Finding Genes with Genemark and Glimmer

Genomes only need to be self consistent

- At the physical level, DNA is a long string of nucleotides that encodes information
  - As long as the organism can get the information it needs, it doesn't matter how it gets that information
- Between different organisms, nothing strictly demands that they encode their information the same way
  - The only reason there are any similarities is that every organism "reads" the genome in a way similar to its ancestors

- This makes gene finding hard.

Genemark and Glimmer predict genes

- Genemark and Glimmer have a tough job: they must identify genes in a general range of organisms, even though those organisms start and stop genes in different ways.

- How they work: <u>trainable</u> algorithms
  - Because of evolution, related organisms tend to start and stop genes in the same way
  - Genemark and Glimmer calibrate on many examples of genes from related organisms, and then make predictions on a new sequence

A general workflow: Genemark/Glimmer

1) Gathering a training set
   1) A large number of sequences where starts and stops are encoded similiarly, and annotated, thank to existing work
2) The training set is fed to Genemark/Glimmer examples with labeled gene starts and finishes

```
ACTAGGATGCATC....  AAAACAGTAGATCGA
CCATGGATGCTTC....  AATCAGATAGATCGA
ACTAAAATGCAAT....  CCTACGATAGATCGA
```

An example with Genemark

less /proj/cse308/Project2/genemarkDemo/hiv.genBank

```
gene            336..1838
                /gene="gag"
CDS             336..1838
                /gene="gag"
                /note="Pr55"
                /codon_start=1
                /product="Gag"
                /protein_id="AAC82593.1"
                /db_xref="GI:2801504"
                /translation="MGARASVLSGGELDRWEKIRLRPGGKKKYKLKHIVWASRELERF
                AVNPGLLETSEGCRQILGQLQPSLQTGSEELRSLYNTVATLYCVHQRIEIKDTKEALD
                KIEEEQNKSKKKAQQAAADTGHSNQVSQNYPIVQNIQGQMVHQAISPRTLNAWVKVVE
                EKAFSPEVIPMFSALSEGATPQDLNTMLNTVGGHQAAMQMLKETINEEAAEWDRVHPV
                HAGPIAPGQMREPRGSDIAGTTSTLQEQIGWMTNNPPIPVGEIYKRWIILGLNKIVRM
                YSPTSILDIRQGPKEPFRDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTIL
                KALGPAATLEEMMTACQGVGGPGHKARVLAEAMSQVTNSATIMMQRGNFRNQRKIVKC
                FNCGKEGHTARNCRAPRKKGCWKCGKEGHQMKDCTERQANFLGKIWPSYKGRPGNFLQ
                SRPEPTAPPEESFRSGVETTTPPQKQEPIDKELYPLTSLRSLFGNDPSSQ"
```

Gene Start and stop

Gene Name

Protein Sequence

This part of the GenBank file tells you about a specific gene, notably, where it starts and stops.

Getting genes from the geneBank genome

```
ORIGIN
        1 ggtctctctg gttagaccag atctgagcct gggagctctc tggctaacta gggaacccac
       61 tgcttaagcc tcaataaagc ttgccttgag tgcttcaagt agtgtgtgcc cgtctgttgt
      121 gtgactctgg taactagaga tccctcagac ccttttagtc agtgtggaaa atctctagca
      181 gtggcgcccg aacagggacc tgaaagcgaa agggaaacca gaggagctct ctcgacgcag
      241 gactcggctt gctgaagcgc gcacggcaag aggcgagggg cggcgactgg tgagtacgcc
      301 aaaaattttg actagcggag gctagaagga gagagatggg tgcgagagcg tcagtattaa
      361 gcgggggaga attagatcga tgggaaaaaa ttcggttaag gccagggggga aagaaaaaat
      421 ataaattaaa acatatagta tgggcaagca gggagctaga acgattcgca gttaatcctg
      481 gcctgttaga aacatcagaa ggctgtagac aaatactggg acagctacaa ccatcccttc
      541 agacaggatc agaagaactt agatcattat ataatacagt agcaaccctc tattgtgtgc
      601 atcaaaggat agagataaaa gacaccaagg aagctttaga caagatagag gaagagcaaa
      661 acaaaagtaa gaaaaaagca cagcaagcag cagctgacac aggacacagc aatcaggtca
      721 gccaaaatta ccctatagtg cagaacatcc aggggcaaat ggtacatcag gccatatcac
      781 ctagaacttt aaatgcatgg gtaaaagtag tagaagagaa ggctttcagc ccagaagtga
      841 tacccatgtt ttcagcatta tcagaaggag ccaccccaca agatttaaac accatgctaa
      901 acacagtggg gggacatcaa gcagccatgc aaatgttaaa agagaccatc aatgaggaag
      961 ctgcagaatg ggatagagtg catccagtgc atgcagggcc tattgcacca ggccagatga
     1021 gagaaccaag gggaagtgac atagcaggaa ctactagtac ccttcaggaa caaataggat
     1081 ggatgacaaa taatccacct atcccagtag gagaaattta taaaagatgg ataatcctgg
     1141 gattaaataa aatagtaaga atgtatagcc ctaccagcat tctggacata agacaaggac
     1201 caaaggaacc ctttagagac tatgtagacc ggttctataa aactctaaga gccgagcaag
     1261 cttcacagga ggtaaaaaat tggatgacag aaaccttgtt ggtccaaaat gcgaacccag
     1321 attgtaagac tattttaaaa gcattgggac cagcggctac actagaagaa atgatgacag
     1381 catgtcaggg agtaggagga cccggccata aggcaagagt tttggctgaa gcaatgagcc
     1441 aagtaacaaa ttcagctacc ataatgatgc agagaggcaa ttttaggaac caaagaaaga
     1501 ttgttaagtg tttcaattgt ggcaaagaag ggcacacagc cagaaattgc agggccccta
     1561 ggaaaaaggg ctgttggaaa tgtggaaagg aaggacacca aatgaaagat tgtactgaga
     1621 gacaggctaa ttttttaggg aagatctggc cttcctacaa gggaaggcca gggaattttc
     1681 ttcagagcag accagagcca acagccccac cagaagagag cttcaggtct ggggtagaga
     1741 caacaactcc ccctcagaag caggagccga tagacaagga actgtatcct ttaacttccc
     1801 tcaggtcact ctttggcaac gacccctcgt cacaataaag ataggggggc aactaaagga
     1861 agctctatta gatacaggag cagatgatac agtattagaa gaaatgagtt tgccaggaag
     1921 atggaaacca aaaatgatag ggggaattgg aggttttatc aaagtaagac agtatgatca
     1981 gatactcata gaaatctgtg gacataaagc tataggtaca gtattagtag gacctacacc
     2041 tgtcaacata attggaagaa atctgttgac tcagattggt tgcactttaa attttcccat
```

Gag gene:
336-1838

A note on cutting and pasting genBank files

```
ORIGIN
        1 ggtctctctg gttagaccag atctgagcct gggagctctc tggctaacta gggaacccac
       61 tgcttaagcc tcaataaagc ttgccttgag tgcttcaagt agtgtgtgcc cgtctgttgt
      121 gtgactctgg taactagaga tccctcagac ccttttagtc agtgtggaaa atctctagca
      181 gtggcgcccg aacagggacc tgaaagcgaa agggaaacca gaggagctct ctcgacgcag
```

- The position of the first nucleotide in every row is numbered on the left.
- Nucleotides are divided into groups of 10
- If you want 336 to 1838, this formatting makes it easy to cut and paste what you need.

- If you want to get rid of spaces, use nano. “Ctrl-w, Ctrl-r” allows you to replace strings with other strings: replace spaces with nothing.

I've cut and pasted this gene out for you

cd /home/chen/308test/Project2/genemarkDemo

less geneMarkDemo.fasta

```
> hiv gag gene
atgggtgcgagagcgtcagtattaagcgggggagaattagatcgatgggaaaaaattcggttaaggccagggggaaagaaaaaatataaattaaaacatatag
tatgggcaagcagggagctagaacgattcgcagttaatcctggcctgttagaaacatcagaaggctgtagacaaatactgggacagctacaaccatcccttca
gacaggatcagaagaacttagatcattatataatacagtagcaaccctctattgtgtgcatcaaaggatagagataaaagacaccaaggaagctttagacaag
atagaggaagagcaaaacaaaagtaagaaaaaagcacagcaagcagcagctgacacaggacacagcaatcaggtcagccaaaattaccctatagtgcagaaca
tccaggggcaaatggtacatcaggccatatcacctagaactttaaatgcatgggtaaaagtagtagaagagaaggctttcagcccagaagtgatacccatgtt
ttcagcattatcagaaggagccaccccacaagatttaaacaccatgctaaacacagtgggggggacatcaagcagccatgcaaatgttaaaagagaccatcaat
gaggaagctgcagaatgggatagagtgcatccagtgcatgcagggcctattgcaccaggccagatgagagaaccaaggggaagtgacatagcaggaactacta
gtacccttcaggaacaaataggatggatgacaaataatccacctatcccagtaggagaaatttataaaagatggataatcctgggattaaataaaatagtaag
aatgtatagccctaccagcattctggacataagacaaggaccaaaggaacccttttagagactatgtagaccggttctataaaactctaagagccgagcaagct
tcacaggaggtaaaaaattggatgacagaaaccttgttggtccaaaatgcgaacccagattgtaagactattttaaaagcattgggaccagcggctacactag
aagaaatgatgacagcatgtcagggagtaggaggacccggccataaggcaagagttttggctgaagcaatgagccaagtaacaaattcagctaccataatgat
gcagagaggcaattttaggaaccaaagaaagattgttaagtgtttcaattgtggcaaagaagggcacacagccagaaattgcagggcccctaggaaaaagggc
tgttggaaatgtggaaaggaaggacaccaaatgaaagattgtactgagagacaggctaatttttagggaagatctggccttcctacaagggaaggccaggga
attttcttcagagcagaccagagccaacagccccaccagaagagagcttcaggtctggggtagagacaacaactccccctcagaagcaggagccgatagacaa
ggaactgtatcctttaacttccctcaggtcactctttggcaacgacccctcgtcacaataa
```

This could also be one of the contigs from Project 2, except that we know this is exactly one gene.

In Project 2, a contig could have part of a gene, no genes, multiple genes.

Lets go to the Genemark website
http://exon.biology.gatech.edu/heuristic_hmm2.cgi

Heuristic Approach for Gene Prediction in Prokaryotes (Reload this page)

Reference: Besemer J. and Borodovsky M., Heuristic approach to deriving models for gene finding, **NAR**, 1999, Vol. 27, No. 19, pp. 3911-3920.
[Download PDF]

GeneMark.hmm 2.0 and GeneMark 2.4 use model parameters estimated as described in the paper mentioned above. Please note that the program output for sequences larger than 1 MB is sent by email.

Input Sequence

Title (optional):

Sequence:

To find genes from fasta sequences, paste them here.

Run Default Start GeneMark.hmm

Start GeneMark here

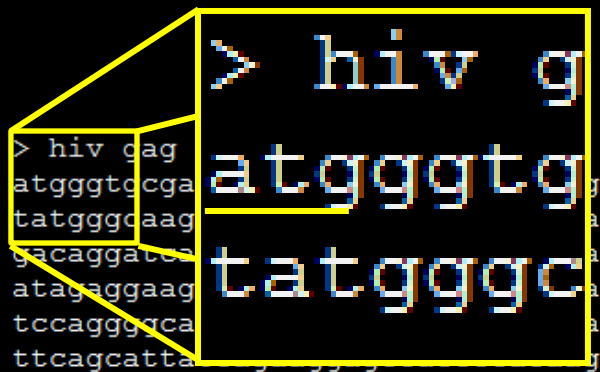
Paste genemarkDemo.txt into GeneMark

Parse predicted by GeneMark.hmm 2.0

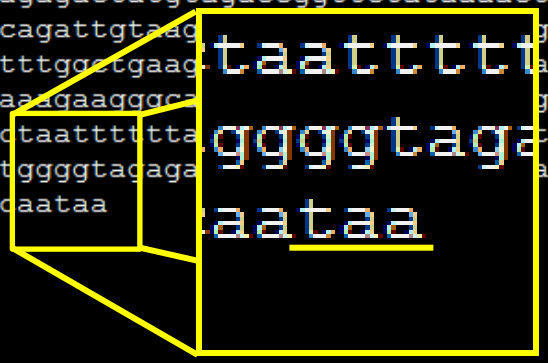
```
GeneMark.hmm PROKARYOTIC (Version 2.8)
Date: Thu Mar 17 07:56:08 2011
Sequence file name: sequence
Model file name: heuristic_no_rbs.mat
RBS: N
Model information: Heuristic_model_for_genetic_code_11_and_GC_30

FASTA definition line: Thu Mar 17 07:56:08 EDT 2011
Predicted genes
   Gene     Strand     LeftEnd     RightEnd        Gene      Class
    #                                            Length
    1         +           1          1503          1503         1
```

Here GeneMark says that our exact sequence is also a gene

We can see this is possible, visually, too.



Classic Start Codon



Classic Stop Codon

Glimmer works much the same way

www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi

NCBI Microbial Genomes

HOME | SEARCH | SITE MAP | Genome Project | Genome | Prokaryotic Projects | Collaborators | gMap | ProtMap | TaxPlot | BLAST | FTP | Contact us

Microbial Genome Annotation Tools

GLIMMER is a system for finding genes in microbial DNA, especially the genomes of bacteria, archaea, and viruses. GLIMMER (Gene Locator and Interpolated Markov ModelER) uses interpolated Markov models to identify coding regions.

- Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER, *Nucleic Acids Research* 27:23 (1999), 4636-4641.
- Salzberg S, Delcher A, Kasif S, White O. Microbial gene identification using interpolated Markov models, *Nucleic Acids Research* 26:2 (1998), 544-548.

Download GLIMMER from the Center for Bioinformatics and Computational Biology.

Genomes
- Genome Projects
- Prokaryotic Projects
- Microbial Genomes
- Home
- Complete Genomes
- Draft Assemblies
- Registered
- Plasmids
- Entrez Genome

Submit a Genome
- Sequin
- Submission Guide
- Register a Project
- Submit a Genome
- Submit Traces

Tools

Resources

Sequencing Centers

Collaborators

Statistics

Scroll down a bit to submit fasta sequences

Upload your sequence from file:

Choose File No file chosen

Or copy/paste your sequence FASTA here:

```
> HIV COMPLETE
GGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGAACCCAC
TGCTTAAGCCTCAATAAAGCTTGCCTTGAGTGCTTCAAGTAGTGTGTGCCCGTCTGTTGT
GTGACTCTGGTAACTAGAGATCCCTCAGACCCTTTTAGTCAGTGTGGAAAATCTCTAGCA
GTGGCGCCCGAACAGGGACCTGAAAGCGAAAGGGAAACCAGAGGAGCTCTCTCGACGCAG
GACTCGGCTTGCTGAAGCGCGCACGGCAAGAGGCGAGGGGCGGCGACTGGTGAGTACGCC
AAAAATTTTGACTAGCGGAGGCTAGAAGGAGAGAGATGGGTGCGAGAGCGTCAGTATTAA
GCGGGGGAGAATTAGATCGATGGGAAAAAATTCGGTTAAGGCCAGGGGGAAAGAAAAAAT
```

Additional parameters:

Genetic code:

- (●) 11 (Bacteria, Archaea)
- ( ) 4 (Mycoplasma/Spiroplasma)

Topology:

- ( ) circular
- (●) linear

Run GLIMMER v3.02

I used this file:

/proj/cse308/Project2/genemarkDemo/hiv.fasta

Make sure to set topology to linear

Glimmer finds several genes in all of HIV

Submit new Data

GLIMMER (ver. 3.02; iterated) predictions:

orfID	start	end	frame	score
>HIV COMPLETE				
orf00001	106	255	+1	5.81
orf00004	336	1838	+3	3.29
orf00005	1904	4642	+2	3.07
orf00006	4608	5165	+3	4.20
orf00011	5771	8341	+2	3.20
orf00013	8554	8336	-2	3.54
orf00015	8913	8740	-1	5.59
orf00016	9145	8975	-2	10.62

From Glimmer

gene	336..1838
gene	<1631..4642
gene	4587..5165
gene	5105..5341
gene	5377..7970
gene	5516..8199
gene	5608..5856
gene	5771..8341
gene	8343..8714

GenBank data

Glimmer and GeneMark are not perfectly accurate as you can see here, but they often get close.

Examples for Sequence Alignment

```
[08:42 AM][chen@jupiter sampleSequences] cat HomologsShort.aln
CLUSTAL W (1.83) multiple sequence alignment


Mouse           GTATCCAAC
Fly             GTATCAAAT
                ***** **
[08:42 AM][chen@jupiter sampleSequences] cat HomologsMedium.aln
CLUSTAL W (1.83) multiple sequence alignment


Mouse           GTATCCAACGGTTGTGTGAGTAAAATTCT
Fly             GTATCAAATGGATGTGTGAGCAAAATTCT
                ***** ** ** ******** ********
[08:43 AM][chen@jupiter sampleSequences] cat Homologs.aln
CLUSTAL W (1.83) multiple sequence alignment


Shark           GTGTCCAACGGTTGTGTCA-----GTAAAATCCTGGGC------AGATACTATGAAACAG GATCCATCAGA
MutShark        GTGTCCACCGGTTGTGTCA-----CTAATTTCCTGGGC------AGATACTATGAAACAG GATCCATCAGA
Mouse           GTATCCAACGGTTGTGTGA-----GTAAAATTCTGGG------CAGGTATTACGAGACTG GCTCCATCAGA
InsMouse        GTATCCAACGGTTGTGTGACAGATGTAAAATTCTGGGGCATCACAGGTATTACGAGACTG GCTCCATCAGA
Fly             GTATCAAATGGATGTGTGA-----GCAAAATTCTCGGG------AGGTATTATGAAACAG GAAGCATACGA
DelFly          GTATCAAATGGATGTGTGA-----GCAAAATTCTCGGG------AGGTATTATGAAACAG GAAGCATACGA
Squid           GTCTCCAACGGCTGCGTTA-----GCAAGATTCTCGGA------CGGTACTATGAGACGG GCTCCATAAGA
Flatword        GTGTCTAATGGTTGTGTTA-----GTAAAATACTTTGC------CGATATTATGGAACAG GTTCTATTAAA
                ** ** *  ** ** ** *       **  * **  *        * ** ** *  ** **    **   *
```

These files are located here:

/proj/cse308/Project2/sampleSequences