Published in final edited form as:
*Nat Chem Biol*. 2010 January ; 6(1): 4–5. doi:10.1038/nchembio.288.

# The Evolution of Gene Annotation

**Simon Kasif**[1,2,3] and **Martin Steffen**[1,4]

Simon Kasif: kasif@bu.edu; Martin Steffen: steffen@bu.edu

[1]Department of Biomedical Engineering, Boston University, Boston, MA USA

[2]Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology, Boston, MA USA

[3]Center for Advanced Genomic Technology, Boston University, Boston, MA USA

[4]Department of Pathology and Laboratory Medicine, Boston University School of Medicine, Boston, MA USA

## Abstract

Complete and accurate annotation of gene function is an essential starting point for genome interpretation and a host of systems and synthetic biology endeavors. Detecting errors in existing annotation now has an important new tool.

The explosion of DNA sequence data has advanced many fields of science, but it has done so unevenly. There has been no corresponding explosion in our knowledge of gene function. If one were to sequence a newly discovered microbe today, the percentage of genes of unknown function would be very similar to that obtained a decade ago. In the near future, this problem will likely worsen, as the new ultra-high throughput sequencing methods become more widely used, and "terabase" becomes the relevant unit for most discussion. The "new sequence" juggernaut stresses our current methods of functional annotation, and improvements are needed to both enhance the accuracy of predictions and increase the rate of experimental validation. A new method from Hsiao *et al.* now comes at this problem from a different angle: that of correcting mistaken annotations[1].

The initial wave of high-throughput genomic annotation methods began as a marriage of accurate identification of the coding sequences with automated prediction of the function of their protein products based on homology and sequence similarity to genes of known function (Fig. 1)[2]. The second generation of automated annotation systems evolved into a taxonomic organization of the protein space as clusters, stored in databases such as PFAM[3] and NCBI Protein Clusters[4], and this led to improved classification accuracy of remotely homologous proteins.

These state-of-the-art sequence based methods can now annotate a large fraction of genes in a sequenced genome. Various studies suggest that higher than 40%–60% sequence identity is required to reliably transfer broad functional annotations. However, detailed functional assignment and prediction of substrate specificity of a given protein, evaluation remains in many ways, an art[5]. In addition these methods fail to classify roughly 30% or more of a typical microbial orfeome. The third wave of systematic function prediction programs based on genomic context methods, which use information complementary to sequence homology, such as phylogenetic profiles, co-expression, chromosomal gene clustering, and protein fusion, has emerged in the past decade[6]. As an example, two genes that have adjacent chromosomal locations in multiple genomes may be predicted to be functionally related. Different measurements produce different correlations, thus the fourth generation of functional annotation methods use machine learning methods to train an integrative

classifier that predicts function using fusion of the correlations to other genes that are already associated with fully annotated genes[7].

An intriguing new direction on genomic context methods is developed by Vitkup and coworkers, who have devised a new method which recognizes errors in existing metabolic network annotation by identifying discrepancies in the fused correlation of genes that are annotated as neighbors in the metabolic network. Such methods are urgently needed as estimates of annotation error can be high[8]. The key insight from the new paper is that if two genes interact in a metabolic network but do not share a significant "fused functional correlation," one might suspect an error. For instance, three metabolic genes that are involved in a metabolic sub-system might form an operon in a genome and therefore should be expected to be co-expressed and show a strong correlation in their mRNA profiles across a set of perturbations. The "gene policing" proposal is a new application of the idea that genome context can be effectively integrated with the local topology of a metabolic network to accurately predict gene function[9]

As a proof of concept, they apply their method to correct misannotations in the leucine degradation pathway of *B. subtilis*. During sporulation, when nutrients may be limited, leucine can be degraded to acetyl-CoA and enter into the TCA cycle, providing energy for the developing spore. Several genes in the *yng* gene cluster had conflicting and/or disparate functional annotations, and were identified as having poor genomic correlations with their network neighbors. Furthermore, their algorithm predicted the functions for three of the genes (*yngI*, *yngF*, and *yngE*), as participating in leucine degradation, resulting in a contiguous set of six genes performing six consecutive enzymatic reactions in leucine degradation. The investigators go on to support these annotations experimentally, demonstrating that knockouts in any of these genes inhibited the formation of acetyl-CoA from leucine during sporulation, and impressively, that formation of acetyl-CoA from the closely related isoleucine was unaffected.

It is especially noteworthy that their approach is capable of highly detailed predictions, specifying all four components of enzyme commission (EC) numbers. In general, the majority of prediction methods have been evaluated only in the context of broad functional assignments, and it remains a relatively open question as to how well computational methods will perform when being benchmarked by their predictions of detailed metabolic functions, sub-systems components[10] and predicted substrate specificities.

While increasing the accuracy of predictions is essential, so too is the formidable task of increasing the pace of function validation. While new approaches offer promise, high-throughput experimental methods, such as transcription profiling with microarrays, generally provide insufficient information for the definitive assignment of molecular function, despite proving invaluable for hypothesis generation. How best, then, to accelerate the biochemical characterization of genes of unknown function, a task that often requires artisanal attention to minute experimental detail? One potential solution proposes the establishment of a broad community collaboration between computational biologists and experimentalists, in which high quality predictions are matched with biochemical expertise[11]. Such an effort is now gaining momentum, with a website interchange "SciBay" destined to appear soon, showcasing carefully vetted functional predictions and offering modest funds to experimentalists as an incentive to attempt validation. It is probable that predictions made by Vitkup and collaborators with the method described here, labelled as "putative erroneous assignments," would be among the most promising targets for experimental testing.

## References


1. Hsiao TL, Revelles O, Chen L, Sauer U, Vitkup D. Automatic policing of biochemical annotations using genomic correlations. Nat Chem Biol 2010;6:34–40. [PubMed: 19935659]
2. Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402. [PubMed: 9254694]
3. Finn RD, et al. The Pfam protein families database. Nucleic Acids Res 2008;36:D281–D288. [PubMed: 18039703]
4. Klimke W, et al. The National Center for Biotechnology Information's Protein Clusters Database. Nucleic Acids Res 2009;37:D216–D223. [PubMed: 18940865]
5. Murali TM, Wu CJ, Kasif S. The art of gene function prediction. Nat Biotechnol 2006;24:1474–1475. author reply 1475-6. [PubMed: 17160037]
6. Galperin MY, Koonin EV. Who's your neighbor? New computational approaches for functional genomics. Nat Biotechnol 2000;18:609–613. [PubMed: 10835597]
7. Jansen R, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 2003;302:449–453. [PubMed: 14564010]
8. Godzik A, Jambon M, Friedberg I. Computational protein function prediction: are we making progress? Cell Mol Life Sci 2007;64:2505–2511. [PubMed: 17611711]
9. Zheng Y, Szustakowski JD, Fortnow L, Roberts RJ, Kasif S. Computational identification of operons in microbial genomes. Genome Res 2002;12:1221–1230. [PubMed: 12176930]
10. Overbeek 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]
11. Roberts RJ. Identifying protein function--a call for community action. PLoS Biol 2004;2:E42. [PubMed: 15024411]

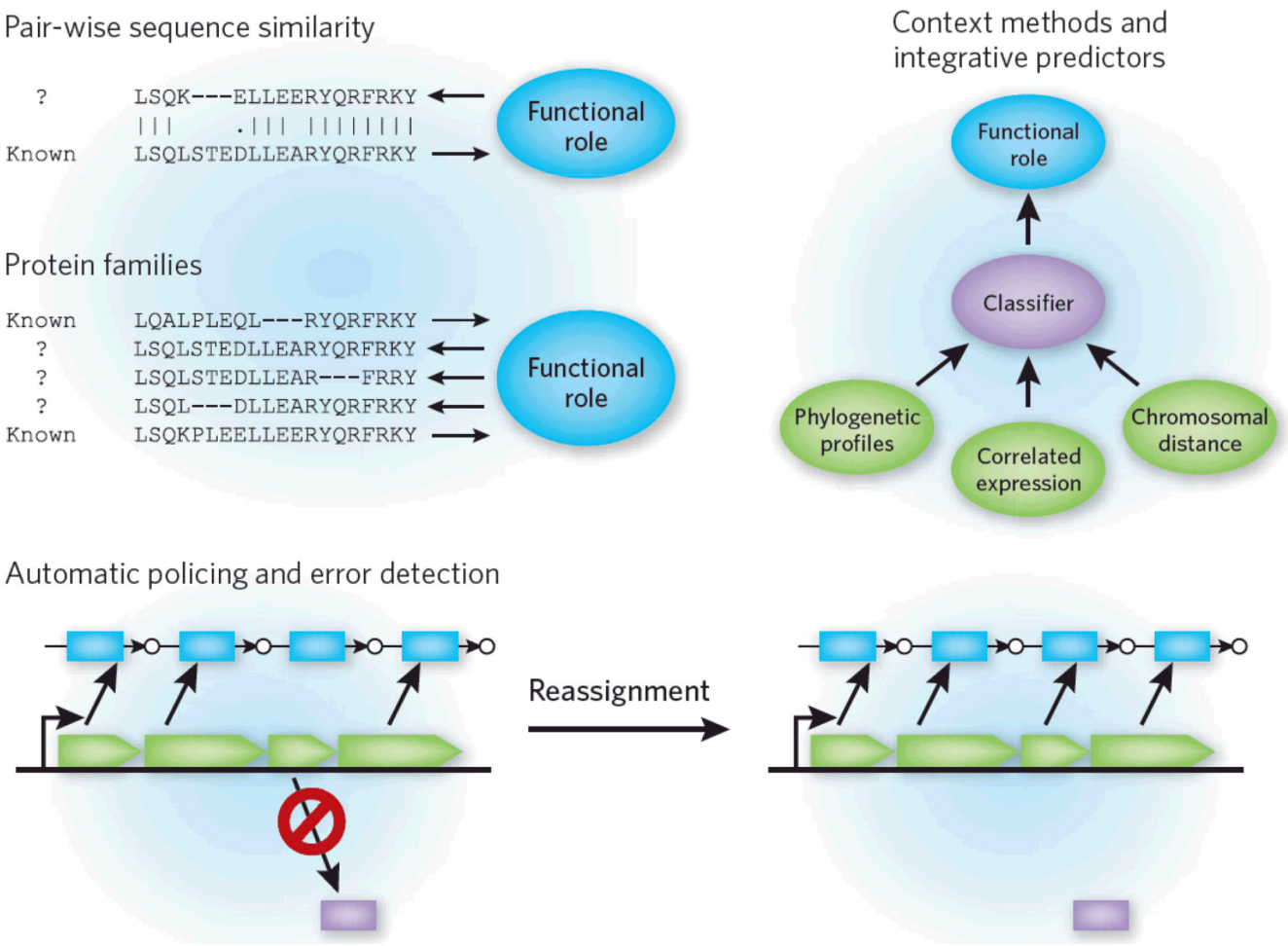


**Figure 1.**
Overview of gene function prediction methods. The earliest methods were based primarily on sequence similarity. Newer methods have improved accuracies on proteins with low sequence similarity by including orthogonal types of genomic data. The method by Vitkup and coworkers specializes in identifying mistaken annotations, employing a combination of sequence-based and integrative methods.