Structural Annotation
Overview

Sucheta Tripathy
Virginia Bioinformatics Institute,
Blacksburg
Gene Finding -- outline

- A quick review
- Prediction of coding regions in eukaryotic and prokaryotic genomes
- Prediction of translation starts of genes
- Prediction of splice junctions in eukaryotic genomes
  - donor site prediction
  - acceptor site prediction
- Information fusion – combining multiple pieces of information for gene prediction
- How to predict genes in a newly sequenced genome
- Popular gene prediction programs
The basic idea of pattern recognition

- How do kids learn to distinguish “dogs” from “cats”?
  - were “trained” by being told “A is a dog”, “B is a cat”, “C is another dog”, ….
  - they learn to “extract” common features (patterns) among animals they were told to be “dogs” and “cats”
  - then apply these extracted features to identify new dogs and cats

- Pattern recognition is generally done by
  - providing “training sets” which are individually labeled “positives” versus “negatives”, or “good” versus “bad”, etc.
  - learning the general rules that separate the “positives” from “negatives” or “good” from “bad”, ….
  - applying the learned rules to new situations
Gene finding through learning

- Is gcgatgcgcgtgcgtgagacgtaggcccgaagagagagatgtagga
ggaaggtttaggtgttagatgattgttagtttgtagtcatgtagtgatgatcgtag a gene?

- Remember “dogs”, “cats” ....

- but the “patterns” here are much more hidden and more complex than the distinguishing features between “dogs” and “cats”

We need to study the basic structures of genes first ....!
Basic Gene Structures

[Diagram showing the structure of a gene, including promoter, enhancer, exons, introns, and mRNA processing steps such as transcription, capping, splicing, and translation.]
Gene Structure – open reading frame (ORF)

- Open reading frame (ORF): a segment of DNA with one in-frame start codon and one in-frame stop codon at the two ends and no in-frame stop codon in the middle

How many genes can an ORF have inside it?

Answer: one because an ORF has only one stop
Gene Structure -- open reading frame (ORF)

- Generally true: all long (> 300 bp) orfs in prokaryotic genomes encode genes

But this may not necessarily be true for eukaryotic genomes

- Coding region –
  - gene in prokaryotic genomes
  - exon in eukaryotic genomes
Gene Structure

- Each coding region (exon or whole gene) has a fixed translation frame.
- A coding region always sits inside an ORF of same reading frame.
- All exons of a gene are on the same strand.
- Neighboring exons of a gene could have different reading frames.
Gene Structure – reading frame consistency

- Now … we are talking about a little more “complex” features

- Neighboring exons of a gene should be frame-consistent

\[
\text{b} = (\text{m} - \text{j} - 1 + \text{a}) \mod 3
\]

```
inexon1 [i, j] in frame a and exon2 [m, n] in frame b are consistent if

1 \mod 3 = 1
2 \mod 3 = 2
3 \mod 3 = 0
4 \mod 3 = 1
5 \mod 3 = 2
```

6/5/2009
Codon Frequencies

- Coding sequences are translated into protein sequences
- We found the following – the dimer frequency in protein sequences is NOT evenly distributed

The average frequency is ¼%

Some amino acids prefer to be next to each other

Some other amino acids prefer to be not next to each other
<table>
<thead>
<tr>
<th></th>
<th>ALA</th>
<th>ARG</th>
<th>ASN</th>
<th>ASP</th>
<th>CYS</th>
<th>GLU</th>
<th>GLN</th>
<th>GLY</th>
<th>HIS</th>
<th>ILE</th>
<th>LEU</th>
<th>LYS</th>
<th>MET</th>
<th>PHE</th>
<th>PRO</th>
<th>SER</th>
<th>THR</th>
<th>TRP</th>
<th>TYR</th>
<th>VAL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td>.99</td>
<td>.5</td>
<td>.27</td>
<td>.45</td>
<td>.13</td>
<td>.52</td>
<td>.34</td>
<td>.51</td>
<td>.19</td>
<td>.38</td>
<td>.83</td>
<td>.46</td>
<td>.2</td>
<td>.31</td>
<td>.37</td>
<td>.73</td>
<td>.56</td>
<td>.09</td>
<td>.18</td>
<td>.65</td>
</tr>
<tr>
<td><strong>R</strong></td>
<td>.5</td>
<td>.5</td>
<td>.2</td>
<td>.3</td>
<td>.1</td>
<td>.4</td>
<td>.3</td>
<td>.4</td>
<td>.18</td>
<td>.25</td>
<td>.63</td>
<td>.37</td>
<td>.14</td>
<td>.22</td>
<td>.26</td>
<td>.54</td>
<td>.34</td>
<td>.08</td>
<td>.17</td>
<td>.46</td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>.31</td>
<td>.19</td>
<td>.11</td>
<td>.2</td>
<td>.05</td>
<td>.23</td>
<td>.23</td>
<td>.26</td>
<td>.07</td>
<td>.13</td>
<td>.27</td>
<td>.16</td>
<td>.07</td>
<td>.11</td>
<td>.15</td>
<td>.24</td>
<td>.16</td>
<td>.04</td>
<td>.08</td>
<td>.27</td>
</tr>
<tr>
<td><strong>D</strong></td>
<td>.54</td>
<td>.32</td>
<td>.17</td>
<td>.47</td>
<td>.08</td>
<td>.51</td>
<td>.19</td>
<td>.42</td>
<td>.13</td>
<td>.25</td>
<td>.48</td>
<td>.26</td>
<td>.12</td>
<td>.20</td>
<td>.24</td>
<td>.40</td>
<td>.29</td>
<td>.06</td>
<td>.14</td>
<td>.5</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td>.14</td>
<td>.11</td>
<td>.05</td>
<td>.09</td>
<td>.04</td>
<td>.09</td>
<td>.06</td>
<td>.13</td>
<td>.04</td>
<td>.08</td>
<td>.14</td>
<td>.08</td>
<td>.03</td>
<td>.06</td>
<td>.07</td>
<td>.14</td>
<td>.10</td>
<td>.02</td>
<td>.04</td>
<td>.13</td>
</tr>
<tr>
<td><strong>E</strong></td>
<td>.57</td>
<td>.43</td>
<td>.22</td>
<td>.42</td>
<td>.09</td>
<td>.59</td>
<td>.30</td>
<td>.33</td>
<td>.16</td>
<td>.28</td>
<td>.64</td>
<td>.40</td>
<td>.17</td>
<td>.20</td>
<td>.21</td>
<td>.44</td>
<td>.36</td>
<td>.06</td>
<td>.16</td>
<td>.44</td>
</tr>
<tr>
<td><strong>Q</strong></td>
<td>.34</td>
<td>.31</td>
<td>.11</td>
<td>.20</td>
<td>.06</td>
<td>.27</td>
<td>.29</td>
<td>.20</td>
<td>.12</td>
<td>.15</td>
<td>.45</td>
<td>.21</td>
<td>.10</td>
<td>.13</td>
<td>.17</td>
<td>.29</td>
<td>.22</td>
<td>.05</td>
<td>.10</td>
<td>.29</td>
</tr>
<tr>
<td><strong>G</strong></td>
<td>.50</td>
<td>.39</td>
<td>.22</td>
<td>.37</td>
<td>.11</td>
<td>.37</td>
<td>.21</td>
<td>.50</td>
<td>.16</td>
<td>.28</td>
<td>.50</td>
<td>.33</td>
<td>.14</td>
<td>.23</td>
<td>.21</td>
<td>.54</td>
<td>.35</td>
<td>.07</td>
<td>.17</td>
<td>.46</td>
</tr>
<tr>
<td><strong>H</strong></td>
<td>.21</td>
<td>.17</td>
<td>.07</td>
<td>.14</td>
<td>.04</td>
<td>.16</td>
<td>.10</td>
<td>.17</td>
<td>.08</td>
<td>.09</td>
<td>.22</td>
<td>.10</td>
<td>.05</td>
<td>.09</td>
<td>.12</td>
<td>.17</td>
<td>.11</td>
<td>.03</td>
<td>.06</td>
<td>.21</td>
</tr>
<tr>
<td><strong>I</strong></td>
<td>.37</td>
<td>.25</td>
<td>.13</td>
<td>.27</td>
<td>.08</td>
<td>.27</td>
<td>.15</td>
<td>.27</td>
<td>.09</td>
<td>.15</td>
<td>.34</td>
<td>.22</td>
<td>.08</td>
<td>.14</td>
<td>.18</td>
<td>.29</td>
<td>.21</td>
<td>.04</td>
<td>.11</td>
<td>.32</td>
</tr>
<tr>
<td><strong>L</strong></td>
<td>.79</td>
<td>.65</td>
<td>.30</td>
<td>.53</td>
<td>.16</td>
<td>.62</td>
<td>.45</td>
<td>.50</td>
<td>.25</td>
<td>.31</td>
<td>.97</td>
<td>.47</td>
<td>.19</td>
<td>.32</td>
<td>.44</td>
<td>.71</td>
<td>.49</td>
<td>.10</td>
<td>.22</td>
<td>.67</td>
</tr>
<tr>
<td><strong>K</strong></td>
<td>.43</td>
<td>.41</td>
<td>.19</td>
<td>.26</td>
<td>.08</td>
<td>.35</td>
<td>.24</td>
<td>.26</td>
<td>.14</td>
<td>.20</td>
<td>.49</td>
<td>.41</td>
<td>.13</td>
<td>.17</td>
<td>.20</td>
<td>.37</td>
<td>.32</td>
<td>.07</td>
<td>.15</td>
<td>.33</td>
</tr>
<tr>
<td><strong>M</strong></td>
<td>.23</td>
<td>.17</td>
<td>.09</td>
<td>.17</td>
<td>.04</td>
<td>.19</td>
<td>.12</td>
<td>.14</td>
<td>.06</td>
<td>.10</td>
<td>.25</td>
<td>.15</td>
<td>.07</td>
<td>.08</td>
<td>.11</td>
<td>.20</td>
<td>.17</td>
<td>.03</td>
<td>.06</td>
<td>.17</td>
</tr>
<tr>
<td><strong>F</strong></td>
<td>.30</td>
<td>.20</td>
<td>.11</td>
<td>.22</td>
<td>.07</td>
<td>.22</td>
<td>.13</td>
<td>.26</td>
<td>.08</td>
<td>.14</td>
<td>.33</td>
<td>.15</td>
<td>.08</td>
<td>.14</td>
<td>.14</td>
<td>.27</td>
<td>.18</td>
<td>.04</td>
<td>.10</td>
<td>.28</td>
</tr>
<tr>
<td><strong>P</strong></td>
<td>.35</td>
<td>.28</td>
<td>.13</td>
<td>.23</td>
<td>.05</td>
<td>.27</td>
<td>.16</td>
<td>.24</td>
<td>.10</td>
<td>.17</td>
<td>.39</td>
<td>.19</td>
<td>.10</td>
<td>.15</td>
<td>.26</td>
<td>.42</td>
<td>.29</td>
<td>.04</td>
<td>.09</td>
<td>.33</td>
</tr>
<tr>
<td><strong>S</strong></td>
<td>.68</td>
<td>.51</td>
<td>.26</td>
<td>.46</td>
<td>.14</td>
<td>.43</td>
<td>.26</td>
<td>.55</td>
<td>.17</td>
<td>.33</td>
<td>.68</td>
<td>.38</td>
<td>.18</td>
<td>.28</td>
<td>.36</td>
<td>.93</td>
<td>.55</td>
<td>.09</td>
<td>.17</td>
<td>.56</td>
</tr>
<tr>
<td><strong>T</strong></td>
<td>.51</td>
<td>.36</td>
<td>.19</td>
<td>.29</td>
<td>.10</td>
<td>.31</td>
<td>.21</td>
<td>.37</td>
<td>.12</td>
<td>.25</td>
<td>.52</td>
<td>.32</td>
<td>.13</td>
<td>.21</td>
<td>.31</td>
<td>.54</td>
<td>.42</td>
<td>.07</td>
<td>.13</td>
<td>.39</td>
</tr>
<tr>
<td><strong>W</strong></td>
<td>.07</td>
<td>.08</td>
<td>.05</td>
<td>.06</td>
<td>.02</td>
<td>.06</td>
<td>.05</td>
<td>.06</td>
<td>.03</td>
<td>.06</td>
<td>.12</td>
<td>.07</td>
<td>.03</td>
<td>.04</td>
<td>.04</td>
<td>.09</td>
<td>.07</td>
<td>.02</td>
<td>.03</td>
<td>.07</td>
</tr>
<tr>
<td><strong>Y</strong></td>
<td>.21</td>
<td>.16</td>
<td>.09</td>
<td>.15</td>
<td>.05</td>
<td>.16</td>
<td>.09</td>
<td>.17</td>
<td>.06</td>
<td>.10</td>
<td>.23</td>
<td>.11</td>
<td>.06</td>
<td>.10</td>
<td>.1</td>
<td>.17</td>
<td>.13</td>
<td>.03</td>
<td>.07</td>
<td>.2</td>
</tr>
<tr>
<td><strong>V</strong></td>
<td>.69</td>
<td>.42</td>
<td>.24</td>
<td>.46</td>
<td>.13</td>
<td>.48</td>
<td>.31</td>
<td>.42</td>
<td>.18</td>
<td>.29</td>
<td>.72</td>
<td>.37</td>
<td>.16</td>
<td>.27</td>
<td>.33</td>
<td>.55</td>
<td>.42</td>
<td>.07</td>
<td>.18</td>
<td>.61</td>
</tr>
</tbody>
</table>

6/5/2009
<table>
<thead>
<tr>
<th>Name</th>
<th>ala</th>
<th>arg</th>
<th>asn</th>
<th>asp</th>
<th>cys</th>
<th>glu</th>
<th>gln</th>
<th>gly</th>
<th>his</th>
<th>ile</th>
<th>leu</th>
<th>lys</th>
<th>met</th>
<th>phe</th>
<th>pro</th>
<th>ser</th>
<th>thr</th>
<th>trp</th>
<th>tyr</th>
<th>val</th>
</tr>
</thead>
<tbody>
<tr>
<td>ala</td>
<td>9.5</td>
<td>4.1</td>
<td>4.3</td>
<td>5.3</td>
<td>1.2</td>
<td>6</td>
<td>4.8</td>
<td>6.5</td>
<td>2</td>
<td>6.5</td>
<td>11.5</td>
<td>6</td>
<td>2.6</td>
<td>3.7</td>
<td>3.5</td>
<td>6.2</td>
<td>5</td>
<td>1.1</td>
<td>2.7</td>
<td>6.5</td>
</tr>
<tr>
<td>arg</td>
<td>7.9</td>
<td>5.5</td>
<td>3.9</td>
<td>5.3</td>
<td>1.1</td>
<td>6</td>
<td>5.5</td>
<td>5.9</td>
<td>2.6</td>
<td>6.5</td>
<td>11.4</td>
<td>5</td>
<td>2.2</td>
<td>4.7</td>
<td>3.6</td>
<td>5.5</td>
<td>4.4</td>
<td>1.4</td>
<td>4</td>
<td>6.6</td>
</tr>
<tr>
<td>asn</td>
<td>9.6</td>
<td>4.9</td>
<td>4.2</td>
<td>4.9</td>
<td>1</td>
<td>5.3</td>
<td>5.6</td>
<td>7.4</td>
<td>2.3</td>
<td>6</td>
<td>10</td>
<td>4.9</td>
<td>2</td>
<td>3.5</td>
<td>5.1</td>
<td>6.1</td>
<td>5.5</td>
<td>1.5</td>
<td>3.1</td>
<td>6.1</td>
</tr>
<tr>
<td>asp</td>
<td>9.3</td>
<td>4</td>
<td>4.7</td>
<td>5.1</td>
<td>1</td>
<td>6.7</td>
<td>2.9</td>
<td>7</td>
<td>1.8</td>
<td>7.1</td>
<td>9.6</td>
<td>6.3</td>
<td>2.3</td>
<td>4.3</td>
<td>3.9</td>
<td>5.9</td>
<td>5.1</td>
<td>1.6</td>
<td>3.6</td>
<td>6.6</td>
</tr>
<tr>
<td>cys</td>
<td>8.4</td>
<td>4.8</td>
<td>3.3</td>
<td>5.4</td>
<td>1.7</td>
<td>5.6</td>
<td>5.2</td>
<td>8.1</td>
<td>4.3</td>
<td>5.4</td>
<td>10.2</td>
<td>3.8</td>
<td>1.8</td>
<td>4.1</td>
<td>4.5</td>
<td>6.3</td>
<td>4.3</td>
<td>1.6</td>
<td>3.4</td>
<td>6.8</td>
</tr>
<tr>
<td>glu</td>
<td>9.4</td>
<td>5.8</td>
<td>3.6</td>
<td>4.5</td>
<td>0.8</td>
<td>4.9</td>
<td>7</td>
<td>5.8</td>
<td>2.6</td>
<td>5.9</td>
<td>12.7</td>
<td>5</td>
<td>2.4</td>
<td>4</td>
<td>3.5</td>
<td>5.4</td>
<td>5</td>
<td>1.1</td>
<td>2.8</td>
<td>6.8</td>
</tr>
<tr>
<td>gln</td>
<td>10.3</td>
<td>4.9</td>
<td>3</td>
<td>4.4</td>
<td>0.9</td>
<td>4.5</td>
<td>6.8</td>
<td>7</td>
<td>2.7</td>
<td>5.5</td>
<td>12.8</td>
<td>4.1</td>
<td>2</td>
<td>3.9</td>
<td>3.8</td>
<td>5.8</td>
<td>5.3</td>
<td>1.4</td>
<td>3</td>
<td>6.9</td>
</tr>
<tr>
<td>gly</td>
<td>8.1</td>
<td>4.8</td>
<td>3.9</td>
<td>5.1</td>
<td>1.2</td>
<td>6</td>
<td>4.6</td>
<td>6.4</td>
<td>2.4</td>
<td>6.8</td>
<td>10.5</td>
<td>5.8</td>
<td>2.7</td>
<td>4.8</td>
<td>2.4</td>
<td>5.8</td>
<td>5.1</td>
<td>1.4</td>
<td>3.7</td>
<td>7.5</td>
</tr>
<tr>
<td>his</td>
<td>7.3</td>
<td>4.7</td>
<td>4</td>
<td>4.8</td>
<td>1.5</td>
<td>4.9</td>
<td>5.6</td>
<td>6.9</td>
<td>3</td>
<td>6.2</td>
<td>10.8</td>
<td>4.8</td>
<td>1.6</td>
<td>5</td>
<td>5.2</td>
<td>6.8</td>
<td>4.9</td>
<td>1.7</td>
<td>4.2</td>
<td>5.1</td>
</tr>
<tr>
<td>ile</td>
<td>11</td>
<td>4.7</td>
<td>4.9</td>
<td>6.5</td>
<td>1.1</td>
<td>6.9</td>
<td>3.6</td>
<td>7.2</td>
<td>2.1</td>
<td>5.3</td>
<td>8.6</td>
<td>5.3</td>
<td>1.8</td>
<td>4.2</td>
<td>7</td>
<td>5.6</td>
<td>0.9</td>
<td>2.9</td>
<td>6.1</td>
<td></td>
</tr>
<tr>
<td>leu</td>
<td>10.4</td>
<td>4.2</td>
<td>4.3</td>
<td>5.2</td>
<td>1.1</td>
<td>5.2</td>
<td>3.7</td>
<td>6.8</td>
<td>2</td>
<td>5.6</td>
<td>10.6</td>
<td>5.3</td>
<td>2.3</td>
<td>3.8</td>
<td>4.5</td>
<td>7.4</td>
<td>6.2</td>
<td>1</td>
<td>2.6</td>
<td>6.6</td>
</tr>
<tr>
<td>lys</td>
<td>10.6</td>
<td>5.2</td>
<td>3.8</td>
<td>5.2</td>
<td>0.5</td>
<td>5.3</td>
<td>5.9</td>
<td>6.6</td>
<td>2.6</td>
<td>5.2</td>
<td>11.3</td>
<td>4.7</td>
<td>1.9</td>
<td>2.8</td>
<td>4.6</td>
<td>6</td>
<td>5.5</td>
<td>1.2</td>
<td>2.6</td>
<td>7.6</td>
</tr>
<tr>
<td>met</td>
<td>10.8</td>
<td>4.8</td>
<td>3.8</td>
<td>4.6</td>
<td>0.7</td>
<td>4.6</td>
<td>4.9</td>
<td>7</td>
<td>1.7</td>
<td>4.7</td>
<td>11.4</td>
<td>5.2</td>
<td>2.8</td>
<td>3.3</td>
<td>5.1</td>
<td>7.4</td>
<td>6.3</td>
<td>0.9</td>
<td>2</td>
<td>6.8</td>
</tr>
<tr>
<td>phe</td>
<td>9.6</td>
<td>3.7</td>
<td>5.2</td>
<td>6.5</td>
<td>1.2</td>
<td>6.4</td>
<td>2.7</td>
<td>7.9</td>
<td>1.9</td>
<td>6.7</td>
<td>7.4</td>
<td>5</td>
<td>2.5</td>
<td>3.9</td>
<td>3.6</td>
<td>8</td>
<td>5.8</td>
<td>1.3</td>
<td>3.3</td>
<td>6.3</td>
</tr>
<tr>
<td>pro</td>
<td>8.4</td>
<td>3.6</td>
<td>4.6</td>
<td>5.4</td>
<td>0.7</td>
<td>7.6</td>
<td>5.2</td>
<td>5.4</td>
<td>2.3</td>
<td>6.1</td>
<td>11.2</td>
<td>5.5</td>
<td>2.4</td>
<td>4.2</td>
<td>2.8</td>
<td>6.5</td>
<td>5.4</td>
<td>1.4</td>
<td>2.9</td>
<td>7.5</td>
</tr>
<tr>
<td>ser</td>
<td>9.1</td>
<td>4.6</td>
<td>3.7</td>
<td>5</td>
<td>1</td>
<td>5.4</td>
<td>5.2</td>
<td>7.2</td>
<td>2.6</td>
<td>6</td>
<td>11.6</td>
<td>4.5</td>
<td>2.2</td>
<td>4.1</td>
<td>4.1</td>
<td>6.5</td>
<td>5</td>
<td>1.2</td>
<td>3.2</td>
<td>6.8</td>
</tr>
<tr>
<td>thr</td>
<td>9.1</td>
<td>4.2</td>
<td>3.7</td>
<td>5.6</td>
<td>0.9</td>
<td>5.7</td>
<td>5.7</td>
<td>7.5</td>
<td>2.2</td>
<td>5.5</td>
<td>12</td>
<td>4.2</td>
<td>2</td>
<td>3.5</td>
<td>5.5</td>
<td>6.2</td>
<td>5.3</td>
<td>1.1</td>
<td>2.6</td>
<td>6.7</td>
</tr>
<tr>
<td>trp</td>
<td>7.1</td>
<td>6.3</td>
<td>3.2</td>
<td>4.8</td>
<td>1.3</td>
<td>3.9</td>
<td>8.5</td>
<td>6.6</td>
<td>3.6</td>
<td>5</td>
<td>14.2</td>
<td>3.2</td>
<td>2.4</td>
<td>4.6</td>
<td>3.9</td>
<td>5.8</td>
<td>4.3</td>
<td>1.3</td>
<td>3</td>
<td>6.1</td>
</tr>
<tr>
<td>tyr</td>
<td>7.9</td>
<td>6.5</td>
<td>3.6</td>
<td>4.9</td>
<td>1.2</td>
<td>4.5</td>
<td>7</td>
<td>7.1</td>
<td>2.6</td>
<td>5</td>
<td>11.7</td>
<td>4</td>
<td>1.6</td>
<td>4.7</td>
<td>4.9</td>
<td>6.4</td>
<td>4.6</td>
<td>1.5</td>
<td>3.4</td>
<td>5.7</td>
</tr>
<tr>
<td>val</td>
<td>9.6</td>
<td>4.1</td>
<td>4.4</td>
<td>5.9</td>
<td>1</td>
<td>6.2</td>
<td>3.4</td>
<td>6.4</td>
<td>1.8</td>
<td>6.5</td>
<td>10.2</td>
<td>5.2</td>
<td>2.5</td>
<td>3.7</td>
<td>3.8</td>
<td>7.2</td>
<td>6.1</td>
<td>1.1</td>
<td>2.7</td>
<td>7.1</td>
</tr>
</tbody>
</table>
Dicodon Frequencies

- Believe it or not – the biased (uneven) dimer frequencies are the foundation of many gene finding programs!

- Basic idea – if a dimer has lower than average dimer frequency; this means that proteins prefer not to have such dimers in its sequence;

Hence if we see a dicodon encoding this dimer, we may want to bet against this dicodon being in a coding region!
Dicodon Frequencies

- Relative frequencies of a di-codon in coding versus non-coding
  - frequency of dicodon X (e.g., AAAAAA) in coding region, total number of occurrences of X divided by total number of di-codon occurrences
  - frequency of dicodon X (e.g., AAAAAA) in non-coding region, total number of occurrences of X divided by total number of di-codon occurrences

In human genome, frequency of dicodon “AAA AAA” is ~1% in coding region versus ~5% in non-coding region

Question: if you see a region with many “AAA AAA”, would you guess it is a coding or non-coding region?
Basic idea of gene finding

- Most dicodons show bias towards either coding or non-coding regions; only fraction of dicodons is neutral
- Foundation for coding region identification

Regions consisting of dicodons that mostly tend to be in coding regions are probably coding regions; otherwise non-coding regions

- Dicodon frequencies are key signal used for coding region detection; all gene finding programs use this information
Prediction of Translation Starts

- Certain nucleotides prefer to be in certain position around start “ATG” and other nucleotides prefer not to be there.

- The “biased” nucleotide distribution is information! It is a basis for translation start prediction.

- Question: which one is more probable to be a translation start?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4</td>
<td>17.36</td>
<td>16.53</td>
<td>46.28</td>
<td>19.83</td>
</tr>
<tr>
<td>-3</td>
<td>19.01</td>
<td>28.93</td>
<td>29.75</td>
<td>22.31</td>
</tr>
<tr>
<td>-2</td>
<td>17.36</td>
<td>57.85</td>
<td>19.01</td>
<td>5.79</td>
</tr>
<tr>
<td>-1</td>
<td>48.76</td>
<td>5.79</td>
<td>42.98</td>
<td>2.48</td>
</tr>
<tr>
<td>+1</td>
<td>28.93</td>
<td>39.67</td>
<td>14.88</td>
<td>16.53</td>
</tr>
<tr>
<td>+2</td>
<td>15.70</td>
<td>50.41</td>
<td>18.98</td>
<td>7.44</td>
</tr>
<tr>
<td>+3</td>
<td>21.49</td>
<td>22.31</td>
<td>32.23</td>
<td>13.22</td>
</tr>
<tr>
<td>+4</td>
<td>23.14</td>
<td>38.84</td>
<td>35.88</td>
<td>14.05</td>
</tr>
<tr>
<td>+5</td>
<td>19.83</td>
<td>23.97</td>
<td>32.23</td>
<td>22.31</td>
</tr>
<tr>
<td>+6</td>
<td>21.49</td>
<td>31.40</td>
<td>25.62</td>
<td>19.01</td>
</tr>
</tbody>
</table>

CACC ATG GC

TCGA ATG TT

6/5/2009
Prediction of Translation Starts

- Mathematical model: $F_i (X)$: frequency of $X$ (A, C, G, T) in position $i$
- Score a string by $\Sigma \log (F_i (X)/0.25)$

<table>
<thead>
<tr>
<th>CACC</th>
<th>ATG</th>
<th>GC</th>
<th>TCGA</th>
<th>ATG</th>
<th>TT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.37 + 0.29 + 0.20 + 0.30 + 0.24 + 0.29</td>
<td>-(0.62 + 0.62 + 0.22 + 0.55 + 0.28 + 0.25)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>= 1.69</td>
<td>= -2.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The model captures our intuition!
Prediction of Splice Junction Sites

- A start exon starts with a translation start and ends with a donor site.
- An internal exon starts with an acceptor site and ends with a donor site.
- An terminal exon starts with an acceptor site and ends with a stop codon.

Accurate prediction of exons/genes requires accurate prediction of splice junctions.
Prediction of Acceptor Sites

- Nucleotide distribution in the flanks of acceptors

\[
Y_{75} Y_{72} Y_{78} Y_{79} Y_{77} Y_{80} Y_{66} Y_{78} Y_{85} Y_{84} NC_{68} AG | G_{63}
\]

<table>
<thead>
<tr>
<th></th>
<th>-14</th>
<th>-13</th>
<th>-12</th>
<th>-11</th>
<th>-10</th>
<th>-9</th>
<th>-8</th>
<th>-7</th>
<th>-6</th>
<th>-5</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>11.1</td>
<td>12.7</td>
<td>3.2</td>
<td>4.8</td>
<td>12.7</td>
<td>8.7</td>
<td>16.7</td>
<td>16.7</td>
<td>12.7</td>
<td>9.5</td>
<td>26.2</td>
<td>6.3</td>
<td>100</td>
<td>0.0</td>
<td>21.4</td>
</tr>
<tr>
<td>C</td>
<td>36.5</td>
<td>30.9</td>
<td>19.1</td>
<td>23.0</td>
<td>34.9</td>
<td>39.7</td>
<td>34.9</td>
<td>40.5</td>
<td>40.5</td>
<td>36.5</td>
<td>33.3</td>
<td>68.2</td>
<td>0.0</td>
<td>0.0</td>
<td>7.9</td>
</tr>
<tr>
<td>G</td>
<td>9.5</td>
<td>10.3</td>
<td>15.1</td>
<td>12.7</td>
<td>8.7</td>
<td>9.5</td>
<td>16.7</td>
<td>4.8</td>
<td>2.4</td>
<td>6.3</td>
<td>13.9</td>
<td>0.0</td>
<td>0.0</td>
<td>100</td>
<td>62.7</td>
</tr>
<tr>
<td>U</td>
<td>38.9</td>
<td>41.3</td>
<td>58.7</td>
<td>55.6</td>
<td>42.1</td>
<td>40.5</td>
<td>30.9</td>
<td>37.3</td>
<td>44.4</td>
<td>47.6</td>
<td>27.0</td>
<td>25.4</td>
<td>0.0</td>
<td>0.0</td>
<td>7.9</td>
</tr>
</tbody>
</table>

Multiple positions have high “information content”

Information content: \( \sum F(X) \log \left( \frac{F(X)}{0.25} \right) \)

If every nucleotide has 0.25 frequency in a position, then the position’s information content is ZERO.

Use “information content as a criterion for determining the length of flanks”
Prediction of Acceptor Sites

- Mathematical model: $F_i(X)$: frequency of $X$ (A, C, G, T) in position $i$
- Score a segment as a candidate acceptor site by $\sum \log \left( \frac{F_i(X)}{0.25} \right)$
- For each candidate acceptor sequence, apply the model and get a score
- If the score if larger than zero, predict it is an “acceptor”; the higher score, the higher the probability the prediction is true
Prediction of Donor Sites

- Nucleotide distribution in the flanks of donors

<table>
<thead>
<tr>
<th></th>
<th>-3</th>
<th>-2</th>
<th>-1</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>34.0</td>
<td>60.4</td>
<td>9.2</td>
<td>0.0</td>
<td>0.0</td>
<td>52.6</td>
<td>71.3</td>
<td>7.1</td>
<td>16.0</td>
</tr>
<tr>
<td>C</td>
<td>36.3</td>
<td>12.9</td>
<td>3.3</td>
<td>0.0</td>
<td>0.0</td>
<td>2.8</td>
<td>7.6</td>
<td>5.5</td>
<td>16.5</td>
</tr>
<tr>
<td>G</td>
<td>18.3</td>
<td>12.5</td>
<td>80.3</td>
<td>100</td>
<td>0.0</td>
<td>41.9</td>
<td>11.8</td>
<td>81.4</td>
<td>20.9</td>
</tr>
<tr>
<td>U</td>
<td>11.4</td>
<td>14.2</td>
<td>7.3</td>
<td>0.0</td>
<td>100</td>
<td>2.5</td>
<td>9.3</td>
<td>5.9</td>
<td>46.2</td>
</tr>
</tbody>
</table>

- Mathematical model: $F_i(X)$: frequency of $X$ (A, C, G, T) in position $i$
- Score a segment as a possible donor site by $\Sigma \log (F_i(X)/0.25)$
Prediction of Donor Sites

- For each candidate donor sequence, apply the model and get a score
- If the score is larger than zero, predict it is a “donor”; the higher score, the higher the probability the prediction is true
Prediction of Donors/Acceptors

- Position specific weight matrix model

<table>
<thead>
<tr>
<th></th>
<th>-3</th>
<th>-2</th>
<th>-1</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>34.0</td>
<td>60.4</td>
<td>9.2</td>
<td>0.0</td>
<td>0.0</td>
<td>52.6</td>
<td>71.3</td>
<td>7.1</td>
<td>16.0</td>
</tr>
<tr>
<td>C</td>
<td>36.3</td>
<td>12.9</td>
<td>3.3</td>
<td>0.0</td>
<td>0.0</td>
<td>2.8</td>
<td>7.6</td>
<td>5.5</td>
<td>16.5</td>
</tr>
<tr>
<td>G</td>
<td>18.3</td>
<td>12.5</td>
<td>80.3</td>
<td>100</td>
<td>0.0</td>
<td>41.9</td>
<td>11.8</td>
<td>81.4</td>
<td>20.9</td>
</tr>
<tr>
<td>U</td>
<td>11.4</td>
<td>14.2</td>
<td>7.3</td>
<td>0.0</td>
<td>100</td>
<td>2.5</td>
<td>9.3</td>
<td>5.9</td>
<td>46.2</td>
</tr>
</tbody>
</table>

- Build a “position specific weight matrix model”
  - collect known {donor, acceptor} sequences and align them so that the GT or YAG are aligned
  - Calculate the percentage of each type of nucleotide at each position

- There are more sophisticated models for capturing higher order relationships between positions
Prediction of Exons

- For each orf, find all donor and acceptor candidates by finding GT and YAG motifs

- Score each donor and acceptor candidate using our position-specific weight matrix models

- Find all pairs of (acceptor, donor) above some thresholds

- Score the coding potential of the segment [donor, acceptor], using the hexmer model
Prediction of Exons

- For each segment [acceptor, donor], we get three scores (coding potential, donor score, acceptor score)

- Various possibilities
  - all three scores are high – probably true exon
  - all three scores are low – probably not a real exon
  - all in the middle -- ??
  - some scores are high and some are low -- ??

- What are the rules for exon prediction?
Prediction of Exons

- A “classifier” can be trained to separate exons from non-exons, based on the three scores
- Closer to reality – other factors could also help to distinguish exons from non-exons

- Exon length distribution
- Coding density is different in regions with different G+C contents

A practical gene finding software may use many features to distinguish exons from non-exons.
Prediction of Exons

Each box represents a predicted exon

A true exon typically has more than one predicted candidates, overlapping with each other
Gene Prediction in a New Genome

- We have assumed that some genes and non-genes are known before starting training a “gene finder” for that genome!

- What if we want to develop a gene finder for a new genome that has just been sequenced?

- Suggestions??
Gene Prediction in a New Genome

- One way is to identify a set of genes in the new genome through homology search against known genes in GenBank
  - BLAST, FASTA, Smith-Waterman

- So we can get some “coding” regions for training a gene finder

- How about non-coding regions?

  We can use the intergenic or inter-exon regions if we can identify any

- With these known genes and non-genes, we can train a gene finder just like before ….
Computational Gene Finding

- Making the call: coding or non-coding and where the boundaries are

- Need a training set with known coding and non-coding regions
  - select threshold(s) to include as many known coding regions as possible, and in the same time to exclude as many known non-coding regions as possible

If threshold = 0.2, we will include 90% of coding regions and also 10% of non-coding regions
If threshold = 0.4, we will include 70% of coding regions and also 6% of non-coding regions
If threshold = 0.5, we will include 60% of coding regions and also 2% of non-coding regions
Known Gene Finders

- GeneScan
- GeneMarkHMM
- FgenesH
- GlimmerHMM
- GeneZilla
- SNAP
- PHAT
- AUGUSTUS
- Genie
Challenges of Gene finder

- Alternative splicing
- Nested/overlapping genes
- Extremely long/short genes
- Extremely long introns
- Non-canonical introns
- Split start codons
- UTR introns
- Non-ATG triplet as the start codon
- Polycistronic genes
- Repeats/transposons
Evaluation of Gene prediction

- Sensitivity = No. of Correct exons/No. of actual exons
- Specificity = No. of Correct exons/No. of predicted exons
Codon preference

- Uneven usage of amino acids in the existing proteins
- Uneven usage of synonymous codons.
- S = AGGACG, when read in frame 1, it results in the sequence

\[
\begin{align*}
C_1^1 &= AGG \\
C_2^1 &= ACG
\end{align*}
\]

\[
P^1(S) = P(C^1) = F(AGG)F(ACG)
\]

\[
P^1(S) = P(C^1) = 0.022 \times 0.038 = 0.000836
\]

\[
F_0(c) = 1/64 = 0.0156
\]

\[
P_0(S) = P_0(C) = F_0(C_1)F_0(C_2)\cdots F_0(C_m)
\]

\[
P_0(C) = 0.0156 \times 0.0156 = 0.000244
\]

\[
LP^1(S) = \log(0.000836/0.000244) = \log(3.43) = 0.53
\]

6/5/2009
In coding sequences 2 Ts are separated by $2+3n$ (i.e.; $2, 5, 8, 11$)

- $A1 =$ Number of A's in positions $1, 4, 7, 10, \ldots$
- $A2 =$ Number of A's in positions $2, 5, 8, 11, \ldots$
- $A3 =$ Number of A's in positions $3, 6, 9, 12, \ldots$
- $\text{MAX}(A1, A2, A3)/\text{MIN}(A1, A2, A3)+1$ ----4 position parameters

- A,T,G,C content of the sequence form the content parameter – 4 content parameters

- Weight is assigned to each parameter
- Value above a certain threshold is coding.
Problem: Find Exon-Intron boundaries and start, stop codon for this sequence.

!!NA_SEQUENCE 1.0 EMBOSS_001 Length: 645 Type: N Check: 3485 ..

1   aggttttgtcatgacgatgaaacagtgagctagaagcctgttatatcgact
 51  acaatagacgacgacaggaagcctttggaccacggagaaggaagattaatg
101 atgagagattggaacctccccgtttccaaagagatccagctgccgctccat
151 tgctactggattgaacttttcagacagacttttggtgtgcaagttagcat
201 atgcgaactcagtcggcttttttaaagttgacagcactttgcaagggcggtta
251 cgcataatgatctttttACTAOTCAATgtaaaatcttttcacattcaattggtc
301 gtctccggactggttcgtggtttcctttatattaatatatgcttggttccttgc
351 atgcaggttagccccagcgccctttccacacattccgagccgagcaca
401 ggaaattgcttggtgctcagacctaaaagtgagagcggtggcagttcttga
451 cgctcgaagagcaagtgggttgctgacacacttggacggggcaccaggagctaagg
501 attgaaacaccgactgtagcgtcgtccagtacggtttgtccacgaatct
551 actgcttatgtgttagtttattttgctactaggtttgtttgtgatctatatttcag
601 ttctttggtcccccatcagggaaagcgatcgcagctcggtttgtgtgatga
Steps:

1. Build PSSMs for start, donor, acceptor sites.
2. Mark all donor, acceptor, start, stop sites.
3. Eliminate unlikely donor, acceptor sites.
4. Score each site from PSSM.
5. Check frame compatibility.
6. Run a Blastx to nr database.
7. Translate and check if peptide sequence follows the dipeptide frequency norm.

Web Resources:  http://vmd.vbi.vt.edu
Thank You!