GUEST COMMENTARY

Digging with Experimental Pick and Computational Shovel: a New Addition to the Histidine Kinase Superfamily

Igor B. Zhulin*

School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30332-0230

Experimental microbiology in the postgenomic era has survived the first wave of uncertainty: a perspective of being replaced by in silico microbiology. Although it is clear now that experimental science is irreplaceable, we are facing yet another wave: the frustration of having a half million microbial genes in databases but lacking simple ways of getting biologically relevant information. Therefore, it is reassuring to see some clear water between the waves: examples of experimental research that takes advantage of a comparative genomic approach one step at a time. The paper by Karniol and Vierstra in this issue (10) describing a new family of histidine kinases shows how the use of simple bioinformatics tools by microbiologists can (i) identify new targets for experimental work and (ii) provide important feedback for improving these tools.

HISTIDINE KINASES: MAJOR ENVIRONMENTAL SENSORS IN BACTERIA

Environmental sensing by histidine kinases is a fundamental property of a microbial cell. Extensive research on this topic during the last fifteen years has been recently summarized in two books (7, 8), a major review (15), and dozens of more specialized reviews. Histidine kinases appear to be the major class of environmental sensors in bacteria. A histidine kinase is a perfect sensor because of its modular architecture. The sensing capabilities lie within an input module (14), which contains one or more sensory domains that detect various physicochemical parameters. The input module communicates the information to a transmitter module within the same protein (14), which in turn sends the signal in the form of a phosphoryl group to another protein, a cognate response regulator. The activated response regulator triggers a cellular response, usually on the level of transcription. Figure 1 shows a domain representation of FixL from Bradyrhizobium japonicum (5), a typical histidine kinase where the input module contains two sensory PAS domains and the transmitter module contains dimerization (HisKA) and ATP-binding (HATPase_c) domains.

HISTIDINE KINASES: EASY TARGETS IN MICROBIAL GENOME ANNOTATION

Histidine kinases are easy targets for genome annotators because of the significant sequence conservation within the dimerization and ATP-binding domains. Profile hidden Markov models (HMMs) were designed for these domains, enabling their rapid detection and visualization in protein sequences in two primary domain databases, Pfam (2) and SMART (11). Recent implementation of Pfam and SMART domains into the conserved domain database (12) and InterPro (13) tools results in identification of histidine kinases in any routine similarity search of primary protein sequence databases, nr-NCBI and SWISS-PROT. Using SMART and Pfam domain models and conventional BLAST (1) searches, hundreds of histidine kinases have been identified in completely sequenced prokaryotic genomes (6). (As well as many other experimental and computational scientists) was under impression that if you have a protein sequence, you will know in a few seconds whether or not it is a histidine kinase.

THE HWE HISTIDINE KINASE FAMILY: A DIFFICULT CASE

The paper by Karniol and Vierstra (10) describes a new family of histidine kinases exemplified by a well-studied protein, the BphP2 light sensor histidine kinase from Agrobacterium tumefaciens (9). The family is named HWE after uniquely conserved histidine, tryptophan, and glutamate residues. The family consists of dozens of homologs, and many of them have no obvious characteristics of histidine kinases. Figure 1 shows that scanning a protein SMa2063 from Sinorhizobium meliloti, a member of the newly identified HWE family (10), against the SMART database results in a prediction of “protein of unknown function” because no known domain can be detected. SMART is a professionally curated domain database specialized in signal transduction; therefore, its opinion regarding this protein can be viewed as an expert one. For example, SMART easily detects all domains, not only those that are well conserved, such as HATPase_c, but also those that are poorly conserved, such as HisKA and PAS, in an experimentally characterized histidine kinase FixL (Fig. 1). Searching the Pfam database with the SMa2063 sequence results in ambiguous results, where the HATPase_c domain is detected at the borderline of statistical significance and is overlapped with another unrelated domain predicted with a similar unreliable statistical score. Both SMART and Pfam utilize the HHMMer
program for domain detection (3). Changing the program to reverse-position-specific BLAST (1), a domain search tool implemented in the conserved domain database (13), does not improve the situation: no conserved SMART or Pfam domains can be detected in SMa2063. Thus, it comes as no surprise that although 37 histidine kinases were annotated in the completely sequenced genome of S. meliloti, no conserved SMART or Pfam domains in its N-terminal half and HisKA and HATPase_c domains in the C terminus. Therefore, FixL is clearly predicted to be a sensor histidine kinase, even if its function would not have been known. In contrast, no known domains or motifs are predicted in the SMa2063 protein, except for a short low-complexity region (a purple rectangle) in the central portion of the protein. The study by Karniol and Vierstra (10) not only predicts that the protein of unknown function Sma2063 is a histidine kinase but also proves it experimentally.

Nevertheless, Karniol and Vierstra (10) were able to convincingly predict that SMa2063 is a histidine kinase. This has been done by detecting the SMa2063 protein in BLASTP searches (1) initiated with the ATP-binding domain of the BphP2 light sensor histidine kinase from A. tumefaciens (9) followed by a thorough analysis of its alignment with homologous domains. Special attention was paid to conserved motifs that have a functional role in the histidine kinase activity. The reader is referred to the paper for details of this analysis and experimental results demonstrating that SMa2063 and other proteins similar to BphP2 are indeed histidine kinases.

The importance of the results obtained by Karniol and Vierstra for microbial signal transduction is obvious. The new family of histidine kinases, which includes many previously unrecognizable members, is a big step forward. These sensor molecules initiate important regulatory cascades, and their discovery will facilitate experimental research aimed at understanding cellular properties that are controlled by two-component systems in a given microbial species. New findings also prompt new questions for experimental research. How significant is the deviation of the kinase module in terms of structure and function? Is there any distinct feature in cognate response regulators, etc.? The impact of this finding on bioinformatics is less obvious, but it is as important. It shows that current domain detecting tools need significant improvement when it comes to signal transduction proteins. The failure of SMART and Pfam to recognize a version of a conserved domain, such as HATPase_c, clearly calls for adjustments to current HMMs. New (for the HWE family) and improved (for the entire superfamily) models will ensure better automated detection of histidine kinases in all available and newly sequenced genomes.

**HOW TO STORE KNOWLEDGE IN THE SILICON AGE**

The main difference between biological research in the pre- and postgenomic eras is in the numbers. Traditionally, an exciting finding of a novel function for a given protein meant that it was for this protein only. Nowadays, this finding can be extrapolated to numerous homologs that have never been (and most of them will never be) in the hands of experimentalists and exist mainly in the virtual realm of databases. It is vitally important for the future of biology to make sure that such extrapolation is (i) applied and (ii) applied correctly. The second point is a subject of serious discussions and debates; however, even the first one is not very clear. Postgenomic biology is experiencing a normal disease of growth, where experimental and genomic information exist in parallel, rarely interacting worlds (Fig. 2). The scientific community has a traditional way of learning from reports published in peer-reviewed journals. Searching genomic databases with a BLAST program (1) became the second way of learning for many biologists (the BLAST paper has been cited >10,000 times, and many more publications refer to BLAST without citation). However, the real problem is that much (if not most) biological information in the databases is not only poorly peer reviewed but also annotated by people who cannot possibly be experts in all areas of biological research or are not biologists at all. Experimental scientists have little control over this process. For example, how will the important finding reported by Karniol and Vierstra (10) make it into the primary databases? One way is that curators at the National Center for Biotechnology Information, Swiss-Prot, Pfam, and SMART will find the paper and take time to connect each relevant record in the database to their publication (current automated tools usually fail in doing so) or convert printed alignments into models and make appropriate descriptions for domain databases. What if they miss the paper or are not willing to deal with the alignments from scratch (currently, most curators will ask for a ready-to-go alignment file)? Well, then SMa2063 and dozens of other proteins in these databases will remain hypothetical. There is a way out of this situation. For example, authors who wish to publish a new protein family can be asked to submit their alignments and descriptions that they feel are appropriate to primary domain databases if the paper is accepted for publication. This will ensure that the new finding, which has been peer reviewed, finds its place in the database and that its description will be provided by experimentalists themselves and not by database curators.

Another attempt to bridge experimental and genomic information is the development of specialized knowledge databases rather than sequence databases. At this time, most of such databases are organism oriented. For example, EcoCyc (http://ecocyc.org/) and CYORF (http://cyano-genome.ad.jp/) serve...
FIG. 2. Relationship between experimental and genomic information in the area of signal transduction. The current status of the information flow is shown by solid lines. The scientific community gets information by reading the results of experimental work and searching genome sequence databases for similarities. Future developments are shown by dashed lines. Computational biologists can retrieve relevant information from primary genomic databases and carry out more-specific annotation than that given in the primary database. This can be stored in a separate database (knowledge environment) accessible to all scientists interested in signal transduction. They can carry out online annotation of proteins they study, link them to the appropriate peer-reviewed publications, and communicate their suggestions on database improvements directly to the database developers. Thus, information will be presented in a form which suits the scientific community the best.

REFERENCES

The views expressed in this Commentary do not necessarily reflect the views of the journal or of ASM.