

# A stochastic model for creation and loss of genomic sequence blocks

Alexander Alekseyenko <sup>1</sup>, Christopher Lee <sup>2</sup>

**Keywords:** stochastic modeling, sequence evolution, creation and loss, alternative splicing, immigration-death

## 1 Introduction

We propose a model for evolution of large genomic segments, which captures the stochastic nature of creation and loss process. The question of whether a segment of DNA sequence was created or lost during evolution, comes up very often in comparative genomics analysis. When we observe difference in presence of a DNA block in two or more organisms we often are interested in learning what caused this difference. One can sometimes successfully infer creation and loss events from the phylogenetic tree of the organisms making parsimony assumptions. Such methods, however, totally ignore the stochastic nature of the process underlying the creation and loss of sequence blocks. We use immigration-death process to model these events and provide a framework for estimation of the corresponding rates. Such model might have immediate application to analysis of gene acquisition and loss [1] or single exon creation and loss [3]. One can also imagine using the proposed immigration-death model to allow creation and loss of coevolving sequence blocks in analysis of phylogenies.

## 2 Model for sequence creation and loss

In practical problems dealing with sequence creation and loss we often directly observe conservation patterns of a certain sequence in two or more modern organisms. A set of such observations constitutes observed data for our model. Formally, if we observe  $K$  distinct conservation patterns in a set of  $N$  organisms, we let  $\mathbf{D} = (\mathbf{d}_1, \dots, \mathbf{d}_K)^t$  be our data matrix, where each pattern  $\mathbf{d}_i = (X_1, \dots, X_N)$  consists of presence or absence indicators  $X_j \in \{+, -\}$ . An example of such data matrix can be seen in Table 1. Each pattern is observed  $n_i$  times, and typically the number of distinct patterns is equal to  $2^N - 1$  (the pattern with all  $-$ 's is unobservable). We also assume that the phylogenetic relationship of the organisms is known and is represented by a rooted phylogenetic tree  $\tau$ .

We view the evolution of each sequence on each branch as an independent realization of the immigration-death process [2], conditional on the ancestral state. At the branching points the process produces a probabilistic copy of itself and the two identical processes evolve independently along the corresponding edges of the tree. The immigration-death model postulates that sequences are created according to Poisson process with an intensity  $\lambda$  and die at a rate  $\mu$ . Since each sequence evolves independently and observation of each pattern is independent of the rest of the data we can factor the likelihood function into

$$f(D, \tau | \lambda, \mu) = \prod_{i=1}^K p(d_i, \tau | \lambda, \mu)^{n_i}, \quad (1)$$

---

<sup>1</sup>Department of Biomathematics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095, USA. E-mail: [shuriko@ucla.edu](mailto:shuriko@ucla.edu)

<sup>2</sup>Department of Chemistry and Biochemistry, University of California Los Angeles, Los Angeles, CA 90095-1570, USA. E-mail: [leec@mbi.ucla.edu](mailto:leec@mbi.ucla.edu)

where  $p(x, \tau | \lambda, \mu)$  is the probability of observing pattern  $x$  and can be computed by integrating over the unobserved parental states in tree topology  $\tau$ . After assuming independent exponential priors with mean 1 on  $\lambda$  and  $\mu$ ,  $p(\mu) \propto e^{-\mu}$  and  $p(\lambda) \propto e^{-\lambda}$ , the posterior distribution of these parameters amounts to

$$\Pr(\lambda, \mu | \tau, D) \propto f(D, \tau | \lambda, \mu) p(\mu) p(\lambda) \quad (2)$$

In order to infer the model parameters and test hypotheses about evolutionary forces acting on sequence blocks, we construct MCMC sampler to simulate from the posterior distribution 2. We perform a simulation study of the effects of ascertainment and artificial thinning on estimation of our model parameters.

### 3 Application: Exon creation and loss

We will apply this model to a large dataset of alternatively spliced exons in four mammalian genomes to infer the contribution of creation and loss processes to the divergences observed in these data. The input patterns are obtained by conditioning on an exon being present in one of the organisms (i.e. organism 1), restricting our observed pattern matrix to always contain a + in the column representing that organism (see Table 1 for structure of the data). Therefore, each pattern represents the conservation of the corresponding exons in the homologous genes of other organisms. We will analyze these data as an example of practical application of our model. We expect to obtain statistical evidence for increased exon creation rate associated with alternative splicing, suggested previously by our empirical analysis of this dataset (unpublished data).

Patterns				Counts
1	2	3	4	
+	+	+	+	$n_1$
+	+	+	-	$n_2$
		⋮		⋮
+	-	-	-	$n_8$
-	+	+	+	0
		⋮		⋮
-	-	-	-	0

Table 1: Structure of the input data for alternative exon creation and loss analysis

## References

- [1] Lake, J. A. and Rivera, M. C. 2004. Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. *Molecular Biology and Evolution*, 21:681–690.
- [2] Lange K. 2003. *Applied Probability*. New York: Springer Verlag.
- [3] Modrek, B. and Lee, C. 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased rate of exon creation / loss. *Nature Genetics*, 34:177-180