doi:10.1016/j.jmb.2003.09.016 J. Mol. Biol. (2003) 333, 885–892

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

COMMUNICATION

A 'PolyORFomic' Analysis of Prokaryote Genomes Using Disabled-homology Filtering Reveals Conserved But Undiscovered Short ORFs

Paul M. Harrison[1]*, Nicholas Carriero[2], Yang Liu[1] and Mark Gerstein[1,2]

[1]*Department of Molecular Biophysics and Biochemistry Yale University, 266 Whitney Avenue, P.O. Box 208114, New Haven, CT 06520-8114, USA*

[2]*Department of Computer Science, Yale University, 266 Whitney Avenue, P.O. Box 208114, New Haven, CT 06520-8114, USA*

**Corresponding author*

Prokaryote gene annotation is complicated by large numbers of short open reading frames (ORFs) that arise naturally from genetic code design. Historically, many hypothetical ORFs have been annotated as genes in microbes, usually with an arbitrary length threshold (e.g. greater than 100 codons). Given the use of such thresholds, what is the extent of genuine undiscovered short genes in the current sampling of prokaryote genomes? To assess rigorously the potential under-annotation of short ORFs with homology, we exhaustively compared the polyORFome—all possible ORFs in 64 prokaryotes (53 bacteria and 11 archaea) plus budding yeast—to itself and to all known proteins. The novelty of our analysis is that, firstly, sequence comparisons to/between both annotated and un-annotated ORFs are considered, and secondly a two-step disabled-homology filter is applied to set aside putative pseudogenes and spurious ORFs. We find that un-annotated homologous short ORFs (uhORFs) correspond to a small but non-negligible fraction of the annotated prokaryote proteomes (0.5–3.8%, depending on selection criteria). Moreover, the disabled-homology filter indicates that about a third of uhORFs correspond to putative pseudogenes or spurious ORFs. Our analysis shows that the use of annotation length thresholds is unnecessary, as there are manageable numbers of short ORF homologies conserved (without disablements) across microbial genomes. Data on uhORFs are available from http://pseudogene.org/polyo

Keywords: gene annotation; bioinformatics; pseudogenes; hypothetical ORFs

We have now entered the era of "polygenomics", with the sequencing of a microbial genome a commonplace event and the rate of completion of genomes increasing rapidly each year.[1] Hypothetical ORFs are open reading frames that are annotated during microbial genome analysis that do not have any supporting functional information or experimental evidence of expression, or any sequence homology to known proteins motifs or domains. Large numbers of such hypothetical ORFs are annotated in the prokaryote genomes, with many annotators typically using an arbitrary minimum ORF length cut-off for inclusion in the final annotation (e.g. 60 codons for *Lactococcus lactis*[2] or 100 codons for *Aeropyrum pernix*[3,4]). In all of the sequenced archaea and bacteria (and budding yeast) an anomalous peak is observed in distributions of ORF lengths for hypothetical ORFs that is attributable to the use of such thresholds.[5] However, the trend for sequence lengths of known genes and those that are homologous to known genes does not show this behavior.[5] This peak phenomenon is related to the fact that many shorter ORFs of 200 codons or less that have been annotated as genes, are actually "generated" ORFs that arise from the design of the genetic code.[6–8] Substantial reductions in numbers of annotated genes (of up to 30%) for microbes can be derived from analysis of known protein homologies, stop codon frequencies, and nucleotide composition analysis.[5,6,8]

Abbreviation used: ORF, open reading frame.

E-mail address of the corresponding author: harrison@csb.yale.edu

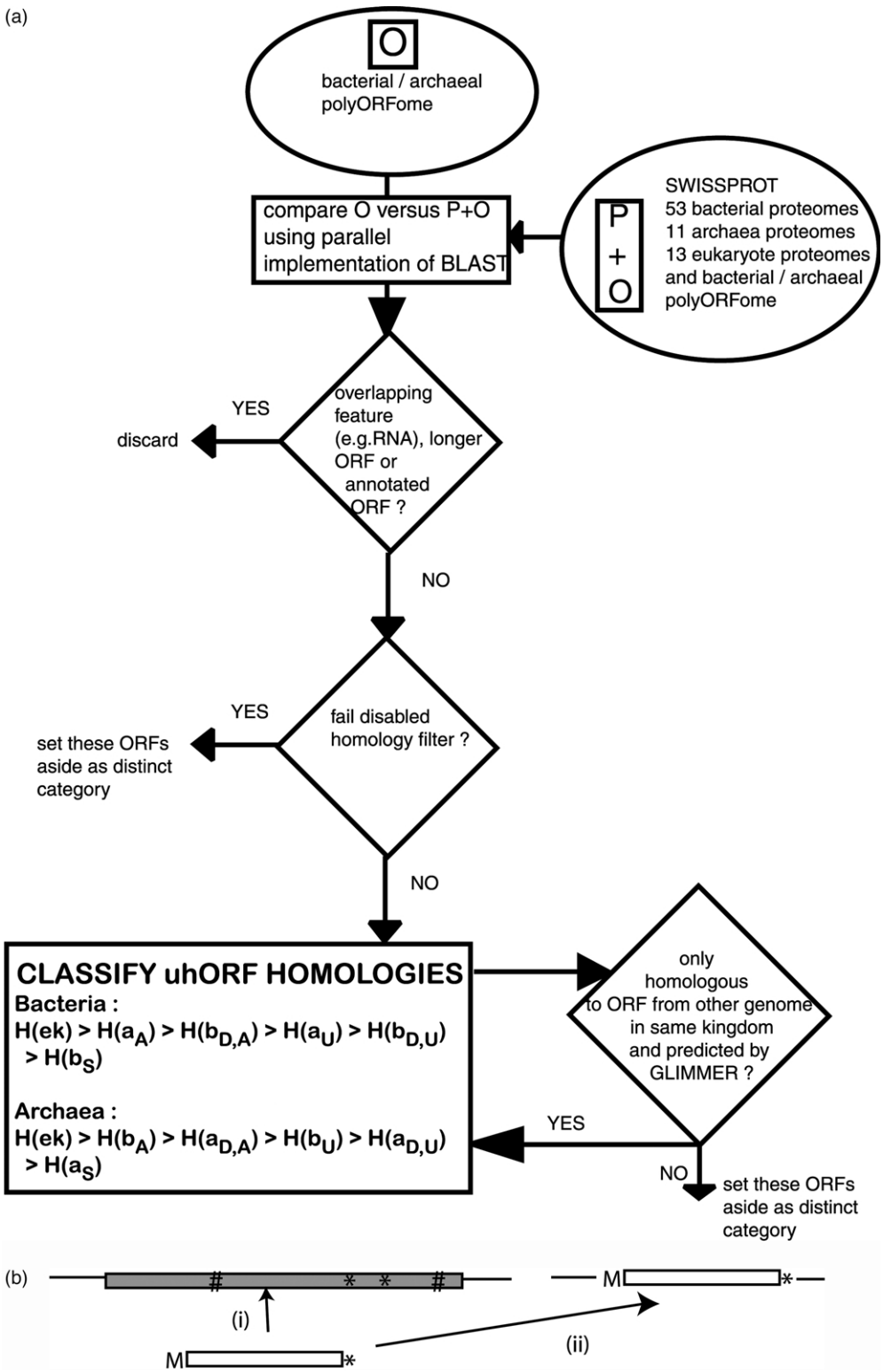

Figure 1 (*legend opposite*)

Conversely, many genuine small ORFs may be lost in genome annotation because of the aforementioned threshold strategy. To help to address the under-estimation of short ORF numbers in microbial genome annotation here, we use large-scale polygenomic sequence comparison, to make a homology-based assessment of potential short genes across a large number of microbial genomes. To do this, we derive the "polyORFome" of all possible ORFs of >15 codons in 64 prokaryotes

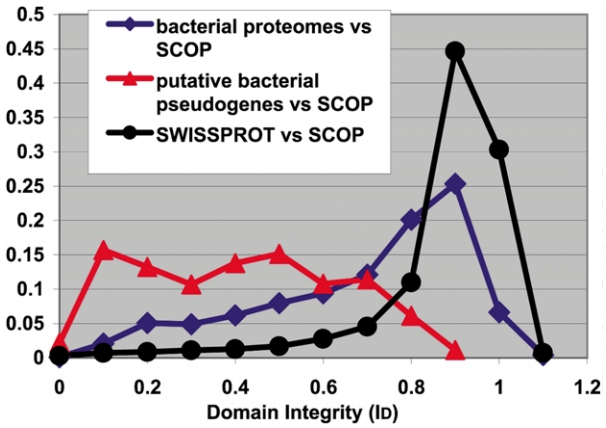

Figure 1. (a) Protein-level homology filtering scheme for uhORFs. We downloaded the genomes and gene annotations for 53 bacteria, 11 archaea and the eukaryote *S. cerevisiae* from http://www.ebi.ac.uk/genomes at the EBI. From these 65 microbial genome sequences, we generated the file of all possible open reading frames (ORFs, i.e. sequence stretches going from a start codon to a stop codon) >15 codons long (3,243,782 in total: 2,580,955 from bacteria, 535,151 from archaea and 127,676 from budding yeast). This is termed the polyORFome. We performed all-against-all sequence comparisons of the polyORFome ORFs in translation, and we compared the polyORFome to the prokaryotic proteomes plus 12 proteomes from completely sequenced eukaryotes, and SWISSPROT,[26] applying a parallel implementation of BLAST 2.2.5 and *e*-value cut-off $= 10^{-4}$, run on a cluster of 12 dual 2.4 GHz Xeon processor nodes with an *ad hoc* combination of scripts and manual intervention. The cluster load was assessed periodically to identify a list of under-used nodes, which was then fed into a launching script along with an identifier for a group of splits (i.e. the set of query files arising from one sequence file) and a starting split. The launching script started one BLAST run on each of the listed nodes, selecting a different split as the file of queries for each blast run and using the entire list of sequence files as the databases to be searched. A progress script scanned output files to provide an estimate of the amount of progress made. *PolyORFome homology filtering scheme to derive refined list of uhORFs*: protein-level sequence homologies for the polyORFome are filtered as shown. uhORFs (un-annotated homologous ORFs) are defined as un-annotated ORFs that have homology to a known protein, or to an annotated or un-annotated ORF from another kingdom, or to an annotated or un-annotated ORF from the same kingdom that is predicted as a gene. These uhORFs are filtered for overlap with other genomic features, including RNA (or sequences homologous to RNA, when translated), for overlap with longer ORFs and with annotated ORFs. They are then passed through a disabled-homology filter. If uhORFs are found to (i) overlap a longer disabled homology to an annotated protein or (ii) have multiple disabled homologs in the same genome (and no orthologs), they are labeled as likely to be non-coding (either pseudogenes or spurious ORFs). Annotation of putative pseudogenes will be described in detail elsewhere. A standard gene prediction program (GLIMMER[12]) is used to assess uhORFs that are homologous only to ORFs (either annotated or un-annotated) in a genome from the same kingdom of life. This program predicts many more potential genes for ORF lengths of <100 codons than plus budding yeast (which was studied individually in this way).[9–11] We use the simple principle of protein-level sequence homology to survey for uhORFs (un-annotated homologous ORFs) in this polyORFome. uhORFs are defined as protein-level ORF sequence homologies either to known proteins, or to annotated/un-annotated ORFs from another kingdom of life, or to annotated/un-annotated ORFs in other genomes from the same kingdom of life, that are predicted

---

are usually annotated during standard prokaryote annotation pipelines.[12] We did not require detection of a Shine–Dalgarno sequence, since for many prokaryotes a large proportion of known genes do not have a detectable one.[27] *uhORF homology classification*: the uhORFs are classified by their profile of protein-level sequence homologies. For bacteria, the uhORFs resulting from the homology filter scheme are classified as follows: (i) homologous to eukaryotic proteins or to known proteins of any sort (denoted H(ek)); otherwise (ii) homologous to annotated archaeal ORFs (denoted $H(a_A)$); otherwise (iii) homologous to annotated ORFs from a different bacterial species that are well-predicted by GLIMMER (denoted $H(b_{D,A})$); otherwise (iv) homologous to un-annotated archaeal ORFs (denoted $H(a_U)$); otherwise (v) homologous to un-annotated ORFs from a different bacterial species that are well-predicted by GLIMMER (denoted $H(b_{D,U})$);otherwise (vi) homologous to any ORF in the same genome (denoted $H(b_S)$). This last category is a catch-all for any ORF that has no verifying homology to a known protein or to an ORF in a different organism. A similar classification is used for the archaea and for the annotated proteomes. For the bacterial and archaeal annotated proteomes, the last category contains all annotated proteins not having an ortholog. It is labelled $\sim H(ekb_Aa_{D,A})$ for archaea, and $\sim H(eka_Ab_{D,A})$ for bacteria. In the box at the end of the flow-chart, entitled Classify uhORF homologies, the symbol > here means otherwise. The strings such as H(ek) > $H(a_A)$ > …, etc. thus signify the order of precedence of the different homology classes. (b) Disabled-homology filtering. ORFs or uhORFs are set aside if they are part of a larger disabled homology to another annotated protein, or have multiple disabled homologies in the same genome (and no orthologs in other species). Frameshifts are represented by the symbol # and stop codons by ∗. (c) Examination of protein domain integrity supports the assignment of disabled homologies to known proteins, as pseudogenes. This shows the distribution of domain integrity (I_D) for different sequence sets for the bacterial proteomes. DI is defined as the completeness of the highest-scoring structural match to a known protein domain (from SCOP[20]), in a sequence *S*, and is given by, $I_D = M_D/L_D$, where M_D is the largest length of matching sequence to the domain (undisrupted by stop codons and frameshifts, in the case of putative ψg) that corresponds to sequence *S* in a FASTA alignment, and L_D is the length of the protein domain sequence. The I_D distributions are derived for SCOP domain matches to: putative prokaryote ψgs (diamond), SWISSPROT v40 (filled circle), and to the total pooled prokaryote proteomes (bacterial + archaeal) (filled square). Discontinuous protein domains (and their homologs) are omitted when deriving these data.

as a gene.[12] The key novel points of our analysis are: (a) a two-step disabled-homology filter is applied to remove any potential pseudogene (ψg) sequences or disabled spurious ORFs, and (b) consideration of homologies between un-annotated ORFs in distinct genomes. The number of apparently conserved short ORF-like homologies is manageably low, corresponding to between about 0.5 and 4% of the size of the annotated proteomes, depending on the criteria used for selection.

uhORFs in bacteria and archaea

Firstly, the numbers of bacterial and archaeal uhORFs found in the polyORFome are overviewed. Secondly, we show that a major problem with such ORFs is their potential relationship with pseudogenes or spurious ORFs. Thirdly, as trends in sequence length are so critical in analyzing genome ORF annotations, we discuss uhORF tendencies for sequence length, comparing these to annotated ORF sequence lengths.

Numbers of uhORFs

uhORFs were derived as described in Figure 1 for 64 microbial genomes. The uhORFs for bacteria were tallied as shown in Table 1A. The homology H categories are explained in the legend to Figure 1. Few uhORFs were found in the polyORFome for the 53 bacterial species surveyed (Table 1A), with 614 uhORFs corresponding to just 0.5% of the total combined size of the annotated bacterial proteomes. If additional uhORFs that are only homologous to un-annotated ORFs in other bacterial genomes are allowed, the uhORF total increases to 921 (0.7% of total annotated bacterial proteomes), and to 2370 (1.8%) if ORFs only homologous to other ORFs in the same genome are included.

Comparable results are obtained for the archaeal genomes. From the polyORFomic sequence comparisons (to both annotated and un-annotated ORFs), we estimate that there are between 206 (0.9% of annotated archaeal proteomes) and 900 (3.8%).

As a specific example, we picked out the genome of the bacterium *Lactococcus lactis*.[2] This genome has large amounts of intergenic DNA in its current annotations, compared to other bacteria and archaea (15.3% for *L. lactis*).[13] We find values similar to those for the aggregates, with between 13 (0.6% of the size of the annotated proteome, 2224 proteins) and 60 (2.7%) uhORFs for *L. lactis*.

There is some sensitivity to the BLAST threshold used, in detecting these uhORFs; for example, for the homology class H(ek), the total tally reduces to 448 for *e*-value = 10^{-5}, and 429 for 10^{-6}; for H($b_{D,U}$), the values are 290 for 10^{-5}, 270 for 10^{-6}, etc. However, because of the manner of BLAST probability calculation,[14] such mild *e*-value

Table 1. PolyORFomic analysis of microbial genomes

Annotated proteomes		Un-annotated ORFs in the polyORFome				
Homology	Total	Homology	Total	After disabled-homology filter (1)	Predicted by GLIMMER (2)	(1) and (2)
A. *Bacteria*						
H(ek)	88,190	H(ek)	1123	488	535	190
Otherwise H(a_A)	4112	Otherwise H(a_A)	61	19	12	8
Otherwise H($b_{D,A}$)	23,080	Otherwise H($b_{D,A}$)	1286	819	232	99
~H(eka$_A$b$_{D,A}$)	16,087	Otherwise H(a_U)	9	8	1	1
Total	132,189	Otherwise H($b_{D,U}$)	9276	8663	378	**307**
		Otherwise H(b_S)	29705	28652	1607	1546 [**1449**][a]
		Total uhORFs = 488 + 19 + 99 + 8 = 614 (0.5% of size of annotated proteomes) + **307** + **1449** = **2370** (1.8% of size of annotated proteomes)				
B. *Archaea*						
H(ek)	12,055	H(ek)	46	28	19	9
Otherwise H(b_A)	2031	Otherwise H(b_A)	27	18	10	8
Otherwise H($a_{D,A}$)	4514	Otherwise H($a_{D,A}$)	896	578	220	155
~H(ekb$_A$a$_{D,A}$)	5363	Otherwise H(b_U)	9	5	4	1
Total	23,963	Otherwise H(a_U)	162	154	59	**55**
		otherwise H(a_S)	11,615	10,879	714	**639**
		Total extra short ORFs = 28 + 18 + 155 + 5 = 206 (0.9% of size of annotated proteomes) + **55** + **639** = **900** (3.8% of size of annotated proteomes)				

The first two columns of parts A and B of the Table show the breakdown of the annotated proteomes of bacteria and archaea, respectively, into homology classifications as described in the legend to Figure 1. The remaining columns show each of the homology classifications for the uhORFs, but broken down into Total uhORFs, uhORFs that remain after disabled-homology filter (i) (labelled (1)), uhORFs that are well-predicted by the program GLIMMER (labelled (2)), and the intersection of these sets ((1) and (2)). At the bottom of each section are tallied the uhORFs that give the lower bound estimates described for uhORF numbers described in the text (double underlined).

[a] The value in square-brackets gives the number of uhORFs when those violating disabled-homology filter criterion (ii) are removed.

threshold sensitivity is expected for short alignments, which require higher levels of sequence identity to maintain as high a BLAST probability as longer sequences.

Disabled-homology filtering

Disabled homology to a protein is characterized by disruptions from frameshifts and mid-sequence stop codons. On the basis of our previous analyses of putative pseudogenes,[10,11,15–17] we filtered the uhORF data for involvement in disabled protein-level homologies in two ways: (i) uhORFs were set aside that were part of a larger disabled homology to annotated proteins; (ii) uhORFs were set aside that had multiple disabled protein-level homologies elsewhere in the same genome, or in another sequenced strain of the genome, and no orthologs (Figure 1(b)). These procedures remove ORFs that are part of ψg's or are likely to be spurious. This is similar to procedures employed in the recent large-scale sequencing of *Saccharomyces* species;[18,19] however, unlike these annotation efforts, we have not used the disabled-homology filtering criterion (ii) to assess conservation between close species within the same genus, as it is unclear whether disabled homologies in this situation are due to the spurious nature of an ORF, or are genuine pseudogenes. Previously, we have found that disabled ORFs (dORFs) for both known and hypothetical proteins show similar chromosomal distributions, suggesting that a large proportion of these dORFs to hypothetical proteins are genuine pseudogenes.[9]

In the total combined bacterial genomes, using a disabled-homology based method,[10,11,15–17] 6064 putative pseudogenes were assigned, of which 1990 (30%) overlap or entail annotated ORFs.[20] Similarly, 831 pseudogenes were assigned in the archaeal genomes, with 328 (39%) of these interfering with annotated ORFs. Detection and analysis of these prokaryotic pseudogenes is described in detail elsewhere.[20] This data set of putative pseudogenes was used for criterion (i) above. For those putative pseudogenes that match a known structural protein domain (from the SCOP database[21]), we have calculated a measure of protein domain integrity (I_D) (shown in Figure 1(c)). I_D is the largest fraction of a protein domain match that is undisrupted by frameshifts and stop codons. From this graph, it is clear that the potential to code for a protein for this population of sequences is compromised severely, with 56% having $I_D < 0.4$, compared to only 18% for bacterial genes. This supports our strategy for assigning them as putative pseudogenes, and setting aside (uh)ORFs that overlap them.

Using criterion (i) of the disabled-homology filter, an additional 1039 uhORFs were detected for bacteria but disallowed by the disabled-homology filter for pseudogenes and spurious ORFs (i.e. approximately 31% of candidate uhORFs were set aside in this way). This is much larger than the proportion of existing bacterial ORF annotations (2.0%) that can be re-annotated as

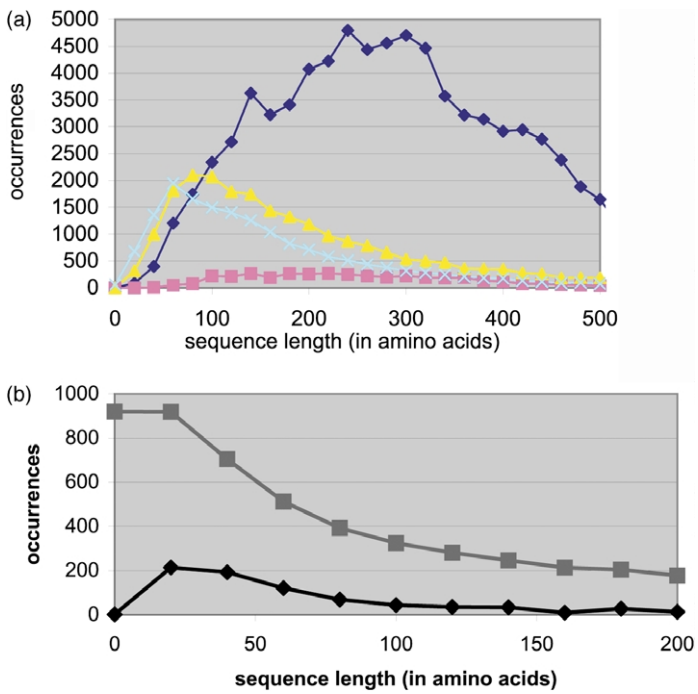

Figure 2. Length distributions of annotated ORFs and uhORFs for the 53 bacterial genomes in aggregate. (a) The plot shows the length distribution for all existing annotated bacterial ORFs with the H(ek) homology classification (dark blue line), all those otherwise homologous to archaeal proteins ($H(a_A)$, pink line), all those otherwise homologous to proteins from other bacteria ($H(b_{D,A})$, yellow line), then those otherwise not homologous to any protein from another genome ($\sim H(eka_Ab_{D,A})$, cyan line). All bins labelled x contain all ORFs between lengths x and $x + 20$. (b) The upper line (square) shows the length distribution of all bacterial uhORFs that are in the following categories in summation: $H(ek) + H(a_A) + H(b_{D,A}) + H(a_U) + H(b_{D,U})$. The lower line (diamond) shows the corresponding backwardly cumulative distribution. The small number of un-annotated ORFs of longer than 200 amino acid residues in this Figure and in Figure 3(a) are due to ambiguous endpoints in existing annotations or, in rare cases, simply due to missing blocks of annotation in Genbank/EMBL files. All bins labelled x contain all ORFs between lengths x and $x + 20$.

putative pseudogenes.[20] Interestingly, the proportion of potential uhORFs for archaea that are disallowed by the disabled-homology filter is much smaller (16%). This may arise because of distinct overall mechanisms and rates of gene disablement/decay/deletion for both kingdoms.[11,17] Criterion (ii) of the disabled-homology filter was applied to the bacterial genomes, and results in the removal of a small number of uhORFs (97/1549) in the H(b_S) homology category (Table 1A). Interestingly, only about 10% of the uhORFs set aside using criterion (i) would also set aside by criterion (ii) (data not shown); this may be due to the small size of the ORFs studied.

Length distributions for bacterial and archaeal annotated ORFs and uhORFs

What are the ORF length tendencies for the existing annotated ORFs and for the uhORFs? The existing ORF annotations demonstrate very different length distributions, depending on their protein-level sequence homologies. This is shown for bacteria, in aggregate (Figure 2). The distributions for ORFs that are homologous to known proteins or eukaryotic proteins (classified as H(ek)), or to archaeal proteins (H(a_A)), peak in the range 150–300 codons. However, distributions for ORFs that are homologous only to annotated proteins in other bacteria (denoted H($b_{D,A}$)), peak in the 60–100 codon range. This is the range of the thresholds that are used commonly in single-genome annotation for otherwise unsupported ORFs (Figure 2). This observation implies that many of the H($b_{D,A}$) ORFs are artefactual, as in similar less-detailed observations by others.[5] This tendency is even more noticeable for ORFs that have no homology or which are homologous only to other annotated ORFs in their own genome ($\sim$H(eka$_A$b$_{D,A}$)). These anomalous peaks are even more obvious for the existing archaeal genome ORF annotations in aggregate (Figure 3). Also, comparable trends are found for annotated ORFs in the eukaryote budding yeast (Figure 4).

It is likely that this homology-dependent behavior for the lengths of existing ORF annotations is artefactual.[5,6,8,11] Therefore, in our present analysis of uhORFs, we have conservatively considered only un-annotated homologies to ORFs from organisms in different kingdoms, or that are predicted as genes by the program GLIMMER (Table 1).[12] For bacteria, the uhORFs peak in the 30–50 codon range, whereas for archaea they tend to be longer (peaking in the 60–80 range) (Figures 2(b) and 3(b)). The numbers of uhORFs found represent a very small fraction of the number of possible ORFs in this length range. For example, for the bacterial genomes studied, the uhORFs in the range 60–80 codons length comprise $<0.4\%$ of all the possible ORFs. This shows how selective, for shorter ORF lengths, the application of sequence homology as an annotation principle is, in addition to its potency for existing ORF annotations (Figures 2–4).

Conclusions

There are manageably few undetected homologous short ORFs (uhORFs) in the sequenced prokaryotes, given the very large number of possible ORFs at such short ORF lengths. Depending on the type of sequence homology studied, we

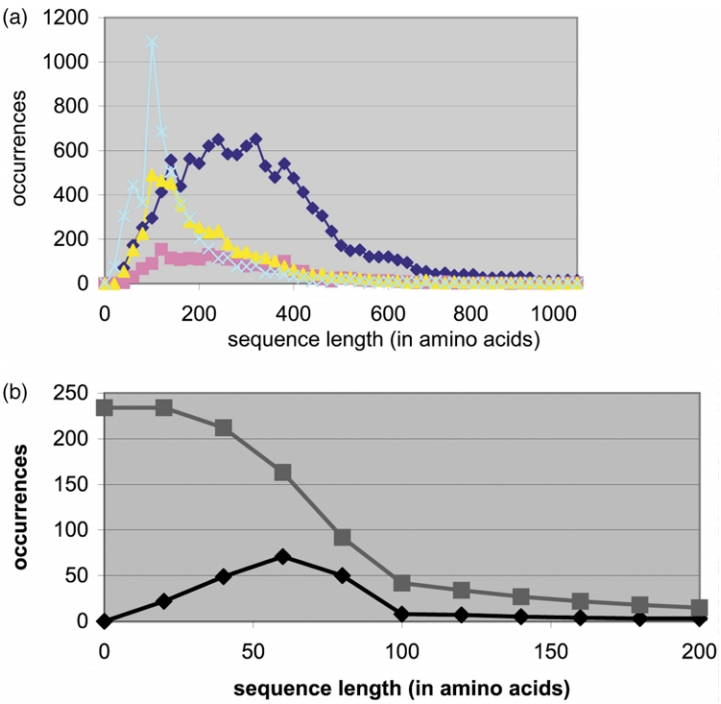

Figure 3. Length distributions of annotated ORFs and uhORFs for the 11 archaeal genomes in aggregate. (a) The plot shows the length distribution for all existing annotated archaeal ORFs with the H(ek) homology classification (dark blue line), all those otherwise homologous to bacterial proteins (H(b_A), pink line), all those otherwise homologous to proteins from other archaea (H($b_{D,A}$), yellow line), then those otherwise not homologous to any protein from another genome ($\sim$H(eka$_A$b$_{D,A}$), cyan line). All bins labelled x contain all ORFs between lengths x and $x+20$. (b) The lower line (diamond) shows the length distribution of all archaeal uhORFs that are in the following categories in summation: H(ek) + H(b_A) + H($a_{D,A}$) + H(b_U) + H($a_{D,U}$). The upper line (square) shows the corresponding backwardly cumulative distribution. All bins labelled x contain all ORFs between lengths x and $x+20$.

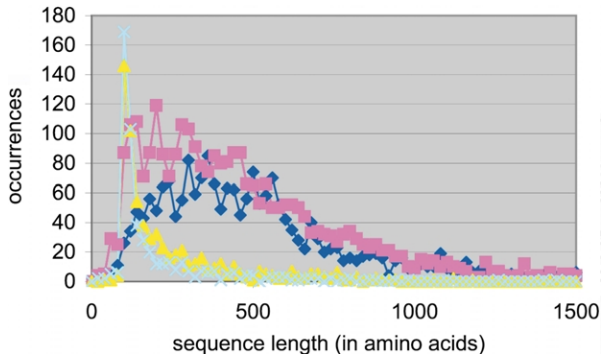

Figure 4. Length distributions of annotated ORFs for budding yeast. The plot shows the length distribution for all existing annotated budding yeast ORFs homologous to proteins from archaea or bacteria (dark blue line), all those otherwise homologous to proteins from other eukaryotes (pink line), all those otherwise homologous to annotated ORFs in budding yeast (yellow line), then those otherwise not homologous to any other annotated protein (i.e. singletons) (light blue line). All bins labelled x contain all ORFs between lengths x and $x + 20$.

estimate that they correspond to between about 0.5% and 4% of the size of the current annotated prokaryote proteomes. This is a scale of magnitude lower than the 10–30% of genes that are discarded in microbial genome annotations in another recent polygenomic analysis.[5] Our data thus represents the other half of the "equilibrium" in the microbial gene re-annotation process, and demonstrates the restrictive power of sequence homology at shorter ORF lengths. It is possible that some of the newly discovered short ORFs may have leader peptide functions, or are the truncated form of pseudogenes; this remains to be investigated. This study shows that the use of thresholds in annotation is unnecessary, and introduces the use of disabled-homology filtering for assignment of putative pseudogenes and disabled homologs of spurious ORFs.

The present analysis does not include in its estimates genes for which there are no detectable orthologs or paralogs. There may exist a distinct population of fast-evolving short ORFs, which would be difficult to detect by conventional sequence alignment procedures.[11] The existence of such ORFs in *Drosophila* species has been deduced from examination of randomly picked cDNAs.[22] Such proteins may be non-globular, or disordered in the native state; disordered proteins have been shown to have a tendency for apparently anomalous or positive selection patterns.[23] Families of divergent species-specific membrane proteins are observed.[17] Fast-evolving short ORFs are implied by a recent analysis of synonymous and non-synonymous codon substitution patterns in bacteria.[24] Most such short ORFs can be detected only from comparison to the complete sequences of closely related organisms; it was shown recently through large-scale sequencing of multiple *Saccharomyces* species[19] and *Saccharomycetes*,[25] that 1–2% of the *Saccharomyces cerevisiae* proteome could be detected only in this way. In tandem with such sequencing, more sophisticated analysis of patterns of divergence may be needed to distinguish lineage-specific families that have large numbers of genuine pseudogenes, from clusters of spurious ORFs.

References

1. Bernal, A., Ear, U. & Kyrpides, N. (2001). Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucl. Acids Res.* **29**, 126–127.
2. Bolotin, A., Wincker, P., Mauger, S., Jaillon, O., Malarme, K., Weissenbach, J. *et al.* (2001). The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. *lactis* IL1403. *Genome Res.* **11**, 731–753.
3. Kawarabayasi, Y., Hino, Y., Horikawa, H., Yamazaki, S., Haikawa, Y., Jin-no, K. *et al.* (1999). Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Res.* **6**, 83–101.
4. Kawarabayasi, Y., Hino, Y., Horikawa, H., Yamazaki, S., Haikawa, Y., Jin-no, K. *et al.* (1999). Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1 (supplement). *DNA Res.* **6**, 145–152.
5. Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D. & Krogh, A. (2001). On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet.* **17**, 425–428.
6. Das, S., Yu, L., Gaitatzes, C., Rogers, R., Freeman, J., Bienkowska, J. *et al.* (1997). Biology's new Rosetta stone. *Nature*, **385**, 29–30.
7. Merino, E., Balbas, P., Puente, J. L. & Bolivar, F. (1994). Antisense overlapping open reading frames in genes from bacteria to humans. *Nucl. Acids Res.* **22**, 1903–1908.
8. Mackiewicz, P., Kowalczuk, M., Gierlik, A., Dudek, M. R. & Cebrat, S. (1999). Origin and properties of non-coding ORFs in the yeast genome. *Nucl. Acids Res.* **27**, 3503–3509.
9. Kumar, A., Harrison, P. M., Cheung, K. H., Lan, N., Echols, N., Bertone, P. *et al.* (2002). An integrated approach for finding overlooked genes in yeast. *Nature Biotechnol.* **20**, 58–63.
10. Harrison, P., Kumar, A., Lan, N., Echols, N., Snyder, M. & Gerstein, M. (2002). A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. *J. Mol. Biol.* **316**, 409–419.
11. Harrison, P. M., Kumar, A., Lang, N., Snyder, M. & Gerstein, M. (2002). A question of size: the eukaryotic proteome and the problems in defining it. *Nucl. Acids Res.* **30**, 1083–1090.
12. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucl. Acids Res.* **27**, 4636–4641.
13. Mira, A., Ochman, H. & Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17**, 589–596.
14. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of

protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
15. Harrison, P. M., Hegyi, H., Balasubramanian, S., Luscombe, N. M., Bertone, P., Echols, N. *et al.* (2002). Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.* **12**, 272–280.
16. Zhang, Z., Harrison, P. & Gerstein, M. (2002). Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res.* In the press.
17. Harrison, P. M. & Gerstein, M. (2002). Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J. Mol. Biol.* **318**, 1155–1174.
18. Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J. *et al.* (2003). Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, **31**, 71–76.
19. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
20. Liu, Y., Harrison, P. & Gerstein, M. (2003). Polygenomic analysis of prokaryotes reveals widespread proteome decay, and degradation of putatively horizontally-transferred genes. *Genome Res.* In the press.
21. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
22. Schmid, K. J. & Tautz, D. (1997). A screen for fast evolving genes from *Drosophila*. *Proc. Natl Acad. Sci. USA*, **94**, 9746–9750.
23. Brown, C. J., Takayama, S., Campen, A. M., Vise, P., Marshall, T. W., Oldfield, C. J. *et al.* (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.* **55**, 104–110.
24. Ochman, H. (2002). Distinguishing the ORFs from the ELFs: short bacterial genes and the annotation of genomes. *Trends Genet.* **18**, 335–337.
25. Blandin, G., Durrens, P., Tekaia, F., Aigle, M., Bolotin-Fukuhara, M., Bon, E. *et al.* (2000). Genomic exploration of the hemiascomycetous yeasts: 4. The genome of *Saccharomyces cerevisiae* revisited. *FEBS Letters*, **487**, 31–36.
26. Bairoch, A. & Apweiler, R. (2000). The SWISSPROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45–48.
27. Ma, J., Campbell, A. & Karlin, S. (2002). Correlations between Shine–Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J Bacteriol.* **184**, 5733–5745.

Edited by F. E. Cohen

(Received 6 June 2003; accepted 10 September 2003)