Yves Brun,
Systems Biology/Microbiology Faculty Search,
Department of Biology,
Indiana University,
Jordan Hall 142, 1001 E 3rd St,
Bloomington IN 47405-7005

Dear Dr. Brun,

I am writing to briefly introduce my candidacy for the tenure-track faculty position at Indiana University. I am currently a research assistant professor at Boston University, running a dynamic computational biology research laboratory. The lab, comprised of two graduate students under my supervision, is involved in independent and collaborative projects with theoretical and experimental colleagues. I am also teaching an introductory Computational Biology course.

I began my education at the University of Illinois, Urbana-Champaign, majoring in Computational Biophysics. Immediately after receiving my Bachelor's degree, I went on to join one of the first graduate programs in Bioinformatics, at Boston University. I was awarded the NSF pre-doctoral fellowship to study molecular evolution of protein domains using graph-theoretic techniques. I received my Ph.D. in May, 2004 under the guidance of Prof. Charles Delisi and stayed on at Boston University as a research assistant professor. Since starting graduate school four years ago, I published fifteen papers in peer-reviewed journals.

In my future research, I plan to build on the significant progress in the area of protein evolution to guide investigations into the mutational dynamics of upstream regulatory regions. To accomplish this, I plan to expand the current effort in transcription factor binding site identification in complex eukaryotic systems. Next, I plan to use the data generated from *de-novo* whole-genome binding site identification to define cellular control networks, and to understand how those networks emerge as a result of molecular evolution. I hope that this work will yield insights across a wide spectrum of biological problems from structure-function relationship in proteins to emergence of pathways and speciation.

Attached please find my CV, a brief description of my research interests (Sec A,B), a teaching statement, and a detailed outline of future projects(Sec C). I believe that my interests and goals perfectly complement the excellence of teaching and research missions of the department. I am especially excited to collaborate with the laboratories of Profs. Michael Lynch and Matthew Hahn. I am looking forward to meeting the distinguished faculty at the Department of Biology and sharing my enthusiasm for research and teaching. If you have any questions, please don't hesitate to contact me directly.

Sincerely,
Boris Shakhnovich

Boris Shakhnovich
Research Assistant Professor,
Bioinformatics Department,
24 Cummington St.,
Boston University,
Boston, MA 02215
Email: borya@acs.bu.edu;
Laboratory Website: http://romi.bu.edu

# Teaching Statement

*Boris E Shakhnovich*

I believe that mentorship, whether in the classroom or in the lab, is a sacred trust. Students place the weight of their future at our doorstep with no thought of recourse. We have to live up to that trust as teachers by inspiring and guiding the students on their chosen career path. While my current position is on a research track, I have requested extended teaching responsibilities. In fact, the most rewarding aspect of my promotion to a faculty position has been a more direct role in teaching courses and mentoring graduate students. During my tenure as a research assistant professor, I have supervised two graduate, two rotation, and four undergraduate students. I have taught 3 classes at Boston University: two seminar courses, and an additional course I developed that covers introductory topics in computational biology (to download the lectures and homework from that class please visit ***romi.bu.edu/bf527***). I have also been extensively involved in helping to plan and amend the curriculum of the Bioinformatics program.

Students who choose to join multidisciplinary programs are a brave and industrious bunch. Since I, myself, have gone through the curriculum at an interdisciplinary graduate program, I have first-hand experience of the kinds of challenges facing both students and instructors in computational biology. First, the multifaceted nature of the field increases the difficulty of identifying the common core of the science. Second, the subject matter is rapidly evolving and so are the skills and knowledge needed to succeed. Finally, the novelty of the field limits the choices in teaching materials and resources available to instructors. Thus, I think that in choosing courses, students and instructors should focus on subjects where fluency will have sustained impact by providing a suitably general framework applicable to a variety of research areas.

To give a sense of the wide array of constitute parts that make up the core of computational biology, we can consider the canonical example of sequence alignment. The solution to that problem is a fusion of relatively standard computer science, coupled to rigorous statistics, illuminated by a flash of insight about the role of evolution. Most importantly, students need to have a firm grasp of the basic tenets of evolution and quantitative population genetics, such as origins of variation, selection, and drift in order to put their own research into a proper perspective. The knowledge of statistics is indispensable as evidenced by the fact that application of extremal probability theory to sequence alignment made computational biology the requisite tool for every biologist and broadly accessible to the scientific community. Students should be well versed in chemistry and physics as these skills are necessary for understanding the basics of DNA, RNA and protein function. What makes curricular development difficult in this area is the need for students to understand both the fundamentals of quantitative disciplines and their application to real biological problems. In my opinion, their knowledge of biology should span the gamut from the basic anatomy of the gene, to prokaryotic and eukaryotic genome organization and mechanisms. Students must understand biology on a level such that they can collaborate with experimental colleagues, ask relevant questions, and guide their own research.

Many of the above subjects belong to the realm of well-established disciplines (e.g. physics, math, biology, chemistry), and are partially covered on their home departments. However, to learn the amount of material necessary for research in computational biology, students must take a prohibitively large number of these courses. Also, the traditional approaches to presenting these subjects rarely provide the students with computational biology problems either as motivation, examples, or homework. Thus, students miss out on the subtleties of how to properly ask biologically interesting and relevant questions when taking courses in statistics, graph-theory, or algorithms, and computationally relevant questions when taking courses in chemistry or molecular biology. I believe that students will benefit from a more topical approach. For example, in my own course, I assign homework that encourages students to apply fundamental skills and knowledge to interdisciplinary problems and a final project where they can explore current research areas.

While this statement outlines only some of my philosophy about teaching, I am excited to discuss these issues in person. I would be comfortable teaching courses in computational and systems biology, bioinformatics, molecular biology, evolution and population genetics. Specifically, I am interested in incorporating multidisciplinary problems into the course material and introducing students to a broad perspective on the subject. I am also interested in developing a concise curriculum that prepares students for research in their area of interest. I look forward to participating in both teaching and curricular development.

## A. Introduction:

The increasing rate of data deposition from high-throughput experimental assays challenge and undermine status-quo approaches to biology. Beginning with whole-genome sequencing, high-throughput acquisition of biological data has recently been extended to protein structure, transcription factor mediated regulation, and functional properties of genetic mutations. Along with static data, there are now techniques to survey the phenotypic state of whole cell systems such as microarray-based expression profiling, protein abundance, and metabolite quantification. While the boon of experimental techniques represents a paradigm shift in the way scientific communities perceive biology, a comprehensive theory describing the relationship between these data and the evolutionary processes that may have led to its generation is still largely lacking. Aside from its importance as inspiration for future studies, the lack of a holistic framework increases the difficulty in assessing the gathered data for fidelity and possible bias. For the past year, with the help of a small group of dedicated students, and in close collaboration with both experimental and computational scientists, I was involved in research addressing some of these issues.

The broad theme of my research proposal is the *elucidation of the coordinated evolution of open reading frames and upstream regulatory regions*. The proposed research projects build on my early work in molecular evolution and on recent success in *de-novo*, computational identification of transcription factor binding sites and cis-regulatory modules. The proposal is further divided into five inter-related areas: (i) *Evolution of Transcriptional Control Regions*, (ii) *Extending our Understanding of the Structure-Function Relationship*, (iii) *High-performance Computational Approaches to Transcription Factor Binding Site Discovery* (iv) *Defining System-Level Organization of Regulatory Mechanisms* and (v) *Publication of Data, Resources and Tools*. Section **B** outlines a brief overview of the goals and methods in each area. Section **C** provides an in-depth discussion of the preliminary evidence and proposed projects. The research outlined here will help understand how pathways evolve through duplication of open reading frames and mutational re-wiring of control regions. Along the way, I plan to continue developing tools, innovative frameworks and open-source databases that aid experimental biologists in research defining, testing and validating eukaryotic regulation models and structure-function relationships in proteins.

While the proposal may seem ambitious and expansive, the goals and methods are well defined, many of the projects are well underway and preliminary results are encouraging enough to fully support feasibility. Beyond the challenges outlined here, the comprehensiveness of the proposed approach — promises to elucidate aspects of the inter-relationship between structural, functional and regulatory elements of cellular organization. Specifically, by choosing to investigate biological systems at different levels of granularity, I hope to shed light on novel evolutionary mechanisms responsible for changing the phenotype of the cell.

## B. Outline of major areas of interest

### *B.1 Correlated evolution of genes and controlling structures*

One of the most exciting emerging areas of computational biology is concerned with the study of both the evolutionary[1, 2] and synthetic[3-5] rewiring of transcriptional control. There is now significant evidence that everything from altering gene function to speciation[6] can occur as a result of mutations in the upstream regulatory regions. For organisms, altering transcriptional control may be a faster way to change gene function or invent a new pathway without relying on major genomic rearrangements or the relatively slow evolution of open reading frames. However, there is also significant evidence for conservation of upstream regions, especially in the positions responsible for transcription factor binding[7-9]. In fact, conservation is the most common assumption when searching for transcription factor binding sites (TFBS) using computational techniques[10-12].

*I am interested in understanding the dynamics between conservation and rewiring in the evolution of transcriptional control.* However, despite increasing numbers of whole-genome sequences and high-throughput experimental techniques mapping transcription factor (TF) binding to upstream regions[12-15], our ability to predict functional positions is limited.[16] This rather limited understanding of the genome outside the open reading frame complicates prediction of regulatory mechanisms, modeling molecular evolution, and annotating gene function.

Even if functional positions were characterized with high fidelity, models describing the impact of mutations in the upstream region on the phenotype of the cell are as of yet inadequate. The strength of selection, and the probability associated with fixation of mutations affecting transcriptional control are often hard to estimate. First, transcriptional activation is often context dependent, and mutation in the upstream region may influence not only the strength of binding between the TF and the DNA, but also the gating logic[17] that defines the temporal expression of the downstream product. Furthermore, successful adaptation to change in regulation control requires correlated mutations of the open reading frames. Other considerations include the potential of other genes and upstream regions in the genome to mutate in a coordinated fashion to create a new pathway. Another set of unknowns are the environmental constraints influencing the impact of these mutations on the fitness of the organism. All these variables are integral in defining the strength of selection on mutations in upstream promoter regions. Many are at present impossible to measure with accuracy, but the role of some can be elucidated using recent results and proposed projects outlined in this proposal.

*One of the major goals of this proposal is to capitalize on the significant progress in our understanding of protein evolution to guide investigations into evolution of upstream promoter regions.* Specifically, I am interested in the mutational events, dynamics, and mechanisms responsible for changing protein function. Regulatory elements in upstream regions play an integral role in placing gene function in cellular context. Thus, successful investigation into the co-evolution of upstream regions and open reading frames must take into account  all constitute parts and requires (i) *Accurate mapping of functional TF binding sites in upstream regions* because false positives or negatives introduce noise into evolutionary models; (ii) *Understanding transcriptional control in context* (e.g. multi-TF control via cis-regulatory modules controlling pathways). This is useful for defining the phenotypic consequences of mutations; and (iii) *Exploration of the evolutionary dynamics of open reading frames.* The latter can be used to elucidate the genetic potential to adapt to change of function. My lab has been working intensely in all three avenues, laying the foundation for a cohesive research program exploring the relationship between evolution of transcriptional control and evolution of protein sequence, structure, and function. My progress and future plans for tackling each of these problems is briefly outlined below and in detail in section **C**.

### *B.1.1 Mapping transcription factor binding sites*

A number of recent publications describe mutational dynamics of upstream regions. However, the interpretation of these results is often difficult. For example, some models describe a gradual mutation process of functional sites in upstream regions leading to change in TF binding specificity[2], while others describe a sudden and remarkably simultaneous exchange of multiple binding sites across the genome resulting in sudden emergence of a novel pathway[1]. High fidelity TFBS identification methods are an indispensable prerequisite for assessing accuracy of evolutionary models and for future studies into the evolution of upstream regions. Determining the inter-relationships between TFs and genes can then be used to guide experimental studies probing control and regulation[15, 18].

Computational approaches aimed at identifying binding positions can be divided into two categories: those that look for conserved *w*-mers in upstream regions of hypothetically co-regulated genes, and others that search for conservation between orthologous upstream regions. Current approaches are shown to work with limited success on easy input sets. When some of the input upstream sequences do not share the TFBS, (modeling experimental error or incorrect co-regulation hypothesis), or there is more than one binding site present, the effectiveness of the computational methods quickly degenerates. Independent assessments of these algorithms report accuracy between 10-40% for the most easily identifiable TF binding motifs in *S. Cerevisiae* [12, 19-22]. These performance limits render current algorithms essentially useless for *de-novo* identification of binding sites in whole genomes or when regulation mechanisms are complex and poorly understood (e.g. Humans). Finally, current implementations of algorithms are plagued by many false-positive predictions and their application is unsuitable to modeling evolutionary processes in upstream regions.

One strategic goal of this proposal is to develop and implement a systematic computational framework for *de-novo* discovery of transcription factor binding sites (TFBS) (See **C.3.1**). The primary objective is to make the edifice applicable to both mapping TFBS in whole-genomes (See **C.4.1**) and elucidation of complex regulatory models of transcription often found in eukaryotes. (See **C.3.6**)

*First,* I describe the GibTig algorithm currently under development in the lab. (See **C.3.1-2**) The algorithm is a Gibbs-sampling inspired approach and uses positional conservation in conjunction with parallelized computing strategies to increase the number of true-positive and almost eliminate false-positive predictions of binding sites. I propose several theoretical and practical improvements to the current implementation of the algorithm. The final product will be deployed in a distributed computational environment and possess improved predictive capabilities. (See **C.3.3**)

*Our second goal* is to create a global map of TFBS in the Yeast genome. (See **C.4.1**) We aim to use the CHIP-Chip data[13] along with paralogous gene families likely sharing functional attributes (See **C.1.0**) and clustering of genes related by more remote homology (See **C.2.4**) to create sets of hypothetically co-regulated genes for input into the GibTig algorithm. (See **C.3.1**) The improved sensitivity from the GibTig TFBS identification method in conjunction with the computational discovery of co-regulated genes should allow for significant improvement in sensitivity, specificity and coverage of TFBS identification in Yeast. We can use these results to iteratively improve the GibTig algorithm with training from retrospective and prospective experimental studies. Since this is the most computationally intensive aspect of the proposal, I aim to leverage our close collaboration with the BlueGene team at IBM.

*As a third goal*, I plan to take advantage of our promising results in the GABA model system (See **C.3.4-5**) to further elucidate the mechanisms of regulation for the GABA receptor in human neo-cortical cells. (See **C.3.6**) Application of computational methods to identify TFBS in mammalian genomes presents unique challenges (See **C.3.1**) due to increased length of promoter regions, complex regulation models, and relative dearth of experimental information identifying co-regulated genes. Despite these difficulties, we were able to use GibTigs to successfully predict and experimentally validate DNA sequences specifically bound by proteins in neuronal nuclear extracts. Proposed projects include finding tissue-specific regulatory elements and predicting mutations likely to significantly affect binding affinity. (See **C.3.6**) This project will also serve to validate and refine our TFBS and CRM (see **C.4.2**) identification strategies. (See **C.3.6**)

*Finally*, I plan to disseminate all validated and hypothetical binding sites and regulation models using a dynamic, query-driven database. The database design will allow the user to build and interact with models describing complex transcriptional control in eukaryotic genomes. (See **D.5.2-3**)

### *B.1.2 Regulation in Context*

Gene expression is modulated by combinations of transcription factors that bind all or some subset of upstream promoter sites in a condition dependent manner [23]. Since a particular transcription factor will generally be able to bind upstream regions and participate in the modulation of several genes, the relationship between transcription factors and genes is many to many. A more comprehensive understanding of the role that individual binding positions play in regulating transcription requires a model that incorporates other factors involved in modulating expression of the downstream gene.

Correlated expression of genes controlled by cis-regulatory modules (CRMs) is a major mechanism directing biochemical, signaling, and other pathways.[24-26] Defining the TF composition of CRMs will shed light on the biological switches that control coordinated expression of genes involved in common pathways. Thus, integration of individual TF binding sites into CRMs (e.g. relating groups of TFs to the genes they regulate) is an important and necessary step in discovery of pathways, higher order regulation processes, and the cellular control network. Furthermore, understanding and delineating subsets of genes controlled by CRMs may prove to be an invaluable addition to the growing repertoire of techniques aimed at *de-novo* prediction of individual gene function.[12, 27-30] Finally, one of the foremost challenges in understanding the potential impact of mutations in upstream regions is the detailed elucidation of the relationships between subsets of TFs controlling correlated expression of subsets of genes.

We have already enjoyed initial successes in reconstructing major control switches and pathways using a neural network trained on data from the partial mapping between TFs and genes from CHIP-Chip experiments[13]. (See **C.4.0**) I briefly outline a computational strategy for *de-novo* reconstruction of the cellular control network in *S. Cerevisiae.*(See **C.4.2**) First, I will use newly developed measures of functional distance (See **C.2.5**) to identify potentially co-regulated genes (See **C.4.1**). I will then use GibTigs (See **C.3.1**) to create a whole-genome transcription factor binding map (See above and **C.4.1**). Once the map of functional positions is completed, we plan to use the already implemented neural net based on adaptive resonance theory (ART)[31, 32] to identify sets of TFBS shared by subsets of genes. These binding profiles will define cis-regulatory modules. The genes they regulate are candidates for sharing common pathway function. We can then use CRMs to explore protein interactions between TFs, redundancies in the cellular control network, and the prevalent gating logics regulating transcriptional control.

Next, I will make an effort to redefine gene function using TFBS profiles (See **C.4.3**) and begin exploring the relationships between CRMs and evolution of open reading frames (See **C.1.1**). Since we understand some of the impact of functional constraint on the dynamics of gene duplication and divergence (See **C.1.0**), we can use that knowledge to guide research into the relationship between CRMs and selection acting on open reading frames. We can also use the same kinds of methods to research the relationships between CRMs and molecular evolution.

## *B.2 Understanding protein evolution*

The study of molecular and organismal evolution depends on a proper understanding of protein evolution[33-35]. Mutations of upstream regions have to be understood in perspective of the open reading frames they help to regulate (i.e. changing the regulation of the protein makes sense only if that protein can adapt to its new function). Since proteins represent the most basic phenotype of the cell, they are often the substrates of selection. Genes are selected for, or against, because the proteins they code are more or less fitted to the environment. The crux of research into molecular evolution, or any study of evolution, is the study of change. Partly, work in the lab has been, and is going to continue to be concerned with the detailed study of changes in proteins with respect to their sequence, structure, function, and phylogenetic distribution.

### *B.2.1 Relationships between constraint, gene duplication and divergence*

Gene duplication and divergence has long been credited as the primary mechanism behind innovation in molecular evolution.[36] Currently, three pathways are hypothesized to affect the divergence of two loci after a duplication event. The most common outcome of a duplication event is non-functionalization, when one copy becomes a pseudogene[37, 38] while the parental locus is retained in its original form. The less prevalent, but nonetheless important alternatives are neo-functionalization, when one gene copy retains the original function while the other is free to find a new function undergoing nearly neutral evolution[36], and sub-functionalization, when both copies retain purifying selection through sub-division of parental pleiotropy.[39-41] *One of my primary interests involves investigating the inter-relationships between functional constraint, purifying selection, and preference for a duplication pathway.*

Insights from this research will contribute to our understanding of the relationship between organismal pressure and molecular evolution not only on the level of individual proteins, but also on the level of gene families and pathways.

Recent studies in the lab have shown that the strength of purifying selection does not necessarily depend on any phenotypic determinant of the gene. Instead, selection is characteristic of membership in a gene family (See **C.1.0**). Surprisingly, while mutations accumulate uniformly slower for all genes in families under strong selection, paralogs in those families diverge farther in sequence. Observation of fewer pseudogenization events in families under strong selection supports a model where paralogs preferentially divide ancestral function after duplication. The observed homogeneity of purifying selection on all genes in paralogous families could be used as a signature of the relationship between function and evolutionary pressure. Recently duplicated paralogs are likely to perform similar functions, and their divergence in function space is limited[42]. In turn, limitation on functional divergence may impose corollary constraints on the variability of strength of purifying selection. This line of reasoning implies that dynamics of molecular evolution are inherently contextual[12, 43]. Since gene families not only share a constrained set of functions[42], but also a characteristic strength of purifying selection, any gene's ability to fix after the duplication event, diverge, neo-functionalize, or sub-functionalize is determined by membership of that gene in a family, and in turn that family's function in the organism[43].

I plan to use results outlined above to connect evolutionary dynamics of the gene family to the potential for functional change from mutations in the open reading frame and the upstream region. I plan to use GibTigs (See **C.3**) to explore the relationship between selection acting on members of paralogous gene families and transcription factor binding sites. This involves creating a model describing the likely mutational events contributing to the evolution of regulatory control. (See **C.1.1**) Does strong selection on a particular gene family mean that change in transcription is slower for those genes, or does the farther separation of paralogs in sequence imply a faster exchange of transcriptional regulation? Furthermore, understanding the impact of functional constraint on the dynamics of divergence after duplication will help in creating a TFBS map of the *S. Cerevisiae* genome (See **C.4.1**). Finally, a comprehensive framework describing duplication and divergence of paralogs will aid in further elucidating the relationship between sequence, structure, and function. (See **C.2.4**) –

### *B.2.2 The prevalence of convergent vs. divergent evolution*

While often forming the basis of evolutionary inquiries, documentation of differences alone is not sufficient for a comprehensive understanding of evolutionary processes. For example, some structural and sequence similarities can be attributed not to common evolutionary ancestry, but to favorable interactions of the protein backbone or particularly good packing arrangements and certain chain topologies[44-51]. An informed decision about the evolutionary relationship between two proteins requires not only the ability to identify structural, functional, and phylogenetic similarities, but also an in-depth understanding of the mechanisms and environmental pressures that could have lead to their generation[52-56]. *One fundamental question at the heart of molecular evolution is the role of historical versus physical factors in the observed distribution of sequences, structures, and functions.*

We have shown that some structural characteristics of proteins (e.g. designability approximated by the maximal eigenvalue of the contact matrix) correlate with sequence entropy of protein families[57]. A question open to debate is the relative contribution of the physical characteristics versus historical factors that underlie this correlation. I plan to continue exploring this aspect of protein evolution. Understanding the potential for sequence and functional diversity inherent in protein structure will help to further separate the relative contribution of history, physics and selection in molecular evolution. I will use the data from our studies on the dynamics and constraints affecting duplication (See **B.2.1**), to guide the exploration of mathematical models describing sequence divergence. For example, one suitable choice of model can be based on quasi-species descriptions first pioneered by Eigen.[58, 59] The traditional model can be easily amended to incorporate uniform selection on paralogs (See **B.2.1**) and structural determinants of designability.[57]

## *B.3 Functional annotation via structure homology modeling*

Almost equivalent to solving the problem of protein evolution is the problem of annotation through homology modeling [60-62]. Since most measures of similarity are based on evolutionary relationship, ability to accurately annotate new genes in recently sequenced genomes hinges in part on our ability to understand divergence of sequence, structure, and function. The problem is complicated because even very modest divergence from the closest homologue can carry with it the possibility of functional change[61, 63]. With sequence alignment nearing the limit of resolution and with fewer new genes amenable to functional annotation using standard homology techniques, new functional inference will come from our increased understanding of the structure-function relationship. The solution to the problem hinges in part on finding the integration of evolutionary pressures that yields the most precise measure of similarity between two arbitrarily chosen proteins[64, 65].

### ***B.3.1 Quantifying distance in function space***

In an attempt to increase generality and applicability, researchers have struggled to move away from comparisons largely driven by intuition and define quantitative distances between homologs in sequence[66], structure[67], and function[42, 68]. While rigorous measures for sequence and structure are now well established, the problem of defining functional distance has been particularly daunting. Existing database methods of describing functions using ontologies are not *a-priori* well-suited for calculating functional distances [68]. However, using mostly anecdotal evidence, researchers have shown that sequences sharing key structural characteristics often display common function[64]. Thus, there is general consensus on the principle of the structure-function relationship, but not its quantification.

We have previously shown[69] that functional distances between domains can be successfully quantified using Euclidian metrics applied to the gene ontology(GO)[70]. We are currently working on extending the simple Euclidian based measure of functional distance to a more accurate and sensitive kernel-based method. (See **C.2.4**). Briefly, we propose to use a diffusion kernel to implicitly describe local distances between gene ontology classifications. Having a precise distance in function space will enable us to quantitatively relate sequence and structure divergence to functional similarity. We can use the new functional distance to create a statistical framework assessing accuracy of functional annotation from homology modeling. Furthermore, genes with small functional distance can be clustered to yield seed sets for *de-novo* identification of transcription factor binding sites (see **C.4.1)** and expression modules[71]. – Finally, the same kernel-based methods can be generalized to measure distances on other biologically inspired graphs such as protein-protein interactions and genetic networks.

### *B.3.2 Putting the structure-function relationship in phylogenetic and transcriptional context*

Relating structural homology to function has also been complicated by numerous examples of folds performing many unrelated functions. This many-to-many relationship between structure and function has been linked to fundamental biological processes and characteristics such as adaptation, specialization, pleiotropy, or differential regulation.[72-74] Since protein function often depends on genomic context, defining predominant trends in the coalescent evolution of organisms and proteins may be instrumental in improving our understanding of the structure-function relationship[42].

We have recently shown that considering organismal context vastly improves our ability to infer function using structural homology modeling[69]. First, I plan to extend this work redefining function using binding profiles. (See **C.4.3**) Function will then carry information not only about the catalytic or binding potential of the structure, but also about its cellular and pathway context. Next, I plan to explore the relationship between structure and function with respect to regulatory context. Does constraint on the functional potential imposed by structure have a corollary constraint imposed on transcriptional regulation? Understanding the dynamics of divergence between structure and function [75, 76] will be aided by determining the role of regulation in influencing the dynamics of molecular evolution. Finally, we will explore the role of environment (e.g. using phylogenetic profiles) in influencing this redefined notion of structure function relationship. For example, we can explore whether the separation of orthologs by function among different phylogenetic profiles[77] is complemented by the co-evolution of prevalent transcriptional control models.

### *Conclusions*

The goal of this research proposal is to explore biology at different levels of granularity. The proposed work is going to allow us to build a multi-dimensional model of evolution. Using tools, techniques and findings outlined above, we can begin exploring evolution from a series of vantage points. Using increased understanding of the evolution of coding regions coupled with tools that identify functional positions in upstream regions, we can begin hypothesis-driven investigations into the correlated evolution of gene families and control sequences. The ability to trace the correlated mutations in the coding and upstream regions that change both the biochemical function and expression pattern of the protein could uncover novel evolutionary mechanisms. The goal is to build a comprehensive model describing the inter-relationships between mutational dynamics of binding sites, duplication and divergence of open reading frames, and appearance of novel pathways. The results from this research could be instrumental in furthering our understanding of how genes function, or new pathways evolve. The forthcoming research promises insight to some of the most interesting and urgent problems in biology, such as the origins of speciation, transcriptional mechanisms, and complexity.

## C Research Proposal in Detail: Preliminary Evidence and Future Directions

I describe the proposed projects from a top-down perspective: first the global, long-term goals, afterwards the progress and proposals for the constitute parts. In this first section, I describe proposed research into evolution of functional elements in upstream promoter regions. Mutations in functional sequences inside the upstream regions have been shown to be responsible for changes in protein function or even speciation[1, 2, 78]. Since mutations in the upstream region change the transcriptional control of the open reading frames, successful models will use the significant knowledge of the evolution of coding sequences to guide this research. (See **B.1**)

One challenge of studying control mechanisms and their evolution is the ability to determine positions in the upstream region that are functional with a low false positive rate. (See **B.1.1**) Error in defining functional positions in upstream promoter regions, like incorrect sequence alignment for studying evolution of open reading frames, introduces noise in any evolutionary model. To solve this problem, we will utilize the GibTig algorithm which has been shown to minimize the false positive rate in TFBS identification (See **C.3**). Another difficulty is determining the phenotypic impact of mutations (See **B.1**). We will address this problem from three directions. First, I have been studying the origin and impact of constraint on the dynamics of molecular evolution (See **C.1.0**). I plan to use these results to guide investigations into the relationship between selection on the open reading frame and the upstream regulatory region. Second, I have been studying the relationship between sequence, structure and function. This work will be extended to relate function and regulation (See **C.2**) more concretely and quantitatively. Finally, I have been investigating higher-order regulatory constructs. This will be used to formulate the interaction network between functional sites in upstream regions and further elucidate the inter-relationships between regulatory mechanisms.(See **C.4**)

## *C.1 Evolution of Transcriptional Control Regions*

The nature of the link between organismal and molecular evolution remains a fundamental question in evolutionary biology. The relationship between evolution of the gene and the organism can be characterized by the change in fitness suffered by the organism from a mutation in the gene or its upstream region.[79, 80] While the formulation is relatively simple, quantifying this effect is often difficult. For example, benefit from mutation in one environment may have an opposite effect in another.[81, 82] Driven by the recent availability of sequenced genomes along with high-throughput functional assays, researchers observed a number of significant correlations between intrinsically functional characteristics of gene sequences such as essentiality[83-85], number of protein-interaction partners[86], or expression level[87], and the strength of purifying selection. However, the relative importance of each characteristic has been a subject of vigorous debate.[88-90] Furthermore, the observed correlations have had limited impact on our understanding of dynamics behind gene duplication and divergence. Finally, there has been no study of the relationship between characteristics of the upstream region controlling the gene and the constraint put on the evolution of the downstream gene.

### ***C.1.0 Background: The origin and impact of constraint in molecular evolution***

In this section, I describe the recently observed inter-relationships between functional constraint imposed by the organism, purifying selection and dynamics of duplication. While not necessarily essential in rich media conditions, genes from families with paralogs performing essential functions, are under stronger purifying selection than genes in families with no paralogs fulfilling essential functions (*Table 1).* Thus, analogously to the approach taken with individual genes[83-85], we can use existence of an essential paralog as a marker for *families* under strong selection. For brevity, we will call families under strong selection *Exigent (demanding)* and families with no essential genes *Peregrine (wandering)*. We argue that selection is a function of gene family membership because the observed differences are independent of other functional characteristics known to correlate with purifying selection such as CAI and protein abundance.(Data not shown) The uniformity of selection on all genes inside families, independent of their functional characteristics, suggests that gene families may be the appropriate context to use when investigating the relationship between functional constraint, selection and molecular evolution.

*Table1: Comparison of SNP density. Both essential and non-essential genes in families containing ~~at~~ least one essential gene show stronger purifying selection than genes in families with no paralogs fulfilling essential functions.*

| | *Essential Genes* | *Nonessential Genes* | P-Val |
|---|---|---|---|
| All genes including singletons | .01567 | .02158 | 1e-20 |
| | *Families containing essential genes* | *Families containing no essential genes* | |
| Only nonessential genes in families | .012 | .027 | 1e-40 |

Comparing the sequence distributions between the two types of families, I find that both are skewed towards higher sequence divergence (lower identity) for paralogs in *exigent* families. (*Fig 1a,b*) For example, in *S. Cerevisiae,* the average amino-acid sequence identity between paralogs in *exigent* families is 40% while the average in *peregrine* families is 73%. We can observe qualitatively similar results when considering paralogous families in *E. Coli* and *C. Elegans* genomes. Thus, we show that paralogs from *exigent* families are farther separated in sequence space. Observations of slower divergence rate (*Table 1*) and greater separation in sequence space between paralogs (*Fig 1*) in *exigent* families are seemingly contrary to each other. This paradox can be resolved if we hypothesize that paralogs from *exigent* families survive longer before pseudogenization. A skewed distribution of pairs with more synonymous substitutions in *exigent* families is consistent with longer average survival time of both duplicates before non-functionalization.[38, 43] If we assume that synonymous substitution rate has been approximately uniform, duplicates from *exigent* families survive on average 3x longer (mean Ks = 3.25) than duplicates from *peregrine* families (mean Ks =1.15; P value of difference <1e-40). Another way to assess preference for a duplication pathway is to measure the non-functionalization rate. I observe that the fraction of pseudogenes[91] that can be attributed to duplication events in *exigent* families is ~7x lower than would be expected at random e.g. if we assume an equal probability of pseudogenization, proportional to the number of genes in each family type. (*Table 2*).

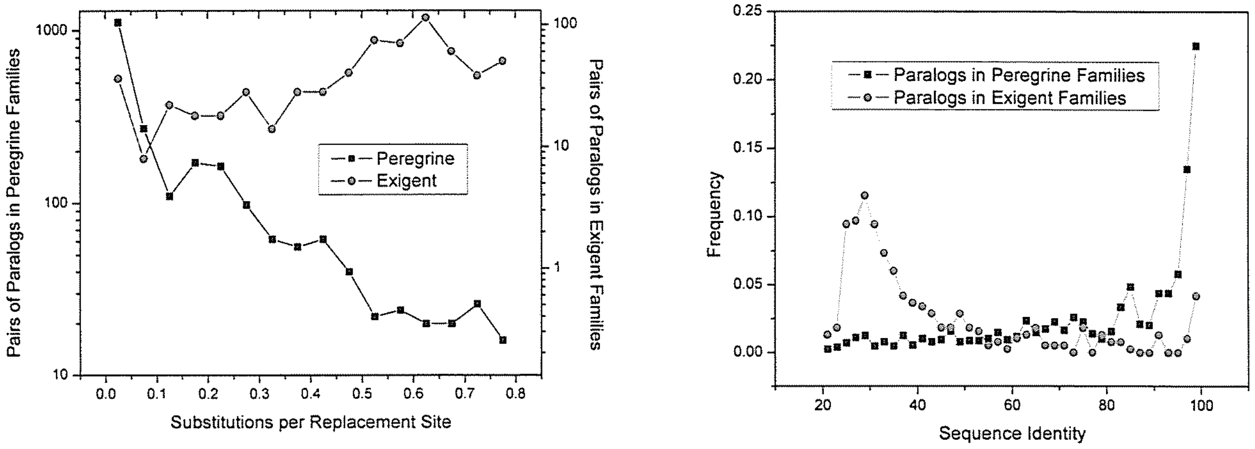


*Fig 1a. The distribution of the substitutions per replacement site (Ka) for pairs of paralogs in peregrine (squares) and exigent (circles) families. Mean Ka for peregrine~.14, for exigent~.50 $P_{val}$<1e-50. b. Distribution of sequence identity defined using BLAST. The distribution is drawn for pairs of paralogs in peregrine families (squares) and exigent families (circles). The mean sequence identity for paralogs in peregrine families is 73% while the mean for pairs of paralogs in exigent families is 40%. (Pval<1e-40).*

*Table 2. The number of genes and pseudogenes in both types of families (exigent and peregrine). The ratio of pseudogenes/genes is 7x less in exigent families with a probability P <1e-20 that this occurred by chance. The same calculation can be done by comparing the ratio of genes in the two types of families to that of pseudogenes.*

| | Genes | Pseudogenes | Pseudogene/Gene Ratio |
|---|---|---|---|
| Exigent | 278 | 4 | .014 |
| Peregrine | 656 | 62 | .095 |
| *Exigent/Peregrine* Ratio | .42 | .06 | P-val<1e-20 |

Here, I present evidence that gene family membership and selection, probably through functional constraints imposed by the organism, play a pivotal role in the progression of molecular evolution by influencing the fate of duplication events and subsequent divergence of paralogs. In fact, essentiality may represent constraints complementary to those imposed by membership in a gene family. Observed homogeneity in purifying selection on all genes in families of paralogs could be insightful of the relationship between function and evolutionary pressure. Recently duplicated paralogs are likely to perform similar functions and their divergence in functional space is limited[42]. In turn, limitation on functional divergence may impose corollary constraints on the variability of strength of purifying selection. This line of reasoning leads to the conclusion that dynamics of molecular evolution are inherently contextual[12, 43]. Since gene families not only share a constrained set of functions[42] but also a characteristic strength of purifying selection, any gene's ability to fix after the duplication event, diverge, neo-functionalize or sub-functionalize is determined by membership of that gene in a family, and in turn that family's function in the organism[43].

*Thus, the functional constraint on the open reading frames is partly determined by their regulation, is there a relationship between functional elements in the upstream region and the constraint on the gene? Furthermore, if functional constraint on gene families has such a profound impact on the dynamics of evolution of the genes, what impact does it have for the evolution of the upstream regions controlling the expression of these genes? Finally, is there a difference between evolution of regulation for paralogous and orthologous genes? We propose the following approaches to tackling the above questions:*

### *C.1.1 Proposed research: Modeling Evolution of Upstream Regions*

There are two major challenges in modeling evolutionary mechanisms in upstream regions. First, we have to determine the evolutionary precursors or direct descendants to the promoter sequences. Second, we have to model the likely sequence of mutational events leading to the observed set of differences between closely related upstream regions. The first problem: identifying closely related sequences is easier when considering evolution of orthologous upstream regions. For example, sequence alignment of ORFs across closely related species e.g. the seven species of Yeast[92] will yield a set of upstream regions likely related directly through common descent. On the other hand, identifying homology in upstream regions of paralogous open reading frames may prove more difficult. (See **C.1.1.2**)

Thus, the first part of this project will conglomerate sets of open reading frames and their upstream regions that are likely related through evolution. These upstream regions do not have to be aligned, instead, the GibTig algorithm, will be used to define functional positions. (See **C.3**). Once we know which positions are functional, we can subject the functional elements to standard evolutionary analysis. Next, we propose to divide the evolutionary process into discrete mutational events. This division will enable precise modeling of evolutionary paths based on parsimony, distance or maximum likelihood.

Initially, we would propose the following set of discrete evolutionary events:

1. Addition of another copy of the functional element
2. Deletion of a copy of the functional element
3. The functional element may change in sequence e.g. point mutation
4. The element may change position with respect to the upstream binding site
5. Since binding sites may be read 3'-5', 5'-3, complement or reverse complement, the binding may transmute from one variant into another.

After defining the basic model, we have to consider the problems of evolution of orthologous regions and paralogous regions separately.

*C.1.1.1 Proposed Research: Correlated Evolution of ORFs and TFBS in Orthologous Upstream Regions*

As mentioned above (See **C.1.1**), we start by identifying sets of orthologous regions related via common descent. This can easily be done using sequence alignment in conjunction with synteny analysis.[93] Next, we can use the sets of orthologous upstream regions to define the TFBS using the GibTig algorithm (See **C.4.1**). After defining functional positions, we can use the separation of evolutionary processes into discrete events outlined above in **C.1.1** to calculate the relative likelihood of each mutation. Using a maximum likelihood approach[94] (in collaboration with S. Sunyaev from Harvard) and the standard tree of the Yeast species [92], we can infer the most parsimonious scenario of evolution for each set of elements in the upstream region[95-97].

We will try to relate our discoveries about the evolution of the upstream regions to the better-studied evolution of open reading frames. (See **C.1.0**) The evolution of open reading frames can be characterized by evolutionary pressure e.g. non-synonymous mutations - Ka, synonymous mutations-Ks and the ratio-Ka/Ks) (See **C.1.0**). Open reading frames under higher evolutionary pressure are likely to diverge less e.g. Ka/Ks is smaller. With the model of likely evolutionary events and changes in the upstream region in hand, we plan to correlate evolutionary pressure on the open reading frames and the evolution of the upstream region.

Specifically, for every gene in the S. Cerevisiae genome, we plan to compile the average Ka/Ks ratio between orthologous open reading frames (in the six sequences Yeast genomes). We proceed to identify the evolutionary model describing the most parsimonious series of events explaining the differences between the upstream regions. Using these two datasets, we can start asking quantitative questions about the relationship between pressure imposed on the gene and the evolutionary model governing mutation of its upstream region. For example, we would like to investigate whether the functional elements in the upstream regions of proteins under high selection e.g. ribosomal proteins are restricted to only short-term changes e.g. point mutations as opposed to rearrangements such as additions or deletion of functional elements (See **C.1.1**).

### *C.1.1.2 Proposed Research : Evolution of functional elements in paralogous upstream regions*

Research into the evolution of upstream regions related by paralogy is more difficult than comparing upstream regions related by orthology (See **C.1.1.1**). Unlike orthologous regions, paralogous upstream regions could emerge as result of more complex sequence rearrangement events e.g. transpositions, whole-genome duplications etc. Some of these scenarios yield parallel duplication of the upstream region along with the open reading frame, while others do not. Another problem is the determination of time since the last common ancestor of the duplicated pair. Unlike orthologous upstream regions which could be tied to the speciation event, duplications and genome rearrangements have no fixed divergence times. This problem is complicated further by the variable mutation rate across the genome[98]. We will attempt to overcome these problems by leveraging our knowledge of the consensus-based graph-theoretical approach analogous to the one used to measure divergence and diffusion of open reading frames. (See **C.1.0**)

Briefly, we first divide the genome into families of genes (See **C.1.0**) by building a sequence comparison graph. Finding strongly connected components as in (**C.1.0 or C.2.1**) will enable identification of closely related gene families. Using this graph, we are in a position to start evaluating which upstream regions are related through common ancestry e.g. via transposition or whole-genome duplication events. First, we will try to align every pair of upstream regions in the gene family using approximate alignment algorithms commonly used for sequences with low identity e.g. LAGAN[99, 100]. If whole-upstream region alignment proves fruitless, we will use the GibTig algorithm (See **C.3.1**) to search for common motifs. Furthermore, we can leverage our TFBS map from **C.4.1** as additional evidence for positions of functional sites.

Once we have the positions and functional sites, we would like to evaluate whether upstream regions are also divided into broad evolutionary categories akin to the open reading frames (See **C.1.0**). In **C.1.0**, we note that essential genes can be used as markers of paralogous families under strong evolutionary pressure. We plan to use these results to compare the conservation of functional elements in upstream regions of paralogous families under strong versus weak evolutionary pressure.(See **B.1**) Furthermore, we can then evaluate the time-frame needed to accomplish specific changes in the functional elements of the upstream regions by comparing the divergence of the functional elements to the divergence of the rest of the upstream promoter region. We saw in **C.1.0** that membership in a gene family may influence the speed of divergence of the gene. We can explore whether functional elements e.g. specific TFBS also influence speed of divergence. Finally, we plan to correlate the divergence of the upstream regions to the functional divergence of the open reading frames. (See **C.4.2**)

## *C.2 Extending our understanding of Structure-Function Relationship*

In this section, we present projects that further our understanding of the structure-function relationship. These studies are mostly direct extensions from **C.2.1** and **C.2.2** and [75, 76, 101, 102]. There is one other section in the proposal that deals with structure-function relationships: **C.4.2**. There, function is redefined using transcription factor binding sites from whole-genome TFBS identification in **D.4.1**. The majority of the projects in this section capitalize on the highly successful graph-theoretic approach outlined in **C.2.1** and [101].

### *C.2.1 Background: PDUG: Protein Domain Universe Graph: Developing a unified view of sequence-structure-function space*

Evolution is, at its core, a science of comparison. In order to study evolution, we need to create a computational framework to represent our current body of knowledge. We chose to approach the problem from a graph-theoretic prospective where nodes are domains and edges are comparison measures. Aside from providing a unified framework, evolutionary graphs like these provide a means for organization of the diverse glut of experimental data that has become the cornerstone of bioinformatics research. We have been very successful in applying this framework to both theoretical investigations of protein evolution [103-105] and applied investigations of structure-function relationship[75, 76, 106]. An example of a large cluster of protein domains connected by structural homology is shown in **Fig 2**.

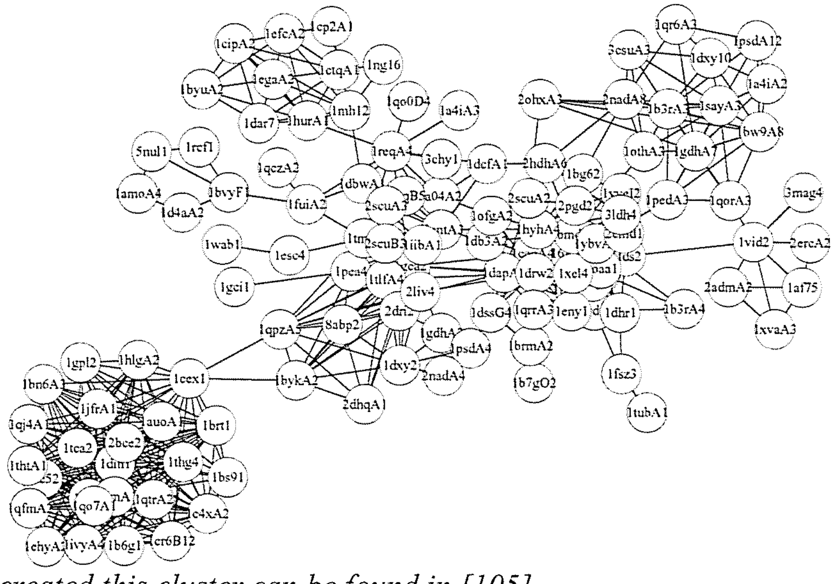


*Fig 2 An example of large cluster of TIM-barrel fold protein domains. Protein domains whose DALI[107] similarity Z-score is greater than $Z_{min} = 9$ are connected by lines. An in-depth discussion of how we created this cluster can be found in [105].*

### *C.2.2 Background: Correlated Divergence of Structure and Function*

Using structural similarity based clustering of protein domains: Protein Domain Universe Graph (PDUG), and a hierarchical system of functional annotation: Gene Ontology (GO) as two evolutionary lenses, we find that each structural cluster (domain fold) can be characterized by a unique distribution of functions[75]. These functional distributions are like "functional fingerprints" specific to a cluster and vary from cluster to cluster. Furthermore, as structural similarity threshold for domain clustering in PDUG is relaxed we observe an influx of earlier-diverged domains into clusters. These domains join clusters without destroying the functional fingerprint. The uniqueness of the functional fingerprint is not destroyed (does not become random), but is complemented with similar functions. This preservation of unique functional fingerprints through evolutionary dynamics further highlights the close relationship between structure and function. These results can be understood in light of a divergent evolution scenario that posits correlated divergence of structural and functional traits in protein domains from one or few progenitors.

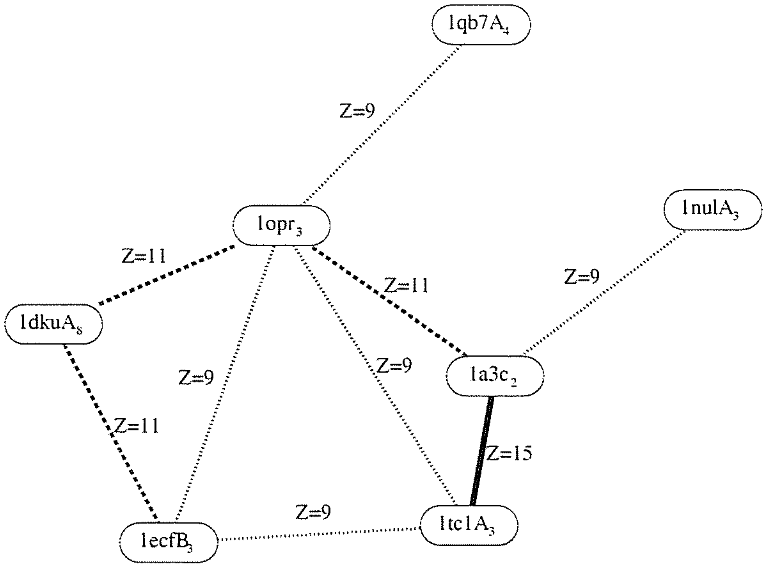


**Fig 3.** *A schematic picture of how protein domains diffuse into structural clusters and define them at different thresholds. Nodes represent protein domains, links represent structural similarity at above threshold. Thick line represents domains connected at $Z_{min}$ = 15, dotted lines represent domains connected at $Z_{min}$ = 11 and thin lines represent domains connected at $Z_{min}$=9.*

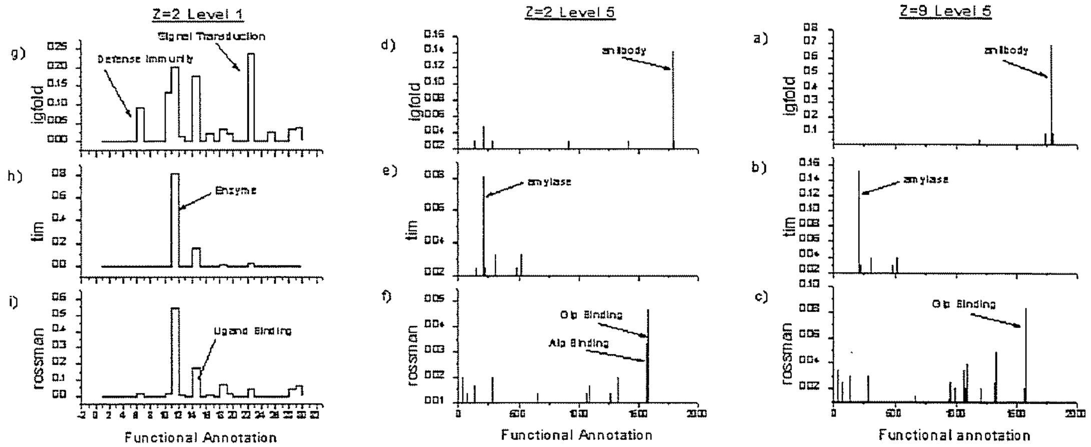


**Fig 4.** *Examples of some functional fingerprints of folds. X-axis is the functional annotation category. Y-axis is the normalized number of proteins annotated with that function. (a-c) Functional annotation at the fifth level of GO ontology for $Z_{min}$ = 9. Notably, each cluster has its own, distinct functional fingerprint that is observably different than those of other clusters. (d-f) At the fifth level of annotation after proteins joined their ancestral clusters, for $Z_{min}$= 2. The fingerprints are more diverse, however still differ significantly from each other. (g-i) At the first level of annotation with $Z_{min}$ = 2, the fingerprints overlap significantly, and hard to distinguish one from the other*

### *C.2.3 Proposed Research: Co-Evolution of enzymes and metabolic networks*

In our previous, published work[76, 108], we introduced a general framework for investigating relationships between protein structure and function. In those studies we chose to define function through evidence of biochemical activity annotated using the gene ontology (GO, [70]) system. In general, function is a poorly defined concept that can be defined in various ways. For example, enzymatic function can be defined by inclusion and role in specific metabolic pathways. KEGG (Kyoto Encyclopedia of Genes and Genomes) is one such pathway annotation ontology[109]. We can use previously developed techniques (See **C.2.1** and **C.2.2**) and the graph-theoretic frameworks of PDUG and KEGG to probe for the existence of a structure-pathway relationship. This project would be based on the hypothesis that divergence of structure proceeded by a parallel incorporation into an ever-widening enzymatic pathway. Thornton and co-workers found anectodal evidence of this phenomenon, but stopped short of investigating its prevalence[110]. I would like to develop a more systematic understanding of the correlated evolution of protein structure and enzymatic pathways.

*C.2.3.1 Proposed Research: Divergence of pathways using a graph theoretic approach*

We start with two graphs defined by nodes representing protein domains. In one graph, the edges are path-lengths on KEGG pathway maps between domains performing enzymatic reactions. This graph (Metabolic Expanse Graph: MEG) is defined as G=[*v*,*e*] where *v*=(set of vertices representing non-redundant protein domains as defined by DALI[76, 101], *e*=(set of edges weighted by distance between two domains on the KEGG map). In the other graph, we have the same definition for nodes, but weigh the edges by the structural similarity between the domains, Z score (See **C.2.1**).

The objective is to find and quantify both the static and dynamic relationship between these graphs. We can use cutoffs to define strongly connected components in both MEG and PDUG (See **C.2.1**[101] and **C.1.0**). The cutoffs which we will call $M_c$ for the MEG graph and $Z_c$ for the PDUG are used to vary the average similarity between nodes and represent functional divergence in MEG and structural divergence in PDUG.

We go on to compare the strongly connected components between PDUG and MEG using the Jaccard distance outlined in [76]. We aim to find the cutoff values that produce the best overlap between the two clustered graphs. This is equivalent to asking: How far diverged in structure are protein domains that take part in common metabolic pathways? Intuitively we would expect the relationship between structure and pathway inclusion to hold for very small distances. For instance, we would expect domains working on closely matching metabolites to have similar active sites and mechanisms, thus sharing structural elements. However, if metabolites are significantly different in their chemical structure the structural similarity should decrease.

*C.2.3.2 Proposed research: Measuring functional distance using similarity measures between substrates (with Prof. Minoru Kanehisa)*

If we are successful in finding the correlating divergence levels using KEGG and PDUG, we plan to expand this research using the newly developed, more refined distance metrics measuring chemical distance between metabolites. Dr. Kanehisa and co-workers have developed a graph-theoretic framework for comparing chemical moieties. [111-114] Using this distance metric, we could define the Metabolic

Moiety Similarity Graph (MMSG). MMSG would have nodes representing domains and edges would be weighed based on the distance between the carbohydrate chains e.g. substrates for the two domains. Using the same framework as above (See **C.2.3.1**), we could probe for the correlated evolution of protein — structures and carbohydrate substrates.

Results from this work would be very helpful in identifying the primary determinants of protein domain structures responsible for substrate selection. Furthermore, by following both the expansion of the structural repertoire and the metabolic substrate repertoire, we will be able to understand how and in what order evolution "discovered" the multitude of folds and superfamilies needed to carry out the necessary metabolic functions. A deeper understanding of the differences in domains that enable manipulation of different chemical moieties could further our understanding of the residues needed for function versus those needed for folding.

### *C.2.4 Proposed Research: Measuring distance in function space using kernel mapping on GO (with Prof. Gilad Lerman from University of Minnesota)*

Due to the complex nature of the structure-function relationship and the non-linear progression of protein evolution [42, 65, 115], relationships between proteins are currently assessed qualitatively on the basis of many shared characteristics such as sequence, structure and function. This annotation schema yields a hierarchical organization where proteins sharing sequence are closest, followed by those sharing function and then structure in descending degree of proximity[110, 116]. While quantitative comparison measures exist for sequence and structure, distance between protein functions is yet to be developed.

In [69] we show that functional distances can, in principle, be quantified using the GO annotation system. However, one problem is that approach yields a low level of coverage. We aim to generalize that notion of function distance [69] to a more exact, kernel-based method. Having an accurate measure of distance in function space is integral to creating hypothetically co-regulated sets of genes for our de-novo TFBS mapping (See **C.4.1**), for identifying the relationship between evolution of functional elements and function, (See **C.1.1**) for relating CRMs to functions (See **C.4.2**), for quantifying sequence-structure-function relationships and for assessing accuracy of homology modeling methods.

Since GO[70] is hierarchical, the top level of hierarchy is, by design, less precise than the bottom level. This can be seen easily by considering that there are only twenty possible annotations at the top, and more than two thousand on the fifth level of the Gene Ontology. The basic idea behind building an — appropriate kernel is that edges at the bottom of the ontology shared by few domains will be assigned small local distances or equivalently high values of local similarities. Alternatively, edges appearing at a top of a tree and have none or few siblings will be assigned large local distances or equivalently small values of local similarities.

We propose various techniques for embedding of the Gene Ontology into Euclidean spaces. The first two techniques share a similar idea. They assume a metric space $M$ represented by the graph, where only local distances are specified and an unknown embedding $P$ from $M$ to a finite dimensional space $E$. The actual mapping $P$ is not needed for recovering the Euclidean distances, but only the kernel $\mathbf{K}(x,y) = \langle P(x),P(y)\rangle$. The Euclidean distance between point x and y in the embedded space is $d^2(x,y) = \mathbf{K}(x,x)+\mathbf{K}(y,y)-2\mathbf{K}(x,y)$. We can propose three embedding methods that differ in their choice of kernel. The first method applies a variant of the Laplacian embedding (equivalently the diffusion embedding) as described in [117-119]. The second method applies a variant of locally linear embedding[120]. The last method estimates directly the shortest path distance. It can also be approximated by the isomap kernel[121].

### *C.2.5 Proposed Research: An interesting example: Evolution of vancomycin resistance.*

The resistance to vancomycin is due to a cluster of five genes. [122] Enterococci gain resistance to vancomycin by the accumulation of the *vanHAX* genes. The *vanH* enzyme, encoded by the gene of the same name, is involved in the creation of a new pathway of enzymes that produce D-lactate from pyruvate. The enzyme *vanA* adds a D-lactate moiety to the end of the peptide cross-bridge, rather than another D-alanine. The peptide strand therefore ends with a D-Ala-D-Lac, rather than the usual D-Ala-D-Ala. Finally, *vanX* hydrolyzes the D-Ala-D-Ala moiety so that the new peptide moiety is more common. Although the change in the last peptide from Ala to Lac does not have any effect upon the quality of the peptidoglycan layer or its ability to cross-link the glycan strands, it does lower the binding affinity of the vancomycin to its target by 1000-fold. Vancomycin, therefore, has a tougher time in binding to its substrate and is rendered ineffective. [123]

Recently, there has been an emergence of resistance to vancomycin and other antibacterial drugs in enterococci bacteria. [122, 124] This poses a major health problem not only due to an increased threat from the proliferation of disease, but also because of a possibility that enterococci will spread the resistance to other bacteria. The main mechanism of vancomycin action is due to blocking of bacterial cell wall formation. We would capitalize on the recent functional evolution of the D-ala-D-ala ligase as a model for studying functional evolution in general. We would like to address the following questions. How many structures in general are implicated in acid-D-amino ligation? What is the most likely metabolic pathway where the original D-ala-D-ala ligase was involved? How did that protein evolve into a D-ala-D-lac ligase? What are the most likely ancestor functions for the rest of the genes in the vancomycin resistance cluster? Can we generalize our findings to predict future resistance from just knowing the pressure and can we predict the most likely "new" metabolic network that will evolve?

Preliminary studies on structural neighbors of vanA (1iow) suggest that the "new" function of adding a D-Lactate moiety is not too far diverged from the functions of its structural and sequence neighbors. Almost half of the sequence neighbors are implicated in D-ala-D-ala ligation while, the other half is implicated in ligating D-ala-D-lac moieties. Furthermore, there are five known non-redundant domain structures that are topologically similar to the template of vanA. The functional fingerprints of these templates (See **C.2.2**) are similar to vanA.

This system represents a unique model for study because the evolution of the resistance network is very recent and the divergence from ancestor proteins is relatively small. In order to assess the evolutionary path, we will use our understanding of the dynamics of divergence of structure and function (See **C.2.1** and [75, 76]) to find the closest homologous proteins for the ones involved in vancomycin resistance. We can use value decomposition methods to assess which characteristics (evolutionary pressures) are most likely responsible for the novel function of the VanA gene as well as the conglomeration of the five genes into a pathway responsible for vancomycin resistance.

### *C.2.6 Conclusion and Outlook:*

Structure-function relationships are inherently contextual and must be understood in terms of evolutionary pressures acting on the divergence of the domain or protein. PDUG is a powerful tool for studying the correlated evolution of structure and function (See **C.2.1**). Furthermore, the simple graph representation may be extended to a hypergraph [69] to include other pressures e.g. phylogenetic or metabolic (See **C.2.3**). This approach has already yielded a number of insights into the relationship between structure and function and the origins of functional fingerprints of folds (See **C.2.2**). This new paradigm of looking at protein domain universe can be used in many ways, from increasing the sensitivity and specificity of homology based annotation [76] to understanding the range of functions possible for a given structure. (See **C.2.2**) Furthermore, it significantly simplifies formulation of problems and investigations into evolutionary dynamics governing structural divergence e.g. pressures needed to evolve the domain from one function to another (See **C.2.5**). Finally, our proposed quantification of distance in function space will provide an invaluable tool applicable to many areas of computational biology research( See **C.2.4**) We believe that our approach along with parallel research outlined in this proposal will, for the first time provide a comprehensive and quantitative theory of structure-function relationship in proteins, in evolutionary prospective.

## *C.3* High-performance Computational Approaches to Transcription Factor Binding Site Discovery

The DNA sequence upstream from the gene consists of *a transcription start site and two regulatory regions: a core promoter* that is located within about 40 bp of the start site and an *"upstream" regulatory region,* extending over as many as 1kbp farther upstream, that transduces core promoter binding events into stable interactions that alter gene transcription[125-127]. Additional sites involved in enhancing and silencing expression may extend up to 10kb upstream of the transcription start site, especially in genomes of more complex eukaryotic species such as Drosophila and H.Sapiens[128-130].

Computational methods for recognizing binding sites, have performance limits that seriously impair their potential. The obstacles to identifying functional sites with high fidelity are both biological and algorithmic in nature. For example:

(i) Multiple transcription factors often play a key role in the regulation of a single gene. Thus an algorithm has to choose between many quantitatively "equally good" solutions.

(ii) A single factor may display significant variability in both the width and composition of the binding sequence, while the makeup of the allowable variations is not well understood.

(iii) TFBS may be located quite far from the coding region they control, either upstream or downstream or in the introns.

(iv) Other highly conserved DNA sequences such as transcription start sites, transposable elements, LINEs, SINEs, RNAi and tandem repeat elements serve to obscure the identification of TFBS due to their widespread presence and conservation.

When trying to identify binding positions in the upstream regions of functionally related genes, the input is the set of hypothetically co-regulated genes[131]. The necessary data often comes from chromatin immunoprecipitation (ChIP)[13-15]. Most recently, ChIP has been combined with microarray based expression analysis (ChIP-Chip) to associate some 206 yeast TFs to the genes they modulate[13, 15]. However, Young and colleagues report statistically significant binding for just over 50% of the genes in the genome. At the same time, they were able to identify binding specificity for only 65 of the 206 TFs[13].

The relatively poor coverage for both the identification of binding positions and mapping between TFs and genes, underlines the need for a more sophisticated approach at computational identification of functional sites in upstream regions.

We propose an algorithm that solves many of the issues outlined above. First, using functional comparison methods (See **C.2.4**) and gene family identification (See **C.1.0**), we can use computational approaches instead of CHIP-Chip to identify potentially co-regulated sets of genes thus improving coverage. In the tests we have run, the GibTig algorithm performs significantly better than competing approaches. In fact the level of false-positives is low enough that we hope to create a *de-novo* TFBS map for the whole Yeast genome using this algorithm (See **C.4.1**).

### *C.3.1 Background: Description of the GibTig algorithm*

Gibbs sampling and other algorithmic approaches to motif detection follow two strategies to discover repeating sequence patterns in DNA: enumeration[132] and probabilistic sequence modeling. Enumeration strategies rely on sequence counting to find over-represented *w-mers[133]*. Model-based methods represent a set of *w-mers* as a position weight matrix(PWM). The PWM describes nucleotide base multinomial probabilities for each position. The statistical significance of the PWM is evaluated with respect to background frequencies of nucleotides in sequences outside the ones used to build the PWM. The algorithm tries to maximize the difference between the PWM and the background model using Monte Carlo or Gibbs sampling[131, 134, 135].

In order to improve on the false positive rate currently plaguing transcription factor identification algorithms[136], we developed a different algorithm called GibTigs. The inspiration for the algorithm comes from the hypothesis that binding between the TF and the binding site imposes spatial constraints related to the polymerase assembly and looping. Thus, we impose an additional requirement of conservation in position when searching for functional TFBS by Gibbs sampling. This enables identification of "peaks" (GigHits) that correspond to functional TFBS (Fig 5). Briefly, we consolidate sampling runs analogously to a shotgun sequencing approach. At every position, we calculate the distribution of frequencies where sampling converged on that nucleotide position. We then build consensus matrices from the conglomeration of runs (Fig 5).

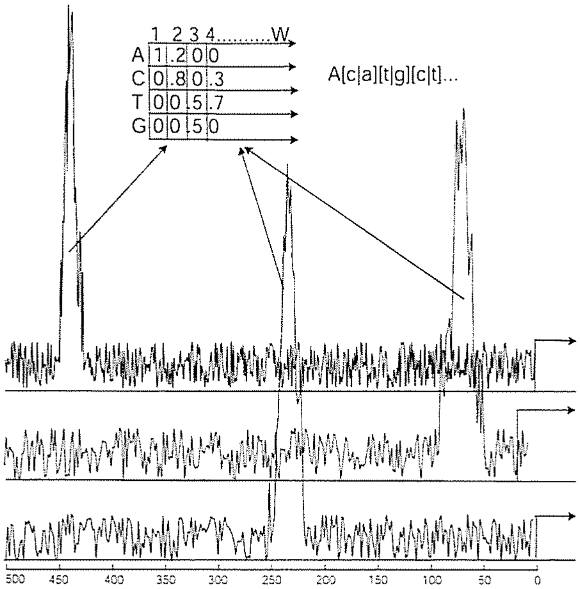


**Fig 5.** *A cartoon illustration of 3 GibTigs and accompanying GibHits. Each line represents an upstream region from a gene. The transcription start site is represented with an arrow. The Y-axis over the upsteram regions represent the counts of the number of times gibbs sampling converged on that nucleotide position. The peaks are the GibHits, the baseline noise are the local minima where the Gibbs sampler converged. We conglomerate all the GibHits into a single PWM*

### *C.3.2 Background: Predictions in Yeast: Comparison of GibTigs with existing algorithms.*

In this section, we outline a comparison between our current implementation of GibTigs and the state-of-the-art TFBS identification algorithms using a set of upstream regions shown to bind STE12 from CHIP-Chip data[15]. We quantify false positive results by comparing to a random set of upstream regions from the yeast genome. We report the results of GibTigs as compared to the most commonly used algorithm: BioPropsector by Lawrence, Liu et. al. [12, 131]. While we report our results with respect to BioProspector, we found that in accordance to previously published comparisons[12, 136, 137] BioProspector performs superior to most other TFBS identification algorithms (data not shown).

To evaluate the performance of both algorithms with minimal bias, we require a scoring scheme capable of quantifying the "right answer." We use an algorithm originally developed by Pietrokovski[138] in the context of aligning amino acid motifs, and later adapted to nucleotide PWMs by Hughes et al.[20]. We use the PWMs generated from GibHits and BioProspector runs and compare those to the experimentally defined STE12 and TEC1 binding profiles. Our scoring scheme will identify the percentage of computationally defined motifs comparable to the experimentally defined STE12 PWM[139, 140].

Using the 49 intergenic regions shown to bind to STE12 by Chip-Chip we increase the percentage of reported sites matching the STE12 PWM from 68% using Bioprospector to 94% using the GibTig algorithm. Furthermore, we decrease the probability of getting a spurious PWM that matches STE12 from random upstream sequences from 12% to 2%. We also compared our results to the TEC1 PWM TEC1. TEC1 is a protein known to be involved in regulation of some of the same genes. Using intergenic regions as the input set, the number of reported GibHits with similarity to the TEC1 PWM falls from 7% (repor~~ted~~ by Bioprospector) to 0% and the number expected at random using a sampling of intergenic regions as input is also 0%. Thus, using Gibtigs we can significantly increase the number of true positives and decrease the number of false positives in this input set. (Fig 6 A,B)

However, we glean the results most illuminating of the improvements from the GibTig method when we disregard the transcription start sites of the upstream ORF. In this experiment, we implicitly assume that binding sites can occur in the upstream open reading frame. From this input set, the number of TEC1 matching GibTigs rises from 0% (found using the intergenic set) to 56% with 0% expected from random intergenic sequences. The reason for this becomes clear as an afterthought when we observe that almost forty percent of TEC1 sites predicted by GibTigs are in ORFs of the upstream gene.

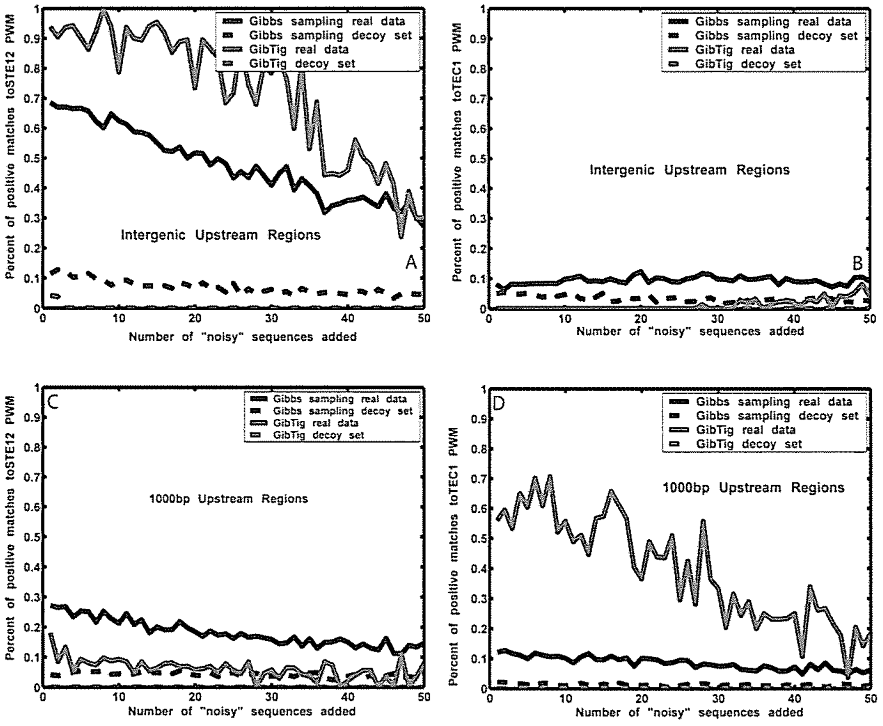


**Fig6 A** *Comparison between BioPropsector[131, 137] results and GibTig generated GibHits matching the STE12 PWM. Black lines are the percentage of BioProspector runs that returned matrices that match the STE12 PWM. The other lines are percentage of GibTigs matching STE12. Solid lines are runs done on the input set containing the 49 upstream promoters known to bind STE12 from ChIP-chip data, dashed lines are the results using 49 upstream regions drawn randomly from the yeast genome. X axes count over the addition random sequences* ***B****. The same as A above except both the BioProspector and GibTig results were compared to the Tec1 PWM.* **C,D** *The same as A except the analysis was performed using equal sized upstream regions of 1kb. The percentage of Tec1 matching GibHits rises 5 fold to 56%. The number of GibHits that generated from the random set of upstream regions decreases to 0% with respect to matching to either STE12 or TEC1 PWM.*

In summary, our GibTig algorithm reduces the chance for a false positive hit by decreasing positive matches to the PWM from random upstream sequences. At the same time, we report significantly improved percentages of true-positive hits. Perhaps, most important, is our ability to handle longer upstream regi~~on~~s. In this case, this ability yields a new biological hypothesis e.g. that TEC1 binds 40% of the time in the ORFs of the upstream genes. This hypothesis was later, independently verified by R. Young (Personal communication and [141]

### *C.3.3 Proposed Research: Future Development of GibTigs*

Unlike solutions proposed previously[136], the GibTig algorithm (See **C.3.1**) with improvements proposed in this section is explicitly targeted at whole genome transcription factor binding site (TFBS) identification and elucidation of complex regulatory networks(See **C.3.4-5**)[12, 15, 142-144]. Our philosophy is, whenever necessary, sacrifice computational speed for a dramatic increase in sensitivity and specificity. We recognize that biological experiments are en-masse more time-consuming and, computer speeds continue to increase extremely rapidly—what's computationally demanding today becomes routine tomorrow. Thus, we assume a multi-pronged approach meant to maximize applicability and ease of verification by wet-bench biology. We have shown in **C.3.2** that our GibTig approach works better than existing methods at identifying known TFBS in yeast. In later sections, we also show experimental validation of novel binding sites in human GABA receptors(See **C.3.4-5**).

While already a useable tool, the GibTig algorithm represents an ongoing research project. In this section, we present some ideas for improving the algorithm. First, the current implementation is very CPU intensive, we need to improve convergence speed and create a fully functional implementation for use in specialized and highly distributed environments. Secondly, we are interested in developing a more robust statistical framework. This framework will enable us to compare various parameter scenarios to each other with greater accuracy. (See **C.3.3.1**) Thirdly, the current algorithm is clumsy with respect to handling sets of orthologous upstream sequences. However, it is clear that phylogenetic footprinting is a growing asset in the field and should be incorporated. Finally, the model does not currently support a-priori known physical information on the various types of binding models. Recent research has shown that incorporation of physical models of protein:dna binding aids significantly in TFBS detection.[145, 146] Thus, we propose some additions and future research aimed at capitalizing on the strengths of the GibTig algorithm while trying to significantly improve performance. Due to space constraints, I only outline a subset of proposed improvements in sections **C.3.3.1-3**.

*C.3.3.1 Proposed Research: Improving the statistical framework*

The most widely known and utilized statistical model in computational biology (theKarlin-Altschul statistic) has transformed sequence alignment by creating a ubiquitous measure of significance used to compare different alignments[147] and even different algorithms[148-150]. Our aim is to create an analogous unified statistic for TFBS identification. Current algorithms use either a heuristically derived statistic based on a reshuffling the null model[12, 131, 137] or a simple maximum a posteriori (MAP) score.[136] An analytical, uniform scoring scheme would increase accuracy and provide an ability to compare predictions of TFBS from different input sets and for different parameter values e.g. different widths. (See **B.1.1**)

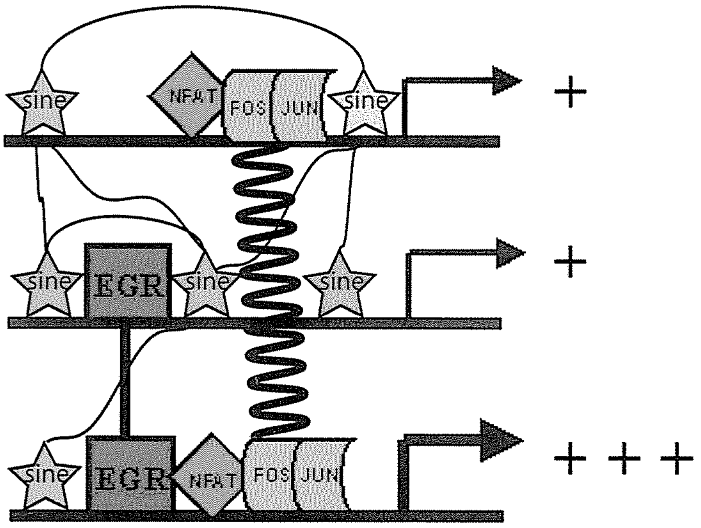
We will verify the performance of our statistical formulation using simulations with random sequences and real intergenic sequences enriched with a motif. We can compare distributions of the developed scoring functions from sequences enriched with a known TFBS to those from random sets of genes akin to the approach taken in **C.3.2**. Finally, we can use a uniform statistic to quantify improvement to the robustness of the algorithm with addition of random sequence. We hope that this methodology will set a uniform standard of significance that will enable quantitative comparison of TFBS predictions with different numbers of upstream regions, various lengths and widths.

*C.3.3.2 Proposed Research: Separating Functional from Conserved Sites Using KL Divergence*

One of the challenges in creating a robust TFBS identification algorithm is separating functional from simply conserved sequences. Conserved *w-mers* can occur for a variety of reasons such as the presence of repeat elements: SINEs and LINEs[80, 151, 152] or euchromatin binding regions[15, 153] or simply due to a random fluctuations in the presence of a random motif (data not shown or see **B.1.1**).

We empirically observed that functional TFBS conserve position in the upstream region. (See **C.3.1-2**) Furthermore, there are underlying biological reasons for why the distributions of TFBS and other conserved DNA elements across upstream regions would differ. *Fig 7* Since TFs operate by transducing-the polymerase assembly, proper functionality is sensitive to concentration effects[12, 154-157]. To ensure expression of the downstream gene at the proper rate yielding correct mRNA levels, the sum of the TF:DNA binding affinities has to be conserved[12, 154-157]. Some researchers theorize that there is an inherent tradeoff between the variability in the PWM and the compensating increase in their number[12, 154-157]. Current measures of statistical over-representation are not necessarily sensitive enough to distinguish between classes of conserved *w-mers*.

***Fig*** *7 An illustration of conservation of position for functional TFBS versus other repeat elements such as "sines". Functional TFBS are under stoichiometric constraints where binding of TF to genes results in different transcription levels. Non-functional elements can mutate with no evolutionary constraint allowing for many combinations of positions yielding similar estimation of statistical over-representation*



Every nucleotide position in each GibHit maps a set of related positions in the input regions $p_i \rightarrow \{p_i^n; i=1...L_i; n=1..N\}$ specified by repeated local convergences of the Gibbs sampler (See **C.3.1**) and used to construct the PWM from that collection of GibHits. We can calculate the Kullback-Leibler divergence on this set defined by $D=\sum f_i^n Log\left[\frac{f_i^n}{b_i^n}\right]$ where $f_i^n$ is the frequency of $p_i^n$ in the GibHit, and $b_i^n$ is the background distribution generated from random sequences with the same average length. Thus, GibHits that have a unique mapping of positions different from the background distribution of other conserved *w-mers* will have high *D* scores.

While GibHits that show no preference in position across any of the upstream regions will show low *D* scores. If the D scores prove useful, we will incorporate those into the uniform statistic proposed in **C.3.3.1**.

### *C.3.3.3 Conclusions*

The improvements to the GibTig algorithm outlined in the sections above represent a small subset of a number of ideas that have potential not only to advance the field of TFBS identification, but also significantly improve our ability to elucidate mechanisms of transcriptional regulation. There are a number of other ideas which had to be excluded due to space constraints including building multithreading support, faster convergence through conservation of momentum and a more sophisticated inclusion of orthologous upstream regions through markov-modelling of di- tri- and quad nucleotides in collaboration with S. Sunyaev[94]. One of the strengths inherent in this methodology is the development and implementation of a systematic null-model framework (see **C.3.3.1**) for quantifiable assessment of improvements introduced by each proposed modification. Through close collaborations with high-performance computational facilities and with experimental scientists, we are in a unique position to further develop and improve on an already strong computational framework for identifying functional sites in upstream promoter regions.

### ***C.3.4 Preliminary Evidence: Application of GibTigs to TFBS identification in GABA receptors***

GABA is the major neurotransmitter in the central nervous system (CNS) and its regulated release is controlled by the activity of distinct cells, referred to as GABAergic neurons[158]. The axonal processes of these neurons are situated close (opposite the synaptic cleft that separates them) to the dendritic processes or cell bodies of neighboring neurons that contain GABA receptors (GABA$_A$, the integral chloride ion channels[159]; GABA$_B$, the seven-transmembrane spanning heteromeric receptors that are G-protein coupled to various effectors such as GIRK $K^+$ channels)[160]. Over 19 different genes (GABRs) code for the pentameric GABA$_A$ receptor and there are pharmacologically distinct forms of receptors that change their expression levels either during development or disease.

First, we used the GibTig algorithm to identify conserved motifs in the intergenic regions. We classified significantly scoring GibHits as TFBS. In total, we predicted 15 TFBS distributed across 7 intergenic sequences. Out of 15 predicted TFBS, 50% were previously characterized motifs, as determined by matching the sequences to PWMs of known human transcription factors from TRANSFAC[140]. The other 50% represented TFBS hypothesized to bind to uncharacterized TF. We used EMSA to test predicted sequences for binding by proteins present in neuronal nuclear extracts. To date, we attempted validation ~~for~~ 6 of the 15 predicted TFBS, including both previously characterized and putatively novel predictions.

### *C.3.5 Preliminary Evidence: Experimental Validation of GABA Predictions (collaboration with Prof. Shelley Russek)*

As a first step to determine the applicability of GibTig analysis to identify novel TFBS in GABRs, we tested the ability of several radiolabeled oligonucleotides containing GibTig sequences to specifically bind nuclear extracts of neocortical neurons or fibroblasts. As shown in (Figs 8,9) below, using electromobility shift assays (EMSA), to our surprise 6/6 GibTigs predicted oligos examined displayed specific binding as measured by cold competition. In addition to specific binding, using fibroblast extracts for comparison, a GibTig in GABRA4 (Figure 9A) displays neural specific binding, of especial interest given that expression of GABRs is restricted to the nervous system.

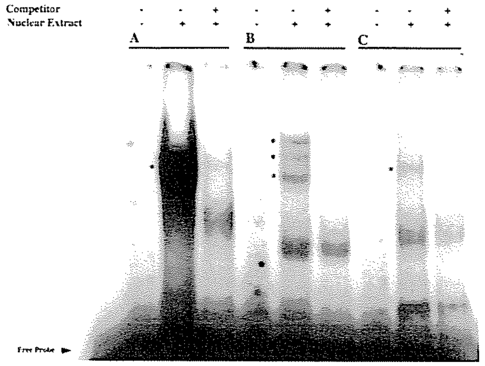


**Fig 8.** *Three putative transcription factor binding sites form DNA-protein complexes in neocortical nuclear extracts. Neocortical nuclear extracts from E18 rat embryos were incubated with three $^{32}P$ radiolabeled probes from human GABRB1, GABRD and GABRB3. Cold wildtype oligonucleotides were used to define specificity through competition. Cold oligonucleotides were added at 100-fold excess over probe. The conditions for each lane are as indicated. Specific binding complexes are shown with a (*). The probe sequences are as follows: A) GABRB1: AATACGGTCCCTACT, B) GABRD: ACTTAATTTGATTCCAT and C) GABRB3: CGTGCCGGGGCGCGGCGGA.*

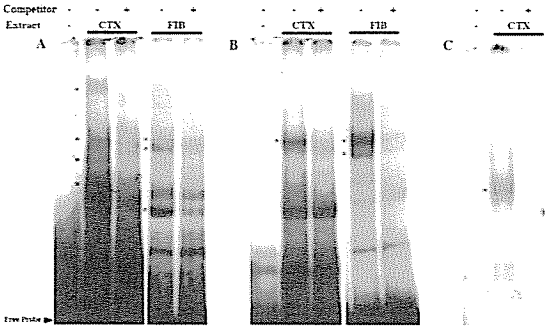


**Fig 9.** *Another three putative transcription factor binding sites form DNA-protein complexes in neocortical and fibroblast nuclear extracts. Neocortical (CTX) and fibroblast (FIB) nuclear extracts from E18 rat embryos were incubated with three $^{32}P$ radiolabeled probes from human GABRA4 and GABRD. Cold wildtype oligonucleotides were used to define specificity through competition. Cold oligonucleotides were added at 100-fold excess over probe. The conditions for each lane are as indicated. Specific binding complexes are shown with a (*). The probe sequences are as follows: A) GABRA4:AGCGCGGGCGAGTGTGAGCGCGAGTGTGCGCACGCCGCGGG, B) GABRA4: GTGCACACACACGCCCACCGCGGCTCGGG, and C) GABRD: TGACCGTAGTAGA.*

Because EMSA binding activity is measured *in vitro,* positive results in these assays do not always correlate with function in living neurons. As a first step to determine whether GibTigs are functional in living neurons, we exposed cortical cultures to decoys containing either the GibTig analyzed in Figure 10 or an unrelated regulatory sequence that contains the binding site for the cAMP regulatory element binding protein (CREB). As can be seen in Figure 10, application of the GibTig determined oligonucleotide inhibited GABRA4 mRNA levels while the CRE control oligo exhibited no effect.

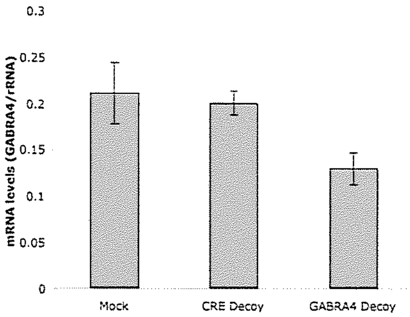


**Fig 10.** Decoy analysis of a GibTig in GABRA4. Primary cultures of rat neocortical neurons were treated with DOTAP (N-[1-(2,3-dioleoyloxy)propyl]-N, N,N-trimethylammonium methylsulfate) alone (Mock) or with DOTAP and phosphothioate oligonucleotides from either a cAMP response element (CRE Decoy) or a GibTig sequence from the GABRA4 promoter (GABRA4 Decoy). (GTGCACACACACGCCCACCGCGGCTCGGG). mRNA was harvested after 24 hr, and real-time PCR was performed with GABRA4 specific primers. Data was normalized to rRNA levels, and expressed as mRNA levels relative (GABRA4/rRNA). Results are shown as mean ± SD.

Taken together, these results suggest that even though the analysis was done on a small set of predictions, GibTigs can be successfully applied to solving the problem of de-novo TFBS identification in human upstream promoter regions. As we outline previously, the increased size of the human upstream regions makes detection of TFBS difficult. Until now, this has been a very difficult problem, with no known solution. Clearly, more work, has to be done in order to characterize the complex regulatory machinery of GABA$_A$ receptors. However, these are very promising results suggestive of a capable and robust framework with which to guide experimental research into the complex regulatory mechanisms of GABR genes and other regulatory mechanisms in *H. Sapiens.*

### *C.3.6 Proposed Research: Extending Discovery of TFBS and Regulation Mechanisms to Other Systems in the H. Sapiens*

Since initial results from the GibTig analysis applied to the upstream regions of GABRs were very encouraging (See **C.3.5**) we plan to expand this research to identifying regulation of sub-complexes, target identification and tissue-specific regulation in GABA and other systems in the Human genome. The aim is to improve sensitivity of the analysis and discover novel eukaryotic regulation models. For example, since the GABA-A receptor exhibits tissue specific expression of sub-complexes of the heteromeric receptor, we can use this system to develop a strategy for computational discovery of tissue-specific regulation mechanisms. Next, we can test the system further by applying the same pipeline of tools to discover control mechanisms in genes related to aging. (See **C.3.7**) [161, 162].

#### *C.3.6.1 Proposed Research: GABA receptor pathways (with Shelley Russek)*

First, we will test co-regulation of an exhaustive set of hypothetical sub-complexes. We plan to supplement existing data with phylogenetic footprint information from orthologous genes in the chimp, baboon, mouse and rat genomes, if available. We have shown that putative TFBS for GABR genes are conserved in other species. (See **C.3.5**) For each set of input genes, we will define TFBS using our GibTig approach as described in detail in **C.3.1-2**. This analysis will result in a set of binding sites upstream of every gene included in the variant of the GABA receptor. Each TFBS can be mapped back to the variant subsample that generated it. This procedure will identify mapping between the sub-complexes of GABA genes and sets of TFBS that co-regulate coordinated expression of these sub-complexes. The predictions can then be tested for both in-vitro and in-vivo activity as described in **C.3.5**

*C.3.6.2 Proposed Research: Estimating the false-positive rate of TFBS predictions*

To increase specificity and separate functional TFBS from other conserved elements (See **C.3.3.2**) we plan to perform a set of control runs. Subsets of upstream regions from the human genome will be chosen at random with size distributions matching those of the subsets discussed above. We will then estimate the false-positive PWM identification from this random sampling of upstream regions. We can compare each PWM originating from a subset of upstream regions of GABA genes to the expected occurrence at background level from our random set of upstream regions. (See **C.3.2-3**). This procedure will quantify the false-positive and true-positive rates of TFBS identification in the GABA genes. Furthermore, PWM in common with random sets of upstream regions probably constitute repetitive elements not specific for GABRs or those that lack transcriptional functionality altogether. (See **C.3.3.2**)

*C.3.6.3 Proposed Research: Predicting tissue-specific binding*

To identify tissue specific regulation, we will perform the search using our GibTig strategy on sets of upstream sequences from paralogous GABA genes. PWM that are common between GABA genes and paralogs or those sharing a common phylogenetic profile[163] could potentially be TFBS that are not specific to GABRs but regulate a large range of functions common to the "family" of ligand-gated ion channel genes to which GABRs belong. We are particularly interested in identifying TFBS that are unique to upstream regions of GABRs alone. Thus, by subtracting the set of positions identified in common with paralogs and other members from the same gene family (See **C.1.0**), we hope to identify the TFBS that are functional only in the neo-cortical tissue. (See **C.3.5**) Unique TFBS not found in upstream regions of other genes in the genome can be tested experimentally using cold-competition gel-shift assays and decoy analysis in non neo-cortical tissues such as the fibroblast (See **C.3.5**).

*C.3.6.4 Proposed Research: Creating hypothetical mutation hypotheses*

Finally, we plan to use the PWMs built from GibHits to create mutation hypotheses which will be used to create decoy oligos (See **C.3.5**) for functional analysis. The identification of mutations predicted to decrease the affinity of binding between the TF and the oligonucleotide will be estimated based on the probabilities in the computationally-derived binding affinity matrix. (See **C.3.1**) Any change in the original *w-mer* not matching the TFBS model identified by GibTIgs can be considered as a potential disruptive mutation. We can then test the effectiveness of the predictions using decoy analysis outlined in **C.3.5**.

### *C.3.7 Proposed Research: Age related TFs (based on work by Stuart Kim)*

Recent microarray experiments done in the lab of Stuart Kim [161, 162] have identified a set of genes that change expression patterns in correlation with aging in human kidney and brain. While the results were very interesting, no clear regulation mechanisms were proposed to explain the change in transcriptional activity. We plan to take the set of genes identified in those studies and subject them to the same pipeline of tools as described in **C.3.6**.

First, we will attempt to pick subsamples of the genes that show common TFBS in the upstream regions identified by the GibTig algorithm. Next, we will compare the binding sites predicted to those which would be expected if the sets of genes were picked randomly from the human genome. We will keep only those predictions which would not be expected by chance from a random sample of equal length upstream regions. The predicted *w*-mers would then be candidates for further testing and experimental validation for binding and activity.

## C.4 Defining System Level Organization of Regulatory Mechanisms

The combinatorial potential of subsets of TFs (Cis-Regulatory Modules, CRMs) controlling correlated expression of subsets of genes is immense (See **B.1.2**). Clearly, performing experimental assays aimed at validation of all combinations is impossible. However, recent experiments in *S. Cerevisiae* have revealed the "parts-list" of this network by detailing the mapping between genes and the TF likely to bind in the upstream region: location data. The results from these experiments have already expedited biological insights through integrative treatment of location data with existing high-throughput expression and annotation data.[15, 164, 165] However, a comprehensive picture defining CRMs and elucidating the mechanisms of their collaborative control and regulation of sets of genes is still largely lacking.

Furthermore, there is a direct relationship between understanding regulation by CRMs and building evolutionary models. Evolutionary pressure can be defined in terms of the consequence of mutation on the organism. Understanding the complex interdependency of interactions between transcription factors and genes is necessary for quantifying the strength of selection on mutations in upstream regions. (See **B.1.2**) Thus, mapping the full regulatory network and understanding how sub-networks are selected by environmental conditions[23, 126, 166, 167] is central not only to understanding the life-cycle of the cell[17, 168-171] but also in modeling the effect of mutations in binding sites on the phenotype of the cell.

### *C.4.0 Background: Using neural nets to identify regulatory modules*

The *cellular control network* describes subsets of transcription factors (TFs) acting together to facilitate correlated binding, transduction and co-expression of subsets of genes[125-127]. We would like to generalize the knowledge of the relationship between transcription factors and genes, into a higher-order model describing the organization of the genome into pathways. We used a neural network based on Adaptive Resonance Theory (ART) [31, 32] to classify genes into groups based on publicly available location data (from CHIP-Chip[13]). Briefly, the 3-layer neural net works by assigning and separating sets of feature vectors in multi-dimensional space. The algorithm proceeds by iteratively placing each vector $V_G$ into a long-term memory trace (LTM). The neural net can be influenced by modifying the sensitivity parameter $\rho$=[0,1]. At one extreme, when $\rho$=1, all LTMs are different e.g. genes will be clustered with other genes that share exactly their unique feature vector, when $\rho$=0, all genes belong to one LTM.

We used the ART neural network to identify sets of TFs involved in co-regulation of sets of genes. Since we can expect the co-regulated sets of genes to form pathways, we compare each subset of genes defined by ART to current annotation of genes into pathways by KEGG [109, 172]. To assess the level of similarity between sets of genes controlled by CRMs and those implicated in the same pathway by KEGG, we use a modified multi-set Jaccard.[173] Partitioning of genes into identical subsets by KEGG and the neural net will yield Jaccard =1. If the partitions do not overlap Jaccard = 0. To assess the statistical significance of the overlaps, we convert Jaccard coefficients into normalized Z-scores using a permutation test e.g. we create clusters in the null-model by randomly placing genes into sets with the same distribution of sizes as KEGG pathways.

### *C.4.0.1 Background: Comparing ART2 predicted CRMs to KEGG pathways*

We report two results. First, classification of genes into modules by ART2 closely resembles the partitioning of genes into KEGG, pathways (Fig 11). While the maximum raw Jaccard value is .375, the probability that this happened from a random classification of genes is infinitesimally small (Z>90 by a permutation test). The result also suggests that our ART defined CRMs describe sets of TFs that act in unison as major biological switches. These switches coordinate the regulation of genes in pathways.

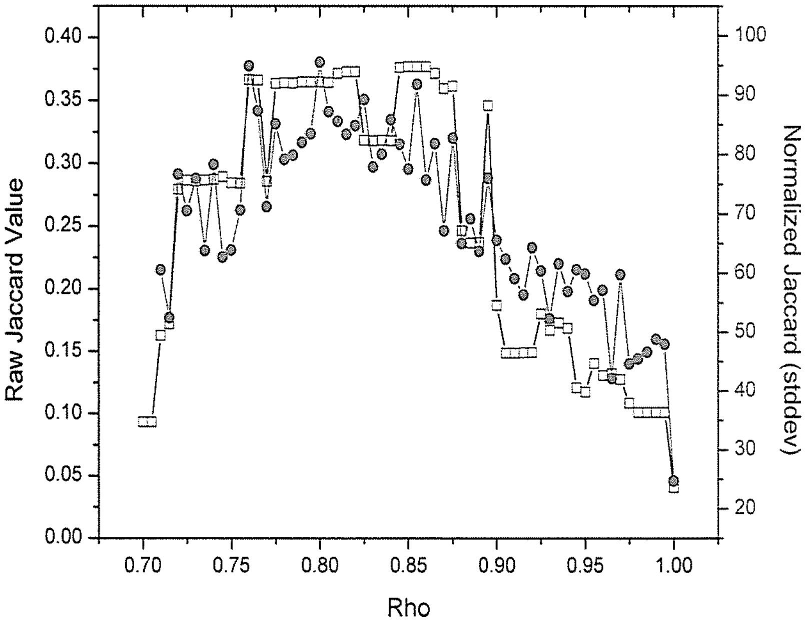


**Fig 11 .** *Jaccard overlap values between gene content of the cis-regulatory modules defined by ART and pathways as defined by KEGG. Circles denote normalized values with respect to a permutation null-model with the Y axis on the right. Squares are real Jaccard values with the Y axis on the left. Simulated random binding data shows no significant overlap with KEGG annotation (data not shown)*

*C.4.0.2 Background: Coherence of expression in ART defined CRMs*

Since genes in the same pathway are expected to share coherent expression patterns, we go on to evaluate the homogeneity of expression inside ART-defined modules. We compare ART defined modules based on location data alone to modules identified through clustering of expression data. We find that genes inside CRM-controlled modules are very likely to be co-expressed[174, 175]. Furthermore, we find (Fig 12) that gene sets controlled by CRMs consisting of a larger number of TFs (e.g. learned by the neural net at lower sensitivity threshold $\rho$) are more similar to expression modules at higher expression cutoff. Intuitively, this suggests that expression level is a function of the number of TFs in the CRM. The cellular control network achieves higher average expression of genes inside the module by combining smaller sets of TFs for additive effect on gene expression. These results also indicate that we may use this technique to investigate the hierarchy of control using ART deveined CRMs.

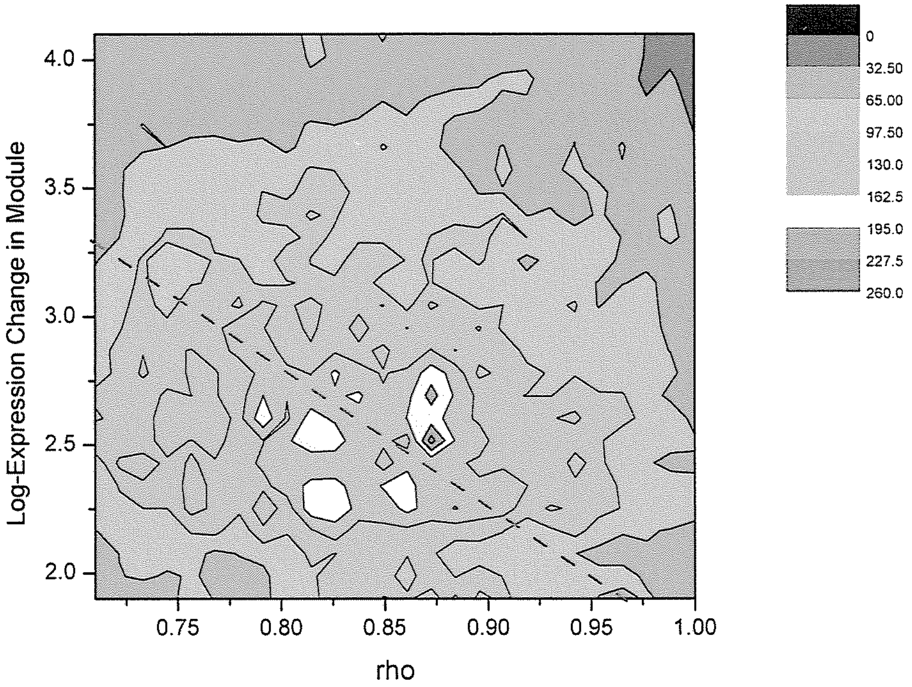


**Fig 12**. *The normalized correlation landscape between CRM modules and expression modules. The X axis is the $\rho$ sensitivity parameter used by the neural net. Larger values of $\rho$ describe more restrictive CRMs consisting of fewer TFs regulating smaller sets of genes. The Y axis is the average over-expression Ec inside modules derived from expression data[174, 175]. The color code is the standard deviations away from random estimated by a permutation test. The line illustrates inverse dependency of over-expression and $\rho$.*

An insight readily gleaned from this study, is that we can identify pathways and sets of transcription factors used to regulate those pathways (CRMs) from mapping between transcription factors and genes alone. Thus, we use these results to argue that our ability to determine a global map between TF and the genes they control (See **C.4.1**) will also yield significant insights into the function of other genes in the genome. By employing methods described here, we can use the relationships from whole genome TFBS identification to identify and elucidate the genetic parts-lists of major metabolic and signaling pathways (See **C.4.2**). As more location experiments are performed, or de-novo computational methods to predict TFBS mature (See **C.4.1**), we can expect the same approach to yield more coverage and higher specificity (See **C.3**).

### *C.4.1 Proposed Research: Whole-Genome identification of TFBS using GibTigs in yeast (with the BlueGene Team from IBM)*

Here, we outline an approach for the creation of a global "map" of TFBS upstream of all genes in the *S. Cerevisiae* genome[176]. The first challenge in computationally determining TFBS upstream of ~6000 open reading frames (ORFs) is identifying the hypothetical sets of co-regulated genes used as input into any TFBS identification algorithm (See **C.3.1**). Naive, exhaustive sampling of upstream regions to find enriched sets that share common TFBS would not be productive. For example, the number of possible sets of *5* genes including all subsets is on the order of $2^{40}$. However, in **C.3.2** we showed that CHIP-Chip data provides reasonable hypotheses for choosing sets of co-regulated genes to use in computational TFBS identification. We will combine our initial success at identifying potential co-regulated genes from CHIP-Chip with computational methods e.g. clustering based on functional distance (See **C.2.4**) or gene family identification (See **C.1.0**). These seeds would then be used to grow the whole-genome TFBS map.

#### *C.4.1.1 Proposed Research: Using computational methods and high-throughput experimental data to create sets of hypothetically co-regulated genes.*

We start by iteratively building the sets of upstream regions from putatively co-regulated genes using CHIP-Chip data[13, 15]. We will first create non-redundant sets where genes show binding to only one transcription factor and then identify sets that share binding by more than one TF. First, we plan to take all upstream regions for single-TF-bound gene sets (yielding 206 sets of upstream regions) and use these as input into the GibTig (**See C.3.1-3**) algorithm to find TFBSs. We predict that this kind of filtering to exclusion upstream sequences without extraneous motifs coupled with the increased sensitivity of the GibTig algorithm will increase the identification of binding specificities to well above the 31% of TFs currently reported by Young et al.[13] (see **C.3**) For example, when a similar filtering procedure was done to include only upstream regions that bind STE12, the recall (true positive matches to experimentally defined STE12 PWM) increased from ~60% (See **C.3.2**) to 100%. (data not shown). Due to the high computational complexity of the problem, we plan sequential deployment on BlueGene. Applying more computational power afforded by the IBM supercomputer can decrease the number of false-positives and further refine the sensitivity of the predictions.

We plan to complement the high-throughput experimental techniques such as CHIP-Chip and microarray experiments with computational approaches such as clustering based on the functional distance measure developed in **C.2.4** and paralogous gene families (See **C.1.0**). In both cases, we plan to leverage the already developed graph-theoretic clustering techniques to find the most homogeneous set of upstream regions(See **C.2.1** and **C.1.0**). Briefly, we start by defining genes as nodes and edges weighed by the computational distance metric (e.g. functional distance outlined in **C.2.4**). Then, using a clustering cutoff, we will find all strongly connected components in the graph. The cutoff value will be defined by the phase transition in the largest component [101]. These computational approaches are expected to yield TFBS for well beyond the 50% of upstream regions currently achieved with high-throughput CHIP-Chip[13].

### *C.4.1.2 Proposed Research: Including longer upstream regions*

In preliminary studies (See **C.3.2**), we find that TFBS occur not only in intergenic regions but also inside ORFs upstream of the regulated gene. Furthermore, we can show that the ability to identify the TFBS signal depends on the abundance and conservation of the conserved *w-mer* (data not shown). For example, the strength of the signal for the TEC1 PWM grows three-fold when parts of the upstream ORFs are included in the input set into the GibTig algorithm (**C.3.2**). To increase the probability of finding TFBS for as yet uncharacterized TFs and maximize the number of TFBS found for known TFs, we plan to include upstream regions from .5kb to 2kb in increments of .5kb disregarding the start sites of the next ORF region. This approach will ensure inclusion of TFBS that exist farther upstream or in ORFs such as those observed for TEC1. These studies require ability to handle longer upstream regions and robustness to noise demonstrated by our GibTig TFBS identification algorithm (See **C.3.1-3**).

*C.4.1.3 Proposed Research: Immediate implications for the data from whole-genome TFBS map in yeast*

Using the global TFBS map in yeast, we can also start asking questions about distributions of TFs that bind in ORFs. For example, we plan to undertake a study detailing the scatter in position of binding for every TF[139]. Using the data from this analysis in combination with microarray studies[1, 2, 175, 177, 178], we also plan to compare regulation mechanisms for TFs that bind in ORFs versus ones that bind in intergenic regions. However, the most important aspect of this research is that the resulting data will reveal a much more comprehensive map of TFBS in the *S. Cerevisiae* genome that can be used in conjunction with the planned database (See **C.5.2**) to identify targets for wet-lab experiments, create models of regulation and elucidate cis-regulatory modules.(See **C.5.3**) We expect that this data will prove invaluable to the scientific community.

***C.4.2 Synthesis: Reconstruction of major pathways using the computationally derived TFBS Map***

One obvious application of the whole-genome TFBS map derived in **C.4.1** is prediction of genes involved in common pathways. We show in **C.4.0** that using the partial map of TFBS from CHIP-Chip data, we were able to reconstruct KEGG pathways with significant precision. The accuracy of the pathway prediction can be used for improving the Gibtig algorithm (See **C.3.1-3**). We plan to make the pathway data available through our CRMer database (See **C.5.3**) and through our collaboration with Prof. Minoru Kanehisa by supplementation of KEGG (See **C.4.2.2**).

*C.4.2.1 Proposed Research: Using the ART neural net to predict pathways from whole-genome TFBS data*

We start by defining TFBS profiles for each gene using data from **C.4.1**. The profile would describe the set of predicted TFBS upstream of every ORF in the yeast genome. We can then turn these profiles into vectors where the dimensions represent TFs (See **C.4.0**). Every ORF in the genome would have an associated vector describing TFs that bind in the upstream region. Analogously to the p-value vector built from CHIP-Chip (See **C.4.0**), we would use the statistical significance of the TFBS prediction (developed in **C.3.3.1**) as the numerical approximation for the strength of binding.

Vectors from all genes would then be classified using the ART neural net. [31, 32] As shown in **C.4.0**, the neural net divides vectors into sets using the long-term memory trace (LTM). The LTM describes the relative importance of each TF in controlling the CRM. Since data from **C.4.1** is expected to have more coverage and better accuracy than the one obtained from CHIP-Chip data, we hope that this CRM prediction procedure will benefit from the increased quantity and quality of data. The predicted sets of genes would be compared to existing databases of pathways to assess the overlap in annotation.

*C.4.2.2 Proposed Research: Comparing to and extending existing pathway maps (with Minoru Kanehisa)*

Results from the whole genome TFBS identification using computational seeds (See **C.4.3.1**) would be compared with results from **C.4.0.** We can assess the improvement in coverage from building sets using computational methods over high-throughput experimental techniques such as CHIP-Chip and microarrays for building sets. However, since computational discovery of TFBS in Yeast is expected to yield a larger number of predictions, we will need a more careful curation of predicted results. Dr Kanehisa's group has been spearheading the manual annotation of genes into pathways and depositing the results into KEGG (Kyoto Encycplopedia of Genes and Genomes[109]). We plan to collaborate with Minoru Kanehisa's group in Kyoto University to assess the prospective value of our predicitons.

As a first pass, we can perform an automatic comparison of pathways predicted using ART and those annotated in KEGG using the Jaccard coefficient (See **C.4.0-1**). Since the neural net predicts only sets of genes, but does not describe the internal inter-relationship between members in those sets, we will need to refine the annotation further using orthology or manual assignment. This would likely be undertaken by professional curators at Kyoto. Results from this study would be deposited both in the KEGG database as well as into our own CRMer database (See **C.5.3**). As we extend the GibTig methodology to other, more complex eukaryotic organisms (See **C.3.6**), the same predictive procedure could be used to determine pathways in Human, Fly, Worm and other eukaryotic genomes. Furthermore, we can combine this technique with existing homology based methods (COG and KOG) to redefine protein function (See **C.4.3**) and begin studying evolution of pathways and control mechanisms on a genome-wide scale. (See **C.1.1, B.1**)

### *C.4.3 Proposed Research: Redefining Gene Function using TFBS profiles*

Gene function is a poorly defined concept (See **B.3** and [75, 76]). On one hand, gene products are often proteins that have several potential enzymatic or biochemical functions (See **C.2.2**). On the other, the same gene is subject to transcriptional control which places function in context (See **B.1.2** and **C.2.3**) e.g. along with other proteins in the same pathway. Furthermore, expression controls not only concentrations of partner proteins but also substrates needed for catalysis[179, 180]. At the same time, the protein is confined to a specific cellular component which further complicates a generalizeable definition of function. Currently, ontologies attempt to emphasize this point by dividing function into three parts: Biochemical, Pathway and Cellular Localization based on experimental data[70]. However, this annotation is often effuse especially for the "pathway" and "cellular localization" annotations.

Using data from our whole-genome TFBS map in *S. Cerevisiae* (See **C.4.1**), we will be in a position to redefine gene function using TFBS profiles. The transcriptional control of a gene represents a conglomeration of all aspects of its function: biochemical, pathway and cellular localization. We show in **C.4.0** that using a partial map based on CHIP-Chip derived TFBS-gene mapping, genes can be divided into common pathways. We would like to extend this concept further by completely redefining gene function with respect to elements controlling that gene's transcription. The function of a gene would be redefined as a unique fingerprint of TFBSs, strengths of binding and positions in the upstream region describing not only control of the specific gene, but also implicating the gene in a pathway. (See **C.4.2**)

#### *C.4.3.1 Proposed Research: Structure-Function relationship using TFBS profiles*

Using the above redefinition of function in terms of TFBS profiles, we would like to re-explore the relationship between structure and function. We show in preliminary evidence (See **C.2.1-2** and [42]) that structural similarity has a parallel in function space. Analogously, we could imagine that divergent evolution on a smaller evolutionary time-scale due to duplication and divergence (See **C.1.0**) could result in a similar correspondence between structural similarity and transcriptional control.

We will use the PDUG formalism (See **C.2.1**) and the recently published data from Skolnick et. al. (Personal communication and [181, 182]) to describe the similarity in structure between all threaded ORFs in the Yeast genome. In parallel, we can use the ART neural net (See **C.4.0**) to describe distances between TFBS profiles. Using this distance measure we will build the BPG (Binding Profile Graph). In both graphs (PDUG and BPG) the open reading frames would represent nodes and edges would be weighted by – structural similarity Z score[107] or distance between TFBS profiles (See **C.4.0**) . We will then proceed to calculate the strongly connected components of both graphs at varying cutoffs analogously to [108].

The purpose of this study would be to determine the mapping between the structural classes such as Family, Fold and Superfamily[76] and the binding profiles. Preliminary evidence shows that during duplication and divergence of genes, the biochemical function is constrained by structural determinants (See **C.1.0**). Alternatively, the pathway assignment is constrained by biochemical function. Finally, evolution coordinates nearly simultaneous expression of a set of genes constituting a pathway by selecting a common TFBS profile (See **B.1.2**) [1]. We would like to follow this line of logic to find the relationship between constraint placed on biochemical function by structural determinants and the TFBS controlling coordinated expression of paralogous genes (See **C.1.0**) involved in common pathways (See **B.3.2**).

### *C.4.3.2 Proposed Research: TFBS profiles in phylogenetic context*

In [69] we saw that inclusion of genomic context improves the precision of structure-function relationship when function is broadly defined as biochemical activity. Given that we find a corresponding relationship between binding profiles and structural similarity from **C.4.3.1**, we would like to further explore the relationships between phylogenetic context, structural similarity and transcriptional regulation. If structure represents a coarse description of the potential of a gene, this potential may be fulfilled by disparate TFBS profiles. However, each profile corresponds to a set of genes acting in unison along some pathway. Genes in common pathways are well-known to travel together along the evolutionary tree. [77, 163, 183] Thus, if a set of proteins share a particular distribution on an evolutionary tree, they are more likely to be involved in the same pathway, and therefore share a common TFBS profile.

First, we would develop a eukaryotic phylogenetic profile scheme analogous to one used in prokaryotes[69] using the KOG database[184, 185]. We would then regress the structure-TFBS profile relationship along phylogenetic profiles. We hope that proteins occurring in similar set of genomes would further improve the precision of the relationship between the new definition of function and structure analogously to results reported in [69].One possible issue includes a limited number of genomes from which to draw phylogenetic profiles. However, we hope that with an ever-increasing compendium of sequenced organisms, the number of genomes will be sufficient. Furthermore, we can use a stringent threshold that minimizes the false-positive rate if the number of eukaryotic genomes remains too low for predictive purposes[163].

## *C.5* Publication of Data, Resources and Tools

The predictions and theoretical arguments resulting from the studies outlined in **C.1-4** above have to be made freely available to the outside world for maximum effect. I believe that presentation of the data is often as important as the data itself. With hundreds of thousands of possible predictions, the usefulness of the data hinges on proper organization and presentation. The data should be packaged into intelligent databases allowing for dynamic querying and concise overview of the results. We plan to continue providing and updating our successful ELISA database[102, 106] describing our results in structure-function investigations.

### *C.5.1 ELISA : Database of Evolutionary functional Lineage Inferred from Structural Analysis.*

ELISA stores information on all PDUG nodes(See C.2.1), their characteristics and connections to other nodes. Each domain (node) has structure, sequence and taxonomic data recorded, as well as SCOP fold name and PDUG cluster information. It also includes structure comparison and sequence comparison data to other nodes. We allow searches using functional, sequence, structural and genomic information.

We provide a dynamic web interface for the underlying relational database. The web interface is divided into three levels. The first level asks for the query protein sequence, function, taxonomy or fold description. The function and fold description should match those used in GO and SCOP respectively. The sequence query approach will perform a sequence alignment against the database to identify all domains matching the query protein. For example, one of the possible queries at the first level could ask to see all nodes are present in a specific genome. A whole neighborhood may also be delineated by common taxonomic representation or common function and structure annotation.

In this way ELISA can be used to break new sequences into domains and identify structures and possible functions of those domains. For example, let the query protein contain a domain that was labeled as a tRNA ligase. Assume that the researcher would like to know of other ligases in humans with similar structural characteristics. This could be useful in understanding possible interactions for a small molecules or drug target. ELISA shows that there are close to forty domains that are involved in ligase activity in the human. Not surprisingly, there is a large overlap between ligases and ATP binding proteins. What is somewhat surprising, is how few structures involved in ligase activity are recycled for other functions. There are almost no other functions present in close structural proximity. Using ELISA, we can also say something about the structure of human ligases, all human ligases are alpha and beta proteins. Thus, user-defined combinations of characteristics can limit the divergence of a protein set to describe related proteins as in the example above. Through this logical "limitation of divergence", the researcher can find the prevailing pressures in the evolution of a protein family.

### *C.5.2 Publishing data from whole genome binding site mapping and CRM identification*

To facilitate usage of our TFBS prediction data by biologists and other researchers, we plan to develop a dynamic database using results from large-scale GibTig calculations done in high-performance computing environments. The specific role of the GibTig Interface is twofold. First, the interface will act as a browser for GibHit results from both whole genome yeast experiments (see **C.4.1**) and GABA receptor experiments (see **C.3.6**). Second, we plan to incorporate our preliminary results with the ART neural network (see **C.4.0**) to enable dynamic definition of cis-regulatory modules based on user-defined input sets. The interface will provide an engine to generate regulation models involving sets of genes. The interface and the underlying data will be freely available to the public on the world-wide-web.

At the core of the GibTig Interface is a relational database repository of GibHits either from genome wide GibTig mapping of TFBS in yeast (See **C.4.1**) or from $GABA_A$ receptor GibTig experiments (See **C.3.6**). The database will use the model of the Ensembl repository of genomic information[186]. Briefly, Ensembl combines eukaryotic genomic sequence information with various genomic features such as genes, pseudogenes, transcripts, exons, and introns in a single relational database. Moreover, Ensembl is easily extended to combine additional relational tables with existing sequence and feature data. Ensembl is therefore an ideal foundation upon which to build the GibTig Interface. However, Ensembl does not maintain repositories for all sequenced genomes. Notable exceptions include all strains of yeast. In these cases, it will be necessary to populate a minimal set of tables with genomic information prior to including GibTig information.

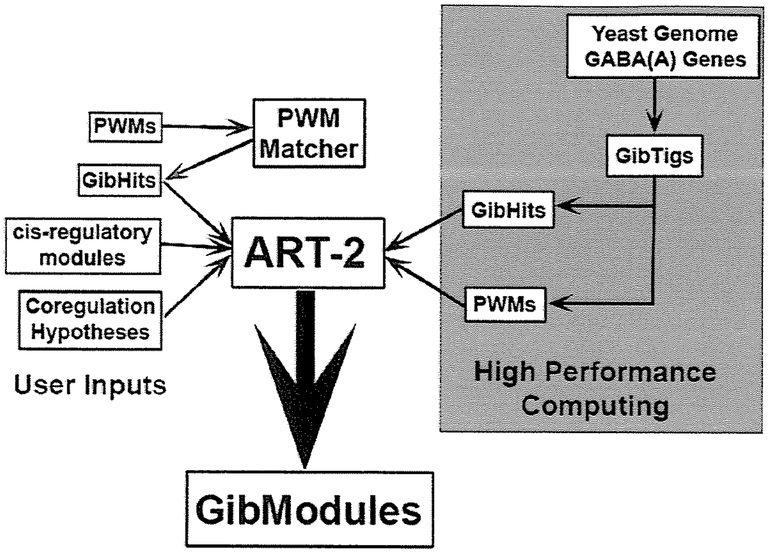


*Fig 13: Schematic of GibTig Interface design. The GibTig Interface will serve as a front end to provide researchers access to- results from genome wide GibTig analyses, and to ART generated cis-regulatory modules (GibModules).*

The purpose of the GibTig Interface is to facilitate research based on our pre-computed GABA$_A$ and yeast genome GibTigs. In particular, scientists will be able to use the interface to predict cis-regulatory motifs and TFBS in the promoter regions of their favorite gene. (See **C.4.1**), Furthermore, they will be able to identify motifs that correspond to known TFs, identify motifs for novel TFs, predict loss-of-function point mutations (See **C.3.6.4**), and cis-regulatory modules (See **C.4.2**).

### *C.5.3 CRMer—Database of predicted Pathways and Control Structures in Yeast*

The GibTig Interface will also couple the yeast and GABA predictions with our neural network classifier (See **C.4.0** and **C.4.2**) to help researchers generate regulation models describing sets of genes controlled by sets of transcription factors (i.e. cis-regulatory modules, or CRMs). In preliminary evidence (See **C.4.0**), we trained a neural network (ART) to identify biologically meaningful CRMs using ChIP-chip profiles. Regulatory modules, termed GibModules, can be generated by researchers as needed in one of three ways. Researchers interested in a set of potentially co-regulated genes will use the classifier to identify sets of GibHits that describe common binding profiles of those genes. On the other hand, researchers interested in set of PWMs will use the classifier to identify a set of genes combinatorially controlled by these GibHits. For example, researchers interested in CRMs involving STE12 could input a gene set of interest e.g. a set of genes that are expressed during filamentous growth to the ART classifier. ART would then find the most likely CRM for that input set e.g it would find that the genes are also regulated by TEC1. Then, the researcher could input STE12 and TEC1 PWMs to identify a larger module of genes coordinately regulated by both STE12 and TEC1.

Thus, a third approach is iterative generation of GibModules i.e. using a set of genes to predict ~~a~~ set of factors, and then using those factors to predict a new set of genes. Iterative application of the ART-2 GibTig Interface will generate larger GibModules that will put a gene or a set of genes in transcriptional context. Using the previous example, an iterative approach might reveal genes controlled by both STE12 and TEC1 perform a specific set of functions that differ from genes regulated by STE12 alone or TEC1 alone. (See **C.3.2**)

Researchers can also use GibModules to predict functions for uncharacterized or incompletely characterized genes. Since GibHits are generated independent of experimental conditions, GibModules can reveal pathways and gene functions that are not observed with commonly assayed experimental conditions. As an example, applying the classifier to $\text{GABA}_A$ related GibHits might reveal novel modes of co-regulation within the $\text{GABA}_A$ receptor (See **C.3.4-5**). In some cases, those novel modes of regulation may be predictive of a disease phenotype, such as complexes that are implicated with some forms of epilepsy[187], and Alzheimer's disease[188]. In the case of $\text{GABA}_A$ receptors, understanding GibModules connected to epilepsy can potentially identify a set of transcription factors that can be targeted to treat the disease.

**References:**

1. Ihmels, J., et al., *Rewiring of the yeast transcriptional network through the evolution of motif usage.* Science, 2005. **309**(5736): p. 938-40.
2. Tanay, A., A. Regev, and R. Shamir, *Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast.* Proc Natl Acad Sci U S A, 2005. **102**(20): p. 7203-8.
3. Basu, S., et al., *A synthetic multicellular system for programmed pattern formation.* Nature, 2005. **434**(7037): p. 1130-4.
4. Hasty, J., D. McMillen, and J.J. Collins, *Engineered gene circuits.* Nature, 2002. **420**(6912): p. 224-30.
5. Hasty, J., et al., *Synthetic gene network for entraining and amplifying cellular oscillations.* Phys Rev Lett, 2002. **88**(14): p. 148101.
6. Gale, C., et al., *Candida albicans Int1p interacts with the septin ring in yeast and hyphal cells.* Mol Biol Cell, 2001. **12**(11): p. 3538-49.
7. Li, X. and W.H. Wong, *Sampling motifs on phylogenetic trees.* Proc Natl Acad Sci U S A, 2005. **102**(27): p. 9481-6.
8. Arvidson, D.N., et al., *The tryptophan repressor sequence is highly conserved among the Enterobacteriaceae.* Nucleic Acids Res, 1994. **22**(10): p. 1821-9.
9. Hajra, A., et al., *DNA sequences in the promoter region of the NF1 gene are highly conserved between human and mouse.* Genomics, 1994. **21**(3): p. 649-52.
10. Prakash, A., et al., *Motif discovery in heterogeneous sequence data.* Pac Symp Biocomput, 2004: p. 348-59.
11. Wang, T. and G.D. Stormo, *Combining phylogenetic data with co-regulated genes to identify regulatory motifs.* Bioinformatics, 2003. **19**(18): p. 2369-80.
12. Giaever, G., et al., *Functional profiling of the Saccharomyces cerevisiae genome.* Nature, 2002. **418**(6896): p. 387-91.
13. Harbison, C.T., et al., *Transcriptional regulatory code of a eukaryotic genome.* Nature, 2004. **431**(7004): p. 99-104.
14. Zeitlinger, J., et al., *Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling.* Cell, 2003. **113**(3): p. 395-404.
15. Lee, T.I., et al., *Transcriptional regulatory networks in Saccharomyces cerevisiae.* Science, 2002. **298**(5594): p. 799-804.
16. King, D.C., et al., *Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences.* Genome Res, 2005. **15**(8): p. 1051-60.
17. Buchler, N.E., U. Gerland, and T. Hwa, *On schemes of combinatorial transcription logic.* Proc Natl Acad Sci U S A, 2003. **100**(9): p. 5136-41.
18. Hartwell, L.H., et al., *From molecular to modular cell biology.* Nature, 1999. **402**(6761 Suppl): p. C47-52.
19. Tavazoie, S., et al., *Systematic determination of genetic network architecture.* Nat Genet, 1999. **22**(3): p. 281-5.
20. Hughes, J.D., et al., *Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae.* J Mol Biol, 2000. **296**(5): p. 1205-14.
21. Roth, F.P., et al., *Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.* Nat Biotechnol, 1998. **16**(10): p. 939-45.
22. Spellman, P.T., et al., *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.* Mol Biol Cell, 1998. **9**(12): p. 3273-97.
23. Howard, M.L. and E.H. Davidson, *cis-Regulatory control circuits in development.* Dev Biol, 2004. **271**(1): p. 109-18.

24. Herskowitz, I., *MAP kinase pathways in yeast: for mating and more.* Cell, 1995. **80**(2): p. 187-97.
25. Roeder, R.G., *The role of general initiation factors in transcription by RNA polymerase II.* Trends Biochem Sci, 1996. **21**(9): p. 327-35.
26. Estruch, F., *Stress-controlled transcription factors, stress-induced genes and stress tolerance in budding yeast.* FEMS Microbiol Rev, 2000. **24**(4): p. 469-86.
27. Hieter, P. and M. Boguski, *Functional genomics: it's all how you read it.* Science, 1997. **278**(5338): p. 601-2.
28. Lockhart, D.J. and E.A. Winzeler, *Genomics, gene expression and DNA arrays.* Nature, 2000. **405**(6788): p. 827-36.
29. Boguski, M.S., *Biosequence exegesis.* Science, 1999. **286**(5439): p. 453-5.
30. Cliften, P., et al., *Finding functional features in Saccharomyces genomes by phylogenetic footprinting.* Science, 2003. **301**(5629): p. 71-6.
31. Cao, Y. and J. Wu, *Projective ART for clustering data sets in high dimensional spaces.* Neural Netw, 2002. **15**(1): p. 105-20.
32. Linares-Barranco, B. and T. Serrano-Gotarredona, *A Modified ART 1 Algorithm more Suitable for VLSI Implementations.* Neural Netw, 1996. **9**(6): p. 1025-1043.
33. Cavalier-Smith, T., *The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification.* Int J Syst Evol Microbiol, 2002. **52**(Pt 1): p. 7-76.
34. Ohta, T., *Evolution by gene duplication and compensatory advantageous mutations.* Genetics, 1988. **120**(3): p. 841-7.
35. Ohta, T., *Role of gene duplication in evolution.* Genome, 1989. **31**(1): p. 304-10.
36. Ohno, S., *Evolution by gene duplication.* 1970, Berlin, New York,: Springer-Verlag. xv, 160.
37. Nei, M., Roychoudhury, A. K., *Probability of fixation of nonfunctional genes at duplicate loci.* Am. Nat, 1973. **107**(156): p. 590-605.
38. Petrov, D.A. and D.L. Hartl, *Pseudogene evolution and natural selection for a compact genome.* J Hered, 2000. **91**(3): p. 221-7.
39. Force, A., et al., *Preservation of duplicate genes by complementary, degenerative mutations.* Genetics, 1999. **151**(4): p. 1531-45.
40. Lynch, M. and A. Force, *The probability of duplicate gene preservation by subfunctionalization.* Genetics, 2000. **154**(1): p. 459-73.
41. He, X. and J. Zhang, *Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution.* Genetics, 2005. **169**(2): p. 1157-64.
42. Tanella, A., Reddy, R.,Shakhnovich, B., *Using Neural Networks to Define Cis Regulatory Modules Controlling Major Biological Pathways in Yeast.* In Preparation.
43. Jordan, I.K., Y.I. Wolf, and E.V. Koonin, *Duplicated genes evolve slower than singletons despite the initial rate increase.* BMC Evol Biol, 2004. **4**(1): p. 22.
44. Efimov, A.V., *Structural similarity between two-layer alpha/beta and beta-proteins.* J Mol Biol, 1995. **245**(4): p. 402-15.
45. Doolittle, R.F., *Convergent evolution: the need to be explicit.* Trends Biochem Sci, 1994. **19**(1): p. 15-8.
46. Saier, M.H., Jr., *Convergence and divergence in the evolution of transport proteins.* Bioessays, 1994. **16**(1): p. 23-9.
47. Salemme, F.R., M.D. Miller, and S.R. Jordan, *Structural convergence during protein evolution.* Proc Natl Acad Sci U S A, 1977. **74**(7): p. 2820-4.
48. Charbonnier, J.B., et al., *Structural convergence in the active sites of a family of catalytic antibodies.* Science, 1997. **275**(5303): p. 1140-2.
49. Chothia, C., et al., *Protein folds in the all-beta and all-alpha classes.* Annu Rev Biophys Biomol Struct, 1997. **26**: p. 597-627.
50. Chothia, C., M. Levitt, and D. Richardson, *Structure of proteins: packing of alpha-helices and pleated sheets.* Proc Natl Acad Sci U S A, 1977. **74**(10): p. 4130-4.

51. Hunt, N.G., L.M. Gregoret, and F.E. Cohen, *The origins of protein secondary structure. Effects of packing density and hydrogen bonding studied by a fast conformational search.* J Mol Biol, 1994. **241**(2): p. 214-25.
52. Makarova, K.S., et al., *A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis.* Nucleic Acids Res, 2002. **30**(2): p. 482-96.
53. Cheek, S., H. Zhang, and N.V. Grishin, *Sequence and structure classification of kinases.* J Mol Biol, 2002. **320**(4): p. 855-81.
54. Leonard, C.J., L. Aravind, and E.V. Koonin, *Novel families of putative protein kinases in bacteria and archaea: evolution of the "eukaryotic" protein kinase superfamily.* Genome Res, 1998. **8**(10): p. 1038-47.
55. Russell, R.B., et al., *Evolutionary relationship between the bacterial HPr kinase and the ubiquitous PEP-carboxykinase: expanding the P-loop nucleotidyl transferase superfamily.* FEBS Lett, 2002. **517**(1-3): p. 1-6.
56. Leipe, D.D., E.V. Koonin, and L. Aravind, *Evolution and classification of P-loop kinases and related proteins.* J Mol Biol, 2003. **333**(4): p. 781-815.
57. Shakhnovich, B.E., et al., *Protein structure and evolutionary history determine sequence space topology.* Genome Res, 2005. **15**(3): p. 385-92.
58. Eigen, M., *Naturwissenshaften.* Vol. 58. 1971.
59. Eigen, M., Schuster, P, *The hypercycle: A principle of natural self-organization.* 1979, Berlin: Springer Verlag.
60. Ponting, C.P. and N.J. Dickens, *Genome cartography through domain annotation.* Genome Biol, 2001. **2**(7): p. Comment 2006.
61. Wilson, C.A., J. Kreychman, and M. Gerstein, *Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores.* J Mol Biol, 2000. **297**(1): p. 233-49.
62. Ison, J.C., *Exploring protein domain structure.* Brief Bioinform, 2000. **1**(3): p. 305-12.
63. Buchan, D.W., et al., *Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database.* Genome Res, 2002. **12**(3): p. 503-14.
64. Orengo, C.A., et al., *The CATH protein family database: a resource for structural and functional annotation of genomes.* Proteomics, 2002. **2**(1): p. 11-21.
65. Kinch, L.N. and N.V. Grishin, *Evolution of protein structures and functions.* Curr Opin Struct Biol, 2002. **12**(3): p. 400-8.
66. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
67. Dietmann, S. and L. Holm, *Identification of homology in protein structure classification.* Nat Struct Biol, 2001. **8**(11): p. 953-7.
68. Lord, P.W., et al., *Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.* Bioinformatics, 2003. **19**(10): p. 1275-83.
69. Shakhnovich, B.E., *Improving the precision of the structure-function relationship by considering phylogenetic context.* PLoS Comput Biol, 2005. **1**(1): p. e9.
70. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
71. Shakhnovich, B.E., et al., *Comparisons of predicted genetic modules: identification of co-expressed genes through module gene flow.* Genome Inform Ser Workshop Genome Inform, 2004. **15**(1): p. 221-8.
72. Fraser, A.G. and E.M. Marcotte, *A probabilistic view of gene function.* Nat Genet, 2004. **36**(6): p. 559-64.
73. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J Mol Biol, 1995. **247**(4): p. 536-40.

74. Eisenstein, E., et al., *Biological function made crystal clear - annotation of hypothetical proteins via structural genomics.* Curr Opin Biotechnol, 2000. **11**(1): p. 25-30.

75. Shakhnovich, B.E., et al., *Functional fingerprints of folds: evidence for correlated structure-function evolution.* J Mol Biol, 2003. **326**(1): p. 1-9.

76. Shakhnovich, B.E. and J. Max Harvey, *Quantifying structure-function uncertainty: a graph theoretical exploration into the origins and limitations of protein annotation.* J Mol Biol, 2004. **337**(4): p. 933-49.

77. Pellegrini, M., et al., *Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.* Proc Natl Acad Sci U S A, 1999. **96**(8): p. 4285-8.

78. Teichmann, S.A. and M.M. Babu, *Gene regulatory network growth by duplication.* Nat Genet, 2004. **36**(5): p. 492-6.

79. Keightley, P.D. and A. Eyre-Walker, *Terumi Mukai and the riddle of deleterious mutation rates.* Genetics, 1999. **153**(2): p. 515-23.

80. Drake, J.W., et al., *Rates of spontaneous mutation.* Genetics, 1998. **148**(4): p. 1667-86.

81. Macarthur, R. and R. Levins, *Competition, Habitat Selection, and Character Displacement in a Patchy Environment.* Proc Natl Acad Sci U S A, 1964. **51**: p. 1207-10.

82. Levins, R., *Evolution in changing environments; some theoretical explorations.* 1968, Princeton, N.J.,: Princeton University Press. ix, 120.

83. Hirsh, A.E. and H.B. Fraser, *Protein dispensability and rate of evolution.* Nature, 2001. **411**(6841): p. 1046-9.

84. Yang, J., Z. Gu, and W.H. Li, *Rate of protein evolution versus fitness effect of gene deletion.* Mol Biol Evol, 2003. **20**(5): p. 772-4.

85. Hurst, L.D. and N.G. Smith, *Do essential genes evolve slowly?* Curr Biol, 1999. **9**(14): p. 747-50.

86. Fraser, H.B., et al., *Evolutionary rate in the protein interaction network.* Science, 2002. **296**(5568): p. 750-2.

87. Pal, C., B. Papp, and L.D. Hurst, *Highly expressed genes in yeast evolve slowly.* Genetics, 2001. **158**(2): p. 927-31.

88. Jordan, I.K., Y.I. Wolf, and E.V. Koonin, *No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly.* BMC Evol Biol, 2003. **3**(1): p. 1.

89. Drummond, A.D., A. Raval, and C.O. Wilke, *A single determinant for the rate of yeast protein evolution.* arxiv.org, 2005(q-bio.PE/0506011).

90. Wall, D.P., et al., *Functional genomic analysis of the rates of protein evolution.* Proc Natl Acad Sci U S A, 2005. **102**(15): p. 5483-8.

91. Harrison, P., et al., *A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution.* J Mol Biol, 2002. **316**(3): p. 409-19.

92. Kellis, M., et al., *Sequencing and comparison of yeast species to identify genes and regulatory elements.* Nature, 2003. **423**(6937): p. 241-54.

93. Waterston, R.H., et al., *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): p. 520-62.

94. Ogurtsov, A.Y., S. Sunyaev, and A.S. Kondrashov, *Indel-based evolutionary distance and mouse-human divergence.* Genome Res, 2004. **14**(8): p. 1610-6.

95. Felsenstein, J., *Evolutionary trees from DNA sequences: a maximum likelihood approach.* J Mol Evol, 1981. **17**(6): p. 368-76.

96. Wolf, M.J., et al., *TrExML: a maximum-likelihood approach for extensive tree-space exploration.* Bioinformatics, 2000. **16**(4): p. 383-94.

97. Whelan, S., P.I. de Bakker, and N. Goldman, *Pandit: a database of protein and associated nucleotide domains with inferred trees.* Bioinformatics, 2003. **19**(12): p. 1556-63.

98. Simon, C., et al., *Large differences in substitutional pattern and evolutionary rate of 12S ribosomal RNA genes.* Mol Biol Evol, 1996. **13**(7): p. 923-32.

99. Brudno, M., et al., *LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA.* Genome Res, 2003. **13**(4): p. 721-31.

100. Brudno, M., et al., *Glocal alignment: finding rearrangements during alignment.* Bioinformatics, 2003. **19 Suppl 1**: p. i54-62.

101. Dokholyan, N.V., B. Shakhnovich, and E.I. Shakhnovich, *Expanding protein universe and its origin from the biological Big Bang.* Proc Natl Acad Sci U S A, 2002. **99**(22): p. 14132-6.

102. Shakhnovich, B.E., J.M. Harvey, and C. Delisi, *ELISA: a unified, multidimensional view of the protein domain universe.* Genome Inform Ser Workshop Genome Inform, 2004. **15**(1): p. 213-20.

103. Tiana, G., et al., *Imprint of evolution on protein structures.* Proc Natl Acad Sci U S A, 2004. **101**(9): p. 2846-51.

104. Deeds, E.J., B. Shakhnovich, and E.I. Shakhnovich, *Proteomic traces of speciation.* J Mol Biol, 2004. **336**(3): p. 695-706.

105. Dokholyan, N.V. and E.I. Shakhnovich, *Understanding hierarchical protein evolution from first principles.* J Mol Biol, 2001. **312**(1): p. 289-307.

106. Shakhnovich, B.E., et al., *ELISA: Structure-Function Inferences based on statistically significant and evolutionarily inspired observations.* BMC Bioinformatics, 2003. **4**(1): p. 34.

107. Dietmann, S., et al., *A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3.* Nucleic Acids Res, 2001. **29**(1): p. 55-7.

108. Shakhnovich, B.E., N.V. Dokholyan, and E.I. Shakhnovich, *Functional fingerprints of folds: Evidence for correlated structure function evolution.* J. Mol. Biol, 2003. **326**: p. 1-9.

109. Kanehisa, M., *The KEGG database.* Novartis Found Symp, 2002. **247**: p. 91-101; discussion 101-3, 119-28, 244-52.

110. Todd, A.E., C.A. Orengo, and J.M. Thornton, *Evolution of function in protein superfamilies, from a structural perspective.* J Mol Biol, 2001. **307**(4): p. 1113-43.

111. Aoki, K.F., et al., *A score matrix to reveal the hidden links in glycans.* Bioinformatics, 2005. **21**(8): p. 1457-63.

112. Aoki, K.F., et al., *Application of a new probabilistic model for recognizing complex patterns in glycans.* Bioinformatics, 2004. **20 Suppl 1**: p. I6-I14.

113. Aoki, K.F., et al., *Efficient tree-matching methods for accurate carbohydrate database queries.* Genome Inform Ser Workshop Genome Inform, 2003. **14**: p. 134-43.

114. Aoki, K.F., et al., *KCaM (KEGG Carbohydrate Matcher): a software tool for analyzing the structures of carbohydrate sugar chains.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W267-72.

115. Chothia, C., et al., *Evolution of the protein repertoire.* Science, 2003. **300**(5626): p. 1701-3.

116. Thornton, J.M., et al., *Protein folds, functions and evolution.* J Mol Biol, 1999. **293**(2): p. 333-42.

117. Kondor, R.I. and J. Lafferty, *Machine learning : proceedings of the Nineteenth International Conference (ICML 2002) : University of New South Wales, Sydney, Australia, July 8-12, 2002.* 2002, San Francisco, Calif.: Morgan Kaufmann Publishers. x, 706.

118. Belkin, M. and P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation.* Neural Computation, 2003. **15**(6): p. 1373-1396.

119. Coifman, R.R., et al., *Geometric diffusions as a tool for harmonic analysis and structure definition of data: Multiscale methods.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(21): p. 7432-7437.

120. Roweis, S.T. and L.K. Saul, *Nonlinear dimensionality reduction by locally linear embedding.* Science, 2000. **290**(5500): p. 2323-+.

121. Tenenbaum, J.B., V. de Silva, and J.C. Langford, *A global geometric framework for nonlinear dimensionality reduction.* Science, 2000. **290**(5500): p. 2319-+.

122. Arthur, M. and P. Courvalin, *Genetics and mechanisms of glycopeptide resistance in enterococci.* Antimicrob Agents Chemother, 1993. **37**(8): p. 1563-71.

123. Walsh, C., *Molecular mechanisms that confer antibacterial drug resistance.* Nature, 2000. **406**(6797): p. 775-81.

124. Walsh, C.T., et al., *Bacterial resistance to vancomycin: five genes and one missing hydrogen bond tell the story.* Chem Biol, 1996. **3**(1): p. 21-8.

125. Alberts, B., *Molecular biology of the cell.* 4th ed. 2002, New York: Garland Science. xxxiv, 1463, [86].

126. Alberts, B., *Essential cell biology.* 2nd ed. 2004, New York, NY: Garland Science Pub. xxi, 740, [102].

127. Stillman, B., *Mechanisms of transcription.* Cold Spring Harbor symposia on quantitative biology, v. 63. 1998, Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press. xxx, 679.

128. Kuhn, E.J. and P.K. Geyer, *Genomic insulators: connecting properties to mechanism.* Curr Opin Cell Biol, 2003. **15**(3): p. 259-65.

129. Bell, A.C., A.G. West, and G. Felsenfeld, *Insulators and boundaries: versatile regulatory elements in the eukaryotic.* Science, 2001. **291**(5503): p. 447-50.

130. Cai, H.N. and P. Shen, *Effects of cis arrangement of chromatin insulators on enhancer-blocking activity.* Science, 2001. **291**(5503): p. 493-5.

131. Lawrence, C.E., et al., *Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.* Science, 1993. **262**(5131): p. 208-14.

132. Brazma, A., et al., *Approaches to the automatic discovery of patterns in biosequences.* J Comput Biol, 1998. **5**(2): p. 279-305.

133. Bussemaker, H.J., H. Li, and E.D. Siggia, *Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis.* Proc Natl Acad Sci U S A, 2000. **97**(18): p. 10096-100.

134. Bailey, T.L. and C. Elkan, *Fitting a mixture model by expectation maximization to discover motifs in biopolymers.* Proc Int Conf Intell Syst Mol Biol, 1994. **2**: p. 28-36.

135. Neuwald, A.F., J.S. Liu, and C.E. Lawrence, *Gibbs motif sampling: detection of bacterial outer membrane protein repeats.* Protein Sci, 1995. **4**(8): p. 1618-32.

136. Tompa, M., et al., *Assessing computational tools for the discovery of transcription factor binding sites.* Nat Biotechnol, 2005. **23**(1): p. 137-44.

137. Liu, Y., et al., *A suite of web-based programs to search for transcriptional regulatory motifs.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W204-7.

138. Pietrokovski, S., *Searching databases of conserved sequence regions by aligning protein multiple-alignments.* Nucleic Acids Res, 1996. **24**(19): p. 3836-45.

139. Zhu, J. and M.Q. Zhang, *SCPD: a promoter database of the yeast Saccharomyces cerevisiae.* Bioinformatics, 1999. **15**(7-8): p. 607-11.

140. Wingender, E., et al., *TRANSFAC: an integrated system for gene expression regulation.* Nucleic Acids Res, 2000. **28**(1): p. 316-9.

141. Bao, M.Z., et al., *Pheromone-dependent destruction of the Tec1 transcription factor is required for MAP kinase signaling specificity in yeast.* Cell, 2004. **119**(7): p. 991-1000.

142. Luscombe, N.M., et al., *Genomic analysis of regulatory network dynamics reveals large topological changes.* Nature, 2004. **431**(7006): p. 308-12.

143. Brown, A.J. and N.A. Gow, *Regulatory networks controlling Candida albicans morphogenesis.* Trends Microbiol, 1999. **7**(8): p. 333-8.

144. Chen, H.C., et al., *Quantitative characterization of the transcriptional regulatory network in the yeast cell cycle.* Bioinformatics, 2004. **20**(12): p. 1914-27.

145. Moses, A.M., et al., *Position specific variation in the rate of evolution in transcription factor binding sites.* BMC Evol Biol, 2003. **3**(1): p. 19.

146. Kechris, K.J., et al., *Detecting DNA regulatory motifs by incorporating positional trends in information content.* Genome Biol, 2004. **5**(7): p. R50.

147. Karlin, S. and S.F. Altschul, *Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.* Proc Natl Acad Sci U S A, 1990. **87**(6): p. 2264-8.
148. Shpaer, E.G., et al., *Sensitivity and selectivity in protein similarity searches: a comparison of Smith-Waterman in hardware to BLAST and FASTA.* Genomics, 1996. **38**(2): p. 179-91.
149. Jaroszewski, L., L. Rychlewski, and A. Godzik, *Improving the quality of twilight-zone alignments.* Protein Sci, 2000. **9**(8): p. 1487-96.
150. Webber, C. and G.J. Barton, *Estimation of P-values for global alignments of protein sequences.* Bioinformatics, 2001. **17**(12): p. 1158-67.
151. Hickey, D.A., *Evolutionary dynamics of transposable elements in prokaryotes and eukaryotes.* Genetica, 1992. **86**(1-3): p. 269-74.
152. Jurka, J., *Repeats in genomic DNA: mining and meaning.* Curr Opin Struct Biol, 1998. **8**(3): p. 333-7.
153. Pirrotta, V., *Chromatin-silencing mechanisms in Drosophila maintain patterns of gene expression.* Trends Genet, 1997. **13**(8): p. 314-8.
154. Djordjevic, M., A.M. Sengupta, and B.I. Shraiman, *A biophysical approach to transcription factor binding site discovery.* Genome Res, 2003. **13**(11): p. 2381-90.
155. Ronen, M., et al., *Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics.* Proc Natl Acad Sci U S A, 2002. **99**(16): p. 10555-60.
156. Sengupta, A.M., M. Djordjevic, and B.I. Shraiman, *Specificity and robustness in transcription control networks.* Proc Natl Acad Sci U S A, 2002. **99**(4): p. 2072-7.
157. Bulyk, M.L., P.L. Johnson, and G.M. Church, *Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors.* Nucleic Acids Res, 2002. **30**(5): p. 1255-61.
158. Jones, E.G., *GABAergic neurons and their role in cortical plasticity in primates.* Cereb Cortex, 1993. **3**(5): p. 361-72.
159. Kosaka, T., et al., *GABAergic neurons containing the Ca2+-binding protein parvalbumin in the rat hippocampus and dentate gyrus.* Brain Res, 1987. **419**(1-2): p. 119-30.
160. Bormann, J., *The 'ABC' of GABA receptors.* Trends Pharmacol Sci, 2000. **21**(1): p. 16-9.
161. Rodwell, G.E., et al., *A transcriptional profile of aging in the human kidney.* PLoS Biol, 2004. **2**(12): p. e427.
162. Lund, J., et al., *Transcriptional profile of aging in C. elegans.* Curr Biol, 2002. **12**(18): p. 1566-73.
163. Wu, J., S. Kasif, and C. DeLisi, *Identification of functional links between genes using phylogenetic profiles.* Bioinformatics, 2003. **19**(12): p. 1524-30.
164. Segal, E., et al., *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.* Nat Genet, 2003. **34**(2): p. 166-76.
165. Bar-Joseph, Z., et al., *Computational discovery of gene modules and regulatory networks.* Nat Biotechnol, 2003. **21**(11): p. 1337-42.
166. Brivanlou, A.H. and J.E. Darnell, Jr., *Signal transduction and the control of gene expression.* Science, 2002. **295**(5556): p. 813-8.
167. Zhou, Q. and W.H. Wong, *CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling.* Proc Natl Acad Sci U S A, 2004. **101**(33): p. 12114-9.
168. Davidson, E.H., D.R. McClay, and L. Hood, *Regulatory gene networks and the properties of the developmental process.* Proc Natl Acad Sci U S A, 2003. **100**(4): p. 1475-80.
169. Davidson, E.H., et al., *A genomic regulatory network for development.* Science, 2002. **295**(5560): p. 1669-78.
170. Schroeder, M.D., et al., *Transcriptional control in the segmentation gene network of Drosophila.* PLoS Biol, 2004. **2**(9): p. E271.
171. Kalir, S. and U. Alon, *Using a quantitative blueprint to reprogram the dynamics of the flagella gene network.* Cell, 2004. **117**(6): p. 713-20.

172. Kanehisa, M., et al., *The KEGG resource for deciphering the genome.* Nucleic Acids Res, 2004. **32 Database issue**: p. D277-80.

173. Cormen, T.H. *Introduction to algorithms, second edition.* [Text] 2001 [cited; 2nd:[Available from: http://www.books24x7.com/marc.asp?isbn=0262032937 Click here for the electronic version.

174. Ihmels, J., S. Bergmann, and N. Barkai, *Defining transcription modules using large-scale gene expression data.* Bioinformatics, 2004. **20**(13): p. 1993-2003.

175. Ihmels, J., et al., *Revealing modular organization in the yeast transcriptional network.* Nat Genet, 2002. **31**(4): p. 370-7.

176. Mewes, H.W., et al., *Overview of the yeast genome.* Nature, 1997. **387**(6632 Suppl): p. 7-65.

177. Duggan, D.J., et al., *Expression profiling using cDNA microarrays.* Nat Genet, 1999. **21**(1 Suppl): p. 10-4.

178. Hughes, T.R., et al., *Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer.* Nat Biotechnol, 2001. **19**(4): p. 342-7.

179. Massingham, T., L.J. Davies, and P. Lio, *Analysing gene function after duplication.* Bioessays, 2001. **23**(10): p. 873-6.

180. True, J.R. and E.S. Haag, *Developmental system drift and flexibility in evolutionary trajectories.* Evol Dev, 2001. **3**(2): p. 109-19.

181. Zhang, Y. and J. Skolnick, *Automated structure prediction of weakly homologous proteins on a genomic scale.* Proc Natl Acad Sci U S A, 2004. **101**(20): p. 7594-9.

182. Kihara, D. and J. Skolnick, *Microbial genomes have over 72% structure assignment by the threading algorithm PROSPECTOR_Q.* Proteins, 2004. **55**(2): p. 464-73.

183. Galperin, M.Y. and E.V. Koonin, *Who's your neighbor? New computational approaches for functional genomics.* Nat Biotechnol, 2000. **18**(6): p. 609-13.

184. Koonin, E.V., et al., *A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes.* Genome Biol, 2004. **5**(2): p. R7.

185. Tatusov, R.L., et al., *The COG database: an updated version includes eukaryotes.* BMC Bioinformatics, 2003. **4**: p. 41.

186. Hubbard, T., et al., *Ensembl 2005.* Nucl. Acids Res., 2005. **33**(suppl_1): p. D447-453.

187. Loup, F., et al., *Selective alterations in GABAA receptor subtypes in human temporal lobe epilepsy.* The Journal Of Neuroscience: The Official Journal Of The Society For Neuroscience, 2000. **20**(14): p. 5401-5419.

188. Grünblatt, E., S. Hoyer, and P. Riederer, *Gene expression profile in streptozotocin rat model for sporadic Alzheimer's disease.* Journal of Neural Transmission, 2004. **111**(3): p. 367-386.

**List of references**
**Boris E. Shakhnovich**

**Dr. Eugene Koonin**
National Center for Biotechnology
Information (NCBI)
National Institutes of Health (NIH)
Bldg. 38A, Room 5N503
8600 Rockville Pike
Bethesda, MD 20894, USA
Tel: (301) 435-5913
Fax: (301) 435-7794 or (301) 480-9241
Email: koonin@ncbi.nlm.nih.gov

**Dr. Charles DeLisi**
Department of Biomedical Engineering
Boston University
44 Cummington Street
Boston, MA 02215
Tel: (617) 353-1122
Fax: (617) 353-4814
Email: delisi@bu.edu

**Dr. Jeffrey Skolnick**
Center for Excellence in Bioinformatics
University of Buffalo
901 Washington Street, Ste. 300,
Buffalo, NY 14203-1199
Tel: (716) 849-6711
Email: skolnick@buffalo.edu

**Dr. Minoru Kanehisa**
Bioinformatics Center, Institute for
Chemical Research
Kyoto University
Gokasho, Uji, Kyoto 611-0011, Japan
Tel: +81-774-38-3270
Fax: +81-774-38-3269
Email: kanehisalab@kuicr.kyoto-u.ac.jp

**Dr. Gilad Lerman**
Department of Mathematics
University of Minnesota
534 Vincent Hall
206 Church St. SE
Minneapolis, MN 55455
Tel: (612) 624-5541
Email: lerman@math.umn.edu

**Dr. Shamil Sunyaev**
Division of Genetics,
Harvard Medical School
New Research Building, Room 466B
77 Avenue Louis Pasteur,
Boston, MA 02115
Tel: (617) 525-4735
Fax: (617) 525-4705
email: ssunyaev@rics.bwh.harvard.edu