Debraj GuhaThakurta
Statement of Research Interests

Elucidation of transcription regulatory mechanisms and networks through the development and application of computational, statistical and experimental methods

Background and significance

Identification and annotation of all the functional elements in genomes, including genes and regulatory elements, is a fundamental challenge in this current era of biomedical research in which the complete genomic sequences of a number of prokaryotes and eukaryotes are available. Less than a third of the mammalian genome that is estimated to be under evolutionary selection pressure is coding (Chiaromonte et al, 2003; Waterston et al, 2002); the remaining is suggested to belong to untranslated regions, non-coding genes, chromosomal structural elements, and regulatory elements which control a variety of biological processes including gene expression, translation, chromosomal replication and condensation. However, in contrast to the advances made in the identification of protein coding sequences, identification of regulatory elements remains an unsolved problem in genome annotation. Whereas the number of genes in many of the sequenced prokaryotes and eukaryotes can now be reasonably estimated, there is no clear estimate of the number of functional regulatory elements in these genomes, especially in higher eukaryotes. The number of coding genes in eukaryotes, ranging from flies or worms to mammals, are not dramatically different (between ~14,000 to ~29,000), and it is now thought that organismal complexity may be attributed to phenomena like alternative splicing, DNA rearrangement, and increased number of transcription regulatory elements as well as transcription factors which regulate gene expression (Levine and Tjian, 2003). The identification of *cis*-regulatory elements controlling gene expression and the transcription factors (TFs), which bind to these elements, thus lie not only at the very heart of elucidating the network of gene interactions at the cellular level but also explaining the origins of organismal complexity and development. Only a small fraction of the eukaryotic transcription factors, DNA binding sites and regulated genes are currently known (TRANSFAC database, Matys et al, 2003). Therefore, identifying the transcription factors (TFs) and deciphering the function of the regulatory elements is likely to be an important research area in genomics and computational biology for years to come.

Identification and annotation of the regulatory elements in eukaryotes is a difficult problem. In higher eukaryotes and mammalian genomes the regulatory elements could be spread over a large area of sequence around the genes, sometimes over 100 Kb (Loots et al, 2000). Comparative sequence analysis is a promising approach that has been used successfully to identify regulatory elements and regions of conserved regulatory function (Cooper and Sidow, 2003). Despite the utility of comparative genomics, the identification and modeling of the individual *cis*-regulatory elements remains a challenging problem because transcription factor binding sites (TFBS) are usually short (~6 to 25 nucleotide) and degenerate sequences that are difficult to detect and align.

Computational and high-throughput experimental (e.g. ChIP-chip) methods for identification and modeling TFBS have been developed over the past two decades (Bulyk, 2003; Stormo 2000). Some of the computational methods have successfully been used to discover functional *cis*-regulatory elements in sequences that are thought to be regulated by a common transcriptional mechanism (e.g. GuhaThakurta et al, 2002; Huges et al, 2000). Where models for DNA binding sites are available, genome-wide predictions for the regulatory sites can be made (Chekmenev et al, 2005; Kel et al, 2003), but these usually suffer from a large false positive rate. In addition, many of the binding sites are functional *in vivo* only in certain temporal or spatial contexts (developmental or disease states, specific tissues etc.) with other nearby sites that form *cis*-regulatory modules (Davidson 2001; Howard and Davidson, 2004; Kel-Margoulis et al, 2002). This means individual binding sites predictions may not be biologically functional. Modeling composite DNA regulatory elements, which are the target sites for transcription factors that bind DNA in a synergistic manner, is therefore essential in understanding eukaryotic transcription regulation. A few computational and statistical methods have been developed recently for the identification of *cis*-regulatory modules (e.g. Aerts, et al, 2003; Frith et al, 2003; GuhaThakurta and Stormo, 2001; Johansson et al, 2003).

These methods have been valuable in expanding our knowledge of transcription factor binding sites in the genome. However, significant obstacles still exist in elucidation of regulatory elements in higher eukaryotes and mammalian genomes since the sequence space in which these elements may exist is large and the complexity of regulatory protein-DNA interactions is likely to be higher. Even when regulatory elements are identified, methods are needed to determine which elements are functional under specific temporal or spatial contexts *in vivo* and to identify the TFs which bind to these sites. Hence both computational and experimental methods to address the problem of transcription regulatory element detection, modeling, and annotation will need to be developed and improved.

One of the most important benefits of characterizing the *cis*-regulatory elements is their use in construction of gene regulatory networks. Several studies have utilized the information on *cis*-regulatory elements for network reconstruction,..

either in isolation or in combination with other orthogonal sources of information, e.g. microarray expression data (Bolouri and Davidson, 2002; Haverty et al, 2004; Pilpel et al, 2001). Recently, expression data has also been combined with genetic variation in segregating populations to create regulatory networks (Zhu et al, 2004). These approaches demonstrate how multiple sources of data can be integrated to create gene networks, not only in unicellular organisms but complex multicellular organisms. Although promising, these methods are new and there is significant need for investigation, improvement, and validation.

Research directions in the lab

My lab would be involved in the investigation of transcription regulatory mechanisms and networks in eukaryotes through computational and experimental methods. Experimental validation of biological hypothesis made through computational work will be a key to elucidation of the regulatory mechanisms. Hence, my lab would develop the experimental methods that would compliment our computational research. In addition to validation work, the experimental arm of the lab would be utilized to generate raw data that would be used in training the computational methods. Lessons learned from experiments would provide valuable insights on how to improve the computational methods. The problems that my lab would like to address are:

i) Development and application of computational, statistical, and experimental methods for identification of functional transcription regulatory modules.

There is a considerable need for further research this area. We have previously developed a computational method for identification of composite regulatory elements (GuhaThakurta and Stormo 2001) which can be improved and applied to different data-sets followed by experimental validation to determine the performance.

ii) Identification of genomic binding sites and the set of regulated genes for transcription factors.

Several strategies can be employed in this area, including:

(a) Computational predictions followed by experimental validation using mRNA profiling or transgenic studies *in vitro* or *in vivo*.

(b) Perturbation of transcriptional pathways in model organisms or cell-lines (using RNAi mediated knockdown or other treatments), followed by expression profiling. The genes with altered mRNA levels at early time points are likely candidates for direct regulation by the transcription factors. Application of DNA pattern recognition methods on these genes could define the binding sites for the perturbed transcription regulators (e.g. GuhaThakurta et al, 2002).

iii) Reconstruction of transcription regulatory networks by combining data from mRNA profiling, *cis*-regulatory elements and genetic variation of gene expression in segregating populations.

DNA sequence variations affecting gene expression can be thought of as natural perturbations occurring in a segregating population which can be utilized to make inferences about gene networks (Schadt et al, 2005; Zhu et al, 2004). Expression levels can be treated as quantitative traits and mapped to the genome using standard linkage analysis methods. DNA-variation and expression data can thereby be integrated to order the genes in an acyclic network graph. The underlying principle is if two variables (e.g. expression levels of two genes) in a segregating population are genetically linked to the same locus on the genome, then the one with stronger linkage is causal for (in other words, upstream in the network relative to) the one with weaker linkage (after having controlled for other necessary factors). This has been applied to specific tissues in mouse, where doing crosses and breeding animals are expensive and time-consuming. The method can also be applied to elucidate regulatory networks in simpler organisms like *S. cerevisiae* or *C. elegans*, where performing crosses and *in-vivo* manipulations is much easier in terms of the methods, time and cost. Genetic variation of gene expression in *S. cerevisiae* has been studied recently in some detail (Brem and Kruglyak, 2005), hence it can be used as a model to study unicellular organisms. Linkage mapping has been done in *C. elegans* before using recombinant inbred lines (Shook et al, 1996; Ayyadevara et al, 2001), and progress has been made in high-throughput characterization of gene expression in specific tissues (Roy et al, 2002; Zhang et al, 2002). Tissue specific expression profiling in *C. elegans* is still challenging and has limitations in terms of quantitative accuracy. However if the methods can be improved to obtain quantitatively reproducible tissue-specific expression profiles, then the genetic variation of gene expression, combined with the information from *cis*-regulatory elements, can be a powerful tool for elucidation of regulatory networks in simple multicellular organisms. Therefore, in addition to mouse, we will investigate simpler model organisms in our laboratory. I have been directly involved in the research and development in this field at Rosetta Inpharmatics, hence our lab would be in a good position to work in this new and promising area of network reconstruction research.

All of the computational and much of the experimental work, including that with transgenics and simple model organisms, could be done in our lab. I have considerable experience working on TFBS and *cis*-regulatory modules through the analysis of sequence and expression data (GuhaThakurta et al, 2002; Pramila et al, 2002). I have developed software and statistical methods for detection of composite DNA regulatory elements (GuhaThakurta and Stormo, 2001; GuhaThakurta et al, 2002; GuhaThakurta et al 2004). I also have prior experience in experimental characterization of protein-nucleic interactions and structural aspects of protein-nucleic acid recognition (Dunstan et al, 2005; GuhaThakurta et al, 2000; GuhaThakurta et al, 1999; Xing et al, 1997), and elucidating regulatory mechanisms in *C. elegans*. As done previously, we would pursue research collaborations with other laboratories that are interested in specific regulatory pathways. This would help in characterizing a few key pathways in greater depth and detail, information from which would be utilized in improving the computational methods.

Identification of causal genes and polymorphisms for complex diseases using integrative genomics approaches

Background and significance

The true value of sequencing the human genome is realized when causal genes for diseases that affect human health and life are identified through genomic and genetic approaches. The availability of numerous polymorphic markers makes it possible to perform genome-wide linkage analysis to identify disease loci in humans as well as animal models like the mouse and rat. Since its introduction about a decade ago, the DNA microarray technology has revolutionized both basic and applied biomedical research (Shoemaker and Linsley 2002). Although extremely valuable, both linkage and microarray methods have limitations for identification of causal disease genes as explained below. Often the linkage regions on the genome are large containing hundreds of genes, from which identifying the causal gene by fine-mapping or positional cloning is a difficult and time-consuming undertaking (e.g. Encinas and Kuchroo, 2000; Symula et al, 1999). Comparison of mRNA profiles between normal and disease states can identify genes that are associated with the disease but cannot distinguish the genes that are causal for the disease from the ones that are reactive to (i.e. changes in response to) the disease.

Recently, genetic linkage mapping and gene expression profiling have been integrated to distinguish the causal versus the reactive genes for complex diseases (Schadt et al, 2005). As mentioned earlier, the causal or reactive relationship between two variables, say a disease trait and a gene-expression profile, that share common genetic linkage, can be established by systematically testing whether the variation at the linkage loci leads to change in the expression that causes the disease. This appears to be a powerful technique for the identification of causal genes for complex diseases.

The central dogma of genetics is that genomic polymorphisms underlie susceptibility to (or protection from) diseasess by affecting biological processes at the molecular level including protein structure, transcription, and splicing. Linkage mapping is frequently performed to associate disease susceptibility to genomic loci which is sometimes followed by identification of causal genes through positional cloning and fine-mapping. However, the more fundamental question of what the specific sequence polymorphisms are that qualitatively or quantitatively perturb the molecular functions of those causal genes is frequently left unanswered. With the full sequencing of several mammalian genomes and efforts to characterize intraspecific sequence variations (The International HapMap Consortium, 2003; Lindblad-Toh et al, 2000), this fundamental question in genetics can now be addressed.

Quantitative genetic variation in gene expression has been used fruitfully in many ways, e.g. to identify causal genes for complex diseases (Schadt et al, 2005), segregate populations into disease subtypes (Schadt et al, 2003), and build networks of gene interactions (Zhu et al, 2004). There are now several examples where polymorphisms in the promoter regions, and those causing expression changes in specific genes, are associated with diseases (Pastinen and Hudson 2004; Knight et al, 2005). In order to understand the molecular basis of many diseases it will therefore be important to identify and annotate the causal regulatory polymorphisms. It is estimated that a sizeable fraction of promoter polymorphisms in human genes may affect expression (Buckland et al, 2005; Hoogendoorn et al, 2003). However, the patterns of polymorphisms which underlie heritable variation of gene expression, and methods for isolating the causal regulatory polymorphisms starting from linkage data have not been investigated in any detail.

Research directions in the lab

My lab would focus on the following problems:

i) Identification of novel causal genes for diseases by integrating data from molecular profiling and genetics.

ii) Characterizing causal regulatory polymorphisms for gene expression variation and disease.

I would be keen in collaborating with other laboratories that are studying the genetic causes of diseases. The experimental part of performing the intercrosses (in case of mouse), sample collection, clinical trait measurement and genotyping would be done in the collaborating lab (or through a vendor). I have been intimately involved in the investigation and development of the methods for identification of causal genes using genetics and profiling approaches. Our lab would therefore be fully equipped to undertake the statistical data analysis, database development, and identification of causal candidates. There are several avenues for improvements in the analysis methods, which would be pursued in our lab.

We would invest in developing methods to identify and annotate polymorphisms that cause variations in mRNA levels by affecting the transcription regulatory machinery. This will lead to insights into the molecular basis of how regulatory polymorphisms predispose individuals towards disease. Our expertise in characterization of *cis*-regulatory elements will be helpful in pursuing such studies.

Reference

Aerts S,V an Loo P, Thijs G, Moreau Y, De Moor B, Computational detection of *cis*-regulatory modules. (2003) *Bioinformatics*,**19** , Suppl 2:II5-II14.

Ayyadevara S, Ayyadevara R, Vertino A, Galecki A, Thaden JJ, Shmookler Reis RJ, Genetic loci modulating fitness and life span in *Caenorhabditis elegans*: categorical trait interval mapping in CL2a x Bergerac-BO recombinant-inbred worms. (2003) *Genetics*, **163**, 557-570.

Bolouri H and Davidson EH, Modeling DNA sequence-based cis-regulatory gene networks. (2002) *Developmental Biol.*, **246**, 2-13.

Brem RB and Kruglyak L, The landscape of genetic complexity across 5,700 gene expression traits in yeast. (2005) *Proc Natl Acad Sci USA*, **102**, 1572-1577.

Buckland PR, Hoogendoorn B, Coleman SL, Guy CA, Smith SK, O'Donovan MC, Strong bias in the location of functional promoter polymorphisms. (2005) *Hum Mutat.*, **26**, 214-223.

Bulyk ML, Computational prediction of transcription-factor binding site locations. (2003), *Genome Biol*,**5** , 201.

Chekmenev DS, Haid C, Kel AE, P-Match: transcription factor binding site search by combining patterns and weight matrices. (2005) *Nucleic Acids Res*,**33** , W432-437.

Chiaromonte F, Weber RJ, Roskin KM, Diekhans M, Kent WJ, Haussler D., The share of human genomic DNA under selection estimated from human-mouse genomic alignments. *Cold Spring Harb Symp Quant Biol.*, (2003) **68**, 245-254.

Cooper GM and Sidow A, Genomic regulatory regions: insights from comparative sequence analysis. (2003) *Curr Opin Genet Dev.*, **13**, 604-610.

Davidson EH, Genomic regulatory systems: Development and evolution, *Academic Press, New York*, (2001).

Dunstan MS, Guhathakurta D, Draper DE, Conn GL, Coevolution of protein and RNA structures within a highly conserved ribosomal domain. (2005) *Chem Biol*, **12**, 201-206.

Encinas JA and Kuchroo VK, Mapping and identification of autoimmunity genes. (2000) *Curr Opin Immunol.*, **12**, 691-697.

Frith MC, Spouge JT, Hanse U and Weng Z; Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. (2002) *Nucleic Acids Res*, **30**, 3214-3224.

GuhaThakurta D and Draper DE, Protein-RNA sequence covariation in a ribosomal protein-rRNA complex. (1999) *Biochemistry*, **38**, 3633-3640.

GuhaThakurta D and Draper DE, Contributions of basic residues to ribosomal protein L11 recognition of RNA. (2000), *J Mol Biol.*, **295**, 569-580.

GuhaThakurta D and Stormo GD, Identifying target sites for cooperatively binding factors. (2001) *Bioinformatics*,**17** , 608-621.

GuhaThakurta D, Palomar L, Stormo GD, Tedesco P, Johnson TE, Walker DW, Lithgow G, Kim S, Link CD, Identification of a novel cis-regulatory element involved in the heat shock response in *Caenorhabditis elegans* using microarray gene expression and computational methods. (2002) *Genome Res.*,**12** , 701-712.

GuhaThakurta D, Schriefer LA, Waterston RH, Stormo GD, Novel transcription regulatory elements in *Caenorhabditis elegans* muscle genes. (2004) *Genome Res.*,**14** , 2457-2468

Howard ML and Davidson EH, *cis*-Regulatory control circuits in development. (2004) *Developmental Biol*, 109-117.

Haverty PM, Hansen U, Weng Z, Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. (2004) *Nucleic Acids Res.*, **32**, 179-188.

Hughes JD, Estep PW, Tavazoie S, Church GM, Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae.(*2000) *J. Mol. Biol.*, **296,** 1205-1214.

Hoogendoorn B, Coleman SL, Guy CA, Smith K, Bowen T, Buckland PR, O'Donovan MC., Functional analysis of human promoter polymorphisms. (2003) *Hum Mol Genet.*, **12**, 2249-2254.

Johansson O, Alkema W, Wasserman WW, and Lagergren J, Identification of functional cluster of transcription factor binding sites in genome sequences: The MSCAN algorithm. (2003) *Bioinformatics*, **19,** i169-i176.

Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E, MATCH: A tool for searching transcription factor binding sites in DNA sequences. (2003) *Nucleic Acids Res*, 31, 3576-3579.

Kel-Margoulis OV, Kel AE, Reuter I, Deineko IV, Wingender E, TRANSCompel: a database on composite regulatory elements in eukaryotic genes. (2002) *Nucleic Acids Res.*, **30**, 332-324.

Knight JC, Regulatory polymorphisms underlying complex disease traits (2005) *J. Mol Med*,**83** , 97-109.

Levine M and Tjian R, Transcription regulation and animal diversity. (2003) *Nature*, **424**, 147-151.

Lindblad-Toh K et al, Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. (2000) *Nat Genet.*, **24**, 381-36.

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA, Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. (2000) *Science*, **288**, 136-140.

Matys V et al, TRANSFAC: transcriptional regulation, from patterns to profiles. (2003) *Nucleic Acids Res*, **31**, 374-378.

Pastinen T, Hudson TJ, Cis-acting regulatory variation in the human genome. (2004) *Science*,**306** , 647-650.

Pilpel Y, Sundarsanam P, Church GM, Identifying regulatory networks by combinatorial analysis of promoter elements. (2001) *Nat. Genet.*, **29**, 153-159.

Pramila T, Miles S, GuhaThakurta D, Jemiolo D, Breeden LL, Conserved homeodomain proteins interact with MADS box protein Mcm1 to restrict ECB-dependent transcription to the M/G1 phase of the cell cycle. (2002) *Genes Dev*,**16** , 3034-3045.

Roy PJ, Stuart JM, Lund J and Kim S, Chromosomal clustering of muscle-expressed genes in *C. elegans*. (2002) *Nature*,**418** , 975-979.

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH, Genetics of gene expression surveyed in maize, mouse and man. (2003) *Nature*,**422** , 297-302.

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L, Castle J, Zhu H, Kash SF,Drake TA,Sachs A, Lusis AJ, An integrative genomics approach to infer causal associations between gene expression and disease. (2005) *Nat Genet*, **37**, 710-717.

Shoemaker DD and Linsley PS, Recent developments in DNA microarrays. (2002) *Curr Opin Microbiol*, **5**, 334-337.

Shook DR, Brooks A, Johnson TE, Mapping quantitative trait loci affecting life history traits in the nematode *C. elegans*. (1996) *Genetics*, **142**, 801-817.

Stormo GD, DNA binding sites: representation and discovery. (2000) *Bioinformatics*,**16** , 16-23.

Symula DJ, Frazer KA, Ueda Y, Denefle P,Stevens ME, Wang ZE, Locksley R, Rubin EM Functional screening of an asthma QTL in YAC transgenic mice. (1999) *Nat Genet.*, **23,** 241–244.

The International HapMap Consortium, The International HapMap Project. (2003) *Nature*, **426**, 789-796.

Xing Y, Guha Thakurta D, Draper DE, The RNA binding domain of ribosomal protein L11 is structurally similar to homeodomains. (1997) *Nat Struct Biol*,**4** , 24-27.

Waterston RH et al, Initial sequencing and comparative of the mouse genome. *Nature*, (2002) **410**, 520-562.

Zhang Y, Ma C, Delohery T, Nasipak B, Foat BC, Bounoutas A, Bussemaker HJ, Kim SK, Chalfie M, Identification of genes expressed in *C. elegans* touch receptor neurons. (2002) *Nature*,**418** , 331-335.

Zhu J, Lum PY, Lamb J, GuhaThakurta D, Edwards SW, Thieringer R, Berger JP, Wu MS, Thompson J, Sachs AB, Schadt EE, An integrative genomics approach to the reconstruction of gene networks in segregating populations. (2004) *Cytogenet Genome Res*, **105**, 363-374.

A STATEMENT OF TEACHING PHILOSOPHY AND INTERESTS

DEBRAJ GUHATHAKURTA

Philosophy

Today, education in a wide diversity of scientific areas is critical in order to have a successful career in research, teaching, management and leadership in biomedical sciences. In addition to the basic principles of molecular biology, cell biology and genetics, a biomedical scientist needs to have a reasonably good understanding of genome science, computational biology, basic statistics, and the familiarity with programming plus a wide variety of analyses software, to generate useful hypotheses and make discoveries. As revolutionary new discoveries and methods change the way future scientific discoveries are made, the subject matter being taught in class at all levels needs to evolve to appropriately prepare the students for their careers.

My philosophy as an educator is not only to provide the students with the good foundation they need in the basics of biological and genome science, but also to evolve the materials being taught depending on the breakthrough discoveries. I am keen to impart the research ideas that are likely to re-shape the landscape of biological science in the years to come. I will also strive to give the students a clear view of the 'big picture' and convey to them, in a transparent way, how the material they are learning in the classroom can help them in making critical contributions to improving human health and life.

Teaching Interests

The fields of genomics and bioinformatics currently encompass a vast area. Being interdisciplinary in nature, they can be taught in many different departments and specializations within an institute. From an information technology perspective, bioinformatics involves the development of database and software tools for managing, curating, retrieving, and visualizing biomedical data. On the other hand, from a scientific perspective, it provides the means for making important discoveries, and generating testable hypotheses through statistical and computational analysis of biological data. I am interested in teaching courses that provide the students the opportunity to learn how to make discoveries and inferences through statistical and computational analysis of complex biological data. I will strive to convey to the students the interdisciplinary nature of genomic science through key examples that utilizes the practices and methods from disparate scientific disciplines like genetics, molecular biology, statistics, artificial intelligence, etc.

Since biomedical science is now critically dependant on high-throughput technologies and the analyses of large quantities of data, there is a need for students, at all levels, who are appropriately trained in genome science and bioinformatics. Courses could be offered at undergraduate and graduate levels, with general material (overview of methods and examples) presented at lower undergraduate levels and specialized material presented at senior undergraduate or graduate levels. I am committed to teaching both graduate and undergraduate courses, in order to prepare the students at all levels.

Below are a few specific ways I would like to contribute to the teaching programs in an institution:

i) Teaching a general course to undergraduates which provides an overview of the basic computational biology and genomic methods.

ii) Teaching a specialized course in genome sequence analysis, gene regulation and networks.

iii) Teaching sections of a course where the instructors are brought together from different departments or research specializations. In such a course, I would be interested in teaching computational methods for modeling transcription regulation, methods for sequence analysis, basics of genetic linkage mapping, or the application of genetics and gene-expression profiling in biomedical science.

iv) Coordinating a seminar course which reviews key findings and discoveries from recent literature in genomics and computational biology.

Although the development and applications of the methods for genome analysis are best described in original research and review articles, there are now several good text books available to introduce these materials is a systematic way in the classroom (e.g.: http://www.iscb.org/bioinformaticsBooks.shtml). Some of these text books could be used in the courses along with the research and review articles in specific areas.

Finally, since genomics and computational biology are essentially interdisciplinary fields, I would be eager to collaborate with other professors within the institute (both intra- and inter-departmental), and identify the best ways to train the students for successful careers in the biomedical field.

A Statement of Past Research Accomplishments

Debraj GuhaThakurta

I have been involved in research in genomics, computational and molecular biology, and biophysical chemistry over the past eleven years. One of my key strengths is the understanding and experience in a diverse array of research areas. Especially, my experience in both computational and experimental biology has helped me to be effective in developing methods and making discoveries through the iterative use of both strategies. Some of the scientific problems that I have addressed over the past several years and the key achievements are described below.

The identification and annotation of the *cis*-regulatory elements controlling gene-expression lie not only at the very heart of elucidating the network of gene interactions at the cellular level but also explaining the origins of organismal complexity and development. However, to date this remains largely an unsolved problem genome annotation, especially in higher eukaryotes. I have been involved in addressing this important and challenging problem for the past several years. Using computational and experimental methods, I have identified and elucidated the function of novel regulatory elements in eukaryotic genomes. Through collaborative efforts with laboratories, I discovered regulatory elements which guide expression in the muscle tissue, in heat-shock treatment, and in cell cycle. I have developed software and statistical methods for detection of composite DNA regulatory elements, which is a hallmark of eukaryotic gene regulation. Specifically, I have developed a novel computational method for identification of DNA binding sites for cooperatively acting transcription factors. It is one of the first to computationally detect novel composite regulatory elements using a machine learning approach. Since its publication and release, this software has been widely used and referenced by other scientists in the field. These studies demonstrate how novel regulatory elements for specific cellular transcriptional pathways can be identified with computational methods and validated using transgenic strategies.

Prior to the development of computational and statistical methods for analysis of gene regulation, I studied structural aspects of protein-nucleic acid recognition. Protein-RNA complexes are crucial in performing many necessary cellular functions, including translation of message to protein. However, not many protein-RNA complexes are well characterized at the structural level, and the mechanisms of protein-RNA recognition is poorly understood in comparison to the more thoroughly studied protein-DNA recognition. I was intrigued by this problem and decided to pursue it during my graduate studies. Ours was one of the first studies to demonstrate that evolutionary co-variation of nucleotide and amino-acid residues can be utilized to identify key structural interactions between RNA and protein in a complex. In addition, I did the first extensive study investigating how residues on a ribosomal protein contribute to ribosomal RNA recognition in a conserved protein-nucleic acid complex. These studies shed light on the structural aspects of protein-RNA recognition.

At Rosetta Inpharmatics LLC (Merck & Co.), where I am currently employed, I have been a part of a research team that pioneered the use DNA microarrays and genetic variation of gene expression in identification of disease genes and building regulatory networks. I have been involved in two main areas of research and development, viz. annotation of the coding genes in the human genome with microarray technology and target gene identification for complex human diseases using a combination of mouse genetics and molecular profiling. I contributed significantly to the development of novel methodologies for analyses, most of which have been published in reputed journals. I have successfully used integrative genomic techniques (integrating genomic, molecular profiling and genetics data) to identify new drug targets for cardiovascular disease and obesity. Several of these targets have entered the pipeline for extensive validation and drug development with high-throughput screening. My colleagues and I have also utilized the genetic variation of gene expression to reconstruct regulatory network in mouse tissues. I developed and maintained databases for genomic, genetic and profiling data, and wrote programs for data analysis in Perl, R, and Matlab. In addition, I have used available software like JMP and Spotfire for data analysis. Thus, at Rosetta, I have become familiar with the pharmaceutical R&D process, and obtained significant experience in integrative genomics, and aspects of data management and analysis in a collaborative research environment.

While at Rosetta Inpharmatics, I served as a thesis advisor to one MS student from Indiana University School of Informatics. He worked as an extended intern under my guidance to investigate an intriguing issue in genetics, viz. the patterns of polymprhisms that underlie genetic variation of gene-expression in segregating populations. At Rosetta, I have directly supervised two employees including one Ph.D. senior scientist. In addition, during my post-doctoral work at Washington University, I had the opportunity to supervise graduate and undergraduate students in their internships or laboratory rotation.

Due to my expertise and experience in various areas of computational and molecular biology, I am frequently asked to serve as reviewer in reputed journals. I have also been invited to review international research grant applications in computational biology and genomics. I have served as an associate guest editor for two special issues of the IEEE Intelligent Systems magazine on bioinformatics.

I am confident that my experience in various aspects of computational and molecular biology, along with the ability to collaborate, and guide students and employees will help me to establish a vibrant and successful research laboratory.