Decoding of Hidden Markov Models
with Applications to Sequence Alignment

Michal Nánási

na získanie akademického titulu *philosophiae doctor*
v odbore doktorandského štúdia: 9.2.1 Informatika

Bratislava, 2014
Dizertačná práca bola vypracovaná v dennej forme doktorandského štúdia na Katedre informatiky Fakulty matematiky, fyziky a informatiky Univerzity Komenského v Bratislave.

**Predkladateľ:** Michal Nánási  
Katedra informatiky  
Fakulta matematiky, fyziky a informatiky  
Univerzita Komenského  
Mlynská dolina  
842 48 Bratislava

**Školiteľ:** doc. Mgr. Bronislava Brejová, PhD.  
Katedra informatiky  
FMFI UK, Bratislava

**Oponenti:**  
-----------------

-----------------

-----------------

-----------------

-----------------

-----------------

-----------------

-----------------

-----------------


Obhajoba dizertačnej práce sa koná dňa ................ o ................ hod.

pred komisiou pre obhajobu dizertačnej práce v odbore doktorandského štúdia vymenovanou predsedom odborovej komisie dňa ..........................  

v študijnom odbore 9.2.1 Informatika  

na Fakulte matematiky, fyziky a informatiky Univerzity Komenského v Bratislave,  

Mlynská dolina, 842 48 Bratislava.

**Predsedá odborovej komisie:** prof. RNDr. Branislav Rovan, PhD.  
Katedra informatiky  
FMFI UK, Bratislava
Self Report (Autoreferát)

Introduction

New sequencing technologies are producing more and more biological data, including genomic sequences of many species. Therefore it is important to develop tools for automated analysis of such data. In dissertation we focus on computational methods for sequence annotation and sequence alignment.

In the sequence annotation problem, we want to label parts of the sequences according to their function, or meaning. We call such a labeling an annotation. For example, we can label each symbol of a genomic sequence base on whether it is part of a gene or not as in the following example (g is a label representing genes and n is a label for non-gene parts).

Sequence: ACGGTGGTTAGCTGCTGTCTGATCTAGCTAGT
Annotation: nnnnnnnngggggggggggggggggggggggnnnnnnnnngg

The sequence alignment is a data structure that characterizes similarity or shared origin of two or more sequences. We insert gap symbols (dashes) so that corresponding parts of the sequence are in the same column as in the following example.

Sequence X: CTGCTAGCTACGT--GTGT
Sequence Y: ---------ACGTGGAT--

Both annotation and alignment are fundamental bioinformatics problems. The first stages of analysis of newly sequenced genomes typically include aligning new sequences with the genomes of related species, that are already sequenced, and searching for known structures (like genes) inside new genomes. The search for know structures is done using sequence annotation methods. Many subsequent methods for analysing genomes rely on sequence annotation and sequence alignment. To avoid artefacts in the results of these downstream methods, there is need to develop algorithms for producing sequence annotation and alignment with as low error rate as possible. Tools for both sequence annotation and alignment are often based on hidden Markov models (HMMs). We propose new techniques for use of HMMs in these domains and also give proofs of NP-hardness for several related problems.

We work with generative probabilistic models, hidden Markov Models (HMM) and their variants. In general, an HMM is a state machine that generates a sequence (string) along with a sequence of states
(called state path). Since an HMM is a probabilistic model, it also defines the probability of sequences and state paths. The state path contains information about the structure of the generated sequence. In practice we are often given the generated sequence and the state path is hidden. The goal of the decoding algorithm is to reverse the generation process and obtain the original state path or at least its approximation.

While the sequence annotation and sequence alignments seem to be very different problems, the unifying element in dissertation is the use of HMMs, and in particular the selection of decoding algorithms. Selection of appropriate decoding algorithm is often neglected in practice, and usually most of the development is focused on the structure of an HMM. However, the right selection of a decoding algorithm can lead to significant improvements int the model prediction. We summarize the main contributions in two following sections.

**Sequence Annotation**

We study a special type of decoding algorithms: two-stage algorithms. In the first stage, the algorithm infers important aspects of the annotation, and in the second stage it fills remaining details in a way consistent with the first-stage results. We test this approach on the HIV recombination detection problem. We extend the Viterbi algorithm and the HERD algorithm to two-stage algorithms, and we show that two-stage algorithms can improve the accuracy of decoding (as far we know, such algorithms were previously used only for reducing the running time). Then we study the computational complexity of several decoding criteria appropriate for the first stage of two-stage algorithm, and we show NP-hardness results for obtaining the optimal annotations using these criteria.

We study computational complexity problems related to footprints and sets of a state path. Footprint of state path $\pi$, denoted as $f(\pi)$, is maximal subsequence of $\pi$ such that it does not contain two same consecutive states. Set of a state path, denoted as $S(\pi)$ is the set of states that are in the state path $\pi$. For example, if $\pi = uuuuuuuuuuuu$, then $f(\pi) = uuuuu$ and $S(\pi) = \{u, v\}$. The probability of set/footprint $K$ given sequence $X$ is the sum of the probabilities of state paths generating $X$ that have set/footprint $K$. We studied three following problems.

**Definition 1** (The most probable set problem). Given an HMM $H$, sequence $X$ of length $n$ and a number $p \in [0, 1]$, decide if there exists a set of states $S$ such that $\Pr(S(\pi) = S, X | H) \geq p$.

**Definition 2** (The most probable restriction problem). Given an HMM $H$, sequence $X$, integer $l$ and number $p \in [0, 1]$, decide if there exists a subset of states $S$ of size $l$, such that $\Pr(S(\pi) \subseteq S, X | H) \geq p$.

**Definition 3** (The most probable footprint problem). Given an HMM $H$, sequence $X$ of length $n$ and a number $p \in [0, 1]$, decide if there exists a footprint $F$ such that $\Pr(f(\pi) = F, X | H) \geq p$.

We showed that the most probable footprint problem and the most probable restriction problems are NP-complete and that the most probable set problem is NP-hard. It is not clear if the most probable
Alignment | Repeat | Block
---|---|---
3-state VA (baseline) | 4.78% | 
3-state PD | 4.41% | 
Context | 5.98% | 
Muscle | 5.62% | 
SFF MPD | 3.37% | 95.97% | 97.78% | 43.07% | 44.87% | 
SFF PD | 3.53% | 95.86% | 97.87% | 42.70% | 47.37% | 
SFF BPD | 3.51% | 93.09% | 98.07% | 36.50% | 41.67% | 
SFF BVA | 3.91% | 93.26% | 97.96% | 35.77% | 40.66% | 
SFF VA | 4.04% | 95.29% | 97.85% | 42.70% | 48.95% | 

Tabuľka 1: Accuracy of decoding methods on simulated data. Best result in each metric is bold. Upper part of the table contain standard 3-state HMM for aligning sequences with the Viterbi algorithm and Posterior decoding. Additionally we also include alignment software Muscle and Context. Standard algorithms methods do not provide repeat annotation and therefore only error rate is available for them.

set problem is in NP, because even deciding if the probability of a set is non-zero is NP-hard. We show different proofs for these theorems, and then we show the modification of proof of the most probable footprint problem to proofs of the other two theorems.

**Sequence Alignment**

In sequence alignment, the goal is to search for corresponding parts of the sequences and arrange them into same position in the alignment. To choose the biologically correct alignment, we usually optimize some scoring scheme. We will consider scoring schemes which are defined using pair hidden Markov models (pHMM). A pHMM generates pairs of sequences along with their alignment (an alignment is defined by the state path). This model is an extension of HMM.

We proposed a tractable method for aligning sequences with tandem repeats. A tandem repeat consists of consecutive copies (not exact) of a certain motif (short genomic sequence). Tandem repeats cause problems with sequence alignments because it is hard to distinguish between individual copies of the motif. We extend a traditional pHMM model for sequence alignment by additional states modeling tandem repeats. We call this model Sunflower field model (SFF). We propose new decoding algorithms tailored to this model, namely block Viterbi algorithm (BVA), and block posterior decoding (BPD). The method was tested on simulated data that resembles human-dog alignment. We run SFF with standard decoding algorithms, the Viterbi algorithm (VA), posterior decoding (PD), marginalized posterior decoding (MPD), and with the two new decoding algorithms. One such comparison is in table 1 where we compare
algorithms using different accuracy measures. The alignment error is the fraction of incorrectly predicted columns of an alignment. Repeat sensitivity and specificity measures the accuracy of prediction of repeat annotation for individual bases. The block sensitivity and specificity measures the accuracy of finding the tandem repeats with exact boundaries. The sensitivity is the number of correctly predicted features divided by the number of correct features. The specificity is the number of correctly predicted features divided by the number of all predicted features. The results showed that our new model and decoding methods decreased the error rate. In particular, method increased accuracy near the border of tandem repeats (such data are not shown here).

Paper about this method was presented on WABI 2013 conference, journal version was published in Algorithms in Molecular Biology. Journal version contain also experimental evaluation on real sequences.
Bibliography


Abstract

We study two important problems in computational biology: sequence annotation and sequence alignment. In the thesis we concentrate on the use of hidden Markov models (HMMs), well established generative probabilistic models.

In the first part, we study the sequence annotation problem, specifically the two-stage HMM decoding algorithms and the computational complexity of related problems. In particular, we demonstrate that two-stage algorithms can be used to increase the accuracy of decoding, and we prove the NP-hardness for three problems appropriate for the first stage: the most probable set problem, the most probable restriction problem and the most probable footprint problem.

The second part of the thesis focuses on alignment of sequences that contain tandem repeats. Tandem repeats are highly repetitive elements within genomic sequences that cause biases in alignments. To address this issue, we introduce a new HMM that models alignments containing tandem repeats, combine it with existing and new decoding algorithms, and evaluate our approach experimentally.

In both problems, we use the decoding algorithms to improve the accuracy of HMM predictions. Decoding algorithms are often neglected, and most of the development is focused on the structure of an HMM. However, a proper selection of a decoding method can lead to significant improvements in the predictions.
Abstrakt

Zaoberáme sa dvomi dôležitými bioinformatickými problémami: anotáciou sekvencií a zarovnávaním sekvencií. V práci sa sústredíme na využitie skrytých Markovových modelov (HMM), dobre známych generatívnych pravdepodobnostných modelov.

V prvej časti študujeme anotáciu sekvencií, konkrétne dvojstupňové dekóдовacie algoritmy a výpočtové problémy, ktoré s nimi súvisia. Ukážeme, že dvojstupňové algoritmy môžu zlepšiť presnosť dekódovania a dokážeme, že tri problémy vhodné pre prvý stupeň výpočtu sú NP-ťažké: problém najpravdepodobnejšej mnóžiny, problém najpravdepodobnejšej reštrikcie a problém najpravdepodobnejšej stopy.

Druhá časť sa zaoberá zarovnávaním sekvencií, ktoré obsahujú tandemové opakovania. Tandemové opakovania sú opakujúce sa časti genomických sekvencií, ktoré často spôsobujú chyby v zarovnaniach. Aby sme vyriešili tento problém, vyvinuli sme nový HMM, ktorý modeluje zarovnania obsahujúce tandemové opakovania a skombinovali sme ho s existujúcimi ako aj novými dekó dovacími algoritmami. Náš prístup sme vyhodnotili experimentálne.

V oboch problémoch sme používali dekóдовacie algoritmy na zlepšenie presnosti predikcií HMM. Dekódovacie algoritmy sú často podceňované a väčšina vývoja ide do vytvárania topológie HMM. Avšak správnym výberom dekóдовacej metódy môžeme dosiahnuť významné zlepšenie predikcií.
Vlastné publikácie autora


  
  