Poster C-45

On the origin of microbial ORFans: Quantifying the strength of the evidence for viral lateral transfer

Authors:
Yanbin Yin (*University at Buffalo*)
Daniel Fischer (*University at Buffalo*)

Short Abstract: We provide evidence suggesting that lateral transfer from viruses alone is unlikely to explain the origin of the majority of microbial ORFans in the majority of prokaryotes. Consequently, other, not necessarily exclusive, mechanisms are likely to better explain the origin of the increasing number of ORFans.

Long Abstract:
ORFans are defined as ORFs (Open Reading Frames) having no sequence homologs in other genomes. Since we coined the term “ORFan” in 1999, accumulating evidence has demonstrated that most ORFans correspond to real, functional proteins, and not to errors in the ORFs annotation. Previous studies in our group, and subsequently repeated by others have shown that as more genomes are being sequenced, the number of ORFans continues to grow. However, the origin of microbial ORFans is still one of the unexplained puzzles of the post-genomic era. Several hypothesis on the origin of ORFans have been suggested in the last few years, all based on selected, relatively small, subsets of ORFans, but to the best of our knowledge, no comprehensive, genome-wide ORFan study has been carried out. One of the hypotheses for the origin of ORFans is that they have been acquired thru lateral transfer from viruses. Here we carry out a genome-wide study on the origins of ORFans to quantify the strength of current evidence supporting this hypothesis.

To obtain an ORFans collection we downloaded the 277 microbial genomes available at the NCBI on Nov. 03, 2005. We carried out all vs. all (820,768 ORFs) BLASTP searches to identify homologs for each ORF. To identify ORFans, we developed a novel method that defines two featured values for each microbial ORF. The first is the “H value”, which is simply the number of genomes (including the residing genome) that contain at least one homolog of this ORF. The second value we define is the “U value”, which is a measure of the “uniqueness” of the ORF, and is a normalized sum of the ORFs homologs, weighed by the overall genomic distance between the residing genome and the genomes of the ORF’s homologs. Thus, ORFs with H=1 and one BLASTP hit were classified as singleton ORFans; ORFs with H=1 and more than 1 BLASTP hits were classified as paralogous ORFans; ORFs with H>1 and U<=0.1 were classified as orthologous ORFans. In total, we collected 110,186 ORFans (13.4% of all ORFs in the 277 genomes) of which 64,324 (7.8%) are singleton ORFans, 10,419 (1.3%) are paralogous ORFans and 35,443 (4.3%) are orthologous ORFans.

To identify viral homologs, for each proteome in our 277 genome database, we performed a BLASTP search against the public virus proteins database and computed two percentages: ORFans-VH%, the percentage of ORFans having homologs in viruses and non-ORFans-VH%, the percentage of non-ORFans having homologs in viruses. One way to

quantify the strength of the hypothesis that the origin of ORFans is viral is to compare the value of ORFans-VH% with that of non-ORFans-VH%. A significantly higher ORFans-VH% would suggest that viral transfer is more common among ORFans than among non-ORFans (which corresponds to the overall detectable baseline of transfer). Surprisingly, we found that out of the 277 genomes studied, only 22 (7.9%) had ORFans-VH% > non-ORFans-VH%. Eighteen of these 22 genomes are members of Firmicutes, and 4 belong to Gamma-proteobacteria. Taking all the 277 genomes together we found that only 2.8% of all ORFans have viral homologs while the percentage for all non-ORFans is 7.8% (p-value 2.2e-16). These findings suggest that the evidence based on current homology to viruses is very weak in general.

When comparing the non-ORFans-VH% versus the ORFans-VH% values of various prokaryotic groups, we found that Firmicutes (66 genomes) and Gamma-proteobateria (63 genomes) have significant higher percentages than the other 148 genomes ("Others"). This suggests that the current virus database may be biased towards those viruses attacking Firmicutes and Gamma-proteobacteria, and on average Firmicutes (with some notable exceptions) tend to be more attackable by viruses.

Thus, we can only claim that the evidence today is weak in general, and future sampling of the viral genomes may provide stronger evidence. With more viruses sampled, at least for some prokaryote genomes, the percentage of ORFans with homologs in viruses can become very high. However, it is questionable whether a full sampling of viral genomes will provide homologs to 100% of the ORFans or only to a fraction of them. For instance, we have found 44 genomes with no viral homologs for any of their ORFans, and with a percentage of non-ORFans with virus homologs significantly lower than the rest of the genomes. This suggests that they may correspond to genomes immune to viral attack, as is the case for obligatory species. Even if lateral transfer from viruses turns out to be the main origin of microbial ORFans, one is still left with the need to explain the origin of the also abundant viral ORFans (in our viral genome database, 27% of the viral proteins correspond to viral ORFans, i.e. ORFs with no homologs). Most likely, as is the case in so many cases in evolutionary biology, the origin of ORFans will turn out to be non-exclusive, and may include other yet-unknown mechanisms.