An Expectation-Maximization Algorithm for Reconstructing Heterogeneous Gene-Content Evolution

Wataru Iwasaki
iwasaki@cb.k.u-tokyo.ac.jp

Toshihisa Takagi
tt@k.u-tokyo.ac.jp

Department of Computational Biology, Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8568, Japan

Keywords: expectation-maximization, gene content, heterogeneous evolution, reconstruction

1 Introduction

Genome projects have determined more than 400 genomes, and over 1,600 projects are still ongoing. The sequence information provides us with opportunities to study the evolution of gene content, i.e., the set of genes an organism has. Most preceding methods for reconstructing the gene-content evolution assume that rates of gene gain/loss are homogeneous over a phylogenetic tree [1]. However, biological observations have suggested that the gene-content evolution is heterogeneous. For instance, parasitization frequently results in a massive gene loss, and genome duplication results in a massive gene gain: both cause the heterogeneous evolution of gene content. A few probabilistic models that account for the heterogeneity have been proposed [2]; nevertheless, effective and efficient reconstruction algorithms applicable to the hundreds of the genomic data have not been developed. Here, we have developed a novel algorithm for the reconstruction of gene-content evolution based on the assumption of its heterogeneity. The algorithm estimates the rates of gene gain/loss effectively and efficiently, using the framework of the expectation-maximization (EM) algorithm.

2 Method

This algorithm requires an input dataset $D$: phylogenetic topology for organisms to be analyzed, and an “ortholog table”, which represents gene-content that each organism possesses, by counting gene numbers belonging to each ortholog group. We formulated gene-content evolution as the finite-state continuous-time Markov process. In the process, there are four states 0, 1, 2, and 3, respectively corresponding to the number of genes in an ortholog group. We ignored those genes that have more than three copies in an ortholog group, because such high-copy-number genes often have special characteristics (e.g. transposons) and give less information on the evolution of gene repertoire.

Transition probabilities from a parent to its child along each edge $n$ of the phylogenetic tree are given by a 4x4 matrix $M_n$, where the $(i,j)$-th element of $M_n$ represents the transition probability from the state $i$ to $j$. We assumed that the ortholog groups evolve independently, as is the case for the typical point-mutation models for analyzing sequence evolution. The matrix $M_n$ is defined by a following differential equation:

$$M_n(p + dp) = (E + R_n dp)M_n(p),$$

where $p$ is a parameter from the parent to the child along each edge ($0 \leq p \leq 1$), $E$ is an 4x4 unit matrix, and $R_n$ is a transition rate matrix defined by $\alpha_n$ and $\beta_n$ that are rates of gain and loss of genes, respectively, per one ortholog along the edge $n$. That is, for example, large $\alpha_n$ indicates massive gene gain on the edge $n$ by gene genesis, horizontal gene transfers, gene duplications, and/or other mechanisms. We did not model evolutionary time ($t_n$) in explicit form, because of the mathematical redundancy and the ambiguity for estimating $t_n$; instead, if given $t_n$, dividing $\alpha_n$ and $\beta_n$ by $t_n$ simply gives the evolutionary rates in time unit. The values $\alpha_n$ and $\beta_n$ will be different among edges, which implies the heterogeneity of the gene-content evolution. By following the above differential equation, we solved the matrix $M_n$ analytically by linear
algebraic procedure and represented $M_n$ by using $\alpha_n$ and $\beta_n$.

Therefore, if we had $\alpha_n$ and $\beta_n$ for all edges (i.e. parameter $\theta$), we could calculate $M_n$ for all edges, and reconstruct the most likely evolutionary history by applying the maximum-likelihood method. However, this is not the case. In order to estimate likely parameter $\theta$, we utilized the framework of the EM algorithm to overcome this difficulty. We calculated $Q(\theta|\phi') = \sum_{h} P(h|D, \theta') \log P(D, h|\theta)$ for the present evolution model, and solved $\arg \max_{\theta} Q(\theta|\phi')$ by the Lagrange multiplier method.

There were two problems upon the implementation. The first problem was computational complexity. To solve this problem, we applied the inside-outside algorithm for syntax analysis of the stochastic context free grammar. By applying this algorithm, the complexity was reduced, so that up to thousands of genomic data can be handled in reasonable time. The other was initial-condition dependence of the EM algorithm. Here, we employed the maximum-parsimony method, which minimizes transitions, to set the initial values.

3 Results and Discussion

We first tested the performance of the above algorithm on simulation datasets. Gene content comprising 20,000 ortholog groups were evolved along the binary phylogenetic trees (each has 16, 32, 64, 128, and 256 leaves, respectively), by following the Markov evolution model. Our algorithm showed better performance than the maximum-parsimony based method, which assumes uniformity of gene-content evolution. It is notable that more improvement was observed for larger datasets, considering the recent expansion of the genomic data.

Then, we applied the algorithm to the actual dataset (179 genomes from all the three domains of life with 26,200 ortholog groups) to reconstruct the gene-content evolution across the universal tree (Figure 1). The estimated parameters varied over the entire tree, indicating the wide-spread heterogeneity of the gene-content evolution. As a general tendency, eukaryotes had higher $\alpha_n$ than prokaryotes. This tendency would reflect difference in survival tactics between the two forms of life: prokaryotes with small genomes have proliferated rapidly, whereas eukaryotes have increased their genome size to acquire various abilities to survive. Among prokaryotes, the gamma-proteobacteria group was characterized by high $\alpha_n$. This result is consistent with previous molecular phylogenetic studies that suggested that the gamma-proteobacteria has acquired many of their genes by horizontal gene transfers [3].

Until today, many mathematical analyses have been performed to understand the evolution of biological systems, assuming the homogeneity of gene-content evolution (e.g. the preferential attachment model for gene-network evolution). Our results suggest that the heterogeneity of gene-content evolution must be considered to understand the evolution of biological systems, which are based on the cooperative function of multiple genes comprised in the gene content.

References

