Chapter 4: Hidden Markov Models

4.4 HMM Applications

Overview

- Alignment with HMM
- Gene predictions
- Protein structure predictions
Sequence Analysis Using HMM

- Step 1: construct an HMM model
  - Design an HMM generator for the observed sequences
  - Assign hidden states to underlying sequence regions
  - Model the questions to be answered in terms of hidden-pathway

- Step 2: train the HMM
  - Supervised/unsupervised

- Step 3: analyze sequences
  - Viterbi decoding: compute most likely hidden-pathway
  - Forward/Backward: compute likelihood of sequence (& p/Z-score)
Is the target sequence $S$ generated by profile $H$?

- Viterbi: use $H$ to compute the most probable alignment of $S$
- Forward: compute probability of $S$ generated by $H$

Example: searching for globins

- Create profile HMM from known globin sequences
- Use the forward algorithm to search SWISS-PROT for globin
- Compare against the null hypothesis (random sequence)
- Result: globins are highly distinct
Pairwise Alignment With HMM

- Key idea: recast DP as Viterbi decoding
- "Alignment" = sequence of pairs \{<X,Y>,<X,>_ ,<_ ,Y>\}
- Consider an HMM that emits pairs (pair HMM):

Viterbi Decoding \rightarrow Pairwise Alignment

- Viterbi decoding: optimum path of pairs
- Forward step: select among \{XY,-Y,X-\}
Comparing With Standard DP

- Compute relationships scores $\Leftrightarrow$ probabilities
- Consider an HMM that emits pairs (pair HMM)

$$<\delta, \epsilon, \tau, P, q> \Rightarrow <s(x, y), d, e>$$

e.g., $e = -\log\left[\frac{\epsilon}{1 - \tau}\right]$
Schematics

- Prokaryotic genes
- Eukaryotic genes

Modeling Gene Structure With HMM
- Gene regions are modeled as HMM states
- Emission probabilities reflect nucleotide statistics
Simple Gene Models

- Consider a region of random length $L$
  - Markov model $\Rightarrow$ $L$ is geometrically distributed $P[L=k]=p^k(1-p)$
  - $E[L]=p/(1-p)$
- How do we model ORF?
  - Codons can be modeled as higher-order states

![Gene Model Diagram]

VEIL: Viterbi Exon-Intron Locator
(Henderson, Salzberg, & Fasman 1997)

- Contains 9 hidden states
- Each state is a Markovian model of regions
  - Exons, introns, intergenic regions, splice sites, etc.

![VEIL Architecture Diagram]
Genie (Kulp 96)

- Uses a generalized HMM (GHMM)
- Edges in model are complete HMMs
- States are neural networks for signal finding

• J5’ – 5’ UTR
• EI – Initial Exon
• E – Exon, Internal Exon
• I – Intron
• EF – Final Exon
• ES – Single Exon
• J3’ – 3’ UTR

GeneScan (Burge & Karlin 97)

- Base models overall parse of the gene
  - 5’UTR/3’UTR; Promoter; Poly A...
  - Exon-Intron-Exon structure of gene
- Intron states model 3 scenarios:
  - I0: Intron is between two codons
  - I1: Intron is right after first codon base
  - I2: Intron is right after second codon base
- Exons model respective scenarios
Extending The HMM Model

Problem: the HMM does not model nature
- Sequence lengths are not geometrically distributed
- Some regions have special features (e.g., splice sites use GT…)
- ……
- Challenge: generalize HMM while retaining algorithms
GeneScan: Extended Sequence Generator Model

- Add to state k a length distribution $f_k(L)$
  - Semi-Markov model
  - Extend the Viterbi DP recursion to reflect this
- Incorporate special regions structures
  - Weight matrix model for splice site
- ….
- Preserve the recursion structure

Measuring Classifier Performance

- The challenge: how do we evaluate performance of a classifier?
  - E.g., consider a test to decide whether a person is sick (p) or healthy (n)
- Measured Performance
  - $TP = \text{True Positive}; \ TN = \text{True Negative}; \ FP = \text{False Positive}; \ FN = \text{False Negative}$
  - Sensitivity: $SN = TP/(TP+FN)$
    - Sensitivity = percentage of correct predictions when the actual state is positive
  - Specificity: $SP = TN/(TN+FP)$
    - Specificity = percentage of correct predictions when the actual state is negative
  - Receiver Operating Characteristics (ROC) curves
    - Represent tradeoff between sensitivity & specificity

\[
\begin{array}{c|c|c}
\text{Actual} & \text{Prediction} & \text{Model} \\
\hline
P & TP & N' \\
N & FP & N' \\
\hline
\end{array}
\]
## Performance Comparisons

### A. Comparison of GENSCAN with other gene prediction programs

<table>
<thead>
<tr>
<th>Program</th>
<th>Sequences</th>
<th>Accuracy per nucleotide</th>
<th>Accuracy per exon</th>
<th>WE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENSCAN</td>
<td>570 (8)</td>
<td>0.75 0.78 0.71 0.74 0.71 0.75 0.71 0.72 0.78</td>
<td>0.78 0.71 0.78 0.71 0.75 0.71 0.72 0.78</td>
<td>0.83 0.83 0.83 0.83 0.83 0.83 0.83 0.83</td>
</tr>
<tr>
<td>FGENESH</td>
<td>560 (22)</td>
<td>0.77 0.78 0.78 0.76 0.71 0.74 0.78 0.71 0.72</td>
<td>0.78 0.71 0.78 0.71 0.74 0.78 0.71 0.72</td>
<td>0.80 0.80 0.80 0.80 0.80 0.80 0.80 0.80</td>
</tr>
<tr>
<td>GeneID</td>
<td>570 (22)</td>
<td>0.63 0.64 0.67 0.65 0.64 0.64 0.65 0.64 0.64</td>
<td>0.64 0.64 0.64 0.64 0.64 0.64 0.64 0.64</td>
<td>0.70 0.70 0.70 0.70 0.70 0.70 0.70 0.70</td>
</tr>
<tr>
<td>Genie</td>
<td>570 (65)</td>
<td>0.76 0.77 0.77 0.80 0.74 0.74 0.77 0.74 0.77</td>
<td>0.77 0.74 0.77 0.74 0.74 0.77 0.74 0.77</td>
<td>0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90</td>
</tr>
<tr>
<td>Genleng</td>
<td>570 (90)</td>
<td>0.72 0.79 0.79 0.76 0.71 0.74 0.79 0.71 0.72</td>
<td>0.79 0.71 0.79 0.74 0.74 0.79 0.71 0.72</td>
<td>0.93 0.93 0.93 0.93 0.93 0.93 0.93 0.93</td>
</tr>
<tr>
<td>GeneDecoy2</td>
<td>562 (5)</td>
<td>0.66 0.70 0.70 0.65 0.68 0.65 0.70 0.68 0.65</td>
<td>0.70 0.68 0.70 0.65 0.65 0.70 0.68 0.65</td>
<td>0.86 0.86 0.86 0.86 0.86 0.86 0.86 0.86</td>
</tr>
<tr>
<td>ORAS2</td>
<td>570 (25)</td>
<td>0.72 0.75 0.75 0.76 0.74 0.74 0.75 0.74 0.76</td>
<td>0.76 0.74 0.76 0.74 0.74 0.76 0.74 0.76</td>
<td>0.88 0.88 0.88 0.88 0.88 0.88 0.88 0.88</td>
</tr>
<tr>
<td>SORFIND</td>
<td>561 (6)</td>
<td>0.91 0.85 0.85 0.72 0.84 0.84 0.85 0.84 0.85</td>
<td>0.85 0.84 0.85 0.84 0.84 0.85 0.84 0.85</td>
<td>0.95 0.95 0.95 0.95 0.95 0.95 0.95 0.95</td>
</tr>
<tr>
<td>Xpose</td>
<td>590 (50)</td>
<td>0.61 0.67 0.67 0.60 0.63 0.63 0.67 0.63 0.67</td>
<td>0.63 0.63 0.63 0.63 0.63 0.63 0.63 0.63</td>
<td>0.79 0.79 0.79 0.79 0.79 0.79 0.79 0.79</td>
</tr>
<tr>
<td>GenieDecoy</td>
<td>478 (1)</td>
<td>0.91 0.91 0.91 0.88 0.88 0.88 0.91 0.88 0.88</td>
<td>0.88 0.88 0.88 0.88 0.88 0.88 0.88 0.88</td>
<td>0.81 0.81 0.81 0.81 0.81 0.81 0.81 0.81</td>
</tr>
<tr>
<td>GeneDecoy2</td>
<td>578 (1)</td>
<td>0.88 0.91 0.91 0.88 0.88 0.88 0.91 0.88 0.88</td>
<td>0.88 0.88 0.88 0.88 0.88 0.88 0.88 0.88</td>
<td>0.77 0.77 0.77 0.77 0.77 0.77 0.77 0.77</td>
</tr>
</tbody>
</table>

### B. GENSCAN accuracy for sequences grouped by C+G content and by organism

<table>
<thead>
<tr>
<th>Subset</th>
<th>Sequences</th>
<th>Accuracy per nucleotide</th>
<th>Accuracy per exon</th>
<th>WE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sequences</td>
<td>Sc  Sc  Sc  Sc  Sc  Sc  Sc  Sc  Sc</td>
<td>Sc  Sc  Sc  Sc  Sc  Sc  Sc  Sc  Sc</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00</td>
<td>0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00</td>
<td></td>
</tr>
<tr>
<td>C+G &lt;60</td>
<td></td>
<td>200 (1)</td>
<td>0.94 0.94 0.94 0.94 0.94 0.94 0.94 0.94 0.94</td>
<td>0.94 0.94 0.94 0.94 0.94 0.94 0.94 0.94 0.94</td>
</tr>
<tr>
<td>C+G &lt;50</td>
<td></td>
<td>100 (0)</td>
<td>0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90</td>
<td>0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90</td>
</tr>
<tr>
<td>C+G &lt;40</td>
<td></td>
<td>50 (0)</td>
<td>0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90</td>
<td>0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90</td>
</tr>
<tr>
<td>Primates</td>
<td></td>
<td>200 (1)</td>
<td>0.94 0.94 0.94 0.94 0.94 0.94 0.94 0.94 0.94</td>
<td>0.94 0.94 0.94 0.94 0.94 0.94 0.94 0.94 0.94</td>
</tr>
<tr>
<td>Rodents</td>
<td></td>
<td>100 (0)</td>
<td>0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90</td>
<td>0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90</td>
</tr>
<tr>
<td>Non-mamm.</td>
<td></td>
<td>50 (0)</td>
<td>0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90</td>
<td>0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90</td>
</tr>
</tbody>
</table>

Identifying Transmembrane Proteins
Transmembrane Proteins

- Transmembrane proteins are the key tools for cells interactions
  - Signaling, transport, adhesion...
- A common structural motif: core helices
- Example: receptors
  - Function: receive signals, activate pathways
  - Operations:
    1. Signal recognition
    2. Signal transduction
    3. Pathway activation
    4. Down regulation

Example: The Insulin Pathway

- Insulin binds to InsR receptor
- Glut-4 is transported to membrane to generate glucose influx
- Metabolic network processing

From Wikipedia
TMHMM [Sonhammer et al 98]

- A HMM to identify transmembrane proteins
- Architecture:
  - The helix core is key
  - Globular and loop domains
  - Modeling challenge: variable length

Training
- The Maximization step of Baum-Welch uses simulated annealing
- 3-stage training

---

TMHMM Performance

<table>
<thead>
<tr>
<th>Method</th>
<th>Training set size</th>
<th>Stage of training</th>
<th>Correct topology</th>
<th>Correct location</th>
<th>Single TM sensitivity</th>
<th>Single TM specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMHMM</td>
<td>85</td>
<td>1</td>
<td>66 (71.3%)</td>
<td>61 (74.1%)</td>
<td>95.6%</td>
<td>95.6%</td>
</tr>
<tr>
<td></td>
<td>85</td>
<td>2</td>
<td>65 (75.9%)</td>
<td>66 (83.1%)</td>
<td>96.2%</td>
<td>97.5%</td>
</tr>
<tr>
<td></td>
<td>85</td>
<td>3</td>
<td>64 (72.1%)</td>
<td>69 (83.1%)</td>
<td>96.2%</td>
<td>97.6%</td>
</tr>
<tr>
<td>MEMSAT</td>
<td>85</td>
<td>1</td>
<td>61 (75.5%)</td>
<td>67 (80.7%)</td>
<td>96.8%</td>
<td>94.6%</td>
</tr>
<tr>
<td>PPhilm</td>
<td>85</td>
<td>1</td>
<td>61 (75.5%)</td>
<td>67 (80.7%)</td>
<td>96.8%</td>
<td>94.6%</td>
</tr>
<tr>
<td></td>
<td>85</td>
<td>2</td>
<td>64 (72.1%)</td>
<td>69 (83.1%)</td>
<td>96.2%</td>
<td>97.6%</td>
</tr>
<tr>
<td></td>
<td>85</td>
<td>3</td>
<td>64 (72.1%)</td>
<td>69 (83.1%)</td>
<td>96.2%</td>
<td>97.6%</td>
</tr>
<tr>
<td>TMEMMM</td>
<td>160</td>
<td>1</td>
<td>106 (66.3%)</td>
<td>122 (76.1%)</td>
<td>95.4%</td>
<td>97.3%</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>2</td>
<td>120 (73.9%)</td>
<td>133 (83.1%)</td>
<td>96.8%</td>
<td>97.5%</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>3</td>
<td>123 (76.9%)</td>
<td>134 (83.8%)</td>
<td>97.1%</td>
<td>97.5%</td>
</tr>
<tr>
<td>MEMSAT</td>
<td>160</td>
<td>1</td>
<td>108 (67.5%)</td>
<td>118 (73.9%)</td>
<td>93.3%</td>
<td>95.6%</td>
</tr>
</tbody>
</table>
Proteins
Structure Prediction

Proteins Are Formed From Peptide Bonds

Ramachandran Regions
Levels of Protein Structure

Structure Formation
Helix is Formed Through H Bonds

Tertiary Structures: β-Barrel Example
Why Is Structure Important?

Protein functions are handled by structural elements

Structure Determination

- Map: sequence $\rightarrow$ structure
- Structure determination is handled through crystallography
- X-Ray; NMR
- PDB: a database of structures
### The Problem Of Structure Prediction

- Derive conformation geometry from sequence
  - Ab-initio techniques
  - Homology techniques
- Secondary structure is organized in folds
  - $\alpha$-helix, $\beta$-sheets…

### The Structure Prediction Problem

- Given: an amino acid sequence
- Output: structure annotation \{H,B…\}
HMM Model

- Hidden states: folds
- Observable: AA sequence
- Viterbi decoding: find most likely annotation
- Training: supervised learning use known structures
- Performance: HMM works well for limited protein types
  - E.g., TMHMM...
  - For general structure predictions Neural-Nets are superior (e.g., PHD, R. Burkhard 93)
- Why?
  - HMM works as long as the underlying conformation is consistent with the Markovian model.
  - Protein conformations may depend on complex long-range non-Markovian interactions. E.g.,
    consider the impact of a single AA change on hemoglobin conformation in sickle-cell anemia.

Example (Chu et al, 2004)

- Step 1: align sequences to partition into regions
- Step 2: build extended HMM (semi-markov length)

www.gatsby.ucl.ac.uk/~chuwei/paper/icml04_talk.pdf
Performance

<table>
<thead>
<tr>
<th></th>
<th>CASP2 (20 chains)</th>
<th>CASP3 (36 chains)</th>
<th>CASP4 (40 chains)</th>
<th>CASP5 (56 chains)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q_0$</td>
<td>73.40%</td>
<td>71.12%</td>
<td>74.32%</td>
<td>74.03%</td>
</tr>
<tr>
<td>$Q_{0}^{\text{As}}$</td>
<td>76.62%</td>
<td>73.12%</td>
<td>80.22%</td>
<td>80.43%</td>
</tr>
<tr>
<td>$Q_{0}^{\text{B}}$</td>
<td>61.29%</td>
<td>56.35%</td>
<td>57.81%</td>
<td>59.52%</td>
</tr>
<tr>
<td>$Q_{0}^{\text{P} \text{col}}$</td>
<td>77.73%</td>
<td>78.88%</td>
<td>78.00%</td>
<td>76.81%</td>
</tr>
<tr>
<td>$Q_{0}^{\text{P} \text{col}}$</td>
<td>79.71%</td>
<td>74.91%</td>
<td>81.33%</td>
<td>76.95%</td>
</tr>
<tr>
<td>$Q_{0}^{\text{E} \text{col}}$</td>
<td>76.48%</td>
<td>78.39%</td>
<td>76.19%</td>
<td>78.10%</td>
</tr>
<tr>
<td>$Q_{0}^{\text{E} \text{col}}$</td>
<td>67.36%</td>
<td>65.99%</td>
<td>67.28%</td>
<td>69.88%</td>
</tr>
</tbody>
</table>

$Q_0$ is the overall accuracy, $Q_{0}^{\text{As}} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative} + \text{False Positive}}$

and $Q_{0}^{\text{E} \text{col}} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$

Challenges long-range interactions (e.g., $\beta$-sheets)

Conclusions

- HMM are very useful when a Markovian model is appropriate
- Solve three central problems:
  - Decoding sequences to classify their components
  - Computing likelihood of the classification
  - Training
- May be extended to handle non-Markovian elements
- But can lose their predictive power when Markovian assumptions diverge from nature