Jeffrey E. Barrick
12-12-01

Functional Gene Annotation: The Evolution of Description

Certain Ambiguity

Current systems for annotating gene function in biological databases are unable to capture the richness and complexity of the molecular interactions existing in a cell [1-3]. The function of a gene is an elusive concept because it can be correctly defined on many different biological levels (i.e. molecular: serine protease, cellular: vesicle formation, organism: inflammation response). Further, a single polypeptide may have multiple functions on the same level (binds DNA and binds lactose) or may only have a function as part of a macromolecular assembly (proteasome subunit). Neither a short descriptive phrase associated with a structure in a database, nor a clear-cut categorization scheme applied to all of the genes in a genome adequately conveys these subtleties. In 1998, Monica Riley suggested extending the description of gene function into "more dimensions" as a solution to these problems. She imagined metabolism, regulation, transport, and structure as examples of extra attributes necessary for complete annotation of a gene [1]. Modern systems for annotation based on dynamic linkage of database objects promise to provide these dimensions within a flexible framework for categorizing and annotating gene function.

A Metabolic Viewpoint

How do you completely and accurately describe a gene's function? One approach is to define gene function in terms of interactions with other cellular entities. On a molecular level, this view is inspired by the fantastically complicated Metabolic Pathways chart depicting a well-defined web of enzyme catalyzed reactions linking a host of metabolites [4]. Each connection between metabolites with an enzyme name above it represents a *function* of that polypeptide. As the companion Cellular and Molecular Processes wall chart implies, this formalism need not be restricted to purely chemical transformations. For example, an ion channel's function may be to "change" K^+ inside the cell to K^+ outside the cell. In fact, it is possible to be even more general and define the entities on a genetic level when their exact molecular nature is as yet unknown. A gene discovered in a genetic screen where only a cellular phenotype is known might be described as preventing the formation of actin filaments. As these examples illustrate, it is natural to define any cellular process at any level of resolution in terms of a web of interacting entities. One can imagine that by creating enough relational links and decomposing processes to a fine enough resolution this web could perfectly represent all known molecular interactions and functions in a cell.

Toward a More Realistic, Abstracted View

Toward a web with more links is precisely the direction that the major comprehensive databases of gene function are headed. The WIT (What Is There?) project for

functionally analyzing multiple genomes evolved from the more specialized MPW (Metabolic Pathways Database) [5]. One of their recent goals has been the "construction of more abstract scalable overviews, which should eventually cover all of cellular functionality" [6]. EcoCyc and MetaCyc collate metabolic pathways in *E. coli* and a collection of primarily microbial organisms respectively [7]. They have been extended with data on transcriptional regulation pathways and transporter proteins. The KEGG (Kyoto Encyclopedia of Genes and Genomes) database has a similar scope and goals as WIT [8]. It contains maps of several protein complexes in addition to metabolism and has recently begun incorporating an abstract protein-protein interaction network.

The shift from fixed descriptions of the relationships between biological entities toward flexible links and data abstraction parallels the development of modern programming languages. Although the implementation is apparently incomplete as of this writing, the published design of the aMAZE database is a nice example of complete conversion to this paradigm [9]. It incorporates a linked entity-relationship object model wherein biochemical interactions and entities are stored as separate objects. This elegantly solves all problems with genes having multiple functions on different levels: one need simply create multiple interaction objects referring to the gene. If you need to later expand an interaction into more detail after learning more about it (for example, if an mRNA is found to have multiple splice variants) one simply creates new entities and interactions and explains the new interactions in this area.

Extra layers of information are added by linking these objects to descriptors, like the cellular location where an interaction takes place. These descriptors themselves are linked together to form a hierarchy of categorizations that can be easily updated. This situation is identical to the organization of the extensible controlled vocabulary for biological terms created by the Gene Ontology (GO) Consortium [10]. Descriptors can also provide links to other important forms of information: references to the literature, what algorithm was used to predict gene function during automatic annotation, or the biophysical parameters of an interaction. Fundamentally, the innovation of using links between data objects allows a facile description of gene function and interactions that can always adapt to incorporate new forms of information as they become available.

Dreams and Nightmares of Complexity

Although huge linked webs of objects are an elegant way of representing gene function, they are difficult to visualize and exceptionally tricky to debug if something goes astray. New visualization tools and algorithms for gleaning data from such webs will definitely be necessary. Fortunately, since each individual interaction in the web is straightforward, one might imagine that very simple "digital groundskeeper" algorithms could easily wander through the web annotating it or reporting parts that have become incompatible due to new data. Once the interaction maps become detailed enough, all of the database groups hope to use them to simulate cells on the computer. This could be extremely useful for testing drugs and examining how cells integrate information from their environment. However, it will require either radically new *in vivo* high-throughput technologies or extremely accurate structure/function predictions to adequately obtain the

huge number of biophysical constants required to describe the interactions of all of the components in a cell. Even then, because of small uncertainties in our measurements we can expect the complex model's behavior to rapidly diverge from reality – just like predicting the weather.

1. Riley M: **Systems for categorizing functions of gene products**. *Current Opinion in Structural Biology* 1998, **8**:388-392.
2. Gerstein M, Jansen R: **The current excitement in bioinformatics - analysis of whole- genome expression data: how does it relate to protein structure and function?** *Current Opinion in Structural Biology* 2000, **10**:574-584.
3. Wilson CA, Kreychman J, Gerstein M: **Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores**. *Journal of Molecular Biology* 2000, **297**:233-249.
4. Biochemical Pathways Wall Chart on World Wide Web URL: http://www.expasy.ch/cgi-bin/search-biochem-index
5. Selkov E, Jr., Grechkin Y, Mikhailova N, Selkov E: **MPW: the Metabolic Pathways Database**. *Nucleic Acids Research* 1998, **26**:43-45.
6. Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov E, Kyrpides N, Fonstein M, Maltsev N: **WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction**. *Nucleic Acids Research* 2000, **28**:123-125.
7. Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, Pellegrini-Toole A: **The EcoCyc and MetaCyc databases**. *Nucleic Acids Research* 2000, **28**:56-59.
8. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Research* 2000, **28**:27-30.
9. van Helden J, Naim A, Mancuso R, Eldridge M, Wernisch L, Gilbert D, Wodak SJ: **Representing and analysing molecular and cellular function using the computer**. *Biological Chemistry* 2000, **381**:921-935.
10. Ashbuner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al.: **Gene Ontology: tool for the unification of biology**. *Nature Genetics* 2000, **25**:25-29.