What is functional annotation and why would we do it?

Functional annotation is the process of associating sequence data with sequence ontology useful to the question at hand. Sequences of nucleotides or amino acids are only abstract representations of the molecular composition of a polymer. The information contained in these sequences is sufficient to compare the polymer sequences to other molecules in a database, to extract statistics pertaining to chemical composition of the polymer, calculate lengths, masses and so on. So from sequence information we can get more information by doing calculations and comparisons. However, biology is not a science of converting one form of information into another. The goal in obtaining these sequences is to arrive at an understanding of the biological processes they are associated with. In other words, we need to convert this sequence information into biological knowledge.

But what kind of knowledge do we seek? This depends on the particular biological problem at hand. Nevertheless, in general we imagine that each piece of the genome, be it a protein or RNA coding sequence, a promoter region, is involved in some kind of reaction, which does some work for the cell. So, at the very least, we would like to associate names with the putative reactions (or more general functions) that these regions are associated with. This process of naming already
brings us one step closer to the knowledge we might be interested in. Associated with this name are the details of the function at hand. These details can span a number of different levels of abstraction. At the lowest layer, one might be interested in protein structures and binding and or catalytic activity that these proteins might have. One level above this, one might be interested in other molecules (protein, nucleic acids, etc.) that a protein of interest binds to, since many biological processes are mediated by complexes of molecules rather than individual molecules. At an even higher level of abstraction, one can consider annotating protein sequences according to the cellular locations one expects to find the protein products, the structural or metabolic networks inside the cell which they participate, or even the roles that they play outside of the cell in interaction with the cell’s external environment and associated ecology. All of these kinds of data endow the sequence with additional meaning that informs the process of interpreting the sequence in terms of biological reality.

**sequence information**

CCAAAAGATATCCTTCACAAACCAAGTAATATATTCATCTCATACTCTCTCTAGAATTGATGAA
AAACCTTCTCCGAGATAATGCACACTAAAGGAATGCTGGGGGATTGATTCTCTTACCAAAAGCAGAT
CCGCTATAGATCCAATTAATCATCAGTTAAATTGATGAAATGAAGGCAAATTACTACGGTTGGATTCTAC
GGTTAAGCTATATTAAATCTCACTACATCAATAAAAATAAACAGTTATCTGGTATAGTAAACTGAATAC
TAGTTGATTGGATCCATTCAACTGAAATCATACGAATTCTGGAAAATTTGGGTGAGTCAATGATG

↓

**biological form, function, context, etc.**

![Diagram showing aspartate + citrulline leading to arginine-succinate and then to arginine, with labels for ArgG and ArgH enzymes.]

**Figure 1.** The essence of functional annotation is the interpretation of information contained in sequence data in terms of biological relevant ontological associations.

Having introduced the basic general goal of functional annotation, we can begin to present the specific problem relevant to our project, which is the characterization of emergent pathogenic strains in the genus *Haemophilus*. The *Haemophilus* genus includes a number of pathogenic species, the most famous of which includes various *H. influenzae* strains capable of causing pneumonia and meningitis in children (it is a human pathogen; there are no known animal reservoirs). However, an initial series of biotyping experiments has indicated that the clinical strains we have obtained in infection belong to the species *H. haemolyticus*. This is interesting and of practical importance since *H. haemolyticus* has traditionally not been associated with disease. When new pathogens emerge in
this way there is much interest in the development of antibodies unique to the disease causing strains both for use in typing protocols in a clinical context and as part of the process of developing vaccines. To this end, functional annotation of genome sequences can rapidly pave the way towards progress in this direction. For example, typing data from the CDC indicates differences between strains manifested in assays for outer membrane proteins P2 and Haemophilus D. Since these proteins are presented on the surface of the cell, one can try to find antibodies for typing and vaccine development. By obtaining data on the sequence coding these proteins or the sequences associated with the machinery directing their export to the outer membrane we gain a high resolution picture of the genetic basis of any experimental completed so far or planned in the future. We can compare genes and gene circuits not only for these particular proteins, but also all proteins we think could be exported to the outer membrane (more on how this is done later).

Figure 2. The discovery and characterization of antibodies to discriminate between strains and develop vaccines is of particular interest. Experimental data from the CDC contains measurement for the presence of Haemophilus protein D and outer membrane protein P2, but it will also be possible to identify candidate antigens by cataloging proteins with favorable binding characteristics targeted for export through the periplasm to the cell surface.

There are other problems that functional annotation of sequences can address beyond the explanation of biotyping data and informing the process of vaccine development. The name *H. haemolyticus* is derived from the hemolytic activity (e.g. ability to lyse blood cells) that these strains exhibit, distinguishing them from nonhemolytic *H. influenzae* strains. However, not all *H. haemolyticus* strains in our data set display hemolytic activity. The reason for this is unknown, as are the molecular mechanisms (e.g. coding sequence of a hemolysin) and their regulation. Functional annotation of sequences in these genomes may leads not only to the identification of the molecules involved in this process but may also inform an analysis of any possible relationship between hemolytic activity and virulence.

Functional annotation of sequences extends beyond the naming of genes, there regulatory elements and coding products. Sometimes pieces of the small genome deserve annotation because of their own special physical or chemical
properties. For example, inverted repeats or other kinds of sequences will influence the physical structure of DNA which can have dramatic influence on biological function. In our system the eight basepair sequence ‘AAGTGC GG’ serves as a signature for competence machinery involved in uptake of DNA and transformation. As a result our genomes are expected to be littered with overrepresentations of this sequence, with the particular patterns of the sequence distribution within and between genomes serving as a potential marker of modes of gene acquisition.

![Figure 3](image)

*Figure 3. The genome *H. influenzae Rd KW20* is enriched with the 8bp sequence as a result of insertion of extracellular DNA after uptake by competence machinery. The dots are guides to the eye to mark the scale. Adapted from Smith et al.1.*

Blah, blah, blah... another example? Or shall we get to the computational aspects already?

**How do we do functional annotation?**

Homology based approaches
*Intelligent thoughts here.*

*Ab initio* approaches
*Intelligent thoughts here.*

Combining *ab initio* and homology based approaches
*Intelligent thoughts here.*
Typing data available for our strains

Several typing assays were performed on our samples from the CDC. Explaining the genetic basis for the results of these tests is one our key challenges. As such it is important that one understands the biological significance of each of these tests to illuminate the potential genetic information relevant to these results.

Outer membrane binding protein P2
We have data for probing the presence or absence of outer membrane binding protein P2. Outer membrane proteins are exported by Gram positive bacteria to the outer membrane to perform a variety of functions like secretion or uptake. A hypothetical protein structure of the *H. influenzae* P2 protein is available through Uniprot [http://www.uniprot.org/uniprot/P43839](http://www.uniprot.org/uniprot/P43839). Since P2 is bound integrally to the outer membrane it is useful for typing studies. There are experiments that demonstrate and adapted resistance of P2 to antibodies in *H. influenzae*.

Hemolytic activity
*H. influenzae* has traditional been distinguished from *H. haemolyticus* by its lack of hemolytic activity. However, in our samples and elsewhere one finds *H. haemolyticus* strains lacking hemolytic activity. To this effect their is interest in developing further techniques to resolve the two. An effort to characterize the relationships between *H. influenzae* commensal hemolytic *H. haemolyticus* and commensal nonhemolytic *H. haemolyticus* has been reported in. Both hemolytic and nonhemolytic *H. haemolyticus* strains can exhibit pathogenic behavior (from our CDC data). Nevertheless, since the method of hemin utilization may correlate with virulence and since hemolysis presumably increases hemin availability, the absence of hemolytic activity might be a factor contributing to virulence. Further circumstantial evidence for this hypothesis comes from the observation of the network of distributed heme-aquiring genes in non typable *H. influenzae* facilitated by lateral gene transfer, which is thought to correlate with virulence. For a clinical study of hemolytic microflora and where they can be found in the human host see Branson. There is a lot of variation in hemolytic activity in H species and the *H. haemolyticus* hemolysins are still unknown. Can we find them? The answer to this question is known for related strains.

*Haemophilus* protein D
There is a Uniprot entry available [http://www.uniprot.org/uniprot/P71351](http://www.uniprot.org/uniprot/P71351). It is a surface lipoprotein (see various papers).
Fuculose-kinase
There is a UniProt entry available http://www.uniprot.org/uniprot/P44399. And structural information (predicted) http://modbase.compbio.ucsf.edu/modbase-cgi/model_details.cgi?queryfile=1297632150_834&searchmode=default&displaymode=moddetail&referer=yes&snpflag=& .

Hydrogen sulfide production
Hydrogen sulfide production is a key indicator of the activation of certain metabolic pathways (it is an electron acceptor).

ND Identification strips
These are basically kits for all kinds of analytical chemistry for metabolism based typing. Can anyone interpret the CDC data here?
