Regression analysis for automated detection of horizontal gene transfer

Alexander Y. Panchin, Yuri V. Panchin

Institute for Information Transmission Problems, RAS, alexpanchin@yahoo.com

Alexander I. Tuzhikov

Institute for Information Transmission Problems, RAS alexander.tuzhikov@gmail.com

Genes within a single genome may have different phylogenetic histories: some passed down vertically from the direct ancestors of an organism, while others may have been acquired via horizontal gene transfer (HGT) from other species. HGT is widely accepted as an important factor in the evolution of prokaryotic organisms [1]. A number of HGT events between prokaryotes and eukaryotes and even between different eukaryotic species have been described, and in some cases, the molecular mechanisms that facilitate such events have been uncovered [2]. There is ongoing controversy about the extent in which HGT plays a role in eukaryotic evolution. Numerous examples of HGT were found in different eukaryotic species, including humans by Crisp et al [3]. However, a study by Ku et al concluded that even gene transfer from bacteria to eukaryotes is episodic and coincides with major evolutionary transitions such as the origin of chloroplasts and mitochondria [4].

We developed a novel approach to identify horizontal gene transfer (HGT) events in genomic sequences. Predicted protein-coding sequences from pairs of genomes are mixed together and logistic regression models are trained on a sub-sample of these sequences to separate them. The models use arrays of normalized BLAST bit scores obtained for each sequence by comparison with its closest hit from a number of other genomes, covering different domains of life. Regression models that passed validation on independent subsamples are used to identify sequences that cluster with sequences from other genomes (predicted HGT events). Using the genomes of 61 species, we confirmed some previously reported cases of HGT (such as the transfer of ankyrin-encoding sequences from eukaryotes to their symbiotic bacteria or the acquisition of taumatins by *Caenorhabditis*) and predicted novel cases of HGT (such as the acquisition of MaoC dehydratase domain containing proteins by *Lokiarchaeota*).

Consider protein sequence X from a mix of sequences encoded by genomes S1 and S2. Suppose it has the normalized score X1 against its BH (Best Hit) from genome A, X2 against its BH from genome B e.t.c., up to Xn, where n = 59 (61 genomes minus genomes S1 and S2). Logistic regression analysis can optimize the following prediction function:

$$\ln(Y/(1-Y))=a+b1X1+b2X2+b3X3 \ldots +bnXn$$

Here, Y (with values ranging from 0 to 1) is the estimated probability that the gene belongs to a given cluster (S1 or S2). This prediction function is optimized on a training set of sequences (a random 75% subsample) and verified on the remaining set of sequences. R2 values were used to estimate if the regression explains the variance in the data and is capable of separating sequences from genomes S1 and S2.

We have selected an R2 = 0.73 cut-off as a criteria for the successful separation of genes from two genomes. In most cases, this corresponds to > 90% true positive predictions. The models were bootstrapped (N=1000). If the separation is successful, the genes placed in the incorrect genomic clusters with high confidence (95% probability according to the bootstrapped model) are considered outsider genes. Their presence in the wrong genomic cluster could indicate that they underwent HGT or are present in a genome assembly because of contamination. Core sequences within each genome are defined as sequences that were found in the correct cluster in all cases when the regression model was successful at separation. “Outsider once” sequences are defined as sequences that were found in the correct cluster in all but one case when the regression model was successful at separation. “Outsider multiple” sequences are defined as sequences that were found in the incorrect cluster in at least two cases when the regression model was successful at separation.

Each protein-coding sequence with hits in 3 or more genomes was analyzed using PFAM. For each genome, we compared the number of genes with each discovered PFAM domain between the core and outsider multiple sequences. We used a chi-square test to assess whether any domains were significantly over- or underrepresented among outsider multiple sequences. We considered p-values of 5.5 x 10-7 to be statistically significant (this is because a Bonferroni correction for multiple comparisons was applied to account for 89688 domain/species combinations used in the analysis). Note that technically this is not a strict

statistical procedure, since similar genes will have similar arrays of bit scores and our classification of these genes will not be independent. Similarly, this does not distinguish multiple HGT events of sequences containing the same PFAM domain from a single HGT event followed by multiple gene duplications. Because of this, all main findings require manual evaluation as a second step.

We performed 61*60/2 = 1830 intergenomic comparisons involving a total of 296639 predicted protein-coding sequences with a hit in at least 3 genomes. A total of 67% of the obtained regression models were successfully validated and thus allowed the distinction between the sequences of two genomes.

The majority of sequences (288012, 97.1%) from each analyzed genome were classified as core sequences in all metagenomic experiment with validated regression models. 4279 (1.4%) of the sequences were classified as “outsider once”. The remaining 4348 (1.5%) sequences were classified as “outsider multiple” and are the most likely candidates for HGT.

We used a PFAM analysis to check for any PFAM domains are significantly overrepresented in “outsider multiple” sets of sequences. A list of top-ranked findings (ranked by significance) is presented in Table 1.

The top two findings in our analysis appear to confirm previously known cases of HGT: Ankyrins in Wolbachia endosymbiotic bacteria [5] and Taumatins in nematodes [6]. The other cases have not been described in the literature before.

In conclusion: regression analysis can be a useful tool to discover candidate sequences with unusual phylogenetic histories. It appears that HGT is not uncommon in both prokaryotic and eukaryotic genomes. One important limitation of our approach is that exhaustive calculations are involved, leading to a relatively small number of genomes we can practically use. Due to these limitations, some cases of HGT could remain unidentified.

Acknowledgments

This work was supported by RFBR grant № 15-04-06113.

PFAM Domain	P-value	Species
Thaumatin	P < E-15	*Caenorhabditis elegans*
Ank_2	4,55E-15	*Wolbachia endosymbiont of Culex quinquefasciatus*
Ank_5	4,57E-14	*Wolbachia endosymbiont of Culex quinquefasciatus*
Ank_4	2,05E-12	*Wolbachia endosymbiont of Culex quinquefasciatus*
Transposase_mut	2,36E-11	*Wolbachia endosymbiont of Culex quinquefasciatus*
MULE	2,72E-10	*Wolbachia endosymbiont of Culex quinquefasciatus*
Ank_5	4,39E-10	*Wolbachia endosymbiont of Drosophila simulans*
Ank_4	9,95E-10	*Wolbachia endosymbiont of Drosophila simulans*
Ank_2	4,30E-09	*Wolbachia endosymbiont of Drosophila simulans*
Proton_antipo_M	4,62E-09	*Fusarium oxysporum*
HMGL-like	2,01E-08	*Allomyces macrogynus*
MGAT2	3,00E-08	*Belgica antarctica*
MaoC_dehydratas	3,44E-08	*Archaeon Loki Lokiarch*
PALP	4,15E-08	*Caenorhabditis Elegans*
Beta_elim_lyase	4,42E-08	*Oikopleura dioica*
Epimerase_2	8,21E-08	*Capsaspora owczarzaki*
Actin	1,23E-07	*Archaeon Loki Lokiarch*
EF-hand_6	1,45E-07	*Culex pipiens*
CBFD_NFYB_HMF	1,45E-07	*Culex pipiens*
Thaumatin	1,85E-07	*Tribolium castaneum*
Thiolase_C	2,54E-07	*Culex pipiens*
Ank	5,07E-07	*Wolbachia endosymbiont of Drosophila simulans*

Table 1. PFAM domains that are overrepresented in “outside multiple” sequences in the 61 genomes.

1. H. Ochman et al (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405(6784):299-304.

2. S. Soucy S et al (2015) Horizontal gene transfer: building the web of life. Nat Rev Genet, 16(8):472-482.

3. A. Crisp et al (2015) Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. Genome Biol, 16:50.

4. C. Ku et al (2015) Endosymbiotic origin and differential loss of eukaryotic genes. Nature, 524(7566):427-432.

5. K. Jernigan, S. Bordenstein (2014): Ankyrin domains across the Tree of Life. PeerJ, 2:e264.

6. A. Brandazza et al (2004): Plant stress proteins of the thaumatin-like family discovered in animals. FEBS Lett, 572(1-3):3-7.