Applications of Generalized Pair Hidden Markov Models to Alignment and Gene Finding Problems

Lior Pachter, Marina Alexandersson, Simon Cawley
(Roman Plášil)

March 16, 2009
Motivation

Applications of Generalized Pair Hidden Markov Models to Alignment and Gene Finding Problems

Lior Pachter, Marina Alexandersson, Simon Cawley (Roman Plášil)
Motivation

CCCTGTACTTTTCGAGTTTTTGGTGTAATTGTAATCTATGATAGACAAA

hmm...

CCCTGTACTTTTCGAGTTTTTGGTGTAATTGTAATCTATGATAGACAAA
• doing both at once
• ...because conserved sequences are likely to be coding
• Overall better results for both
• This paper is not the first to do this
• ...but SLAM is first to use a model incorporating both tasks
• Hidden Markov Model → Generalized Pair Hidden Markov Model
Hidden Markov Models

...informatics

- Probability model
- Cryptoanalysis, speech/handwriting/gesture recognition, gene finding
Hidden Markov Models

- States, observables, transition probability, observation probabilities
- Markov property
Definition (HMM)

\[ S = \{s_1, s_2, \ldots, s_N\} \] states

\( O \) observables

\( X_1, X_2, \ldots, X_L \) states which generated an observation

\( \pi_i \) probability of state \( i \) being initial

\( a_{ij} \) probability of transition from \( i \) to \( j \)

\( b_i(o) \) probability of observing \( o \) in state \( i \)
Hidden Markov Models

Evaluation

Problem 1

Pr( ★★★★★ ) = 0.0000001

Probability of a particular observation for given HMM.
→ Forward algorithm
validating model correctness

Problem 2

Find the most likely sequence of states that could have generated given output.
→ Viterbi algorithm
gene finding

Applications of Generalized Pair Hidden Markov Models to Alignment and Gene Finding Problems
Lior Pachter, Marina Alexandersson, Simon Cawley (Roman Plášil)
Forward algorithm

• Given observation \( o_1, o_2, \ldots, o_L \)
• Its probability is...

\[
\sum_{X_1, \ldots, X_L} \pi_{X_1} \cdot b_{X_1}(o_1) \cdot a_{X_1, X_2} \cdot b_{X_2}(o_2) \cdots a_{X_{L-1}, X_L} \cdot b_{X_L}(o_L)
\]

• Calculating using dynamic programming
• \( \alpha_l(i) \) probability that after \( l \) steps the observation will be \( o_1, o_2, \ldots, o_l \) and the last state is \( i \)

\[
\alpha_l(i) = \Pr(o_1, o_2, \ldots, o_l \wedge X_l = i)
\]
Forward algorithm

• Given observation $o_1, o_2, \ldots, o_L$
• Its probability is...

$$
\sum_{X_1, \ldots, X_L} \pi_{X_1} \cdot b_{X_1}(o_1) \cdot a_{X_1, X_2} \cdot b_{X_2}(o_2) \cdots a_{X_{L-1}, X_L} \cdot b_{X_L}(o_L)
$$

• Calculating using dynamic programming
• $\alpha_l(i)$ probability that after $l$ steps the observation will be $o_1, o_2, \ldots, o_l$ and the last state is $i$

$$
\alpha_l(i) = \Pr(o_1, o_2, \ldots, o_l \land X_l = i)
$$
Forward algorithm

initialization $\alpha_1(i) = \pi_i \cdot b_i(o_1)$ for every state $i$

iteration $\alpha_{l+1}(j) = \left( \sum_{i=1}^{N} \alpha_l(i) \cdot a_{i,j} \right) b_j(o_{l+1})$ for every state $j$ a

$1 \leq l \leq L - 1.$

finalization $\Pr(o_1, o_2, \ldots, o_L) = \sum_{i=1}^{N} \alpha_L(i)$
Viterbi algorithm

• Again an observation is given (and fixed)
• Procedure similar to Forward algorithm, but
• ...$\sum \rightarrow \text{max}$
• ...in addition to this we store which state yielded the maximum
Viterbi algorithm

**Initialization**\( \delta_1(i) = \pi_i \cdot b_i(o_1) \) for every state \( i \)

**Iteration** \( \delta_l(j) = \max_{1 \leq i \leq N} (\delta_{l-1}(i) \cdot a_{i,j}) \cdot b_j(i_l) \) for every state \( j \) a
\[ 2 \leq l \leq L \]
\( \psi_l(j) = i, \) for which the expression above is maximal

**Finalization** \( \Pr(O \land X) = \max_{1 \leq i \leq N} (\delta_L(i)) \)
\( \psi_{L+1} = i, \) for which the expression above is maximal
finally we construct the state sequence using \( \psi \).
Aligning of two sequences

Use Pair HMM which generates two observations at the same time.

State duration

- In plain HMM single state generates a single output
- Using a self-loop we get a geometric distribution for output length
- The solution is to generate $n$ observations in a state
- Distribution of probability for $n$ can be given
Generalized Pair HMM

Aligning of two sequences
Use Pair HMM which generates two observations at the same time.

State duration
- In plain HMM single state generates a single output
- Using a self-loop we get a geometric distribution for output length
- The solution is to generate $n$ observations in a state
- Distribution of probability for $n$ can be given
Evaluating GPHMM

- More variables $\Rightarrow$ larger, more complex equations
- ...principle the same
- variable length of observations in state $\Rightarrow$ must calculate for all possible lengths
The model

- Blue states are classical, with self-loop
- Green states are generalized (geometric distribution unfit for exon lengths)
- ...alternating (→ more effective algorithm)
- Intron phase 0, 1, 2 relative to codon

\[ E_{i,j} \] exon at the beginning followed by intron phase \( j \)
\[ E_{F,j} \] exon at the end preceded by intron phase \( j \)
\[ E_{i,j} \] exon after intron phase \( i \) followed by intron phase \( j \)

**Deleted** deleted aminoacid

**Insert** ...

**Match** ...

\[ I_z \] generating nucleotides on strand \( Z \)
\[ I_y \] generating nucleotides on strand \( Y \)

**CNS** conserved noncoding sequence
Applications of Generalized Pair Hidden Markov Models to Alignment and Gene Finding Problems
Lior Pachter, Marina Alexandersson, Simon Cawley (Roman Plášil)
Complexity

- Time $O(D^4 N^2 TU)$
- Space $O(NTU)$
  - $D$: maximum exon length
  - $N$: number of states
  - $T, U$: input sequences lengths
- Improved using an approximate alignment (3 bp)
Results

\( a > b \)

Compared programs

- GENSCAN, classical program, uses single sequence
- ROSETTA, using two syntenic sequences
- SGP-1, using two syntenic sequences, searching in both
- SGP-2, searching in one sequence, using a database of other sequences
- TWINSCAN, searching in one sequence, using a database of other sequences
Results

\[ a \neq b \]

- SLAM is better in some areas
- ...and not much worse in other
- Better to use just two syntenic sequences than a whole database (too many false positives)
- Adding the CNS state helped a lot
  - fewer false positives (conserved \( \not\Rightarrow \) exon)
  - can recognize biologically significant CNS (UTR, regulatory areas)
- Preprocessing necessary
Summary

“It’s over, finally”

- Hidden Markov Model
- Forward and Viterbi algorithm
- generalizations
- a moderate improvement over existing methods