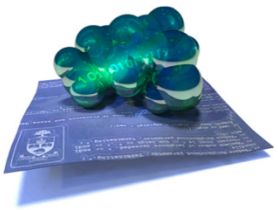
A
Bioinformatics
Course

Function Concepts



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto

Function

"Function" is often equated to the activity of a biological molecule.

However, a particular activity may not be relevant for the biological system under study ("test-tube function? vs. biological function).

The same activity in one cell may serve a significantly different purpose in another cell (function in context vs. isolated function).

Proteins may have more than one activity ("moonlighting" proteins.

Striking examples of "moonlighting" proteins include **epsilon-crystallin/lactate dehydrogenase B4** in ducks, which is the same protein, and the small electron shuttle of the mitochondrial inner membrane, **cytochrome c**, which is also a potent apoptosis signal.

Speaking of Function

Experimentally *observable properties* of a protein like sequence or structure are straightforward to abstract, store, retrieve and interpret.

Aspects of a *protein's behaviour* like conservation, localization, interactions, regulation of expression etc. require more context – but don't pose problems of a principal nature.

However "Function" is a *concept.* It is not an observable. The concept aims to integrate a a large variety of observable aspects of a gene, and its role in its molecular and cellular context. There is no "natural" set of categories and values that presents itself to reason about function.

Thus that aspect of biology that is most relevant to our ultimate goal of understanding is elusive: we can't even properly *talk* about function. A crucially important task of function analysis therefore is knowledge engineering: to define a "language" in which we may categorize, collect and compare functions.

In general, this is not sufficiently appreciated.

"Function" may be the most elusive concept we deal with in bioinformatics, which is surprising since we use the word in nearly everything we write about biology. But the problem is: in biology there really is no such thing as *function*, only activities, behaviour – and *function* is a concept that we apply to summarize and categorize observed behaviour.

What is function?
What do we mean by "function" of a biomolecule?

How do we represent function?

How do we annotate function?

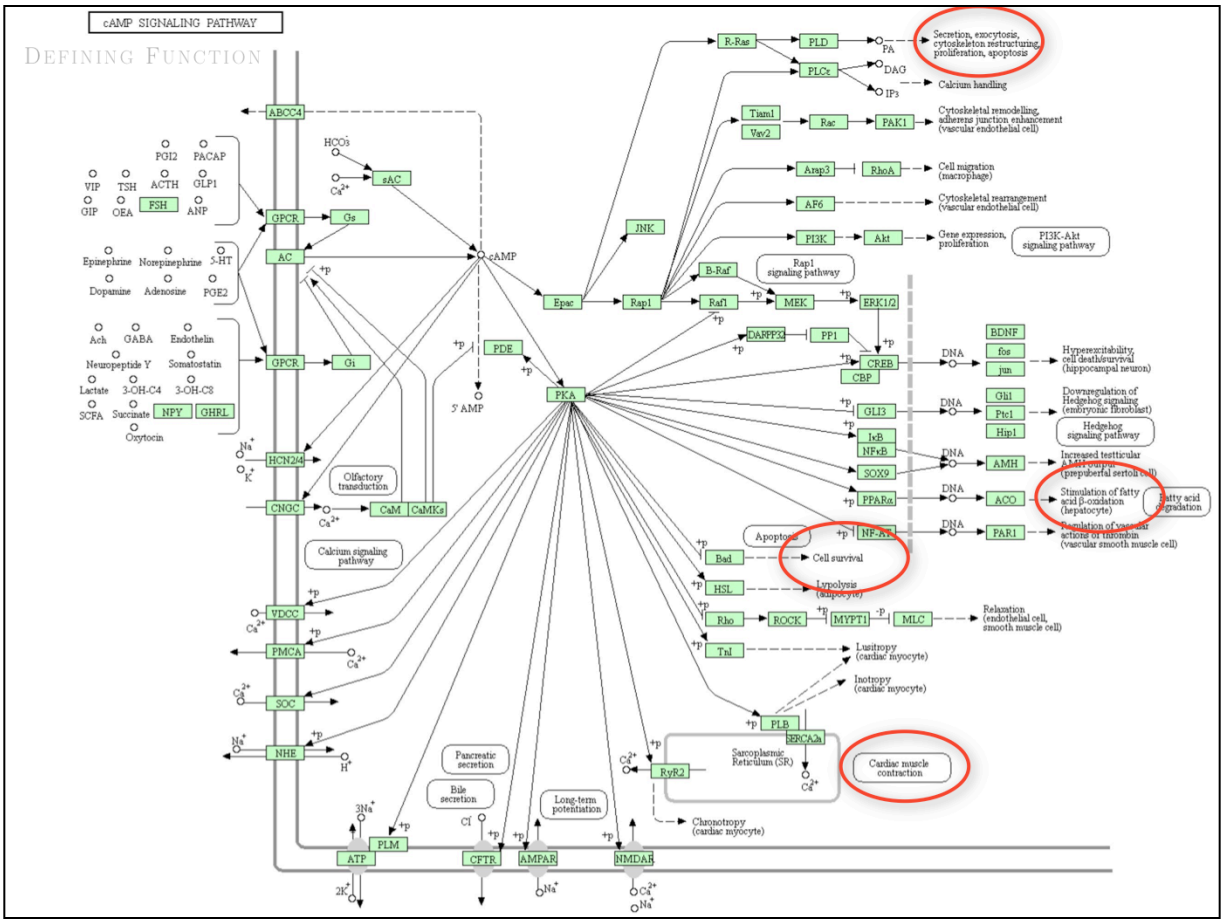
How do we predict function?

Let us focus on epistemology first by enumerating some concepts that we usually associate with the word *function*;

Then we need to talk about ways to define abstractions of functions, and use them in bioinformatics;

Next we need to discuss how known functions are associated with particular biomolecules;

Finally we address strategies of predicting function for unannotated molecules.



Function can be defined informally at many levels including experimental observation, general roles, generic modulation and/or global objectives.

In this KEGG pathway diagram of the human cAMP signalling pathway, aspects of *function* are given as text. These include:

- high-level obejctives, such as *cell survival*;
- generic modulation, such as *hyperexcitability*;
- biological processes such as *secretion, exocytosis, cytoskeleton restructuring, proliferation, apoptosis*;
- specific roles such as *stimulation of fatty acid β-oxidation*.

It is not guaranteed that these terms and concepts are constrained between pathway representations.

Evolution works on sequence, but selects function.

Adaptive sequence change is often localized to individual domains, clusters of mutually interacting amino acids. Domains can be defined from sequence patterns, or from the analysis of 3-D structure. They often fold independently and can thus be duplicated and rearranged independently as modules of function.

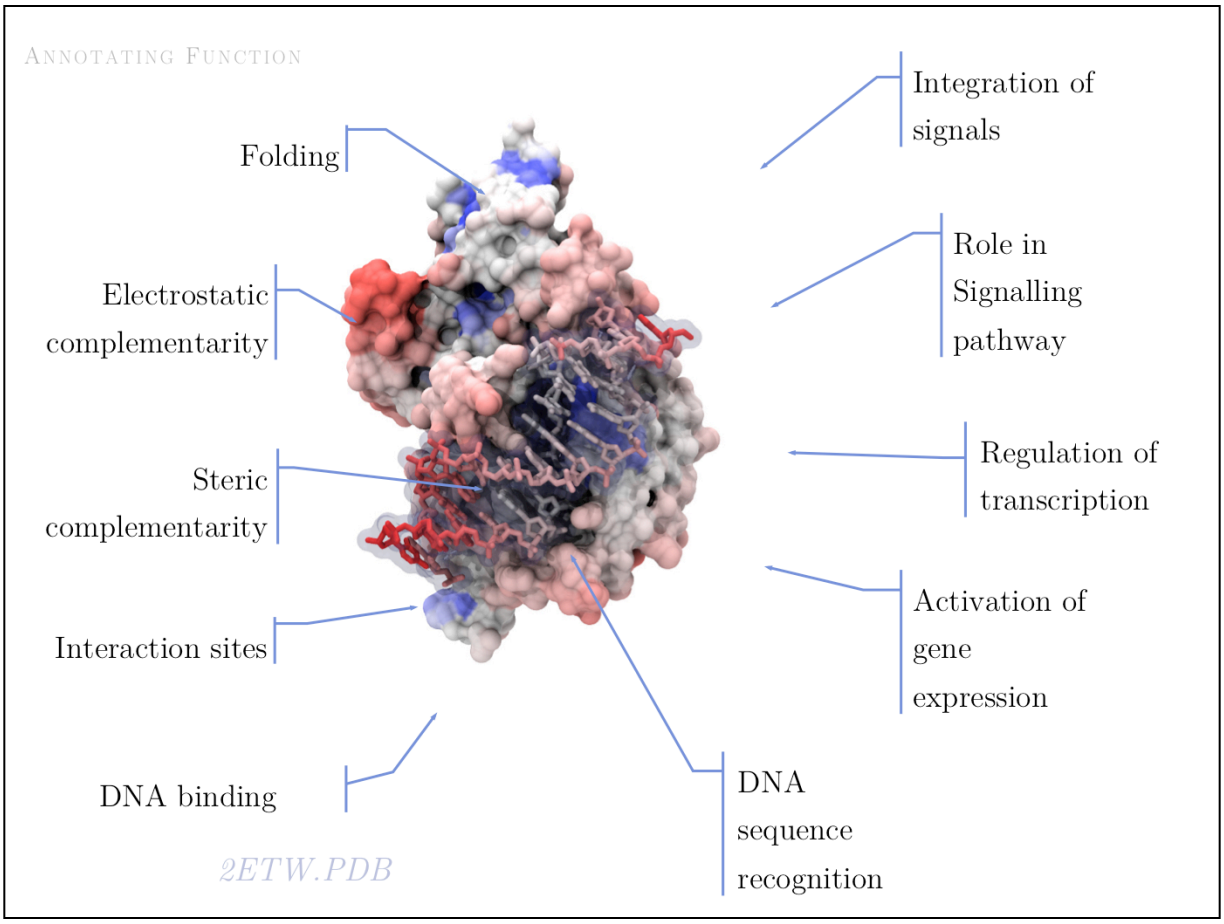
One corollary is that sequence searches and multiple sequence alignments etc. that are done with individual domains are much more specific than when the entire sequence of a protein is used.

Once structural domains have been defined, sequence profiles, HMMs, or other computational procedures can be used to discover more members of the domain family from the database.

Perhaps the most rigorous statement that can be made about *function* is that it is intimately associated with evolution. Even though we know that most evolutionary change actually corresponds to **neutral** drift, in those cases where we identify adaptive evolution, by oserving that sequences are under prurifying selection in response to selective pressure, we can associate *function* with an observable!

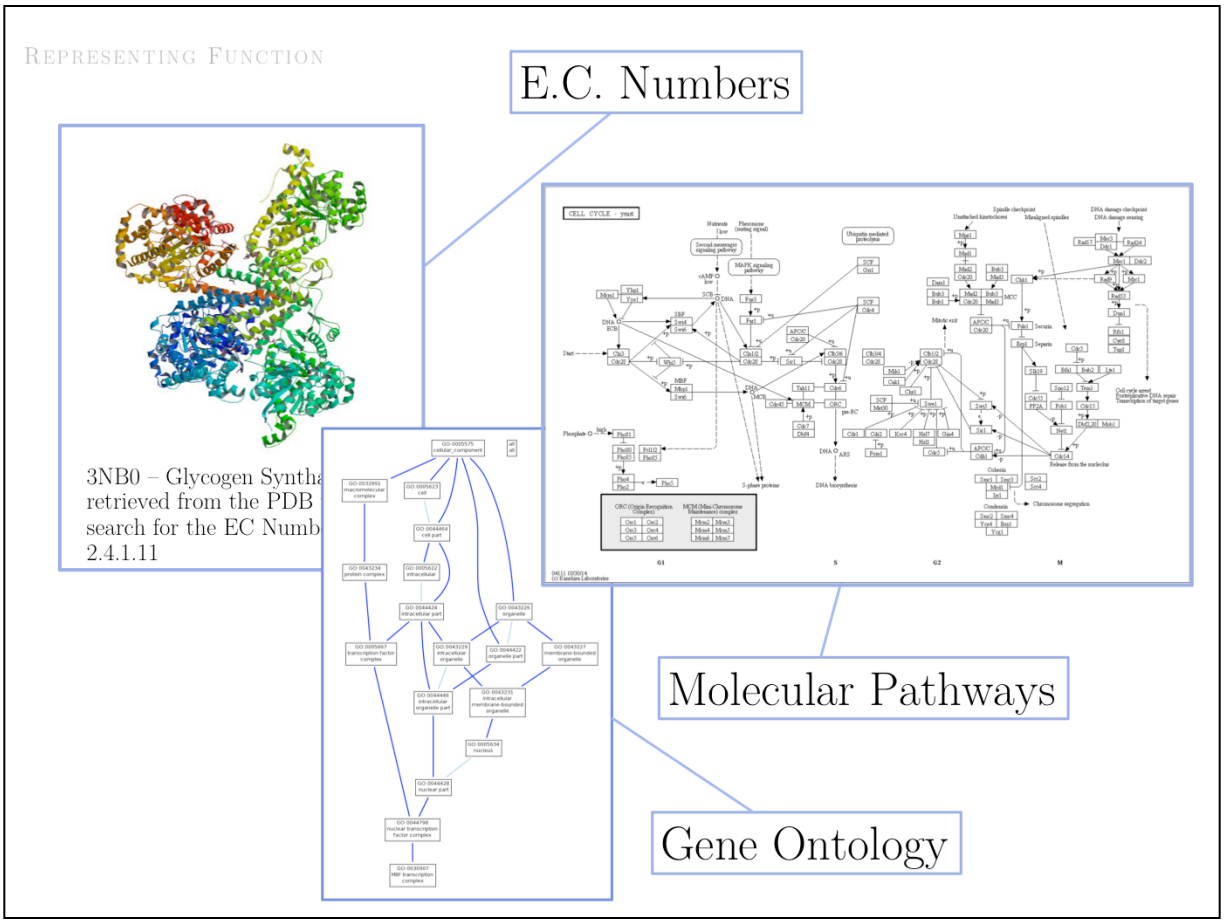
We can define: the function of a molecule is that aspect of its behaviour that is under purifying selection. At least in principle this provides perspectives on experimental design.

This provides a ready explanation for observations that evolutionary change segrgates with folding units i.e. structural domains.



Just as function can be defined on different levels, the anotation of specific biomolecules with function information will depend on the database context as much as on the molecule itself.

The "function" of individual amino acids in a sequence can be defined with respect to structure formation and the 3D-conformation of an active site. Regions may contribute to such active sites, as well as to allosteric regulation or other interaction sites. Sequence-specific DNA binding is an example of DNA binding. Its outcome may be the activation of gene expression (in a transcription factor), which is an example of transcriptional regulation, which is a role of elements in a signalling pathway. Such relationships like "example of", "takes part in" ... impart a hierarchical structure on our collection of function terms. The signalling pathway may have many outcomes which are useful to the cell, the role of our protein may be described specifically with reference to the mechanics of the regulation process, or more generally as abstract information processing, such as "integration of environmental signals".



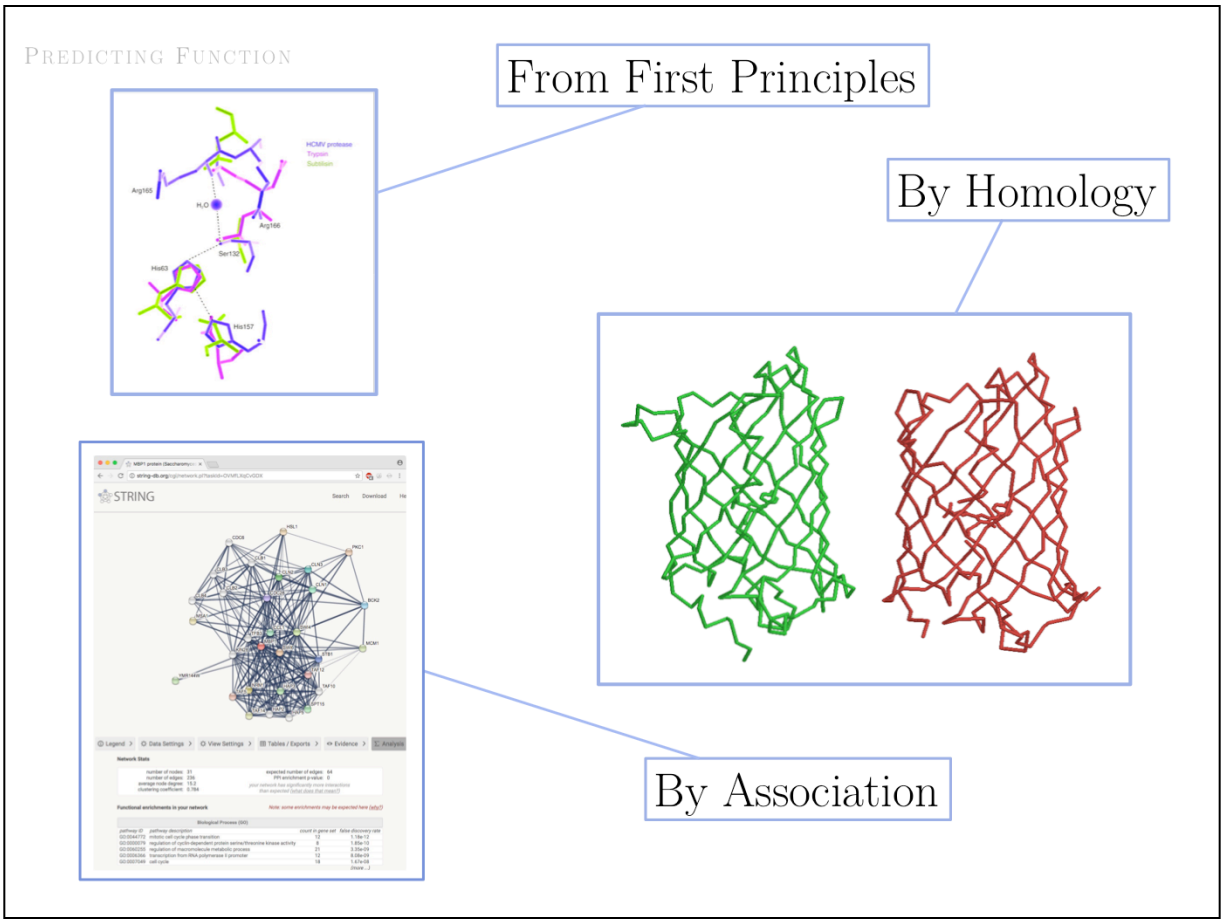
The three most common representations of function are

E.C. numbers, a systematic enumeration of catalytic activity accordint to the nature of the catalyzed reaction;

Pathways, which collect collaborating sets of biomolecules and thier relationships; and

the very extensive GO (Gene Ontology) project and its associated OBO (Open Biology Ontologies) – a large project of knowledge engineering that aims to assemble all **biological processes, molecular functions** and **cellular components** int a unified *ontology* of terms, i.e. associating them hierarchially through **`is_a`** and **`has_a`** relationships.

This means, you can expect such annotations in many databases and use them to establish relationships among database entries.



Prediction of protein function means annotating proteins for no function is known. Three general principles are applied:

Prediction from first principles is sometimes successful if, e.g. the 3D-structure contains structural patterns of residues that comprise an enzyme active site. The textbook example is the protease *catalytic triad*.

Prediction by homology is the most common prediction method: homologous proteins (i.e. proteins that have diverged from a common ancestor) have similar molecular function. That means, to learn about the function of an unnown molecule, list the functions of homologous proteins. This is also called *annotation transfer*. In this case, we need to record and distinguish *inferred* and *observed* function.

Prediction by association applies annotations to members of sets of proteins that are all associated with each other, for example by being part of the same complex, having co-regulated expression, being located in the same genomic operon, etc. Often association properties are used to construct networks, which are then analyzed.

http://steipe.biochemistry.utoronto.ca/abc

B O R I S . S T E I P E @ U T O R O N T O . C A

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA