Errors in genome annotation

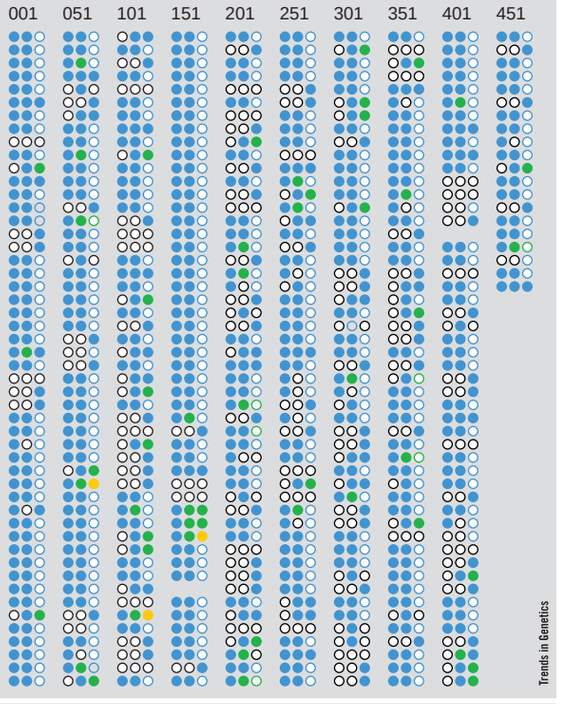
Steven E. Brenner
brenner@hyper.stanford.edu

Department of Structural Biology, Stanford University, Fairchild Building, Stanford, CA 94305-5126, USA.

At the time that Watson and Crick proposed a structure for DNA, a visionary might have suggested that the complete genetic sequence of an organism would eventually be known. However, nobody could have realistically proposed that machines could automatically indicate gene functions. Yet precisely this has been achieved: with no laboratory experiments at all, the roles of most genes in several organisms have been reported.

But how reliable are these functional assignments, upon which we depend for understanding genes and genomes? Without laboratory experiments to verify the computational methods and their expert analysis, it is impossible to know for certain. However, a simple procedure can place a rough upper bound on their accuracy. I have compared three different groups' functional annotation[1–3] for the *Mycoplasma genitalium* genome[1] (Fig. 1). Where two groups' descriptions are completely incompatible, at least one must be in error. In my analysis, there is no penalty for vague or absent functional assignment. Furthermore, I always assume that as many groups as possible have the right description (Fig. 2).

FIGURE 1. Comparison of annotations



Three dots represent (left to right) Frasier *et al.*[1], Koonin *et al.*[2] and Ouzounis *et al.*[3] annotations for each of the 468 *M. genitalium* genes. (Tentative cases from Ouzounis *et al.*[3] were not used.) An open black circle indicates lack of a substantial functional annotation. Compatible annotations are colored identically, while conflicting annotations are in different colors. It is unknown which, if any, of the annotations are actually correct. There are 300 cases where Ouzounis *et al.*[3] simply reported the SWISS-PROT annotation of the same *M. genitalium* gene, indicated by colored open circles. Because Frasier *et al.*[1] annotation played a role in SWISS-PROT descriptions, these Ouzounis *et al.*[3] annotations were not included in this analysis. Though not incorporated in Table 1, the color indicates the compatibility of the functional annotation. The conflict/compatibility analysis here is itself certain to have errors; however, these should not affect the magnitude of the measured annotation error rate.

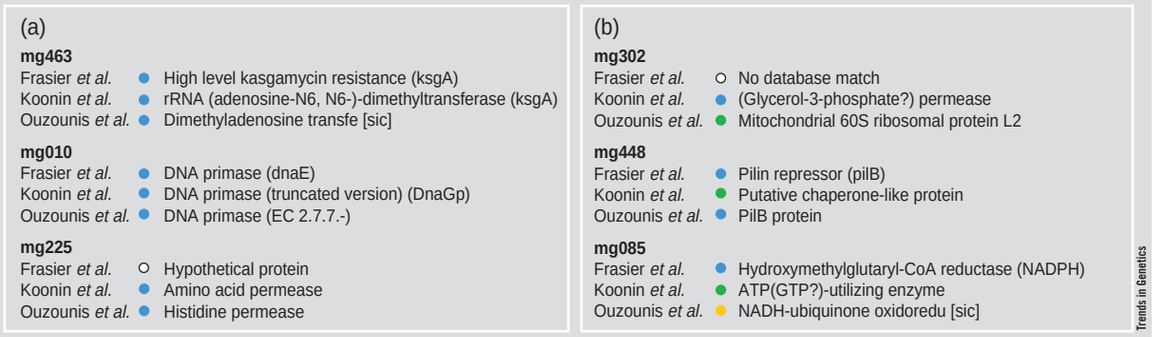
The results are disappointing for those expecting reliable annotation (Table 1). *M. genitalium* was reported to have just 468 genes, many of which are fundamental for all life and therefore easy to analyse. Nonetheless, the error rate is at least 8% for the 340 genes annotated by two or three groups. This value may not be uniform across the three groups, nor does it reflect the overall significance of a group's results. Genes annotated by only one group were not considered, but include such improbable bacterial functions as B-cell enhancing factor, mitochondrial polymerase, and seretonin receptor. This analysis cannot detect those cases where multiple groups arrived at consistent but wrong conclusions – a likely occurrence because all relied on similar methods and data. This evaluation also ignores minor disagreements in annotation, and disparities in degree of specificity (possibly indicating problematic overprediction of function[4]). Therefore, the true error rate must be greater than these figures indicate.

There are several possible reasons why the functional analyses have mistakes, as described at greater length elsewhere[5–8]. For example, it may be that the similarity between the genomic query and database sequence is insufficient to reliably detect homology, an issue solvable by appropriate use of modern and accurate sequence comparison procedures[9,10]. A more difficult problem is accurate inference of function from homology. Typical database searching methods are valuable for finding evolutionarily related proteins, but if there are only about 1000 major superfamilies in nature[11,12], then most homologs must have different molecular and cellular functions.

The annotation problem escalates dramatically beyond the single genome, for genes with incorrect functions are entered into public databases[8]. Subsequent searches against these databases then cause errors to propagate to future functional assignments. The procedure need cycle only a few times without corrections before the resources that made computational function determination possible – the annotation databases – are so polluted as to be almost useless. To prevent errors from spreading out of control, database curation by the scientific community will be essential[4,13].

To ensure that databases are kept usable, the intent of a gene annotation should be clear: does it indicate homolog, ortholog, and/or functional equivalence? Fortunately, some databases already incorporate this information explicitly (e.g. Ref. 14). Errors will, of course, still creep in. To help eliminate the collateral damage, computational assignments should clearly be flagged as such, and they should also indicate their source (which would allow propagation of corrections) and a measure of confidence in their accuracy. This will require new research and development in algorithms and databases, and a broad commitment to maintaining these resources. In short, the accessible documentation needed for reproducibility of a computational function determination should be commensurate with that for a corresponding laboratory bench experiment.

FIGURE 2. Example annotations and analysis



(a) Consistent annotations. Annotations were generally considered consistent for this analysis if either the function or the gene name match (e.g. mg463; mg010). An exception is when one group uses a gene name and another specifically notes that the current gene is a paralog and not identical (consider mg010). Where the descriptions from different groups were compatible, but of different levels of specificity, this was considered a correct assignment (e.g. mg225). The difficulty of reconciling pairs of descriptions to determine whether they reflect compatible functions makes this analysis imprecise. Generally, the approach here is generous and should err on the side of detecting too few errors; it is usually more permissive than Ref. 5. **mg463:** Frasier *et al.*[1] and Koonin *et al.*[2] describe different aspects of function, but give the same gene name. The Ouzounis *et al.*[3] description is compatible with that from Koonin *et al.*[2], but less specific. All three annotations are considered correct for this analysis. **mg010:** Frasier *et al.*[1] and Ouzounis *et al.*[3] agree that this is a DNA primase. Koonin *et al.*[2] use a different gene name and explicitly state that this is a truncated protein. Because of the common functional descriptions, all three are considered correct. However, if Koonin *et al.*[2] had been more explicit in indicating a functional difference, then their annotation would have been marked as conflicting. (Note that mg250 is also annotated as a DNA primase by all three groups.) **mg225:** the Ouzounis *et al.*[3] annotation of histidine permease is more specific than the Koonin *et al.*[2] description of amino acid permease. It may be that histidine permease is an (incorrect) overprediction of function, or it could be correct. The two annotations are considered consistent, and the decision of Frasier *et al.*[1] not to provide a function is not penalized. **(b) Inconsistent annotations. mg302:** lack of a functional assignment from Frasier *et al.*[1] is not penalized. The Koonin *et al.*[2] and Ouzounis *et al.*[3] annotations are wholly inconsistent. This leads to a conflict and a minimum error rate of 50%. Note that the assessment methodology also behaves correctly when two annotators provide different functions for a multi-functional enzyme: each of the annotators is half right and half wrong, and the assessment assigns a 50% error rate. **mg448:** Frasier *et al.*[1] and Ouzounis *et al.*[3] both describe the gene as *pilB*. The encoded protein is involved in pilin formation, and its biochemical function is catalysis of methionine sulfoxide oxidation/reduction in proteins. The Koonin *et al.*[2] annotation, chaperone-like protein, could conceivably be compatible but this is not likely. Because of uncertainty regarding compatibility of the Koonin *et al.*[2] annotation and its qualification as putative, this set of annotations is right on the threshold of consideration. For this analysis, the Koonin *et al.*[2] annotation was considered to be in conflict with the others, giving a minimum error rate of 33%. **mg085:** all three groups provide contradictory functions. The function described by Frasier *et al.*[1] of HMG-CoA reductase is EC 1.1.1.34, while the NADH-ubiquinone oxidoreductase annotated by Ouzounis *et al.*[3] (nu6m_marpo) is EC 1.6.5.3. Neither enzyme uses ATP or GTP, as specified by Koonin *et al.*[2] The analysis assumes one is correct and marks two incorrect. Note: Ouzounis *et al.*[3] annotations equivalent to SWISS-PROT included in these examples are not included in the Table 1 analysis.

Acknowledgements

A previous version of this analysis was performed at the MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK. M. Levitt, C. Chothia, B. Al-Lazikani and P. Koehl provided stimulating discussion.

TABLE I. *M. genitalium* annotations, conflicts and error rates

No. groups annotating gene	No. genes	Annotations per group[a]			Total annotations	No. conflicts	Minimum error rate
		Frasier *et al.*[1]	Koonin *et al.*[2]	Ouzounis *et al.*[3]			
0	33	–	–	–	–	N/A	N/A
1[b]	95	14	15	66	95	N/A	N/A
2	318	279	317	40	636	45	7%
3	22	22	22	22	66	10	15%
Sum (2+3)	340	301	339	62	702	55	8%

Summary of annotations made by each group (Fig. 1), minimal number of conflicting annotations (see Fig. 2), and the resulting minimal fraction of annotations that are erroneous.
[a]Frasier *et al.*[1] data from http://www.tigr.org/tdb/mdb/mgdb/mgdb.html. Koonin *et al.*[2] data from http://www.ncbi.nlm.nih.gov/Complete_Genomes/Mgen. Ouzounis *et al.*[3] data from http://www.embl-heidelberg.de/~genequiz/mycogen.new.html. Instances where Ouzounis *et al.*[3] reported SWISS-PROT annotation of the same gene were removed to avoid duplication with Frasier *et al.*[1] entries. However, even if all of these 300 annotations are included, the minimum annotation error rate drops only to 6%. All annotations were collected in 1996, shortly after the genome was released.
[b]No comparative analysis is possible when only one group made an annotation.

References

1 Frasier, C.M. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* 270, 397–403
2 Koonin, E.V. *et al.* (1996) Sequencing and analysis of bacterial genomes. *Curr. Biol.* 6, 404–416
3 Ouzounis, C. *et al.* (1996) Novelties from the complete genome of *Mycoplasma genitalium*. *Mol. Microbiol.* 20, 898–900
4 Doerks, T. *et al.* (1998) Protein annotation: detective work for function prediction. *Trends Genet* 14, 248–250
5 Galperin, M.Y. and Koonin, E.V. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement, and operon disruption. *In Silico Biol.* 1, 7
6 Smith, T.F. and Zhang, X. (1997) The challenges of genome sequence annotation or 'the devil is in the details'. *Nat. Biotechnol.* 15, 1222–1223
7 Bork, P. *et al.* (1998) Predicting function: from genes to genomes and back. *J. Mol. Biol.* 283, 707–725
8 Bork, P. and Bairoch, A. (1996) Go hunting in sequence databases but watch out for the traps. *Trends Genet.* 12, 425–427
9 Brenner, S.E. *et al.* (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. U. S. A.* 95, 6073–6078
10 Altschul, S.F. *et al.* (1994) Issues in searching molecular sequence databases. *Nat. Genet.* 6, 119–129
11 Chothia, C. (1992) Proteins. One thousand families for the molecular biologist. *Nature* 357, 543–544
12 Brenner, S.E. *et al.* (1997) Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol.* 7, 369–376
13 Smith, T.F. (1998) Functional genomics – bioinformatics is ready for the challenge. *Trends Genet.* 14, 291–329
14 Tatusov, R.L. *et al.* (1997) A genomic perspective on protein families. *Science* 278, 631–637