Stacy Evans
BIOC 218
Computational Molecular Biology
Final Project
Due June 6, 2001

Horizontal Gene Transfer: Effect and Affect on Computational Methods

1- Introduction

Through the efforts of the human genome project, genes horizontally and laterally gene transferred have been implicated as the source of bacterial protein homologs in the human genome. Additionally, similar protein homologs have been identified in archeal and bacterial species. Controversy continually surrounds the estimations of numbers of horizontally transferred genes, the time and distance between transfer events, the identification of horizontally transferred genes, and the impact of horizontally transferred genes on computational methods designed to identify them. I will discuss and critique computational methods used to identify horizontally transferred genes, including similarity and phylogeny comparisons. Using literature examples I will define current methods, and problems challenging those current methods of identification.

As a note, although lateral and horizontal gene transfer appear to be used interchangeably in the literature, horizontal gene transfer defined genes transferred interspecies, while those transferred intraspecies are laterally transferred. I will refer to horizontally transferred gene (HT genes) throughout this paper.

2- Identification of probable HT genes: Computational choices

Several researchers question the ability of computational methods to determine HT genes. For instance, original work conducted in 2001 by the human genome sequencing effort, compared genes identified in the human genome with bacterial genomes. The comparisons were completed using similarity-sequencing methods. Invertebrate and bacterial genomes were searched and compared to the human genome. Any protein sequences not identified in the invertebrate genome, but identified in bacterial genomes were considered a possible remnant of a HT gene from bacterial species to human genome (Andersson et al. 2001). However, several researchers claimed the sequencing effort had overestimated the possible number of gene transfers due to the computational methods used, and their underlying biological underpinning.

Salzberg et al. (2001) directed the discussion, first citing biological argument. Extensive gene loss has been documented for a number of species, including specific examples of invertebrates used in the analysis. Genes not evident in the invertebrate pool, but identified in the bacterial and vertebrate genomes could represent gene loss from invertebrates, rather than HT gene. Second, the authors mathematical compared the sample size of invertebrates. The number of sample genomes from invertebrates was low (n= 5). When sample sets were removed from the comparison analysis successively, the number if HT genes estimated decreased. This implied that as more invertebrate genomes are sequenced, the amount of estimated HT genes will decrease. Third, the invertebrates and vertebrates compared were “crown” species, and only represented a small percentage of the diversity of species, instilling a computational bias of the data into the analysis. The major computational criticism results from the use of similarity searches: they do not accurately describe evolutionarily relatedness and have no inherent ability to challenge rate variations between genes. Non-essential genes might experience more rapid rates of mutation and would allow for differentiation of HT genes. This argument will be discussed more thoroughly in Section 5.0 below. The authors tried to reduce the statistical threshold for genes under consideration, and sidestep the question of evolutionary relationships. However, their statistical change still does not account for the changing rates of mutations needed to decipher the origins of horizontal gene transfer.

In a criticism of the Salzberg et al. paper, Andersson et al. (2001) argued against the findings. The statistical differences in a similarity search can never truly distinguish a gene that has been horizontally transferred from bacteria to a vertebrate animal. Instead the only approach that can distinguish a horizontally transferred gene are phylogeny trees. When a vertebrate gene is nested within a significant branch of bacterial sequences, this could indicate the transfer of gene material across kingdom boundaries. Out of seven genes speculated by Salzberg et al. as examples of horizontally transferred genes, only one (N-acetylneuaminate lyase) was shown to be clustered, and a likely candidate of horizontal transfer (Andersson et al. 2001).

Phylogeny clustering, as hypothesized by Andersson et al. is a solid computational method to develop ideas of which genes have been transferred. However several obstacles still obstruct computational determination of HT genes. Phylogeny analysis also needs a robust sequence dataset for a completed analysis. As shown by Salzburg et al. when compared across kingdoms, analysis must include a wide variety of species and families, and a large sampling population to minimize effects of gene loss, and deter statistical anomalies in the results. Additionally, other questions of homology inherent in phylogeny analysis determine its applicability to identifying HT genes.

3- Phylogenic analysis: Distance methods

Neighbor joining (NJ), and unweighted pair group method with arithmetic mean (UPGMA) are two distance methods for determining phylogeny. Both examine how gene families might have changed over the course of evolution (Mount 2001). In comparison, the neighbor joining method is used to determine the relationship between sequences that have varying rates of evolution, while the UPGMA method is not influenced by variations in changing rates along branches of the tree. Generally, the methods have several drawbacks. To create a tree, involving initial similarity searches, multiple alignments searches, and eventual analysis, is time consuming, and the correct sequence alignments for the phylogeny are critical. To reduce time by combining and automating of all of the above steps, the sensitivity and quality of the results are reduced (Koonin et al. 2001). Additionally, the NJ method specifically assumes a linear relationship of rate sequence change per unit time, which is not necessary correct (Doolittle et al. 1996[1]).

4- Outside Comparisons

Andersson et al. (2001) hypothesized in a well-supported tree, if an actual HT gene from a bacterial species and a vertebrate species are compared, the eukaryotic gene would be surrounded by bacterial species genes. This indicates the gene is related across kingdom boundaries and the gene is a probable candidate of gene transfer. For example, to determine the relatedness of N-acetylneuaminate lyase, Andersson et al. demonstrated the placement of the *Trichomonas vaginalis* protein within a family of bacterial species in a phylogeny. This identified a probable gene transfer from bacteria to protozoan, and phylogenic evidence for horizontal transfer. Additional evidence was determined through a biological analysis. If a computational phylogeny can determine initial candidates of horizontal transfer the resulting species relationship can be compared with: RNA phylogenies, morphological studies, physiological studies (Nesbo et al. 2001), a computationally derived species tree (analyzed with sequences from many parts of the genome), or a phylogeny interpreted through the fossil record (Doolittle et al. 1996). Together these comparisons can provide strong evidence of HT genes.

5- Sensitivity and specificity studied: Orthologs and paralogs

In a phylogeny, a challenge has been to reliably distinguish orthologs from paralogs. With the input of unidentified HT genes, the challenge becomes even more complicated.

Orthologs are a pair of genes with high identity in alignment, theorized to have arisen from a common ancestor but separated by a speciation event (Mount 2001). On the other hand, a gene that has arisen through a duplication event within a single genome, and has diverged in function from the original common gene is a paralog (Mount 2001). Orthologs and paralogs can be distinguished in several ways.

Orthologs are usually chosen from two species by a similarity search of one species(e.g. BLAST), using a gene from the other species. If both species identify the same gene in the other species as the best statistical match, the genes are most likely orthologous to one another (Mount 2001). This relationship can be further analyzed through a subsequent multiple alignment and phylogeny. Another computational method has been developed to more completely assess the comparisons of orthology. This method, an example given by the GeneTree program, develops a reconciled tree from a proposed ortholog phylogeny and a reference species phylogeny. The algorithm works by minimizing the total number of evolutionary events (gene duplications or gene losses) required reconciling the ortholog tree with the original reference species tree (Yaun et al. 1998). Additionally, libraries of orthologs have been developed that comprehensively describe orthologs in bacteria, yeast and mammals (Remm and Sonnhammer 2000).

A common method to determine paralogs is to search the genome of one species with one gene. A cluster of genes, or proteins, is usually identified thought to have arisen from one ancient gene (Mount 2001). Then a multiple alignment and a phylogeny can access the relationship between identified sequences. The comparison of these methods and relationships can be used to distinguish orthologs from paralogs. However, again comparing with HT genes, identity becomes even more complicated.

For example, to determine the evolutionary time since the branching of different kingdoms, Doolittle et al. (1996[1]) compared the phylogeny of 57 enzyme sequences. They could not guarantee that all sequences used were orthologs, and that none of the sequences had resulted from a HT gene event. They proposed the number of sequences analyzed would render these anomalies statistically insignificant. They cross-checked this by omitting different sequences to determine if they had any effect on the determination of tree length (and in their parsimony analysis, tree length is also representative of the evolutionary time since divergence). Specifically, the authors removed the seven fastest changing sets of sequences, and the seven slowest sets of sequences. They also removed the 27 fastest changing species, and then the 27 slowest changing species. Each one had minimal affect on the resulting tree lengths (biggest difference was 10%). Subsequently as a response to literature discussion, Doolittle et al. (1996[2]) postulated their reasons for choosing the slowest and fastest changing enzymes: that the slower half more likely to contained HT genes, while the faster half more likely contained paralogs. The addition of either paralog, or HT genes into the comparison would, if effect "short circuit" the analysis, and the resulting phylogenic tree.

The inclusion of this example addresses several issues. On a smaller comparison scale, for instance comparing one protein with another in a select group of organisms, the phylogenies might allow a researcher to distinguish a orthologous gene from a HT gene through clustering (see Section 3.0 above). If one were to look at a phylogeny as a mapping tool, if the number of unknown HT genes included in an analysis are small compared with a large sample size, it limits the influence of the HT genes in a phylogeny. However, with a small number of species, or a small number of compared sequenced genes, unidentified HT genes could greatly influence the outcome of a phylogeny analysis.

Additionally, Doolittle et al. (1996[2]) postulated that genes with a higher rate of sequence change per unit time would be paralogs while the genes with a smaller evolutionary rate would most likely be HT genes. There is a biological problem in their computational analysis. For general horizontal transfer, it is theorized as least in some cases when a gene is transferred, the gene will undergo accelerated evolution because there is no initial selective pressure to maintain function (Koonin et al. 2001). But as Doolittle states the evolutionary rate is expected to be small. For instance, it is hypothesized that bacteria will only display evidence of HT gene transfer if it conveys some sort of selective advantage to the host, usually some biological function the host does not possess in it's own genome (probably due to limited resources, e.g. the size of the genome). Since the HT gene was now performing a unique function it would probably remain unchanged, with little to no sequence change. However, if a bacterial gene is incorporated into a vertebrate genome, the selective pressure is not as great because for instance, more resources are available. The gene: a) can; but, b) will not necessarily, convey a selective advantage (i.e. uptake is random). Therefore selective pressure could either: a) stay constant (with little change in rate of sequence change); or b) change because no selective pressure will cause it to remain constant (rate of sequence change will increase). The best method to analyze this b) HT gene would be through the NJ method, which would display the protein as a longer branch. However it is unclear which phylogenic distance method would best describe no increase in rate of sequence change.

To further demonstrate the impacts of horizontally transferred genes on a phylogeny, a hypothetical horizontally transferred gene is compared with a group of well-defined orthologous genes. The resulting phylogeny, as speculated by Doolittle et al. (1996[1]) (and as described above) will have a gene from one species, misaligned with genes from other species. In other words, visually compared with a robust species phylogeny, the HT gene will appear to cluster within a group of unrelated species. Proposed by Doolittle et al., this is the ideal method to identifying a HT gene.

In a well-defined paralogous gene grouping, the HT gene would be harder to distinguish. As stated above, a paralogous gene set is contained within one species, and has evolved and differentiated from one theoretical gene. Therefore, there will be several different homologs of a gene within one species, which can be mapped in a phylogeny to show differences and distances within the evolutionary relationship. The rate of sequence change of a HT gene theoretically can be either high or low, while the paralogs rate of sequence change might be unusually high (because it has little selective pressure to remain similar.) Therefore contrary to Doolittle et al., if both rates are high the HT genes and a group of paralogs might be indistinguishable in a visual representation of a phylogeny containing orthologs. The rates of evolution, and the resulting distance calculation might not provide necessary information. If Doolittle is correct, the paralogs will be evolving quickly, and the HT genes will be evolving more slowly, and the distances on phylogeny will discriminate between the two. The HT gene could be distinguished because it should not look like any of the paralogs and would be distantly related shown be an outstanding long branch.

6- Failings summarized

All of the methods of phylogeny above, would fail in certain circumstances, some of which have been discussed. When a similarity search, multiple alignment, or statistical test fails, the phylogeny cannot determine the true relationship among the genes. When all sequences have not been identified, misrepresentation of the genes can occur. For instance if a true ortholog in a genome has not been sequenced, phylogenic comparisons could be compromised by paralog sequences (Remm and Sonnhammer 2000). If paralog and a HT gene have similar rates of sequence change, they will be indistinguishable in a phylogeny. Each of these errors could limit the specificity and sensitivity of the computational analysis.

7- Conclusions

The computational assessment of genes and genomes is an evolving science. New biological comparisons are being adjusted everyday as the sequences and computational methods are refined. As I have discussed, literature criticism is essential to continue conversation and point researchers in the correct direction, to allow honest interpretations of this emerging science. As more analysis is completed of HT genes, researchers will develop better tools to assess their evaluation which move beyond the flaws of the phylogeny distance method today.

Literature Cited

Andersson JO, WF Doolittle, CL Nesbo. 2001. Genomics. Are there bugs in our genome? Science: 292(5523);1903-6.

Doolittle, RF, D Feng, S Tsang, G Cho, E Little. 1996(1). Determining Divergence Times of the Major Kingdoms of Living Organisms with a Protein Clock. Science: 271; 470-7.

Doolittle, RF, D Feng, S Tsang, G Cho, E Little. 1996(2). Dating the Cenancestor of Organisms; Response. Science: 271; 1751-3.

Koonin, EV, KS Makarova, L Aravind. 2001. Horizontal Gene Transfer in Prokaryotes: Qualification and Classification. Annual Review Microbiology: 55; 709-42.

Mount, D. 2001. Bioinformatics; Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Nesbo, CL, S L'Haridon, KO Stetter, and WF Doolittle. 2001. Phylogenetic Analysis of Two "Archaeal" Genes in Thermotoga maritime Revel Multiple Transfers Between Archae and Bacteria. Molecular Biology Evolution: 18(3);362-375.

Remm, M., E Sonnhammer. 2000. Classification of Transmembrane Protein Families in the *Caenorhabditis elegans* Genome and Identification of Human Orthlogs. Genome Research: 10;1679-1689.

Salzberg, SL, O White, J Peterson, JA Eisen. 2001. Microbial Genes in the Human Genome: Lateral Transfer of Gene Loss? Science: 292; 1903-1906.

Yaun, YP, O Eulenstein, M Vingron, P Bork. 1998. Towards detection of orthologues in sequence databases. Bioinformatics; 14(3); 285-289.