Genome Analysis

# On the total number of genes and their length distribution in complete microbial genomes

Marie Skovgaard, Lars Juhl Jensen, Søren Brunak, David Ussery and Anders Krogh

**In sequenced microbial genomes, some of the annotated genes are actually not protein-coding genes, but rather open reading frames that occur by chance. Therefore, the number of annotated genes is higher than the actual number of genes for most of these microbes. Comparison of the length distribution of the annotated genes with the length distribution of those matching a known protein reveals that too many short genes are annotated in many genomes. Here we estimate the true number of protein-coding genes for sequenced genomes. Although it is often claimed that *Escherichia coli* has about 4300 genes, we show that it probably has only ~3800 genes, and that a similar discrepancy exists for almost all published genomes.**

The most reliable method for identifying genes is by similarity to a protein in another organism. Genes with no match to known proteins can be predicted using statistical measures. The most important measure is the codon usage; that is, the difference of the frequency of codons in a gene compared with what would be expected by chance from a given base composition. However, the discriminatory power of a codon usage measure becomes less reliable for shorter open reading frames (ORFs), and this is also the case for other measures of coding potential, such as dicodon or hexamer statistics. Together with the large number of short random ORFs, this tends to give an over-prediction of short genes. Because stop triplets (TAA, TGA, TAG) are AT rich, their frequency is generally higher in AT-rich organisms than in GC-rich organisms, so the likelihood of long ORFs occurring by chance increases with increasing GC content of the organism. Therefore, the problem of discriminating between short proteins and random ORFs is generally less in AT-rich organisms than in GC-rich organisms, as shown in Fig. 1.

The increasing use of sequence databases in molecular biology makes it important to consider the accuracy of the information stored in them. Careful annotators have clearly marked unconfirmed genes as hypothetical. However, many users of the databases assume that all annotated genes indeed correspond to true genes, and this can easily lead to wrong conclusions. An example is a recent study of protein length distributions for the three domains of life[1]. On the basis of the annotation, it was concluded (among other things), that the average or median length of proteins is smaller in Archaea than Bacteria. This is owing to the fact that a very large number of short ORFs are annotated as genes in some of the archaeal organisms. When including only proteins confirmed by a match to a known protein, there seems to be no significant difference in the average (or median) lengths (Fig. 2).

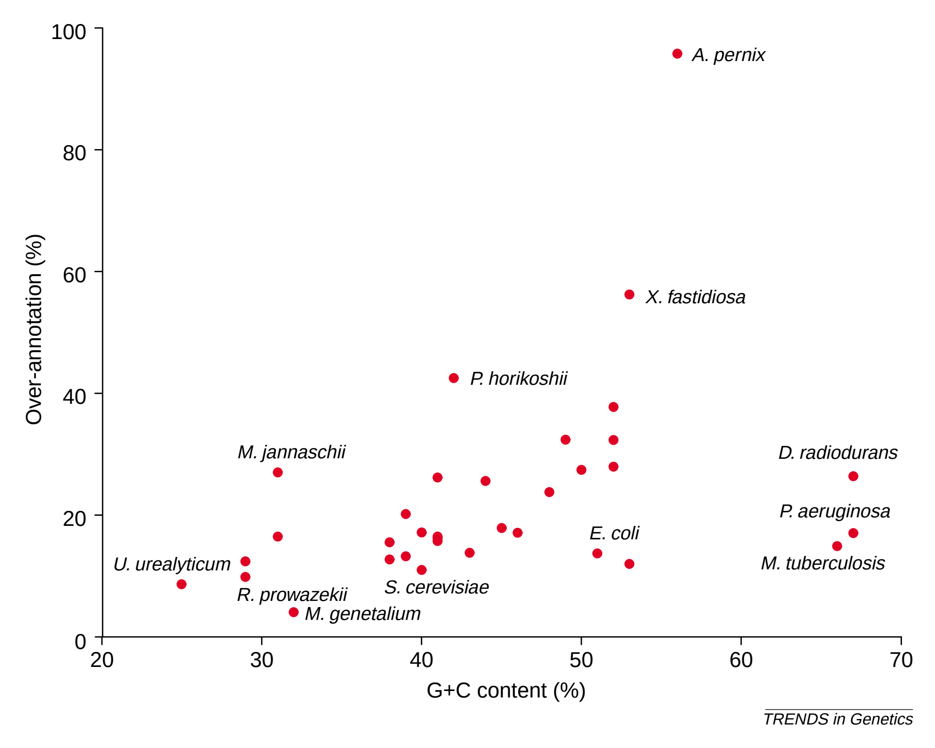

**Fig. 1.** Estimated over-annotation of genes in sequenced genomes. For each organism the SWISS-PROT-based estimate is calculated and the difference to the number of annotated genes shown in percent of the estimated number of genes.

## Length distributions

Shortly after the publication of the complete *S. cerevisiae* sequence, it was shown that there was a systematic error in the coding sequence (CDS) assignments. More than 400 sequences with lengths between 100 and 110 amino acids had no matches to previously assigned proteins[2]. This group stood out as a peak in the length distribution and seemed to be an artefact.

Similarly, we have plotted the distribution of protein lengths for each organism found in GenBank (release 119; http://www.cbs.dtu.dk/krogh/genomes/). Figures 3 and 4 are examples showing the length distribution of the unique dataset confirmed by a match to a protein not marked 'hypothetical' in SWISS-PROT, and the length distribution of those without a match. A very large protein family would result in a peak in the length distribution, which is avoided

http://tig.trends.com 0168–9525/01/$ – see front matter © 2001 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(01)02372-1

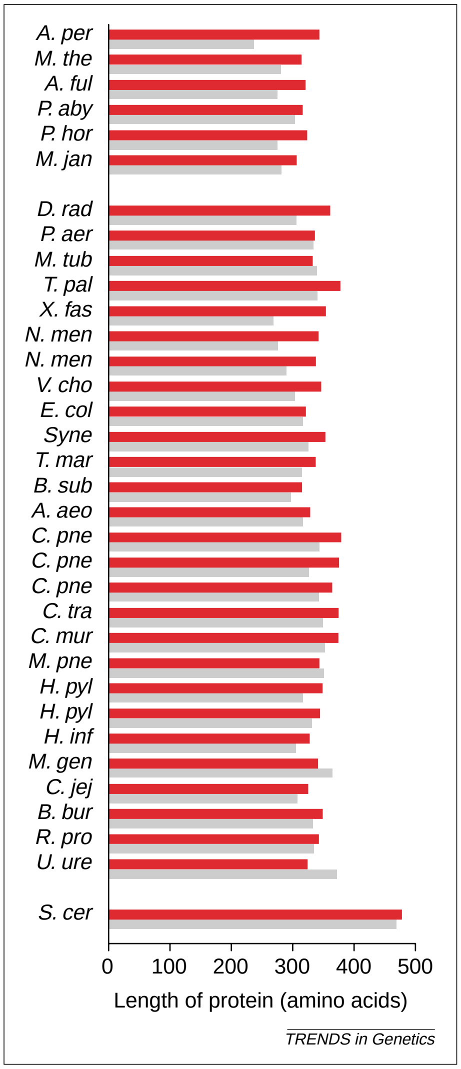

**Fig. 2.** Average length of annotated and confirmed proteins. Red bars show average length of all proteins having matches to non-hypothetical proteins in SWISS-PROT (see Methods) and gray bars show the average length of all annotated proteins. The ordering of organisms is the same as in Table 1.

when using what we will refer to here as the 'unique dataset', in which sequences with similarity to others are taken out (see Methods). At the same time, strong similarity reduction is expected to make the random sequences more dominant in the unique set, because it is not very likely that two random ORFs are similar.

*E. coli* is one of the most studied microbial organisms, and the plot in Fig. 3 reveals a significant difference between the length distributions of unique sequences matching and not matching SWISS-PROT. The sequences not matching SWISS-PROT are generally shorter than the ones matching. Actually, 81% of the 974 proteins that were excluded as 'not matching SWISS-PROT' did have matches to hypothetical proteins in SWISS-PROT, which were not counted. These were mostly from *E. coli* or closely related organisms. The definition for the keyword hypothetical in a SWISS-PROT entry is 'predicted proteins for which there is no experimental evidence that they are expressed *in vivo*.'

The length distribution of the annotated coding sequences in the Archaea *Aeropyrum pernix* is shown in Fig. 4. This is quite extreme, because rather than performing actual gene finding, all ORFs with a length of at least 100 triplets were annotated as coding regions in the GenBank entry[3]. The subset of the unique dataset with matches in SWISS-PROT has a distribution comparable to the distribution seen in other prokaryotic organisms, whereas the length distribution of genes in our dataset that do not match annotated genes has a peak that is not too different from the geometric distribution expected in a random sequence of DNA.

### Estimating the true number of protein-coding genes

The length distributions indicate that too many protein-coding genes are annotated. To obtain an estimate of the true number of proteins in each organism, we have used the proteins in the SWISS-PROT database[4] that are not labeled hypothetical, as a reference. The estimate is based on the assumption that the fraction of proteins with a match in SWISS-PROT is independent of the length of the proteins. Because ORFs longer than 200 amino acids are unlikely to occur by chance (apart from long repetitive sequences), the fraction of those matching SWISS-PROT was used as an estimate of the fraction of the total number of true proteins that match SWISS-PROT. Then, the estimated number of genes is easily obtained by dividing the total number of matching proteins with this fraction. For instance, assume that 1400 of 2000 annotated genes longer than 200 amino acids have a match in SWISS-PROT (70%). If there is a total of 2100 annotated genes with a match in SWISS-PROT, we estimate that the total number of genes is 2100/0.7 = 3000. These estimates are shown in Table 1 and the percentage over-annotation according to the estimate is shown in Fig. 1.

We have argued that some of the short annotated genes are not real. Because of the difficulty in disciminating a short ORF from a truly expressed gene, it is also quite likely that there are short genes that have not been annotated. The alignment-based estimate implicitly assumes that there are quite few of these unannotated genes. This might not hold for all organisms, which would also mean that the estimate is too low. Thus, the alignment-based estimate is really an estimate of how many of the annotated genes are correct.

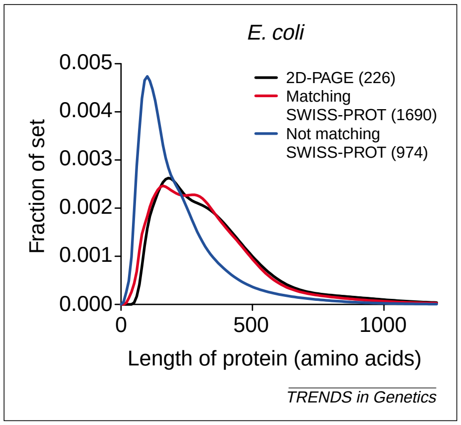

**Fig. 3.** Protein length distributions for *Escherichia coli*. The red and blue lines show the length distributions of unique sequences matching and not matching entries from SWISS-PROT, respectively. The black line shows the distribution of unique proteins from SWISS-2D-PAGE. All length distributions are normalized and have been smoothed by Gaussian kernel density estimation.

It is possible that matching proteins are biased in length, because of the local search method (e.g. due to domain structure) and possibly an inherent length bias in SWISS-PROT, both of which would invalidate our assumption that the fraction of matching genes is approximately the same for long and short genes. If matching proteins are biased towards long proteins (which is perhaps most likely) our estimate is too low. For the genomes we have studied, approximately 15–20% of the matching proteins are shorter than 200 amino acids. To get an idea of the worst-case error, let us assume that proteins longer than 200 amino acids are twice as likely to match SWISS-PROT compared with those shorter than 200 amino acids. Then, the true number of genes would be 15–20% larger than our estimate (equal to the fraction of matching proteins shorter than 200).

As a control of the above alignment estimate, we have calculated another estimate, which is completely independent of database matches.

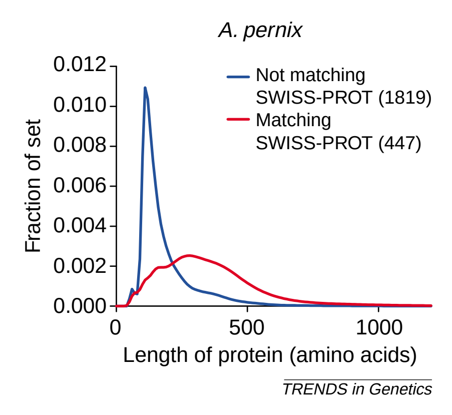

**Fig. 4.** Protein length distributions for *Aeropyrum pernix*. The curves are similar to those in Fig. 3, except that there is no 2D-PAGE data.

**Table 1. Complete microbial genomes from GenBank release 119[a]**

| | A+T (%) | Number of | Number of genes estimated from | | Number |
|---|---|---|---|---|---|
| | | annotated genes | SWISS-PROT | Stop triplet | of COGs |
| **Archaea** | | | | | |
| *A. pernix* | 44 | 2694 | 1376 | 1423 | 1169 |
| *M. thermoautotrophicum* | 50 | 1869 | 1466 | 1535 | 1375 |
| *A. fulgidus* | 51 | 2407 | 1818 | 1927 | 1849 |
| *P. abyssi* | 55 | 1765 | 1497 | 1635 | 1443 |
| *P. horikoshii* | 58 | 2064 | 1448 | 1616 | 1365 |
| *M. jannaschii* | 69 | 1715 | 1350 | 1573 | 1320 |
| **Bacteria** | | | | | |
| *D. radiodurans* | 33 | 2937 | 2323 | 1904 | 2176 |
| *P. aeruginosa* | 33 | 5565 | 4753 | 3508 | 4191 |
| *M. tuberculosis* H37Rv | 34 | 3918 | 3410 | 2537 | 2668 |
| *T. pallidum* | 47 | 1031 | 920 | 820 | 707 |
| *X. fastidiosa* | 47 | 2766 | 1770 | 1792 | 1491 |
| *N. meningitidis* A | 48 | 2121 | 1539 | 1447 | 1455 |
| *N. meningitidis* B | 48 | 2025 | 1530 | 1500 | – |
| *V. cholerae* | 48 | 3828 | 2991 | 2931 | 2745 |
| *E. coli* K-12 | 49 | 4289 | 3771 | 3463 | 3327 |
| *Synechocystis* PCC6803 | 52 | 3169 | 2559 | 2550 | 2113 |
| *T. maritima* | 54 | 1846 | 1576 | 1564 | 1507 |
| *B. subtilis* | 56 | 4100 | 3263 | 3330 | 2803 |
| *A. aeolicus* | 57 | 1522 | 1337 | 1412 | 1317 |
| *C. pneumoniae* CWL029 | 59 | 1052 | 903 | 909 | 647 |
| *C. pneumoniae* AR39 | 59 | 997 | 790 | 913 | – |
| *C. pneumoniae* J138 | 59 | 1070 | 921 | 910 | – |
| *C. trachomatis* | 59 | 894 | 772 | 754 | 631 |
| *C. muridarum* | 60 | 818 | 698 | 763 | – |
| *M. pneumoniae* M129 | 60 | 677 | 610 | 617 | 423 |
| *H. pylori* 26695 | 61 | 1566 | 1303 | 1384 | 1081 |
| *H. pylori* J99 | 61 | 1491 | 1316 | 1351 | 1062 |
| *H. influenzae* Rd | 62 | 1709 | 1479 | 1526 | 1504 |
| *M. genitalium* | 68 | 480 | 461 | 474 | 376 |
| *C. jejuni* | 69 | 1654 | 1420 | 1494 | 1289 |
| *B. burgdorferi* | 71 | 850 | 756 | 772 | 694 |
| *R. prowazekii* | 71 | 834 | 759 | 795 | 674 |
| *U. urealyticum* | 75 | 613 | 564 | 556 | 401 |
| **Eukaryota** | | | | | |
| *S. cerevisiae* | 62 | 6269 | 5560 | 5728 | 2175 |

[a]The table contains the number of annotated proteins, the A+T content, the number of proteins estimated from matches to SWISS-PROT and from stop triplet frequency, and the number of clusters of orthologous genes (COGs) for the organism. The list is ordered by kingdom and A+T content.

We found the maximal number of non-overlapping ORFs longer than 100 triplets and obtained the estimate of the true number of genes by reducing this number by those expected to occur at random. The reduction was estimated from the stop-triplet frequency. These estimates are also shown in Table 1. ORFs shorter than 100 triplets were excluded because relatively few genes are expected, and the estimate becomes ill-behaved because of the huge number of short ORFs. For high AT content, the number of nonoverlapping ORFs is a reasonably good estimate in itself, whereas the corrections become increasingly important as the AT content decreases. The approximation of the corrections is quite crude, and overall the estimate is not very good for low AT content. Indeed, the largest discrepancies between the two estimates is seen for the organisms with low AT, whereas they are quite small for intermediate to high AT content. The stop-triplet-based estimate is usually higher than the alignment-based estimate.

The number of genes that are members of clusters of orthologous genes[5] (COGs) are also shown in Table 1. A COG is defined if a gene is found in at least three lineages, so this should be an approximate lower bound on the number of protein-coding genes in an organism. These numbers are lower than the alignment based estimates except for *Archeoglobus fulgidus* and *Haemophilus influenzae* where they are very close.

It is notable that such a large fraction of the hypothetical *E. coli* ORFs had no significant matches to verified SWISS-PROT entries. The two-dimensional polyacrylamide gel electrophoresis data from SWISS-2DPAGE[6] was used as an independent control of the length distribution. We retrieved 271 *E. coli* sequences from the database, and the length distribution of the similarity reduced set of 226 is shown in Fig. 3. Although it is a fairly small set of genes, it is confirmed experimentally and independent of biasses in alignment. These data support our assumption that the length distribution of SWISS-PROT matches is reasonably unbiased. First, we investigated the possibility of using DNA-expression-array experiments, but they turned out to be unreliable as a control, because a probe for a wrongly annotated gene can be located in an untranslated region of an mRNA from an expressed gene, and will therefore appear to be expressed. Second, probes are often made for the annotated genes, making the results dependent on the annotation.

## Conclusion

Our estimates of the number of real protein-coding genes reduce the number of true proteins by 10–30% for the majority of microbial organisms. The two extremes are represented by *Mycoplasma genetalium*, where the estimates are 1–5% lower, and *A. pernix*, where they are close to 50% lower. The large over-annotation of *A. pernix* has previously been noted[7,8]. Natale *et al.*[7] estimate the correct number of protein-coding genes to be between 1550 and 1700 based on the assumption that the total fraction of confirmed genes should be about the same as for other organisms. However, because the other organisms are also over-annotated, this estimate is perhaps still too high, which explains our slightly lower estimate of about 1400 genes in *A. pernix*. It is also possible that our estimates are slightly low, as discussed above.

The problem with wrongly annotated protein-coding genes is almost entirely due to the difficulty in distinguishing short non-coding ORFs from real genes. The problem cannot be solved at present, but there are several ways in which the situation can be improved until better gene identification methods are developed. First, a measure of statistical significance for gene prediction is needed. Second, an ORF should only be annotated as a CDS if it has either a trustworthy protein match or if it has very high significance. Third, other possible genes should be annotated as ORFs, clearly showing that they are hypothetical.

## Methods

We have analyzed 34 fully sequenced microbial genomes as found in GenBank release 119. For each organism all sequences annotated as 'CDS' in the feature table were extracted and translated to proteins. To generate the unique dataset, these sequences were aligned against themselves using gapped BLASTP[9]. With a threshold of $10^{-3}$ on the expectation scores, we subsequently generated maximal similarity reduced versions of the datasets using the algorithms by Hobohm *et al.*[10]. This procedure reduced the sets by 13–36%.

The full sets were searched against the SWISS-PROT database[4] (releases 38 to 39.7) using BLASTP. Matches to sequences with the keyword 'hypothetical' were discarded. Sequences giving no hits in SWISS-PROT with an expectation score better than $10^{-3}$ were categorized as not matching SWISS-PROT, and sequences were considered to match SWISS-PROT (referred to here as matching proteins) only if at least one match with a score better than $10^{-6}$ was obtained. Sequences for which the best match had an expectation score between $10^{-3}$ and $10^{-6}$ were considered in the 'gray zone' and were not included in any of the categories (typically 3–5%).

Average lengths of all annotated genes and for all matching SWISS-PROT were calculated and used for the histogram in Fig. 2.

Length distributions were calculated for all annotated CDSs, the unique set of annotated CDSs, the unique set having matches to SWISS-PROT, and the ones not matching SWISS-PROT. Rather than plotting raw histograms, we made a Gaussian kernel density estimation of the logarithmic length distribution and log-transformed the distribution back to an ordinary length distribution. The width of the kernel was estimated from the data[11].

An estimate of the true number of genes was calculated by extrapolating from the proportion of annotated genes of length greater than 200 amino acids ($ORF_{200}$) that matches SWISS-PROT entries ($SP_{200}$). The total number of annotated genes matching SWISS-PROT ($SP_{\text{all}}$) was then divided by this ratio to get an estimate of the total number of genes ($SP_{\text{all}}\, ORF_{200}\, / SP_{200}$).

Another estimate of the total number of genes, independent of database matches, can be obtained from the ORF lengths and stop-triplet frequencies:

$$G = \sum_{i=100}^{L} N_{max}(i)\frac{N_{orf}(i) - Ap_{ran}(i)}{N_{orf}(i)}$$

where $L$ is the length of the genome divided by 3, $N_{\text{orf}}(i)$ is the observed number of ORFs of triplet-length $i$, and $N_{\text{max}}(i)$ is the number of these in a set of nonoverlapping ORFs longer than 100 triplets constructed by excluding the shortest ORFs first. $p_{\text{ran}}(i)$ is the probability of finding an ORF of triplet length $i$ at a specific position in the genome, which can be approximated by $p_{\text{ran}}(i) = 2\,p^2_{\text{stop}}(1{-}p_{\text{stop}})^i$, where $p_{\text{stop}}$ is the stop triplet frequency. $A$ is the number of triplets in the genome not occupied by true genes, which can be found by solving the self-consistency equation:

$$L - A = \sum_{i=100}^{L} iN_{max}(i)\frac{N_{orf}(i) - Ap_{ran}(i)}{N_{orf}(i)}$$

The number of ORFs grows exponentially as the length $i$ goes to zero. Therefore, the difference $N_{\text{orf}}(i){-}A\,p_{\text{ran}}(i)$ between the two very large numbers in these formulas is not well determined for short ORFs. This is why we estimate $G$ only from ORFs longer than 100 triplets.

## Acknowledgement

This work was supported by the Danish National Research Foundation.

### References

1 Zhang, J. (2000) Protein-length distributions for the three domains of life. *Trends Genet.* 16, 107–109

2 Das, S. *et al.* (1997) Biology's new rosetta stone. *Nature* 385, 29–30

3 Kawarabayasi, Y. *et al.* (1999) Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Res.* 6, 83–101

4 Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28, 45–48

5 Tatusov, R. *et al.* (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29, 22–28

6 Hoogland, C. *et al.* (2000) The 1999 SWISS-2DPAGE database update. *Nucleic Acids Res.* 28, 286–288

7 Natale, D.A. *et al.* (2000) Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). *Genome Biol.* 1, research/0009.1-19 (http://www.genomebiology.com)

8 Cambillau, C. and Claverie, J.M. (2000) Structural and genomic correlates of hyperthermostability. *J. Biol. Chem.* 275, 32383–32386

9 Altschul, S. *et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402

10 Hobohm, U. *et al.* (1992) Selection of representative protein datasets. *Protein Sci.* 1, 409–417

11 Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis.* Chapman & Hall

**Marie Skovgaard**
**Lars Juhl Jensen**
**Søren Brunak**
**David Ussery**
**Anders Krogh**
Center for Biological Sequence Analysis , BioCentrum-DTU, The Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark.
e-mail: krough@cbs.dtu.dk