

Mathematical Techniques for Gene Finding

Nguyen V.M. MAN

mnguyen@cse.hcmut.edu.vn

Department of Application Softwares, CSE Fac.

University of Technology

Ho Chi Minh City, Vietnam

Pham Manh HUNG

Ho Chi Minh City, Vietnam

February 26, 2007

Abstract

This paper .

Keywords: phylogenetic tree, computational algebra, discrete mathematics, genetics.

Contents

1 Introduction	2
1.1 HMM and gene finding	2
1.2 Few well-known models	3
1.2.1 Genomes	3
1.2.2 A data-driven science	4
1.2.3 Statistical models a biological sequences	4
2 Methods for Gene Finding- Our approach	5
2.1 What are the Jukes-Cantor model on a phylogenetic tree?	6
2.2 A small excursion into the Dynamic Programming approach	6
2.3 An intergrated method	7
3 Applications in genomic sequences	7
4 Summary	7

List of Tables

1 Introduction

This research project aims to investigate algebraic and combinatorial techniques in Biological Sequence Alignments. The paper consists of sections as follows.

HMM and gene finding An introduction

Few well-known models Juke-Cantor and Kimura models

Why Computer Algebra? use of algebraic geometry in formulations

Applications in genomic sequences trees having few leaves, test on real data

1.1 HMM and gene finding

In order to find genes in DNA sequences, it is necessary to identify structural features and sequence characteristics that distinguish genic sequence from non-genic sequence. One of the basic and useful probabilistic models are Hidden Markov Models (HMMs). A few useful notation are the followings.

Definition 1 (*Bayesian networks and HMMs*).

- A Markov chain M is a triple (Q, q, A) in which:
 - a/ Q is a finite set of states (can be identified with an alphabet Σ),
 - b/ p are initial probabilities,
 - c/ A are state transition probabilities, denoted by a matrix $A = [a_{st}]$ in which

$$a_{st} = P(x_i = t | x_{i-1} = s).$$

- d/ And such that the memoryless property is satisfied, i.e.,

$$P(x_i | x_{i-1}, \dots, x_1) = P(x_i | x_{i-1}).$$

- A Bayesian network (a directed graphical model) is a finite directed acyclic graph $G = (V, E) = (\mathbf{X} \oplus \mathbf{Y}, \mathbf{E})$ with two kinds of vertices, observed values $\mathbf{Y} = \{Y_1, Y_2, \dots, Y_n\}$

and hidden variables $\mathbf{X} = \{X_1, X_2, \dots, X_m\}$, where each edge is labeled by a transition matrix whose entries are linear forms in some parameters.

- An hidden Markov model is a directed graphical model in which the Y -vertices can take l possible states, and the X -vertices can take k possible states

A HMM is a probabilistic model that allows for simultaneous modelling of the bases in a DNA sequence of length n and the structural features associated with that sequence. In phylogenomics, usually the Y -vertices have $l = 4$ states, namely $\Omega = \{A, C, G, T\}$. The hidden variables X serve to model features associated with the sequence which is generated by the Y -vertices.

Example 1. A simple scene is $k = 2$, with the set of hidden states being $\{\text{exon}, \text{intron}\}$.

Our primary use of discrete mathematics here is to answer the following question:

Given observations $\mathbf{Y} = \boldsymbol{\sigma}$ and hidden data $\mathbf{X} = \mathbf{h}$, identify all parameter values such that \mathbf{h} is the most likely explanation for the observations $\boldsymbol{\sigma}$.

Definition 2. In general, a graphical model is an algebraic variety that can be represented as the image of a structured polynomial map $f : \mathbb{R}^d \mapsto \mathbb{R}^m$.

Here, \mathbb{R}^d is the space in which coordinates are the *model parameters* s_1, \dots, s_d ; and \mathbb{R}^m is the space in which coordinates $p_{\boldsymbol{\sigma}} = p_{\sigma_1 \sigma_2 \dots \sigma_n}$ are the joint (or marginal) probabilities for the observed random variables \mathbf{Y} .

1.2 Few well-known models

1.2.1 Genomes

Two very first questions arise in this study: why biological sequence analysis is important? And why algebra, statistics and computation are essential in investigating the topic?

For the first one, the obvious and significant answer is that genomes are fundamental objects that carry instructions for the self-assembly of living organisms. The second reason to focus on genomes is the abundance of high fidelity data. Then, statistical methods can be directly applied to **modeling the random evolution of genomes** and to **making inferences about the structure and organization of functional elements**.

Let us clarify fundamental concepts and notion being useful later.

- A *genome*, in every living organism, is a structure made up of DNA (deoxyribo nucleic acids) arranged in a double helix.
- Eukaryotes are organisms whose cells contain a nucleus (as the *human genome*, and prokaryotes are those don't.
- Eukaryotic genomes are divided into *chromosomes*, and the human genome has two copies of each chromosomes.
- Basic *nucleotides* are *A(denine)*, *C(ystosine)*, *G(uanine)* and *T(hymine)*, and the sequence DNA molecules in a genome is represented as a sequence of these letters $\Omega = \{A, C, G, T\}$.
- Certain highly structured subsequences within a genome are *genes*, they play a key role of encoding *proteins* (about $< 5\text{percent}$ of the human genome are genic sequence)
- Proteins are polymers composed of 20 distinct types of *amino acids*.
- *Codons*, within a gene, are triplets of DNA encoding the amino acids.

1.2.2 A data-driven science

Saying that computational biology is a data-driven science comes from the fact that progress in the field is the result of *analyzing data obtained from experiments*. The experiments are performed in individual labs or via collaborations utilizing *high-throughput technologies*. What we are going to study in this project belongs to *mathematical genomics*, a tiny field of *mathematical biology*. Two most concerns in genome analysis are: understand the organization and function of individual genomes, and understand the evolution of genomes and the mechanisms of natural selection. The foremost application that highlighted the use of discrete statistical models for biological sequence analysis is gene finding problem (GFP).

1.2.3 Statistical models a biological sequences

We restrict ourselves to introducing three most relevant terminologies to our study.

Maximum a posteriori inference concerns about inferencing with HMMs. Recall from the set up of Project C, we view the hidden model as a map $F : \mathbb{R}^d \rightarrow \mathbb{R}^{m \times n}$ specified by a matrix of polynomials $F = (f_{ij}(\theta))$, while the *observed model* is the map $\mathbf{f} : \mathbb{R}^d \rightarrow \mathbb{R}^m$ whose coordinates are the row sums of the matrix F .

In MAP inference we assume that one particular observation $i \in [m]$ has been made. The problem is to identify an index $j \in [n]$ that maximizes $f_{ij}(\theta)$; or equivalently we wish to find the best explanation j for the given observation i .

Continuous time Markov process. To work efficiently in GFP the concept of *continuous time Markov process* would be useful.

Definition 3. A *Q-matrix* or *rate-matrix* is a square matrix $Q = (q_{ij})$ with rows and columns indexed by the base set $\Sigma = \{A, C, G, T\}$ (depending on the application, Σ could be binary set or a set of 20 letters of amino acids).

The rate-matrix Q must satisfy that:

- $q_{ij} \geq 0$ for all $i \neq j$,
- $\sum_{j \in \Sigma} q_{ij} = 0$ for all $i \in \Sigma$
- $q_{ii} < 0$ for all $i \in \Sigma$.

A standard example is the Jukes-Cantor rate-matrix

$$Q = \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix}, \quad (1.1)$$

where $\alpha \geq 0$ is a parameter.

2 Methods for Gene Finding- Our approach

We discuss several well-known ways to attack GFP in this section.

1. Use HMM.

2. Use graphical models such as the Jukes-Cantor model and more complex ones.
3. Others like Dynamic Programming, Fourier transform or Hadamard conjugation.

2.1 What are the Jukes-Cantor model on a phylogenetic tree?

The concept of evolutionary model is very basic for the discussions later. An *evolutionary model* is specified by a phylogenetic tree T together with a rate-matrix Q and an initial distribution π for the root of T . The branch lengths t_i [page 152, [7]] of the edges are unknown parameters. The substitution matrix associated with the i th edge of T is $\theta^i = \theta^i(t_i)$, which is defined through the Jukes-Cantor rate-matrix (1.1) of the i th edge.

Definition 4. *The Jukes-Cantor model on a rooted tree T with r edges, denoted e , and n leaves is the polynomial map $\mathbf{f} : \mathbb{R}^r \rightarrow \mathbb{R}^{4^n}$ which is obtained by specializing the transition matrices M_e along the edges to the specific 4×4 matrices θ^i above.*

2.2 A small excursion into the Dynamic Programming approach

Dynamic programming (DP) provides a framework for understanding DNA sequence comparison algorithms, particularly in GFP and inferencing about gene functions.

Since genes can be represented as strings (sequences) of characters A, C, G and T , biological phenomena are based on *sequence similarity* or *distance* analysis such as longest common subsequences (LCS), global sequence alignment (GSA), scoring alignment ... Of course, there are biological strong evidences for each choice of alignment in studying a particular phenomenon.

Sequence alignment, a fundamental task in computational biology, is the alignment of DNA or protein sequences. More clearly, it is a procedure that attempts to provide algorithms that takes DNA sequences from several taxa, line up ‘common positions’ at which *substitutions* may or may not have occurred, and determine where *deletions* and *insertions* have occurred in certain sequences relative to the others. Substitutions, deletions and insertions are called elementary operations. The *Edit distance* (Levenshtein, 1966) between two sequences is the minimal number of elementary operations transforming one sequence into the other.

In those sequence similarity analysis algorithms, that much depend on what kind of distance (Hamming or Edit distance, e.g.) we define between two sequences, and also on various applications, it turns out that DP plays a very important role in reducing the problem complexity.

2.3 An intergrated method

Our approach is combining discrete optimization, algebra formulations, together with statistical inference and IT tools under a biologically meaningful viewpoint. Why is it essential? To name a typical case, up until now we have only tried to align two sequences. What about more than two? And what for? A faint similarity between two sequences becomes significant if present in many. Only DP can't handle the *Multiple alignments* that can reveal subtle similarities that pairwise alignments do not reveal.

3 Applications in genomic sequences

Three major tasks that we will investigate in this project are the followings:

1. Set up necessary and fundamental formulations together with methods to obtain important information for GFP,
2. Create a prototype of database to be a benchmark for methods proposed in (1).

4 Summary

References

- [1] Benny Chor, *Maximum Likelihood Jukes-Cantor Triplets: Analytic Solutions*, 2005, UC Bekerley
- [2] Nguyen, V. M. Man, *Computer-Algebraic Methods for the Construction of Designs of Experiments*, Ph.D. thesis, 2005, Technische Universiteit Eindhoven.
- [3] Nguyen, V. M. Man, *Some New Constructions of strength 3 Orthogonal Arrays*, accepted for publication in the Memphis 2005 Design Conference Special Issue of the Journal of Statistical Planning and Inference, January 2007.
- [4] Glonek G.F.V. and Solomon P.J., *Factorial and time course designs for cDNA microarray experiments*, Biostatistics 5, 89-111, 2004
- [5] Nguyen, V. M. Man and the DAG group at Eindhoven University of Technology, www.mathdox.org/nguyen, 2005,
- [6] Nicholas Eriksson, *Phylogenetic Algebraic Geometry*, preprint July 2004, Berkeley Univ.
- [7] Lior Pachter and Bernd Sturmfels, *Algebraic Statistics for Computational Biology*, <http://bio.math.berkeley.edu/ascb/> 2.1
- [8] T.A. Nguyen and al., *POP-C++ project*, <http://www.eif.ch/gridgroup/popc>, a parallel object-oriented programming system for computing on distributed environments, 2006
- [9] T.A. Nguyen, *An object-oriented model for adaptive high performance computing on the Computational Grid*, Ph.D. thesis, 2004, Swiss Federal Institute of Technology
- [10] Madhav, S. P., iSixSigma LLC, *Design Of Experiment For Software Testing*, isixsigma.com/library/content/c030106a.asp, 2004
- [11] Madhav, S. P., *Quality Engineering using robust design*, Prentice Hall, 1989