Introduction to BioInformatics

BSC4933/5936
Florida State University
Department of Biology
www.bio.fsu.edu

October 2, 2003

Genomics

Steven M. Thompson
Florida State University School of Computational Science and Information Technology (CSIT)

What sort of information can be determined from a genomic sequence?
Lecture Topics —

Easy — restriction digests and associated mapping; e.g. software like the Wisconsin Package’s (Genetics Computer Group [GCG]) Map, MapSort, and MapPlot.

Harder — fragment assembly and genome mapping; such as packages from the University of Washington’s Genome Center (http://www.genome.washington.edu/), Phred/Phrap/Consed (http://www.phrap.org/) and SegMap, and The Institute for Genomic Research’s (http://www.tigr.org/) Lucy and Assembler programs.

Very hard — gene finding and sequence annotation. This will be the bulk of today’s lecture and is a primary focus in current genomics research.

Easy— forward translation to peptides.

Hard again — genome scale comparisons and analyses.

Nucleic Acid Characterization: Recognizing Coding Sequences.

Three general solutions to the gene finding problem:

1) all genes have certain regulatory signals positioned in or about them,

2) all genes by definition contain specific code patterns,

3) and many genes have already been sequenced and recognized in other organisms so we can infer function and location by homology if our new sequence is similar enough to an existing sequence.

All of these principles can be used to help locate the position of genes in DNA and are often known as “searching by signal,” “searching by content,” and “homology inference” respectively.
**URFs and ORFs — definitions**

**URF**: Unidentified Reading Frame — any potential string of amino acids encoded by a stretch of DNA. Any given stretch of DNA has potential URFs on any combination of six potential reading frames, three forward and three backward.

**ORF**: Open Reading Frame — by definition any continuous reading frame that starts with a start codon and stops with a stop codon. Not usually relevant to discussions of genomic eukaryotic DNA, but very relevant when dealing with mRNA/cDNA or prokaryotic DNA.

---

**Signal Searching:**

locating transcription and translation affecter sites.

**One strategy — One-Dimensional Signal Recognition.**

**Start Sites:**

- Prokaryote promoter ‘Pribnow Box,’ TTGACwx{15,21}TAaT;
- Eukaryote transcription factor site database, TFSites.Dat;
- Shine-Dalgarno site, (AGG,GAG,GGA)x{6,9}ATG, in prokaryotes;
- Kozak eukaryote start consensus, cc(A,g)ccAUGg;
- AUG start codon in about 90% of genomes, exceptions in some prokaryotes and organelles.
Signal Searching: locating transcription and translation affecter sites.
One-Dimensional Approaches, cont.

End Sites:
‘Nonsense’ chain terminating, stop codons — UAA, UAG, UGA;
Eukaryote terminator consensus, YGTGTTYY;
Eukaryote poly(A) adenylation signal AAUAAA;
but exceptions in some ciliated protists and due to eukaryote suppresser tRNAs.


Exon/Intron Junctions.

<table>
<thead>
<tr>
<th>Donor Site</th>
<th>Acceptor Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_{64}G_{73}G_{100}T_{108}A_{62}A_{66}G_{64}T_{63}</td>
<td>6P_{y_{74-87}}NC_{65}A_{100}C_{108}N</td>
</tr>
</tbody>
</table>

The splice cut sites occur before a 100% GT consensus at the donor site and after a 100% AG consensus at the acceptor site, but a simple consensus is not informative enough.
Signal Searching:
locating transcription and translation affector sites.
Two-Dimensional Weight Matrices describe the probability at
each base position to be either A, C, U, or G, in percentages.

The Donor Matrix.
CONSENSUS from:  Donor Splice site sequences
from Stephen Mount NAR 10(2) 459;472 figure 1 page 460

<table>
<thead>
<tr>
<th></th>
<th>Exon</th>
<th>cut-site</th>
<th>Intron</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>20</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>A</td>
<td>30</td>
<td>40</td>
<td>64</td>
</tr>
<tr>
<td>U</td>
<td>20</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>C</td>
<td>30</td>
<td>44</td>
<td>11</td>
</tr>
</tbody>
</table>

%G  20  9  11  74  100  0  29  12  84  9  18  20
%A  30  40  64  9  0  0  61  67  9  16  39  24
%U  20  7  13  12  0  100  7  11  5  63  22  27
%C  30  44  11  6  0  0  2  9  2  12  20  28

CONSENSUS sequence to a certainty level of 75 percent.
VMWK©RRGWHH

The matrix begins four bases ahead of the splice site!

Signal Searching:
locating transcription and translation affector sites.
Two-Dimensional Weight Matrices, cont.

The Acceptor Matrix.
CONSENSUS of:  Acceptor.Dat. IVS Acceptor Splice Site Sequences
from Stephen Mount NAR 10(2); 459-472 figure 1 page 460

<table>
<thead>
<tr>
<th></th>
<th>Intron</th>
<th>cut-site</th>
<th>Exon</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>15</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td>A</td>
<td>15</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>T</td>
<td>52</td>
<td>40</td>
<td>54</td>
</tr>
<tr>
<td>C</td>
<td>18</td>
<td>25</td>
<td>30</td>
</tr>
</tbody>
</table>

%G  15  22  10  10  6  7  9  7  5  5  24  1  0  100  52  24  19
%A  15  10  10  15  6  15  11  19  12  3  10  25  4  100  0  22  17  20
%T  52  44  50  54  60  49  48  45  45  57  58  30  31  0  0  8  37  29
%C  18  25  30  21  24  30  34  28  36  35  27  21  64  0  0  18  22  32

CONSENSUS sequence to a certainty level of 75.0 percent at each
position:
BBYHYHYHYDDY©VBH

The matrix begins fifteen bases upstream of the splice site!
Signal Searching:
locating transcription and translation affecter sites.
Two-Dimensional Weight Matrices, cont.

The CCAAT site — occurs around 75 base pairs upstream of the start point of eukaryotic transcription, may be involved in the initial binding of RNA polymerase II.

Preferred region: motif within -212 to -57. Optimized cut-off value: 87.2%.

<table>
<thead>
<tr>
<th></th>
<th>%G</th>
<th>7</th>
<th>25</th>
<th>14</th>
<th>40</th>
<th>57</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>12</th>
<th>9</th>
<th>34</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>%A</td>
<td>32</td>
<td>18</td>
<td>14</td>
<td>58</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>68</td>
<td>10</td>
<td>13</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>%U</td>
<td>30</td>
<td>27</td>
<td>45</td>
<td>1</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>15</td>
<td>82</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>%C</td>
<td>31</td>
<td>30</td>
<td>27</td>
<td>1</td>
<td>3</td>
<td>99</td>
<td>99</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>51</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

CONSENSUS sequence to a certainty level of 68 percent at each position:

HBYRCCAAATSR

Signal Searching:
locating transcription and translation affecter sites.
Two-Dimensional Weight Matrices, cont.

The TATA site (aka "Hogness" box) — a conserved A-T rich sequence found about 25 base pairs upstream of the start point of eukaryotic transcription, may be involved in positioning RNA polymerase II for correct initiation and binds Transcription Factor IID.

Preferred region: center between -36 and -20. Optimized cut-off value: 79%.

<table>
<thead>
<tr>
<th></th>
<th>%G</th>
<th>39</th>
<th>5</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>5</th>
<th>11</th>
<th>40</th>
<th>39</th>
<th>33</th>
<th>33</th>
<th>36</th>
</tr>
</thead>
<tbody>
<tr>
<td>%A</td>
<td>16</td>
<td>4</td>
<td>90</td>
<td>1</td>
<td>91</td>
<td>69</td>
<td>93</td>
<td>57</td>
<td>40</td>
<td>14</td>
<td>21</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>%U</td>
<td>8</td>
<td>79</td>
<td>9</td>
<td>95</td>
<td>8</td>
<td>31</td>
<td>2</td>
<td>31</td>
<td>8</td>
<td>12</td>
<td>8</td>
<td>13</td>
<td>16</td>
</tr>
<tr>
<td>%C</td>
<td>37</td>
<td>12</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>11</td>
<td>35</td>
<td>38</td>
<td>33</td>
<td>30</td>
<td>28</td>
<td>26</td>
</tr>
</tbody>
</table>

CONSENSUS sequence to a certainty level of 61 percent at each position:

STATAWAWSRSSS
Signal Searching:
locating transcription and translation affecter sites.

Two-Dimensional Weight Matrices, cont.

The GC box — may relate to the binding of transcription factor Sp1.

Preferred region: motif within -164 to +1. Optimized cut-off value: 88%.

%G 18 41 56 75 100 99 0 82 81 62 70 13 19 40
%A 37 35 18 24 0 1 20 17 0 29 8 0 7 15
%U 30 12 23 0 0 0 18 1 18 9 15 27 42 37
%C 15 11 2 0 0 0 62 0 1 0 6 61 31 9

CONSENSUS sequence to a certainty level of 67 percent at each position:

WRKGGGGGRGBYK

Signal Searching:
locating transcription and translation affecter sites.

Two-Dimensional Weight Matrices, cont.

The cap signal — a structure at the 5' end of eukaryotic mRNA introduced after transcription by linking the 5' end of a guanine nucleotide to the terminal base of the mRNA and methylating at least the additional guanine; the structure is $^7$MeG$^5$ppp$^5$Np.

Preferred region: center between 1 and +5. Optimized cut-off value: 81.4%.

%G 23 0 0 38 0 15 24 18
%A 16 0 0 95 9 25 22 15 17
%U 45 0 5 26 43 24 33 33
%C 16 100 0 27 31 39 28 32

CONSENSUS sequence to a certainty level of 63 percent at each position:

KCAHBYB
Signal Searching:  
locating transcription and translation affecter sites.  
Two-Dimensional Weight Matrices, cont.

The eukaryotic terminator weight matrix.

Base frequencies according to McLauchlan et al.  
Found in about 2/3’s of all eukaryotic gene sequences.

<table>
<thead>
<tr>
<th>%G</th>
<th>19</th>
<th>81</th>
<th>9</th>
<th>94</th>
<th>14</th>
<th>10</th>
<th>11</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>%A</td>
<td>13</td>
<td>9</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>%U</td>
<td>51</td>
<td>9</td>
<td>89</td>
<td>3</td>
<td>79</td>
<td>61</td>
<td>56</td>
<td>47</td>
</tr>
<tr>
<td>%C</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>29</td>
<td>21</td>
<td>21</td>
</tr>
</tbody>
</table>

CONSSENSUS sequence to a certainty level of 68 percent at each position:

BGTTGYY

Content Approaches. Strategies for finding coding regions based on the content of the DNA itself.

Searching by content utilizes the fact that genes necessarily have many implicit biological constraints imposed on their genetic code. This induces certain periodicities and patterns to produce distinctly unique coding sequences; non-coding stretches do not exhibit this type of periodic compositional bias. These principles can help discriminate structural genes in two ways:

1) based on the local "non-randomness" of a stretch, and
2) based on the known codon usage of a particular life form.

The first, the non-randomness test, does not tell us anything about the particular strand or reading frame; however, it does not require a previously built codon usage table. The second approach is based on the fact that different organisms use different frequencies of codons to code for particular amino acids. This does require a codon usage table built up from known translations; however, it also tells us the strand and reading frame for the gene products as opposed to the former.
Content Approaches, cont.  
"Non-Randomness" Techniques.  
GCG’s TestCode.  

Relies solely on the base compositional bias of every third position base.

The plot is divided into three regions: top and bottom areas predict coding and noncoding regions, respectively, to a confidence level of 95%, the middle area claims no statistical significance. Diamonds and vertical bars above the graph denote potential stop and start codons respectively.

Content Approaches, cont.  Codon Usage Techniques.  GCG’s CodonPreference.  

Genomes use synonymous codons unequally sorted phylogenetically.

Each forward reading frame indicates a red codon preference curve and a blue third position GC bias curve. The horizontal lines within each plot are the average values of each attribute. Start codons are represented as vertical lines rising above each box and stop codons are shown as lines falling below the reading frame boxes. Rare codon choices are shown for each frame with hash marks below each reading frame.
Internet World Wide Web servers.
Many servers have been established that can be a huge help with gene finding analyses. Most of these servers combine many of the methods previously discussed but they consolidate the information and often combine signal and content methods with homology inference in order to ascertain exon locations. Many use powerful neural net or artificial intelligence approaches to assist in this difficult ‘decision’ process.
A wonderful bibliography on computational methods for gene recognition has been compiled by Wentian Li (http://www.nsljj-genetics.org/gene/), and the Baylor College of Medicine’s Gene Feature Search (http://searchlauncher.bcm.tmc.edu/seq-search/gene-search.html) is another nice portal to several gene finding tools.

World Wide Web Internet servers, cont.
Five popular gene-finding services are GrailEXP, Geneld, GenScan, NetGene2, and GeneMark.
The neural net system GrailEXP (Gene recognition and analysis internet link–EXPanded http://grail.lsd.ornl.gov/grailexp/) is a gene finder, an EST alignment utility, an exon prediction program, a promoter and polyA recognizer, a CpG island locator, and a repeat masker, all combined into one package.
Geneld (http://www1.imim.es/software/geneid/index.html) is an ‘ab initio’ Artificial Intelligence system for predicting gene structure optimized in genomic Drosophila or Homo DNA.
NetGene2 (http://www.cbs.dtu.dk/services/NetGene2/), another ‘ab initio’ program, predicts splice site likelihood using neural net techniques in human, C. elegans, and A. thaliana DNA.
GenScan (http://genes.mit.edu/GENSCAN.html) is perhaps the most ‘trusted’ server these days with vertebrate genomes.
The GeneMark (http://opal.biology.gatech.edu/GeneMark/) family of gene prediction programs is based on Hidden Markov Chain modeling techniques; originally developed in a prokaryotic context the programs have now been expanded to include eukaryotic modeling as well.
Homology Inference.

Similarity searching can be particularly powerful for inferring gene location by homology. This can often be the most informative of any of the gene finding techniques, especially now that so many sequences have been collected and analyzed. Wisconsin Package programs such as the BLAST and FastA families, Compare and DotPlot, Gap and BestFit, and FrameAlign and FrameSearch can all be a huge help in this process. But this too can be misleading and seldom gives exact start and stop positions. For example:

805 GCCATGCCCCGGGCGACGGGACGGGCCCAGCTGAGGAGGCC...CT 851
        ||   ||   ||   ||   ||   ||   ||   ||   ||
46 AlaAlaAlsArgCysLysAlsAlsGlULaAlsAlsAspGluProAlsLe  62
852 GAGCTGCTGACGACATGACCCAGCTACCTCCCCTGCGGCCGAACTTG 901
        ||   ||   ||   ||   ||   ||   ||   ||   ||   ||
63 uCysLeuGlnCysAspMetAsnAspCysTyrSerArgLeuArgArgLeuV 79
        ||   ||   ||   ||   ||   ||   ||   ||   ||   ||
902 TACCCGGAGTCTCCGAGAGGCACCCGCTAGCTGCCCTGCCGTAAGGCAGAACATCTACAG 951
        ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||
80 aProThrIleProProAsnLysValSerLysValGluIleLeuLeuGlnLeu 95
        ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||
952 GCGTCTACATGATCTTCGACCTGCTTCTGCTGGAGTGTGCTCAG 990
        ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||   ||
96 HisValIeAspTyrIleLeuAspLeuGlnLeuAlsAlsLeu 108

Summary. The combinatorial approach.

Get all your data in one place. GCG’s SeqLab is a great way to do this due to its advanced annotation capabilities:
The Objective: a complete protein.

Now what?

Beyond just finding genes: Genome scale analyses.

Unfortunately much ‘traditional’ sequence analysis software can’t do it, but there are some very good Web resources available for these types of ‘global view’ analyses. Let’s run through a few examples. NCBI’s Genome pages (http://www.ncbi.nlm.nih.gov/) present a good starting point in North America:
Beyond just finding genes: Genome scale analyses, cont.

That can lead to neat places like the Genome Browser at the University of California, Santa Cruz (http://genome.ucsc.edu/) and the Ensembl project at the Sanger Center for BioInformatics (http://www.ensembl.org):

Beyond just finding genes: Genome scale analyses, cont.

And sites like the the University of Wisconsin’s *E. coli* Genome Project (http://www.genome.wisc.edu/) and The Institute for Genomic Research’s (http://www.tigr.org/) MUMMER package.
## References

A perplexing variety of techniques exist for the identification and analysis of protein coding regions in genomic DNA. Knowing which to use when and how to combine their inferences will go a long way in your genomic analyses!


