Lecture 7:
Gene finding and regulatory motif analysis

December 1, 2015
Gene Finding

Sources:
• Lecture notes of Larry Ruzzo, UW.
• Slides by Nir Friedman, Hebrew U.
• Slides by Chuong Huynh on Gene Prediction, NCBI
• Durbin’s book, Ch. 3
• Pevzner’s book, Ch. 9
Motivation

• ~3Gb human DNA in GenBank
• Only ~1.5% of human DNA is coding for proteins
• 202,237,081,559 total bases in GenBank (10/2015)
• Hundreds of species have been sequenced, thousands to follow
• Total number of species represented in UniProtKB/Swiss-Prot (2015): 13,251
• Need to locate the genes!
• **Goal**: Automatic finding of genes
Reminder: The Genetic Code

1 start, 3 stop codons
Genes in Prokaryotes

- High gene density (e.g. 70% coding in H. Influenza)
- No introns
- most long ORFs are likely to be genes.
Open Reading Frames

- **Reading Frame:** 3 possible ways to read the sequence (on each strand).
  - ACCUUAGCGUA = Threonine-Leucine-Alanine
  - ACCUUAGCGUA = Proline-Stop-Arginine
  - ACCUUAGCGUA = Leucine-Serine-Valine
- **Open Reading Frame (ORF):** Reading frame with no stop codons.
  - ORF is maximal if it starts right after a stop and ends in a stop
- **Untranslated region (UTR):** ends of the mRNA (on both sides) that are not translated to protein.
Finding long ORFs

- In random DNA, one stop codon every $64/3 \rightarrow 21$ codons on average
- Average protein is ~300 AA long
- => search long ORFs
- Problems:
  - short genes
  - many more ORFs than genes
    - In E. Coli one finds 6500 ORFs but only 1100 genes.
    - Call the remaining Non-coding ORF (NORFS)
  - Overlapping long ORFs on opposite strands
Codon Frequencies

• Coding DNA is not random:
  - In random DNA, expect
    • Leucine:Alanine:Tryptophan ratio of 6:4:1
  - In real proteins, 6.9:6.5:1
  - In some species, 3rd position of the codon, up to 90% A or T

• Different frequencies for different species.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th>frequency of usage of each codon (per thousand)</th>
<th>relative freq of each codon among synonymous codons</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gly</td>
<td>GGG</td>
<td>17.08 0.23</td>
<td></td>
<td>Arg</td>
<td>AGG</td>
<td>12.09 0.22</td>
</tr>
<tr>
<td>Gly</td>
<td>GGA</td>
<td>19.31 0.26</td>
<td></td>
<td>Arg</td>
<td>AGA</td>
<td>11.73 0.21</td>
</tr>
<tr>
<td>Gly</td>
<td>GGT</td>
<td>13.66 0.18</td>
<td></td>
<td>Ser</td>
<td>AGT</td>
<td>10.18 0.14</td>
</tr>
<tr>
<td>Gly</td>
<td>GGC</td>
<td>24.94 0.33</td>
<td></td>
<td>Ser</td>
<td>AGC</td>
<td>18.54 0.25</td>
</tr>
<tr>
<td>Glu</td>
<td>GAG</td>
<td>38.82 0.59</td>
<td></td>
<td>Lys</td>
<td>AAG</td>
<td>33.79 0.6</td>
</tr>
<tr>
<td>Glu</td>
<td>GAA</td>
<td>27.51 0.41</td>
<td></td>
<td>Lys</td>
<td>AAA</td>
<td>22.32 0.4</td>
</tr>
<tr>
<td>Asp</td>
<td>GAT</td>
<td>21.45 0.44</td>
<td></td>
<td>Asn</td>
<td>AAT</td>
<td>16.43 0.44</td>
</tr>
<tr>
<td>Asp</td>
<td>GAC</td>
<td>27.06 0.56</td>
<td></td>
<td>Asn</td>
<td>AAC</td>
<td>21.3  0.56</td>
</tr>
<tr>
<td>Val</td>
<td>GTG</td>
<td>28.6  0.48</td>
<td></td>
<td>Met</td>
<td>ATG</td>
<td>21.86 1</td>
</tr>
<tr>
<td>Val</td>
<td>GTA</td>
<td>6.09  0.1</td>
<td></td>
<td>Ile</td>
<td>ATA</td>
<td>6.05  0.14</td>
</tr>
<tr>
<td>Val</td>
<td>GTT</td>
<td>10.3  0.17</td>
<td></td>
<td>Ile</td>
<td>ATT</td>
<td>15.03 0.35</td>
</tr>
<tr>
<td>Val</td>
<td>GTC</td>
<td>15.01 0.25</td>
<td></td>
<td>Ile</td>
<td>ATC</td>
<td>22.47 0.52</td>
</tr>
<tr>
<td>Ala</td>
<td>GCG</td>
<td>7.27  0.1</td>
<td></td>
<td>Thr</td>
<td>ACG</td>
<td>6.8  0.12</td>
</tr>
<tr>
<td>Ala</td>
<td>GCA</td>
<td>15.5  0.22</td>
<td></td>
<td>Thr</td>
<td>ACA</td>
<td>15.04 0.27</td>
</tr>
<tr>
<td>Ala</td>
<td>GCT</td>
<td>20.23 0.28</td>
<td></td>
<td>Thr</td>
<td>ACT</td>
<td>13.24 0.23</td>
</tr>
<tr>
<td>Ala</td>
<td>GCC</td>
<td>28.43 0.4</td>
<td></td>
<td>Thr</td>
<td>ACC</td>
<td>21.52 0.38</td>
</tr>
</tbody>
</table>

Human codon usage

http://genome.imim.es/courses/Lisboa01/slide3.8.html
First Order Markov Model

- Use two Markov models (similar to CpG islands) to discriminate genes from NORFs
- Given a sequence of nucleotides $X_1, \ldots, X_n$ we compute the log-odds ratio:

$$\log \frac{P(X_1, \ldots, X_n \mid G)}{P(X_1, \ldots, X_n \mid R)} = \sum_i \log \frac{A^G_{X_iX_{i+1}}}{A^R_{X_iX_{i+1}}}$$

- Bit= unit when log base 2 is used
First Order Markov Model

- Average log-odds per nucleotide in genes: 0.018
- Average log-odds per nucleotide in NORFs: 0.009
- But the variance makes it useless for discrimination (similar results for 2nd-order MM)

Test on E. Coli data

Durbin et al pp.74
Using codons

- Translate each ORF into a sequence of codons
- Form a 64-state Markov chain
  - Codon is more informative than its translation
- Estimate probabilities in coding regions and NORFs
Using Codon Frequencies

• Assume each codon is iid
• For codon \(abc\) calculate frequency \(f_{abc}\) in coding region
• Given coding sequence \(a_1b_1c_1, \ldots, a_{n+1}b_{n+1}c_{n+1}\)
• Calculate

\[
\begin{align*}
    p_1 &= f_{a_1b_1c_1} \ast f_{a_2b_2c_2} \ast \ldots \ast f_{a_nb_nc_n} \\
    p_2 &= f_{b_1c_1a_2} \ast f_{b_2c_2a_3} \ast \ldots \ast f_{b_nc_na_{n+1}} \\
    p_3 &= f_{c_1a_2b_2} \ast f_{c_2a_3b_3} \ast \ldots \ast f_{c_na_{n+1}b_{n+1}}
\end{align*}
\]

• The probability that the \(i\)-th reading frame is the coding region:

\[
P_i = \frac{p_i}{p_1 + p_2 + p_3}
\]
RNA Transcription

- Not all ORFs are expressed.
- Transcription depends on regulatory signals.
- Minimal regulatory region - core promoter to which RNA polymerase and initiation factors bind to start transcription.
- At the termination signal the polymerase releases the RNA and disconnects from the DNA.
**E. coli** promoters

**consensus sequence:**

\[ \text{nnnTTGACA} \text{nnnnnnnnnnnnnnnnnnnnnnnnnnnnTATAAT} \text{nNNnnnnnn} \text{nnn} \]

-35     \[\text{ mRNA start}\]

-12 

- • “**TATA box**” (or Pribnow Box)
- • Not exact
Positional Weight Matrix (PWM)

- $f_{b,j}$: frequency of base $b$ in position $j$.
- Assumes independence btw positions
- For TATA box:

<table>
<thead>
<tr>
<th>pos:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>95</td>
<td>26</td>
<td>59</td>
<td>51</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>9</td>
<td>2</td>
<td>14</td>
<td>13</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>G</td>
<td>10</td>
<td>1</td>
<td>16</td>
<td>15</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>79</td>
<td>3</td>
<td>44</td>
<td>13</td>
<td>17</td>
<td>96</td>
</tr>
</tbody>
</table>

- $f_b$: background frequency.
Scoring Function

• For sequence $S=B_1 B_2 B_3 B_4 B_5 B_6$

\[
P(S \mid \text{promoter}) = \prod_{i=1}^{6} f_{B_i,i}
\]

\[
P(S \mid \text{non-promoter}) = \prod_{i=1}^{6} f_{B_i}
\]

• Log-likelihood ratio score:

\[
\log \left( \frac{P(S \mid \text{promoter})}{P(S \mid \text{non-promoter})} \right) = \log \left( \frac{\prod_{i=1}^{6} f_{B_i,i}}{\prod_{i=1}^{6} f_{B_i}} \right) = \sum_{i=1}^{6} \log \left( \frac{f_{B_i,i}}{f_{B_i}} \right)
\]
Gene finding: coding density

- As the coding/non-coding length ratio decreases, exon prediction becomes more complex

- Human
- Fugu
- Worm
- E.coli
Eukaryote gene structure

Eukaryotes
Typical structure at DNA level
(not to scale)

- 5' Untranslated region (UTR)
- Intron(s)
- Terminal exon
- Polyadenylation site
- 3' UTR
- Promoter
- Start codon
- Initial exon
- Internal exon
- Donor site
- Acceptor site
- Splice sites
- Transcription start site
Typical figures: vertebrates

- TF binding site: ~6bp; 0-2kbp upstream of TSS
- 5' UTR: ~750 bp, 3' UTR: ~450bp
- Gene length: 30kb, coding region: 1-2kb
- Average of 6 exons, 150bp long
- Huge variance: - dystrophin: 2.4Mb long
  - Blood coagulation factor: 26 exons, 69bp to 3106bp; intron 22 contains another unrelated gene
Splicing

• Splicing: the removal of the introns.
• Performed by complexes called spliceosomes, containing both proteins and snRNA.
• The snRNA recognizes the splice sites through RNA-RNA base-pairing.
• Recognition must be precise: a 1nt error can shift the reading frame making nonsense of its message.
• Many genes have alternative splicing, which changes the protein created.
Exon-intron junctions

Donor site: AGGUAAGU

Branchpoint: CTGAC

Acceptor site: NCAGG

- Pyrimidine [C,T] rich
- 1st approach: positional weight matrices
  - Problematic with weak/short signals
  - Does not exploit all info (reading frames, intron/exon stats...)

⇒ try integrated approaches!
Length Distribution

Since an HMM is a memory-less process, the only length distribution that can be modeled is geometric.

Above is a simple HMM for gene structure.

The length of each exon (intron) has a geometric distribution:

\[ P(\text{exon of length } k) = p^k (1 - p) \]

Since an HMM is a memory-less process, the only length distribution that can be modeled is geometric.
Exon Length Distribution

• **Intron length distribution** seems approximately geometric.
• This is not so for **exons**.
• Length seems to have a functional role on the splicing itself:
  - Too short (under 50bps): the spliceosomes have no room
  - Too long (over 300bps): ends have problems finding each other.
  - But as usual there are exceptions.

=> Need a different model for exons.
Generalized HMM
(Burge & Karlin, J. Mol. Bio. 97 268 78-94)

- Hidden Markov states $q_1, \ldots q_n$
- State $q_i$ has output length distribution $f_i$
- Output of each state can have a separate probabilistic model (weight matrix, MM...)
- Initial state probability distribution $\pi$
- State transition probabilities $T_{ij}$
GenScan Model

Exon
Intron
Exon init/term
5'/3' UTR
Promoter/PolyA

Fig. 2. Gene model

Forward strand

Backward strand

Burge & Karlin JMB 97
GenScan model

- states = functional units on a gene
- The allowed transitions ensure the order is biologically consistent.
- As an intron may cut a codon, one must keep track of the reading frame, hence the three $I$ phases:
  - phase $I_0$: between codons
  - phase $I_1$: introns that start after 1st base
  - phase $I_2$: introns that start after 2nd base
Signal Models

• Genscan uses different models to model the different biological signals
  - Weight Matrix Model
    • Position specific distribution.
    • Each column is independent
  - Used for
    • Translation initiation signal
    • Translation termination signal
    • promoters
    • polyadenylation signals
Splice Sites

• Correct recognition of these sites greatly enhances ability to predict correct exon boundaries.

• Used Weighted Array Model: a generalization of PWM that allows for dependencies between adjacent positions

• Accurate modeling of these sites led to substantial improvement in performance.
**GenScan Performance**

---

### Accuracy of GENSCAN for different signal and exon types.

**(a) Prediction of individual splice sites and translational signals.**

<table>
<thead>
<tr>
<th>Type of signal</th>
<th>Type of exon</th>
<th>Annotated exons</th>
<th>Predicted exons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Number</td>
<td>% Correctly predicted</td>
</tr>
<tr>
<td>Initiation</td>
<td>Initial only</td>
<td>570</td>
<td>66</td>
</tr>
<tr>
<td>Termination</td>
<td>Terminal only</td>
<td>570</td>
<td>78</td>
</tr>
<tr>
<td>5’ splice site</td>
<td>Initial only</td>
<td>570</td>
<td>88</td>
</tr>
<tr>
<td>5’ splice site</td>
<td>Internal only</td>
<td>1510</td>
<td>93</td>
</tr>
<tr>
<td>5’ splice site</td>
<td>Initial and internal</td>
<td>2080</td>
<td>91</td>
</tr>
<tr>
<td>3’ splice site</td>
<td>Terminal only</td>
<td>570</td>
<td>81</td>
</tr>
<tr>
<td>3’ splice site</td>
<td>Internal only</td>
<td>1510</td>
<td>92</td>
</tr>
<tr>
<td>3’ splice site</td>
<td>Internal and terminal</td>
<td>2080</td>
<td>89</td>
</tr>
</tbody>
</table>

**Note:**

- Predicts correctly 80% of exons
- Prediction accuracy per bp > 90%
Fig. 12. GENSCAN PostScript output for sequence HSNCAMX1

Key: □ Initial exon □ Internal exon △ Terminal exon ▽ Single-exon gene
- Annotated exon
- Predicted exon
Sam Karlin, Chris Burge
Regulatory sequence analysis

Slides with Chaim Linhart
Regulation of Transcription

- A gene’s transcription regulation is mainly encoded in the DNA in a region called the **promoter**
- Each promoter contains several short DNA subsequences, called **binding sites (BSs)** that are bound by specific proteins called **transcription factors (TFs)**
Regulation of Transcription (II)

Assumption:

Co-expression

\[ \downarrow \]

Transcriptional co-regulation

\[ \downarrow \]

Common BSs
WH-questions

• √ Why are we looking for common BSs?
• What exactly are we trying to find?
• Where should we look for it?
• How can we find it?
What is the promoter region?

- **Upstream Transcription Start Site (TSS)**
  - Too short → miss many real BSs (false negatives)
  - Too long → lots of wrong hits (false positives)
  - Length is species dependent (e.g., yeast ~600bp, thousands in human)
    - Common practice: ~ 500-2000bp

- **Consider both strands?**
  - Common practice: Yes
What: Models for Binding Sites
(I) Exact string(s)

Example:

\[ BS = \text{TACACC}, \text{TACGGC} \]

CAATGCAGGAT\text{TACACC}GATCGGTA
GGAG\text{TACGGC}AAAGTCCCCCATGTGA
AGGCTGGGACCA\text{TACACC}CTC

In red: hits
(II) String with mismatches

**Example:**

\[ \text{BS} = \text{TACACC} + 1 \text{ mismatch} \]

CAATGCGAGGATTCACCAGATCGGTA
GGAGTACAGCAAGTCCCCATGTGA
AGGCTGGACCAGACTCTACCTA
(III) Degenerate string

Example:

BS = TASDAC \( S = \{ C, G \} \quad D = \{ A, G, T \} \)

CAATGCAGGATACACGATCGGTA
GGAGTAGTACAAGTCCCCATGTGA
AGGCTGGAGCCAGACTCTCCTACGACTA
(IV) Position Weight Matrix (PWM)

a.k.a Position Specific Scoring Matrix (PSSM)

Example:

Score: product of base probabilities.

Need to set score threshold for hits.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.1</td>
<td>0.8</td>
<td>0</td>
<td>0.7</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0.1</td>
<td>0.5</td>
<td>0.1</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.1</td>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>T</td>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
<td>0.3</td>
</tr>
</tbody>
</table>

ATGCAGGATACACCAGATCGGTA  0.0605
GGAGTAGAGCAAGTCCCGTGA  0.0605
AAGACTCTACAATTTATGGCGGT  0.0151
How: Experimental techniques
Protein Binding Microarrays
Berger et al, Nat. Biotech 2006

- Generate an array of double-stranded DNA with all possible k-mers
- Detect TF binding to specific k-mers
Chromatin Immunoprecipitation (ChIP)

DNA-binding proteins are crosslinked to DNA with formaldehyde in vivo.

Isolate the chromatin. Shear DNA along with bound proteins into small fragments.

Bind antibodies specific to the DNA-binding protein to isolate the complex by precipitation. Reverse the cross-linking to release the DNA and digest the proteins.

Identify bound DNA via microarray hybridization or sequencing

http://www.bio.brandeis.edu/haberlab/jehsite/chip.html
How: Analyzing known motifs
**Goal:** Identify enriched TFs = TFs whose BSs are over-represented in promoters of co-regulated genes

- Prepare a dictionary of motif hits
- Compute enrichment of hits in the given promoter set compared to a background set.
Computation of Motif Hits

Computing a threshold for a PWM:

- Compute 2\textsuperscript{nd}-order Markov-Model of background sequences
- Generate random sequences using MM (e.g., 1,000 sequences of length 1,000)
- Set threshold s.t. PWM has \(~5\%\) hits at random.

This “ensures” a pre-defined false-positive rate, but no guarantee on false-negative rate.
Motif Enrichment

Each promoter is hit or not.

Let:  \( B \) = total # of promoters (BG)
      \( T \) = # of target-set promoters
      \( b \) = total # of promoters that are hit
      \( t \) = # of target-set promoters that are hit

Then (hypergeometric distribution assumption):

Prob. for \( t \) hits in target-set:

\[
P(t) = \binom{b}{t} \binom{B-b}{T-t} / \binom{B}{T}
\]

Prob. for at least \( t \) hits:

\[
p-value = \sum_{i=t}^{\min\{b,T\}} P(i)
\]
TF Synergism

Find pairs of TFs that tend to occur in the same promoters

Let: \( T = \# \text{ of promoters in target-set} \)

\( t_1, t_2 = \# \text{ of promoters hit by TF 1,2} \)

\( t_{12} = \# \text{ of promoters hit by both TFs (w/o overlaps!)} \)

Then:

Prob. for co-occurrence of at least \( t_{12} \):

\[
\text{synergism \ p-value} = \sum_{i \geq t_{12}} \binom{t_1}{i} \binom{T - t_1}{t_2 - i} \binom{T}{t_2}
\]
Whitfield et al. (’02) identified 568 genes that are periodically expressed in the human cell-cycle and partitioned them into the 5 phases of the cell-cycle.
PRIMA: results on HCC

PRIMA found 8 enriched TFs in the 568 HCC genes (w.r.t. 13K BG promoters):
Results on HCC (III)

Co-occurring pairs of TFs:
How: Motif finding
Bailey & Elkan ZOOPS model

• $n$ sequences, $m$ possible motif positions per sequence.
• **Assumption**: each sequence contains zero or one occurrence of the motif.
• Prior probability for one occurrence: $\gamma$
• Prior probability for motif in position $j$: $\lambda = \gamma/m$
• *What is the hidden data?*
• *What is the $Q$ function?*
• $Z_{ij}$ indicator for motif at sequence $i$, position $j$.
• $Q_i$ indicator for motif in sequence $i$.

$$\log Pr(X, Z|\theta, \gamma)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{i,j} \log Pr(X_i|Z_{i,j} = 1, \theta)$$

$$+ \sum_{i=1}^{n} (1 - Q_i) \log Pr(X_i|Q_i = 0, \theta)$$

$$+ \sum_{i=1}^{n} (1 - Q_i) \log(1 - \gamma) + \sum_{i=1}^{n} Q_i \log \lambda$$

$Z_{i,j}^{(t)} = \frac{f_i}{f_0 + \sum_{k=1}^{m} f_k}$, where

$f_0 = Pr(X_i|Q_i = 0, \theta^{(t)})(1 - \gamma^{(t)})$, and

$f_j = Pr(X_i|Z_{i,j} = 1, \theta^{(t)})\lambda^{(t)}$, $1 \leq j \leq m$

(ex.)