Ø

Gene prediction
• Simple method (base composition, codon usage)
• Similarity method
• Ab initio methods

Today's topics

Universal Codon Table

<table>
<thead>
<tr>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Met</td>
<td>Val</td>
<td>Leu</td>
<td>Ile</td>
</tr>
<tr>
<td>Val</td>
<td>Leu</td>
<td>Ile</td>
<td>Met</td>
</tr>
<tr>
<td>Leu</td>
<td>Met</td>
<td>Ile</td>
<td>Val</td>
</tr>
<tr>
<td>Ile</td>
<td>Val</td>
<td>Meta</td>
<td>Leu</td>
</tr>
</tbody>
</table>

* Synonymous (silent) substitutions DOES NOT change amino acids
* Nonsynonymous (replacement) substitutions DOES change amino acids

Synonymous Codon Usage Bias

<table>
<thead>
<tr>
<th>Codon (amino acid)</th>
<th>Eschholzia coli</th>
<th>Saccharomyces cerevisiae</th>
<th>Drosophila melanogaster</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Low)²</td>
<td>(High)²</td>
<td>(Low)²</td>
<td>(High)²</td>
</tr>
<tr>
<td>CTT (proline)</td>
<td>1126</td>
<td>619</td>
<td>17.38</td>
</tr>
<tr>
<td>CCT (proline)</td>
<td>1126</td>
<td>619</td>
<td>17.38</td>
</tr>
<tr>
<td>CGT (proline)</td>
<td>1126</td>
<td>619</td>
<td>17.38</td>
</tr>
<tr>
<td>CGT (proline)</td>
<td>1126</td>
<td>619</td>
<td>17.38</td>
</tr>
<tr>
<td>ACT (threonine)</td>
<td>1125</td>
<td>20.59</td>
<td>1584</td>
</tr>
<tr>
<td>ACC (threonine)</td>
<td>1125</td>
<td>20.59</td>
<td>1584</td>
</tr>
<tr>
<td>AGG (threonine)</td>
<td>1125</td>
<td>20.59</td>
<td>1584</td>
</tr>
<tr>
<td>AGA (threonine)</td>
<td>1125</td>
<td>20.59</td>
<td>1584</td>
</tr>
</tbody>
</table>

What causes base/codon usage bias?

> Nonsynonymous sites (1st/2nd codon positions)
  • Amino acid composition
    ➔ Relatively small variation in base composition

> Synonymous sites (3rd codon positions)
  • Mutation bias
    ➔ Base composition consistent with that of introns
  • Translational selection for synonymous codons
    ➔ Even though synonymous substitutions do not change amino acids!
  • Base composition not consistent with that of introns

A recent review: Hanson and Coller (2017)
Highly expressed genes preferentially use synonymous codons that are recognized by the most abundant tRNA species. Natural selection favors using such codons since it will increase translational efficiency and accuracy.

Translational selection affects synonymous codon usage bias. Observed in single cellular organisms and in Drosophila. Such selection is not very strong and may be overcome by strong mutation bias in some organisms (human, some bacteria, etc.)

From Graur and Li (2000) Fundamentals of Molecular Evolution

**Fickett TESTCODE statistic**


The statistic > 0.95 for coding, < 0.74 for noncoding

\[
P_n = \frac{\text{Maximum}[n(1), n(2), n(3)]}{\text{Minimum}[n(1), n(2), n(3)]}
\]

where \( n(1), n(2), \text{ and } n(3) \) are the composition of each nucleotide at codon positions 1, 2, and 3 within the window (default = 200 bp)

\( n = A, T, G, \text{ or } C \)

\( C_n = \text{Base composition of nucleotide } n \text{ in the window} \)

Based on the pre-calculated tables given by Fickett (1982), these 8 values are converted into probabilities of being found in a codon region

The probabilities are multiplied by weighting factor

**Fickett TESTCODE statistic**

D. melanogaster alcohol dehydrogenase (Adh) gene

AF175220: CDS join(44..142,210..453,855..951)
Gribskov's codon preference plot

Based on the frequency of each synonymous codon

- $f_{abc}$: the frequency of codon $abc$ (e.g., $f_{CTG}$)
- $F_{abc}$: the frequency of the synonymous codon family that includes codon $abc$ (e.g., $F_{CTG}$ = $F_{LEU}$)
- $r_{abc}$: the frequency of codon $abc$ in a random sequence
- $R_{abc}$: the frequency of the synonymous codon family that includes codon $abc$ in a random sequence

$$p = f_{abc}/F_{abc}$$
$$P = \left(\frac{1}{w}\right)^{w}$$

Table 1: Gene finding strategies

<table>
<thead>
<tr>
<th>Method</th>
<th>Gene Structure Deduction</th>
<th>Requires Similarity</th>
<th>Requires Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity-based</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Ab initio methods</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Evidenced-based</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Combination or hybrid methods</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Review:
- Goodwin et al. (2012) PLoS ONE 7:e50609

Ab initio and evidence-driven methods

**Choosing and combining**

- $f_{abc}$: the frequency of codon $abc$
- $F_{abc}$: the frequency of the synonymous codon family that includes codon $abc$
- $r_{abc}$: the frequency of codon $abc$ in a random sequence
- $R_{abc}$: the frequency of the synonymous codon family that includes codon $abc$ in a random sequence

$$p = f_{abc}/F_{abc}$$
$$P = \left(\frac{1}{w}\right)^{w}$$

Gene finding strategies

- **Similarity based methods**
  - Gene structure can be deduced based on similarity
  - Requires similar (not too distant) sequences
- **Ab initio methods**
  - Two types of information
    - Compositional information
    - Signal information
- **Evidence-based methods**
  - Use external evidence (e.g., ESTs, RNA-seq) to improve the prediction accuracy
Gene finding: Similarity methods

- **Similarity methods: principles**
  - Coding regions evolve slower than non-coding regions (higher functional constraints in coding regions)
  - Highly similar sequences could have been derived from the common evolutionary origin
    - Gene structure could be conserved
  - Standard pairwise alignment based method can be used (e.g., BLAST, Smith-Waterman)
    - Against genes with known gene structures
    - Genomic vs. cDNA sequences
  - Can be used to confirm prediction inferred by other methods

GeneWise/Wise2
- http://www.ebi.ac.uk/Tools/psa/genewise
  - Compares a genomic DNA sequence to protein sequences or profile HMMs
  - Uses dynamic programming algorithm
  - Sequence errors (e.g., frame shifts) and introns can be considered
  - Searches against the Pfam profile HMM database
  - Stand-alone version: http://www.ebi.ac.uk/~birney/wise2/

Gene finding: Ab initio methods

- **Ab initio methods: principles**
  - Integrate "signal detection" and "coding statistics"
    - Signal detection and coding statistics are obtained from a training data set (known samples)
  - Probabilistic frameworks are used to infer gene structures
  - A scoring system can be used to evaluate gene prediction
  - Various machine learning algorithms are used (e.g., neural networks, HMMs, discriminant analysis)

Coding statistics
- Unequal usage of codons (codon usage bias) in the coding regions → Species specific
- Can be used to differentiate between coding and non-coding regions
- Coding statistics - e.g., a likelihood that the sequence is protein coding
  - Codon usage, hexamer usage, GC content, nucleotide periodicity, etc.

Coding statistics: example
- \( F(c) \): the frequency of codon \( c \) in protein coding genes
- \( S \): a sequence of codons, \( c_1c_2...c_m \)
  - assuming independence between adjacent codons
- \( P(S) \): the probability of finding protein coding sequence \( S \)
  - \( P(S) = F(c_1)F(c_2)...F(c_m) \)
- e.g., \( S = GGC\ ACC \)
  - \( c_1 = GGC \), \( c_2 = ACC \)
  - if \( F(c_1) = 0.027 \) and \( F(c_2) = 0.021 \),
  - \( P(S) = F(c_1)F(c_2) = 0.027 \times 0.021 = 0.000567 \)
Gene finding: Ab initio methods

Coding statistics: example (continued)

- \( F_0(c) \): the frequency of codon \( c \) in a non-coding sequence
- \( P_0(S) \): the probability of finding a non-coding sequence \( S \)
- Assuming the random model of non-coding DNA,
  \[ \log \left( \frac{P_0(S)}{F_0(c_1)F_0(c_2)\ldots F_0(c_m)} \right) = \log \left( \frac{0.000567}{0.0156^2} \right) = \log(2.33) = 0.85 > 0 \Rightarrow S \text{ is coding} \]

Signal Sensors

- Various pattern recognition methods are used for identifying these signals
  - consensus sequences
  - position specific weight matrices (PSSMs)
  - decision trees
  - hidden Markov models (HMMs)
  - neural networks (NN)

Signal detection: example

- Consensus method
  - Obtained by choosing the most frequent base at each position of the multiple alignment
  - consensus sequence: TATAAT
  - consensus (IUPAC): TAT\(^{R}A\)T (R: G/A, N: T/C/A/G)
  - leads to loss of information
  - produces many false positives or false negatives

Signal detection: Consensus method

- Coding statistics can be combined

Signal detection: PSSM (WMM: weight matrix model)

- Assumes independence between adjacent bases
- Sum of the matrix values (or probabilities) for the target sequence (or log-likelihood) is used as its score
- Decision is done based on a cut-off value
Gene finding: Ab initio methods

- Signal detection: Maximal dependence decomposition (MDD; a decision tree) ➜ used in GENSCAN

- Various pattern recognition methods are used
  - Decision trees [e.g., MORGAN, GlimmerM]
  - Discriminant function analysis (LDA, QDA) [e.g., MIZEF, FGENES, HEXON]
  - Hidden Markov Models (HMMs) [e.g., VEIL, GENSCAN, HMMGene, GeneMark.hmm, Glimmer]
  - Neural networks (NN) [e.g., GRAIL II]
  - Combinations with similarity or evidence-based methods [e.g., TwinScan/NSCAN, GenomeScan, CONTRAST, Augustus]
  - Combiner methods [e.g., JIGSAW, GLEAN, GeneComber]

Gene prediction methods

- Combiner methods
  - [e.g., JIGSAW, GLEAN, GeneComber]

Gene finding: GENSCAN

http://genes.mit.edu/GENSCAN.html
Gene prediction by Decision Trees

- **MORGAN**: A decision tree system for finding genes in vertebrate DNA (available but obsolete)
  
  http://www.cbcb.umd.edu/~salzberg/morgan

Gene prediction by Neural Net

- **Grail, Grailexp**: (obsolete; updated to PRODIGAL)

  ![Neural Net Diagram](https://example.com/neural_net_diagram.png)