Genomic ORFans - Past, Present and Future

Siew, N.[1,2] and Fischer, D.[2]

[1] Department of Chemistry, Ben Gurion University

[2] Bioinformatics, Department of Computer Sciences, Ben Gurion University

Sequence ORFans are orphan ORFs (Open Reading Frames) that show no sequence similarity to any other sequence in the databases. ORFans are of particular interest, not only as evolutionary puzzles, but also because little can be learned about them using bioinformatic tools. Thus, the presence of many ORFans does not allow for a full characterization of the genomic content of organisms.

Here we show that the number of ORFans in the first 43 completely sequenced microbial genomes is steadily growing and that after 43 genomes their number is 18,552 (18%) out of a total of 102,114 ORFs. Our data shows that the addition of each new complete genome slowly reduces the number of previous ORFans, but at the same time, the new genome also adds a larger number of new ORFans, and thus the number of ORFans is growing. However, the fraction of ORFans among all ORFs is slowly declining.

Our analysis of size distribution of ORFans and non-ORFans indicates a strong bias towards shorter sequences among ORFans: sequences shorter than 150 residues account for 56% of all ORFans. The fraction of long ORFans in the genomes declines in a rate twice as fast as that of short ORFans. A possible explanation for this bias could be that some short ORFans do not correspond to expressed proteins, or due to limitations of the tools used to identify sequence similarities.

We conclude that the large observed percentage of ORFans reflects a yet unexplained intrinsic property of the genetic material and that further studies aiming at understanding Nature's protein diversity should also include ORFans.