6.047 / 6.878 Computational Biology: Genomes, Networks, Evolution
Fall 2008

Comparative Genomics II: From Genomes to Evolution

Lecture: Manolis Kellis.

November 4, 2008

1 Introduction

This is the second lecture in a series of two lectures on the work being done in Prof. Kellis's lab on comparative genomics. Comparing genomes of related species at different evolutionary distances can teach us both about the genetic code and about evolution. In this lecture we discussed two examples of what we can learn about evolution using comparative genomics.

The topics of this lecture are discussed in greater detail in the papers [1] and [2]. A brief summary is given in these notes with references to figures in lecture slides.

2 Whole genome duplication

Evolution requires infrequent random "errors", i.e. mutations, in cell division to create variation in different species. Genomic duplication is a particular type of such error which can be useful to explain innovations in evolution[3]. Most of the time genomic duplication will lead the daughter cell to be less fit ("sick") and to get selected out. However, if the daughter species survives and adapts, one copy of an original gene can perform its task while the other gene can evolve to gain new function increasing gene content and fitness. In class, we discussed one type of genomic duplication, namely whole genome duplication (WGD), in the study of which comparative genomics has brought in novel information.

2.1 Before comparative genomics

In his 1970 book "Evolution by Gene Duplication" where he postulated the role of genomic duplication in evolution, Ohno also suggested the occurance of whole genome duplications, in particular that the vertebrate genome is the result of one or more whole genome duplications[3]. Such large scale duplications would explain, for example, how there are 4 Hox genes in humans compared to 1 Hox gene in flies.

WGD has been suggested in various other cases but conclusive evidence was not found until 2004. In particular, the possibilty of WGD in the yeast *S. cerevisiae* has been debated since 1997. When *S. cerevisiae* genome was sequenced, large duplication blocks were observed. Wolfe and colleagues suggested that these duplications were due to a WGD[4]. Others have argued that the observed paralogous gene rate of 8% was too small to suggest a WGD and could be explained with independant local duplications[5].

2.2 Comparative evidence for WGD in yeasts

Assume we label a number of neighboring genes in a segment of DNA in order 1-16. Then a WGD occurs and there are two copies of the segment in the genome. As this dual-genome species evolves, some of the redundant genes are lost and some gain new functions such that one of the chromosomes has orthologous genes to gene 1, 3, 4, 6, 9, 10, 12, 13, 14, 16 while the other chromosome has 2, 3, 5, 7, 8, 11, 13, 15. Comparing the two chromosomes we would see paralogues 3 and 13 and we would not be able to tell if these paralogues are from WGD or individual local duplications.

However, if we compare genomes of different species before and after WGD, we would see that a region in the species before WGD will correspond to two regions in the species after duplication. (See slide 5.) Here before and after duplication refer to species that descend from a branch without and with WGD respectively. This is exactly what was observed in Prof. Kellis's lab. Comparative analysis between *S. cerevisiae* and *K. Waltii* gave a less clean signal than the comparison of different *Saccharomyses* species. (See slide 6.) Looking closely at individual matching regions *S. cerevisiae* and *K. Waltii*, the dual match signal that would be expected from a WGD was noticed. (See slide 7.)

During the analysis of the data, sequencing of the *K. Waltii* was completed. It was seen that 16 chromosomes in *S. cerevisiae* corresponds to 8 chromosomes in *K. Waltii*. (See slide 9.) Looking at slide 9, one can see that the correspondence between the chromosomes in the modern species is not perfect. This is due to many chromosome crossing that have happened since the two species branched. (Chromosome crossing is not expected to have any major selective effects so can happen relatively often.)

Another evidence for WGD in the data comes from the positions of chromosome centromeres relative to genes. Assume there is a centromere between genes 6 and 7 in the above example. In the two corresponding *S. cerevisiae* chromosomes, we would expect centromeres to be between labels 6 and 9 and 5 and 7 respectively. This kind of centromere position prediction was seen to hold true in all the chromosomes in the comparison of *S. cerevisiae* and *K. Waltii*. (See slide 10.)

An interesting observation is that WGD event in the yeasts happened approximately 190M years ago. This was also when fruit bearing plants evolved. This was a time when there was an abundence of sugar available in the environment and the first generations of inefficient/sick species of yeasts with too many redundant genes would be able to survive. Also the new functions that would develop with new derived genes could give them useful features needed in these new conditions. (Such as making beer.)

2.3 Post-duplication evolution

K. Waltii has approximately 5000 genes. After genome duplication, the ancestor of *S. cerevisiae* had approximately 10000 genes. Soon after, most of the duplicate genes were lost. *S. cerevisiae* now has 5500 species.

Do the new paralogous genes "share" their old task and evolve at a similar high rate (as proposed by Lynch in 2000)? Or is one better preserved to do the old task while one evolves rapidly to gain new function (as proposed by Ohno in 1970)? Comparing genes in *S. cerevisiae* and *K. Waltii*, it was seen that 95% of the gene pairs showed asymmetric accelaration rates, supporting Ohno's model.

This means we can define "ancesteral" and "derived" functions for the two paralogous genes. Indeed, biological experiments have shown that such a distinction exists. Ancesteral genes are more vital (gene removal is more likely to be lethal), are expressed more abundantly and serve general functions whereas derived genes are used only in specific conditions and in specific tasks and can be removed without causing the organism to die. Also, ancesteral genes have higher network connectivity.

Comparison of *S. cerevisiae* and *S. bayanus* reveals another interesting feature of the evolution of paralogous genes. Both *S. cerevisiae* and *S. bayanus* are derived from the branch where WGD has occured. (Gene ordering data also provides evidence that WGD happened earlier in time compared to species seperation between *S. cerevisiae* and *S. bayanus*.) So the paralogous genes in *S. cerevisiae* have been separated in time longer than orthologous genes between *S. cerevisiae* and *S. bayanus*. However, the paralogous genes are closer in sequence matching than orthologous genes. This observation hints at occurance of gene conversions between paralogues, i.e. when a gene is being duplicated, all or part of it may be replaced by the paralogue.

3 Phylogenomics

In previous lectures we have studies various algorithms to obtain phylogenetic trees for different species. Similar studies can be performed to study phylogeny of orthologous and paralogous genes. These trees contain information from both the evolution of the species and evolution of genes, with gene loss and gene duplication events.

Gene trees can be built for individual genes. Then these genes can be "reconciled" with known species trees to draw complete evolution of the gene. However, accuracy of gene trees built on single genes are usually not very accurate:

- Some gene trees lead to many loss and duplication events which could be explained more simply. See slide 30.

- Gene tree topologies obtained from neighboring genes which evolved together are not robust. See slide 32.
- Simulations of evolution of genes show that the information in single genes may not be enough to get accurate results.

The gene tree accuracy has been shown to be mainly limited by the information from the genes. Trees for longer genes which have higher information content can be reconstructed more successfully. See slide 33. Also, very fast evolving or slow evolving genes carry less information compared to moderately diverging sequences (40-50% sequence identity) which give the best performance. See slide 34. These observations have been made both in Monte Carlo studies and in data from 12 fly species where gene trees were tested for a congruent topolgy to the species tree. Since Monte Carlo and data gave similar results, it was concluded that more information is needed to accurately build gene trees.

It has been observed that mutation rate between two genes in two species can be seperated into a gene-dependant rate and a species-dependant rate. In other words, different gene trees have similar topologies for different species with a gene specific multiplying factor. (See slide 37.) This observation can be used to write a generative model with which different tree topologies can be assigned likelihood values. The species specific substitution rate s_i, which depends on evolutionary dynamics of the species such as population size and generation time and appears to be normal distributed for different genes. The gene specific rate g is constrained by the selective features of it function and is observed to be distributed as a gamma distribution, which can be expected for a rate.

This new information can be used to build more accurate gene trees than previous methods. Assume species tree is known (this can be built using whole genome information.) A training set can be chosen from well known one-to-one orhologies and using congruent trees to the species tree. Using this training set, values of g and s_i can be sampled and the parameters of the corresponding Gamma and Normal distributions ($g \sim G = \Gamma(\alpha, \beta)$, $s_i \sim S_i = N(\mu_i, \sigma_i^2)$) can be inferred by fits.

In the next step, trees are built for remaining genes. A distance matrix M is built for genes in question. Next, trees of different topologies are constructed where the branch lengths b_i are calculated using the matrix M. The likelihood of the topology can be calculated as the probability of observing the branch lengths $b_i = g \times s_i$. The best topology is selected using maximum likelihood.

One example where this method proves better performance compared to previous methods is the gene tree of hemoglobin-β protein shown on slide 40. This method correctly accounts for the fast evolution of the rodent branch.

The method has been also shown to provide accurate results when the gene tree is not congruent to the species tree. Consider the tree containing hemoglobin-α for rat and mouse and hemoglobin-β for dog and human. The correct topology should have dog and human genes seperating after human and rodent since gene duplication happened before relevant speciation events. Indeed the method gives the correct tree containing gene duplication. See Figure 4. in [2] for a more clear description.

In conclusion, observing that gene and species substitution rates can be seperated and calculating species substitution rates, a new gene tree construction algorithms has been developed. This method has been observed to give much better accuracy than previous gene tree construction algorithms.

References

[1] Manolis Kellis, Bruce W. Birren and Eric S. Lander, Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae, Nature 8 April 2004, p617

[2] Matthew D. Rasmussen and Manolis Kellis, Accurate gene-tree reconstruction by learning gene-and species-specific substitution rates across multiple complete genomes, Genome Res. 2007 December; 17(12): 1932-1942 (2007).

[3] Susumu Ohno (1970). Evolution by gene duplication. Springer-Verlag. ISBN 0-04-575015-7.

[4] Wolfe, K. H. and Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387, 708-713 (1997)

[5] Genomic exploration of the hemiascomycetous yeasts: 20. Evolution of gene redundancy compared to Saccharomyces cerevisiae. FEBS Lett. 487, 122-133 (2000)