Applications of HMMs in Computational Biology

BMI/CS 576
www.biostat.wisc.edu/bmi576.html
Mark Craven
craven@biostat.wisc.edu

The protein classification task

Given: amino-acid sequence of a protein
Do: predict the family to which it belongs

GDLSTPDVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLGNGVLCVLAHHFGKEFTPPVQAAYAKVVAAGVANALAHKYH
Protein family - a simplified view

{A C A – – – A T G
 T C A A C T A T C
 A C A C – – A G C
 A G A – – – A T C
 A C C G – – A T C

family

A C A C – – A T C
A A A C – – A T C
T G C T – – A T C

query 1
query 2
query 3

An example from Krogh: An Introduction to HMMs for Biological Sequences, CMMB 1998.

Protein family - HMM

An example from Krogh: An Introduction to HMMs for Biological Sequences, CMMB 1998.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Probability ×100</th>
<th>Log odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consensus</td>
<td>4.7</td>
<td>6.7</td>
</tr>
<tr>
<td>Original sequences</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A C A – – – A T G</td>
<td>3.3</td>
<td>4.9</td>
</tr>
<tr>
<td>T C A A C T A T C</td>
<td>0.0075</td>
<td>3.0</td>
</tr>
<tr>
<td>A C A C – – A G C</td>
<td>1.2</td>
<td>5.3</td>
</tr>
<tr>
<td>A G A – – – A T C</td>
<td>3.3</td>
<td>4.9</td>
</tr>
<tr>
<td>A C C G – – A T C</td>
<td>0.59</td>
<td>4.6</td>
</tr>
<tr>
<td>Exceptional</td>
<td>0.0023</td>
<td>-0.97</td>
</tr>
</tbody>
</table>

An example from Krogh: An Introduction to HMMs for Biological Sequences, CMMB 1998.
Profile HMMs

- profile HMMs are used to model families of sequences

**Delete states** are silent; they account for characters missing in some sequences.

**Insert states** account for extra characters in some sequences.

**Match states** represent key conserved positions.

**Insert and match states** have emission distributions over sequence characters.

Multiple alignment of SH3 domain

Figure from A. Krogh, *An Introduction to Hidden Markov Models for Biological Sequences*
Profile HMMs

- to classify sequences according to family, we can train a profile HMM to model the proteins of each family of interest
- given a sequence \( x \), use Bayes’ rule to make classification

\[
P(c_i \mid x) = \frac{P(x \mid c_i)P(c_i)}{\sum_j P(x \mid c_j)P(c_j)}
\]

- use Forward algorithm to compute \( P(x \mid c_i) \) for each family \( c_i \)
Profile HMM accuracy

See Pfam database for a large collection profile HMMs

- classifying 2447 proteins into 33 families
- $x$-axis represents the median # of negative sequences that score as high as a positive sequence for a given family’s model
The gene finding task

Given: an uncharacterized DNA sequence
Do: locate the genes in the sequence, including the coordinates of individual *exons* and *introns*

Eukaryotic gene structure
Sources of evidence for gene finding

- **signals**: the sequence signals (e.g. splice junctions) involved in gene expression

- **content**: statistical properties that distinguish protein-coding DNA from non-coding DNA

- **conservation**: signal and content properties that are conserved across related sequences (e.g. syntenic regions of the mouse and human genome)

Gene finding: search by content

- encoding a protein affects the statistical properties of a DNA sequence

<table>
<thead>
<tr>
<th>Codon</th>
<th>a.a.</th>
<th>fraction per codon per a.a.</th>
</tr>
</thead>
<tbody>
<tr>
<td>UUU</td>
<td>P</td>
<td>0.46</td>
</tr>
<tr>
<td>UUC</td>
<td>P</td>
<td>0.54</td>
</tr>
<tr>
<td>UUA</td>
<td>L</td>
<td>0.08</td>
</tr>
<tr>
<td>UUG</td>
<td>L</td>
<td>0.13</td>
</tr>
<tr>
<td>CUU</td>
<td>L</td>
<td>0.13</td>
</tr>
<tr>
<td>CUC</td>
<td>L</td>
<td>0.20</td>
</tr>
<tr>
<td>CUA</td>
<td>L</td>
<td>0.07</td>
</tr>
<tr>
<td>CUG</td>
<td>L</td>
<td>0.40</td>
</tr>
<tr>
<td>AUU</td>
<td>I</td>
<td>0.36</td>
</tr>
<tr>
<td>AUC</td>
<td>I</td>
<td>0.47</td>
</tr>
<tr>
<td>AUA</td>
<td>I</td>
<td>0.17</td>
</tr>
<tr>
<td>AUG</td>
<td>M</td>
<td>1.00</td>
</tr>
<tr>
<td>GUU</td>
<td>V</td>
<td>0.18</td>
</tr>
<tr>
<td>GUC</td>
<td>V</td>
<td>0.24</td>
</tr>
<tr>
<td>GUA</td>
<td>V</td>
<td>0.12</td>
</tr>
<tr>
<td>GUG</td>
<td>V</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Homo sapiens data from the Codon Usage Database
The GENSCAN HMM for Eukaryotic Gene Finding [Burge & Karlin ‘97]

Each shape denotes a functional unit of a gene or genomic region and is represented by a submodel in the HMM

Pairs of intron/exon units represent the different ways an intron can interrupt a coding sequence (after 1st base in codon, after 2nd base or after 3rd base)

Complementary submodel (not shown) detects genes on opposite DNA strand

GENSCAN uses a variety of submodel types

<table>
<thead>
<tr>
<th>sequence feature</th>
<th>model</th>
</tr>
</thead>
<tbody>
<tr>
<td>exons</td>
<td>5th order inhomogenous</td>
</tr>
<tr>
<td>introns, intergenic regions</td>
<td>5th order homogenous</td>
</tr>
<tr>
<td>poly-A, translation initiation, promoter</td>
<td>0th order, fixed-length</td>
</tr>
<tr>
<td>splice junctions</td>
<td>tree-structured variable memory</td>
</tr>
</tbody>
</table>
Markov models & exons

- consider modeling a given coding sequence
- for each “word” we evaluate, we’ll want to consider its position with respect to the reading frame we’re assuming

reading frame

```
G C T A C G
C T A C G G
T A C G G A
```

- can do this using an inhomogeneous model

A fifth-order inhomogenous Markov chain

```
GCTAC AAAAA TTTTT CTACG CTACA CTACC CTACT
AAAAA TTTTT TACAG TACAA TACAC TACA T
```

```
transitions
to states in pos 2
```

```
start
AAAAA
CTACA
CTACC
CTACG
CTACT
GCTAC
TTTTT
```

```
position 2
```

```
AAAAA
CTACA
CTACC
CTACG
CTACT
GCTAC
TTTTT
```

```
position 3
```

```
AAAAA
TACAA
TACAC
TACAG
TACAT
GCTAC
TTTTT
```

```
position 1
```
Inference with the gene-finding HMM

given: an uncharacterized DNA sequence
find: the most probable path through the model for the sequence

• this path will specify the coordinates of the predicted genes (including intron and exon boundaries)
• the Viterbi algorithm is used to compute this path

Parsing a DNA sequence

The Viterbi path represents a parse of a given sequence, predicting exons, introns, etc
Other issues in Markov models

- there are many interesting variants and extensions of the models/algorithms we considered here (some of these are covered in BMI/CS 776)
  - separating length/composition distributions with *semi-Markov models*
  - modeling multiple sequences with *pair HMMs*
  - learning the *structure* of HMMs
  - going up the Chomsky hierarchy: *stochastic context free grammars*
  - discriminative learning algorithms (e.g. as in *conditional random fields*)
  - etc.