Genome Analysis

Distinguishing the ORFs from the ELFs: short bacterial genes and the annotation of genomes

Howard Ochman

A substantial fraction of hypothetical open reading frames (ORFs) in completely sequenced bacterial genomes are short, suggesting that many are not genes but random stretches of DNA. Although it is not feasible to authenticate the coding capacity of all such regions experimentally, comparisons of ORFs in related genomes can expose those that encode functional proteins.

The density of genes, coupled with the presence of several gene-specific features and the lack of introns, makes the identification of coding sequences in bacterial genomes a relatively straightforward procedure, at least when compared with the task of gene recognition and annotation of genes in eukaryotes (e.g. Refs [1–5]). But, despite the numerous shared properties of bacterial genes, the thorough and robust annotation of complete bacterial genomes cannot rely solely upon a simple set of pre-established rules. Among the more troublesome tasks is the verification of very small ORFs, which, because of their problematic classification, have been derogatorily named ELFs, short for ‘evil little f...ellows’.

Bacterial genes are short, averaging only ~1000 nucleotides in each of the sequenced genomes. Many annotated ORFs are no more than putative reading frames, as recognized by in-frame start and stop codons separated by an ample distance (also referred to as coding sequences or CDS). Among the shortest ORFs within the *Escherichia coli* genome are well-recognized leader peptides of only 15–30 amino acids, and several ribosomal proteins are less than 100 amino acids in length. But, apart from these few cases, the preponderance of short annotated ORFs are hypothetical, of unknown function, and not likely to be genuine genes.

An insightful analysis on the extent to which microbial genome sequences are over-annotated and the contribution of short putative ORFs to this problem, was performed by Skovgaard *et al.* [6]. Based on the numbers and size distributions of ORFs with matches in the current databases, it was estimated that, for the majority of species, perhaps 10–30% of recognized ORFs do not actually encode proteins. Because the sheer number and pervasiveness of these ORFs preclude their experimental verification, these authors conclude that there is no clear resolution to the problem of the gene misidentification, and they offer several safeguards that would limit annotation to only the most trustworthy ORFs [6]. However, a simple procedure can be applied to determine which ORFs actually encode functional proteins, even in organisms that are not subject to experimental manipulation.

Because selective constraints differ between synonymous and nonsynonymous sites of genes, the reading frame of protein-coding regions can usually be deduced from alignments of homologous sequences in closely related organisms. In that the large proportion of mutations resulting in amino acid replacements is detrimental, divergence at synonymous sites (K_s) greatly exceeds that at nonsynonymous sites (K_a); and thus, for the vast majority of proteins, K_a/K_s ratios are significantly less than one. For regions that lack functional constraints, such as pseudogenes, K_a/K_s ratios are expected to approach unity; and for those few protein-coding regions undergoing adaptive evolution (positive or diversifying selection), K_a/K_s ratios can exceed one [7,8].

Nekrutenko *et al.* [9] have recently championed the use of K_a/K_s ratios as an aid in gene recognition and identification. They showed, through comparing a large set of homologous genes from human and mouse, and by computer simulations, that K_a/K_s ratios can reliably predict protein-coding regions, even those of less than 100 nucleotides. Of course, the K_a/K_s test requires sequences from two rather closely related organisms (i.e. from sequences that are sufficiently divergent but similar enough such that alignments are not confounded by multiple substitutions), and there are currently very few cases among eukaryotes where such data are available. However, the large number of completed bacterial genomes allows scrutiny of the annotated ORFs in several species pairs by this comparative method.

To evaluate the functional status (and authenticity) of recognized ORFs, K_a and K_s values were computed for homologs in pairs of bacterial species applying the following criteria:
(1) at least one of the genomes in a given species pair was fully annotated;
(2) a large proportion of genes had an unequivocal homolog in the two genomes;
(3) the average divergence was sufficient for accurate estimation of K_s and K_a, but not in saturation.

Genome sequences were queried with individual ORFs from a fully annotated reference genome using BLAST similarity searches [10], applying an initial cutoff of 70% sequence identity over at least 80% of ORF length. Homologs were aligned in the reading frame designated in the annotated genome, and K_s and K_a were calculated by the method of Li [11], as implemented in GCG [12]. Sequences and their annotations were obtained from NCBI (http://www.ncbi.nlm.nih.gov), and K_a/K_s data are available upon request from the author.

Twelve species pairs of appropriate divergence were recovered; however, the phylogenetic position of certain strains rendered numerous comparisons redundant. Components of sequence divergence plotted against ORF length for six species pairs are presented in Fig. 1. Note that in all comparisons, ORFs displaying higher K_a/K_s values are generally short, and that the variance in K_a/K_s ratios decreases with ORF length. The identical pattern was observed for the *Neisseria gonorrhea*–*Neisseria meningitidis* comparison, and for homologs in the 26695 and J99 strains of *Helicobacter pylori*, but low amounts of variation yielded less robust estimates of divergence in comparisons of short sequences.

The encouraging finding is that the vast majority of putative ORFs, even those that are short, are genuine protein-coding regions. The extent to which the K_a/K_s ratios test recognizes the noncoding ORFs was evaluated in various ways (Table 1). First, we considered the weakest portion of most

0168-9525/02/$ – see front matter PII: S0168-9525(02)02668-9

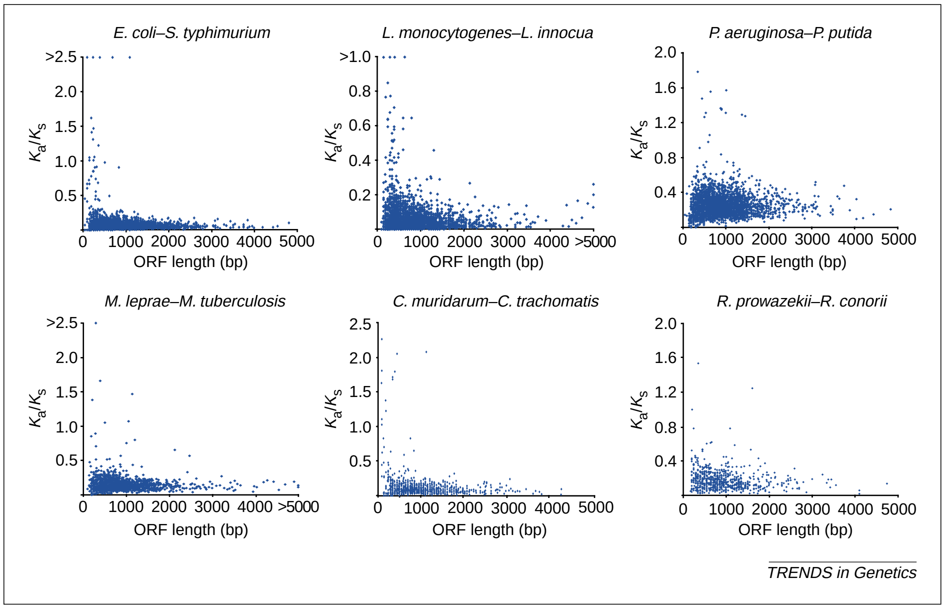

Fig. 1. Relationship between open reading frame (ORF) length and K_a/K_s ratios. Each panel shows the regression for all homologs for a given species pair.

databases (i.e. annotated coding regions under 300 nucleotides in length) and examined the proportion that had K_a/K_s ratios that were at least two standard deviations greater than the mean. In all species pairs, except *Chlamydia*, less than 10% of these short ORFs had anomalously high K_a/K_s ratios. Because the frequency of triplets specifying stop codons is lower in random sequences from GC-rich regions, the degree of over-annotation (the probability of finding long ORFs by chance) is greater in GC-rich genomes. Thus, the Pseudomonads, as anticipated, had the lowest proportion of small, high K_a/K_s ORFs, as well as the lowest proportion of homologs under 300 nucleotides. Although such findings are in line with the expectation that many high K_a/K_s ORFs do not actually encode proteins, there is no simple relationship between base composition and degree of over-annotation, because the quality of annotation varies among genomes.

Certain features of genes complicate the complete reliance on K_a/K_s ratios to detect ORFs. In many bacterial species, the use of synonymous codons is greatly reduced in highly expressed genes, which can experience selection on synonymous sites comparable to that at nonsynonymous sites [13]. Such adaptive codon usage bias will elevate K_a/K_s values (by suppressing K_s); however, comparisons of K_s values among homologs for a given species pair readily identifies such genes. Also, because of the limited number of sites, particularly synonymous sites (which occur at about a third the frequency of nonsynonymous sites), there are large standard errors associated with estimates of K_a and K_s in very short genes. Gene position, strand location and level of transcription can all affect mutation frequencies (and hence, the neutral rate of evolution, as reflected at synonymous sites) [14–17], but such factors are not likely to be restricted to ORFs of a particular length.

The predominance of high K_a/K_s ratios among smaller ORFs in all species pairs could reflect that:

(1) more short ORFs undergo positive selection;

(2) the number of nonsynonymous sites under positive selection is approximately constant over proteins, such that K_a/K_s ratios decrease with ORF length;

(3) codon usage bias is stronger in shorter ORFs, thus elevating their K_a/K_s values;

(4) many short, annotated ORFs are not genuine protein-coding regions.

Although there is general support for the last explanation, one advantage of the K_a/K_s test is that it identifies the specific ORFs exhibiting anomalous patterns of sequence divergence, and the analysis of individual ORFs will often give clues to their authenticity.

Drawing example from the *E. coli–Salmonella typhimurium* genome comparison, for which there exists the most extensive functional and experimental information, ten of the 14 ORFs with K_a/K_s ratios >1.0 are listed as 'hypothetical', and 90% of these are less than 300 nucleotides. Furthermore, none of these 14 ORFs has a K_s value approaching that anticipated for *E. coli–Salmonella* homologs, and the only two genes showing an extreme bias towards codons preferred by *E. coli* genes are, *rpmD* and *prfB*, encoding ribosomal protein L30 and peptide chain release factor RF-2, respectively. Therefore, even in a well-annotated genome such as *E. coli*, most of the ORFs displaying atypical patterns of sequence evolution (as evident by their K_a/K_s ratios) are short, exhibit codon irregularities, and are of unknown function – all indicators that they are not truly protein-coding regions.

Conclusions

The comparative approach, as exemplified by the K_a/K_s test, provides a rapid means

Table 1. Features of annotated ORFs in bacterial genomes

Species pair	%GC[a]	Avg K_s	*n*	K_a/K_s (sd)[b]	% Atypical[c] (avg length)	% Actual <300bp[d]
Rickettsia prowazekii–Rickettsia conorii	31	0.34	701	0.20 (0.14)	3.3% (687 nt)	91% (56/62)
Listeria innocua–Listeria monocytogenes	38	0.62	2529	0.06 (0.09)	2.8% (509 nt)	92% (244/265)
Chlamydia muridarum–Chlamydia trachomatis	41	0.80	823	0.13 (0.21)	2.3% (331 nt)	86% (68/79)
E. coli K12–*Salmonella typhimurium*	53	0.98	2971	0.09 (0.16)	1.2% (303 nt)	92% (248/273)
Mycobacterium leprae–Mycobacterium tuberculosis	63	0.74	1289	0.16 (0.13)	1.6% (757 nt)	92% (95/103)
Pseudomonas aeruginosa–Pseudomonas putida	65	0.78	2613	0.26 (0.14)	2.3% (793 nt)	98% (168/171)

[a] Base composition of all homologs considered, averaged for both species.

[b] Abbreviation: sd, standard deviation.

[c] ORF termed 'atypical' if K_a/K_s ratio is >2 sd above the mean for a given species pair.

[d] ORF termed 'actual' if K_a/K_s ratio is <2 sd above the mean for a given species pair. All actual ORFs had K_a/K_s ratios significantly less than one. Numbers in parentheses denote the actual values from which percentages were derived.

to establish whether an ORF is evolving in a manner typical of a protein-coding region and to evaluate the accuracy of current databases. Although limited to those portions of the genome present in other related organisms, the expansion in the number and phylogenetic distribution of sequenced organisms will allow such methods to have an increasing role in gene recognition and confirmation. It should be noted that, from a functional perspective, the small fraction of annotated ORFs with high K_a/K_s ratios still warrants additional scrutiny. Although a K_a/K_s ratio near unity can denote noncoding sequences or pseudogenes, it is also characteristic of regulatory regions, regions encoding RNAs, and coding regions under extreme codon usage bias. Examination of the relative amount of sequence divergence, as well as features of the sequence itself, permit discrimination among these possibilities and can engender new insights into the role of uncharacterized sequences.

Acknowledgements

I thank Susan Miller and Alex Mira for their expertise in writing and executing computer programs, Nancy Moran and Jeff Lawrence for their helpful comments, and the NIH and DOE for their generous support.

References

1 Kyripides, N.C. and Ouzounis, C.A. (1999) Whole-genome sequence annotation: 'Going wrong with confidence'. *Mol. Microbiol.* 32, 886–887

2 Lewis, S. *et al.* (2000) Annotating eukaryote genomes. *Curr. Opin. Struct. Biol.* 10, 349–354

3 Stormo, G.D. (2000) Gene-finding approaches for eukaryotes. *Genome Res.* 10, 394–397

4 Rogic, S. *et al.* (2001) Evaluation of gene-finding programs on mammalian sequences. *Genome Res.* 11, 817–832

5 Kumar, A. *et al.* (2002) An integrated approach for finding overlooked genes in yeast. *Nat. Biotechnol.* 20, 58–63

6 Skovgaard, M. *et al.* (2001) On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet.* 17, 425–428

7 Li, W-H. (1997) *Molecular Evolution*, Sinauer Associates

8 Nei, M. and Kumar, S. (2000) *Molecular Evolution and Phylogenetics*, Oxford University Press

9 Nekrutenko, A. *et al.* (2002) The K_a/K_s ratio test for assessing the protein-coding potential: an empirical and simulation study. *Genome Res.* 12, 198–202

10 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410

11 Li, W-H. (1993) Unbiased estimation of the rates of synonymous and non-synonymous substitution. *J. Mol. Evol.* 36, 96–99

12 Devereux, J.P. *et al.* (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* 12, 387–395

13 Sharp, P.M. and Li, W-H. (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol. Biol. Evol.* 4, 222–230

14 Eyre-Walker, A. and Bulmer, M.M. (1995) Synonymous substitution rates in enterobacteria. *Genetics* 140, 1407–1412

15 Lobry, J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* 13, 660–665

16 Francino, M.P. and Ochman, H. (2001) Deamination as the basis of mutational strand asymmetry in *Escherichia coli*. *Mol. Biol. Evol.* 18, 1147–1150

17 Sharp, P.M. *et al.* (1989) Chromosomal location and evolutionary rate variation in enterobacterial genes. *Science* 246, 808–810

Howard Ochman
Dept of Biochemistry and Molecular
Biophysics, 233 Life Sciences South,
University of Arizona, Tucson,
AZ 85721, USA.
e-mail: hochman@email.arizona.edu

Human SNP variability and mutation rate are higher in regions of high recombination

Martin J. Lercher and Laurence D. Hurst

Understanding the co-variation of nucleotide diversity and local recombination rates is important both for the mapping of disease-associated loci and in understanding the causes of sequence evolution. It is known that single nucleotide polymorphisms (SNPs) around protein coding genes show higher diversity in regions of high recombination. Here, we find that this correlation holds for SNPs across the entire human genome, the great majority of which are not near exons or control elements. Contrasting with results from coding regions, we provide evidence that the higher nucleotide diversity in regions of high recombination is most likely due, at least in part, to a higher mutation rate. One possible explanation for this is that recombination is mutagenic.

Previous reports, from samples [1–4] of single nucleotide polymorphisms (SNPs) near protein coding regions of human genes, find a strong positive correlation between nucleotide diversity (π) and the local recombination rate. The same is reported in fruit flies [5,6]. But is this true for SNPs sampled throughout the human genome (i.e. not in close proximity to exons or control regions)? To address this, we obtained SNP nucleotide diversity data from a recent analysis of 4.5 million high-quality reads [7]. We took sex-averaged recombination rate for contigs representing 60% of the human genome from a comparison of human genetic and sequence-based maps [8]. To be compatible with the SNP data, we aligned contigs with the 5 September 2000 build of the human genome (http://genome.ucsc.edu). As it is a priori unclear on which length scale correlations exist, we averaged nucleotide diversity and recombination rate over bins of varying sizes, ranging from 2 Mb to 30 Mb. We find a positive correlation between nucleotide diversity and recombination rate for all bin sizes (Table 1; Fig. 1 shows data for 20-Mb bins). The recombination map used in this comparison reflects recent recombination in humans, and might only apply to a limited fraction of the evolutionary time over which polymorphisms were shaped. However, this will only introduce random error, and the 'true' correlation could possibly be stronger than suggested by Table 1.

Why might there be higher nucleotide diversity in regions of high recombination? This correlation is mostly attributed to either background selection [9–11] or hitchhiking with positively selected alleles [12]. Both effects reduce the effective population size of the local genomic region, with the range of this effect determined by the recombination rate. Such effects are expected to dominate in the vicinity of regions in which point mutations are under selection, such as near exons and their associated control elements. However, the polymorphisms studied here are not biased to such regions, and more than 95% will be located in noncoding DNA (intergene spacer or the putative massive introns [13]). Here, then, we explore a

http://tig.trends.com 0168-9525/02/$ – see front matter © 2002 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(02)02669-0