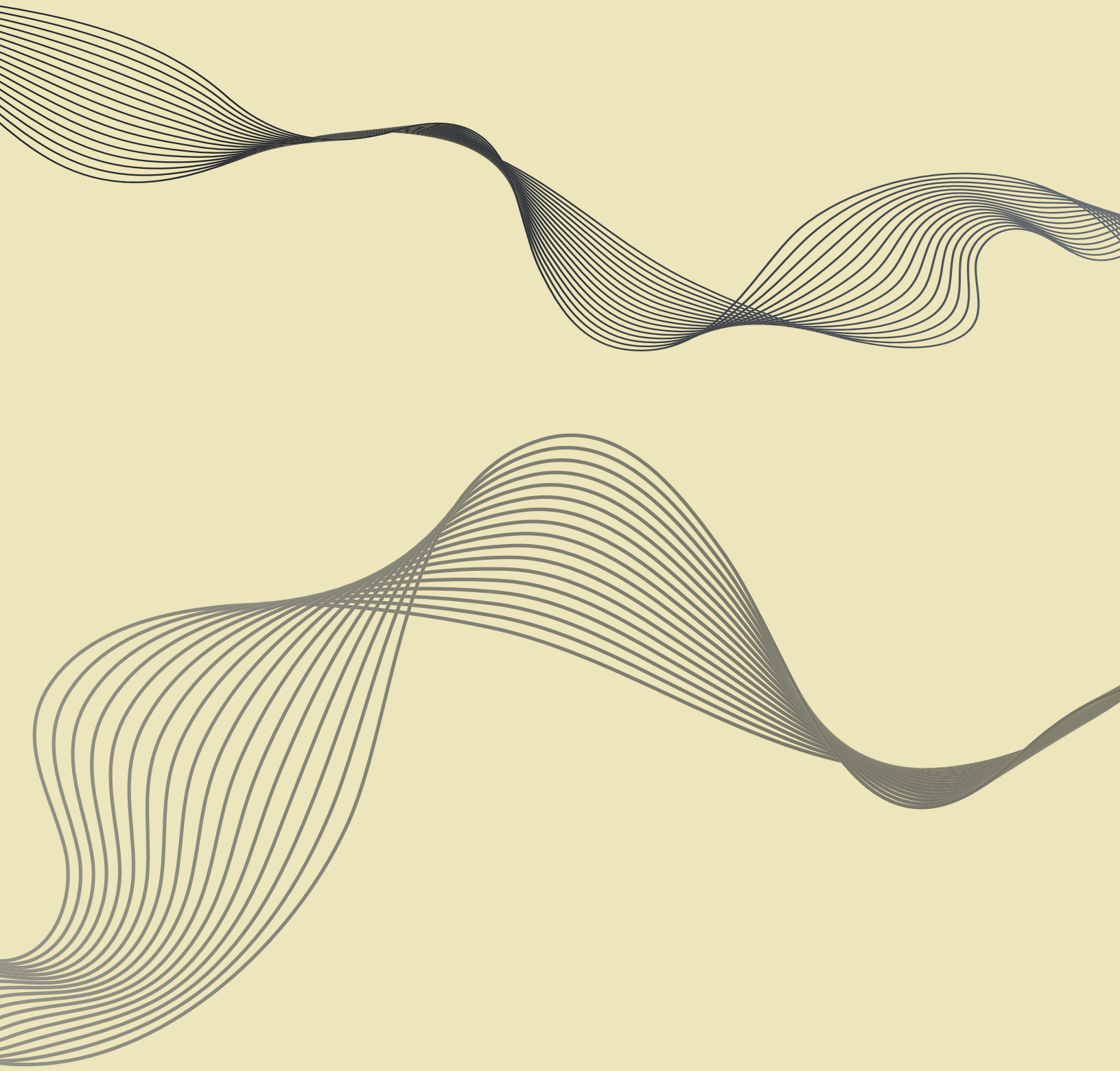


Figure 1: Two-dimensional *Skyline* visualization of results for the query "G-Portein-Coupled receptors". Document publication date is plotted on the x-axis, while semantic query relevance is plotted on the y-axis. Higher values of query relevance correspond to better matches.





Columbia University
Center for Computational Biology and Bioinformatics
1130 St. Nicholas Avenue
New York, NY, 10032

Winter 2009
Issue No. 2