

# A Distance-Based Method for Detecting Horizontal Gene Transfer in Whole Genomes

Xintao Wei<sup>1</sup>, Lenore Cowen<sup>1</sup>, Carla Brodley<sup>1</sup>,  
Arthur Brady<sup>1</sup>, D. Sculley<sup>1</sup>, Donna K. Slonim<sup>1,2</sup>

<sup>1</sup> Department of Computer Science, Tufts University, 161 College Ave., Medford, MA 02155; <sup>2</sup> Department of Pathology, Tufts University School of Medicine, 245 Harrison Ave., Boston, MA 02111

{xwei0a,cowen,brodley,abrady,dsculley,slonim}@cs.tufts.edu

**Abstract.** As the number of sequenced genomes has grown, we have become increasingly aware of the impact of horizontal gene transfer on our understanding of genome evolution. Methods for detecting horizontal gene transfer from sequence abound. Among the most accurate are methods based on phylogenetic tree inference, but even these can perform poorly in some cases, such as when multiple trees fit the data equally well. In addition, they tend to be computationally intensive, making them poorly suited to genomic-scale applications. We introduce a new method for detecting horizontal transfer that incorporates the *distances* typically used by phylogeny-based methods, rather than the trees themselves. We demonstrate that the distance method is scalable and that it performs well precisely in cases where phylogenetic approaches struggle. We conclude that a distance-based approach may be a valuable addition to the set of tools currently available for identifying horizontal gene transfer.

## 1. Introduction

Horizontal or lateral gene transfer, the transfer of genes between genomes rather than by “vertical” inheritance from ancestors, has been known to occur among prokaryotes for many years (Davies 1996) and is increasingly of interest in eukaryotes as well (Doolittle 1998; Hotopp et al. 2007). Horizontally acquired genes can affect how we develop and interpret sequence and functional annotation. The extent and sources of horizontal gene transfer (HGT) in an organism may even affect our ability to reconstruct the entire organism’s evolutionary history (Doolittle 1999). A wide range of algorithms for identifying horizontal gene transfer have been suggested, from sequence composition methods to homology searching to phylogenetic approaches.

Sequence composition methods (Mrazek and Karlin 1999) rely on the observation that sequences transferred from a distant genome retain some of the codon and sequence bias of the original organism, which they lose over time (Lawrence and Ochman 1997). These are among the most efficient and scalable approaches to HGT detection, but they can fail in two important cases: when the transfers are ancient or when they are among sufficiently similar species.

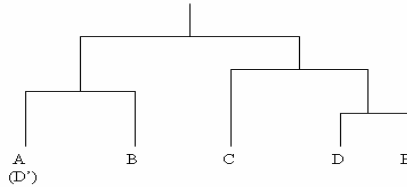
Homology methods, in which conclusions are drawn from the species distributions of the genes' closest neighbors, also scale to whole genomes (Lander et al. 2001; Po-dell and Gaasterland 2007), but annotation errors, incomplete databases, and gene loss raise serious questions about the accuracy of such methods (Salzberg et al. 2001).

In many cases, the best-performing algorithms use phylogenetic approaches to re-construct the evolutionary histories of genomes and individual genes (Eisen 2000). A number of such "tree-based" approaches have been considered, most of which compare inferred trees for individual genes to a "correct" tree showing the overall phylogenetic relationships of the considered species (Robinson 1981; Shimodaira 2002). Such methods are the only ones that incorporate putative evolutionary relationships. Bottom-up tree construction methods, such as the neighbor-joining algorithm (Saitou and Nei 1987), often identify fine structure successfully and so perform relatively well at identifying transfers even between similar species.

However, even tree-based approaches are imperfect. First, they generally require construction of a phylogenetic tree for each gene under consideration. Thus, they are slow and tend to scale poorly to genome-wide applications. In addition, inference of correct phylogenetic trees is a difficult problem, and inferred gene trees can be incorrect, particularly when lineages evolve at different rates (Anderson and Swofford 2004). There are two commonly-used approaches to building the "consensus" tree needed by typical tree-based methods: inferring a phylogenetic tree for each gene and then constructing a consensus of these trees; or else concatenating all the gene sequences together and inferring a tree representing these concatenated sequences. In either case, an incorrect consensus tree will cause errors in the entire algorithm.

We introduce an approach that has many of the strengths of phylogenetic approaches but avoids some of their pitfalls. Specifically, we use the same pairwise distances used by phylogenetic inference algorithms to detect horizontal transfer without building the trees themselves. Since determining the optimal tree topology is the most computationally-intensive part of the tree-based HGT-detection process, a distance-based approach runs much more quickly, allowing scanning of whole genomes. Furthermore, there is no "consensus" tree, so this method doesn't suffer in cases where no single tree that fits all of the data well. Instead, we consider how the *relative* pairwise distances between species in one gene family relate to those relative distances in another. Thus, our method can accommodate genes with different rates of evolution and genes that appear in different sets of species. Because it relies only on the pairwise species distances, we refer to this as the Distance Method.

As an example, consider the tree in Figure 1 showing five species (labeled A – E). One might expect that for most genes, sequence-derived distances between orthologs in D and E would be small, while distances between orthologs in A and D would be larger. However, suppose that a gene from D had relatively recently transferred into A's genome. Then the sequence-derived distance between that gene in D and its closest ortholog in A would be surprisingly small, while the distance between the orthologs in A and B would be surprisingly large. We detect these unusual events using these distances, avoiding the hazards of errors that can be introduced in the tree-construction process and the computational cost of building the trees. Our method compares the pairwise species distances among different gene families and reports the number of unusual-looking distances detected in each family.



**Fig. 1.** A hypothetical tree of five species, A-E. Note that if a gene from D had been transferred into A’s genome, the distances between that gene (D’) in A and E would be surprisingly small, while that between A and B would be surprisingly large. Our method detects HGT by observing these differences.

## 2. Methods

This section first introduces the computational method used to identify horizontally transferred genes. Next, it describes the construction of test data sets used to evaluate the computational approach.

### 2.1 Identification of Horizontally Transferred Genes

Given a species in which we want to find HGT, we start by identifying a set of related species for comparison. In the experiments described here, we followed the example of (Lerat et al. 2003) in selecting *E. coli K12* as our target genome, and a commonly-studied set of 12 other gamma-Proteobacteria (see Table 1) as additional species for comparison. We aim to identify genes from *E. coli* whose evolutionary history *with respect to the other species in our data set* is unusual.

The basic assumption behind our algorithm is that, for a given pair of species, the sequence-derived distances between any two orthologous genes in those two species should be similarly *ranked*, when compared to the distances between other members of the same gene family in other species pairs. Note that this assumption allows for variation in the evolutionary rates of genes.

For example, in Figure 1, for any gene with orthologs in all five species we expect the corresponding sequences in species A and B to be closer than those in A and D. If they are not, it suggests that the evolutionary history of that gene may be atypical. Specifically, if a gene has been recently transferred from another species (whether among those in the data set or outside it), we expect these distance ranks to be unusual for *many* species pairs. Our algorithm identifies such genes. We refer to these HGT candidates as “outlier” genes because of their unusual distance distributions.

**Computing Pairwise Species Distances.** For each gene in the target genome, we identify all orthologous genes in each other species in the data set using BLASTP (Altschul et al. 1997; Schaffer et al. 2001) with an E-value below  $10^{-20}$ . For simplicity, we use the single best BLAST hit in each species to identify orthologous genes, though ultimately more sophisticated approaches may be valuable (Remm et al. 2001;

Podell and Gaasterland 2007). For each gene having at least three detectable orthologs in species other than the target genome, we then construct multiple sequence alignments for the gene family using ClustalW (Thompson et al. 1994). Given these alignments, the `protdist` function in PHYLIP (Felsenstein 2002) calculates the distances between each pair of sequences in the alignment.

We note that multiple sequence alignments can be unreliable, just as inferred phylogenetic trees can, so the distances produced by PHYLIP may be incorrect. However, our method does not need to resolve inconsistencies among the distances by choosing a single tree. Thus, it may be less sensitive than phylogenetic methods to errors or inconsistencies in the inferred distances.

**Detecting “Outlier” Genes.** Different genes evolve at different rates. If we were to rely on raw distances to identify genes whose evolutionary history appears unusual, genes evolving particularly quickly or slowly would be at the top of the list. To avoid this effect, we first normalize the distance data. Specifically, for each gene family and species, we  $z$ -score normalize the set of pairwise distances between the gene in that species and all other species (in which a unique best ortholog for that gene is detectable). For example, in a data set of 190 genes in 13 species (as in Dataset 1, below), we would  $z$ -score normalize 2470 sets (corresponding to 190 genes \* 13 species) of 12 distances each (corresponding to the other orthologs of that gene).

To pick our “outliers” we create a distance vector for each pair of species; in the same example, there would be 13 \* 12 distance vectors, each of length 190 (corresponding to the total number of *E. coli* genes). For each pair of species, we then compute the mean and standard deviation of the values in *that distance vector*, and we identify as outliers any genes that are more than  $c$  standard deviations from the mean.<sup>1</sup> Then, we count up how many of these flagged outliers over all species-pair vectors belong to the same gene. A gene that is flagged as an outlier in this way in more than half the species pairs that include species  $S$  is considered an *outlier gene* for species  $S$ . We then consider species  $S$  an *outlier species* for that gene. Genes with one or more outlier species are reported as having an unusual evolutionary history.

**Clustering Gene Families by Species.** Though normalization is necessary to account for different rates of evolution, an unwelcome side-effect of normalization is that genes existing in only a small number of species are more likely to be chosen as outliers. However, to be applicable on a genome-wide level, our approach must be able to handle genes with detectable orthologs in only a small number of species. This missing-data bias disappears when comparisons are made among sets of genes occurring in roughly similar sets of species. Thus, we pre-process our data by clustering the *E. coli* genes according to the sets of species in which unique orthologs are identifiable. We call this procedure the *Hamming Distance Clustering* step.

To start, we define a *species set* as a set of genes whose orthologs are detectable (by the BLAST method described above) in exactly the same subset of the considered

---

<sup>1</sup> For all experiments in this paper, we choose  $c = 2.326$ , which would correspond to about 2% of the data in each vector if the distances are normally distributed. In practice, they are not, but the top half of the data – that part not constrained by the fact that distances must be non-negative – is close.

species. We call a species set *large* if it contains more than 30 genes, and we assume there are  $k$  species in our data set. Initially, each large species set becomes the core of its own cluster. We now extend these clusters to include the rest of the genes. We do this by an iterative process.

First, for each existing cluster  $C$  in decreasing order by size, let  $v_C$  be a binary vector of length  $k$  indicating the species in which the genes in  $C$  appear. Now, consider in turn each species set  $S$  not already clustered, and create binary indicator vector  $v_S$  for set  $S$ . If the Hamming distance between  $v_C$  and  $v_S$  is at most 2, merge  $S$  into cluster  $C$  (without changing  $v_C$ ). When all  $S$  have been considered, we move on to the next core cluster and repeat the process. Finally, any remaining species sets are assigned to the cluster with the closest core Hamming distance. Once this pre-processing step has been completed, we run our outlier detection algorithm on each cluster and report any genes flagged as outliers in any cluster.

## 2.2 Construction of Test Data Sets

Here we describe the data sets we used to evaluate our approach. We started by downloading thirteen completed gamma-Proteobacteria genomes from the NCBI Genome database in November, 2006. Only the encoded protein sequences were used in this project. Table 1 summarizes the data from these 13 species, which are exactly those chosen by (Lerat et al. 2003). We then constructed three different data sets, which are summarized in Table 2.

**Table 1.** The thirteen gamma-Proteobacterial genomes from which our test data sets were constructed.

Species	Abbrev.	Genome ID	# of proteins
Buchnera aphidicola APS	BA	NC_002528	564
Escherichia coli K12	EC	NC_000913	4,243
Haemophilus influenzae rd	HI	NC_000907	1,657
Pseudomonas aeruginosa PAO1	PA	NC_002516	5,566
Pasteurella multocida Pm70	PM	NC_002663	2,015
Salmonella typhimurium LT2	ST	NC_003197	4,425
Vibrio cholerae	VC	NC_002505,NC_002506	3,835
Wigglesworthia brevialpilis	WB	NC_004344	611
Xanthomonas axonopodis	XA	NC_003919	4,312
Xanthomonas campestris	XC	NC_003902	4,181
Xylella fastidiosa 9a5c	XF	NC_002488	2,766
Yersinia pestis CO92	YC	NC_003143	3,885
Yersinia pestis KIM	YK	NC_004088	4,086

**Dataset 1: Comparison with Lerat’s HGT Method.** The first dataset was designed to determine whether we could identify the same cases of horizontal gene transfer (bioB and mivN) as the consensus-tree approach described in (Lerat et al. 2003). Of the 205 genes in their data set, we were able to identify 189 of them in our database (presumably because the *E. coli* genome annotation has changed somewhat in the intervening years). In fact, only 168 of the 189 genes had orthologs detected by our criteria in all of the 13 species considered. (This is because of differences between their

methods and ours for identifying orthologs.) Given these differences, we therefore added in one other known example of horizontal gene transfer, the *tadA* gene (Planet 2006). In total, Dataset 1 contains 190 genes.

**Dataset 2: Calculating Sensitivity.** This dataset is designed to test the sensitivity of our method. The problem with calculating the sensitivity, specificity, or indeed any measure of accuracy of an HGT detection method is that, for most real data, the right answers are unknown. Specifically, it is impossible to identify the line between true positives and false positives. However, we can take advantage of an idea of Poptsova and Gogarten (Poptsova and Gogarten 2007) to create a small subset of data where we know that the evolutionary history of some specific genes is abnormal, because we’ve “spiked” in those abnormal sequences ourselves.

In this data set, we restricted our attention to genes that were best reciprocal BLAST hits between each pair of species. Thus, only 148 genes from Dataset 1 were selected to form Dataset 2. To simulate horizontal gene transfer between *E. coli* and another species in the data set, we randomly select one of the other species and swap the orthologous gene sequences between *E. coli* and that other species.

In fact, Dataset 2 is really comprised of 10 sub-datasets. Each sub-dataset contains the same 148 genes, but includes 10 different randomly-chosen swapped genes. In total, there are 100 simulated “outlier” genes planted in Dataset 2.

**Dataset 3: A Genomic-Scale Test Case.** There are 4243 *E. coli K12* genes in the genome sequence we downloaded. However, for our distance method to work, we require that genes have more than 3 detectable orthologs among the 13 species. We selected all 2853 *E. coli* genes meeting this criterion, and their orthologs in the other species, to form Dataset 3.

**Table 2.** Summary of test data sets.

	<b>Dataset 1</b>	<b>Dataset 2</b>	<b>Dataset 3</b>
Number of genes	190	148 per sub-dataset	2853
Number of subsets	None	10	None
Known outliers	None	10 per sub-dataset	None
Demonstrates	Feasibility	Sensitivity	Genomic Scale

### 3. Results

#### 3.1 Feasibility

We first ran our algorithm on Dataset 1. The distance method identifies 19 of the 190 genes (10%) as outliers; these are listed in Table 3. Experts disagree on the expected prevalence of horizontal gene transfer in bacterial genomes (Martin 1999), but values between 5 and 15% of the genome are common, so identifying 10% of the input genes in this set seems reasonable. However, because this data set contains only widely-conserved genes, we do not necessarily expect this 10% outlier-detection rate to extend to the whole genome (see Section 3.3).

The 19 gene list includes all three known examples of HGT: *tadA* (Planet 2006) and *mviN* and *bioB* (Lerat et al. 2003). We also note that there are several ribosomal proteins on the list; previous work suggests that horizontal gene transfer is common among ribosomal protein families (Coenye and Vandamme 2005).

Finally, the gene *ileS* appears on this list because of a database error: the *H. influenzae* genome sequence listed in Table 1 lacks the *ileS* (isoleucyl-tRNA synthetase) gene entirely. This is presumably a database error in the NCBI sequence – the gene itself is essential, and the gene appears in other versions of the genome. However, because of this absence, the best BLAST hit of the *E. coli* *ileS* gene in *H. influenzae* turns out to be valyl-tRNA synthetase. Thus, the evolutionary history of the gene appears to the algorithm to be unusual, so this gene is flagged as an outlier. We chose not to correct this error because its presence testifies to the algorithm’s efficacy.

**Table 3.** Genes identified as outliers in Dataset 1. Known examples of HGT and the detected database error are shaded.

rank	# outlier species	# orthologs	Locus	Product name
1	5	5	<i>secE</i>	Translocase
2	4	13	<i>ileS</i>	isoleucyl-tRNA synthetase
2	4	4	<i>rpmD</i>	50S ribosomal protein L30
2	4	4	<i>rpmF</i>	50S ribosomal protein L32
2	4	7	<i>rplO</i>	50S ribosomal protein L15
6	2	13	<i>bioB</i>	Biotin synthase
6	2	13	<i>mviN</i>	Predicted inner membrane protein
6	2	13	<i>ftsZ</i>	cell division protein FtsZ
6	2	9	<i>rplL</i>	50S ribosomal protein L7/L12
6	2	7	<i>rpmG</i>	50S ribosomal protein L33
11	1	13	<i>tadA</i>	tRNA-specific adenosine deaminase
11	1	13	<i>atpD</i>	ATP synthase subunit B
11	1	13	<i>ftsA</i>	cell division protein
11	1	13	<i>gltX</i>	glutamyl-tRNA synthetase
11	1	13	<i>htpX</i>	heat shock protein HtpX
11	1	13	<i>ribA</i>	GTP cyclohydrolase II protein
11	1	8	<i>rplY</i>	50S ribosomal protein L25
11	1	13	<i>rpsJ</i>	30S ribosomal protein S10
11	1	11	<i>yqgF</i>	Holliday junction resolvase-like protein

### 3.2 Sensitivity

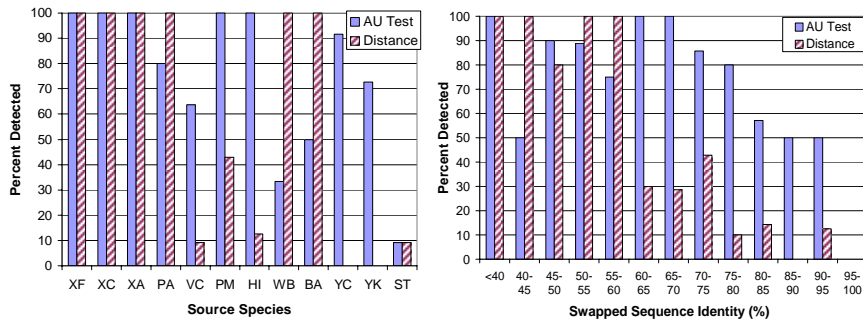
To evaluate the performance of the distance method, we used the simulated anomalies in Dataset 2. We combine the results from each of the ten trials to identify how many of 100 randomly “spiked” anomalous genes we were able to detect. For comparison, we also applied the AU (“approximately unbiased”) test (Shimodaira 2002) to the same data. The AU test is a tree-based method that has been shown to perform well in identifying horizontal gene transfer (Poptsova and Gogarten 2007).

Overall, our distance method did not do as well as the AU test in finding the swapped genes in this data. Only 46 of the 100 swapped genes were identified, com-

pared to 74 under the AU test method. However, a closer analysis of which swaps were found by each method yields some interesting insights.

Figure 2a shows the 100 swapped genes identified by the species with which the *E. coli* representative was swapped. The distance method failed to identify any swaps between *E. coli* and the two *Y. pestis* genomes (YC and YK), which are highly similar to *E. coli*. But it did well on identifying swaps from many other organisms.

The most interesting phenomenon illustrated in Dataset 2 is that the distance method identifies all exchanges between *E. coli* and *B. aphidicola* or *W. brevipalpis*, while the AU test results are much weaker for these genes. These two species are endosymbionts, which are evolving more rapidly than other species in the data set. Their evolutionary relationship to each other and to the rest of the species in the data set is unclear. Some phylogenetic methods suggest that they are closely related to each other (Lerat et al. 2003), but others disagree (van Ham et al. 1997; Spaulding and von Dohlen 1998; Moya et al. 2002). We suspect that for sequences from these species, the tree-based AU test fails because long branch attraction (Anderson and Swofford 2004) creates errors in the consensus tree. However, the distance method does not suffer from this problem at all.



**Fig. 2.** Known outliers detected by the Distance Method and the AU test. a) Breakdown by source species. While the AU test outperforms the distance method overall, this is not the case for all species. In particular, the distance method identifies all spiked sequences deriving from the *B. aphidicola* and *W. brevipalpis* genomes. This is interesting because these symbiotic species are rapidly evolving, so many tree inference methods have trouble placing them correctly. The distance method for detecting outliers avoids this pitfall. b) Breakdown by percent identity of the swapped sequences. Distance outperforms the AU test for dissimilar sequences, but performance of the AU test falls off less dramatically as sequence similarity increases.

Figure 2b shows the Dataset 2 results broken down by the degree of sequence identity between the swapped genes. These results demonstrate that the distance method does well in identifying swapped sequences with only moderate sequence similarity, even in cases where tree inference methods struggle, but it has trouble when the spiked sequences are too similar to those of the host genome.

It is possible that the distance method preferentially identifies rapidly-evolving genes, despite our attempts to account for this via normalization. To eliminate this possibility, we examined the “outlier species” for all 19 genes flagged as outliers in Dataset 1 (i.e., without swapped-in genes). Not a single gene was considered to be an outlier in *B. aphidicola* or *W. brevipalpis*. This is because the variance of the pair-

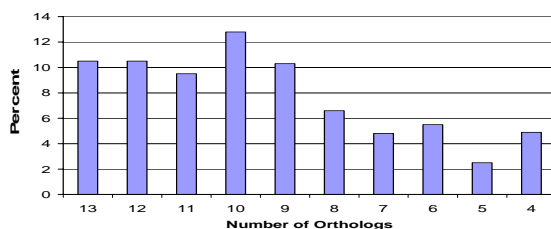
wise distances is large for these species, so we don't identify their genes as being unusually far from the mean. Thus,  $z$ -score normalization appears to be effective, and our ability to detect transfers with these species in Dataset 2 is not an artifact.

Finally, we measured the running times of the two methods on a typical run of Dataset 2 (one set of 148 genes). Both methods use BLAST and ClustalW as pre-processing steps, which took ~5.75 minutes on our 2.4 GHz linux machine. Building the trees in PAML and running the AU test code in CONSEL required 46.5 min., not including time needed to construct a consensus tree (already known for the 13 species involved). In contrast, the distance method required 2.5 min. to calculate pairwise distances in PHYLIP for all pairs of species in all gene families, and then a total of 0.41 seconds to identify outliers in all 148 genes.

### 3.3 A Genome-Scale Application

We ran the Distance Method on Dataset 3 to identify efficacy across a genomic-scale data set. A total of 214 genes (7.5%) were detected as outliers. The full list is available as supplementary data. Figure 3 shows that the probability of a gene's being detected as an outlier is slightly lower for genes detected in few species. This makes sense, because if the sequences exist in fewer species, there are fewer species pairs available to witness the unusual history for that gene. In addition, however, this result demonstrates that the clustering approach successfully overcomes any normalization-induced bias towards selecting genes that appear in few species.

The entire run took 110 minutes on our linux machine; 28 minutes of that was needed to compute distances in PHYLIP, and just under 6 seconds to identify outliers. In other words, the part of the distance method after the pre-processing step of identifying orthologs and aligning them (shared with the AU test) took less real time for the *entire genome* than the unique AU test calculations did for just 148 genes.



**Fig. 3.** Percentage of the 214 outliers from Dataset 3 with detectable orthologs in different numbers of species.

To assess accuracy in this data set, we did not compare our results to tree-based methods, since none to our knowledge is suitable for genome-wide scanning. However, a newly-published method for re-ranking BLAST results has been proposed as a way to find previously-undetected HGT events on a genomic scale (Podell and Gaasterland 2007). We compare our results to theirs. In addition, we can search the literature for validation of our findings, though this is a labor-intensive process.

DarkHorse (Podell and Gaasterland 2007) identifies putative horizontal gene transfer using BLAST to detect closest neighbors, but extending attention beyond the sin-

gle best BLAST hit. Their method has been shown to be applicable on a genomic scale and more sensitive than traditional BLAST searching, and it has already been tested on the *E. coli* genome. We compared our results to those reported as supplemental data in the DarkHorse paper.

Because genes in Dataset 3 must have more than three detectable orthologs among the 13 species in our data set, many of the *E. coli* genes that DarkHorse predicts to be horizontally transferred are not included in Dataset 3. However, of the 2853 genes in Dataset 3, DarkHorse predicts that 31 of them are examples of HGT between *E. coli* and another species. Among our list of predicted outliers, we have only 7 in common with this list of 31: *ygfK*, *ygfO*, *ydcU*, *yjhH*, *yjhG*, *yagE*, and *paaH*.

This result raises two questions. First, how likely is it that we would find that many overlapping genes just by chance? To address this question, we chose 100 random sets of 214 genes from Dataset 3, and measured their intersection with the 31 genes in the DarkHorse list. In none of those 100 cases did we ever see seven intersecting genes, and in only one case did we even see as many as six.

Second, in the cases where the two methods disagree, which is correct? We offer no dispute of the DarkHorse predictions, except the general observation that different evolutionary rates, gene loss, and sequence annotation errors are known to limit the accuracy of homology-based methods (Eisen 2000). However, we manually searched for publications linking 60 of our predicted outliers to horizontal transfer between *E. coli* and another species. We found such evidence in 5 of the 60 cases: *trkG* (Ly et al. 2004), *fliA* and *fliS* (Ren et al. 2005), *agaV* (Charbit and Autret 1998), and *cmtA* (Sprenger 1993). These data suggest that many of our novel predictions may be correct, and that a method that combines multiple approaches might be the best one.

## 4. Conclusions

Our results demonstrate the potential of using distances to detect HGT instead of full phylogenetic methods. The Distance Method described here identifies many known positive examples, including some missed by other methods, but also appears to miss some that other methods detect. Specifically, the Distance Method does particularly well at identifying outlier sequences with only moderate sequence similarity to the host gene, even in cases (such as rapidly evolving symbiotic organisms) where tree-inference methods often fail. On the other hand, the Distance Method struggles to detect transfers between closely related genomes. These transfers are challenging for any HGT method, but the AU Test outperforms the Distance Method here.

These results suggest that, if there were a fast (genomic-scale) tree-based method with accuracy similar to that of the AU Test, the best solution would be to combine that method with the Distance Method. We consider these initial results promising, and we expect that further development of such approaches will yield a scalable HGT-detection method with high accuracy and speed.

This work also has implications for another problem beyond that of HGT-detection: the detection of unnatural genes in the environment. Genetically-modified genomes may appear in the environment by accident, such as when genetically-modified organisms escape containment (Warwick et al. 2007), or by design, such as

the malicious engineering of pathogenic organisms. We are interested in ways to identify signs of such “unnatural” DNA by sequence analysis. If we can reliably find genes that appear to have been derived from a foreign source, content-based methods may help us infer whether the transfer was recent or ancient (Lawrence and Ochman 1997), and functional analysis may suggest whether the transfer occurred naturally or with human intervention. Thus, a distance-based approach to identifying atypical lineages may prove to be a powerful, scalable tool for finding unnatural DNA.

### Acknowledgements

The authors gratefully acknowledge the support of AFOSR/DARPA seedling grant FA9550-06-1-0478. DS and XW were supported in part by NIH grant 1R21LM009411; AB and LC were supported in part by NSF grant (ASE+NHS)(dms)0428715. Thanks to Jonathan Eisen and Sourav Chatterji for extremely valuable discussions.

### References

- Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res* **25**(17): 3389-402.
- Anderson, F. E. and D. L. Swofford (2004). "Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA." *Mol Phylogenet Evol* **33**(2): 440-51.
- Charbit, A. and N. Autret (1998). "Horizontal transfer of chromosomal DNA between the marine bacterium *Vibrio furnissii* and *Escherichia coli* revealed by sequence analysis." *Microb Comp Genomics* **3**(2): 119-32.
- Coenye, T. and P. Vandamme (2005). "Organisation of the S10, spc and alpha ribosomal protein gene clusters in prokaryotic genomes." *FEMS Microbiol Lett* **242**(1): 117-26.
- Davies, J. (1996). "Origins and evolution of antibiotic resistance." *Microbiologia* **12**(1): 9-16.
- Doolittle, W. F. (1998). "You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes." *Trends Genet* **14**(8): 307-11.
- Doolittle, W. F. (1999). "Lateral genomics." *Trends Cell Biol* **9**(12): M5-8.
- Eisen, J. A. (2000). "Horizontal gene transfer among microbial genomes: new insights from complete genome analysis." *Curr Opin Genet Dev* **10**(6): 606-11.
- Felsenstein, J. (2002). PHYLIP (Phylogeny Inference Package), version 3.66. Department of Genetics, University of Washington, Seattle, Washington.
- Hotopp, J. C., M. E. Clark, D. C. Oliveira, J. M. Foster, P. Fischer, M. C. Torres, J. D. Giebel, N. Kumar, N. Ishmael, S. Wang, et al. (2007). "Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes." *Science* **317**(5845): 1753-6.
- Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. (2001). "Initial sequencing and analysis of the human genome." *Nature* **409**(6822): 860-921.
- Lawrence, J. G. and H. Ochman (1997). "Amelioration of bacterial genomes: rates of change and exchange." *J Mol Evol* **44**(4): 383-97.
- Lerat, E., V. Daubin and N. A. Moran (2003). "From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria." *PLoS Biol* **1**(1): E19.

- Ly, A., J. Henderson, A. Lu, D. E. Culham and J. M. Wood (2004). "Osmoregulatory systems of *Escherichia coli*: identification of betaine-carnitine-choline transporter family member BetU and distributions of betU and trkG among pathogenic and nonpathogenic isolates." J Bacteriol **186**(2): 296-306.
- Martin, W. (1999). "Mosaic bacterial chromosomes: a challenge en route to a tree of genomes." Bioessays **21**(2): 99-104.
- Moya, A., A. Latorre, B. Sabater-Munoz and F. J. Silva (2002). "Comparative molecular evolution of primary (*Buchnera*) and secondary symbionts of aphids based on two protein-coding genes." J Mol Evol **55**(2): 127-37.
- Mrazek, J. and S. Karlin (1999). "Detecting alien genes in bacterial genomes." Ann N Y Acad Sci **870**: 314-29.
- Planet, P. J. (2006). "Tree disagreement: measuring and testing incongruence in phylogenies." J Biomed Inform **39**(1): 86-102.
- Podell, S. and T. Gaasterland (2007). "DarkHorse: a method for genome-wide prediction of horizontal gene transfer." Genome Biol **8**(2): R16.
- Poptsova, M. S. and J. P. Gogarten (2007). "The power of phylogenetic approaches to detect horizontally transferred genes." BMC Evol Biol **7**: 45.
- Remm, M., C. E. Storm and E. L. Sonnhammer (2001). "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons." J Mol Biol **314**(5): 1041-52.
- Ren, C. P., S. A. Beatson, J. Parkhill and M. J. Pallen (2005). "The Flag-2 locus, an ancestral gene cluster, is potentially associated with a novel flagellar system from *Escherichia coli*." J Bacteriol **187**(4): 1430-40.
- Robinson, D. a. F., LR (1981). "Comparison of phylogenetic trees." Mathematical Biosciences **53**: 131-147.
- Saitou, N. and M. Nei (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." Mol Biol Evol **4**(4): 406-25.
- Salzberg, S. L., O. White, J. Peterson and J. A. Eisen (2001). "Microbial genes in the human genome: lateral transfer or gene loss?" Science **292**(5523): 1903-6.
- Schaffer, A. A., L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin and S. F. Altschul (2001). "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements." Nucleic Acids Res **29**(14): 2994-3005.
- Shimodaira, H. (2002). "An approximately unbiased test of phylogenetic tree selection." Syst Biol **51**(3): 492-508.
- Spaulding, A. W. and C. D. von Dohlen (1998). "Phylogenetic characterization and molecular evolution of bacterial endosymbionts in psyllids (Hemiptera: Sternorrhyncha)." Mol Biol Evol **15**(11): 1506-13.
- Sprenger, G. A. (1993). "Two open reading frames adjacent to the *Escherichia coli* K-12 transketolase (tkt) gene show high similarity to the mannitol phosphotransferase system enzymes from *Escherichia coli* and various gram-positive bacteria." Biochim Biophys Acta **1158**(1): 103-6.
- Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-80.
- van Ham, R. C., A. Moya and A. Latorre (1997). "Putative evolutionary origin of plasmids carrying the genes involved in leucine biosynthesis in *Buchnera aphidicola* (endosymbiont of aphids)." J Bacteriol **179**(15): 4768-77.
- Warwick, S. I., A. Legere, M. J. Simard and T. James (2007). "Do escaped transgenes persist in nature? The case of an herbicide resistance transgene in a weedy *Brassica rapa* population." Mol Ecol.