CSE182-L8

Regular expressions
Protein Sequence Analysis
Profiles

October 30, 2014
Silly Quiz

Skin patterns
Facial Features
Not all features (residues) are important

Skin patterns
Facial Features
Diverged family members provide key features
Regular expressions as Protein sequence motifs

\[ C-X-[DE]-X^{10,12} - C-X-C--[STYLV] \]
Regular Expressions

- Concise representation of a set of strings over alphabet $\Sigma$.
- Described by a string over
- $R$ is a r.e. if and only if

$$\{ \Sigma, \cdot, *, + \}$$

<table>
<thead>
<tr>
<th>Expression</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R = { \varepsilon }$</td>
<td>Base case</td>
</tr>
<tr>
<td>$R = { \sigma }, \sigma \in \Sigma$</td>
<td>Union of strings</td>
</tr>
<tr>
<td>$R = R_1 + R_2$</td>
<td>Concatenation</td>
</tr>
<tr>
<td>$R = R_1 \cdot R_2$</td>
<td>0 or more repetitions</td>
</tr>
<tr>
<td>$R = R_1^*$</td>
<td>0 or more repetitions</td>
</tr>
</tbody>
</table>
Regular Expression

- **Q:** Let $\Sigma=\{A,C,E\}$
  - Is $(A+C)^*EEC^*$ a regular expression?
  - $*(A+C)$?
  - $AC^*..E$?

- **Q:** When is a string $s$ in a regular expression?
  - $R=(A+C)^*EEC^*$
  - Is $CEEC$ in $R$?
  - $AEC$?
  - $ACEE$?
Every R.E can be expressed by an automaton (a directed graph) with the following properties:
- The automaton has a start and end node
- Each edge is labeled with a symbol from $\Sigma$, or $\varepsilon$

Suppose $R$ is described by automaton $A$
- $S \in R$ if and only if there is a path from start to end in $A$, labeled with $s$. 
Examples: Regular Expression & Automata

- \((A+C)^*EEC^*\)
Constructing automata from R.E

- $R = \{\varepsilon\}$
- $R = \{\sigma\}, \sigma \in \Sigma$
- $R = R_1 + R_2$
- $R = R_1 \cdot R_2$
- $R = R_1^*$
Matching Regular expressions

- A string $s$ belongs to $R$ if and only if, there is a path from START to END in $R_A$, labeled by $s$.
- Given a regular expression $R$ (automaton $R_A$), and a database $D$, is there a string $D[b..c]$ that matches $R_A (D[b..c] \in R)$

- Simpler Q: Is $D[1..c]$ accepted by the automaton of $R$?
Alg. For matching R.E.

• If $D[1..c]$ is accepted by the automaton $R_A$
  - There is a path labeled $D[1]...D[c]$ that goes from START to END in $R_A$
Alg. For matching R.E.

- If $D[1..c]$ is accepted by the automaton $R_A$
  - There is a path labeled $D[1]...D[c]$ that goes from START to END in $R_A$
  - There is a path labeled $D[1]..D[c-1]$ from START to node $u$, and a path labeled $D[c]$ from $u$ to the END
D.P. to match regular expression

**Define:**
- $A[u, \sigma] =$ Automaton node reached from $u$ after reading $\sigma$
- $Eps(u):$ set of all nodes reachable from node $u$ using epsilon transitions.
- $N[c] =$ subset of nodes reachable from START node after reading $D[1..c]$
- $Q: \text{when is } v \in N[c]$
D.P. to match regular expression

• Q: when is \( v \in N[c] \) ?
  • A: If for some \( u \in N[c-1] \), \( w = A[u,D[c]] \),
    • \( v \in \{w\}^+ Eps(w) \)
Algorithm

procedure SearchRegularExpression

/*
| T[c] is the database character at position c
| N(c) is a subset of nodes in regular expression R s.t.
|   - v in N(c) iff there is a path from s to v in R
|   that generates T[1..c]
| The Regular expression is described by the following:
|   - A start node s, and end node t
|   - Transition matrix A[v,X] gives the node that v transitions to upon
|         reading symbol X.
|   - Eps(v) is the set of nodes that are reachable from v using epsilon
|         transitions.
*/

N(0) = \{s\} \cup Eps(s)

for c = 1 to n
    N(c) = \phi
    for all (u \in N(c - 1))
        if (v = A[u, T[c]] \neq \phi)
            N(c) = N(c) \cup \{v\} \cup Eps(v)
    end
    if t \in N(c)
        print “T matches the regular expression”
    end
end

Figure 3: An \(O(mn)\) time algorithm that tests if a prefix of \(T\) matches regular expression \(R\)
The final step

- We have answered the question:
  - Is $D[1..c]$ accepted by $R$?
  - Yes, if $\text{END} \in N[c]$

- We need to answer
  - Is $D[l..c]$ (for some $l$, and some $c$) accepted by $R$
Regular expressions as Protein sequence motifs

- Problem: if there is a mis-match, the sequence is not accepted.
QUIZ!

• Question:
• your ‘friend’ likes to gamble.
• He tosses a coin: TAILS, he gives you a dollar. HEADS, you give him a dollar.
• Usually, he uses a fair coin, but ‘once in a while’, he uses a loaded coin.
• Can you say what fraction of the times he loads the coin?
• Suppose you know that in the loaded coin Pr(H)=0.8. Also, that he switches coins with 30% probability.
Representation 2: Profiles

• Profiles versus regular expressions
  – Regular expressions are intolerant to an occasional mis-match.
  – The Union operation (I+V+L) does not quantify the relative importance of I,V,L. It could be that V occurs in 80% of the family members.
  – Profiles capture some of these ideas.
Profiles

- Start with an alignment of strings of length m, over an alphabet A,
- Build an $|A| \times m$ matrix $F=(f_{ki})$
- Each entry $f_{ki}$ represents the frequency of symbol $k$ in position $i$
Profile

- Start with an alignment of strings of length $m$, over an alphabet $A$,
- Build an $|A| \times m$ matrix $F=(f_{ki})$
- Each entry $f_{ki}$ represents the frequency of symbol $k$ in position $i$

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>K</th>
<th>L</th>
<th>L</th>
<th>S</th>
<th>H</th>
<th>C</th>
<th>L</th>
<th>L</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>F</td>
<td>K</td>
<td>A</td>
<td>F</td>
<td>G</td>
<td>Q</td>
<td>T</td>
<td>M</td>
<td>F</td>
<td>Q</td>
</tr>
<tr>
<td>Y</td>
<td>P</td>
<td>I</td>
<td>V</td>
<td>G</td>
<td>Q</td>
<td>E</td>
<td>L</td>
<td>L</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>P</td>
<td>V</td>
<td>V</td>
<td>K</td>
<td>E</td>
<td>A</td>
<td>I</td>
<td>L</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>K</td>
<td>V</td>
<td>L</td>
<td>A</td>
<td>A</td>
<td>V</td>
<td>I</td>
<td>A</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>E</td>
<td>F</td>
<td>I</td>
<td>S</td>
<td>E</td>
<td>C</td>
<td>I</td>
<td>I</td>
<td>Q</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>K</td>
<td>L</td>
<td>L</td>
<td>G</td>
<td>N</td>
<td>V</td>
<td>L</td>
<td>V</td>
<td>C</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.71</th>
<th>0.14</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td>0.28</td>
</tr>
<tr>
<td>Q</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td></td>
<td>0.14</td>
</tr>
</tbody>
</table>
Given a sequence $s$, does it belong to the family described by a profile? We align the sequence to the profile, and score it. Let $S(i,j)$ be the score of aligning position $i$ of the profile to residue $s_j$. The score of an alignment is the sum of column scores.
Scoring Profiles

\[ S(i,j) = \sum_{k} f_{ki} M[r_k, s_j] \]

Scoring Matrix

<table>
<thead>
<tr>
<th>A</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>P</th>
<th>Q</th>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>-10</td>
<td>-33</td>
<td>0</td>
<td>-15</td>
<td>12</td>
<td>28</td>
<td>-28</td>
<td>-12</td>
<td>25</td>
<td>19</td>
<td>15</td>
<td>10</td>
<td>-24</td>
<td>-26</td>
<td>1</td>
<td>-28</td>
</tr>
<tr>
<td>-8</td>
<td>-18</td>
<td>-33</td>
<td>-25</td>
<td>14</td>
<td>1</td>
<td>-32</td>
<td>-28</td>
<td>-25</td>
<td>27</td>
<td>19</td>
<td>14</td>
<td>-24</td>
<td>-14</td>
<td>6</td>
<td>-27</td>
</tr>
<tr>
<td>3</td>
<td>-10</td>
<td>-8</td>
<td>-8</td>
<td>-7</td>
<td>-2</td>
<td>-21</td>
<td>-16</td>
<td>-1</td>
<td>-24</td>
<td>-28</td>
<td>-19</td>
<td>-12</td>
<td>-24</td>
<td>-17</td>
<td>-9</td>
</tr>
<tr>
<td>-2</td>
<td>-18</td>
<td>-1</td>
<td>-12</td>
<td>-10</td>
<td>-8</td>
<td>-5</td>
<td>-2</td>
<td>-1</td>
<td>-24</td>
<td>-27</td>
<td>-19</td>
<td>12</td>
<td>-22</td>
<td>-17</td>
<td>-9</td>
</tr>
<tr>
<td>0</td>
<td>-24</td>
<td>-25</td>
<td>-25</td>
<td>-19</td>
<td>5</td>
<td>15</td>
<td>16</td>
<td>19</td>
<td>5</td>
<td>19</td>
<td>16</td>
<td>16</td>
<td>19</td>
<td>19</td>
<td>19</td>
</tr>
</tbody>
</table>

October 30, 2014
Domain analysis via profiles

• Given a database of profiles of known domains/families, we can query our sequence against each of them, and choose the high scoring ones to functionally characterize our sequences.

• What if the sequence matches some other sequences weakly (using BLAST), but does not match any known profile?
Psi-BLAST idea

- **Iterate:**
  - Find homologs using Blast on query
  - Discard very similar homologs
  - Align, make a profile, search with profile.
  - Why is this more sensitive?

October 30, 2014
Psi-BLAST speed

- **Two time consuming steps.**
  1. Multiple alignment of homologs
  2. Searching with Profiles.

  1. Does the keyword search idea work?

- **Multiple alignment:**
  - Use ungapped multiple alignments only

- **Pigeonhole principle again:**
  - If profile of length $m$ must score $\geq T$
  - Then, a sub-profile of length $l$ must score $\geq lT/m$
  - Generate all $l$-mers that score at least $lT/M$
  - Search using an automaton

October 30, 2014
Representation 3: HMMs

- Building good profiles relies upon good alignments.
  - Difficult if there are gaps in the alignment.
  - Psi-BLAST/BLOCKS etc. work with gapless alignments.
- An HMM representation of Profiles helps put the alignment construction/membership query in a uniform framework.
- Also allows for position specific gap scoring.

October 30, 2014
QUIZ

• Question:
  • your ‘friend’ likes to gamble.
  • He tosses a coin: HEADS, he gives you a dollar. TAILS, you give him a dollar.
  • Usually, he uses a fair coin, but ‘once in a while’, he uses a loaded coin.
  • Can you say what fraction of the times he loads the coin?

October 30, 2014
The generative model

- Think of each column in the alignment as an automaton generating characters.
- For each column, build a node that outputs a residue with the appropriate distribution.
A simple Profile HMM

- Connect nodes for each column into a chain.
- In each step, generate a character, then move to the next node. The chain generates random sequences.
- What is the probability of generating FKVVGQVILD?
- In this representation
  - \( \text{Prob}[\text{New sequence } S \text{ belongs to a family}] = \text{Prob}[\text{HMM generates sequence } S] \)
- What is the difference with Profiles?

October 30, 2014
Profile HMMs can handle gaps

- The match states are the same as on the previous page.
- Insertion and deletion states help introduce gaps.
- In each step, generate a symbol, then move to one of the possible neighbors.
- A sequence may be generated using different paths.
• Probability [ALIL] is part of the family?
• Note that multiple paths can generate this sequence.
  – $M_1I_1M_2M_3$
  – $M_1M_2I_2M_3$
• In order to compute the probabilities, we must assign probabilities of transition between states
Profile HMMs

- The emitted sequence is $S = S_1S_2\ldots S_m$
- The path traversed is $P_1P_2P_3\ldots$ (hidden)
- Joint probability of seeing a sequence $S$, and path $P$
  - $\Pr[S,P|\mathcal{M}] = \Pr[S|P,\mathcal{M}]\Pr[P|\mathcal{M}]$
  - $\Pr[\text{ALIL AND } M_1I_1M_2M_3|\mathcal{M}]$
    $= \Pr[\text{ALIL}| M_1I_1M_2M_3,\mathcal{M}]\Pr[M_1I_1M_2M_3|\mathcal{M}]$
- $\Pr[\text{ALIL} | \mathcal{M}] = ?$
HMM

- Automaton $M = (Q, T, \pi, \Sigma, e)$
- At first, $M$ goes to initial state $j$ with probability $\pi_j$
- In state $j$, $M$ emits a symbol from $\Sigma$ according to $e_j$, and moves to state $k$ with probability $T[j,k]$. 

$T[j,k]$: probability of moving from state $j$ to state $k$

$\pi$: a set of states

$T$: a matrix of transition probabilities

$\Sigma$: a set of symbols

$e_j$: the probability of emitting $S$ while in state $j$. 

October 30, 2014
Profile HMM

- $e_j(s) = \text{emission probability of symbol } s \text{ in state } P_j$
- Transition probability $T[j,k] : \text{Probability of transitioning from state } j \text{ to state } k$.
- $Pr(P,S|M) = e_{P1}(S_1) \cdot T[P_1,P_2] \cdot e_{P2}(S_2) \ldots$
- What is $Pr(S|M)$?
Two solutions

- An unknown (hidden) path is traversed to produce (emit) the sequence S.
- The probability that $M$ emits S can be either
  - The sum over the joint probabilities over all paths.
    - $\Pr(S|M) = \sum_{P} \Pr(S,P|M)$
  - OR, it is the probability of the most likely path
    - $\Pr(S|M) = \max_{P} \Pr(S,P|M)$
- Both are appropriate ways to model, and have similar algorithms to solve them.
Viterbi Algorithm for HMM

- Let $P_{\text{max}}(i,j|M)$ be the probability of the most likely solution that emits $S_1 \ldots S_i$, and ends in state $j$ (is it sufficient to compute this?)
• $P_{\text{max}}(i,j|M) = \max_k P_{\text{max}}(i-1,k) \ T[k,j] \ e_j(S_i)$ (Viterbi)

• $P_{\text{sum}}(i,j|M) = \sum_k (P_{\text{sum}}(i-1,k) \ T[k,j]) \ e_j(S_i)$
We can use the Viterbi/Sum algorithm to compute the probability that the sequence belongs to the family.

Backtracking can be used to get the path, which allows us to give an alignment.

Path: $M_1 M_2 I_2 M_3$
Summary

- HMMs allow us to model position specific gap penalties, and allow for automated training to get a good alignment.
- Patterns/Profiles/HMMs allow us to represent families and focus on key residues
- Each has its advantages and disadvantages, and needs special algorithms to query efficiently.
Protein Domain databases

- A number of databases capture proteins (domains) using various representations.
- Each domain is also associated with structure/function information, parsed from the literature.
- Each database has specific query mechanisms that allow us to compare our sequences against them, and assign function.

October 30, 2014
HMM ‘fair-coin’ example

\[ E_F(H) = 0.5 \]

\[ E_L(H) = 0.1 \]
H H T T T T is the observed sequence