An Expectation-Maximization Algorithm for Analysis of Evolution of Exon-Intron Structure of Eukaryotic Genes

Liran Carmel, Igor B. Rogozin, Yuri I. Wolf, and Eugene V. Koonin

National Center for Biotechnology Information,
National Library of Medicine,
National Institutes of Health,
Bethesda, Maryland 20894, USA
{carmel, rogozin, wolf, koonin}@ncbi.nlm.nih.gov

Abstract. We propose a detailed model of evolution of exon-intron structure of eukaryotic genes that takes into account gene-specific intron gain and loss rates, branch-specific gain and loss coefficients, invariant sites incapable of intron gain, and rate variability of both gain and loss which is gamma-distributed across sites. We develop an expectation-maximization algorithm to estimate the parameters of this model, and study its performance using simulated data.

1 Introduction

Spliceosomal introns are one of the most prominent idiosyncrasies of eukaryotic genomes. They are scattered all over the eukaryota superkingdom, including, notably, species that are considered basal eukaryotes, such as *Giardia lamblia* [1]. This suggests that evolution of introns is intimately entangled with eukaryotic evolution; thus, the study of evolution of exon-intron structure of eukaryotic genes, apart from being interesting in its own right, might shed some light on the still enigmatic rise of eukaryotes. For example, one of the notorious, long-lasting unresolved issues in evolution of eukaryotic genomes is the intron-early versus intron-late debate. Proponents of the intron-early hypothesis posit that introns were prevalent at the earliest stages of cellular evolution and played a crucial role in the formation of complex genes via the mechanism of exon shuffling [2]. These introns were inherited by early eukaryotes but have been eliminated from prokaryotic genomes as a result of selective pressure for genome streamlining. By contrast, proponents of the intron-late hypothesis hold the view that introns had emerged, de novo, in early eukaryotes, and subsequent evolution of eukaryotes involved extensive insertion of new introns (see, e.g., [3,4]).

Various anecdotal studies have demonstrated certain features of intron evolution. But it was not until the accumulation of genomic information in the recent years that large-scale analyses became feasible. Such analyses yielded at least three different models of intron evolution. One model assumes parsimonious evolution [5]; another assumes a simple gene-specific gain/loss model and

A. McLysaght et al. (Eds.): RECOMB 2005 Ws on Comparative Genomics, LNBI 3678, pp. 35–46, 2005.

analyzes it using Bayesian learning [6]; and yet another one assumes a simple branch-specific gain/loss model on three-species phylogenetic topology and analyzes it using direct maximum likelihood [7]. It seems that none of these models is sufficiently general, and each neglects different aspects of this complex evolutionary process. This is reflected in the major contradictions between the predictions laid out by the three models. For example, the gene-specific model [6] predicts an intron-poor eukaryotic ancestor and a dominating intron gain process; the branch-specific model [7] predicts an intron-rich eukaryotic ancestor and a dominating loss process; while the parsimonious model [5] is somewhat in between, predicting intermediate densities of introns in early eukaryotes, and a gain-dominated kaleidoscope of gain and loss events.

Here, we introduce a model of evolution of exon-intron structure, which is considerably more realistic than previously proposed models. The model accounts for gene-specific intron gain/loss mechanisms, branch-specific gain/loss mechanisms, invariant sites (a fraction of sites that are incapable of intron gain), and rate distribution across sites of both intron-gain and intron-loss. Using data from extant species, we follow the popular approach of estimating the model parameters by way of maximum likelihood. Direct maximization of the likelihood is, however, intractable in this case due to a large number of hidden random variables in the model. These are exactly the circumstances under which the expectation-maximization (EM) algorithm for maximizing the likelihood might prove itself useful. None of the software packages that we are aware of, either using direct maximization or EM, can deal with our proposed model. Hence, we devised an EM algorithm tailored to our particular model. As this model is rather detailed, a variety of biologically-reasonable models can be derived as special cases. For this reason, we anticipate a broad range of applicability to our algorithm, beyond its original use. In the following we describe our model of exon-intron structure evolution and an EM algorithm for learning its parameters.

2 The Evolutionary Model

Suppose that we have multiple alignments of G different genes from S eukaryotic species, and let our observed data be the projection, upon the above alignments, of a presence-absence intron map. That is, at every site in each species we can observe either zero (absence of an intron), one (presence of an intron), or $\star$ (missing value, indicating lack of knowledge about intron's presence or absence). Let us define a *pattern* as any column in an alignment, and let $\Omega \leq 3^S$ be the total number of unique observed patterns, indexed as $\omega_1, \ldots, \omega_\Omega$. We shall use n_{gp} to denote the number of patterns ω_p that are observed in gene g.

Let the rooted phylogeny of the above S species be given by an N-node binary tree, where $S = (N+1)/2$. Let $q_0, \ldots, q_{N-1}$ be the nodes of this tree, with the convention that q_0 is the root node. We use the notations q^L, q^R and q^P to describe the left-descendant, right-descendant and parent, respectively, of node q (left and right are set arbitrarily). Also, let $\mathcal{L}(q)$ stand for the set of terminal nodes (leaves) that are descendants of q. We index the branches of the