Detection of genomic contaminations and horizontal gene transfer with *in silico* metagenomic experiments

Alexander I. Tuzhikov,

Institute for Information Transmission problems, RAS, Moscow, Bolshoi Karetniy, 19/1,
alexander.tuzhicov@gmail.com

Yuri V. Panchin,

Institute for Information Transmission problems, RAS, Moscow, Bolshoi Karetniy, 19/1,
ypanchin@yahoo.com

Alexander Y. Panchin

*Institute for Information Transmission problems, RAS, Moscow, Bolshoi Karetniy, 19/1,*alexpanchin@yahoo.com

Due to a rapid development in the amount of sequenced eukaryotic and prokaryotic genomes and a remarkable increase in the performance of modern computers, it is becoming very popular to use high-throughput genomic data for the purpose of phylogenetic analysis and in other studies in the field of evolutionary biology. However, in these analysis errors may arise when genes acquired via horizontal gene transfer (HGT, the acquisition of genes from an organism other than a direct ancestor) are treated as genes derived from a direct ancestor. Automated detection of genes that underwent HGT or possibly contaminating genomic sequencing could improve the precision of evolutionary genomic analysis. We have formulated and implemented one such approach and developed a test to evaluate its performance.

We used the dataset of 43 complete genomes of organisms from different domains of life (including bacteria, archaea, plants, protists, fungi and animals) to test the idea that it is possible to separate predicted protein-coding genes with different phylogenetic history (putative HGT or contamination) by comparing arrays of protein BLAST bit scores and clustering genes with similar arrays of these scores. Consider gene s1 from genome S. Suppose it has the score s1A against its BH (Best Hit) from genome A, s1B against its BH from genome B e.t.c. The scores of gene s1 can be represented as a dot with the following coordinates in an n-space:

$$(s1A, s1B, s1C \dots s1Z)$$

Similar coordinates can be obtained for all genes from genome S and distances between these dots in an n-space may be calculated. After normalizing the scores, we tried using three distance measures: Spearman’s ranked correlation, Pearson’s correlation and Euclidian distance. Among these three measures, surprisingly, Euclidian distance gave the best results in clustering attempts.

We chose bit scores over other measures of sequence similarity due to several important properties. Bit scores do not depend on the database size and represent a level of alignment similarity that takes alignment size into account. We normalize the bit score for each hit by a value equal to the maximum bit score the query sequence could receive if an identical match was found. We assume that sequences from a single genome share the same evolutionary history (are not obtained via HGT and are not contaminations) if they cluster together in the mentioned Euclidian n-space. We attempted several methods of gene clustering implemented in the R package “hclust” using these distances. To date the best clustering results were obtained using Ward's minimum variance criterion that minimizes the total within-cluster variance [1]. At the initial step, all clusters are singletons (clusters containing a single point) and at each step, the pair of clusters with minimum between-cluster distance is merged. Ward's minimum variance uses squared Euclidian distances.

To test whether the clustering approach can discern contaminations and/or HGT we performed a series of procedures called “*in silico* metagenomic experiments”. Genes obtained from two different organisms are pooled together as if we don’t know which organism they came from and the clustering procedure (blindly) attempts to cluster the sequences based on the normalized arrays of bit-scores. For example, when *Drosophila simulans* and its endosymbiont *Wolbachia* genes were combined and clustered the following results were obtained:

1. The majority of *Wolbachia* sequences and the majority of *Drosophila* sequences were clustered separately.

2. A small number of *Wolbachia* sequences were present in *Drosophila* clusters. These sequences are likely candidates for *Drosophila* to *Wolbachia* horizontal gene transfer.

3. Likewise a small number of *Drosophila* sequences are present in the *Wolbachia* clusters. These sequences are likely candidates for *Wolbachia* to *Drosophila* horizontal gene

transfer.

To further support our findings we decided to analyze the genes putatively transferred from *Drosophila* to *Wolbachia* using PFAM (http://pfam.xfam.org/) domain search. The most impressive finding was that the majority of putative horizontally transferred sequences detecting with our approach encoded ankyrin repeat domains (ANK domains) – previously reported as HGT candidates. Out of a total 26 *Wolbachia* genes that encoded proteins with ankyrin repeats in our dataset, 22 were found in *Drosophila* clusters and only 4 clustered with the majority of other *Wolbachia* genes.

Recently Jernigan and Bordenstein [2] showed that the lifestyle of bacteria, rather than phylogenetic history, is a predictor of ANK repeat abundance. They also showed that phylogenetically unrelated organisms that forge facultative and obligate symbioses with eukaryotes show enrichment for ANK repeats in comparison to free-living bacteria. This observation was especially strong for obligate intracellular bacteria. Ankyrin domains are very common in eukaryotes, but much rarer in bacteria, with the exception of parasites and symbionts. In a paper by Siozios et al. [3] it is concluded that ankyrin genes are likely to be horizontally transferred between strains with the aid of bacteriophages. Al-Khodor et al. [4] also suggests that prokaryotic genes encoding ANK-containing proteins have been acquired from eukaryotes by horizontal gene transfer.

We are thrilled that our approach detected these particular genes as horizontally transferred. Our clustering does not take protein domain information into account so we believe that this provides an argument for the validity of the procedure we are developing and a biological explanation for the obtained results, although additional research is required. We were able to obtain similar clustering results for other pairs of distant organisms with *in silico* metagenomic experiments.

Recently Crisp et al. [5] used a measure called HGT index, h (the difference between the BLAST bitscores of the best non-metazoan and the best metazoan matches) to identify genes horizontally transferred to metazoan genomes from non-metazoan genomes. They argued that HGT is common not only for bacteria, but also for many metazoan species including humans and is likely to have contributed to the biochemical diversity within the animal kingdom. Unexpected sources of contaminating sequences have been shown present in genomic data [6]. We believe that our approach might provide a new way to look for detect HGT and contaminations in genomic data.

Acknowledgments

This study has been supported by the Dynasty foundation fellowship to AP and by RFBR grant 15-04-06113.
We are grateful to Vladimir Aleoshin for his insightful comments and discussions.

References

1. J. Ward (1963) Heirachical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, 58:301
2. K. Jernigan and S. Bordenstein (2014) Ankyrin domains across the Tree of Life. *PeerJ*. 2:e264.
3. S. Siozios (2013) The diversity and evolution of Wolbachia ankyrin repeat domain genes. *PLoS One*. 8(2):e55390
4. S. Al-Khodor et al., (2010) Functional diversity of ankyrin repeats in microbial proteins. *Trends Microbiol*. 18(3):132-9
5. A. Crisp et al. (2015) Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes, *Genome Biology*, 16:50.
6. S. Merchant et al., (2014) Unexpected cross-species contamination in genome sequencing projects. *PeerJ*. 2:e675