Bacterial Signal Transduction Modules: from Genomics to Biology

Bacterial signaling proteins consist of modular domains that combine to yield arrays of sensors, transducers, and responders

Michael Y. Galperin and Mark Gomelsky

Microbial genome sequencing has had only a modest impact on practical microbiology so far. It has not resulted, as many had hoped, in a bonanza of new antimicrobial drugs. The progress in basic understanding of pathogens has also been slow. For example, genomic sequences of the bacteria causing two most common sexually transmitted diseases, Treponema pallidum (syphilis) and Chlamydia trachomatis (vaginitis), were determined more than seven years ago. All their genes have been identified and key metabolic pathways are known, but we still cannot reliably grow either of these pathogens in a cell-free or defined medium. Since every analyzed genome includes numerous genes with unknown functions, a “complete understanding” of any bacterial cell in the framework of systems biology is as yet a murky possibility. In stark contrast to the relatively slow progress in these areas, microbial signal transduction presents a case where genomic data sets have lead to visible progress.

The sheer complexity of signal transduction systems, even in model organisms such as Escherichia coli, Bacillus subtilis, and Saccharomyces cerevisiae, has long hindered their systematic analysis. Studies of specific processes, such as nitrogen regulation, chemotaxis, catabolite repression, and sporulation, provide valuable insights into the mechanisms of sensing and signal transduction. However, there had been no easy way to assemble these diverse regulatory systems into a coherent picture. Comparative genomics has changed that by allowing us to predict all regulatory components in a given organism and to compare their organization in various organisms. It has also revealed how much we still do not know about microbial signal transduction.

Protein Domains: the Building Blocks of Life

One of the major findings of comparative genomics is that microbial signaling systems are far more diverse than we ever expected. For example, of 30 sensor histidine kinases encoded by E. coli K-12 and 36 encoded by B. subtilis, only 3 are shared by both organisms: the citrate sensor CitA/CitS, the dicarboxylate sensor DcuS, and the chemotaxis protein CheA. The remaining kinases are present in either one bacterium or the other. Hence, while many sensor...
kinases of *E. coli* have been extensively studied experimentally, analysis of their sequences has been hampered by the absence of close homologs in other model organisms. With data for more than 20 genomes of *Enterobacteriaceae* members and 13 representatives of *Bacillaceae* available, there is now ample opportunity for cross-genome comparisons.
An alignment of closely related proteins reveals a mosaic of highly conserved and less-conserved regions (Fig. 1). The former are often referred to as protein motifs. These are the easiest to recognize in sequence comparisons. Protein motifs often correspond to secondary structure elements—such as α-helices, β-strands or loops—that carry functionally important (hence, conserved) amino acid residues. However, these elements cannot exist on their own, needing additional structural elements to help them to fold into the proper three-dimensional structures. Such autonomously folding discrete structural units of proteins, which typically contain two or more conserved sequence motifs, are referred to as protein domains. Protein domains are usually 100–200 amino acid residues in length, but can be as short as 25 residues and as long as 500. They evolve as single units, so that when their coding sequences are duplicated, translocated within a genome, or transferred across genomes, the domain structure and associated function, such as catalytic activity or ligand binding, typically do not change. Occasionally, however, protein domains lose their activity or gain new functions.

Thanks to many contributors, the most common protein domains are now catalogued in public databases, such as Pfam, SMART, InterPro, and CDD. Each of them allows researchers to identify known domains in a newly sequenced protein. As new genomic data are added, deduced proteins are automatically analyzed for domain content. Thus, protein domain databases offer a treasure trove of information on domain structure, distribution, protein architectures, and domain-domain interactions. That is why computer-assisted sequence analysis may bring insights beyond what experiments can reveal. Of course, sequence analysis relies on experimental data to assign functions to the newly described domains. Otherwise, they are consigned to the category of domains of unknown function (http://www.sanger.ac.uk/Software/Pfam/browse/DUF.shtml).

Microbial signal transduction proteins typically contain several distinct domains (Fig. 1). This modular organization of these proteins leads to the generation of an impressive diversity of regulatory systems by assembling a limited number of protein domains in a variety of combinations, much as a toddler puts together a series of Lego blocks. Many such combinations do not serve a useful purpose and eventually get lost in the course of evolution. Others get fixed in evolution, providing the host with specific means of sensing and responding to the environment. Since the basic functions are embedded in the protein domains, these domains—and not full-length proteins—should really be considered the building blocks of life.

Lessons from the First Recognized Signaling Domain

The modular architecture of signaling proteins was first recognized about 20 years ago and paved the way to the discovery of two-component signal transduction, a key advance in understanding bacterial behavior and regulation. In 1984, Philip Matsumura and his coworkers at the University of Illinois, Chicago, determined the sequence of the *E. coli* chemotaxis response regulator CheY. At that time it was cumbersome to compare protein sequences even though such information was beginning to accumulate—for example, at the Protein Identification Resource (PIR) database, headed by Margaret Dayhoff of Georgetown University in Washington, D.C.

The breakthrough came the next year, when James Hoch and colleagues at the Scripps Research Institute in La Jolla, Calif., sequenced *B. subtilis* sporulation proteins Spo0A and Spo0F and realized that Spo0F was similar to the N-terminal half of Spo0A. Further, using the just-released FASTP sequence comparison program, they found out that Spo0F was also similar to the N-terminal regions of the previously sequenced transcriptional regulators OmpR and ArcA (at the time known as SfrA or Dye) from *E. coli* and concluded that all these protein fragments were homologous.

By the end of 1985, Ann and Jeffry Stock at Princeton University in Princeton, N.J., and Daniel Koshland at the University of California, Berkeley, sequenced the CheY and CheB proteins from *Salmonella enterica* serovar Typhimurium, identified the CheY-like N-terminal fragment of the chemotaxis methylesterase CheB as its regulatory domain, and further linked this domain to the N-terminal portions of *E. coli* OmpR and ArcA and *B. subtilis* Spo0A. “These homologies suggest an evolutionary and functional relationship between the chemotaxis system and systems that are thought to regulate gene expression in response to changing environ-
mental conditions,” concluded Koreshland and the Stocks. Given that sequence identity between different CheY-like receiver domains was only about 20%, which at that time was generally considered less than convincing, this was a prescient and bold conclusion (Fig. 2).

Within the next two years several groups of researchers proved that various CheY-like domains in their different contexts uniformly act as phosphoacceptor domains. In 1989, the three-dimensional structure of the CheY protein was determined, confirming that it is an autonomously folding compact structural unit and revealing that its few conserved residues are close to each other and form the active site. This made the CheY-like receiver domain the first recognized signaling domain.

Properties of this domain in different proteins from diverse bacteria proved to be very similar, which provided some leeway in choosing the experimental model. Thus, the first structure of an OmpR-like response regulator used a protein from the thermophilic bacterium *Thermotoga maritima*, a poor object for chemotaxis studies but a perfect source of thermostable proteins. The CheY domain story has taught us valuable lessons that are applicable to studies of other signaling domains: (i) This domain was first recognized by a (relatively low) sequence similarity as a common module in otherwise unrelated proteins. (ii) Subsequent biochemical and structural characterization of this domain confirmed conservation of its structure and function. (iii) Characterization of the CheY domain had a profound impact on our understanding of signal transduction systems in all microorganisms where this domain is present.

Many other signal transduction domains are now in the same position that the CheY domain was in 20 years ago—recognized by sequence similarities, while biochemical and structural characterizations lag behind. The history of the CheY domain illustrates the value of “domain-centric” approaches which study sensory transduction modules across the whole phylogenetic space, focusing when need be on those organisms and processes that prove most suitable for particular purposes.

**Impact of Comparative Genomics on Studies of Signal Transduction**

As genome sequences accumulated, cross-genome comparisons showed that histidine kinases and methyl-accepting chemotaxis proteins (MCPs) are not the only types of receptor molecules present in bacteria (Fig. 3). A characteristic receptor architecture, which includes a periplasmic sensory domain, one or more transmembrane segments, and a cytoplasmically located output domain, was predicted for many proteins encoded in the newly sequenced genomes. However, their C-terminal output do-
mains were not histidine kinases or MCPs, but adenylate cyclases, Ser/Thr protein kinases, diguanylate cyclases, or phosphodiesterases (Fig. 4).

Although the role of adenylate cyclase in catabolite repression has been known for many years, adenylate cyclases were not known to be environmentally regulated. However, Masayuki Ohmori at the University of Tokyo (now at Saitama University) showed that activity of cyanobacterial membrane-bound adenylate cyclases can be modulated by red and blue light. Because cAMP is required for transcription of the cyanobacterial motility genes, these adenylate cyclases were recognized as legitimate trans-membrane sensors that regulate cellular motility in a light-dependent manner. Meanwhile, studying acute pneumonia caused by *Pseudomonas aeruginosa* in mice, Stephen Lory and colleagues at Harvard Medical School, Boston, Mass., found that its unique type 3 adenylate cyclase with a MASE2 integral membrane sensory domain plays an important role in infection by regulating expression of type III secretion system genes.

The cellular targets for most prokaryotic Ser/Thr kinases remain poorly characterized. Some of them are DNA-binding proteins, such as AfsR, a global regulator of secondary metabolism in *Streptomyces coelicolor*, EmbR, putative regulator of arabinan biosynthesis in *Mycobacterium tuberculosis*, and HU, a histone-like protein in *E. coli*. Others are metabolic enzymes, including phosphofructokinase in *Mycopoccus xanthus* and phosphoglucomutase from *Streptococcus pneumoniae*. In addition, certain Ser/Thr kinases regulate the bacterial cell cycle. It appears that Ser/Thr protein kinases affect processes on both transcriptional and posttranslational levels.

The diguanylate cyclases and cyclic diguanylate- or c-di-GMP-specific phosphodiesterases apparently act primarily at the posttranslational level. These proteins contain GGDEF and EAL domains, named after their conserved sequence motifs. The product of the diguanylate cyclase activity, c-di-GMP, was first identified by the late Moshe Benziman of the Hebrew University of Jerusalem, Israel, as an allosteric activator of cellulose synthase in *Gluconacetobacter* (formerly *Acetobacter* *)xylinus*. However, the discovery of the GGDEF domain in the *Caulobacter crescentus* developmental regulator PleD by Austin Newton of Princeton University showed that this domain has a much wider role. Bacterial genome sequencing revealed a surprising abundance of GGDEF and EAL domains (Fig. 3), indicating that c-di-GMP-dependent regulation is more widespread than was first thought and establishing c-di-GMP as a novel second messenger in bacteria.

Studies of this novel signal transduction system got a significant boost last year, when Urs Jenal and colleagues at the University of Basel proved that the GGDEF domain of PleD has diguanylate cyclase activity and solved the 3D structure of this protein. Mark Gomelsky and colleagues at the University of Wyoming demonstrated the diguanylate cyclase activity of GGDEF domains from several phylogenetically diverse bacteria, and showed that the EAL domain has c-di-GMP-specific phosphodiesterase (that is, c-di-GMP-degrading) activity. Just in the past year, c-di-GMP was implicated in regulating transition between sessility and motility in *E. coli* and *S. enterica* serovar Typhimurium by...
Ute Römling of the Karolinska Institute in Stockholm, Sweden, in twitching motility in *Pseudomonas aeruginosa* by John Mattick of the University of Queensland, in biofilm formation in *Vibrio cholerae* by Andrew Camilli of Tufts University, Boston, Mass., in *Yersinia pestis* by Robert Perry of the University of Kentucky, Lexington, and in photosynthesis gene expression in *Synechococcus elongatus* by Susan Golden of Texas A&M University, College Station. Thus, after being overlooked for decades, the c-di-GMP-mediated signaling emerges as a very exciting area of research brought to prominence largely due to comparative genome analysis.

Not all cellular receptors sense environmental signals. For instance, some histidine kinases have no transmembrane segments, including the *E. coli* chemotaxis histidine kinase CheA, the nitrate sensor NtrB, and the *B. subtilis* sporulation regulators KinA, KinB, and KinC. In fact, almost a third of all histidine kinases and close to half of the known proteins that contain adenylate cyclase, diguanylate cyclase, c-di-GMP-specific phosphodiesterase, Ser/Thr protein kinase, or protein phosphatase domains are not membrane bound. Enzymatic activities of most of these sensory proteins are apparently modulated by intracellular, rather than extracellular, signals. Indeed, many of them contain N-terminal cytoplasmic sensor domains, often PAS or GAF, that affect the activities of the downstream domains. For example, Marie-Alda Gilles-Gonzalez, now at the University of Texas Southwestern Medical Center in Dallas, and colleagues have shown that c-di-GMP-specific phosphodiesterase activity of the *G. xylinus* protein PdeA1 of PAS-GAF-GGDEF-EAL domain architecture is regulated by oxygen binding to its heme-containing PAS domain. While the exact nature of the ligands of most soluble sensory domains remains obscure, there are indications that they monitor a variety of parameters, including intracellular oxygen, carbon monoxide, reactive nitrogen species, light, and...
Assessing Bacterial IQ

Careful accounting of signal transduction proteins in organisms from different phylogenetic lineages reveals several trends. While the total number of such proteins generally correlates with genome size, microbes with complex life styles tend to have more sophisticated and diverse regulatory systems than do parasites. The fraction of the genome dedicated to signal processing (a measure of bacterial “IQ”) is the largest in highly versatile environmental organisms, such as *Geobacter sulfurreducens*, *Desulfovibrio vulgaris*, or *Chromobacterium violaceum*, which can use a variety of electron acceptors. Many “low-IQ” pathogens are responsible for scores of human diseases, whereas “smart” bacteria are usually less harmful. However, sulfate reducers like *D. vulgaris* are largely responsible for corrosion of metals, inflicting untold economical losses.

In this context, *E. coli* K-12 appears not to be so “smart.” It encodes far fewer signal transduction systems than do some of its γ-proteobacterial relatives. For example, the deep-sea bacterium *Idiomarina loihiensis*, with less than two-thirds of the *E. coli* genome, encodes more diguanylate cyclases and three times as many MCPs. Delta-proteobacterium *Bdellovibrio bacteriovorus*, a predator that infects *E. coli* cells and grows in their periplasmic space, encodes almost twice as many histidine kinases and four times as many MCPs, despite having a smaller genome than *E. coli*.

Some pathogens, such as the spirochetes *Leptospira interrogans* and *TREPONEMA DENTICOLA*, encode fairly sophisticated sensory systems. What are they sensing and responding to? Sequence similarity leaves no doubt that the sole histidine kinase of *T. pallidum* is the chemotaxis protein CheA. To fight this pathogen, it would be helpful to learn what signals are perceived by its four MCPs and the single membrane-bound adenylate cyclase.

Looking Beyond the Two-Component Signal Transduction System

Faced with the incredible diversity of signaling systems encoded in microbial genomes, the classical scheme of the two-component signal transduction system composed of a sensor histidine kinase and its cognate response regulator increasingly appears too rigid and its universality continues to be challenged. For example, some kinases do not sense signals by themselves but rely on interactions with separate sensory proteins, as shown for UhpB/UhpC pair by Robert Kadner at the University of Virginia in Charlottesville and his colleagues. Hybrid kinases (such as ArcB, Fig. 2) contain both kinase and CheY domains and can catalyze both intramolecular and intermolecular phosphotransfer. Many kinases contain more than one sensory domain, such as PAS and GAF. Some response regulators contain more than one output domain. For example, the response regulator VieA from *V. cholerae* contains EAL and a DNA-binding domain as outputs.

The abundance of such deviations from the classical scheme suggests that microbial signal transduction may be better viewed as a network containing numerous sensory domains that interact with numerous signal transduction (output) domains. These sensory and transduction domains can be put together in various combinations. In such a network, a battery of receptor sensory domains receives environmental or intracellular signals and transmits them to produce varied outputs, including changes in gene expression, levels of secondary messengers, protein modifications, and so on—thereby modifying cellular metabolism and behavior.

Signaling Systems: Broad Patterns, Many Outstanding Questions

Looking at microbial signal transduction from the genomics perspective, we can make several generalizations:

- Microbial signal transduction is modular. Many, perhaps the majority, of sensory and output domains operating in prokaryotes have probably been identified, although new domains may be discovered when genomes of microorganisms with complex life styles, such as *Myxococcus xanthus* and *Gemmata obscuriglobus*, are analyzed.

- The major task now is to determine the structures of these domains and to elucidate their functions. Once a structure and function of a signal transduction domain is determined, regardless of the system or organism, it adds to our understanding of signal transduction in the microbial universe.
Among the recently emerged but poorly characterized signal transduction systems, protein phosphorylation by Ser/Thr kinases and c-di-GMP-mediated signaling look particularly intriguing.

We still do not (but need to) understand how different signals are integrated in proteins containing more than one sensory domain and how different outputs are regulated in proteins containing more than one output.

Though often misused, “signaling network” proved to be an accurate rubric. Microbial cells often have several different circuits responding to the same signal. How do these different circuits interact? Is there a hierarchy among the regulatory systems of different kinds? Why are signal transduction systems differentially present in various phylogenetic groups? While in silico approaches of comparative genomics are helping us to pose many of these questions, their answers will depend on experiments conducted in vitro and in vivo.

During the past 20 years, we have learned some common principles pertaining to signal transduction in diverse bacteria, and they provide a solid basis for analyzing signaling networks. Combined in silico, in vitro, and in vivo approaches will be integral to efforts to compile the “parts lists” of bacterial cells and understand how they operate, which is a major goal of systems biology.

Comparative genome analysis is providing us with valuable insights into microbial signal transduction, while also showing how little we really know. Nonetheless, the domain-centric approach is the fastest way to bringing signal transduction domains from sequence to function and to structure. Ultimately, this approach will help us understand how microbial cells sense and react to their environments. This knowledge is important for our ability to manipulate microbes in medical, biotechnological, and agricultural applications. It may also help us learn more about multicellular species, including ourselves, that keep using some of the sensory and output domains that evolved long time ago in their prokaryotic predecessors.

ACKNOWLEDGMENT
M. Gomelsky thanks the following granting agencies that have been supporting research on various signalling systems in his laboratory: NIH (NCRR P20 RR15640), NSF (MCV-0316270), and DOE (BER DE-FG02-01ER63232).

SUGGESTED READING