BioE241 labs

Ian Holmes
374C Stanley Hall
UC Berkeley

Of all natural systems, living matter preserves inscribed in its organization the largest amount of its own past history .... no other system is better aufgehoben: *constantly abolished and simultaneously preserved.* [Pauling and Zuckerkandl, 1963]

Contents

Chapter 1

Introduction

This handout describes a series of labs for BioE241, accompanying the theoretical lectures in that class, which describe various statistical models for biological data.

The class has a (non-exclusive) emphasis on models that describe the evolution of DNA, RNA and amino acid sequences on phylogenetic trees.

Other software tools are also used. The labs also ask you to develop some pseudocode and actual implementations (generally in a programming language of your choice) and to do a small amount of elementary math.

1.1 Useful URLs

BioE241 class homepage	`http://biowiki.org/BioE241`
DART software homepage (used by most of the labs)	`http://biowiki.org/DART`
Class materials repository (lecture notes, these lab handouts, data files)	`https://github.com/ihh/bioe241`
DART source code repository	`https://github.com/ihh/dart`

Chapter 2

Simulate a discrete-state continuous-time Markov chain

Provide pseudocode for more than one of these.

Implement at least one.

PRESENTATION: Video of simulation

Several options available. Generate series of images using e.g. scripting language + library. Use Berkeley MPEG encoder to stitch together into a movie, or use one of various applications on desktop OS's that will do this.

2.1 Simulate from an exponential distribution by inverting the cumulative distribution

Pseudocode to be provided.

2.2 Simulate the general reversible-time nucleotide model over a finite time interval

2.3 Simulate the general reversible-time nucleotide model over a phylogenetic tree

2.4 Simulate Gillespie's algorithm

[Gillespie, 1977]

2.5 Simulate the spatial Lotka-Volterra model on a 2D lattice

Easy: discrete-time version. Hard: continuous-time version.

Collect summary statistics.

2.6 Simulate the 1D Ising model and the methylation-induced-CpG-deamination model

It may be easiest to implement the general nearest-neighbor irreversible-time nucleotide model, as both the Ising and CpG models are a subset of this.

Chapter 3

Use profile HMM training and search tools to build a family of homologous protein domain sequences

Chapter 4

Reconstruct the phylogenetic history of sequences

Chapter 5

Use probabilistic models to align protein sequences and reconstruct ancestors, on a given phylogeny

Chapter 6

Simultaneously reconstruct the phylogeny and the alignment

Chapter 7

Estimate the indel and substitution rates

Chapter 8

Use probabilistic models to predict the structure of a single RNA sequence

Chapter 9

Given two related RNA sequences, simultaneously align them and predict their common secondary structure

Chapter 10

Annotate conserved secondary structure in an RNA multiple alignment and reconstruct ancient RNA sequences

Chapter 11

Develop and fit models that allow for lineage-specific evolutionary effects

Chapter 12

Detect recombination breakpoints in multiple sequence alignments

Chapter 13

Use PRISM to prototype probabilistic models using statistical logic programming

Chapter 14

Simulate and analyze spatiotemporal models of evolution, epidemiology, ecology, and population dynamics

Chapter 15

Analyze and fit data to continuous-valued diffusion processes

Bibliography

[Gillespie, 1977] Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81:2340–2361.

[Pauling and Zuckerkandl, 1963] Pauling, L. and Zuckerkandl, E. (1963). Chemical paleogenetics, molecular "restoration studies" of extinct forms of life. *Acta Chemica Scandinavica*, 17:S9–S16.