PROTEINS: Structure, Function, and Genetics 53:241–251 (2003)

Analysis of Singleton ORFans in Fully Sequenced Microbial Genomes

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

ABSTRACT **Singleton sequence ORFans are orphan ORFs (open reading frames) that have no detectable sequence similarity to any other sequence in the databases. ORFans are of particular interest not only as evolutionary puzzles but also because we can learn little about them using bioinformatics tools. Here, we present a first systematic analysis of singleton ORFans in the first 60 fully sequenced microbial genomes. We show that although ORFans have been underemphasized, the number of ORFans is steadily growing, currently accounting for 23,634 sequences. At the same time, the percentage of ORFans as a fraction of all sequences is slowly diminishing, and is currently about 14%. Short ORFans comprise about 61% of all ORFans. The abundance of short ORFans may be due to a yet unexplained artifact. The data also suggest that the number of longer ORFans may soon diminish as more genomes of closely related organisms become available. To better address the questions about the functions and origins of ORFans, we propose to focus further studies on the longer ORFans, with emphasis on three new types of ORFans: ORFan modules, paralogous ORFans, and orthologous ORFans. We conclude that the large number of ORFans reflects an intrinsic property of the genetic material not yet fully understood. Further computational and experimental studies aimed at understanding Nature's protein diversity should also include ORFans. Proteins 2003;53:241–251.**

Key words: ORFans; complete genomes; evolution; singletons; microbial diversity

INTRODUCTION

Since the sequencing in 1995 of the genome of the first free-living organism, that of *Haemophilus influenzae,*[1] the genomes of over a few dozen organisms have been sequenced, and dozens more are under way. This wealth of continuously growing sequence data contains a large number of protein sequences awaiting interpretation that, once deciphered, will add to a whole new understanding of Nature.

The availability of complete genome sequences of modern organisms has clearly revealed that the genetic material is mainly the result of the basic evolutionary process of descent with modification. Most of the open reading frames (ORFs) in a newly sequenced organism encode proteins belonging to homologous families that are more or less conserved in a number of organisms. Some of these families contain ORFs from most of the known genomes and usually correspond to widely conserved functions essential for life. Other families contain ORFs from organisms belonging to one kingdom only, thus corresponding to functions specific to that kingdom. In addition to these relatively conserved families, the currently fully sequenced genomes also contain a variety of families with decreasing levels of conservation. At the lower end, we observe a non-negligible number of families that contain ORFs of only a few (generally closely related) organisms, or of a single organism only. Surprisingly, a large number of genome sequences belong to single-member families. We refer to such sequences as orphan ORFs or ORFans for short.[2–4] ORFans account for 25–30% of the ORFs of each newly sequenced genome,[5,6] and their percentage can even be as high as 60%,[7] suggesting that sequence diversity in Nature may be greater than previously expected. Because little can be learned about ORFans via homology, only experimental characterization can help elucidate their functions and origin.[8–13] Thus, each ORFan represents a mystery awaiting interpretation.[13,14]

ORFans may correspond to highly divergent sequences that actually belong to known families (but are beyond recognition capabilities of current tools),[2] or to sequences that correspond to new, unique, single-member families.[2,15] Because there is no obvious evolutionary mechanism to account for the origin of single-member families, one might accept the explanation of their origin as extreme divergence. However, even if all ORFans correspond to highly divergent members of known families, a number of puzzling questions arise. For example, how have their sequences diverged to such an extent that no similar sequences are detected today?[16] If evolution works through descent with modification, then why is it that no similar sequences are found in other organisms? Why is it that we

Grant sponsor: United States–Israel Bination Science Foundation (BSF), Jerusalem, Israel; Grant number 1998422.

Grant sponsor: N.S. is supported in part by grants from the Ministry of Science, Israel, and from the Kreitman Foundation Fellowship.

*Correspondence to: Daniel Fischer, Bioinformatics Group, Department of Computer Science, Ben Gurion University, Beer-Sheva 84105, Israel. E-mail: dfischer@cs.bgu.ac.il

Received 31 October 2002; Accepted 3 January 2003

TABLE I. Dynamics in the Percentage of ORFans

			ORFans (%)					
			Total[a]		Short[b]		long[c]	
Genome	Abbr.[d]	ORFs[e]	Initial[f]	Final[g]	Initial[f]	Final[g]	Initial[f]	Final[g]
1. *H. influenzae*	HI	1707	64.0	5.2	87.6	19.4	57.8	1.4
2. *M. genitalium*	MG	479	32.4	0.8	50.6	4.9	28.6	0.0
3. *M. jannaschii*	MJ	1773	42.5	14.3	67.9	31.5	34.1	8.7
4. *Synechocystis sp.*	SN	3167	34.0	20.5	58.1	47.3	27.7	13.4
5. *M. pneumoniae*	MP	677	5.9	3.8	14.5	12.1	4.0	2.0
6. *S. cerevisiae*	SC	6307	37.7	33.4	72.3	71.1	31.1	26.2
7. *H. pylori*	HP	1575	31.5	17.0	62.8	45.6	22.8	8.9
8. *E. coli* K-12	EC	4289	22.8	5.5	52.8	17.7	15.2	2.4
9. *M. thermoautotrophicum*	MT	1869	23.0	16.5	47.6	34.0	14.2	8.5
10. *B. subtilis*	BS	4100	28.0	12.4	63.5	37.6	16.3	4.1
11. *A. fulgidus*	AF	2407	22.8	16.7	46.7	39.0	14.2	8.6
12. *B. burgdorferi*	BB	850	26.9	18.0	57.1	45.3	19.9	11.6
13. *A. aeolicus*	AQ	1522	14.5	9.7	30.1	21.8	11.9	7.7
14. *P. horikoshii*	PH	2064	30.6	26.7	67.1	62.8	16.3	12.7
15. *M. tuberculosis*	MR	3918	23.7	11.1	48.7	30.6	17.8	6.4
16. *T. pallidum*	TP	1031	25.7	22.9	55.6	53.0	18.6	15.7
17. *C. trachomatis*	CT	894	28.5	6.4	53.4	23.3	23.0	2.6
18. *R. prowazekii*	RP	834	18.2	2.4	40.5	8.5	13.2	1.0
19. *C. pneumoniae*	CP	1052	10.8	10.5	31.2	30.2	6.1	5.9
20. *A. pernix*	AP	2694	56.0	52.0	84.7	81.7	36.9	32.2
21. *T. maritima*	TM	1846	16.9	12.7	42.8	35.8	10.9	7.3
22. *D. radiodurans*	DR	3116	29.1	22.6	63.4	55.7	21.3	15.0
23. *C. jejuni*	CJ	1634	14.2	11.4	39.1	35.2	8.3	5.8
24. *N. meningitidis*	NM	2025	23.6	18.6	53.6	47.6	11.0	6.4
25. *X. fastidiosa*	XF	2831	34.0	25.6	66.1	57.9	13.4	4.8
26. *V. cholerae*	VB	3828	22.7	18.1	53.4	48.9	11.8	7.3
27. *P. aeruginosa*	PA	5565	16.4	11.2	45.8	35.6	10.7	6.5
28. *Buchnera sp.*	BC	574	1.6	1.1	7.3	4.6	0.2	0.2
29. *T. acidophilum*	TA	1478	17.5	6.0	42.4	17.7	11.6	3.3
30. *U. urealyticum*	UR	611	19.5	16.7	41.9	40.3	13.8	10.7
31. *Halobacterium sp.*	HB	2605	25.4	23.4	52.8	50.9	14.7	12.7
32. *B. halodurans*	BH	4066	16.3	12.6	46.1	39.8	7.0	4.0
33. *T. volcanium*	TV	1526	9.2	8.3	29.1	27.5	3.1	2.4
34. *M. loti*	MS	7281	20.9	13.7	53.7	40.4	11.5	6.0
35. *M. leprae*	ML	1605	9.1	8.4	38.6	36.5	1.3	1.0
36. *P. multocida*	PM	2014	6.0	4.4	18.8	16.9	3.7	2.1
37. *C. crescentus*	CC	3737	19.5	15.9	46.8	41.5	13.2	9.9
38. *S. pyogenes*	SP	1696	19.0	7.8	49.4	24.7	10.1	2.8
39. *S. aureus*	SA	2748	15.4	12.6	40.3	36.2	7.1	4.6
40. *L. lactis*	LL	2266	14.0	11.1	38.5	31.7	6.2	4.6
41. *M. pulmonis*	MU	782	18.5	17.1	51.6	50.3	10.4	8.9
42. *S. solfataricus*	SF	2977	15.2	14.8	27.9	27.6	11.2	10.8
43. *S. pneumoniae*	SM	2094	18.1	16.2	48.5	45.6	5.1	3.7
44. *S. meliloti*	SL	6205	11.2	8.6	35.7	28.3	5.1	3.7
45. *C. acetobutylicum*	CA	3848	18.8	16.6	51.2	47.6	9.6	7.9
46. *R. conorii*	RC	1374	25.6	25.2	52.8	52.2	4.4	4.2
47. *L. monocytogenes*	LM	2846	10.2	2.7	29.3	8.5	4.9	1.1
48. *L. innocua*	LI	3043	4.7	4.6	12.5	12.5	2.2	2.1
49. *Y. pestis*	YP	4083	11.9	9.8	37.3	32.9	5.1	3.7
50. *S. typhi*	SI	4767	13.0	7.8	33.3	20.9	6.7	3.7
51. *S. typhimurium*	SY	4553	3.5	3.4	9.6	9.6	1.9	1.8
52. *A. tumefaciens*	AT	5301	10.5	9.9	41.5	40.0	3.9	3.4
53. *S. coelicolor*	SR	7897	19.2	18.5	47.7	46.7	12.7	12.1
54. *T. tengcongensis*	TT	2588	13.5	12.9	35.8	35.3	6.3	5.8
55. *X. axonopodis*	XT	4312	15.0	7.1	34.2	14.7	10.8	5.5
56. *X. campestris*	XC	4181	4.3	4.3	11.3	11.3	2.8	2.7
57. *C. tepidum*	CR	2252	27.3	27.2	65.5	65.2	9.5	9.5
58. *O. iheyensis*	OC	3496	10.5	10.4	33.4	33.4	4.0	3.9
59. *S. agalactiae*	SG	2124	11.7	11.7	35.9	35.9	3.6	3.6
60. *B. suis*	BU	3264	18.8	16.6	51.2	47.6	9.7	7.9
Average		2804	20.2	13.6	45.2	35.2	12.5	6.7

[a]Percentage of ORFans out of all ORFs in genome.
[b]Percentage of ORFans shorter than 150 residues out of all short ORFs in genome.
[c]Percentage of ORFans longer than 150 residues out of all long ORFs in genome.
[d]Abbreviation of the genome's name.
[e]Number of ORFs in genome.
[f]Percentage of ORFans in genome at the time of addition to the database.
[g]Percentage of ORFans in genome after 60 genomes.

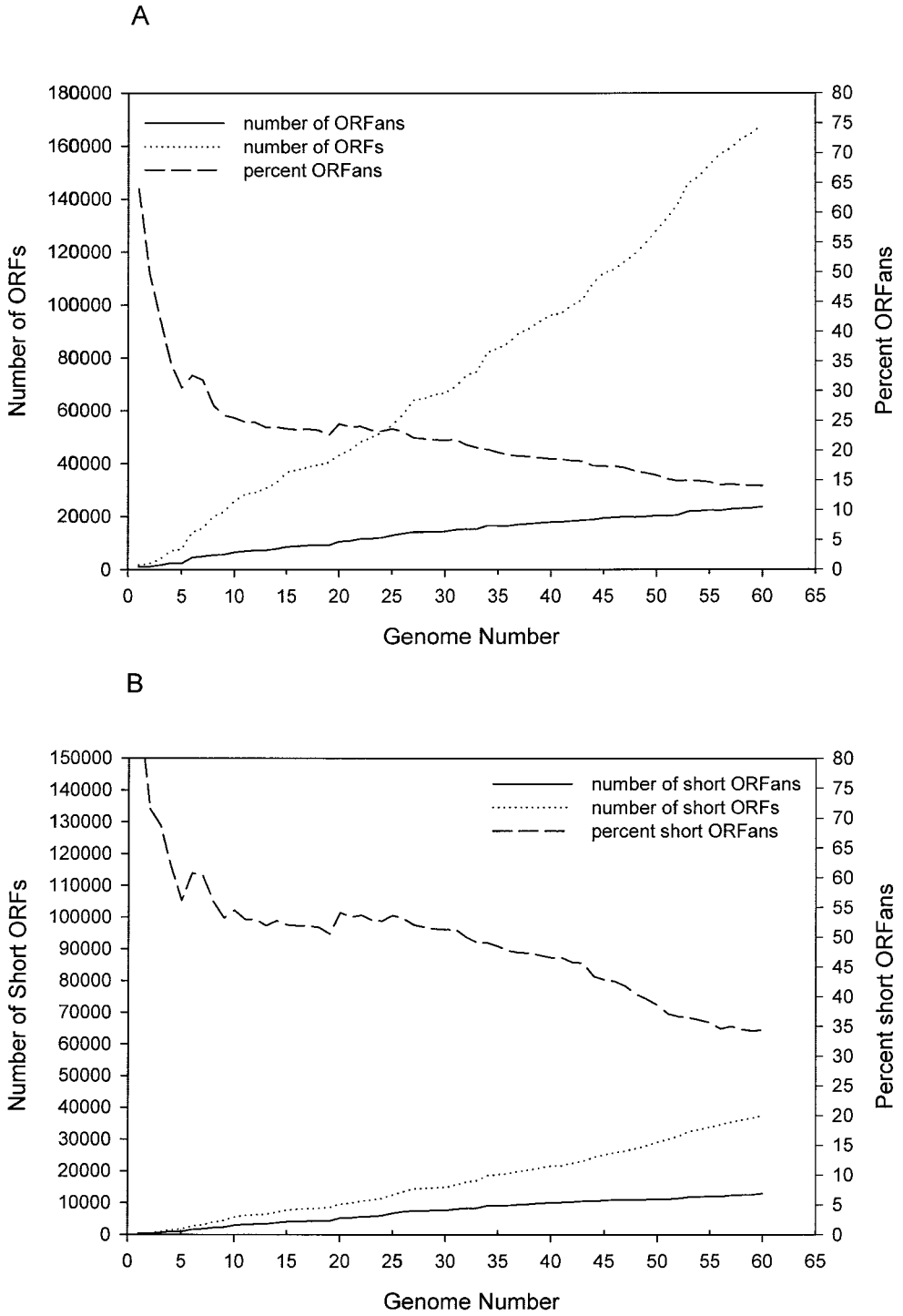


Fig. 1. Dynamics of ORFans in our database. (**A**) With each new sequenced genome (*x* axis, numbers as in Table I), the total numbers of ORFs and ORFans are growing (left-hand scale). At the same time, the fraction of ORFans out of the total number of ORFs is declining (right-hand scale). Each new genome contributes new ORFans of its own, and at the same time adds ORFs that have matches to sequences previously defined as ORFans. Usually, more new ORFans are added than old ORFans turned into non-ORFans; thus, the number of ORFans keeps growing. (**B**) The number of short ORFans is growing fast. After 60 genomes, their fraction of all short ORFs is 38%. (**C**) The number of long ORFans is growing slower. After 60 genomes, their fraction of all long ORFs is 7.0%.

do not find today any of the necessary "intermediate sequences" that must have given rise to these ORFans? Is the origin of these highly divergent sequences due to massive gene loss, or to a process of rapid evolution?[13,17,18]

Although ORFans entail interesting evolutionary puzzles and comprise a significant portion of the genetic material, they have not received proportional attention from the scientific community, and calls for "affirmative action" for

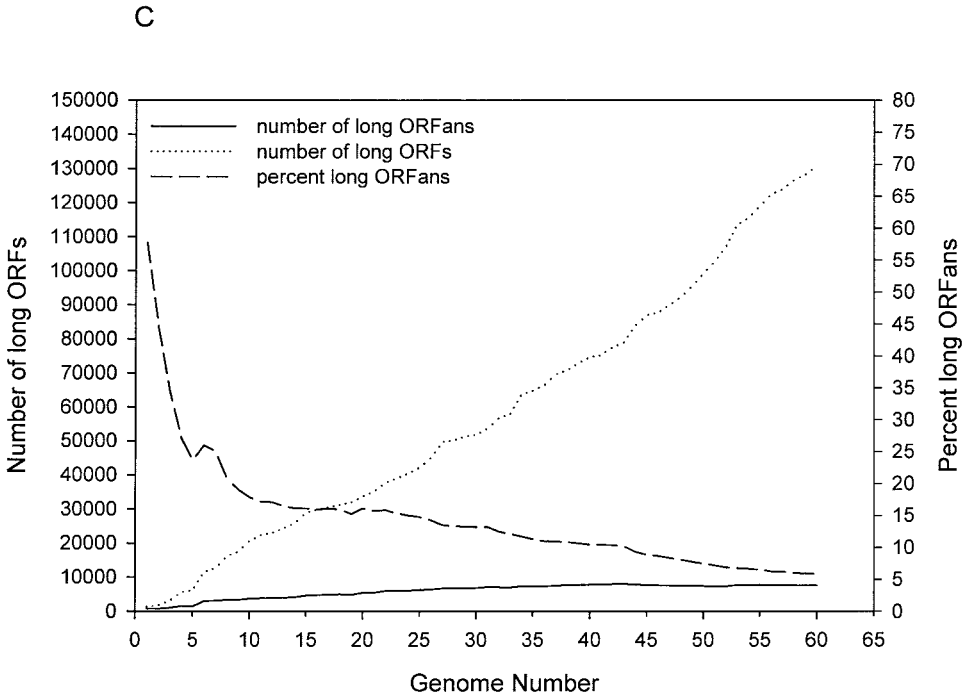


Figure 1. (Continued.)

ORFans have been suggested.[3] Possible explanations of why ORFans have been underemphasized include the following:

1. The high percentage of ORFans is an artifact of sparse sampling of the sequence space; ORFans will disappear as the complete genome sequences of more genomes become available.
2. Many ORFans may not correspond to expressed proteins, but rather, to errors or to incorrectly annotated genes.[2,13,18]
3. ORFans may correspond to nonessential proteins, or to rapidly evolving proteins with nonessential functions,[17,19] some of which may be in the process of extinction.
4. Because experimental characterization is expensive and time-consuming, it is best to focus first on the many sequence families containing homologs from numerous organisms. This is particularly true in Structural Genomics efforts[11,20] aimed at determining the structures of a carefully chosen representative set of proteins, so that relatively accurate computational models can be generated for the majority of the remaining proteins[21]; that is, the goal is to have most of the protein sequences within the so-called "homology-modeling" distance from a representative of known structure.
5. If most ORFans are distant members of known families, undetectable with current tools, in the future, as our knowledge enriches the databases, and as more sensitive tools are developed, we will be able to assign most of them to known families. Thus, ORFans are only transient puzzles that will eventually be elucidated.

Today, with the availability of a few dozen complete genome sequences, we claim that some of the above reasons may not be fully justified, as will be shown below. First, it seems that ORFans are not a mere artifact of sparse sampling; on the contrary, ORFans appear to be an intrinsic phenomenon of the genetic material, and their number continues to grow. Second, although some of the shorter ORFans may correspond to errors, it appears that the majority of the longer ORFans do correspond to expressed proteins. Third, even if we accept that most ORFans correspond to rapidly evolving proteins, then a number of other puzzling questions arise. For example, what are the forces involved in their rapid evolution, and how rapid was this evolution? Did it occur in one, single step, or was it a continuous event that we may be still witnessing today? What do these rapid changes imply about the function and structure of these proteins? Upon the acquisition of important functions, have ORFans stopped their rapid divergence? Do they correspond to the species determinants? Fourth, within 5 to 10 years, the functions and three-dimensional (3D) structures of representatives of the majority of the largest sequence families will probably be known, but little will be known about the large number of small and single-membered families. Finally, it is currently not clear whether most ORFans correspond to distant members of known families or to single-member families with unique functions and structures. Even if ORFans do correspond to the former, then it is not obvious that they can be straightforwardly assigned to their corresponding families. Furthermore, although the combination of experimental characterization and

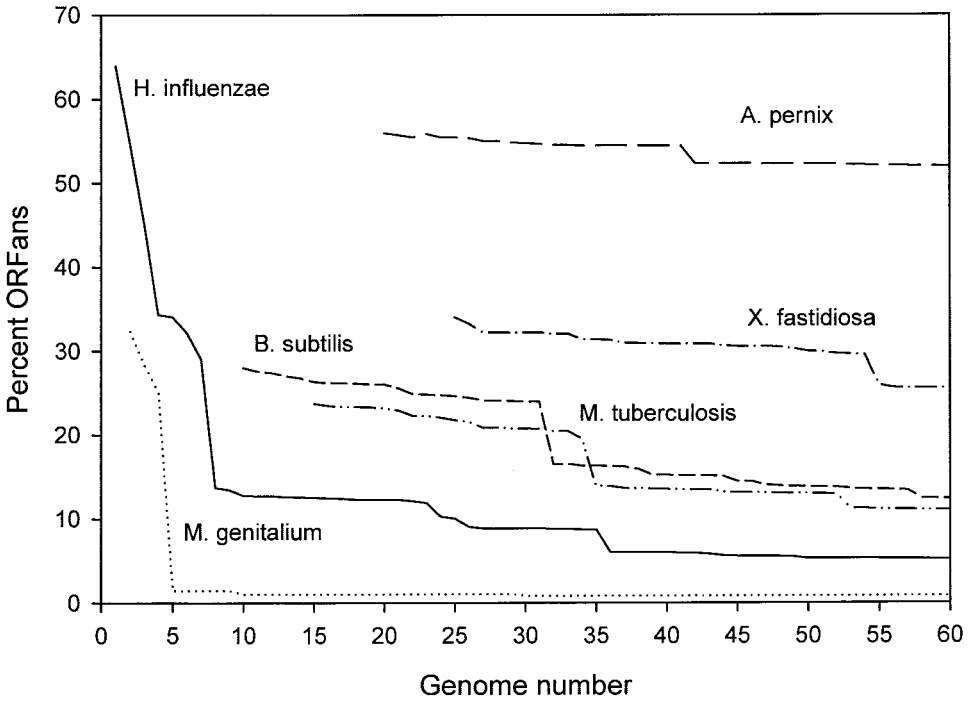


Fig. 2. Dynamics of ORFans in individual genomes. The changes in fraction of ORFans in individual genomes as new genomes are added to the database. The genomes shown are *H. influenzae* (1), *M. genitalium* (2) *B. subtilis* (10), *Mycobacterium tuberculosis* (15), *A. pernix* (20), and *Xylella fastidiosa* (25). The fraction of ORFans drops significantly when a closely related organism is added to the database. Similar to the *M. genitalium* and *B. subtilis* dynamics (see text), the percentage of ORFans in *M. tuberculosis* dropped from 19.5% to 14.0% after the addition of the obligate intracellular pathogen *Mycobacterium leprae* (35). Along the same line, the fraction of ORFans in *H. influenzae* dropped sharply upon the addition of *Escherichia coli* (8), and the small drop in fraction of ORFans that is observed for the crenarcheon *A. pernix* (20) corresponds to the addition of another crenarcheon, *Sulfolobus solfataricus* (42). Finally, the fraction of ORFans of the plant pathogen *X. fastidiosa* (25) dropped slightly after the addition of two other pathogens, *Vibrio cholerae* (26) and *Pseudomonas aeruginosa* (27), and a large drop was observed after the addition of the phylogenetically closer plant pathogen, *X. axonopodis* (55).

more sensitive computational tools will undoubtedly shed light on the functions of ORFans, the questions about the origin of ORFans and the mechanisms of their evolution will still remain.

Thus, both experimental and computational ORFan studies are required to reach a more complete understanding of both the genetic material and the evolutionary puzzles that ORFans entail. Here, we present a systematic census and dynamics analysis of singleton ORFans. We do not attempt to provide answers to all of the questions posed above, nor do we estimate the fraction of ORFans that may correspond to distant members of known families. Our focus is on those sequences that lie beyond the so-called "homology-modeling" distance from all other proteins, which we believe is the appropriate measure to address the theoretical and practical issues raised above. As a first step toward understanding the ORFan phenomenon, we present here a classification and a descriptive analysis of the ORFans in the first 60 fully sequenced microbial genomes.

MATERIALS AND METHODS

We downloaded the genome sequences of the first 60 published, fully sequenced microbial genomes from web-based databases, in the period between March 2000 and October 2002. When more than one strain of an organism was published, we considered only one strain. The genomes were sequentially added to our database in the chronological order in which they were published (see http://www.tigr.org/tdb/mdb/mdbcomplete.html). A complete list of strains, the download sites, and the dates we performed the downloads can be found at http://www.cs.bgu.ac.il/~nomsiew/ORFans.

After the addition of each genome to our genome database, we used gapped BLAST[22] to search for matches between the new genome's ORFs and the other ORFs of previously added genomes. We define a match if the first BLAST hit had an e-value below 10^{-3} (10^{-5} for alignments shorter than 80 residues). An ORF without matches was labeled as an ORFan. If a sequence previously labeled as an ORFan had a match with ORFs in the new genome, then its label was changed to non-ORFan.

To test the sensitivity of our ORFan counts to the definition of a match, we investigated three additional thresholds: (1) 10^{-1}, (2) 10^{-3}, and (3) 10^{-10}. The number of ORFans obtained at these thresholds differed by less than 25% from the number of ORFans counted with the original threshold, which appears to provide a good balance of correct homology detection and false positives. Thus, we concluded that the threshold used is a reasonable measure of homology, and that different threshold values do not significantly affect the qualitative trends of our analysis.

RESULTS

The 60 genomes considered in this study are listed in Table I in chronological order of publication. These genomes contain a total of 168,248 ORFs and include 50 bacteria, 9 archaea, and one eukaryote (*Saccharamyces cerevisiae*), accounting for 85%, 11%, and 4% of the ORFs, respectively.

Each genome was added to our growing database of fully sequenced genomes in the order listed in table I, and ORFans were labeled as described in the Methods section. In Table I, the columns labeled "Initial" list the percentage of ORFans in each genome that were counted at the time the genome was added to our database. The columns labeled "Final" list the percentage of ORFans in each genome after all 60 genomes were added. The total number of singleton ORFans in the database after 60 genomes is 23,634 (14%), of which 17,346 (73%) belong to bacterial genomes, 4180 (18%) belong to archaeal genomes, and the remaining (2108, or 9%) belong to *S. cerevisiae.*

In what follows, we first describe the changes in the number of singleton ORFans as each genome is added to our database, and analyze this dynamic pattern. We then compare the length distribution of ORFans versus that of non-ORFans and present separate dynamics analyses for short and long ORFans.

ORFan Dynamics in the 60 Genomes

As genomes are added one by one to our database of fully sequenced genomes, the total numbers of ORFans and ORFs are calculated. Analysis of the data over time shows two tendencies. The first is that the number of ORFans is steadily increasing (Fig. 1(A), left-hand scale). Each new genome added to the database affects the total number of ORFans in two ways. On the one hand, the new genome contains a number of sequences that have matches with previous ORFans, thus slightly reducing the total number of ORFans. On the other hand, the new genome adds new ORFans. In most genomes, the number of new ORFans added is larger than the number of matches to previous ORFans; thus, the total number of ORFans keeps growing.[4]

The second observed tendency is that the percentage of ORFans out of all ORFs is slowly diminishing (Fig. 1(A), right-hand scale). When only a handful of genomes were considered, the percentage of ORFans was relatively high. At the same time, along with the addition of the first five genomes, the percentage of ORFans dropped sharply from 64% to 31%. This may have suggested that with the sequencing of a few dozen more genomes, ORFans would quickly disappear. However, as Figure 1(A) (right-hand scale) shows, this has not yet happened. The addition of the sixth genome, *S. cerevisiae,* raised the percentage of ORFans to 33%, and since then, it is slowly declining. After the addition of the 11th genome, the percentage of ORFans was 25%; it took the addition of 22 more genomes before it reached 20%. After 60 genomes, the percentage of ORFans is 14%.

The initial percentage of ORFans in each new individual genome at the time it is added to our genome database (Table I) and the number of previously labeled ORFans that become non-ORFans vary widely depending on the evolutionary relationship between the new genome and previously sequenced genomes. For example, relatively large increases in the total number of ORFans are observed after the addition of highly divergent organisms, such as the yeast *S. cerevisiae* (genome 6), the thermophilic archaeon *Methanobacterium thermoautotrophicum* (9), the hyperthermophilic crenarcheon *Aeropyrum pernix* (20), and the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti* (34). On the other hand, the addition of *Chlamydia pneumoniae* (19), closely related to the previously sequenced *Chlamydia trachomatis* (17), caused a temporary drop in the total number of ORFans in the database.

The average initial percentage of ORFans in the individual genomes is 20% (Table I), a slightly lower figure than previous estimates.[2,5] The average for bacteria alone is 19%; for archaea alone, it is higher than the average, 27%, and for *S. cerevisiae,* it is 38%. With the exception of *H. influenzae,* which is the first genome in the database, *A. pernix* (20) has the highest initial percentage of ORFans, 56%.

As new genomes are added to the database, the percentages of ORFans in most previous genomes drop slightly. However, a sharp drop in a genome occurs when a closely related genome is added. This happens because many of the new organisms' ORFs have matches with old ORFans of the previously sequenced close relative. An example is the addition of *Mycoplasma pneumoniae* (genome 5) to the database (Fig. 2). Because almost all ORFs in the previously sequenced *Mycoplasma genitalium* (2) have matches in *M. pneumoniae,* the latter reduced almost all ORFans in the former at once. Furthermore, *M. pneumoniae* added very few new ORFans of its own: Only 6% of its ORFs were initially ORFans (Table I). Notice that the addition of *M. pneumoniae* thus created a number of non-ORFan families containing two ORFs only, one from *M. genitalium* and the other from *M. pneumonia,* but none matching sequences from other organisms. Thus, these new two-member families actually correspond to sequences specific to the Mycoplasma family of related organisms. We refer to these sequences as orthologous ORFans (see Discussion section).

Another example is the gram-positive bacterium *Bacillus subtilis* (genome 10). Its initial fraction of ORFans was 28%, which dropped slowly to 23% along with the addition of 21 more genomes. Upon the addition of the closely related bacterium *Bacillus halodurans* (32), the fraction of ORFans dropped sharply to 17%. Finally, after 60 genomes, *B. subtilis* contained 12% ORFans. Similar tendencies are observed in other families of closely related organisms (see legend, Fig. 2).

After 60 genomes, the final fraction of ORFans in the individual genomes ranges from less than 1% in *M. genitalium* to 33% in *S. cerevisiae* and 52% in *A. pernix,* with an average of 14% per genome (Table I).

To study general trends not affected by the individual variation of a particular genome, we computed the change in percentage of ORFans among groups of 10 consecutive genomes clustered together [Figure 3(A)]. The initial fraction of ORFans in the first group of genomes (1–10)

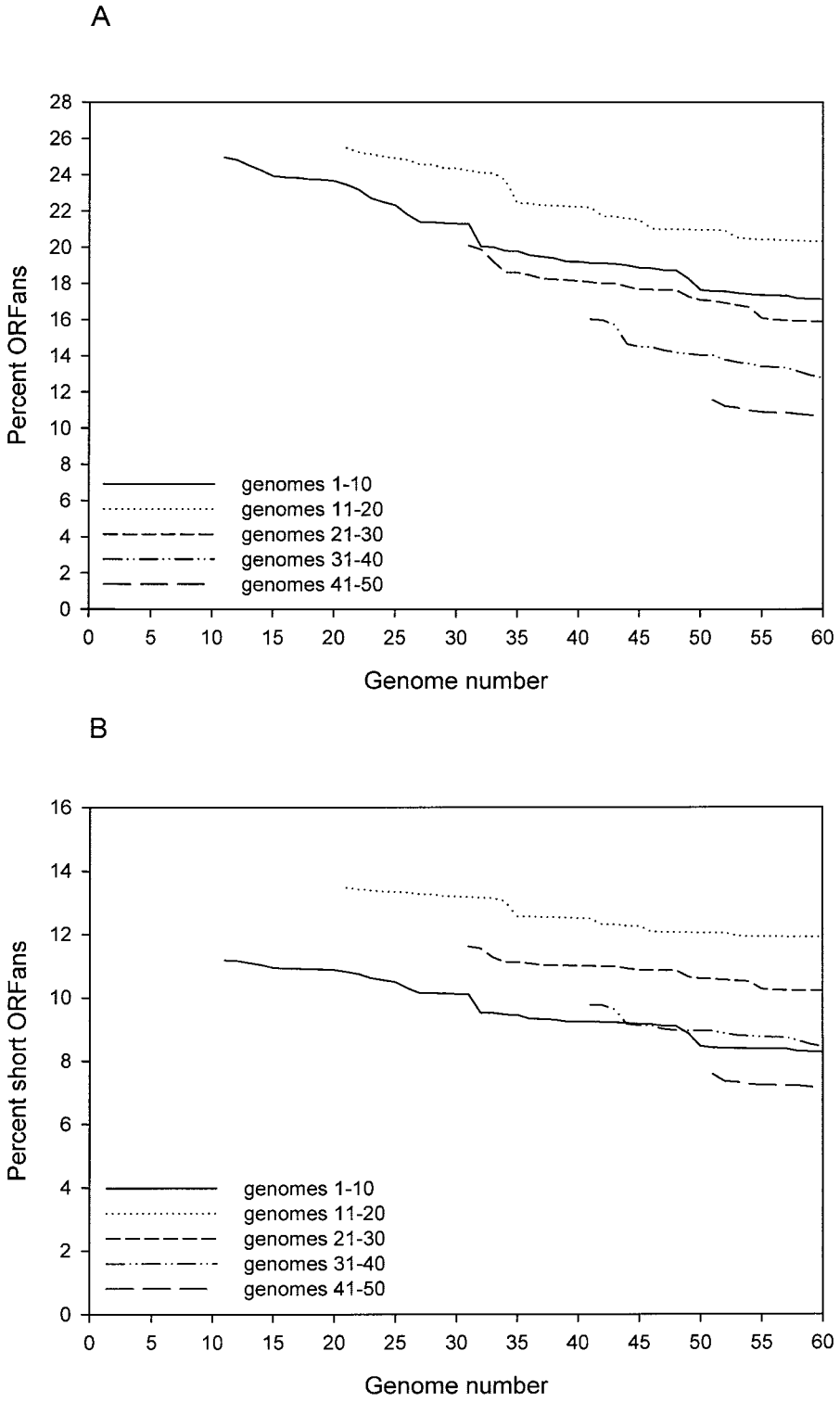


Fig. 3. Decline of ORFan percentage in groups of 10 consecutive genomes. (**A**) Each new genome that is added to the database contributes to a decline in the total number of ORFans of 0.1–0.2%. (**B**) The decline in percentage of short ORFans in groups of 10 genomes clustered together is very slow. (**C**) The decline in percentage of long ORFans in groups of 10 genomes clustered together is twice as fast as the decline in fraction of short ORFans.

dropped from 25% to 17% (after all 60 genomes were added); in the second group (genomes 11–20), the percentage of ORFans dropped from 26% to 20%, and in the third group (genomes 21–30), it dropped from 20% to 16%. In the fourth group (genomes 31–40), the percentage dropped from 16% to 13%, and in the fifth group (genomes 41–50), it

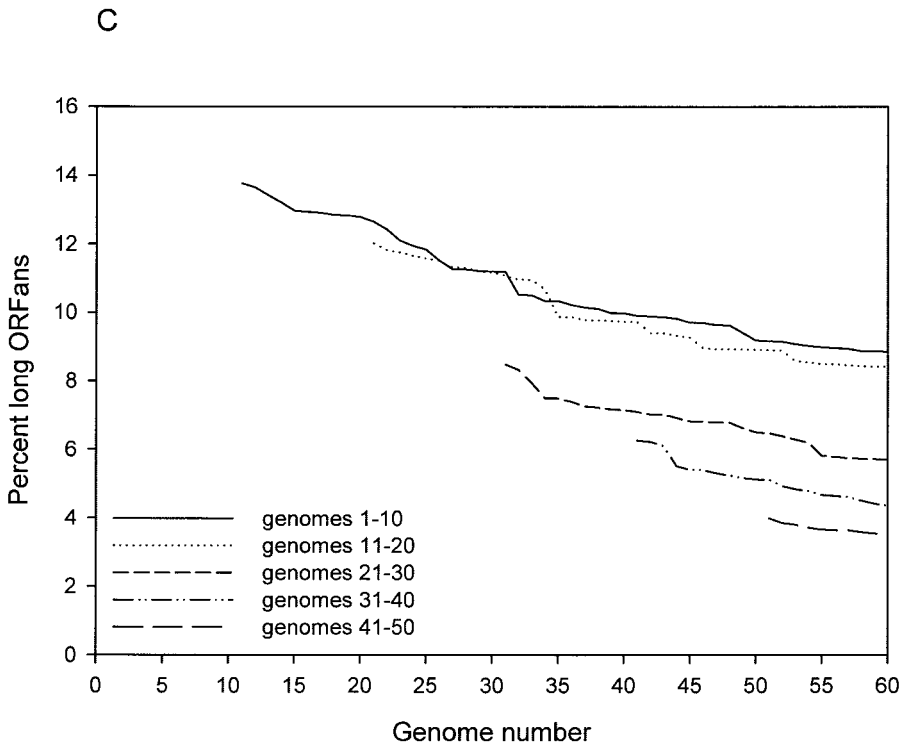


Figure 3. (Continued.)

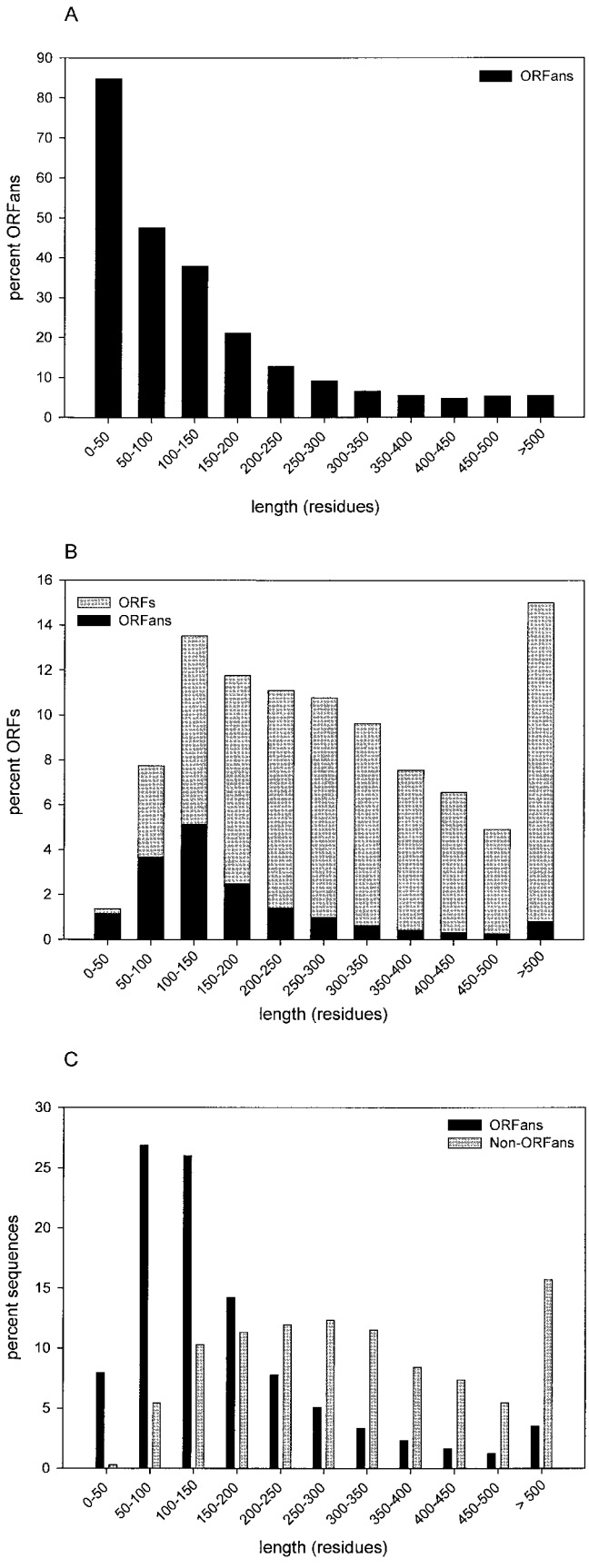
dropped only from 12% to 11%. Notice that the average initial percentage of ORFans for the last two groups is lower than that of the first three groups. On average, each new genome contributes to a decline of 0.1–0.2% in the percentage of total ORFans.

To further understand the growth rate of ORFans, we computed an analytic fit to the data presented in Figure 1(A). A quadratic fit resulted in a correlation of 0.998. An extrapolation of the number of ORFans suggests that the number of ORFans will keep growing along with the addition of a few dozen more genomes, until it reaches a maximum of about 26,000, at which point it will start declining.[4] This extrapolation can be trusted to predict only rough trends for the next few genomes, if these represent similarly diverse organisms as those currently in the database; that is, if the next genomes are closely related to genomes already in the database, the fraction of ORFans will decrease, and, conversely, if the next genomes correspond to highly diverse organisms, the fraction of ORFans will increase. Still, we expect that the percentage of ORFans out of all ORFs will keep declining slowly.

Notice that our counts of ORFans take into account only sequences in the 60 genomes considered here. Thus, if the full-sequence databases were included, some of the ORFans identified here might become non-ORFans, and new ORFans would appear. For example, only about one fourth of *S. cerevisiae*'s ORFans identified here have matches to sequences in the fly and worm genomes. This suggests that most of the remaining yeast ORFans correspond to yeast-specific sequences. Thus, although inclusion of the fly and worm genomes in our database would lower the current number of yeast ORFans, it would not eliminate all ORFans and would also add a significant number of new ORFans. This suggests that although some quantitative differences would exist if we considered the full-sequence databases, the main qualitative findings presented here would not change.

Length Distribution of ORFans and Non-ORFans

It has previously been observed that many of the ORFans correspond to short sequences.[2,23] In our database, the average amino acid residue length of ORFans is 169, and that of non-ORFans is 338. This prompted us to analyze the length distribution of ORFans compared with that of all ORFs and non-ORFans. Figure 4(A) indicates the percentage of ORFans among all ORFs in each length range. Figure 4(B) shows the length distribution of all ORFs (gray bars), with the corresponding percentage of ORFans (black bars) corresponding to the data in Figure 4(A). The figures show that there is a strong bias for ORFans among the shorter sequences. ORFans comprise 81% of all ORFs that are shorter than 50 residues (1881 sequences out of 2316) and 45% of all ORFs of length 50–100 residues (6347 sequences out of 14,208). Only in the length ranges of 150–200 residues (3360 sequences out of 19,727) and 200–250 residues (1837 sequences out of 19,082), the fraction of ORFans is closer to their overall fraction of 14% (17% and 10%, respectively). At higher length ranges, the fraction of ORFans out of all ORFs is significantly lower (4–6% per length range). In total, among the ORFs that are shorter than 150 residues,



ORFans comprise 38% (14,375 sequences out of 37,534), whereas among ORFs longer than 150 residues, ORFans comprise only 7% (9259 sequences out of 130,714). This implies that there is an over-representation of ORFans among the shorter ORFs.

We next compared the length distribution of ORFans [black bars in Figure 4(C)] with that of non-ORFans (gray bars). The percentage of ORFans (non-ORFans) at each length range out of all ORFans (non-ORFans) is shown. Among ORFans, 61% are shorter than 150 residues (the numbers of ORFans at the three first-length ranges are 1881, 6347, and 6147, for a total of 14,375 ORFans out of 23,634), whereas among the non-ORFans only 16% are short (435 + 7,861 + 14,863 = 23,159 non-ORFans out of 144,614). This shows that there is an abundance of short sequences among the ORFans.

Dynamics of Short and Long ORFans

The strong bias for short ORFans prompted us to distinguish between short and long ORFans in our dynamics analyses. We refer to a sequence as short if it has fewer than 150 residues, and as long if it has 150 or more residues. The dynamics of short and long ORFans are presented in Figure 1(B and C), respectively. The number of short ORFans grows faster than that of the long ORFans [Fig. 1(B and C), left-hand scale]. At the same time, the fraction of short ORFans out of short ORFs declines much slower than that of the long ORFans out of long ORFs [Fig. 1(B and C), right-hand scale]. Data extrapolations of the number of short and long ORFans suggest that the number of short ORFans will continue to grow for a few dozen more genomes, whereas the number of long ORFans is already close to its maximum value. Figure 3(B and C) shows in more detail the rates of decline of short and long ORFans in groups of 10 consecutive genomes. The decline in fraction of long ORFans is about twice that of the short ORFans.

In summary, it is clear that the behavior and dynamics of short ORFans are significantly different than those of long ORFans.

DISCUSSION

We have shown that the number of ORFans is currently growing, whereas their fraction among ORFs is slowly diminishing. We have distinguished between ORFans that are shorter and longer than 150 residues. The short ORFans are accumulating at a faster rate than the long

Fig. 4. Length distribution of ORFans. (**A**) ORFans comprise the majority among sequences shorter than 150 residues. The percentage of ORFans out of ORFs in length ranges of 50 residues is shown. Length ranges above 500 residues are shown as one bin. There are 1881 ORFans shorter than 50 residues (81%), and 6347 (45%), 6147 (29%), and 3360 (17%) in the following three length ranges. (**B**) Length distribution of ORFs in each length range out of all ORFs in the database (gray bars) and the corresponding fraction of ORFans in each length range (black bars) show a strong bias toward short sequences among ORFans. (**C**) Length distribution of ORFans among all ORFans (black bars) and of non-ORFans among all non-ORFans (grey bars). There is an over-representation of short sequences among ORFans: 63% of all ORFans are shorter than 150 residues, whereas only 16% of all non-ORFans are short.

ones and are at the same time disappearing more slowly than the long ORFans. Extrapolations suggest that whereas it would take a few dozen more new genomes before the number of short ORFans begins to decline, the number of long ORFans is likely to begin declining sooner.

The fact that more than half of the ORFans are short is puzzling. The slower decrease of short ORFans, and their relative abundance, may be partially due to the presence of short ORFs that do not correspond to expressed proteins. It is probable that some of the short ORFs are the result of random distributions of nucleotides, or of sequencing errors that lead to frame shifts and to wrong stop codons.[13,24,25] This may be especially true for the ORFans that are shorter than 50 residues. Thus, these ORFans do not match other sequences. Nevertheless, the fact that the fraction of short ORFans is declining (albeit slowly) suggests that some short ORFans do correspond to expressed proteins. Indeed, most short ORFans in *M. genitalium* had matches in *M. pneumoniae,* strongly indicating that some short ORFans do correspond to expressed proteins. Thus, we conclude that the abundance of short ORFans cannot be only a consequence of their being mainly nonexpressed proteins. Another possible reason for the abundance of short ORFans could be technical: It may be more difficult for sequence comparison programs such as BLAST[22] to find significant matches for shorter sequences (see the work of Mackiewiez et al.[26] for yet another possible explanation). Still, it remains unclear whether the above-suggested reasons fully explain the strong bias toward short sequences among ORFans.

Conversely, the faster decrease of long ORFans and their lower percentage in genomes suggests that most of them correspond to expressed proteins (also supported by increasing experimental evidence from the Halobacterium NRC-1 structural genomics project (B. Shaanan, J. Eichler, and D. Fischer, unpublished results) and other ORFan structure determination studies.[9,27,28] One explanation for their fast decline may be that long proteins are more conserved among different organisms than short ones. Another explanation could be that longer sequences are often constructed of a few modules. In our computations, two sequences having an above-threshold match with each other are considered non-ORFans, even if some regions along their sequences match no other ORF. These unmatched regions are in fact ORFans. We refer to them as "ORFan modules." Preliminary computations indicate that there are 13,601 ORFan modules longer than 40 residues among the non-ORFans, of which 6669 are longer than 150 residues. This suggests that in addition to the 23,634 singleton ORFans identified above, there are many other ORFan regions awaiting interpretation.

In addition to the ORFan modules, we propose to focus on two different types of ORFans: "paralogous ORFans" and "orthologous ORFans." Paralogous ORFans are defined as sequences that have matches with other ORFs in the same genome, but none with ORFs in other genomes. For example, in our database, there are 26 *B. Subtilis* paralogous families containing 58 sequences having no match in the other organisms. The presence of paralogous ORFans suggests that these ORFans do correspond to functional proteins specific to a single organism. Orthologous ORFans are defined as sequences that have matches only among the family members of related organisms but none outside the family. For example, in our database, there are 49 orthologous families, containing sequences from *B. subtilis* and *B. halodurans,* that match no sequences of other organisms. Thus, in addition to the 1018 (507 + 511) singleton ORFans of *B. subtilis* and *B. halodurans,* a non-negligible number of paralogous and orthologous ORFans in these two organisms do not match ORFs in any other genome. As with the paralogous ORFans, the presence of orthologous ORFans suggests that they do correspond to functional proteins specific to a family of closely related organisms. As more complete genome sequences of closely related organisms are determined, more orthologous ORFans will appear. Families of paralogous and orthologous ORFans can be searched via our ORFan website, at http://www.cs.bgu.ac.il/~nomsiew/ORFans.

In summary, in addition to the singleton ORFans, we propose to focus future studies aimed at understanding the functions and origins of the three other types of ORFans discussed above: (1) ORFan modules; (2) paralogous ORFans; and (3) orthologous ORFans, with emphasis on the longer ones. Although, eventually, few longer singleton ORFans will remain, the other three types of ORFans will continue to present rchallenges to our full understanding of the genetic material.

ACKNOWLEDGMENTS

This work is related in part to an ORFan joint research project with David Eisenberg; we thank him for valuable discussions. We thank Yaniv Azaria for help in developing the ORFan website.

REFERENCES

1. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. Whole-genome random séquencing and assembly of *Haemophilus influenzae* Rd. Science 1995;269:496–512.
2. Fischer D, Eisenberg D. Finding families for genomic ORFans. Bioinformatics 1999;15:759–762.
3. Fischer D. Rational structural genomics: Affirmative action for ORFans and the growth in our structural knowledge. Prot Eng 1999;12:1029–1030.
4. Siew N, Fischer D. Twenty thousand ORFan microbial protein families for the biologist? Structure 2003;11:7–9.
5. Fraser CM, Eisen JA, Salzberg SL. Microbial genome sequencing. Nature 2000;406:799–803.
6. Bloom BR. On the particularity of pathogens. Nature 2000;406: 760–761.
7. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM , Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B. Genome sequence of the human malaria parasite *Plasmodium falciparum.* Nature 2002;419:498–511.
8. Brenner SE. Target selection for structural genomics. Nat Struct Biol 2000;7:967–969.
9. Alimi JP, Poirot O, Lopez F, Claverie J-M. Reverse transcriptase-polymerase chain reaction validation of 25 "orphan" genes from *Escherichia coli* K-12 MG1655. Genome Res 2000;10:959–966.
10. Hutchison CA III, Peterson SN, Gill SR, Cline RT, White O,

Fraser CM, Smith HO, Venter JC. Global transposon mutagenesis and a minimal Mycoplasma genome. Science 1999;286:2165–2169.
11. Vitkup D, Melamud E, Moult J, Sander C. Completeness in structural genomics. Nat Struct Biol 2001;8:559–566.
12. Oliver SG. From DNA sequence to biological function. Nature 1996;379:597–600.
13. Dujon B. The Yeast Genome Project: What did we learn? Trends Genet 1996;12:263–270.
14. Doolittle RF. Microbial genomes multiply. Nature 2002;416:697–700.
15. Coulson AFW, Moult J. A unifold, mesofold, and superfold model of protein fold use. Proteins 2002;46:61–71.
16. Doolittle RF. A bug with excess gastric avidity. Nature 1997;388:515–516.
17. Schmid KJ, Aquadro CF. The evolutionary analysis of "orphans" from the Drosophila genome identifies rapidly diverging and incorrectly annotated genes. Genetics 2001;159:589–598.
18. Wood V, Rutherford KM, Ivens A, Rajandream M-A, Barrell B. A reannotation of the *Saccharomyces cerevisiae* genome. Compar Funct Genomics 2001;2:143–154.
19. Jordan IK, Rogozin IB, Wolf YI, Koonin EV. Microevolutionary genomics of bacteria. Theor Popul Biol 2002;61:435–447.
20. Chance MR, Bresnick AR, Burley SK, Jiang JS, Lima CD, Sali A, Almo SC, Bonanno JB, Buglino JA, Boutlon S, Chen H, Eswar N, He G, Huang R, Ilyin V, McMahan L, Pieper U, Ray S, Vidal M, Wang LK. Structural genomics: A pipeline for providing structures for the biologist. Protein Sci 2002;11:723–738.
21. Fischer D, Baker D, Moult J. We need both computer models and experiments. Nature 2001;409:558.
22. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acid Res 1997;25:3389–3402.
23. Skovgaard M, Jensen LJ, Brunak S, Ussery D, Krogh A. On the total number of genes and their length distribution in complete microbial genomes. Trends Genet 2001;17:425–428.
24. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. The COG database: New developments in phylogenetic classification of proteins from complete genomes. Nucleic Acid Res 2001;29:22–28.
25. Andrade MA, Daruvar A, Casari G, Schneider R, Termier M, Sander C. Characterization of new proteins found by analysis of short open reading frames from the full yeast genome. Yeast 1997;13:1363–1374.
26. Mackiewicz P, Kowalczuk M, Gierlik A, Dudek RM, Cebrat S. Origin and properties of non-coding ORFs in the yeast genome. Nucleic Acids Res 1999;27:3503–3509.
27. Monchois V, Abergel C, Sturgis J, Jeudy S, Claverie JM. *Escherichia coli* ykfE ORFan gene encodes a potent inhibitor of C-type lysozyme. J Biol Chem 2001;276:18437–18441.
28. Goulding CW, Parseghian A, Sawaya MR, Cascio D, Apostol MI, Gennaro ML, Eisenberg D. Crystal structure of a major secreted protein of *Mycobacterium tuberculosis-MPT63* at 1.5-Å resolution. Protein Sci 2002;11:2887–2893.