Structure, Vol. 11, 7–9, January, 2003, ©2003 Elsevier Science Ltd. All rights reserved. PII S0969-2126(02)00938-3

Twenty Thousand ORFan Microbial Protein Families for the Biologist?

Minireview

Naomi Siew[1,2] and Daniel Fischer[2,*]
[1]Department of Chemistry
[2]Bioinformatics Group
Department of Computer Science
Ben Gurion University
Beer-Sheva 84105
Israel

The genomes of most newly sequenced organisms contain a significant fraction of ORFs (open reading frames) that match no other sequence in the databases. We refer to these singleton ORFs as sequence ORFans. Because little can be learned about ORFans by homology, the origin and functions of ORFans remain a mystery. However, in this era of full genome sequencing, it seems that ORFans have been underemphasized. In this minireview, we draw attention to the increasing number of ORFans and to the consequences of this growth to biological research in the postgenomic era.

The genomes of each newly sequenced organism contain an enormous wealth of genomic information. At the same time, each new genome also increases the number of unknowns, as most of them contain 20%–30% ORFs that match no other sequence in the databases [1–5]. We refer to these as sequence ORFans [1, 6].

A few years ago, when only a handful of complete genome sequences were available, a number of possible explanations for the abundance of ORFans were suggested. One explanation was that the relatively high proportion of ORFans may be due to an artifact of sparse sampling of the sequence space, and that with the availability of more genomes, most ORFans would disappear. Another explanation was that many of the ORFans observed at that time may not correspond to expressed proteins, but rather that they correspond to errors or to incorrectly annotated genes [1, 7–9].

Today, with the availability of a few dozen complete genomes, it appears that most longer ORFans do correspond to expressed proteins ([7, 10, 11] and experimental data from the Halobacterium NRC-1 [12] ORFan structural genomics project; B. Shaanan, J. Eichler, and D.F., unpublished results). In addition, their abundance seems not to be a mere artifact of sparse sampling because most new genomes contain ORFans in similar proportions, and the number of ORFans accumulating in the databases continues to grow. Thus, ORFans appear to be an intrinsic phenomenon of genetic material. Because the origin and functions of ORFans are yet to be explained, their presence has been referred to as a mystery [8].

If proteins in different organisms have descended from common ancestral proteins by duplication and adaptive variation, why is it that so many today show no similarity to each other [13]? Why is it that we do not find today any of the necessary "intermediate sequences" that must have given rise to these ORFans? Do most ORFans correspond to rapidly diverging proteins [9, 14]? If so, how rapidly do they diverge, and what are the forces involved in their rapid evolution? Is their rate of change constant or did the rapid changes occur only at specific times? Do these rapidly evolving ORFans correspond to nonessential proteins or to species determinants?

Regardless of their origin, ORFans may be of two types. Some ORFans may correspond to newly evolved proteins (through a yet unknown mechanism; e.g., see [15, 16]) or to unique descendants of ancient proteins, with unique functions and three-dimensional (3D) structures not currently observed in other families [1, 17, 18]. Alternatively, ORFans may correspond to highly divergent members of known protein families, but with functions and/or 3D structures similar to proteins already known [1, 9].

In any case, standard computational methods, which infer function or 3D structure of a protein on the basis of its sequence similarity to a known family, are not suitable for the study of ORFans. Thus, the function of each ORFan will need to be determined by genetic or biochemical approaches [8, 19, 20]. More sensitive computational methods, such as fold recognition [21] or sequence-to-profile comparisons, may succeed in assigning some ORFans to known families, and thus, through their application, some understanding about their roles and functions may be gained. However, even if these methods succeeded in assigning many ORFans to known families, the questions about their rapid divergence would still remain.

Furthermore, even if we accept that most ORFans are distant relatives of known protein families, the fact that they are so divergent has important implications in biology, because the number of different sequence families (ORFans plus non-ORFans) will turn out to be higher than previous estimates [18, 22]. One particularly important implication is within structural biology. Because accurate and reliable 3D models can only seldom be obtained with a lack of significant sequence similarity [23], ORFans lie outside the "homology modeling distance" from known 3D structures. Thus, each ORFan forms a separate single-membered sequence family. A consequence is that a complete coverage of structural genomics projects [17] will require structure determinations for each ORFan [18, 24].

Here we show that most newly sequenced genomes add new ORFans, resulting in an increasing number of sequence families awaiting interpretation; we do not attempt to explain the origin or functions of ORFans, but merely focus on describing the dynamics and the magnitude of the ORFan puzzle, based on data from the first 60 fully sequenced microbial genomes.

We have analyzed the ORFan content in the first 60 fully sequenced microbial genomes (strains not included). The genomes were added to our database of

*Correspondence: dfischer@cs.bgu.ac.il

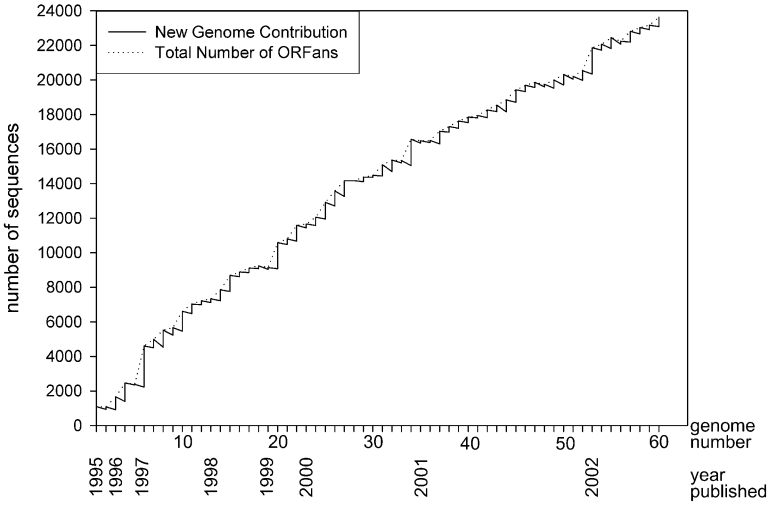

Figure 1. The Total Number of ORFans in Microbial Fully Sequenced Genomes Continues to Grow

Each new genome contains a number of sequences that match previous ORFans (nearly horizontal lines). These matches slightly reduce the total number of ORFans. At the same time, each new genome adds a larger number of new ORFans (vertical lines), and thus the total number of ORFans keeps growing. After 60 genomes, there are 168,248 ORFs, out of which 23,634 are ORFans. This trend is likely to continue for the next genomes, and thus, it seems unlikely that the total number of ORFans will soon drop significantly. But even after a dense sampling of genome space, when only few singleton ORFans will remain, the number of sequence families lying beyond the homology modeling distance from proteins of known structure is likely to be high.

The *x* axis corresponds to the 60 genomes considered, in chronological order of their publication, and according to the year of their publication: 1995: *H. influenzae*, *M. genitalium*; 1996: *M. jannaschii*, *Synechocystis sp.*, *M. pneumoniae*; 1997: *S. cerevisiae*, *H. pylori*, *E. coli* K-12, *M. thermoautotriphicum*, *B. subtilis*, *A. fulgidus*, *B. burgdorferi*; 1998: *A. aeolicus*, *P. horikoshii*, *M. tuberculosis*, *T. pallidum*, *C. trachomatis*, *R. prowazekii*; 1999: *C. pneumoniae*, *A. pernix*, *T. maritima*, *D. radiodurans*; 2000: *C. jejuni*, *N. meningitidis*, *X. fastidiosa*, *V. cholerae*, *P. aeruginosa*, *Buchnera sp.*, *T. acidophilum*, *U. urealyticum*, *Halobacterium sp.*, *B. halodurans*, *T. volcanium*, *M. loti*; 2001: *M. leprae*, *P. multocida*, *C. crescentus*, *S. pyogenes*, *S. aureus*, *L. lactis*, *M. pulmonis*, *S. solfataricus*, *S. pneumoniae*, *S. Meliloti*, *C. acetobutylicum*, *R. conorii*, *L. monocytogenes*, *L. innocua*, *Y. pestis*, *S. typhy*, *S. typhimurium*, *A. tumefaciens*; 2002: *S. coelicolor*, *T. tengcongensis*, *X. axonopodis pv. citri*, *X. campestris pv. campestis*, *C. tepidum*, *O. iheyensis*, *S. agalactiae*, *B. suis*.

Notice that our ORFan counts are conservative because here we did not consider non-ORFans (1) with only partial matches, that is, ORFan domains within the non-ORFans (ORFan modules), (2) with matches within the same organism only, that is, "paralogous-ORFans," or (3) with matches only to closely related organisms, that is, "orthologous-ORFans." These three additional types of ORFans emphasize the ORFan puzzle and are the subject of our ongoing research.

fully sequenced genomes in chronological order of publication (see http://www.tigr.org/tdb/mdb/mdbcomplete.html), and matches to other ORFs in the database were searched using gapped BLAST [25]. We define a match if BLAST detects a hit with an e-value of at most 10^{-3} (for alignments shorter than 80 residues, a stricter value of 10^{-5} is used). ORFs with no matches are labeled as ORFans. If a sequence once classified as an ORFan matches any sequence of a newer genome, then its status is changed to non-ORFan.

Figure 1 shows that each new genome has two effects on the total number of ORFans. On the one hand, the new genome contains ORFs that match older ORFans, converting the latter into non-ORFans. On the other hand, the new genome contains new ORFans, creating new single-membered sequence families. Thus, the change in the total number of ORFans after each genome is added to our database depends on its evolutionary distance from previously sequenced genomes. A relatively sharp drop in the number of ORFans occurs when a genome of a closely related organism is added to the database. This is so because many of the new organism's ORFs match old ORFans of the previously sequenced close relative. The fact that older ORFans find homologs in the genome of a close relative suggests that these ORFans (and their homologs) do correspond to functional, expressed proteins. Nevertheless, the number of new ORFans that each new genome contributes is usually larger than the number of older ORFans that become non-ORFans. Consequently, the total number of ORFans is growing.

After 60 genomes, our database contains 168,248 ORFs, of which 23,634 (14%) are ORFans. When using different thresholds to define matches (e-values of 0.1 to 10^{-10}), the shape of the curve of Figure 1 does not change significantly and the total number of ORFans varies within 30% of the figure reported here.

Figure 1 suggests that the number of ORFans will not drop significantly in the near future. Extrapolation of the data of Figure 1 is inaccurate and depends, among other assumptions, on the diversity of newly sequenced organisms; nonetheless, it can give a rough estimate for the general trends expected for the next few dozen genomes. A quadratic fit (correlation 0.998) suggests that at the one-hundredth genome, there will be over 25,000 ORFans. Even well beyond this point, it is likely that the genome of any new organism, sufficiently divergent from previously sequenced genomes, will contribute a significant number of ORFans. For example, 3,208 out of the 5,268 ORFs (60%) of the recently determined sequence of the eukaryote *P. falciparum* [26] correspond to ORFans. Undoubtedly, the number of ORFans may diminish when a denser sampling of organisms is achieved, but even then it is likely that we will be left with a considerable number of sequence families containing only proteins of closely related organisms (which we refer to as orthologous ORFans; see the legend of Figure 1). Furthermore, the number of current ORFans suggests that even after dense sampling, the total number of sequence families will remain high, and the puzzles of the origin and the functions of these ORFan families will remain.

There are two noteworthy observations about our ORFan database. The first is that over half of the ORFans are shorter than 150 residues. Possible explanations for this bias could be that some of the shorter ORFans may

not correspond to expressed proteins [27], or that their abundance is a result of a limitation of computational sequence comparison; it is harder for current tools to detect sequence similarity for short sequences. Interestingly, the dynamics of long and short ORFans are different. Separate quadratic fits for long and short ORFans suggest that while the number of short ORFans may continue to grow up to the two-hundredth genome, the current number of long ORFans may be close to its maximum.

The second observation is that our counts of ORFans are based only on the first 60 microbial, complete genomes. If the full sequence databases were considered, then some of the ORFans would become non-ORFans, and new ORFans would appear. For example, our database contains 2,108 ORFans from *S. cerevisiae* (SC in Figure 1). These can be considered as eukarya-specific sequences, because *S. cerevisiae* is the only eukaryote in our microbial database. When searching the fly and worm genomes, we found matches for approximately one quarter of the yeast ORFans. The majority of the remaining ORFans may correspond to yeast-specific sequences [8]. Thus, if our database contained the fly and worm genomes, then our counts of *S. cerevisiae* ORFans would be lower, but obviously, many new fly and worm ORFans would be added. In conclusion, we believe that the growing number of ORFans observed in our database is not merely a consequence of considering only microbial, fully sequenced genomes.

We conclude that the increasing number of ORFans suggests that our knowledge of nature's sequence diversity continues to grow, that ORFans may entail an intrinsic phenomenon in evolution, and that a global view of the protein world needs to consider the ORFan sequence families in addition to the large sequence families containing proteins conserved in numerous organisms.

Acknowledgments

This work is related in part to an ORFan joint research project with David Eisenberg; we thank him for discussions and inspiration. This joint research project was partially supported by grant number 1998422 from the United States-Israel Binational Science Foundation (BSF), Jerusalem, Israel. Thanks to Samir Genaim for his help with the figure. N.S. is partially supported by a grant from the Ministry of Science, Israel, and by the Kreitman Foundation Fellowship.

Selected Reading

1. Fischer, D., and Eisenberg, D. (1999). Bioinformatics *15*, 759–762.
2. Fraser, C.M., Eisen, J.A., and Salzberg, S.L. (2000). Nature *406*, 799–803.
3. Bloom, B.R. (2000). Nature *406*, 760–761.
4. Wren, B.W. (2000). Nat. Rev. Genet. *1*, 30–39.
5. Boucher, Y., Camilla, N.L., and Doolittle, W.F. (2001). Curr. Opin. Microbiol. *4*, 285–289.
6. Fischer, D. (1999). Protein Eng. *12*, 101–102.
7. Dujon, B. (1994). Nature *369*, 371–377.
8. Dujon, B. (1996). Trends Genet. *12*, 263–270.
9. Wood, V., Rutherford, K.M., Ivens, A., Rajandream, M.-A., and Barrell, B. (2001). Compar. Funct. Genom. *2*, 143–154.
10. Hutchison, C.A., III, Peterson, S.N., Gill, S.R., Cline, R.T., White, O., Fraser, C.M., Smith, H.O., and Venter, J.C. (1999). Science *286*, 2165–2169.
11. Monchois, V., Abergel, C., Sturgis, J., Jeudy, S., and Claverie, J.M. (2001). J. Biol. Chem. *276*, 18437–18441.
12. Ng, W.V., Kennedy, S.P., Mahairas, G.G., Berquist, B., Pan, M., Shukla, H.D., Lasky, S.R., Baliga, N.S., Thorsson, V., Sbrogna, J., et al. (2000). Proc. Natl. Acad. Sci. USA *97*, 12176–12181.
13. Doolittle, R.F. (1997). Nature *388*, 515–516.
14. Schmid, K.J., and Aquadro, C.F. (2001). Genetics *159*, 589–598.
15. Pellegrini, M., and Yeates, T.O. (1999). Proteins *37*, 278–283.
16. Mackiewicz, P., Kowalczuk, M., Gierlik, A., Dudek, M.R., and Cebrat, S. (1999). Nucleic Acids Res. *27*, 3503–3509.
17. Vitkup, D., Melamud, E., Moult, J., and Sander, C. (2001). Nat. Struct. Biol. *8*, 559–566.
18. Coulson, A.F.W., and Moult, J. (2002). Proteins *46*, 61–71.
19. Oliver, S.G. (1996). Nature *379*, 597–600.
20. Alimi, J.-P., Poirot, O., Lopez, F., and Claverie, J.-M. (2000). Genome Res. *10*, 959–966.
21. Fischer, D., Rice, D., Bowie, J., and Eisenberg, D. (1996). FASEB J. *10*, 126–136.
22. Chothia, C. (1992). Nature *357*, 543–544.
23. Tramontano, A., Leplae, R., and Morea, V. (2001). Proteins Suppl. *5*, 22–38.
24. Frishman, D. (2002). Protein Eng. *15*, 169–183.
25. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Nucleic Acids Res. *25*, 3389–3402.
26. Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., et al. (2002). Nature *419*, 498–511.
27. Skovgaard, M., Jensen, L.J., Brunak, S., Ussery, D., and Krogh, A. (2001). Trends Genet. *17*, 425–428.