Comparative Genomics by way of Relative Abundance Analysis

Paul Welander
498BIO Assignment #3
November 16, 2001

Dr. Samuel Karlin and colleagues at Stanford University have developed a method for assessing genomic similarities based on relative abundances of short nucleotide chains. The goal of such a method is to eliminate the need for homologous sequences that have been previously aligned by another procedure. The approach taken by Dr. Karlin deviates from previous methods of genomic analysis and evolutionary reconstruction by utilizing information derived from the entire genome rather than from specific subsequences. However, the genomic comparisons are generally in agreement with accepted phylogenies.

Relative abundance methods were applied to genomic sequences acquired from GenBank. Studies were restricted to non-redundant nucleotide sequences of species for which at least 100 kb (thousand base pairs) were available. In many cases, the genomes used contained several contiguous sequences of more than 10 kb. The particular genomes studied in the series of papers reviewed here are fairly comprehensive, including species from all six kingdoms: Protists, Fungi, Animals, Plants, Eubacteria, and Archaebacteria.

The method employed works by assigning relative abundance values to the genomic sequence of a given species. Let f_X denote the frequency of a particular nucleotide X (A, C, G, or T) and f_{XY} the frequency of a dinucleotide XY. A standard method for determining the dinucleotide bias of a specific sequence is to employ the odds ratio,

$$\rho_{XY} = f_{XY}/f_X f_Y \, .$$

The formula ρ_{XY} is then modified to account for the double stranded nature of DNA, such that one can define symmetric frequencies,

$$f^*_X = f^*_{X'} = (f_X + f_{X'})/2 \text{ and } f^*_{XY} = f^*_{Y'X'} = (f_{XY} + f_{Y'X'})/2 \, ,$$

where X and X' are complementary nucleotides (eg. A and T). The modified odds ratio is then,

$$\rho^*_{XY} = \rho^*_{Y'X'} = f^*_{XY}/f^*_X f^*_Y \, .$$

The set of ρ^*_{XY} values for a particular genomic sequence is referred to as the *dinucleotide relative abundance profile*. Finally, one can calculate the *dinucleotide relative abundance distance* between two sequences g and h,

$$\delta^*(g,h) = (1/16) \, \Sigma_{XY} \left| \rho^*_{XY}(g) - \rho^*_{XY}(h) \right| \, ,$$

where the sum extends over all possible nucleotide pairs.

Interpreting the odds ratio can be done as follows. From statistical theory, the dinucleotide relative abundance may be described as suppressed if $\rho^*_{XY} < 0.78$, and over-expressed if $\rho^*_{XY} > 1.23$. With respect to these dinucleotide biases, the following trends have been observed: First, the dinucleotide TA is broadly underrepresented in both prokaryotes and eukaryotes, with typical ρ^*_{TA} values ranging from 0.5 – 0.8. A possible

explanation for this may be the low thermodynamic stacking energy of the TA dinucleotide, the lowest among all possible pairs. Second, CG is the most suppressed dinucleotide in vertebrates with ρ^*_{XY} = 0.23 – 0.37, while at the same time being significantly over-expressed in some bacterial forms.

On the other hand, distance values are utilized to construct genomic relationships between species and can generally be categorized as follows:

Label	**Distance ($\delta^*(g,h)$) Range**	**Example**
Random	0.000 – 0.018	
Very Close	0.020 – 0.030	Pig and Bovine
Close	0.035 – 0.050	Human and Bovine
Moderately Related	0.055 – 0.075	Frog and Mouse
Weakly Related	0.080 – 0.115	Human and Trout
Distantly Related	0.120 – 0.150	Human and Yeast
Distant	0.160 – 0.200	Human and Fruit Fly
Very Distant	> 0.200	Human and *E. coli*

The breadth of the studies performed by Dr. Karlin and colleagues has led to a few surprising results. One such finding was that vertebrate sequences are generally more similar to those of fungi than to either the protists or invertebrates. The genomic distance between fungi and vertebrates, albeit large, is much smaller than the distance between fungi and invertebrates. These results lend credence to particular protein-derived phylogenetic trees.

Other observations include the following: First, δ^*-differences within a single species are, with very few exceptions, much lower than those between species. This reflects a certain level of robustness in the genome signature. In general, the most homogeneous genomes occur among fungi, while protist genomes are much more divergent. Second, δ^*-differences were determined for 15 large human sequences from different chromosomes and known genes. Difference values within the sequences fell in the range 0.013 – 0.046, while those between the sequences varied from 0.020 to 0.081. These differences indicate a strong to moderate relationship between the various human sequences.

A third finding was that the mammals could be split into two groups: rodents and non-rodents. Mutual δ^*-differences within groups were vary close while between groups the differences revealed only a moderate similarity. That insects and protists form very diverse groups was the fourth noteworthy result. Typical δ^*-differences within these groups ranged from 0.07 to 0.12; at best the insects and protists studied are weakly related. On the other hand, fungi constitute a relatively coherent group with differences in the range from 0.035 to 0.075, close to moderately similar. Finally, in the group of plants, the monocots and dicots are only moderately related, while relationships within these classes are very close. All plants studied are closely related to yeast, but only weakly related to other fungi.

In summary, Dr. Karlin's methods attempt to infer genomic relationships on the basis of entire genomes without directly comparing DNA sequences. Comparisons within and between species sample sequences are based on dinucleotide relative

abundance distances, and generally support the use of such a method. Through his analysis of genomes, Dr. Karlin has deduced relationships among a wide range of species. In most cases these findings simply confirm earlier statements, but in a few cases, they reveal similarities that had been previously dismissed or unquestioned. What remains to be seen is how the ignorance of long-range sequence specificity affects the conclusions presented. In a sense, the strength of relative abundance analysis is also its weakness. It seems that without the details of a genomic sequence, relative abundance analysis can at best provide only a rough guide to the relationships among different species. Regardless of this fault, it is evident that relative abundance analysis has become a powerful tool for analyzing the vast amounts of genomic data now available.

References:

Campbell, A., Mrazek, J. and Karlin, S. "Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA." *Proc. Nat. Acad. Sci.* **96**, 9184 (1999).

Karlin, S. and Ladunga, I. "Comparisons of eukaryotic genomic sequences." *Proc. Nat. Acad. Sci.* **91**, 12832 (1994).

Karlin, S. and Mrazek, J. "Compositional differences within and between eukaryotic genomes." *Proc. Nat. Acad. Sci.* **94**, 10227 (1997).