Xu Y. Computational challenges in deciphering genomic structures of bacteria. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 25(1): 53–70 Jan. 2010

Computational Challenges in Deciphering Genomic Structures of Bacteria

Ying Xu (徐 鹰)

Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, The University of Georgia, Athens, GA 30602, U.S.A.

BESC BioEnergy Science Center, U.S.A.

College of Computer Science and Technology, Jilin University, Changchun 130012, China

E-mail: xyn@bmb.uga.edu

Received October 1, 2009; revised November 16, 2009.

Abstract This article addresses how the functionalities of the cellular machinery of a bacterium might have constrained the genomic arrangement of its genes during evolution and how we can study such problems using computational approaches, taking full advantage of the rapidly increasing pool of the sequenced bacterial genomes, potentially leading to a much improved understanding of why a bacterial genome is organized in the way it is. This article discusses a number of challenging computational problems in elucidating the genomic structures at multiple levels and the information that is encoded through these genomic structures, gearing towards the ultimate understanding of the governing rules of bacterial genome organization.

Keywords bioinformatics, microbial genomics, genome structure, comparative genome analysis, biological pathways

1 Cellular Machinery and Genome

Bacteria are the simplest free-living organisms on Earth. They are unicellular. A typical cell (see Fig.1(a)) has a diameter of ∼0.5 micrometer, about one twentieth of the diameter of a human cell. While tiny, they are equipped with rather sophisticated capabilities for doing some amazing things. A bacterial cell consists of multiple cellular systems for carrying out the basic house-keeping functions such as DNA replication during multiplication of a cell, gene transcription and translation for the generation of functional molecules such as proteins and RNAs, and metabolic processes to handle nutrient processing and energy conversion in support of its basic needs as a living organism. On top of these, a bacterial cell also has certain information processing capabilities such as environmental sensing, signaling and circadian rhythms, as well as a complex network of regulatory elements that controls the timing and the amount of different types of biomolecules to be created under specific conditions. In a complex environment where the living conditions such as the temperature, the pH level and the amount of available nutrients constantly change, the bacteria there have learned to adapt to the changing environment by turning on and off the relevant cellular systems to best cope with the current environmental condition. Some bacterial cells are capable of moving themselves by using their flagella. Like all living organisms, bacterial cells can reproduce, often at an astonishing rate. Bacterial reproduction is generally asexual, which involves only one parent and the offspring are exact replicates of the parent.

While very small, bacteria play many essential roles in keeping our environment livable as well as in keeping each of us healthy, among many other things that they are capable of doing. For example, cyanobacteria, the largest as well as perhaps one of the oldest groups of bacteria on Earth, is considered the architect of our atmosphere since they were responsible for creating the oxygen atmosphere under which our lives and many other forms of lives are on going. For each healthy human being, we have approximately 10^{14} bacterial cells inside our intestines, some of which help us to process certain foods, such as vegetables, that are otherwise not (easily) digestible by ourselves among other functions they do inside our intestines.

While individual bacterial cells can execute very complex biological functions, bacterial communities

Survey

The work is supported in part by the NSF of USA (Grant Nos. DBI-0354771, ITR-IIS-0407204, DBI-0542119, CCF0621700), NIH of USA (Grant Nos. 1R01GM075331 and 1R01GM081682) and the grant for the BioEnergy Science Center.

could do substantially more. For example, it is known that bacterial cells can self-organize into hierarchically structured colonies under natural growth conditions[1-4]. Fig.1(b) shows two such complex patterns of bacterial colonies, which emerge through interplays among individual bacteria coordinated through their inter-cellular communications. While much is yet to be understood about the formation of such complex patterns, it is known that the bacterial colonies provide the degree of plasticity and flexibility required for better durability and adaptability of a whole bacterial community to the changing environment.

As we understand now, all these capabilities are encoded in the genome of the bacterial cell. Specifically, the bacterial genome encodes the components of the cellular machinery capable of carrying out all these capabilities, along with the regulatory information about the conditions under which each such component will be activated. A typical bacterial genome ranges from 1 million base pairs (bps) of nucleotides to ∼10 million bps long. The genome of the widely known *E. coli* K12 is ∼4 million bps long and encodes ∼4000 protein genes. In addition to the protein genes, a bacterial genome may also encode functional RNA genes, called *non-coding RNAs* although we do not have a clear idea yet about the number of such RNA genes that a typical bacterial genome may encode. Under a particular cellular condition, a specific group of genes may be activated to produce protein and/or RNA molecules through a chain of events such as sensing, signaling, regulation, transcription (and translation), and together they will carry out their designed functions after being transported to the desired sub-cellular locations. Then the physical and chemical laws on Earth will take over so these functional molecules will fold into the correct structural conformations, form molecular associations, say, to activate proteins through post-translational modifications such as phosphorylation, to catalyze chemical reactions, or to transport ions in or out of a cell. The combined effect of these molecular functions and interactions could be some rather complex cellular functions such as converting nutrient transported from the environment to some form of energy that the cell can utilize. In order to keep a cell to function properly, all individual molecular functions have to be done in the proper order, at the right time, under the right conditions and for the right duration of time, controlled by the relevant regulatory information that is also encoded in the genome of the bacterial cell.

1.1 What is Known About Bacterial Genomes in General

Haemophilus influenzae was the first sequenced bacterial genome, completed in 1995[7]. Since then, over 879 bacterial genomes have been sequenced (as the writing of this article), and ∼2000 additional bacterial genomes are in the pipeline being sequenced or to be sequenced[8]. A bacterial genome is typically circular, and encodes proteins on both the leading and the lagging strands of its chromosome[9-10]. On average, a bacterial genome encodes one protein gene per ∼1000 bps. Bacterial genes are organized as *operons*[11-14], the basic transcriptional units, and then further grouped into a higher level organization called *regulons*[15-17], the basic units of the cellular response system. In addition, a genome encodes various regulatory elements such as *cis* regulatory motifs for transcription regulation. Bacterial genomes are constantly changing their composition and organization as they may move some of their genetic material from one location to another within a genome or exchange genetic material with other organisms. Such genetic material is generally called *mobile genetic elements*[18-23], which provide the basic material for genomes to gain new functionalities during evolution.

Numerous computational techniques have been developed to characterize and identify these and other biological entities encoded in a bacterial genome, some

Fig.1. (a) A cartoon of bacterial cell[5]. (b) Patterns formed by *Paenibacillus dendritiformis* bacteria[6].

of which have matured to a level that they are currently being widely used for studying bacterial genomes.

Protein-Encoding Genes. Protein-encoding genes account for ∼80% of a typical bacterial genome. Identification of such genes in a genome represents one of the areas where computational techniques have played a key role. The basic information for identification of protein-encoding genes in a genome is that the protein-encoding regions generally have distinguishing di-codon (two codons with each consisting of three nucleotides) frequencies[24] from those in non-coding regions. In addition, the start of a protein-encoding gene also has detectable signals[25-26]. By combining these two pieces of information, various computational techniques have been developed to find genes in a sequenced bacterial genome. The best gene-prediction programs can reach over 90% of identification accuracy at the nucleotide level[27].

Functions of Protein-Encoding Genes. The functionality of a protein is not directly encoded in the genome; instead it is determined by the three-dimensional shape of its folded structure, which is ultimately decided by its amino acid sequence and the physical laws on Earth. Numerous computational techniques have been developed for prediction of molecular and cellular functions of proteins, mostly based on comparison to proteins with known (experimentally characterized) functions. Sequence homology-based approaches represent the most widely-used class of methods for protein function prediction. Structural homology-based approaches, such as protein threading[28-29], allow identification of more distant homologues that are not easily identifiable by sequence-based approaches. In addition, there are prediction methods based on functional motifs in protein sequences[30-32]. The current estimate is that ∼70% of the proteins encoded in a newly sequenced bacterial genome can have some level of functional prediction[33] using computational techniques.

Non-Coding RNA Genes. Non-coding RNA (ncRNA) genes are involved in a variety of cellular processes ranging from regulation of gene transcription to genome modification and editing[34-36]. Various types of ncRNA genes have been observed in bacterial genomes except for microRNAs[37]. It has been observed that homologous RNA genes are generally not as well conserved at the sequence level as their protein counterparts while their secondary structures are generally better conserved, suggesting that the key functional information is encoded in the secondary structures. A number of computer programs have been developed for identification of ncRNA genes using both sequence and secondary structure conservation information[38-41]. This class of methods is mostly limited to prediction of ncRNA genes whose homologous genes have been characterized before.

Operons. Unlike eukaryotic genomes, bacterial genes are organized into operons[11-12], in which genes are arranged in tandem on the same strand of a genome and share a common promoter and a common terminator. The average operon size in a bacterial genome is 2∼3 genes. Genes in the same operon generally encode proteins that work in the same biological process, and hence are functionally related. Various observations have been made about genes of the same operons. For example, the distance between adjacent genes within an operon tends to be shorter than the distance between adjacent genes in adjacent operons; the close proximity relationships among genes in the same operons are often kept across multiple (related) genomes. Based on these and other observations, a number of computer tools have been developed to predict operons[13-14,42-46]. Currently the best prediction programs have slightly better than 80% accuracy in predicting if two adjacent genes are in the same operon.

Promoter Regions and cis Regulatory Elements. The transcription of a gene, the first step in making a gene into its functional form, is initiated through having the RNA polymerase (RNAP) bind to the promoter of the gene (actually its operon) while such binding generally requires a sigma factor[47-49] and additional regulatory proteins to recruit the RNAP to the right location. Such regulatory proteins, called *transcription regulators*, will first bind to their *cis regulatory elements* around the promoter and then recruit the RNAP. Operons that are regulated by the same transcription regulator generally have *cis* regulatory elements with similar sequences, i.e., these *cis* elements or motifs are *conserved* at the sequence level. So identification of such *cis* regulatory elements has been generally done through identification of conserved short sequences across the upstream regions of operons considered to be transcriptionally co-regulated. While scientists have been working on this problem since the late 1980's, the problem remains an unsolved and challenging problem. The best programs will have no better than 50% of prediction accuracy for large scale applications in bacterial genomes[50-51].

Mobile Elements. Mobile genetic elements (MGEs) can be broadly defined as regions of a genome that are able to move themselves within a genome. A dominating class of the MGEs is the *transposable elements*. Some genes can move from one organism to another, called *horizontally transferred genes* (HTGs). These mobile elements form the basis for gene transfer, genomic rearrangement or deletion during the evolution of a genome. Various computational techniques have

been developed for characterizing and identifying these mobile elements[19-20,52-54]. While the prediction of the MGEs has reached a good level of maturity, reliable prediction of HTGs remains a very challenging problem[55-58].

1.2 What Is Not Known About Bacterial Genomes

We can consider a bacterial cell as a complex machine consisting of three intertwined systems: (i) a metabolic system that carries out all the essential cellular functionalities, (ii) a regulatory system that determines under what conditions to carry out which functionalities, and (iii) a signaling system that detects "need attention" signals inside the cell or in the surrounding of the cell, transmits signals across different regulatory elements and changes the functional states of the relevant components of the metabolic system. Each system is made of interacting molecules, mostly proteins, DNAs and RNAs plus other molecules such as lipids and small molecules like water. While the house-keeping molecules are constitutively active to maintain the basic functions of a cell, most of the functional molecules are only activated upon request. In a sense, the genome of a cell can be viewed as an instruction book about the composition and the organization of each functional molecule and the regulatory switch that controls the timing of the activation of the molecule and its abundance. The switch gets turned on and off by the cell releasing specific regulatory protein(s) to its vicinity, typically triggered by a chain of signaling events executed through a sequence of interactions among signaling molecules. Essentially, a cell is like a dynamically assembled machine with a constantly changing structure, which is determined by the interplays between the cellular environment and the activated functional molecules, dictated by the physical and chemical laws.

One challenge in "reading" the genome is that our understanding about the language in which the book is written is very limited. For a typical bacterial genome, we know (or are able to find out) the majority of the proteins it encodes, the regulatory switches for some proteins and their corresponding regulatory proteins for only a few of them. We also know a few RNA genes but know very little about their regulatory elements. We know that some of the encoded molecules will work directly with each other when activated, e.g., proteins encoded in the same operons. From other sources of information, we also know that under certain conditions, some of these encoded molecules will be activated with a specific abundance to work together or in subsequent steps to accomplish a designed function after being put into certain sub-cellular compartments, and then be degraded upon completion of their task. In a sense, we have some general knowledge about the basic parts and some interacting parts of the cellular machine of a bacterial cell, and have made some observations about which parts will react to what conditions[59-60]. But we are clearly far from understanding the machine as a functional system.

First we do not have any knowledge yet about the designing principles of the machine. There is an emerging field called *systems biology*[61-67] that studies issues of biological systems from a system's point of view. Scientists in that field study known biological networks and processes, attempting to decipher the designing principles of such systems and link so derived information back to the genome. In a sense, they are using a top-down approach to derive how biological systems are encoded in a genome, while we are here attempting to study the same problem using a bottom-up approach.

Second, while we know numerous individual metabolic processes, a few signaling pathways and regulatory networks in some specific organisms such as *E. coli*, we do not yet know how such individual "components" fit together to form a higher level organization. Actually we do not even know if these components constitute *natural* subsystems of a cellular system, in the sense that they can be replaced by equivalent subsystems following some assembly rules. Intuitively we would imagine that the genes encoding each pathway or network should be arranged in tandem or at least in close proximity in the genome to facilitate efficient transcription but that is apparently not the case when we examine the actual distributions of such genes/operons. To the best of our knowledge, there have not been any published studies about why the genes encoding each pathway are arranged in the genome the way they are (beyond individual operons).

Third, it is well known that the circular DNA of a bacterial chromosome is folded into a complex shape in a cellular location called *nucleoid*, which consists of multiple loops, each of which is a topologically independent domain, possibly formed through connecting

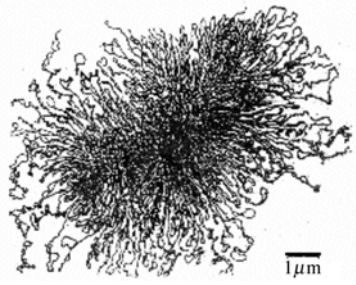

Fig.2. Electronic microscopy image of folded *E. coli* genophore[68].

parts of them to histone-like proteins H-NS[68], as shown in Fig.2. Each loop is probably further folded into a compact structure (independent of other loops) that may be opened up during the transcription of the genes encoded in the loop and folded up again after the completion of the transcription[68]. While researchers already knew the folded structure as shown in Fig.2 30+ years ago[69], not till very recently did people start to study the detailed structure of the folded DNA and associated DNA-binding proteins. Still, very little is known about how these loops correlate with genes in the same operons, regulons or biological pathways.

Fourth, the cellular machinery of a bacterium not only has a dynamic structure but also has a dynamic parts list. It is now generally understood that bacteria exchange genetic material among related organisms constantly through horizontal gene transfers[70-71]! So genes come and go! The scale of such exchange of material is surprisingly high. For example, among three closely related *E. coli* strains: MG1655, EDL933 and CFT073, only 39.2% of the genes are shared by the three organisms[72]. Our current understanding about how foreign genes adapt to the new host and develop a new function there after getting into the host genome is very limited.

Fifth, while we know most of the proteins encoded in a genome, we know very little about their interaction partners other than the ones sharing a common operon. Actually it may be impossible to predict such interaction relationships based on genome sequence alone. As we know, proteins are put into the designed cellular compartments after their creation. It will depend on the geometric shapes and the physical properties of their folded structures whether two proteins may interact, which clearly cannot be predicted easily from the sequence alone. Even more challenging is to predict under what conditions two proteins may interact. We know even less about interactions between RNAs, as well as between proteins and RNAs, which are all essential parts of the cellular machinery that is encoded in a genome.

Sixth, while we know very little about how the whole cellular machinery is assembled from its parts list, we know even less about the dynamic behavior of the machinery, much of which is probably determined by the physical and chemical laws when the proteins and RNAs along with other molecules are put into the same sub-cellular locations. Currently, studies of the dynamic behaviors of biological systems are generally conducted in the fields of biophysics and computational chemistry, not necessarily using any genomic information. Ultimately studies of genomes and such dynamic behaviors of biological systems will need to be merged into one framework that takes into consideration all the relevant information to realistically model biological systems.

Seventh, while bacteria are unicellular organisms, they do communicate with other cells of the same organism and even with other organisms. Actually the majority of the bacterial cells of different organisms need to co-exist in order for them to survive in nature. For example, hundreds of, possibly much more, different bacteria co-exist in human intestines, totaling $\sim 10^{14}$ bacterial cells. The majority of these bacterial cells are not free-living, indicating that they need metabolites generated by other bacterial cells to survive. Such a bacterial community, together with their living environment, forms a very complex metabolic system. Without any key component of the community, the whole system will collapse. While researchers have started looking into such bacterial communities using probing techniques like metagenome sequencing[73-75], our general understanding about such bacterial community, as well as their genomes, is very limited.

1.3 Information Potentially Derivable That Can Help to Bridge the Knowledge Gap

There is clearly a substantial gap between the information encoded in bacterial genomes and what we know about it. The challenge does not only lie in the reality that we do not know how to read the language yet, in which the genome is written, but also lies in another (often forgotten) reality that this is not a self-contained book — it is missing all the information about the physical and chemical laws. So we need to constantly remind ourselves about this when attempting to bridge the gap!

In the rest of the article, I will discuss the information that is potentially derivable from genome sequence alone in the foreseeable future, which will need to be integrated, when attempting to understand the whole picture, with information derivable based on other experimental data such as (a) microarray gene expression data[76-77] for inference of the transcription subsystem and associated regulatory subsystem in a cell, (b) tiling arrays[78-79] for identification of *cis* regulatory elements of operons, (c) ChIP on chip data[80] for identification of interaction partners between transcription regulators and their *cis* regulatory elements, (d) proteomic data measuring the presence and the quantities of proteins under specific conditions typically collected using mass spectrometry techniques[81-82], (e) metabolomic data measuring the metabolites as the results of metabolic reactions and their quantities using mass spectrometry or nuclear magnetic resonance (NMR) techniques[83-84], (f) protein interaction data generated

using techniques like yeast two-hybrid arrays[85] or pull-down approaches[86-87], (g) protein and complex structures generated using X-ray crystallography[88-89], NMR or electronic cryo-microscopy techniques[90-91], which can provide detailed information such as how a protein executes a particular reaction, and (h) imaging data for tracing the movements of bio-molecules inside a cell, as well as information derived through systems level modeling and simulational studies as outlined earlier.

So what information is potentially derivable from the genome sequence alone, other than the one that people have already derived? The key in answering this question lies in the genomes themselves, not just individual genomes but all the (available) genomes as a whole. A substantial amount of information could be derived through genome comparisons! By genome comparisons, people have observed various "rules" for arranging genes in a genome such as genes working in the same pathways tend to group into operons[92]. The general belief is that these observed patterns of genomic arrangements are probably due to functional reasons. As organisms evolve to adapt to their changing environments, the genomic structures and their gene lists change. This is illustrated in Fig.3, which shows substantially rearranged genomic blocks across three related cyanobacterial genomes. From the figure, we see that while the global gene arrangements of the three related genomes are quite different, many of the local structures are well preserved. For example, genes encoding a specific metabolic pathway will generally be arranged into a set of similar operons across related organisms. Such functional constraints on genomic structures provide the basis for the powerful comparative

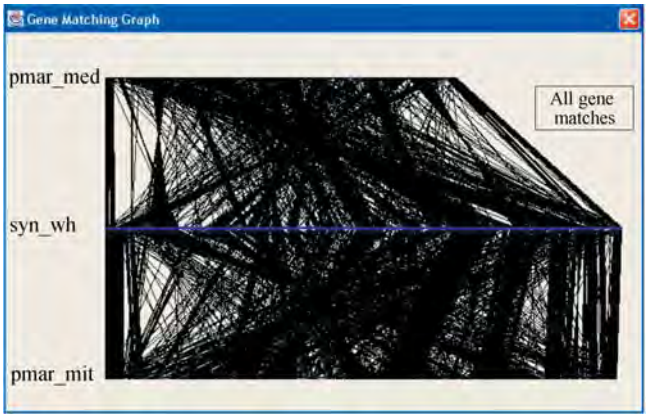

Fig.3. Orthologous gene mapping across three cyanobacterial genomes, MED4 (top), WH8102 (middle) and MIT9313 (bottom), each represented by a horizontal line[94]. The line connecting two dots across two genomes represents an orthologous gene pair.

genome analyses[93], which can be used to reveal unknown genomic structures by detecting conserved structures, local or global, across multiple genomes.

The most essential idea of comparative genome analyses is to discover conserved (or preserved) patterns, such as the preserved gene blocks in Fig.3 across multiple genomes, and map such discovered patterns to well-understood biological concepts. Using such information, people have developed computational methods for prediction of orthologous genes (intuitively, equivalent genes in different organisms) across multiple genomes[95], for inference of HTGs in a genome[57], for discovery of a type of previously unknown genomic structure, called uber-operons[96-97], and for prediction of *cis* regulatory elements of genes (or operons)[98], just to name a few.

2 Comparative Genome Analyses — Knowledge Discovery Through Comparison

The idea of comparative genome analyses is to identify the equivalent (genomic) elements across multiple genomes and then to discover patterns among the orthologous elements with statistical significance. Originally, the idea was introduced for identification of *orthologous* genes, referring to genes in different organisms that have evolved from a common ancestral gene through speciation only[93,95]. This is in contrast to *paralogous* genes, which refer to genes related by duplication within a genome[99]. Generally, orthologs retain the same function in the course of evolution, whereas paralogs evolved to adopt new functions. Identification of orthologous genes has become the basis for many genome analysis strategies. For example, identification of *cis* regulatory motifs for transcription regulation relies on identification of promoters of orthologous genes[100] across related organisms, which has been widely used for eukaryotic genomes. Other applications include (a) prediction of protein functions[101], (b) operon prediction[102], (c) uber-operon identification[97], (d) co-evolutionary analyses of genes[103-104], (e) co-occurrence analyses of genes[105-106], (f) genome rearrangement analyses[107], (g) prediction of protein-protein interactions[108], (h) biological pathway mapping[93] among other applications.

Whereas the importance of finding orthologous genes is obvious, there has not yet been a generally accepted algorithm for solving the problem. This is because the definition of orthologous genes given above, as an evolutionary concept, does not give rise to an operational definition. The existing methods are generally based on finding genes with the highest sequence similarities, either one way[101], two-ways[109-110], or three-ways[111].

While practically useful, these methods are not very satisfying due to the lack of a solid basis for prediction of orthology. By definition, orthologous genes refer to genes having the same functions when working in the same biological process across different organisms; hence their prediction should also rely on information of their working partners. A few efforts have been attempted along this direction[93], i.e., to predict orthologous genes through first identification of the working partners of a query gene in its genome and then checking if the homologs of the partner genes are also the working partners of the homolog of the query gene in the target genome. However such prediction methods have relatively low coverages, and do not work particularly well for genomes that are remotely related. The field of comparative genome analyses can definitely use a theoretically sound and computationally effective method for orthologous gene prediction.

The above idea can also be applied to other biological entities such as ncRNA genes, promoters, and even higher-level organizations like operons, regulons and biological pathways. There is really a need for developing a general framework for mapping various genomic elements as outlined above to their counterparts (orthologs) across genomes and in support of more effective ways for information discovery through comparative analyses of bacterial genomes. We believe that some of the more global structures of genomes may only become apparent through analyses of such large correspondence maps.

3 Genome Visualization in Support of Knowledge Discovery

Up till very recently, most of the genome analysis studies have been to answer specific biological questions such as finding protein-encoding genes or identification of operons based on identified characteristics of previously known similar elements. The availability of the rapidly increasing pool of bacterial genomes allows computational scientists to go beyond this type of analysis to carry out discovery-driven analyses of genomes. One such example is the discovery of uber-operons[97,112]. During the analysis of predicted operons across multiple bacterial genomes, researchers found that whereas operons are in general not conserved across genomes, some of their unions are conserved, which led to the discovery of uber-operons[97]. We would expect that many similar discoveries will be made as more genomes become available and as our ability to "mine" genomes continues to improve. One thing that computational scientists can do to facilitate such discoveries is through development of effective visualization techniques of genomes to make many of the genomic features visually apparent.

We have recently developed a simple but yet very effective method for visualizing genomes[55], through which numerous non-trivial genomic features become visually obvious. The basic idea of this genome-visualization technique is that for a given genome, we partition its sequence into a series of non-overlapping and equal-sized fragments① of M bps; then for each k-mer, for any fixed integer k between 1 and 7, we calculate the combined frequency of the k-mer and its reverse complement within each partitioned fragment. We define a matrix of $N(k)$ columns and $genome\_length/M$ rows, with each element representing the frequency of the corresponding combined k-mer within the corresponding sequence fragment, where $N(k)$ is the number of unique combined k-mers. Note that $N(k) = 4^k/2$ or $(4^k + 4^{k/2})/2$, depending on if k is odd or even. For example, $N(4) = 136$. Our first observation is that *the combined k-mer frequency distribution is highly stable across the whole genome, for any fixed k-mer*; and *this is true for any sequenced genome, prokaryotic or eukaryotic, chromosomal or organelle*[55].

The discovery of this property led to an image representation of a genome, where we map the frequency of each k-mer to a grey level so that higher frequencies are mapped to brighter grey levels (we refer the reader to [55] for the detailed mapping information). We noted that the stable frequency distribution for each combined k-mer gives rise to a vertical line with a generally consistent grey level across the whole image, and hence we term the image *the barcode* of the genome. Fig.4 shows the mapped grey-level images of four genomes and a random DNA sequence. We also noted that this interesting barcode property of a genome is mostly due to the 5th order Markov chain property that the coding regions of a genome, which typically account for 85% of the whole genome, have[55].

By visualizing these and the other computed genome barcodes, we have made the following observations. (a) The majority of the sequence fragments in a genome share highly similar barcodes while the fragments with distinct barcodes in the genome generally represent horizontally transferred genes or highly expressed genes[55,113]; and (b) different classes of genomes, such as eukaryotic, prokaryotic, mitochondrial, plasmid and plastic, each have their unique and identifiable characteristics, as shown in Fig.5. Hence we should be able to tell if a piece of genomic segment is from which class of genomes such as eukaryotic, prokaryotic or plastids.

The barcode representation of a genome has made discoveries of some genomic features relatively easy,

①Possibly except for the last fragment that might be shorter than M bps.

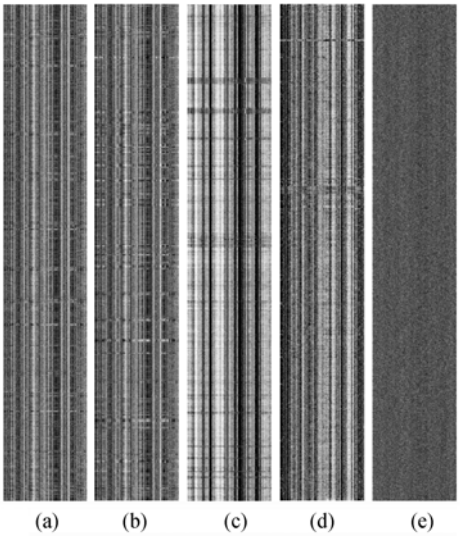

Fig.4. Grey-level representations of k-mer frequency distributions. (a) *E. coli* K12. (b) *E. coli* O157. (c) Chromosome 1 of *B. pseudomallei* K96243. (d) Archaean *P. furiosus* DSM3638. (e) Random nucleotide sequence generated using a zero-th order Markov chain model. The x-axis for each representation is the list of all (combined) 4-mers arranged in the alphabetical order, and the y-axis is the genome axis with each pixel representing a fragment of $M = 1000$ bps long.

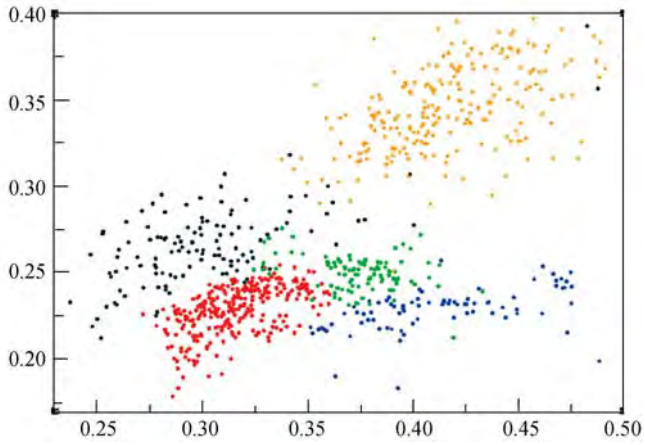

Fig.5. Barcodes in feature space. The x-axis is the average of variations of the 4-mer frequencies across a whole genome across all 4-mers, and the y-axis measures the similarity level among all 1000-bps partitioned fragments of the genome, each represented as a 136-dimensional vector of 4-mer frequencies; specifically, for each genome, we build a minimum spanning tree[114] based on the 4-mer frequency vectors for its sequence fragments and their distances. The y-axis is the averaged weight (distance) of all edges in the minimum spanning tree. The green dots represent prokaryotes (586 genomes), the blue ones for eukaryotes (83 chromosomes), the red ones for plastids (101 genomes with lengths $>$ 20000 bps), and the black for mitochondria (120 genomes with lengths $>$ 20000 bps).

including native *versus* recently transferred genetic elements and mRNAs that can bind to a particularly class of proteins[55]. We believe that a visualization capability like this could substantially speed up knowledge discoveries about genomic structures and information encoded in them. The barcode scheme utilizes only one specific property of the genomes, i.e., the stable k-mer frequency distributions across a genome and the uniqueness of the collection of such k-mer distributions for each genome, which has already led to some very exciting discoveries. We expect that different ways to visualize bacterial genomes, utilizing other features of the genomes, may lead to even more effective and more general visualization tools to facilitate knowledge discoveries. The challenge lies in identifying genomic features useful to make such visualization tools effective.

4 Identification of Operons and *cis* Regulatory Motifs — Going Beyond Individual Genes

Genes in bacterial genomes are organized into operons, which are transcribed together. While there have been debates about the driving force for the formation of operons[92,115], one probable reason is that operons are to facilitate efficient co-transcription of genes working in the same biological processes. It is this feature of bacterial pathways that makes their elucidation much easier than their counterparts in eukaryotes. By utilizing this feature, researchers have developed computational tools for derivation of gene components of pathways based on predicted operons and other information[116].

Numerous operons have been experimentally studied, particularly in model organisms like *E. coli* and *B. subtilis*. As of now, 690 operons in *E. coli*[117] and 992 operons in *B. subtilis*[118] have been elucidated experimentally. In addition, a few other organisms also have experimentally determined operons such as *Shewanella oneidensis*[119], *Pediococcus pentosaceus* [120] and *Lawsonia intracellularis*[121]. Researchers have made a number of observations about these known operons, including: (a) the intergenic distance within an operon is generally shorter than the inter-operonic distance; (b) the adjacency relationship among genes within an operon is often conserved across related genomes, and (c) genes in the same operon are generally functionally related. The other information also found useful for operon prediction includes (i) the *cis* motifs for the general transcription regulators that recruit the RNA polymerases, such as the motifs for sigma factors[123], (ii) the different length distributions of operons on the two (leading and lagging) strands of a genome[124], and (iii) termination signals for some classes of operons, specifically the rho-independent terminators[125].

Based on these and other observations, researchers have developed a number of operon prediction programs such as DOOR[13-14], MicrobesOnline[46] and GeneRegulation[122].

Fig.6. Schematic for two operons and their promoter sequences in a genome. Each rectangular box represents a gene and each star represents a regulatory motif in the promoter region.

The state of the art in operon prediction is about ∼92% accuracy in determining if a pair of adjacent genes on the same genomic strand represents an operon boundary or not when trained on data from the same organism, and ∼84% when trained on data from a different organism[14]. While respectable, this level of accuracy could possibly be improved through better utilization of the genomic sequence data. There are a few areas where further exploration could lead to improved signals for operon boundaries. For example, none of the existing operon-prediction programs have fully utilized the information about promoter sequences, which have traditionally been considered as a separate problem, although they are really the two sides of the same problem. We anticipate that a strategy that attempts to solve the two problems together may lead to more effective ways for solving both problems more accurately.

Recent studies suggest that operons may have substructures that can be transcribed independent of the other genes in the same operon[126]. Our analyses of 380 sets of microarray gene-expression data[126] for *E. coli* K12 support this observation, which suggest that the majority of the multi-gene operons in *E. coli* K12 have more than one unique transcript (unpublished data). Our data suggest that there could be regulatory elements inside operons that can control the transcription of subsets of genes within operons. Further analyses along this direction could lead to discoveries about such substructures and their associated regulatory elements, and the results could fundamentally change our understanding about operons as we know.

The *cis* regulatory motifs for each operon are located in their promoter region, and they can bind to their transcription regulators (proteins) released by the cell under designed conditions, to adjust the transcription status of the operon. Prediction of *cis* regulatory motifs for transcription regulation represents one of the classical problems in bioinformatics, which has been receiving considerable attention since the late 1980's. The first generation of prediction algorithms mostly focused on solving the following problem: given a set of promoter sequences of possibly co-expressed genes, find sequence segments in the promoters, which have high information content when aligned[127]. The co-expression information could come from gene-expression data or other experiments. Numerous algorithms have been developed to tackle this problem[128-129]. The next major advancement in *cis* motif finding came when the phylogenetic foot-printing technique[130-131] was developed, which does not require the initial guess or information about co-expression. The basic idea of the technique, originally developed for motif finding in eukaryotic genomes, is to find orthologous genes across related organisms, and then find conserved sequence motifs across the promoters of the orthologous genes. The assumption for the technique to work is that orthologous genes of closely related organisms are regulated by orthologous regulators that have highly similar DNA-binding domains, and hence have similar *cis* regulatory motifs. While this technique has proved to be useful for eukaryotic genomes, its application in bacterial genomes has been limited mainly because it has to be generalized to find orthologous operons, which may not exist in general across related organisms. Clearly further thinking is needed to effectively extend the idea of finger-printing to bacterial genomes. An important and challenging application of finding co-transcribed operons is to solve the regulon problem as detailed in Section 6.

5 Searching for Functionally Associated Genes — Basic Components of Cellular Machinery

We (loosely) refer genes working in the same biological (metabolic, regulatory or signaling) pathway as *functionally associated*. The question we are interested in addressing is if so-defined functionally associated genes can be identified based on genomic sequence information alone. Clearly having such a capability is very important to elucidation of biological pathways as it can provide the component list of a pathway. From predicted operons, some portions of such a list can be derived but the question now is "Can we derive functionally associated genes beyond the ones encoded in the same operons?"

We have previously developed a computational scheme[132-133] for predicting such functionally associated genes, mainly based on two types of information of genes: (a) co-evolutionary information and (b) co-occurrence information. There have been a number of computational methods developed to derive co-evolutionary relationships of genes. Among them, the phylogenetic profile analysis was shown to be effective[132-133]. Genes are considered to be *co-occurred* if they appear in the same neighborhood of a genome substantially more frequent than by chance. Using such information, one can predict if two genes in

a genome are functionally associated. When applying to all the gene pairs encoded in a genome, one can build a *functional linkage map* consisting of all gene pairs predicted to have functional associations using the above idea[133]. From such a functional linkage map (represented as a graph with genes as nodes and a predicted functional association between two genes as an edge), one can identify subgraphs whose intra-subgraph edge density is substantially higher than the density of edges connecting the subgraph with the rest of the graph. A few research groups have developed and published computational methods for predicting if two genes are functionally associated essentially using such information such as AVID[135] and PIPA[136].

Our analysis of the so derived subraphs, which were termed *functional modules*, led to an interesting observation that the majority of these functional modules correspond to component genes of known biological pathways[93,132-133] as shown in Fig.7. This suggests that by fully utilizing co-evolutionary and co-occurrence information of genes, one can possibly derive component information of biological pathways. So the general question is "Can we possibly determine which class of biological pathways could possibly have the majority of their component genes derived using this type of method?" To fully address this problem, we probably need more sophisticated methods for identification of "functional modules?" and carefully benchmark the to-be-developed methods against known pathways in databases such as KEGG [137] or MetaCyc[138]. New insights from such studies could lead to improved ways to derive biological pathways encoded in bacterial genomes.

6 Identification of Regulons — Working Towards Elucidation of Cellular Response Systems

A *regulon* typically consists of a few, in some cases many, transcriptionally co-regulated operons, whose protein products work together to accomplish a high-level function in response to a particular stimulus, extra-cellular or intra-cellular. For example, genes involved in the response system to nitrogen in a bacterial cell may include genes that encode transcription regulators coupled through two-component systems[139-140] with proteins for sensing the availability of nitrogen in various forms; transporter proteins that can transport the detected forms of nitrogen from the environment into the cell; enzymes that can break the up-taken nitrogen (possibly in compound forms) into a form that the cell can utilize; enzymes that can convert the nitrogen to energy; and possibly genes encoding some secondarily related cellular processes triggered by the above activities. Operationally, all the genes involved in such a nitrogen uptake and assimilation regulon should be transcriptionally regulated by a common transcription regulator or a group of transcription regulators, and hence they should have conserved *cis* regulatory motifs in their promoters that can bind to the same regulatory protein(s). Hence computational prediction of regulons can be formulated as to group operons into (possibly overlapping) clusters, each of which contains a maximal set of operons sharing at least one common *cis* regulatory motif. Intuitively this problem seems to be quite solvable but the reality is that there is not a single prediction program developed and publicly available

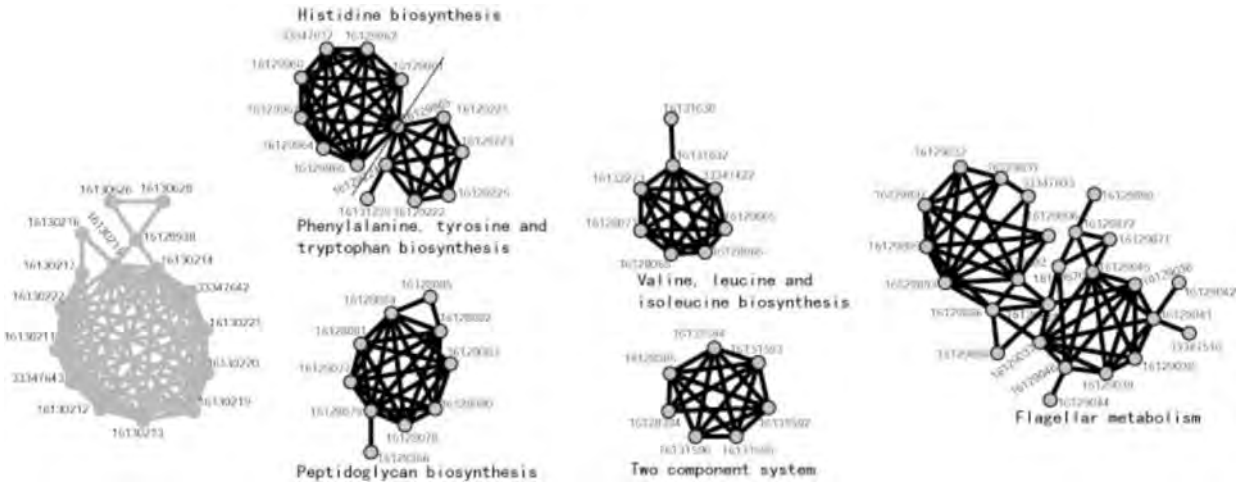

Fig.7. Examples of predicted functional modules in *E. coli*. Modules marked in black are consistent with known pathways in KEGG, and the modules marked in grey have highly similar GO numbers.

for solving this problem for bacterial genomes in general!

The challenge in solving this problem comes mainly from the reality that neither operons nor *cis* regulatory motifs can be predicted very accurately; what makes the situation even worse is that some of the *cis* regulatory motifs are not well conserved although they all bind to the same transcription regulators (see Section 5). This is particularly the case for *cis* motifs of global transcription regulators such as the sigma factors[141-142] in general or CRP[143-144] in *E. coli.* This makes the prediction of co-regulated genes very challenging, particularly at a genome scale. There have been numerous attempts[145-148] gearing towards prediction of regulons, using both sequence-conservation information across promoters of the same genome and conservation information of orthologous promoters across related genomes. In addition, people have tried to predict regulons using the assumption that true *cis* regulatory motifs tend to cluster with other *cis* motifs while false ones may not[149]. Even with all these efforts, accurate prediction of regulons based on genome sequences alone has been very challenging although theoretically the problem should be solvable!

7 Metabolic Pathways — What Determines Their Gene Arrangement in a Genome?

Each of the three intertwined cellular systems, namely, metabolic, regulatory and signaling, is made of a collection of molecular interactions, somewhat artificially partitioned into "pathways". A *pathway* basically is a collection of functional molecules that work together to form a molecular complex or work in consecutive steps in some chemical reactions. Interacting pathways form *networks*, which are capable of accomplishing more complex functions. Often in the biological literature, pathways and networks are used interchangeably although networks tend to refer to more complex pathways. The executions of pathways or networks are generally referred as *biological processes*. As discussed earlier, genes encoding a metabolic pathway tend to group into operons[150]. A natural question is "Do operons encoding a pathway tend to cluster together in a genome?" Intuitively we would imagine so. However our simple analyses of the genomic distributions of operons of all the well characterized 123 pathways of *E. coli* K12 in KEGG[151] indicate otherwise as we found that operons encoding a pathway are typically scattered across the whole genome, as shown in Fig.8. So the more interesting and more challenging question is "Are there any rules that dictate the genomic arrangement of operons across all metabolic pathway-encoding genes?" Clearly this is a fundamental question about bacterial genomes. To the best knowledge of the author, there have not been any published studies that attempt to address this question.

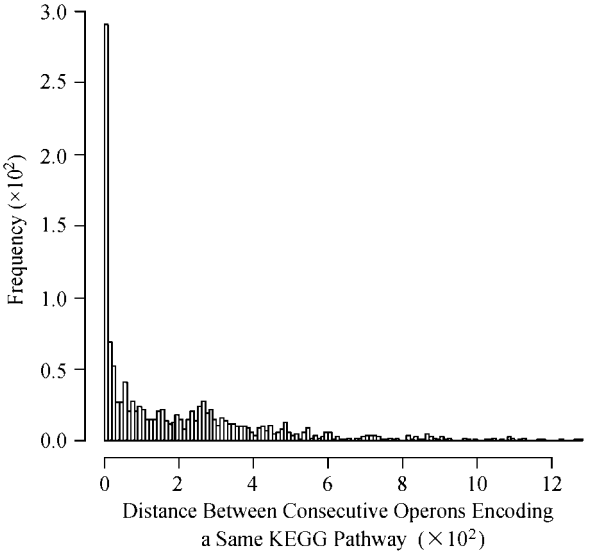

Fig.8. Distribution of the average distance between two "consecutive" operons encoding a pathway across all KEGG metabolic pathways.

We have recently carried out a study attempting to address this issue[152]. We first noticed that many of the *E. coli* operons are shared by multiple pathways. For example, in *E. coli* K12, 56% of operons are shared by at least two pathways and on average each operon is shared by 2.09 pathways among all operons encoding the 120+ metabolic pathways. So we speculate that it is the overlaps among the pathways that might have prevented, at least in part, operons of the same pathways from clustering together in a genome. We then examined the expression patterns of all the 123 *E. coli* K12 metabolic pathways on a large microarray gene-expression dataset, collected under 380 experimental conditions[137,151,153], which covers a wide range of conditions under which *E. coli* can survive. We noted that more frequently expressed (hence more frequently used that we assume) pathways tend to have their operons more closely clustered together, measured using a "scattering" score of a pathway's operons, than operons in less frequently expressed pathways. Further analyses led to a very interesting and potentially profound finding — that is operons of the 123 known *E. coli* pathways are so arranged in the genome that the total *effort* to locate all operons in each pathway during its transcription is very close to being minimized among all alternative arrangements of the involved operons, when taking into consideration of the frequencies of all individual pathways being used during the life time of

the cell; here we assume that our estimated activation frequency of each pathway based on the 380 sets of microarray data is generally accurate. Fig.9 highlights this discovery, which is detailed in [152].

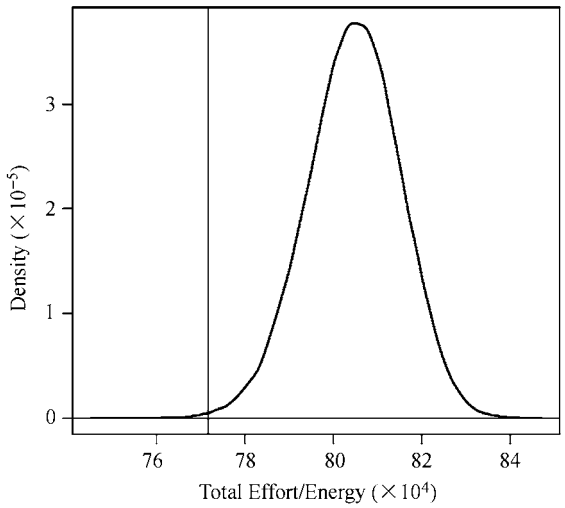

Fig.9. Energy distribution over one million randomly permuted arrangements of the operons covered by 123 *E. coli* pathways. The vertical line shows the energy level of the current arrangement of operons in *E. coli* genome.

This analysis result suggests the feasibility in studying the global properties of bacterial genomes, which has been rarely done before if ever. Using similar or more general ideas, one can possibly start to address other general issues about bacterial genomes (and possibly beyond): (a) what other factors, other than what we outlined in this section, may have constrained the genomic arrangement of operons? (b) how each cellular subsystem is encoded in a bacterial genome and why so? (c) whether the rules to-be fully derived for the genomic arrangement of operons determine the assembly rules of cellular (sub)systems in a bacterial cell, just to name a few. We anticipate that computational analyses of genomic sequence data as well as gene expression (and protein expression) could play key roles in addressing these and other questions about the general properties of bacterial genomes (and possibly beyond).

8 Identification of Mobile Elements — Understanding the Dynamic Nature of a Genome

Genomes are not static; actually they change constantly in terms of the composition of their genetic material as well as the genomic locations of the genetic materials. There are two basic mechanisms for a genome to change: (a) having foreign genetic material inserted into the genome, through horizontal gene transfers[57-58]; and (b) having one piece of genetic material to move from one location to another within the same genome. For either class, a mobile genetic element (MGE) needs to be first excised or transcribed from the host genome into either an RNA or a DNA; and then it may be transmitted and integrated into another location within the genome for the second class or from the host genome to another genome through horizontal gene transfers for the first class[154]. The transposition activities of these MGEs alter the organization as well as the composition of the host genome, which may affect the functionalities and/or the livelihood of the host organism. It is the selection process of evolution that determines which cells with altered genomes may survive.

Identification of horizontally transferred genes (HTGs) in a host genome represents a highly important problem as it will not only inform which genes are from foreign species but also, more importantly, provide the key information about how organisms on Earth have evolved[155] as well as a theoretical basis for inference of orthologous genes in general. Various models have been proposed to infer HTGs, among which two are the most popular ones: (a) *surrogate methods* that infer HTGs based on their atypical nucleotide compositions compared to the other genes in the same genome; and (b) *phylogenetic-incongruence*-based methods that infer HTGs based on their atypical gene trees compared to the gene trees of the "typical" genes in the genome. Generally speaking, the first class of methods is more suited for finding recent HTGs with the advantage being that they require only relatively simple computational procedures and associated data analyses. The second class of methods is generally more accurate in identification of HTGs but requires expensive computation of phylogenetic trees and associated tree analyses[155]. While much of the fundamental work on inference of HTGs has been done by evolutionary biologists, a challenging problem to computational scientists is how to convert such evolutionary work to effective computer algorithms and computer programs that are accessible to biologists in general.

Transposable elements (TEs) represent the majority of the second class of MGEs, and they generally fall into two classes: DNA transposable elements and (RNA) retrotransposable elements[19-20,52,156-159]. As of today only DNA transposable elements have been found in bacterial genomes[159]. Different from the retrotransposable elements that utilize a "*copy and paste*" strategy to increase their population in a genome, DNA transposable elements employ a "*cut and*

paste" strategy to create a new copy while losing the old one. However sometimes, a DNA TE may create a new copy while keeping the original one by accident, i.e., the removed original copy might be brought back by the cellular DNA repair mechanism[160]. A direct result in having different transposition mechanisms is that while a (eukaryotic) genome may have thousands up to hundreds of thousands of copies of the same retrotransposable element, a DNA transposable element in general has substantially lower number of copies in a genome if any.

Identification of TEs in a genome represents a relatively simple problem compared to the identification of HTGs since TEs have more identifiable signals. Specifically each TE generally has inverted repeats (though possibly degenerative) at its two terminals, and it encodes a transposase and possibly other genes between the two terminal signals. One computational challenge lies in deriving the roles that the TEs have played during the evolution of the host genome, particularly in gaining or losing functionalities due to the transpositions of genetic materials caused by the TEs.

9 Developing Tools to Facilitate Experimental Biologists to Mine Genomes — a Practical Consideration

Sequencing technology has advanced to such a level that large sequencing centers such as the Joint Genome Institute can sequence a bacterial genome in a day. To date, 800+ bacterial genomes have been sequenced and 2000+ are in the pipeline being or to be sequenced. It is foreseeable that we could see over 10 000 sequenced bacterial genomes within the next few years. While the rapidly increasing genomic sequence data have provided unprecedented opportunities for biologists to study the underlying organisms, it has also raised some very challenging problems to the computational scientists as there is clearly a large and widening gap between what is available or easily derivable from the annotated genomes and what a typical experimental biologist can realistically get from the annotated genomes. This gap has slowed down the information transmission from the genome sequencing and annotation centers to experimental biologists in general, posing an urgent need for development of more effective computational techniques that can help to close the gap. The need for biologists to directly mine the genomes is clear as directly working on the genomes and seeing returned intermediate results on computer screens could inspire a biologist to ask related, broader and deeper questions, which may not happen by working through some computer programmers.

We envision that some novel, effective and easy-to-use computational problem-solving environments are needed to assist experimental biologists to mine and analyze annotated bacterial genomes without the need of them knowing any computer programming languages. Such a computational problem-solving environment should allow a biologist user to directly send instructions to the system to carry out desired data mining analyses of genomes, particularly comparative genome analyses, using a language general and flexible enough that can represent sophisticated queries and are interpretable by a computer system. Because of the flexibility needed for such a capability, a query may need to be interpreted as a workflow consisting of low-level prediction, analysis and utility tools that have been pre-implemented within the system. We would imagine that mapping a query to a workflow will require an ontology that defines the basic operations (by computational tools) and the relationships among the desired results and the available tools. Using such ontology, a query can be possibly automatically translated into a dynamically composed workflow, which will be ultimately executed through calling pre-implemented analysis and utility tools and database retrievals. The following could represent a scenario of a user using such a problem-solving environment for genome mining, which could potentially guide the development of such a system.

Suppose an experimental biologist Steve is interested in studying why *Lactococcus* is pathogenic and its relative *Streptococcus* is not through an integrated computational and experimental study. He wants first to mine the annotated genomes, including these two, using the planned problem-solving environment to collect as much information as possible before he designs his experimental research plan. Through our planned user interface, Steve may pose the following query to get things started: "*Give me all the unique genes that Streptococcus has but none of the genomes in the same genus of Lactococcus have*".

The system should know how to interpret the term "genes" and "orthologous genes" in another organism implied in this query based on the to-be-developed ontology. The term "genus" will need to be included in the ontology so the system will be able to interpret "in the same genus of *Lactococcus*" as all the genomes that are from the same genus of *Lactococcus*. The term "unique" will be interpreted through applying a utility tool, which will check if genomes other than *Streptococcus* have the equivalents of *Streptococcus genes*. After looking through the returned unique genes, Steve may pose the following query: "*What metabolic pathways do these genes encode*?"

This query may first trigger a pathway-assignment tool to assign all the KEGG metabolic pathways to the *Streptococcus* genome. Then a utility tool will be called to find all the "matched" pathways based on the found unique genes. All the matched pathways will be displayed on the user interface. After Steve goes through the displayed pathways, he might start thinking about his experimental design. He might want to block one particular pathway from functioning through knocking out the main transcription regulator of the pathway to check if this pathway is possibly responsible for the pathogenicity. To do that, Steve might want to find out if genes in the target pathway might be transcriptionally co-regulated by asking: *"Give me the operons covered by this pathway*", which will trigger an operon prediction tool.

After looking through the returned operons, Steve may follow with another query: *"Give me the possible cis regulatory sites that appear in the promoter regions of these operons"*. This may trigger invocation of a program to extract all the promoter sequences of the relevant operons and then calling a motif-finding tool to find the highest scoring conserved *cis* regulatory motif among the collected promoters. Since a substantial portion of the transcription regulators in bacterial genomes are self-regulated, Steve may want to try his luck to see if the relevant transcription regulator is also self-regulated by posing the following query: "*Give me all transcription regulators that share the same cis regulatory element just found"*.

This will trigger a set of tools that will first search for matching motifs in all promoter regions in the target genome, and then select the ones that are followed by a transcription-regulator gene (or an operon containing such a gene). Assuming that Steve is lucky, the program returns two such transcription regulators. Through further investigation of the literature, Steve decides to knock out one transcription regulator to check if knocking it out will affect the pathogenicity of the organism.

We believe that a capability like this could substantially improve our ability to transmit the information retrievable and easily derivable from the annotated genomes. Computer scientists have to help to develop such a sophisticated system.

10 Concluding Remark

Very little is known about the general rules that govern the genomic arrangements of genes in a bacterial genome. With our current knowledge about a large number of detailed pathways in model organisms like *E. coli*, as well as the availability of a large number of bacterial genomes and functional data collected under many conditions, we believe that it is the right time to study bacterial genomes gearing towards understanding of their global properties and the associated governing rules. This is a different kind of biology compared with the biology that we have been learning from text books in the past decades. In a sense this is really an information science as our goal is really about finding out how information is encoded in a genome, the popularly known *genome deciphering problem*.

Acknowledgement The author would like to thank Dr. Fengfeng Zhou and Dr. Yanbin Yin for their help in preparing this manuscript, and also thank Xiaojia Tang, Xizeng Mao of the CSBL Lab at the University of Georgia, Professor Yanchun Liang of Jilin University, China, and Professor Jonathan H. Badger of Venter Institute for helpful discussions and insightful comments on the early drafts of the manuscript.

References

[1] Karatan E, Watnick P. Signals, regulatory networks, and materials that build and break bacterial biofilms. *Microbiol. Mol. Biol. Rev.*, 2009, 73(2): 310-347.

[2] An D, Parsek M R. The promise and peril of transcriptional profiling in biofilm communities. *Curr. Opin. Microbiol.*, 2007, 10(3): 292-296.

[3] Hoffman L R, D'Argenio D A, MacCoss M J, Zhang Z, Jones R A, Miller S I. Aminoglycoside antibiotics induce bacterial biofilm formation. *Nature*, 2005, 436(7054): 1171-1175.

[4] Hall-Stoodley L, Costerton J W, Stoodley P. Bacterial biofilms: From the natural environment to infectious diseases. *Nat. Rev. Microbiol.*, 2004, 2(2): 95-108.

[5] How Deep is the Gene Pool? *Astrobiology Magazine European Edition*, 2008, http://www.astrobio.net/amee/summer_2008/Interviews/AnthonyPooleInterview.php.

[6] Ben-Jacob E. Bacterial know how: From physics to cybernetics. *PhysicaPlus*, 2006, 7, http://physicaplus.org.il/zope/home/en/1124811264/1145390912_eshel_en.

[7] Fleischmann R D, Adams M D, White O, Clayton R A, Kirkness E F, Kerlavage A R, Bult C J, Tomb J F, Dougherty B A, Merrick J M *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 1995, 269(5223): 496-512.

[8] Pruitt K D, Tatusova T, Klimke W, Maglott D R. NCBI Reference sequences: Current status, policy and new initiatives. *Nucleic Acids Res.*, 2009, 37(Database Issue): D32-D36.

[9] Rocha E P. The replication-related organization of bacterial genomes. *Microbiology*, 2004, 150(Pt 6): 1609-1627.

[10] Mackiewicz D, Mackiewicz P, Kowalczuk M, Dudkiewicz M, Dudek M R, Cebrat S. Rearrangements between differently replicating DNA strands in asymmetric bacterial genomes. *Acta Microbiol. Pol.*, 2003, 52(3): 245-260.

[11] Reznikoff W S. The operon revisited. *Annu. Rev. Genet.*, 1972, 6: 133-156.

[12] Ames B N, Martin R G. Biochemical aspects of genetics: The operon. *Annu. Rev. Biochem.*, 1964, 33: 235-258.

[13] Mao F, Dam P, Chou J, Olman V, Xu Y. DOOR: A database for prokaryotic operons. *Nucleic Acids Res.*, 2009, 37(Database Issue): D459-D463.

[14] Dam P, Olman V, Harris K, Su Z, Xu Y. Operon prediction using both genome-specific and general genomic information. *Nucleic Acids Res.*, 2007, 35(1): 288-298.

[15] Su Z, Olman V, Xu Y. Computational prediction of Pho regulons in cyanobacteria. *BMC Genomics*, 2007, 8: 156.

[16] Claverys J P, Prudhomme M, Martin B. Induction of competence regulons as a general response to stress in gram-positive bacteria. *Annu. Rev. Microbiol.*, 2006, 60(1): 451-475.

[17] Yasbin R E, Cheo D L, Bayles K W. Inducible DNA repair and differentiation in Bacillus subtilis: Interactions between global regulons. *Mol. Microbiol.*, 1992, 6(10): 1263-1270.

[18] Zhou F, Xu Y. RepPop: A database for repetitive elements in Populus trichocarpa. *BMC Genomics*, 2009, 10: 14.

[19] Zhou F, Olman V, Xu Y. Insertion sequences show diverse recent activities in Cyanobacteria and Archaea. *BMC Genomics*, 2008, 9: 36.

[20] Zhou F, Tran T, Xu Y. Nezha, a novel active miniature inverted-repeat transposable element in cyanobacteria. *Biochem. Biophys. Res. Commun.*, 2008, 365(4): 790-794.

[21] Hayes F. Transposon-based strategies for microbial functional genomics and proteomics. *Annu. Rev. Genet.*, 2003, 37: 3-29.

[22] Hamer L, DeZwaan T M, Montenegro-Chamorro M V, Frank S A, Hamer J E. Recent advances in large-scale transposon mutagenesis. *Curr. Opin. Chem. Biol.*, 2001, 5(1): 67-73.

[23] Izawa T, Ohnishi T, Nakano T *et al.* Transposon tagging in rice. *Plant Mol. Biol.*, 1997, 35(1/2): 219-229.

[24] Noguchi H, Park J, Takagi T. MetaGene: Prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.*, 2006, 34(19): 5623-5630.

[25] Nielsen P, Krogh A. Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics*, 2005, 21(24): 4322-4329.

[26] Hannenhalli S S, Hayes W S, Hatzigeorgiou A G, Fickett J W. Bacterial start site prediction. *Nucleic Acids Res.*, 1999, 27(17): 3577-3582.

[27] Solovyev V, Kosarev P, Seledsov I, Vorobyev D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.*, 2006, 7(Suppl. 1): S10.1-12.

[28] Ellrott K, Guo J T, Olman V, Xu Y. Improving the performance of protein threading using insertion/deletion frequency arrays. *J. Bioinform. Comput. Biol.*, 2008, 6(3): 585-602.

[29] Guo J T, Xu Y. Towards modeling of amyloid fibril structures. *Front. Biosci.*, 2008, 13: 4039-4050.

[30] Marnef A, Sommerville J, Ladomery M R. RAP55: Insights into an evolutionarily conserved protein family. *Int. J. Biochem. Cell Biol.*, 2009, 41(5): 977-981.

[31] Finn R D, Tate J, Mistry J, Coggill P C, Sammut S J, Hotz H R, Ceric G, Forslund K, Eddy S R, Sonnhammer E L *et al.* The Pfam protein families database. *Nucleic Acids Res.*, 2008, 36(Database Issue): D281-D288.

[32] Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche B A, de Castro E, Lachaize C, Langendijk-Genevaux P S, Sigrist C J. The 20 years of PROSITE. *Nucleic Acids Res.*, 2008, 36(Database Issue): D245-D249.

[33] Bork P. Powers and pitfalls in sequence analysis: The 70% hurdle. *Genome Res.*, 2000, 10(4): 398-400.

[34] Aravin A A, Hannon G J. Small RNA silencing pathways in germ and stem cells. *Cold Spring Harb. Symp. Quant. Biol.*, 2008, 73: 283-290.

[35] Mattick J S, Amaral P P, Dinger M E, Mercer T R, Mehler M F. RNA regulation of epigenetic processes. *Bioessays*, 2009, 31(1): 51-59.

[36] Stricklin S L, Griffiths-Jones S, Eddy S R. C. elegans noncoding RNA genes. *WormBook*, 2005, 1-7.

[37] Goodrich J A, Kugel J F. From bacteria to humans, chromatin to elongation, and activation to repression: The expanding roles of noncoding RNAs in regulating transcription. *Crit. Rev. Biochem. Mol. Biol.*, 2009, 44(1): 3-15.

[38] Bradley R K, Uzilov A V, Skinner M E, Bendana Y R, Barquist L, Holmes I. Evolutionary modeling and prediction of non-coding RNAs in Drosophila. *PLoS One*, 2009, 4(8): e6478.

[39] Childs L, Nikoloski Z, May P, Walther D. Identification and classification of ncRNA molecules using graph properties. *Nucleic Acids Res.*, 2009, 37(9): e66.

[40] Voss B, Georg J, Schon V, Ude S, Hess W R. Biocomputational prediction of non-coding RNAs in model cyanobacteria. *BMC Genomics*, 2009, 10: 123.

[41] Song D, Yang Y, Yu B, Zheng B, Deng Z, Lu B L, Chen X, Jiang T. Computational prediction of novel non-coding RNAs in Arabidopsis thaliana. *BMC Bioinformatics*, 2009, 10(Suppl 1): S36.

[42] Wang S, Wang Y, Du W, Sun F, Wang X, Zhou C, Liang Y. A multi-approaches-guided genetic algorithm with application to operon prediction. *Artif. Intell. Med.*, 2007, 41(2): 151-159.

[43] Tran T T, Dam P, Su Z, Poole F L, 2nd, Adams M W, Zhou G T, Xu Y. Operon prediction in Pyrococcus furiosus. *Nucleic Acids Res.*, 2007, 35(1): 11-20.

[44] Zhang G Q, Cao Z W, Luo Q M, Cai Y D, Li Y X. Operon prediction based on SVM. *Comput. Biol. Chem.*, 2006, 30(3): 233-240.

[45] Price M N, Arkin A P, Alm E J. OpWise: Operons aid the identification of differentially expressed genes in bacterial microarray experiments. *BMC Bioinformatics*, 2006, 7: 19.

[46] Alm E J, Huang K H, Price M N, Koche R P, Keller K, Dubchak I L, Arkin A P. The MicrobesOnline Web site for comparative genomics. *Genome Res.*, 2005, 15(7): 1015-1022.

[47] Loewen P C, Hengge-Aronis R. The role of the sigma factor sigma S (KatF) in bacterial global regulation. *Annu. Rev. Microbiol.*, 1994, 48: 53-80.

[48] Errington J. Bacillus subtilis sporulation: Regulation of gene expression and control of morphogenesis. *Microbiol. Rev.*, 1993, 57(1): 1-33.

[49] Stragier P, Losick R. Cascades of sigma factors revisited. *Mol. Microbiol.*, 1990, 4(11): 1801-1806.

[50] Prakash A, Tompa M. Discovery of regulatory elements in vertebrates through comparative genomics. *Nat. Biotechnol*, 2005, 23(10): 1249-1256.

[51] Tompa M, Li N, Bailey T L, Church G M, De Moor B, Eskin E, Favorov A V, Frith M C, Fu Y, Kent W J *et al.* Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, 2005, 23(1): 137-144.

[52] Chen Y, Zhou F, Li G, Xu Y. A recently active miniature inverted-repeat transposable element, Chunjie, inserted into an operon without disturbing the operon structure in Geobacter uraniireducens Rf4. *Genetics*, 2008, 179(4): 2291-2297.

[53] Xu Z, Wang H. LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.*, 2007, 35(Web Server Issue): W265-W268.

[54] Feschotte C, Keswani U, Ranganathan N, Guibotsy M L, Levine D. Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. *Genome Biol. Evol.*, 2009, pp.205-220.

[55] Zhou F, Olman V, Xu Y. Barcodes for genomes and applications. *BMC Bioinformatics*, 2008, 9(1): 546.

[56] Whitaker J W, McConkey G A, Westhead D R. Prediction of horizontal gene transfers in eukaryotes: Approaches and challenges. *Biochem. Soc. Trans.*, 2009, 37(Pt 4): 792-795.

[57] Fournier G P, Huang J, Gogarten J P. Horizontal gene transfer from extinct and extant lineages: Biological innovation and the coral of life. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 2009, 364(1527): 2229-2239.

[58] Huang J, Gogarten J P. Ancient gene transfer as a tool in phylogenetic reconstruction. *Methods Mol. Biol.*, 2009, 532: 127-139.

[59] Taylor R, Singhal M. Biological network inference and analysis using SEBINI and CABIN. *Methods Mol. Biol.*, 2009, 541: 551-576.

[60] Schadt E E, Zhang B, Zhu J. Advances in systems biology are enhancing our understanding of disease and moving us closer to novel disease treatments. *Genetica*, 2009, 136(2): 259-269.

[61] Kreutz C, Timmer J. Systems biology: Experimental design. *FEBS J.*, 2009, 276(4): 923-942.

[62] Iyengar R. Computational biochemistry: Systems biology minireview series. *J. Biol. Chem.*, 2009, 284(9): 5425-5426.

[63] van Gend C, Snoep J L. Systems biology model databases and resources. *Essays Biochem.*, 2008, 45: 223-236.

[64] Sauro H M, Bergmann F T. Standards and ontologies in computational systems biology. *Essays Biochem.*, 2008, 45: 211-222.

[65] Brul S, Mensonides F I, Hellingwerf K J, Teixeira de Mattos M J. Microbial systems biology: New frontiers open to predictive microbiology. *Int. J. Food Microbiol.*, 2008, 128(1): 16-21.

[66] Davidov E, Holland J, Marple E, Naylor S. Advancing drug discovery through systems biology. *Drug Discov. Today*, 2003, 8(4): 175-183.

[67] Ideker T, Galitski T, Hood L. A new approach to decoding life: Systems biology. *Annu. Rev. Genomics. Hum. Genet*, 2001, 2: 343-372.

[68] Griswold A. Genome packaging in prokaryotes: The circular chromosome of E. coli. *Nature Education*, 2008, 1(1).

[69] Mason D J, Powelson D M. Nuclear division as observed in live bacteria by a new technique. *J. Bacteriol.*, 1956, 71(4): 474-479.

[70] Gogarten J P, Townsend J P. Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.*, 2005, 3(9): 679-687.

[71] Koonin E V, Makarova K S, Aravind L. Horizontal gene transfer in prokaryotes: Quantification and classification. *Annu. Rev. Microbiol.*, 2001, 55: 709-742.

[72] Lawrence J G, Hendrickson H. Genome evolution in bacteria: Order beneath chaos. *Curr. Opin. Microbiol.*, 2005, 8(5): 572-578.

[73] Preidis G A, Versalovic J. Targeting the human microbiome with antibiotics, probiotics, and prebiotics: Gastroenterology enters the metagenomics era. *Gastroenterology*, 2009, 136(6): 2015-2031.

[74] Petrosino J F, Highlander S, Luna R A, Gibbs R A, Versalovic J. Metagenomic pyrosequencing and microbial identification. *Clin. Chem.*, 2009, 55(5): 856-866.

[75] Hattori M, Taylor T D. The human intestinal microbiome: A new frontier of human biology. *DNA Res.*, 2009, 16(1): 1-12.

[76] Sivachenko A Y, Yuryev A, Daraselia N, Mazo I. Molecular networks in microarray analysis. *J. Bioinform. Comput. Biol.*, 2007, 5(2B): 429-456.

[77] Wade J T, Struhl K, Busby S J, Grainger D C. Genomic analysis of protein-DNA interactions in bacteria: Insights into transcription and chromosome organization. *Mol. Microbiol.*, 2007, 65(1): 21-26.

[78] Tian F, Shah P K, Liu X, Negre N, Chen J, Karpenko O, White K P, Grossman R L. Flynet: A genomic resource for Drosophila melanogaster transcriptional regulatory networks. *Bioinformatics*, 2009, 25(22): 3001-3004.

[79] Kaufmann K, Muino J M, Jauregui R, Airoldi C A, Smaczniak C, Krajewski P, Angenent G C. Target genes of the MADS transcription factor SEPALLATA3: Integration of developmental and hormonal pathways in the Arabidopsis flower. *PLoS Biol.*, 2009, 7(4): e1000090.

[80] Gilchrist D A, Fargo D C, Adelman K. Using ChIP-chip and ChIP-seq to study the regulation of gene expression: Genome-wide localization studies reveal widespread regulation of transcription elongation. *Methods*, 2009, 48(4): 398-408.

[81] Lau K W, Jones A R, Swainston N, Siepen J A, Hubbard S J. Capture and analysis of quantitative proteomic data. *Proteomics*, 2007, 7(16): 2787-2799.

[82] Budzikiewicz H, Grigsby R D. Mass spectrometry and isotopes: A century of research and discussion. *Mass Spectrom Rev.*, 2006, 25(1): 146-157.

[83] Walker G S, O'Connell T N. Comparison of LC-NMR and conventional NMR for structure elucidation in drug metabolism studies. *Expert Opin. Drug Metab. Toxicol.*, 2008, 4(10): 1295-1305.

[84] Mesnard F, Ratcliffe R G. NMR analysis of plant nitrogen metabolism. *Photosynth. Res.*, 2005, 83(2): 163-180.

[85] Bruckner A, Polge C, Lentze N, Auerbach D, Schlattner U. Yeast two-hybrid, a powerful tool for systems biology. *Int. J. Mol. Sci.*, 2009, 10(6): 2763-2788.

[86] Lee E J, Hyun S, Chun J, Shin S H, Kang S S. Ubiquitylation of Fe65 adaptor protein by neuronal precursor cell expressed developmentally down regulated 4-2 (Nedd4-2) via the WW domain interaction with Fe65. *Exp. Mol. Med.*, 2009, 41(8): 555-568.

[87] Chun J, Kwon T, Lee E J, Hyun S, Hong S K, Kang S S. The subcellular localization of 3-phosphoinositide-dependent protein kinase is controlled by caveolin-1 binding. *Biochem. Biophys. Res. Commun.*, 2005, 326(1): 136-146.

[88] Warren E M, Huang H, Fanning E, Chazin W J, Eichman B F. Physical interactions between MCM10, DNA, AND DNA polymerase α. *J. Biol. Chem.*, 2009, 284(36): 24662-24672.

[89] Hrmova M, Fincher G B. Functional genomics and structural biology in the definition of gene function. *Methods Mol. Biol.*, 2009, 513: 199-227.

[90] Li H, Thanassi D G. Use of a combined cryo-EM and X-ray crystallography approach to reveal molecular details of bacterial pilus assembly by the chaperone/usher pathway. *Curr. Opin. Microbiol.*, 2009, 12(3): 326-332.

[91] Ritchie D W. Recent progress and future directions in protein-protein docking. *Curr. Protein Pept. Sci.*, 2008, 9(1): 1-15.

[92] Xie G, Keyhani N O, Bonner C A, Jensen R A. Ancient origin of the tryptophan operon and the dynamics of evolutionary change. *Microbiol. Mol. Biol. Rev.*, 2003, 67(3): 303-342.

[93] Mao F, Su Z, Olman V, Dam P, Liu Z, Xu Y. Mapping of orthologous genes in the context of biological pathways: An application of integer programming. *Proc. Natl. Acad. Sci. USA*, 2006, 103(1): 129-134.

[94] Chen X, Su Z, Xu Y, Jiang T. Computational prediction of operons in Synechococcus sp. WH8102. *Genome Inform.*, 2004, 15(2): 211-222.

[95] Fulton D L, Li Y Y, Laird M R, Horsman B G, Roche F M, Brinkman F S. Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics*, 2006, 7: 270.

[96] Li G, Che D, Xu Y. A universal operon predictor for prokaryotic genomes. *J. Bioinform. Comput. Biol.*, 2009, 7(1): 19-38.

[97] Che D, Li G, Mao F, Wu H, Xu Y. Detecting uber-operons in prokaryotic genomes. *Nucleic Acids Res.*, 2006, 34(8): 2418-2427.

[98] Walker A K, See R, Batchelder C, Kophengnavong T, Gronniger J T, Shi Y, Blackwell T K. A conserved transcription motif suggesting functional parallels between Caenorhabditis elegans SKN-1 and Cap'n'Collar-related basic leucine zipper proteins. *J. Biol. Chem.*, 2000, 275(29): 22166-22171.

[99] Musso G, Zhang Z, Emili A. Retention of protein complex membership by ancient duplicated gene products in budding yeast. *Trends Genet.*, 2007, 23(6): 266-269.

[100] Wang T, Furey T S, Connelly J J, Ji S, Nelson S, Heber S, Gregory S G, Hauser E R. A general integrative genomic feature transcription factor binding site prediction method applied to analysis of USF1 binding in cardiovascular disease. *Hum. Genomics*, 2009, 3(3): 221-235.

[101] Conesa A, Gotz S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics.*, 2008, 2008: 619832.

[102] Yan B, Methe B A, Lovley D R, Krushkal J. Computational prediction of conserved operons and phylogenetic footprinting of transcription regulatory elements in the metal-reducing bacterial family Geobacteraceae. *J. Theor. Biol.*, 2004, 230(1): 133-144.

[103] Sharon I, Davis J V, Yona G. Prediction of protein-protein interactions: A study of the co-evolution model. *Methods Mol. Biol.*, 2009, 541: 61-88.

[104] Ventura M, Turroni F, Canchaya C, Vaughan EE, O'Toole PW, van Sinderen D. Microbial diversity in the human intestine and novel insights from metagenomics. *Front. Biosci.*, 2009, 14: 3214-3221.

[105] Jaeger C, Hatziagelaki E, Petzoldt R, Bretzel R G. Comparative analysis of organ-specific autoantibodies and celiac disease—Associated antibodies in type 1 diabetic patients, their first-degree relatives, and healthy control subjects. *Diabetes Care*, 2001, 24(1): 27-32.

[106] Morita M, Shibuya M, Kushiro T, Masuda K, Ebizuka Y. Molecular cloning and functional expression of triterpene synthases from pea (Pisum sativum) new alpha-amyrin-producing enzyme is a multifunctional triterpene synthase. *Eur. J Biochem.*, 2000, 267(12): 3453-3460.

[107] Bader M, Abouelhoda M I, Ohlebusch E. A fast algorithm for the multiple genome rearrangement problem with weighted reversals and transpositions. *BMC Bioinformatics*, 2008, 9: 516.

[108] Jiang X F, Yang J. A novel approach to predict protein-protein interactions related to alzheimer's disease based on complex network. *Protein Pept. Lett.*, Sept. 2009.

[109] Moriya Y, Itoh M, Okuda S, Yoshizawa A C, Kanehisa M. KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.*, 2007, 35(Web Server Issue): W182-W185.

[110] Berglund A C, Sjolund E, Ostlund G, Sonnhammer E L. InParanoid 6: Eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res.*, 2008, 36(Database Issue): D263-D266.

[111] Tatusov R L, Fedorova N D, Jackson J D, Jacobs A R, Kiryutin B, Koonin E V, Krylov D M, Mazumder R, Mekhedov S L, Nikolskaya A N *et al.* The COG database: An updated version includes eukaryotes. *BMC Bioinformatics*, 2003, 4: 41.

[112] Lathe W C, 3rd, Snel B, Bork P. Gene context conservation of a higher order than operons. *Trends Biochem. Sci.*, 2000, 25(10): 474-479.

[113] Karlin S, Mrazek J, Ma J, Brocchieri L. Predicted highly expressed genes in archaeal genomes. *Proc. Natl. Acad. Sci. USA*, 2005, 102(20): 7303-7308.

[114] Cormen T H, Leiserson C E, Rivest R L, Stein C. Introduction to Algorithms, Second Edition. Cambridge, MA: The MIT Press, 2001.

[115] Fani R, Brilli M, Lio P. The origin and evolution of operons: The piecewise building of the proteobacterial histidine operon. *J. Mol. Evol.*, 2005, 60(3): 378-390.

[116] Su Z, Mao F, Dam P, Wu H, Olman V, Paulsen I T, Palenik B, Xu Y. Computational inference and experimental validation of the nitrogen assimilation regulatory network in cyanobacterium Synechococcus sp. WH 8102. *Nucleic Acids Res.*, 2006, 34(3): 1050-1065.

[117] Salgado H, Gama-Castro S, Martinez-Antonio A, Diaz-Peredo E, Sanchez-Solano F, Peralta-Gil M, Garcia-Alonso D, Jimenez-Jacinto V, Santos-Zavaleta A, Bonavides-Martinez C *et al.* RegulonDB (version 4.0): Transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. *Nucleic Acids Res.*, 2004, 32(Database Issue): D303-D306.

[118] De Hoon M J, Imoto S, Kobayashi K, Ogasawara N, Miyano S. Predicting the operon structure of Bacillus subtilis using operon length, intergene distance, and gene expression information. *Pac. Symp. Biocomput.*, 2004, 9: 276-287.

[119] Lin H Y, Bledsoe P J, Stewart V. Activation of yeaR-yoaG operon transcription by the nitrate-responsive regulator NarL is independent of oxygen-responsive regulator Fnr in Escherichia coli K-12. *J. Bacteriol.*, 2007, 189(21): 7539-7548.

[120] Barthelmebs L, Lecomte B, Divies C, Cavin J F. Inducible metabolism of phenolic acids in Pediococcus pentosaceus is encoded by an autoregulated operon which involves a new class of negative transcriptional regulator. *J. Bacteriol.*, 2000, 182(23): 6724-6731.

[121] Dale C J, Moses E K, Ong C C, Morrow C J, Reed M B, Hasse D, Strugnell R A. Identification and sequencing of the groE operon and flanking genes of Lawsonia intracellularis: Use in phylogeny. *Microbiology*, 1998, 144(Pt 8): 2073-2084.

[122] Bockhorst J, Craven M, Page D, Shavlik J, Glasner J. A Bayesian network approach to operon prediction. *Bioinformatics*, 2003, 19(10): 1227-1235.

[123] Kowarz L, Robbe-Saule V, Norel F. Identification of cis-acting DNA sequences involved in the transcription of the virulence regulatory gene spvR in Salmonella typhimurium. *Mol. Gen. Genet.*, 1996, 251(2): 225-235.

[124] Mrazek J, Karlin S. Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl. Acad. Sci. USA*, 1998, 95(7): 3720-3725.

[125] Yachie N, Arakawa K, Tomita M. On the interplay of gene positioning and the role of rho-independent terminators in Escherichia coli. *FEBS Lett.*, 2006, 580(30): 6909-6914.

[126] Bockhorst J, Qiu Y, Glasner J, Liu M, Blattner F, Craven M. Predicting bacterial transcription units using sequence and expression data. *Bioinformatics*, 2003, 19(Suppl 1): i34-i43.

[127] Stormo G D, Hartzell G W, 3rd. Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad .Sci. USA*, 1989, 86(4): 1183-1187.

[128] Bailey T L, Boden M, Buske F A, Frith M, Grant C E, Clementi L, Ren J, Li W W, Noble W S. MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res.*, 2009, 37(Web Server Issue): W202-W208.

[129] Liu X, Brutlag D L, Liu J S. BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, 2001, 6: 127-138.

[130] Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen B A, Johnston M. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science*, 2003, 301(5629): 71-76.

[131] Blanchette M, Tompa M. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, 2002, 12(5): 739-748.

[132] Wu H, Mao F, Olman V, Xu Y. On application of directons to functional classification of genes in prokaryotes. *Comput. Biol. Chem.*, 2008, 32(3): 176-184.

[133] Wu H, Mao F, Olman V, Xu Y. Hierarchical classification of functionally equivalent genes in prokaryotes. *Nucleic Acids Res.*, 2007, 35(7): 2125-2140.

[134] Bowers P M, Cokus S J, Eisenberg D, Yeates T O. Use of logic relationships to decipher protein network organization. *Science*, 2004, 306(5705): 2246-2249.

[135] Jiang T, Keating A E. AVID: An integrative framework for discovering functional relationships among proteins. *BMC Bioinformatics*, 2005, 6: 136.

[136] Yu C, Zavaljevski N, Desai V, Johnson S, Stevens F J, Reifman J. The development of PIPA: An integrated and automated pipeline for genome-wide protein function annotation. *BMC Bioinformatics*, 2008, 9: 52.

[137] Aoki-Kinoshita K F, Kanehisa M. Gene annotation and pathway mapping in KEGG. *Methods Mol. Biol.*, 2007, 396: 71-91.

[138] Caspi R, Foerster H, Fulcher C A, Hopkinson R, Ingraham J, Kaipa P, Krummenacker M, Paley S, Pick J, Rhee S Y*et al.* MetaCyc: A multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, 2006, 34(Database Issue): D511-D516.

[139] Buckler D R, Zhou Y, Stock A M. Evidence of intradomain and interdomain flexibility in an OmpR/PhoB homolog from Thermotoga maritima. *Structure*, 2002, 10(2): 153-164.

[140] Perez E, Samper S, Bordas Y, Guilhot C, Gicquel B, Martin C. An essential role for phoP in Mycobacterium tuberculosis virulence. *Mol. Microbiol.*, 2001, 41(1): 179-187.

[141] Hengge R. The two-component network and the general stress sigma factor RpoS (sigma S) in Escherichia coli. *Adv. Exp. Med. Biol.*, 2008, 631: 40-53.

[142] Campbell E A, Westblade L F, Darst S A. Regulation of bacterial RNA polymerase sigma factor activity: A structural perspective. *Curr. Opin. Microbiol.*, 2008, 11(2): 121-127.

[143] Germer J, Becker G, Metzner M, Hengge-Aronis R. Role of activator site position and a distal UP-element half-site for sigma factor selectivity at a CRP/H-NS-activated sigma(s)-dependent promoter in Escherichia coli. *Mol. Microbiol.*, 2001, 41(3): 705-716.

[144] Colland F, Barth M, Hengge-Aronis R, Kolb A. Sigma factor selectivity of Escherichia coli RNA polymerase: Role for CRP, IHF and lrp transcription factors. *EMBO J.*, 2000, 19(12): 3028-3037.

[145] Kivistik P A, Kivi R, Kivisaar M, Horak R. Identification of ColR binding consensus and prediction of regulon of ColRS two-component system. *BMC Mol. Biol.*, 2009, 10: 46.

[146] Munch R, Hiller K, Grote A, Scheer M, Klein J, Schobert M, Jahn D. Virtual footprint and PRODORIC: An integrative framework for regulon prediction in prokaryotes. *Bioinformatics*, 2005, 21(22): 4187-4189.

[147] Yellaboina S, Ranjan S, Chakhaiyar P, Hasnain S E, Ranjan A. Prediction of DtxR regulon: Identification of binding sites and operons controlled by Diphtheria toxin repressor in Corynebacterium diphtheriae. *BMC Microbiol.*, 2004, 4: 38.

[148] Dombrecht B, Marchal K, Vanderleyden J, Michiels J. Prediction and overview of the RpoN-regulon in closely related species of the Rhizobiales. *Genome Biol.*, 2002, 3(12): RESEARCH0076.

[149] Smith A D, Sumazin P, Xuan Z, Zhang M Q. DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc. Natl. Acad. Sci. USA*, 2006, 103(16): 6275-6280.

[150] Jacob F, Monod J. On the regulation of gene activity. *Cold Spring Harbor Symposia on Quantitative Biology*, 1961, 26: 193-211.

[151] Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P, Goto S, Kanehisa M. KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res.*, 2008, 36(Web Server Issue): W423-W426.

[152] Yin Y, Zhang H, Xu Y. A governing rule for gene arrangement at a global scale in bacterial genomes. submitted, 2009.

[153] Faith J J, Driscoll M E, Fusaro V A, Cosgrove E J, Hayete B, Juhn F S, Schneider S J, Gardner T S. Many microbe microarrays database: Uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.*, 2008, 36(Database Issue): D866-D870.

[154] Langille M G I, Zhou F, Fedynak A, Hsiao W W L, Xu Y, Brinkman F S L. Mobile Genetic Elements and Their Prediction. Computational Methods for Understanding Bacterial and Archaeal Genomes, Xu Y, Gogarten J P (eds.), London: Imperial College Press, 2008, pp.113-136.

[155] Gogarten J P, Zhaxybayeva O. Horizontal Gene Transfer: Its Detection and Role in Microbial Evolution. Computational Methods for Understanding Bacterial and Archaeal Genomes, Xu Y, Gogarten J P (eds.), London: Imperial College Press, 2008, pp.137-152.

[156] Vitte C, Panaud O. LTR retrotransposons and flowering plant genome size: Emergence of the increase/decrease model. *Cytogenet Genome Res.*, 2005, 110(1-4): 91-107.

[157] Craig N L, Craigie R, Gellert M, Lambowitz A M. Mobile DNA II. Washington DC: American Society for Microbiology, 2002.

[158] Bestor T H. Transposons reanimated in mice. *Cell*, 2005, 122(3): 322-325.

[159] Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: The reference centre for bacterial insertion sequences. *Nucleic Acids Res.*, 2006, 34(Database Issue): D32-D36.

[160] Chandler M, Mahillon J. Insertion Sequences Revisited. 2nd Ed, Washington DC: American Society of Microbiology, 2002.

Ying Xu is the Regents-Georgia Research Alliance Eminent Scholar Chair and Professor in the Department of Biochemistry and Molecular Biology and the Institute of Bioinformatics, the University of Georgia. He received his Ph.D. degree in theoretical computer science from the University of Colorado at Boulder in 1991. He was a visiting assistant worked for Oak Ridge National Laboratory from 1993 to 2003, where he was a senior staff scientist and a group leader. His current research interests include (a) computational and systems biology relevant to human cancer and early detection, (b) microbial genomes and encoded pathways, and (c) plant genomes and plant cell walls. He has published over 200 research articles and four books covering different areas of bioinformatics and systems biology.