Gene Structure Prediction
(Gene Finding)

Yuzhen Ye (yye@indiana.edu)
School of Informatics & Computing, IUB
Contents

- Gene prediction problem
- Approaches
  - Similarity search based
  - De novo approaches considering sequence patterns of different elements (exons, introns, etc)
  - Approaches that integrate other evidences (e.g., RNA-seq)
Gene prediction

aatgcataactgcggctatgctaataagcaatgcggctatgctaagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
tagcataatgcggctatgctaataagcaatgcggctatgctaagctgggatccga
What’s a gene

- What is a gene, post-ENCODE?
  - Gerstein et al. Genome Res. 2007 17: 669-681
  - ENCODE consortium: characterization of 1% of the human genome by experimental and computational techniques

- Definitions:
  - Definition 1970s–1980s: Gene as open reading frame (ORF) sequence pattern
  - Definition 1990s–2000s: Annotated genomic entity, enumerated in the databanks
  - The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products
Post-ENCODE definition

The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products.
Can we still do gene prediction?

- **Prokaryotic genes**

- **Eukaryotic genes**
Gene predictors (for protein-coding genes)

- Similarity based predictors
- Statistical approaches (ab initio gene predictors)
  - Predictions depend on analysis of a variety of sequence patterns that are characteristic of exons, intron-exon boundaries, upstream regulatory regions, etc.
  - Hidden Markov models (a probabilistic sequence model), Neural networks (NN) and other methods are used to combine sequence content and signal classifiers into a consistent gene structure
De novo (ab initio) gene predictors

- **GLIMMER**: uses interpolated Markov model (for prokaryotic gene prediction)
- **GeneMark**: GeneMark GeneMark.hmm (for both prokaryotic and eukaryotic gene prediction)
- **Eukaryotic genes predictors**
  - **GENSCAN**: uses Hidden Markov Models (HMMs)
  - **TwinScan (GenomeScan)**
    - Uses both HMM and similarity (e.g., between human and mouse genomes)
Gene prediction is easier in microbial genomes

- Microbial genome tends to be gene rich (80%-90% of the sequence is coding)
- Often no introns
- Highly conserved patterns in the promoter region, transcription and translation start site
Prokaryote gene structure

- **Transcribed region**
  - **start codon**
  - **stop codon**

- **Coding region**
- **Untranslated regions**

- **Promoter**
- **Transcription start site**
- **Transcription stop site**

- `-k` denotes \( k \)th base before transcription, \( +k \) denotes \( k \)th transcribed base
Open reading frame (ORF)

ORF is a sequence of codons which starts with start codon, ends with an end codon and has no end codons in-between.

*Searching for ORFs – consider all 6 possible reading frames: 3 forward and 3 reverse (next slide)*

**Is the ORF a coding sequence?**

1. Must be long enough (roughly 300 bp or more)
2. Should have average amino-acid composition specific for the given organism.
3. Should have codon use specific for the given organism.
Six frame translation of a DNA sequence

- stop codons – TAA, TAG, TGA
- start codons - ATG
The genetic code

UAA, UAG and UGA correspond to 3 Stop codons that (together with Start codon ATG) delineate Open Reading Frames

The Genetic Code
Codon usage

- Codon: 3 consecutive nucleotides
- $4^3 = 64$ possible codons
- Genetic code is degenerative and redundant
  - Includes start and stop codons
  - An amino acid may be coded by more than one codon (degeneracy & wobbling pairing)
  - Certain codons are more in use
- Uneven use of the codons may characterize a real gene
## Codon usage in Mouse genome

<table>
<thead>
<tr>
<th>AA</th>
<th>codon</th>
<th>/1000</th>
<th>frac</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ser</td>
<td>TCG</td>
<td>4.31</td>
<td>0.05</td>
</tr>
<tr>
<td>Ser</td>
<td>TCA</td>
<td>11.44</td>
<td>0.14</td>
</tr>
<tr>
<td>Ser</td>
<td>TCT</td>
<td>15.70</td>
<td>0.19</td>
</tr>
<tr>
<td>Ser</td>
<td>TCC</td>
<td>17.92</td>
<td>0.22</td>
</tr>
<tr>
<td>Ser</td>
<td>AGT</td>
<td>12.25</td>
<td>0.15</td>
</tr>
<tr>
<td>Ser</td>
<td>AGC</td>
<td>19.54</td>
<td>0.24</td>
</tr>
<tr>
<td>Pro</td>
<td>CCG</td>
<td>6.33</td>
<td>0.11</td>
</tr>
<tr>
<td>Pro</td>
<td>CCA</td>
<td>17.10</td>
<td>0.28</td>
</tr>
<tr>
<td>Pro</td>
<td>CCT</td>
<td>18.31</td>
<td>0.30</td>
</tr>
<tr>
<td>Pro</td>
<td>CCC</td>
<td>18.42</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leu</td>
<td>CTG</td>
<td>39.95</td>
<td>0.40</td>
</tr>
<tr>
<td>Leu</td>
<td>CTA</td>
<td>7.89</td>
<td>0.08</td>
</tr>
<tr>
<td>Leu</td>
<td>CTT</td>
<td>12.97</td>
<td>0.13</td>
</tr>
<tr>
<td>Leu</td>
<td>CTC</td>
<td>20.04</td>
<td>0.20</td>
</tr>
<tr>
<td>Ala</td>
<td>GCG</td>
<td>6.72</td>
<td>0.10</td>
</tr>
<tr>
<td>Ala</td>
<td>GCA</td>
<td>15.80</td>
<td>0.23</td>
</tr>
<tr>
<td>Ala</td>
<td>GCT</td>
<td>20.12</td>
<td>0.29</td>
</tr>
<tr>
<td>Ala</td>
<td>GCC</td>
<td>26.51</td>
<td>0.38</td>
</tr>
<tr>
<td>Gln</td>
<td>CAG</td>
<td>34.18</td>
<td>0.75</td>
</tr>
<tr>
<td>Gln</td>
<td>CAA</td>
<td>11.51</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Codon frequency

Input sequence

- frequency in coding region
- frequency in non-coding region

Compare

Coding region or non-coding region
A simple calculation assuming independence of nucleotides

<table>
<thead>
<tr>
<th>codon position</th>
<th>A</th>
<th>C</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28%</td>
<td>33%</td>
<td>18%</td>
<td>21%</td>
</tr>
<tr>
<td>2</td>
<td>32%</td>
<td>16%</td>
<td>21%</td>
<td>32%</td>
</tr>
<tr>
<td>3</td>
<td>33%</td>
<td>15%</td>
<td>14%</td>
<td>38%</td>
</tr>
<tr>
<td>frequency in genome</td>
<td>31%</td>
<td>18%</td>
<td>19%</td>
<td>31%</td>
</tr>
</tbody>
</table>

\[
\frac{P(x|\text{ORF})}{P(x|\text{random})} = \prod_i \frac{P(A_i \text{ at ith position})}{P(A_i \text{ in the sequence})}
\]

Score of AAAGAT:

\[
\frac{.28 \times .32 \times .33 \times .21 \times .26 \times .14}{.31 \times .31 \times .31 \times .31 \times .31 \times .19}
\]
Coding region prediction using hexmer frequencies

Coding potential – hexmer frequencies in coding versus in non-coding regions $\sum \log \left( \frac{C_i (X)}{N_i (X)} \right)$
Markov chain model

- Sometimes we need to model dependencies between adjacent positions in the sequence
  - There are certain regions in the genome, like TATA within the regulatory area, upstream a gene.
  - The pattern CG is less common than expected for random sampling.

- Such dependencies can be modeled by Markov chains.
Markov chains

- A Markov chain is a sequence of random variables with Markov property, i.e., given the present state, the future and the past are independent.

- A famous example of Markov chain is the “drunkard's walk”—at each step, the position may change by +1 or −1 with equal probability.
  - \( \Pr(5 \rightarrow 4) = \Pr(5 \rightarrow 6) = 0.5 \), all other transition probabilities from 5 are 0.
  - these probabilities are independent of whether the system was previously in step 4 or 6.
1\textsuperscript{st} order Markov chain

An \textbf{integer time stochastic process}, consisting of a set of \( m>1 \) states \( \{s_1, \ldots, s_m\} \) and

1. An \( m \) dimensional \textbf{initial distribution vector} \( (p(s_1), \ldots, p(s_m)) \)
2. An \( m \times m \) \textbf{transition probabilities matrix} \( M=(a_{s_is_j}) \)

For example, for DNA sequence:
the states are \( \{A, C, T, G\} \) \( (m=4) \)
\( p(A) \) the probability of \( A \) to be the 1\textsuperscript{st} letter
\( a_{AG} \) the probability that \( G \) follows \( A \) in a sequence.
For each integer $n$, a Markov Chain assigns probability to sequences $(x_1 \ldots x_n)$ as follows:

\[
p((x_1, x_2, \ldots, x_n)) = p(X_1 = x_1) \prod_{i=2}^{n} p(X_i = x_i \mid X_{i-1} = x_{i-1}) = p(x_1) \prod_{i=2}^{n} a_{x_{i-1}x_i}
\]
**kth order Markov chain (a Markov chain with memory k)**

- *kth Markov Chain assigns probability to sequences* \((x_1...x_n)\) *as follows:*

\[
p(x_1...x_n) = p(X_1 = x_1, ..., X_k = x_k) \cdot \prod_{i=k}^{n} p(X_i = x_i | X_{i-1} = x_{i-1}, X_{i-2} = x_{i-2}, ..., X_{i-k} = x_{i-k})
\]

| Initial distribution | Transition probabilities |
Inhomogeneous Markov chain for gene finding

Again, the parameters (the transition probabilities, $a$, $b$, and $c$ need to be learned from training samples)
Gene finding using inhomogeneous Markov chain

Consider sequence $x_1x_2x_3x_4x_5x_6x_7x_8x_9...$
where $x_i$ is a nucleotide

let $p_1 = a_{x_1x_2}b_{x_2x_3}c_{x_3x_4}a_{x_4x_5}b_{x_5x_6}c_{x_6x_7}....$
$p_2 = c_{x_1x_2}a_{x_2x_3}b_{x_3x_4}c_{x_4x_5}a_{x_5x_6}b_{x_6x_7}....$
$p_3 = b_{x_1x_2}c_{x_2x_3}a_{x_3x_4}b_{x_4x_5}c_{x_5x_6}a_{x_6x_7}....$

then probability that $i$th reading frame is the coding frame is:

$$P_i = \frac{p_i}{p_1 + p_2 + p_3}$$
Glimmer: Interpolated Markov model based approach

- Made of 2 programs
  - BuildIMM
    - Takes sequences as input and outputs the Interpolated Markov Models (IMMs)
  - Glimmer
    - Takes IMMs and outputs all candidate genes
    - Automatically resolves overlapping genes by choosing one, hence limited
    - Marks “suspected to truly overlap” genes for closer inspection by user
Simple HMM for prokaryotic genes

- **A:** $p_A$
- **C:** $p_C$
- **G:** $p_G$
- **T:** $p_T$

<table>
<thead>
<tr>
<th>States</th>
<th>Transition probability</th>
<th>Emission probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position 1</td>
<td>1</td>
<td>1-$q$</td>
</tr>
<tr>
<td>Position 2</td>
<td>1</td>
<td>$p$</td>
</tr>
<tr>
<td>Position 3</td>
<td>1</td>
<td>1-$p$</td>
</tr>
</tbody>
</table>

Non-Coding Region
Hidden Markov model

- Hidden Markov Model is a Markov model in which one does not observe a sequence of states but results of a function prescribed on states.
- States are hidden to the observers.
- In the simple gene HMM model, states are coding-region and non-coding region. But we observe the sequences of nucleotides.
Assume that at each state a Markov process emits (with some distribution) a symbol from alphabet $\Sigma$.

Rather than observing a sequence of states we observe a sequence of emitted symbols.

Example:
$\Sigma =$\{A,C,T,G\}.
Generate a sequence where A,C,T,G have frequency $p(A) = .33$, $p(G) = .2$, $p(C) = .2$, $p(T) = .27$ respectively

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Emission Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.33</td>
</tr>
<tr>
<td>T</td>
<td>.27</td>
</tr>
<tr>
<td>C</td>
<td>.2</td>
</tr>
<tr>
<td>G</td>
<td>.2</td>
</tr>
</tbody>
</table>

Emission probabilities
HMM: formal definition

HMM is a Markov process that at each time step generates a symbol from some alphabet, $\Sigma$, according to emission probability that depends on state.

$M = (Q, \Sigma, \pi, a, e)$

$Q$ – finite set of states, say n states =\{q_0, q_1, \ldots\}

$a$ – n x n transition probability matrix: $a(i,j) = \Pr[ q_{t+1} = j | g_t = i ]$

$\pi$ – n-vector, start probability vector: $\pi(i) = \Pr[ q_0 = i ]$

$\Sigma = \{ \sigma_1, \ldots, \sigma_k \}$-alphabet

$e(i,j) = \Pr[ o_t = \sigma_j | q_t = i ]$; $o_t$ – t\textsuperscript{th} element of generated sequences = probability of generating $\sigma_j$ in state $q_i$ ($S = o_0, \ldots, o_T$ the output sequence)
Algorithmic problems related to HMM

- Given HMM $M$ and a sequence $S$ compute $Pr(S|M)$ – probability of generating $S$ by $M$.
- Given HMM $M$ and a sequence $S$ compute most likely sequence of states generating $S$.
- What is the most likely state at position $i$.
- Given a representative sample (training set) of a sequence property construct HMM that best models this property.
HMM based gene predictors for microbial genomes

- **GenMark** - [Borodovsky, McInnich – 1993, Comp. Chem., 17, 123-133] 5th order HMM
Additional information for gene prediction

- Upstream regions of genes often contain motifs that can be used for gene prediction.

- ATG
- TATACT
- TTCCAA
- Pribnow Box
- -35
- GGAGG
- Ribosomal binding site
- -10
- Transcription start site
- 0
- 10
- STOP
Promoter structure in prokaryotes

Transcription starts at offset 0.

- Pribnow Box (-10)
- Gilbert Box (-30)
- Ribosomal Binding Site (+10)
Ribosomal binding site

1055 E. coli Ribosome binding sites listed in the Miller book
Eukaryotic gene prediction is more difficult

- In eukaryotes, the gene is a combination of coding segments (exons) that are interrupted by non-coding segments (introns).
- And all the complicating factors as mentioned in the “What is a gene” paper.
- More non-coding regions than coding regions.
- This makes computational gene prediction in eukaryotes even more difficult.
Donor and acceptor sites: GT and AG dinucleotides

- The beginning and end of exons are signaled by donor and acceptor sites that usually have GT and AC dinucleotides.
- Detecting these sites is difficult, because GT and AC appear very often.
### Splicing site signal

#### Donor site

<table>
<thead>
<tr>
<th>Position</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>5'</td>
<td>26</td>
<td>26</td>
<td>25</td>
<td>23</td>
</tr>
<tr>
<td>-8</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>-2</td>
<td>60</td>
<td>15</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>-1</td>
<td>9</td>
<td>5</td>
<td>78</td>
<td>8</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>99</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>98</td>
</tr>
<tr>
<td>2</td>
<td>54</td>
<td>2</td>
<td>41</td>
<td>3</td>
</tr>
<tr>
<td>...</td>
<td>21</td>
<td>27</td>
<td>27</td>
<td>25</td>
</tr>
</tbody>
</table>

#### Table

<table>
<thead>
<tr>
<th>Position</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>5'</td>
<td>26</td>
<td>26</td>
<td>25</td>
<td>23</td>
</tr>
<tr>
<td>-8</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>-2</td>
<td>60</td>
<td>15</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>-1</td>
<td>9</td>
<td>5</td>
<td>78</td>
<td>8</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>99</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>98</td>
</tr>
<tr>
<td>2</td>
<td>54</td>
<td>2</td>
<td>41</td>
<td>3</td>
</tr>
<tr>
<td>...</td>
<td>21</td>
<td>27</td>
<td>27</td>
<td>25</td>
</tr>
</tbody>
</table>
Splice site prediction

- Try to recognize location of splicing signals at exon-intron junctions
  - This has yielded a weakly conserved donor splice site and acceptor splice site
- Profiles for sites are still weak, and lends the problem to the Hidden Markov Model (HMM) approaches, which capture the statistical dependencies between sites
Similarity-based gene finding

- Alignment of
  - Genomic sequence and (assembled) EST sequences
  - Genomic sequence and known (similar) protein sequences (spliced alignment)
  - Two or more similar genomic sequences
Using EST to find exon structure

- Human EST (mRNA) sequence is aligned to different locations in the human genome
- Find the “best” path to reveal the exon structure of human gene
Using similarities to find the exon structure

- The known frog gene is aligned to different locations in the human genome
- Find the “best” path to reveal the exon structure of human gene
Finding local alignments

Use local alignments to find all islands of similarity
Exon chaining problem

- Locate the beginning and end of each interval (2n points)
- Find the “best” path: Given a set of putative exons, find a maximum set of non-overlapping putative exons
Spliced alignment for gene prediction

- **Goal**: Find a chain of blocks in a genomic sequence that best fits a target sequence
- **Input**: Genomic sequences $G$, target sequence $T$, and a set of candidate exons $B$.
- **Output**: A chain of exons such that the global alignment score between the chain of exon and $T$ is maximum among all chains of blocks from $B$. 
Spliced alignment can be solved by DP algorithm.
GENSCAN

- States -- functional units
- “Phase” three frames.
- Two symmetric sub modules for forward and backward strands

Performance: 80% exon detecting (but if a gene has more than one exon probability of detection decrease rapidly).
Aligns two sequences and marks each base as gap ( - ), mismatch (:), match ( | ), resulting in a new alphabet of 12 letters: \( \Sigma \{ A-, A::, A \mid, C-, C::, C \mid, G-, G::, G \mid, T-, T::, T\mid \} \).

Run Viterbi algorithm using emissions \( e_k(b) \) where \( b \in \{ A-, A::, A\mid, \ldots, T\mid \} \).

The emission probabilities are estimated from human/mouse gene pairs.

- Ex. \( e_I(x\mid) < e_E(x\mid) \) since matches are favored in exons, and \( e_I(x-) > e_E(x-) \) since gaps (as well as mismatches) are favored in introns.
- Compensates for dominant occurrence of poly-A region in introns.
What’s new

- What is a gene?
- Prediction of gene fragments in short reads (e.g., metagenomic sequences)
- Gene prediction (refinement) integrating other evidences (e.g., RNA-seq)
  - Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm (NAR, 2014)
- Go beyond protein coding genes
  - ncRNA gene prediction