Goal 2

Characterize Gene Regulatory Networks

Background and Strategy

Gene regulatory networks govern which genes are expressed in a cell at any given time, how much product is made from each one, and the cell's responses to diverse environmental cues and intracellular signals (see figure, p. 31). Among the myriad cellular outputs from such networks are the metabolic capabilities of microbes and the responses of cells to environmental stresses, toxins, and low doses of radiation, all topics of central importance to DOE bioscience missions. Because gene regulatory networks (GRNs) are so central to understanding and manipulating cells, they are critical to DOE's biology missions. For microbial systems, Genomes to Life will widen the scope of network analysis to encompass comprehensive mapping of all regulatory networks, including the "circuitry" that operates without altering gene expression.

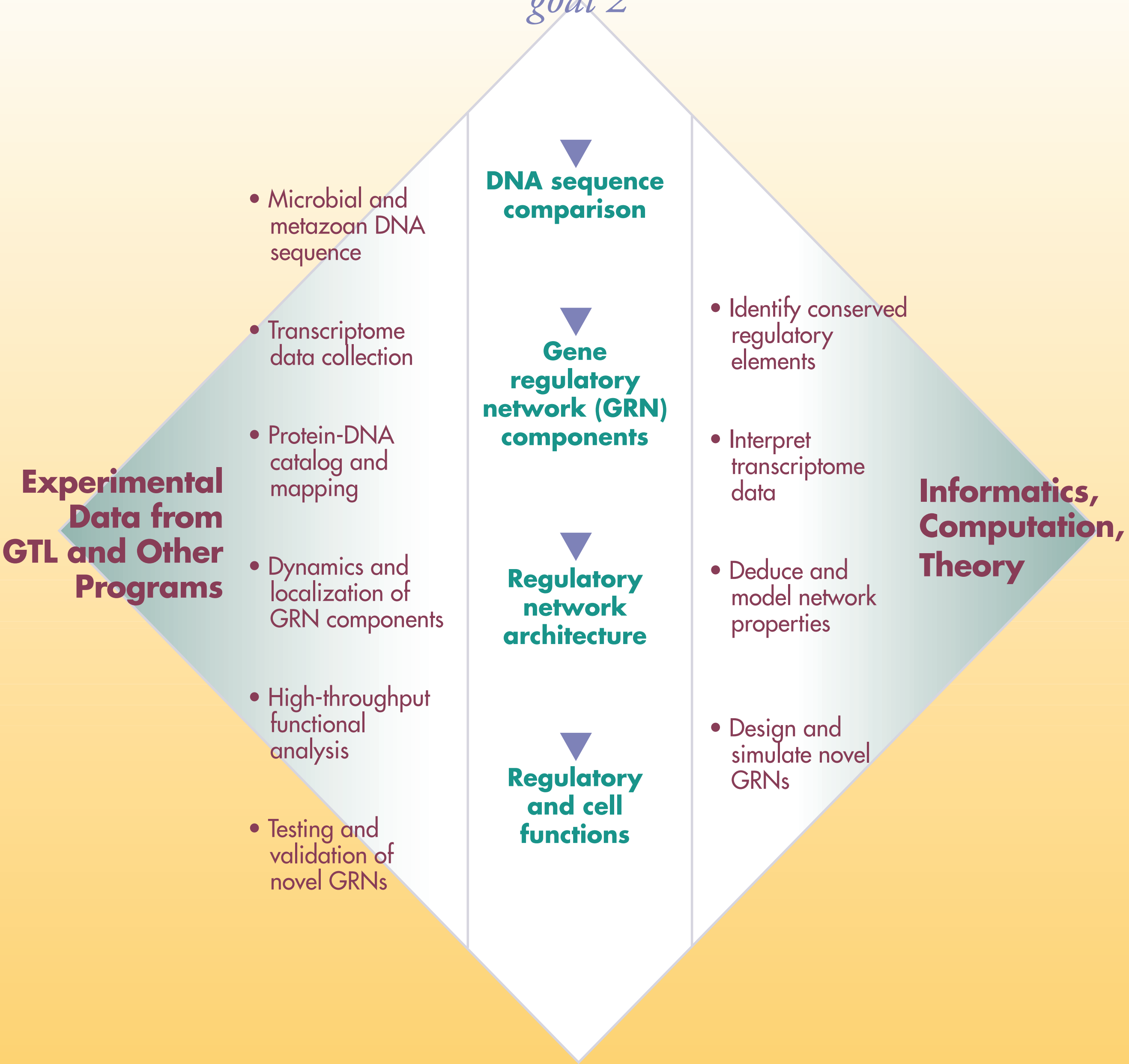
Major objectives for Goal 2 are to discover the architecture, dynamics, and function of regulatory networks; make useful computational models of them; and learn how to adapt and design them. Because the theory and modeling of regulatory networks represent the core of a new discipline, Genomes to Life will also emphasize the recruitment and education of a cadre of regulatory biologists who specialize in the computational modeling and theory of regulatory networks that are intimately coupled with cycles of experimental testing and verification.

Within the network discovery portion of Goal 2, one activity is to map related networks at multiple nodes across phylogeny based on comparison of genome sequences. Knowledge of comparative network structure and function is likely to produce insights into fundamental issues in biology, in addition to providing essential information for later phases of Genomes to Life. One such basic question has emerged from the Human Genome Project: How can a multicellular organism as complex as a human, with all its cell and tissue types and functions, use only 2 or 3 times as many genes (about 30,000) as the simple worm and 5 to 10 times as many as a single-cell microbe? Much of the answer may be in the regulatory network architecture and complexity. The cis-acting regulatory apparatus (see sidebar, pp. 32–33) and, by implication, the gene regulatory networks of which it is a critical part are said to be at the nexus between evolution and development [C. H. Yuh et al., *Science* **279**(5358), 1896–1902 (1998)]. Complex body forms may therefore have emerged during evolution due mainly to the appearance

GENOMES to LIFE

CHARACTERIZE GENE REGULATORY NETWORKS

goal 2



Experimental Data from GTL and Other Programs

- Microbial and metazoan DNA sequence
- Transcriptome data collection
- Protein-DNA catalog and mapping
- Dynamics and localization of GRN components
- High-throughput functional analysis
- Testing and validation of novel GRNs

DNA sequence comparison

Gene regulatory network (GRN) components

Regulatory network architecture

Regulatory and cell functions

- Identify conserved regulatory elements
- Interpret transcriptome data
- Deduce and model network properties
- Design and simulate novel GRNs

Informatics, Computation, Theory

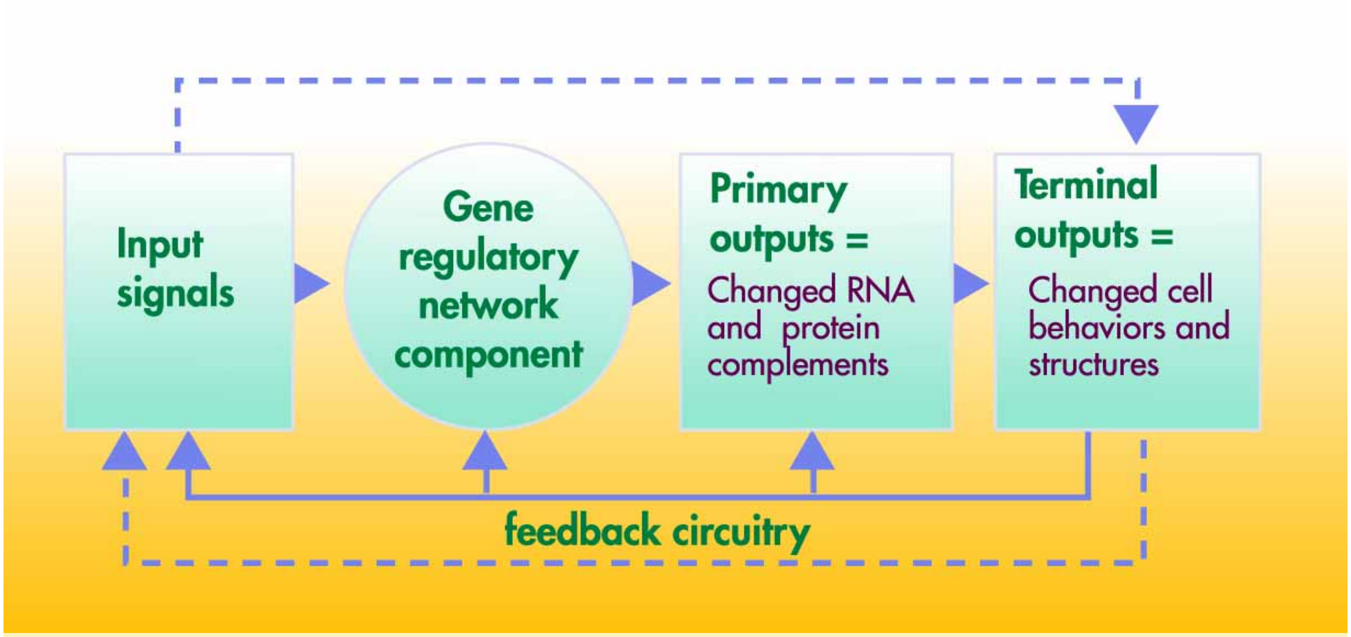
of more complicated and varied GRNs capable of controlling exquisitely complex combinatorial patterns of gene expression while the repertoire of genes itself is rather modestly expanded from much simpler organisms. If proved correct, an intriguing extension of this idea is that changes in the "wiring" of such networks might also dominate the functionally important differences between, for example, humans and our nearest animal relatives, the chimpanzees. Tracing the evolution of regulatory networks rather than individual genes will yield the information needed to probe these possibilities.

Specific Aims

Aim 1. Develop the capability within the next 10 years to comprehensively map microbial and metazoan regulatory circuitries. Use this capability to construct detailed GRN maps for specific subgenomic networks positioned across multiple species and to build comprehensive regulatory circuitry maps at the whole-genome level for selected microbes

Initial tasks for Aim 1 will be to identify and map core gene regulatory network components. In metazoa a major focus will be to identify cis-acting regulatory sequences in the genome and the regulatory proteins that interact with them. Integral to this effort is the task of relating the regulatory apparatus to the groups of target genes they regulate and to whatever is known about the function of those target genes. To map GRNs, several core technologies and approaches will likely be applicable to both microbial and eukaryotic systems, although pilot studies are needed to further define the best approach to use in genomes of varying sizes and structures. One such promising strategy is to use comparative genomics to initiate large-scale GRN component identification, focusing on candidate cis-regulatory sequences and the regulatory proteins with which they interact. Results from comparative sequence analysis would then be integrated with data from other key technologies such as large-scale gene-expression analysis, comprehensive loss-of-function and gain-of-function genetic analyses, and measures of in vivo protein-DNA interactions, and proteome status, among others.

Other critical elements in network mapping will come from activities encompassed by Goal 1 or by specific adaptation of technology developed in that work to regulatory network components. This includes learning the composition of multiprotein complexes that assemble on DNA to regulate gene expression; learning the composition and regulatory actions of protein machinery that govern


Input signals
Gene regulatory network component
Primary outputs = Changed RNA and protein complements
Terminal outputs = Changed cell behaviors and structures
feedback circuitry

Gross anatomy of a minimal gene regulatory network (GRN) embedded in a regulatory network. A regulatory network can be viewed as a cellular input-output device. At minimum, a gene regulatory network typically contains the following components: (1) an input signal reception and transduction system that mediates intra- and extracellular cues (left box; often, more than one signal impinges on a given target gene); (2) a "core component" complex composed of trans-acting regulatory proteins and cognate cis-acting DNA sequences (circle; functionally similar components may be associated with multiple target genes, resulting in similar gene-expression patterns); and (3) primary molecular outputs from target genes, which are RNA and protein (box to right of circle). The net effects are changes in cell phenotype and function (right box). Direct and indirect feedbacks typically are important. More realistic networks often feature multiple tiers of regulation, with first-tier gene products regulating expression of another group of genes, and so on. Beyond GRN boundaries are signaling responses and feedbacks, such as those that drive bacterial chemotaxis, which do not involve regulation of gene expression but instead act directly on proteins and protein machine assemblies (dashed arrows). Some regulatory networks have no embedded GRN component.

post-transcriptional and post-translational regulation, and determining subcellular localization of regulatory proteins and how that localization changes as a function of circuit dynamics.

Vigorous application of a comprehensive genome-wide approach to network mapping in selected microbes has the potential to yield the first complete dissection of the regulatory networks that run a living cell. The most experimentally advanced microbial systems already offer a uniquely powerful combination of approaches that are variously performed in vivo, in vitro, and in silico on a comprehensive genome-wide scale. The result is integrative knowledge of the subsystems and systems contained in bacterial gene regulatory circuitry [M. T. Laub et al., *Science* **290**, 2144–48 (2000) and Goal 2 sidebar, p. 33]. Moreover, recent discoveries clearly show that regulatory networks in both microbes and metazoa employ many mechanisms distinct from both transcription and translation. Examples include active control of protein turnover, dynamic localization of regulatory and structural

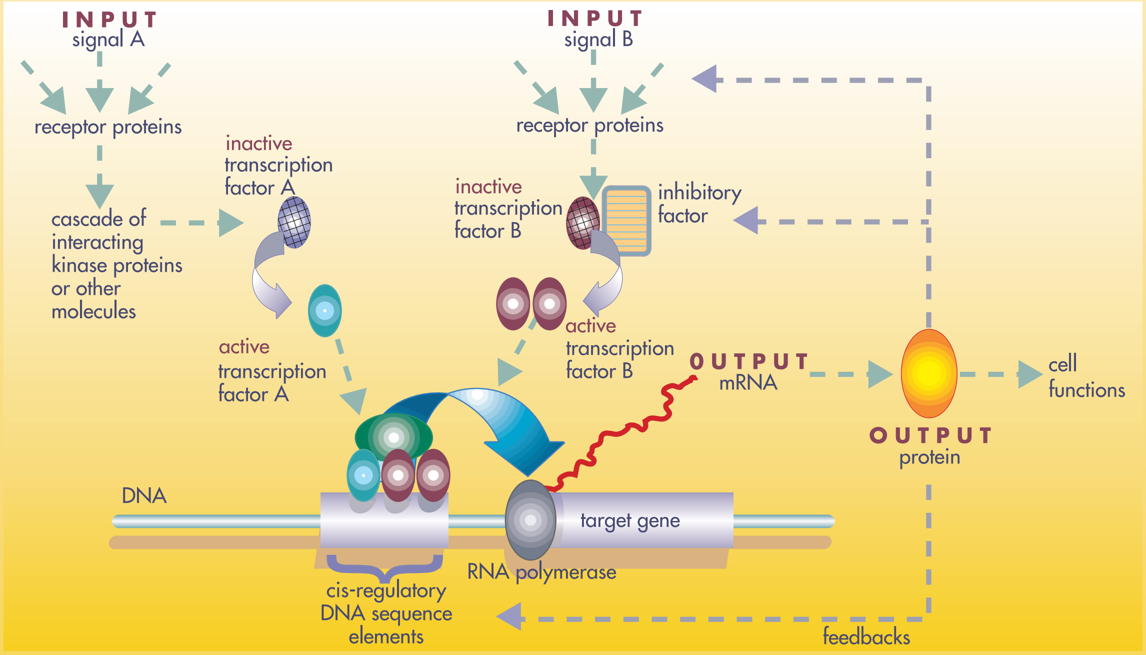
Gene Regulatory Networks

Gene regulatory networks (GRNs) are the on-off switches and rheostats of a cell operating at the gene level. They dynamically orchestrate the level of expression for each gene in the genome by controlling whether and how vigorously that gene will be transcribed into RNA. Each RNA transcript then functions as the template for synthesis of a specific protein by the process of translation. A simple GRN would consist of one or more input signaling pathways, regulatory proteins that integrate the input signals, several target genes (in bacteria a target operon), and the RNA and proteins produced from those target genes. In addition, such networks often include dynamic feedback loops that provide for further regulation of network architecture and output. As indicated in the schematic below, input signaling pathways transduce intracellular and/or extracellular signals to a group of regulatory proteins called transcription factors. Transcription factors activated by the signals then interact, either directly or indirectly, with DNA sequences belonging to the specific genes they regulate. The factors also interact with each other to form multiprotein complexes bound to the DNA.

GRNs act as analog biochemical computers to specify the identity and level of expression of groups of target genes. Central to this computation are DNA recognition sequences with which transcription factors associate. Every gene has its own novel "cis-acting" sequence elements. They vary greatly in complexity from one gene to another and from generally simpler structures in bacteria to more complex structures in multicellular organisms. When active transcription factors associate with the cis-elements of their cognate target genes, they can function to specifically repress (down-regulate) or induce (up-regulate) synthesis of the corresponding RNA. The immediate molecular output of a gene regulatory network is the constellation of RNAs and proteins encoded by network target genes. The resulting cellular readouts are changes in the structure, metabolic capacity, or behavior of the cell mediated by new expression of up-regulated proteins and elimination of down-regulated proteins.

GRNs are remarkably diverse in their structure, but several basic properties are illustrated in the figure below. In this example, two different signals impinge on a single target gene where the cis-regulatory elements provide for an integrated output in response to the two inputs. Signal molecule A triggers the conversion of inactive transcription factor A (green oval) into an active form that binds directly to the target gene's cis-regulatory sequence. The process for signal B is more complex. Signal B triggers the separation of inactive B (red oval) from an inhibitory factor (yellow rectangle). B is then free to form an active complex that binds to the active A transcription factor on the cis-regulatory sequence. The net output is expression of the target gene at a level determined by the action of factors A and B. In this way, cis-regulatory DNA sequences, together

A GENE REGULATORY NETWORK


INPUT
signal A
receptor proteins
cascade of interacting kinase proteins or other molecules
inactive transcription factor A
active transcription factor A
INPUT
signal B
receptor proteins
inactive transcription factor B
inhibitory factor
active transcription factor B
OUTPUT
mRNA
OUTPUT
protein
cell functions
DNA
target gene
RNA polymerase
cis-regulatory DNA sequence elements
feedbacks

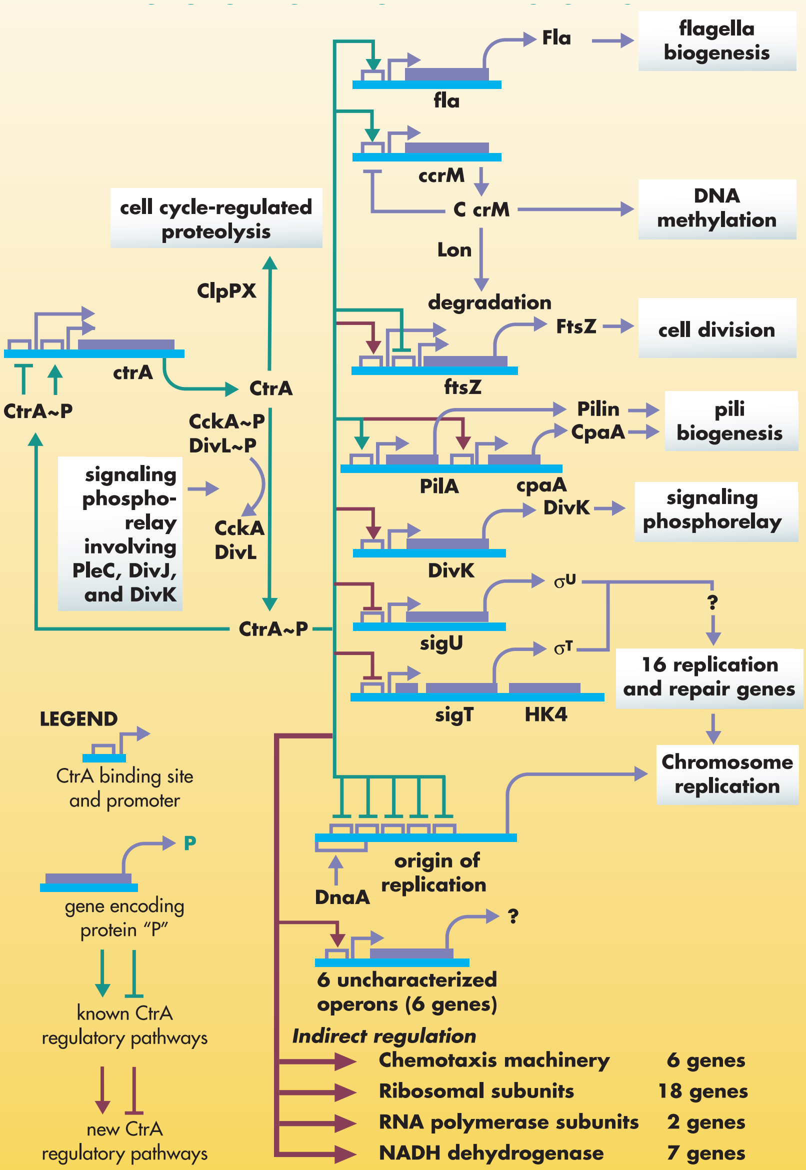
with the proteins that assemble on them, integrate information from multiple signaling inputs to produce an appropriately regulated readout. A more realistic network might contain multiple target genes regulated by signal A alone, others by signal B alone, and still others by the pair of A and B.

Co-regulated target genes often code for proteins that act together to build a specific cell structure or to effect a concerted change in cell function. For example, genes encoding components of the multiprotein proteasome machine (see Goal 1 sidebar, pp. 22–23) are co-regulated at the RNA level. This was shown by microarray gene chip analyses in yeast cells, and each gene was found to possess a similar cis-regulatory DNA sequence that mediates binding of a particular transcription factor. Similarly, a bacterium may respond to a shortage of its preferred energy source by activating expression of genes whose protein products function in a biochemical pathway that allows it to use a different, more abundant source of energy.

Some genes are regulated by a single input mechanism, but, especially in higher organisms, a gene often responds to information from multiple signals via the activity of diverse transcription factors. For example, in human fibroblast cells responding to a "growth factor" impinging on cell surface signal receptors, a platoon of "immediate early" genes is upregulated as the first step in the complex process of cell proliferation. Some of the same genes, though not all of them, can be activated in brain cells by the distinct stimulus of seizure. This network also illustrates that GRNs can be multitiered. The first signal (growth factor) initiates expression of "immediate early" target genes, which include transcription factors such as *c-Fos* and *c-Jun*. These transcription factors then cause a second group of target genes, called "early genes," to be expressed. Among these early genes are other transcription factors such as *c-Myc*. They regulate expression of yet another group of genes of a "delayed early" group. In this way a multitiered GRN cascade can be constructed, replete with feedbacks and crosstalk to other networks.

A major gene regulatory network in the bacterium *Caulobacter* is now beginning to be mapped in a comprehensive manner based on genome-wide expression analyses coupled with genetic methods [M. T. Laub et al., *Science* **290**, 2144–48 (2000)]. *Caulobacter* has about 3000 genes, of which almost 20% were found to be differentially expressed during the cell division cycle. Of the 553 responding genes, 38 are likely to be direct targets of a sequence-specific DNA-binding protein called CtrA and another 144 are indirectly regulated by CtrA. A first-pass connectivity map of the CtrA gene regulatory network derived from this study is summarized in this figure. Green indicates previously known relationships, while red indicates relationships that emerged from this global gene-expression study performed using microarray technology.

A *CAULOBACTER* CELL DIVISION GRN


fla
Fla
flagella biogenesis
ccrM
C crM
DNA methylation
Lon
degradation
cell cycle-regulated proteolysis
ClpPX
ctrA
CtrA
CtrA~P
ftsZ
FtsZ
cell division
CckA~P
DivL~P
PilA
cpaA
Pilin
CpaA
pili biogenesis
signaling phospho-relay involving PleC, DivJ, and DivK
CckA
DivL
DivK
DivK
signaling phosphorelay
CtrA~P
sigU
σU
?
sigT
σT
HK4
16 replication and repair genes
Chromosome replication
origin of replication
DnaA
?
6 uncharacterized operons (6 genes)
Indirect regulation
Chemotaxis machinery 6 genes
Ribosomal subunits 18 genes
RNA polymerase subunits 2 genes
NADH dehydrogenase 7 genes
LEGEND
CtrA binding site and promoter
P
gene encoding protein "P"
known CtrA regulatory pathways
new CtrA regulatory pathways

proteins, and complex phospho-transfer pathways. In addition, the cell membrane appears to be an integral component of essential cell signaling processes. Including nontranscriptional systems is therefore critical for a full understanding of regulatory circuitry in all organisms. Genomes to Life will first capitalize on the relative simplicity of microbes to extend network analysis to include all regulatory circuitry.

Aim 2. Verify regulatory circuit architecture and connect regulatory network properties with their biological outputs

Genomes to Life's ambitious next step will be to map higher-order connectivity between these circuits and tie them to cellular functions and phenotypes. This will begin with experimental verification of network composition and architecture generated by Aim 1. The most effective, efficient, and scalable methods for doing this for both microbes and more complex creatures will be explored in the early years of Genomes to Life. On the output side of the network equation, Genomes to Life will seek to assign cellular function and phenotype to each subsystem, and then to link subsystems to learn higher-order connectivity. This is an enormous challenge, the dimension of which will become apparent only when we gain a much more comprehensive view than we now have. Genomes to Life will therefore draw on the data and resources from this and other DOE programs as well as those generated by the National Institutes of Health and the National Science Foundation. Functional annotation in Genomes to Life will require large bioinformatics and computational components, some aspects of which will make demands considerably greater than that provided by whole-genome sequence assembly of the human genome (see Goal 4, p. 44).

Aim 3. Develop a theoretical framework and associated set of computational modeling tools to predict the dynamic behavior of natural or designed regulatory networks

Developing a comprehensive view of the basic architecture of eukaryotic gene regulatory or microbial regulatory networks will be the culmination of a discovery process that began 40 years ago with the *Escherichia coli* Lac operon. Although that accomplishment will be a grand one, even a complete wiring diagram will not reveal how any regulatory network really works, nor will it provide a solid basis for using them, for modifying them in useful ways, or for designing new ones. To master the complexities of regulatory switches, oscillators, and more complex functions will require a predictive theoretical framework and computational horsepower of teraflop speed. To meet this challenge, Genomes to Life will seek to nurture and accelerate emerging capabilities that include new concepts combined with relevant ideas from engineering, applied mathematics, and other disciplines.

Aim 4. Learn to modify natural networks and design novel ones for DOE mission purposes

A major motivation for mapping regulatory networks and then developing predictive computational models and simulators is to ultimately learn how to prudently use such networks to develop biological solutions to important problems such as bioremediation. This is a long-term goal that will require results from Aims 1–3 above to bring network modification and design onto a firm footing. However, even in the early years of Genomes to Life, developing the capability to design increasingly complex networks with useful control properties will be an important activity.

Computation Needs

The task of elucidating and adapting the circuitry of gene regulatory networks will be dependent on computational methods to identify and characterize regulatory sequences and computer models of the regulatory networks. The computational research tasks for Goal 2 involve developing methods to do the following:

- Extract regulatory elements, including operon and regulon sequences, using sequence-level comparative genomics.
- Simulate regulatory networks using both nondynamical models of regulatory capabilities and dynamical models of regulatory kinetics.
- Predict the behavior of modified or redesigned gene regulatory networks.