B4B Chapter Proposal: Evolutionary History of Repeats

Sergey Nikolenko

August 26, 2012

Outline

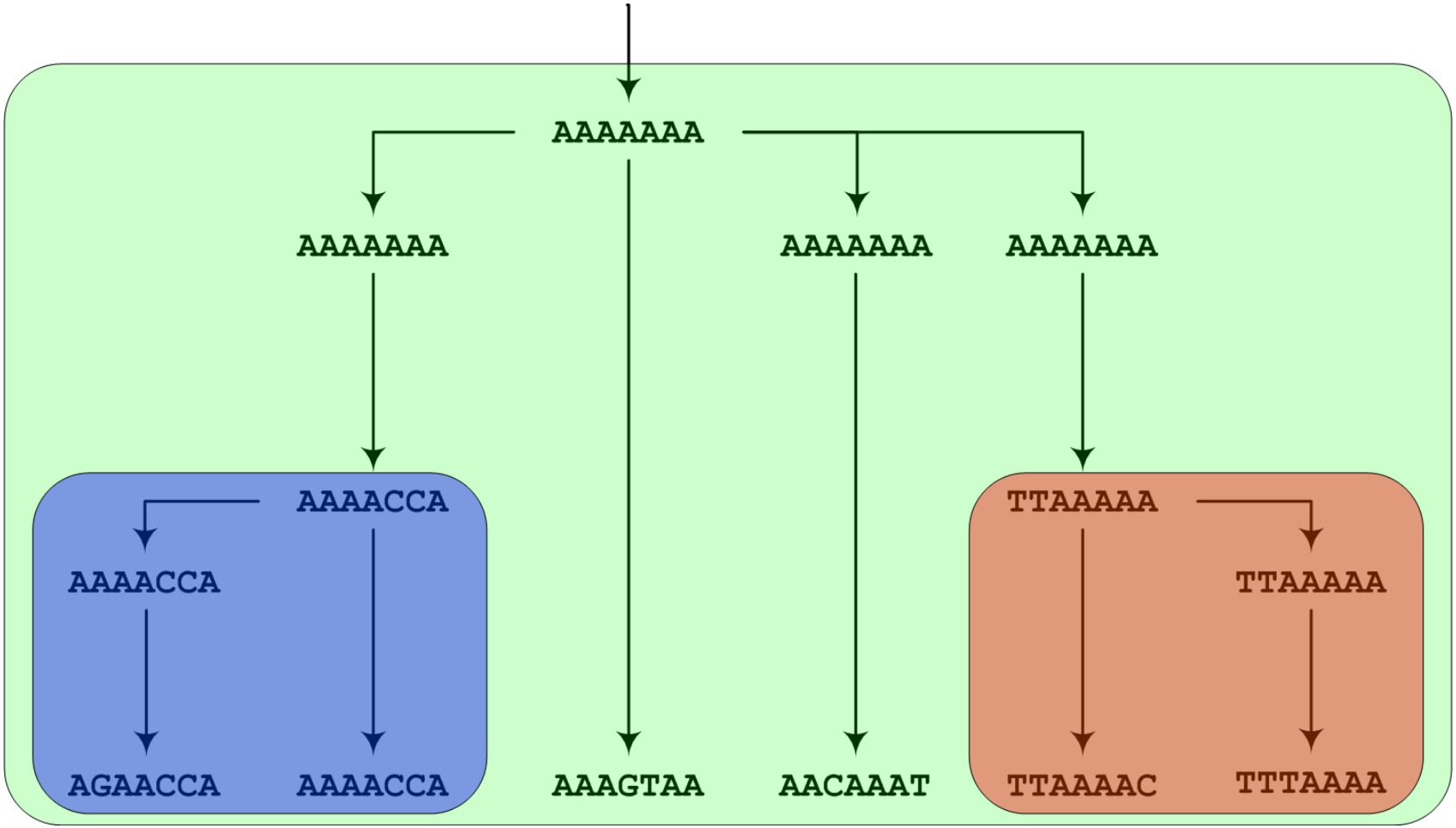
Mobile elements

- Mobile elements are DNA sequences that can move around the genome and copy themselves.
- Here we are mostly interested in LINEs and SINEs – interspersed elements that copy themselves a lot.
- Main example – *Alu* elements: a string of ≈ 300 nucleotides whose "copies" comprise more than 10% of the human genome.

Mobile elements

- At some point, a mobile element becomes active for a relatively short time and copies itself several times, then shuts back down.
- Moreover, mobile elements always undergo random mutations.
- The problem: given a collection of mobile elements, reconstruct its evolutionary history.
- Two main sources:
  - A. Price, E. Eskin, P. Pevzner, Whole-genome analysis of Alu repeat elements reveals complex evolutionary history, *Genome Res.*, 2004, vol. 14, pp. 2245–2252;
  - S. O'Rourke, N. Zaitlen, N. Jojic, E. Eskin, Reconstructing the Phylogeny of Mobile Elements, *RECOMB 2007*, LNBI 4453, pp. 196–210.

Problem setting



1. **Compelling problem**

Formalization

- This problem is easy to formalize:
  - each element is a string over $\{A, C, G, T\}$ of the same length;
  - the random generation process begins with a set of k source elements;
  - at every time step, there is a small probability for a letter to mutate randomly;
  - at every time step, there is an even smaller probability for an element to become active, i.e., copy itself several times without additional errors.
- The model is simple (not a formula in sight so far), but it is a quite accurate model (which is very seldom the case).

Formalization

- Moreover, since we assume that elements never disappear, the model suggests that the problem splits in two:
  - first identify subfamilies in the set of mobile elements;
  - then reconstruct their phylogenetic tree.

2. **Simple but accurate formalization that naturally decomposes the problem**

Outline

Reconstructing the phylogeny

- Let us begin with the second problem: reconstructing the phylogeny in a set of already identified subfamilies.
- There are two ideas here:
  - the "radius" (variance) of a subfamily corresponds to how long its elements had been mutating, so we can estimate which subfamilies are older;
  - then we can, following our model, simply compute the most probable ancestor of a cluster c with source element (consensus) $\mu(c)$ by Bayes theorem:

$$\operatorname{argmax}_{c' : \text{age}(c') > \text{age}(c)} p(c') p(\mu(c) \mid c').$$

Reconstructing the phylogeny

- This application of the Bayes theorem:

$$\mathrm{par}(c) = \mathrm{argmax}_{c':\mathrm{age}(c')>\mathrm{age}(c)} p(c')p(\mu(c) \mid c')$$

  is about as easy as it gets, but it is not some theoretical far removed simplification – it does solve our problem!
- Phylogeny reconstruction in this case allows for a gentle introduction to Bayesian inference (because we know all tree nodes, including intermediate ones, – otherwise structural EM would be needed, which is hardly a good introductory topic).

3. **Phylogeny reconstruction: a real life problem that introduces the very basics of machine learning**

Identifying subfamilies: biprofiles

- Subfamily identification is a more complicated problem, but a lot can be done with simple methods.
- First part:
  - show why distance-based clustering does not work (it can't detect correlations between mutations);
  - show that correlations can be found by collecting *biprofiles* [Price, Eskin, Pevzner, 2004].

Identifying subfamilies: biprofiles

- Biprofiles: to identify correlated mutations,
  - compute the number of times N_{ij} two mutations occur together and
  - compute the p-value of N_{ij} under the null hypothesis that they are independent.

4. **Biprofiles: a simple introduction to hypothesis testing**

Identifying subfamilies: clustering

- Model-based considerations yield a similar but even better way.
- I would begin this part with the EM algorithm – clustering is one of the simplest examples of latent variables, but here it is at least not exactly the mixture of gaussians seen in virtually every single book on this topic.
- The bottom line will be that EM provides good results but is slow, and additional care is needed to find the number of clusters.

5. **A simple example of the EM algorithm which is not a mixture of gaussians :)**

Identifying subfamilies: clustering

- Then we move on to RATS, a randomized clustering algorithm based on biprofiles [O'Rourke, Zaitlen, Jojic, Eskin, 2007]:
  - choose a number of random pairs of mutations (preselected based on p-value);
  - test them for independence;
  - if not independent, split the subfamily and repeat;
  - after splitting, try to join subfamilies back to avoid spurious splits.

Identifying subfamilies: clustering

- Educationally speaking, RATS lets us present:
  - the notion of mutual information between pairs of positions; it follows the gamma distribution for uncorrelated positions, and this leads to the independence test;
  - the computation of the necessary number of pairs to be tested as a function of desired p-value.
- In general, this shows that off-the-shelf algorithms (in this case, EM) don't always work in bioinformatics even when applicable since bioinformatics deals with massive datasets.

6. **A more complicated faster algorithm showing some more in-depth statistics**

Exercises

- Bioinformatics is a very practical field – one needs exercises, including software exercises.
- The proposed chapter abounds with opportunities for both
  - mathematical exercises in elementary probability theory and statistics and
  - software exercises: every algorithm above, including a dataset generator that implements the model of evolution, is relatively simple and straighforward to code.

7. **Many accessible exercises in both mathematics and programming**

Chapter plan

1. Introduction: mobile elements, how they work and why they are interesting.
2. Evolution model for mobile elements; formal problem setting.
3. Reconstructing phylogeny: searching for the MAP hypothesis.
4. Identifying subfamilies: failures of "regular" clustering and biprofiles.
5. Identifying subfamilies: model-based approaches. EM and RATS.

Why this is a good topic

1. Compelling problem.
2. Simple but accurate formalization that leads to a natural decomposition of the problem.
3. Phylogeny reconstruction: a real life problem that introduces the very basics of machine learning (Bayes theorem).
4. Subfamily identification: biprofiles make for a simple introduction to hypothesis testing.
5. Model-based: a simple example of the EM algorithm.
6. A faster algorithm showing some more in-depth statistics.
7. Many accessible exercises in both mathematics and programming.

Thank you!

Thank you for your attention!