Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Microbiology

Using comparative genomics to drive new discoveries in microbiology

Daniel H Haft

Bioinformatics looks to many microbiologists like a service industry. In this view, annotation starts with what is known from experiments in the lab, makes reasonable inferences of which genes match other genes in function, builds databases to make all that we know accessible, but creates nothing truly new. Experiments lead, then biocuration and computational biology follow. But the astounding success of genome sequencing is changing the annotation paradigm. Every genome sequenced is an intercepted coded message from the microbial world, and as all cryptographers know, it is easier to decode a thousand messages than a single message. Some biology is best discovered not by phenomenology, but by decoding genome content, forming hypotheses, and doing the first few rounds of validation computationally. Through such reasoning, a role and function may be assigned to a protein with no sequence similarity to any protein yet studied. Experimentation can follow after the discovery to cement and to extend the findings. Unfortunately, this approach remains so unfamiliar to most bench scientists that lab work and comparative genomics typically segregate to different teams working on unconnected projects. This review will discuss several themes in comparative genomics as a discovery method, including highly derived data, use of patterns of design to reason by analogy, and *in silico* testing of computationally generated hypotheses.

Addresses
J Craig Venter Institute, Rockville, MD, USA

Corresponding author: Haft, Daniel H (danielhhaft@gmail.com)

Current Opinion in Microbiology 2015, **23**:189–196

This review comes from a themed issue on **Genomics**

Edited by **Neil Hall** and **Jay Hinton**

http://dx.doi.org/10.1016/j.mib.2014.11.017

Introduction

In the classic problem in genome annotation, sequence is linked to function directly by experiment for just one or two model proteins. In each newly sequenced genome, it must be decided whether the closest homolog to such an exemplar performs the same function, or does something different. What method should be used to decide whether the new sequence should receive the same functional annotation? In some homology families, all members out to the limits of detection perform the same function. In others, function diverges rapidly once identity falls below 60% [1]. Any blanket rule that relies on fixed criteria for propagating functional annotation from the proven to the unproven is bound to perform badly. Each protein family is different, necessitating different cutoffs. The most similar sequences by BLAST do not always match those closest by recent common ancestry [2]. For most families, in fact, no BLAST score cutoff could separate the functionally equivalent homologs from all other proteins; the two sets interleave. Consequently, missed annotation and misannotation both run rampant in public databases, with overly specific annotation an especially troublesome symptom [3]. Approaches that make one bold computational leap per annotation simply cannot perform well.

The best approach to high quality annotation is an incremental process that advances through large numbers of very modest assumptions. Continual testing that newly assigned annotations in a protein family remain consistent with each protein's species of origin, inferred metabolic background, and neighborhood of adjacent or nearby genes keeps confidence in the annotation process high. One or two characterizations of a histidine biosynthesis enzyme, for example, may suffice to show a typical histidine operon structure, bring up many more sequences from similar contexts, generate multiple sequence alignments and phylogenetic trees, and lead eventually to an almost perfect classifier with near zero false positives and false negatives over all genomes sequenced to date. The resulting entry in the protein family definition database, with its hidden Markov model (HMM) [4] based on a curated seed alignment, together with its cutoff score, and its set of annotations to transfer, becomes a fully automated tool that emulates what the expert biocurator would do, in theory, if asked to annotate the same target gene.

Reasoning through large numbers of small assumptions may seem unreasonable to the protein chemist trained to expect a progressive loss of yield with every additional step. A 500-step protein purification probably would yield very little. But biocuration resembles fitting together a 500-piece jigsaw puzzle rather than purifying a protein. It is true that placing each new piece requires one more new hypothesis, but the fact that the piece fits at all gives strong validation that clears lingering doubt from earlier stages. Once the puzzle is completed, a picture emerges whose obvious self-consistency gives robust confirmation that most or all pieces were placed correctly.

Derived data

The currency for reasoning by comparative genomics to infer molecular functions and biological processes consists mostly of *highly derived data*, often very far removed from the lists of which specific protein sequences have had which functions proven in the lab [5]. We use HMM-based classifications of proteins into families to tell what enzymes are present in a microbe, then combinations of these assignments to assert that an enzymatic pathway or other subsystem is complete, or else completely absent, for any given genome. These assertions are used in turn to generate a list of 1s and 0s, called a phylogenetic profile, to show which species have a given marker, or a whole subsystem, and which do not. Examples of additional highly derived data types include: the list of a genomes with the same apparent hole in an enzymatic pathway, predictions by metabolic modeling that a list of genes all are essential, the list of all species that carry one marker but not another, phylogenetic trees calculated from multiple sequence alignments, inferred gene duplication and gene loss events, domain structures of proteins, predicted signal peptides and transmembrane helices, conserved gene neighborhoods, conserved gene order, or matters as simple as finding where members of two selected protein families are encoded by adjacent genes. Each of these types of observation, far removed from typical laboratory experimental measures of protein activity, can lead annotators to clearer pictures of protein function.

High-dimensional data

Collections of complete genomes contain intrinsically *high-dimensional data*. Numerous data types each reflect on some aspect of a protein family's biology that other methods cannot assess. Mutation rates inferred from a molecular phylogenetic tree, which regions of sequence are best conserved and where the few invariant residues map on the most closely related crystal structure, the frequencies of gene loss, gene duplication, and lateral gene transfer, the conservation of gene neighborhoods and gene order within those neighborhoods, the lists of cofactors synthesized in species with a member of that protein family, the functions of most closely related sequences known to differ in function, which additional markers occur in the same genomes as the family in question and which markers never do, where microbes with the family live, and many other traits carry information that might support a theory of what some protein's role and function might be, or might refute it.

The high-dimensionality of comparative genomics data means bioinformatics can deliver a range of metrics that have a high degree of statistical independence. A detailed hypothesis about the workings of a putative new system, based on analogy to some known system, might lead to a number predictions that should all be jointly true. If the first suggestive finding is a mere statistical anomaly, and not true evidence for the biology proposed, the various other metrics will not lend support. The hypothesis can be dropped. But if multiple statistical measures of independent facets of a proposed biological system are confirmed, the hypothesized new system may become strongly supported well before the first new 'wet lab' experiment is performed.

Bioinformatics journeys

We suggest the term '*bioinformatics journey*' to describe a code-breaking exercise in comparative genomics that starts with some (possibly weak) hypothesis, and by progressively filling in the biological picture, manages to deliver a richly detailed scenario that merits strong confidence for many of its predictions. For example, two proteins are weakly similar, and might be proposed to belong to some still-undefined homology family. Once a sufficient set of true homologs has been collected and shown in a multiple sequence alignment, the hypothesis of homology (descent from a common ancestor) becomes iron-clad. If any members of the seed alignment are removed, an HMM based on the remainder could easily recover them — a powerful form of cross-validation. PSI-BLAST [6] makes this kind of journey almost routine. Meanwhile, the outcome of a bioinformatics journey such as a definition of a new homology domain can provide new information. For example, the multiple sequence alignment may reveal motifs of strong local sequence similarity, an emergent property not apparent in individual pairwise alignments, to show which types of residues in a protein family are most conserved and thus give clues to what the general molecular function might be.

If the nature of a biological question is favorable for computational analysis, then a hypothesis made *in silico* will have important implications that can be tested without recourse to new experiments. Sometimes this means applying purely computational tests. Sometimes this means validation in the rear view mirror, locating an old published report whose results suddenly merit a new interpretation. The C-terminal region of the S-layer glycoprotein of halophilic archaea is a homology domain, the PGF-CTERM domain, and it invariably co-occurs in genomes with archaeosortase A, which was proposed to be an enzyme that cleaves and removes such regions [7]. The PGF-CTERM has a highly hydrophobic transmembrane alpha-helix, sufficient to anchor a protein to the membrane. Why anchor a protein at its C-terminus only to cleave that anchor? A much earlier finding was that a prenyl-derived lipid moiety, large enough to serve as a membrane anchor, was added to this S-layer glycoprotein, somewhere near its C-terminus, but that finding too was odd — why give a surface protein a second, redundant C-terminal membrane anchor [8]? In light of the discovery of archaeosortase, attaching a large lipid group suddenly makes sense, not as redundant anchor to the membrane, but as a replacement. Archaeosortase could