Published in final edited form as:

Curr Opin Microbiol. 2015 February ; 0: 189–196. doi:10.1016/j.mib.2014.11.017.

Using Comparative Genomics to Drive New Discoveries in Microbiology

Daniel H. Haft

J Craig Venter Institute, Rockville, Maryland USA

Daniel H. Haft: danielhhaft@gmail.com

Abstract

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.

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

© 2014 Elsevier Ltd. All rights reserved.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

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's 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 1's and 0's, 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 be a transpeptidase, substituting a large lipid anchor for the C-terminal alpha helix, and not simply a protease.

As with jigsaw puzzle assembly, successive stages of the *in silico* analysis of a subsystem can add new information as they validate earlier suppositions. By the end of one of these bioinformatics journeys, the description of a new protein sorting system [9], an unexpected cofactor dependency [10], or a new natural product [11] can become detailed, specific, and quite trustworthy in advance of the first new experiment.

The subsystem approach

A key to reasoning methods in comparative genomics is seeing genes in context: physically (gene neighborhoods), taxonomically, and metabolically. The one-dimensional “bag-of-genes” view of genome-encoded biology must be abandoned. Clues to biological process and molecular function come from comparing each protein to all its characterized homologs, of course, but also from the source species’ lineage, niche, and phenotype, from all that can be reconstructed of its metabolism, and from the arrangement of nearby genes along the chromosome. The single most useful improvement over the bag-of-genes view is the subsystem approach [12, 13]. If several genes collaborate to run some biological process, they form a subsystem. The *molecular function* ontology, from the Gene Ontology (GO) project [14], gives a low-level view of genome content, a parts list of individual gene products: thirty kinases, six GTPases, seventy regulators, etc. An entirely different GO ontology, *biological process*, provides higher level view based on what gets done: cobalamin biosynthesis, histidine degradation, endospore formation, etc, to given a much more intuitive view of genome content. Analyzing genomes at this level, subsystem by subsystem, gives feedback that improves the reach of and confidence in the collections of protein families that assign molecular function. RAST subsystems [15] or Genome Properties [16] both use the process of “populating” subsystems, enumerating the genes in a given genome that fill the expected set of roles, to refine the collections of protein family definitions, FIGfams [17] and TIGRFAMs [16] respectively, on which their computation is based. The resulting improved protein families help separate known proteins from the

unknown, making both the high-level (subsystems) and low-level (list of genes) views of genome content more accurate and informative.

Bioinformatics methods that begin only after proteins have been assigned to protein families are collectively called “**post-homology methods**.” They include phylogenetic profiling [18], analysis of conserved gene neighborhoods and conserved gene order [19], and “Rosetta stone” gene fusion analysis [20], among others. Data mining using these methods can be a powerful means to deduce that two different proteins work together, and eventually to parlay information from one molecular marker definition (i.e. one protein family) into a broader understanding of a group of protein families and how they might work together [21]. The STRING database server [22] lets even a novice user start with a single protein from any of a large number of reference genomes and, within seconds, see graphical representations of proteins grouping into highly connected networks as computed from evidence from a host of methods: joint presence or joint absence in the same species (i.e. phylogenetic profiling), conserved gene neighborhood, domain fusions, respective homologs that are coordinately regulated or co-precipitated, etc.

Comparative Genomic Reasoning Methods for New Discovery

The post-homology methods, collectively, make it possible to find evidence that one protein family works with another, or that a whole cohort of protein families work together. The next task is to understand what the group of proteins does. Even if each protein is understood only poorly, based on distant homologies to characterized proteins that differ in their specific functions, looking at the whole set together will suggest the nature of the biological process, and therefore help improve the prediction of specific functions.

Non-orthologous Gene Displacement (NOD)

Most subsystems should be complete, or completely absent, over the majority of species. It usually makes no sense for a cell to make nearly all enzymes required by a pathway but omit one for a step in the middle, leaving a hole. If a large number of species all have variants of a pathway with the same hole, a reasonable assumption is that the pathway is complete after all. The seemingly missing activity likely is present, but is not recognized because it is provided by an enzyme from a different protein family. A non-orthologous gene displacement (NOD) event must have occurred [23]. The enzyme histidinol-phosphatase phosphohydrolase (EC 3.1.3.15) performs the penultimate step in histidine biosynthesis. But in different lineages, enzymes from at least three completely unrelated superfamilies catalyze this step. Presumably, a new phosphohydrolase can arrive in a genome by lateral gene transfer and be retained, perhaps for the sake of a secondary function it performs. The original enzyme, now redundant, may be lost, permanently displaced. Using recurring pathway holes to find new protein families that fill the hole, and using hypothesizing NOD to infer a specific molecular function for those families, is now a standard informatics technique.

Patterns of Design

All bacteria live under similar sets of biological constraints. To survive, thrive, and overgrow their competitors, so they can exist to be studied at all, they must obtain nutrients, synthesize proteins, reproduce their DNA, defend against various threats, and grow and divide. Nutrients required inside the cell somehow must enter across a selectively permeable plasma membrane. All twenty common amino acids are required, each of which must be synthesized if it cannot be imported. Proteins are synthesized from the N-terminus to the C-terminus. The interior of a membrane is hydrophobic, and made selectively permeable by specific transporters. We know from first principles that most pathways will be complete, or completely absent; across most genomes. Cells are unlikely to synthesize a cofactor they cannot use, or an enzyme that needs a cofactor it cannot get. Proteins that work together in the same subsystem, and are equally necessary to some biological process, are likely to be jointly present, or jointly absent, across the vast majority of genomes. Genes needed for the same biological process tend to be co-clustered so they can be coordinately regulated. Such truisms set the stage for reasoning with genomic data, and several examples are presented below.

The similarities of one biological problem to another often force Nature to develop similar-looking solutions. Arabinose utilization requires import, then catabolism. The same holds true for mannose, fructose, glycerol, etc. The transport and catabolic operons tend to be near each other, a “uniform functional organization” [24]. Richness in a few classes of enzyme (e.g. carbohydrate kinases) tends to distinguish sugar degradation operons from others. These basic assumptions about parallels from one sugar utilization system to another, taken together with comparative genomics of 19 *Shewanella* species and simple assays of their sugar utilization, led to highly trustworthy functional predictions for over 100 different protein families [24].

Protein-sorting systems

Even for a family of proteins that has no trace of homology to any characterized protein, recognizing the pattern of design of the system to which it belongs, then reasoning by analogy, may make it possible to infer protein function. Protein sorting systems provide a superb example of what can be discovered by comparisons across multiple unrelated bacterial genomes. A recurring pattern of these systems is that a portion of the protein coding region exists only in the precursor form, serves to guide the protein to its proper cellular destination, and then is removed. Similar (often homologous) regions can lead otherwise unrelated proteins through the same sorting pathway. If a short N-terminal or C-terminal homology domain in a group of genomes actually represents a new type of protein-sorting signal, then those same genomes (and only those) should encode machinery dedicated to recognizing and removing the signal as it translocates target proteins.

Pfam [25] model PF00746, “Gram positive anchor”, describes a region found at the C-termini of typically tens of proteins per genome for those species (mostly Firmicutes) that have any at all. This small region of sequence is a C-terminal protein sorting signal [26], and the signal itself has an easily appreciated pattern of design. First, there is a signature motif

(in this case LPXTG). Next (possibly after a short spacer), there is a highly hydrophobic transmembrane (TM) alpha helix. Lastly, there is a cluster of basic residues, lysines and/or arginines. By the “positive-inside rule”, this cluster suggests the C-terminal end of the helix is oriented toward the cytoplasm. The whole region, barely longer than needed to span the membrane, occurs at the extreme C-terminal end of member proteins. The signal is recognized and cleaved by a cysteine protease called sortase [27]. Sortase A (SrtA) cleaves between Thr and Gly in the motif LPXTG, then transfers the N-terminal region a peptidoglycan precursor. The mature protein ends up immobilized on the cell surface by the transpeptidation, covalently linked to peptidoglycan through the Thr residue at its new C-terminus.

Figure 1 shows the LPXTG sorting signal, plus four additional classes of C-terminal domains whose qualitative descriptions sound almost identical. Each is short, with a signature motif, a hydrophobic transmembrane helix, and a cluster of basic residues. All five seem to represent protein-sorting signals, meaning each should have a processing enzyme to perform the requisite C-terminal cleavage. Data mining methods such as Partial Phylogenetic Profiling [9, 21] can discover what would-be protein family most strictly co-occurs with each new class of sorting signal. Peptidase or transpeptidase activity can be predicted, with high confidence, even for new families found this way that are entirely new, with no hint to their function whatsoever from experiments on even their most remote homologs.

A C-terminal protein-sorting signal may represent less than 1 % of total protein length, but knowing the pattern of design for such sorting signals shows where to look and spurs the search for new classes. Finding the key protein families generates hints to what their respective subsystems might be doing. The enzyme that cleaves PEP-CTERM proteins, and also a histidine kinase/response regulator system for regulating their expression, both occur most of the time in extended loci for Extracellular Polymeric Substance (EPS) biosynthesis, closely linked to biofilm formation by bacteria of soils, sediments, hot spring mats, corroding metal surfaces, etc [9]. The PEP-CTERM/exosortase subsystem may help modulate properties of the biofilm matrix material, which may contain protein as well as polysaccharide. Rhombosortase occurs only in species with type II secretion systems (T2SS), and its targets include some known T2SS effectors [28]; the system may modulate periplasmic vs. secreted expression of GlyGly-CTERM proteins, many of which are biodegradative enzymes that make most sense to secrete only under specialized conditions. Archaeosortase cleaves the C-terminal helix from the S-layer glycoprotein of halophilic archaea, a textbook model for protein glycosylation in prokaryotes, and may mark the end of protein maturation [7].

Inferring roles for RiPPs

Ribosomally synthesized and post-translationally modified peptide natural products (RiPPs) have a recurring pattern of design in which the RiPP precursor is small, and typically encoded near its maturase(s). The typical family of maturase can act on more than one family of RiPP precursor. Some families of precursors are substrates for more than one type of maturase. This quirk of RiPP biology allows discovery of new types of systems by

“annotation walking.” A putative peptide maturase, a member of a known family such as the cyclodehydratases, has no identified target, but an open reading frame (ORF) situated nearby is a candidate. Comparative genomics shows homologs to this ORF in similar genomic contexts, and multiple sequence alignment suggests a conserved leader peptide, a cleavage site, and a highly variable core peptide region. But what if the leader peptide of the precursor is conserved primarily to accommodate the cleavage and export machinery? Other classes of maturing enzyme might operate on the core peptide. Members of the new precursor family may show up next to no known family of maturase, but instead in the company of a novel protein that might represent a new class of maturase. Annotation walking has added to both the guild of RiPP precursor families and the guild of maturing enzymes. One branch of the enormous radical SAM enzyme superfamily, those having a C-terminal SPASM domain (named for Subtilisin, PQQ, Anaerobic Sulfatase, and Mycofactocin modification), appears responsible for even more peptide and protein modification than LanM family lanthionine synthases, which are far better known [29].

RiPP precursors and their maturing enzymes bring the discussion of design patterns and reasoning by analogy to a final topic. The gene pair, one the target, the other its modifier, is not a complete subsystem. It is just a **design motif**. Many design patterns are possible; discerning which pattern the motif fits is essential to interpreting the gene pair correctly. Bacteriocins, peptides one microbe uses to kill another, are the most familiar simply because they are the easiest to show by assay. One radical SAM/SPASM enzyme makes subtilisin A, a bacteriocin. But another, PqqE, helps make pyrroloquinoline quinone (PQQ), a cofactor. PQQ biogenesis is widespread, and highly conserved, and lacks a transporter for peptide export, and associates with large paralogous families of enzymes that interact with PQQ. The subsystem of a third radical SAM/SPASM enzyme, MftC, is widespread in Actinobacteria and closely resembles PQQ biogenesis, not subtilisin A-like subsystems, in having wide distribution, a consistent operon structure, no export system, a highly conserved precursor form, and close correlations with several large families of oxidoreductases. The presumed product, mycofactocin [30], seems far more likely to participate in redox biochemistry than in any bacteriocin-like process. It may prove important to survival of *Mycobacterium tuberculosis* inside the human macrophage during latent infection. Although the mycofactocin hypothesis remains unproven, it illustrates just how far biocuration, comparative genomics, and annotation reasoning based on patterns of design can go to generate useful hypotheses for experimental confirmation.

Conclusion

Courses in biochemistry and molecular biology often teach the details of famous model systems, such as the transcriptional repression in phage lambda lysogeny, or ribosome-mediated attenuation of the *trp* operon, to make a more general point about the principles we should learn to recognize in system after system. Spotting an apparent match to one of the classic textbook patterns of design lets the scientist apply reasoning by analogy to propose roles and functions of the new system. A little reflection shows key aspects of many such hypotheses will be testable *in silico*, at least in part, through comparative genomics. How the parts of a system work together will influence how instances of the system will vary from one genome to another. In this era of next-generation sequencing, with just days required to

create a complete genome sequence from a newly isolated microbe, comparative genomics has the potential, for various types of subsystems, to take the lead in discovery and characterization.

Acknowledgements

This work was supported by award 1203831 from the National Science Foundation and awards HHSN272200900007C and 1U19AI110819-01 from the National Institutes of Health.

References

1. Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol.* 2003; 333:863–882. [PubMed: 14568541]
2. Koski LB, Golding GB. The closest BLAST hit is often not the nearest neighbor. *J Mol Evol.* 2001; 52:540–542. [PubMed: 11443357]
3. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol.* 2009; 5:e1000605. [PubMed: 20011109]
4. Eddy SR. Accelerated Profile HMM Searches. *PLoS Comput Biol.* 2011; 7:e1002195. [PubMed: 22039361]
5. Galperin MY, Kolker E. New metrics for comparative genomics. *Curr Opin Biotechnol.* 2006; 17:440–447. [PubMed: 16978854]
6. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* 2001; 29:2994–3005. [PubMed: 11452024]
7. Haft DH, Payne SH, Selengut JD. Archaeosortases and exosortases are widely distributed systems linking membrane transit with posttranslational modification. *J Bacteriol.* 2012; 194:36–48. [PubMed: 22037399]
8. Kikuchi A, Sagami H, Ogura K. Evidence for covalent attachment of diphitynylglycerol phosphate to the cell-surface glycoprotein of *Halobacterium halobium*. *J Biol Chem.* 1999; 274:18011–18016. [PubMed: 10364251]
9. Haft DH, Paulsen IT, Ward N, Selengut JD. Exopolysaccharide-associated protein sorting in environmental organisms: the PEP-CTERM/EpsH system. Application of a novel phylogenetic profiling heuristic. *BMC Biol.* 2006; 4:29. [PubMed: 16930487] Phylogenetic profiling is successful only if protein family definitions are available so their phyletic patterns can be matched to the trait in question, which may be a phenotype, the presence of some other molecular marker, or a computed trait such as a recurring pathway hole. This paper introduces the data-mining algorithm Partial Phylogenetic Profiling (PPP), in which defining families and scoring their concordance with a phyletic pattern occurs simultaneously. PPP scores the potential match of every protein in a genome to a given taxonomic distribution (the query) by testing all possible breadths for a protein family built around that protein, and scoring the optimal breadth. The first use of PPP discovered a protein family, the exosortases, that strictly co-occurs with the novel protein-sorting signal, PEP-CTERM. Analogy to the unrelated LPXTG/sortase system, in the absence of any sequence homology, suggested that exosortase recognizes and cleaves the PEP-CTERM signal. Experimental work from the Pohlschroder lab (see reference 33) on archaeosortase, a distant archaeal homolog of exosortase, has now proven that members of the exosortase/archaeosortase family participate in protein sorting and removal of the sorting signal, confirming the original prediction.
10. Selengut JD, Haft DH. Unexpected abundance of coenzyme F₄₂₀-dependent enzymes in *Mycobacterium tuberculosis* and other actinobacteria. *J Bacteriol.* 2010; 192:5788–5798. [PubMed: 20675471] This study began when Partial Phylogenetic Profiling (PPP - see references 9 and 21) showed that large numbers of enzymes per genome in select Actinobacteria, including *Mycobacterium tuberculosis* have their closest homologs exclusively in species capable of synthesizing the cofactor F₄₂₀. The hypothesis that these flavonoid cofactor-dependent enzymes rely on F₄₂₀ rather than FMN as cofactor, was tested and confirmed *in silico* by a complementary

method, SIMBAL (Sites Inferred by Metabolic Background Assertion Labeling). SIMBAL uses contextual information to divide a protein family into YES and NO training sets, then performs data mining similar to the PPP algorithm to find the sites that most closely follow the partition rule (i.e. F₄₂₀ biosynthesis). Correlations were strongest by far for sites known from crystal structure to be involved in cofactor binding, confirming the PPP result and thus completing the bioinformatics journey.

11. Haft DH, Basu MK, Mitchell DA. Expansion of ribosomally produced natural products: a nitrile hydratase- and Nif11-related precursor family. *BMC Biol.* 2010; 8:70. [PubMed: 20500830]
12. Haft DH, Selengut JD, Brinkac LM, Zafar N, White O. Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. *Bioinformatics.* 2005; 21:293–306. [PubMed: 15347579]
13. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy- Lagard V, Diaz N, Disz T, Edwards R, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* 2005; 33:5691–5702. [PubMed: 16214803]
14. Blake JA, Dolan M, Drabkin H, Hill DP, Li N, Sitnikov D, Bridges S, Burgess S, Buza T, McCarthy F, et al. Gene Ontology annotations and resources. *Nucleic Acids Res.* 2013; 41:D530–D535. [PubMed: 23161678]
15. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, et al. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics.* 2008; 9:75. [PubMed: 18261238]
16. Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E. TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res.* 2013; 41:D387–D395. [PubMed: 23197656]
17. Meyer F, Overbeek R, Rodriguez A. FIGfams: yet another set of protein families. *Nucleic Acids Res.* 2009; 37:6643–6654. [PubMed: 19762480]
18. Kensch PR, van Noort V, Dutilh BE, Huynen MA. Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. *J R Soc Interface.* 2008; 5:151–170. [PubMed: 17535793]
19. Fouts DE, Brinkac L, Beck E, Inman J, Sutton G. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic Acids Res.* 2012; 40:e172. [PubMed: 22904089]
20. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. *Science.* 1999; 285:751–753. [PubMed: 10427000]
21. Basu MK, Selengut JD, Haft DH. ProPhylo: partial phylogenetic profiling to guide protein family construction and assignment of biological process. *BMC Bioinformatics.* 2011; 12:434. [PubMed: 22070167]
22. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 2013; 41:D808–D815. [PubMed: 23203871]
23. Koonin EV, Mushegian AR, Bork P. Non-orthologous gene displacement. *Trends Genet.* 1996; 12:334–336. [PubMed: 8855656]
24. Rodionov DA, Yang C, Li X, Rodionova IA, Wang Y, Obraztsova AY, Zagnitko OP, Overbeek R, Romine MF, Reed S, et al. Genomic encyclopedia of sugar utilization pathways in the *Shewanella* genus. *BMC Genomics.* 2010; 11:494. [PubMed: 20836887] This study exploits comparative genomics and the typical pattern of design of sugar utilization subsystems to turn simple tests of growth on various carbon sources into a highly efficient workflow for matching protein families to specific enzymatic, transport, and regulatory functions. Homologies to carbohydrate catabolism enzymes for previously known pathways help spot operons for uncharacterized carbohydrate utilization pathways (an example of annotation walking), reducing the search space. Growth tests on 18 sugars gave phenotypes for 19 *Shewanella* organisms, after which matching genotype to phenotype, operon by operon, was highly efficient. This one study resulted in over 150 new definitions of single-function protein families, 62 with novel activities and several of those with follow-up experimental confirmation, all added to FIGfams to improve automated genome annotation going forward.

25. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. Pfam: the protein families database. *Nucleic Acids Res.* 2013; 42:D222–D230. [PubMed: 24288371]
26. Schneewind O, Mihaylova-Petkov D, Model P. Cell wall sorting signals in surface proteins of gram-positive bacteria. *Embo J.* 1993; 12:4803–4811. [PubMed: 8223489]
27. Ton-That H, Liu G, Mazmanian SK, Faull KF, Schneewind O. Purification and characterization of sortase, the transpeptidase that cleaves surface proteins of *Staphylococcus aureus* at the LPXTG motif. *Proc Natl Acad Sci U S A.* 1999; 96:12424–12429. [PubMed: 10535938]
28. Sikora AE, Zielke RA, Lawrence DA, Andrews PC, Sandkvist M. Proteomic analysis of the *Vibrio cholerae* type II secretome reveals new proteins, including three related serine proteases. *J Biol Chem.* 2011; 286:16555–16566. [PubMed: 21385872]
29. Haft DH, Basu MK. Biological systems discovery in silico: radical Sadenosylmethionine protein families and their target peptides for posttranslational modification. *J Bacteriol.* 2011; 193:2745–2755. [PubMed: 21478363] This work continues the use of bioinformatics grammars (patterns of design plus patterns of signal from comparative genomics studies) to show additional examples of peptide modification systems that do not fit the default assumption that most products are toxins or pheromones. The SCIFF system is almost perfectly conserved in a branch of the Firmicutes, far more than the capacity to sporulate, suggesting a housekeeping for SCIFF, perhaps in translation or secretion. The His-Xaa-Ser repeats system, sporadically distributed, is closely linked to multiple markers of DNA mobility, suggesting a possible role in mobilizing DNA from cell to cell. A lesson from this paper is that relying on bacteriocin activity assays to learn about peptide-derived natural products, and ignoring patterns of design, is "looking under the street lamp", searching where it is easy to perform experiments while missing the chance to find new types of biological process.
30. Haft DH. Bioinformatic evidence for a widely distributed, ribosomally produced electron carrier precursor, its maturation proteins, and its nicotinoprotein redox partners. *BMC Genomics.* 2011; 12:21. [PubMed: 21223593] This study used comparative genomics methods, including annotation walking and protein family definition by hidden Markov model, to show that the SPASM domain (TIGRFAMs model TIGR04085) is one of the most common markers for post-translational modification of small peptides. A look at design patterns for different types of subsystems that feature modified peptides showed that a predicted natural product in *Mycobacterium tuberculosis*, **mycofactocin**, almost certainly acts in redox enzyme pathways within the cytosol, like the cofactor PQQ, rather than in cytotoxicity following secretion, like subtilosin A and other bacteriocins.
31. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res.* 2004; 14:1188–1190. [PubMed: 15173120]
32. Pathak DT, Wei X, Bucuvalas A, Haft DH, Gerloff DL, Wall D. Cell contact-dependent outer membrane exchange in myxobacteria: genetic determinants and mechanism. *PLoS Genet.* 2012; 8:e1002626. [PubMed: 22511878]
33. Abdul Halim MF, Pfeiffer F, Zou J, Frisch A, Haft D, Wu S, Tolic N, Brewer H, Payne SH, Pasa-Tolic L, et al. *Haloferax volcanii* archaeosortase is required for motility, mating, and C-terminal processing of the S-layer glycoprotein. *Mol Microbiol.* 2013; 88:1164–1175. [PubMed: 23651326]
34. Haft DH, Varghese N. GlyGly-CTERM and rhombosortase: a C-terminal protein processing signal in a many-to-one pairing with a rhomboid family intramembrane serine protease. *PLoS One.* 2011; 6:e28886. [PubMed: 22194940]

Highlights

We describe progress in using comparative genomics to make new discoveries.

Computational genomics provides for a code-cracking approach to annotation.

Patterns of Design let biocurators reason by analogy about unknown proteins.

Figure 1. Sequence logos for five C-terminal sorting signals

A pattern of design is seen in the domain architecture: a signature motif, then a transmembrane alpha helix, then a cluster of basic (positively charged) residues. As a rule, these sorting signals occur in many times per genome if they occur at all, toward the extreme carboxyl-terminal ends of proteins with predicted N-terminal signal peptides. Sequence logos [31] are shown to show information content, in bits, for multiple alignments derived from Pfam [25] or TIGRFAMs [16] database seed alignments after removal of gappy columns and of uninformative sequence N-terminal to the defining motif. A) PF00746 shows sorting signals

led off by the LPxTG motif, spaced a small distance from the start of the transmembrane helix. Proteins are cleaved between the 4th and 5th positions of the motif by sortase, and in most cases transferred to a peptidoglycan precursor. B) The MYXOCTERM predicted sorting signal as modeled by TIGR03901, restricted to the Myxococcales. Its processing enzyme is still unknown [32]. C). The PEP-CTERM sorting signal, modeled by TIGR02595, predicted target of exosortases (TIGR02602) in Gram-negative bacteria [9]. D) PGF-CTERM, modeled by TIGR04126, target of archaeosortase A (TIGR04125) [7, 33]. E) GlyGly-CTERM, modeled by TIGR03501, target of rhombosortase (TIGR03902) [34].