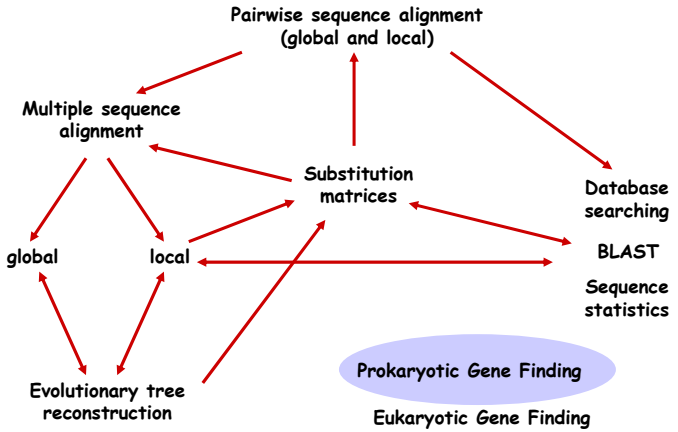
- Tues, Nov 29:
  Gene Finding 1
  Online FCE's: Thru Dec 12
- Thurs, Dec 1:
  Gene Finding 2
- Tues, Dec 6:
  PS5 due
  Project presentations 1 (see course web site for schedule)
- Thurs, Dec 8
  Final papers due
  Project presentations 2
- Monday Dec 19
  1pm - 4pm Final Exam, Room: HH B131



What is a Gene?

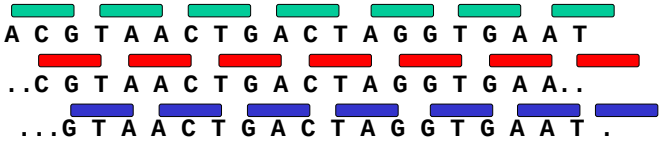
Snyder and Gerstein, Science 2003

- Something that encodes a heritable trait
- One gene, one enzyme
- One gene, one polypeptide
- One gene, one product (include RNA products)
- "A complete chromosomal segment responsible for making a functional product"
  - coding region
  - regulatory region
  - expressed product
  - functional product

Prokaryotic Gene Finding

- Identify Open Reading Frames (ORFs)
- Coding Statistics
- Identify individual gene architecture features
- Assemble an integrated gene description
- Homology

Reading Frames



- Each grouping of the nucleotides into consecutive triplets constitutes a reading frame.
- Three reading frames in the 5'->3'direction
- Three in the reverse direction on the opposite strand.

Open Reading Frames

An ORF is a contiguous set of codons, each specifying an amino acid (starting with ATG).

GGGAGCATGGTGCACCTGACTCCTGAGGTGACTTAGAC

M V H L T P E V T Stop

All coding sequences are ORF's, but not all ORF's encode proteins

Prokaryotic Gene Finding

- Identify Open Reading Frames (ORFs)
- Coding Statistics
- Identify individual gene architecture features
- Assemble an integrated gene description
- Homology

Coding Statistics

Fickett and Tung,1992
Guigo and Fickett,1995
(Electronicreserves)

- Codon usage
  - Determine codon (triplet) frequencies in known coding regions
  - Compare with codon frequencies in sliding window

ccgcctggcgtcgcggtttgttttcatctcttcatctgca

CodingStatistics

Fickett and Tung,1992
Guigo and Fickett,1995
(Electronicreserves)

- Codon usage — Species specific
- Codon pair preference — Species specific

ccgcctggcgtcgcggtttgttttcatctctcttcatctgca

CodingStatistics

Fickett and Tung,1992
Guigo and Fickett,1995
(Electronicreserves)

- Codon usage — Species specific
- Codon pair preference — Species specific
- Amino acid usage — Species specific

Gly Val Ala Val Cys Phe Ser

ccgcctggcgtcgcggtttgttttcatctctcttcatctgca

CodingStatistics

Fickett and Tung,1992
Guigo and Fickett,1995
(Electronic reserves)

- Codon usage — Species specific
- Codon pair preference — Species specific
- Amino acid usage — Species specific
- Amino acid pair preference — Species specific

Gly Val | Ala Val | Cys Phe | Ser Ser

ccgcctggcgtcgcggtttgttttcatctctcttcatctgca

CodingStatistics

Fickett and Tung,1992
Guigo and Fickett,1995
(Electronic reserves)

- Codon usage — Species specific
- Codon pair preference — Species specific
- Amino acid usage — Species specific
- Amino acid pair preference — Species specific
- Third position — Any organism
  - 3rd base tends to be the same much more often than chance

ccgcctggcgtcgcggtttgttttcatctctcttcatctgca

Coding Statistics continued

Fickett and Tung,1992
Guigo and Fickett,1995
(Electronic reserves)

CG content — Species specific

In *E. coli:*

Coding regions are embedded in segments of uniform, 53% G+C, about 1000 bases long

Non-coding regions are embedded in segments of uniform, 46% G+C, about 500 bases long

aa, at, ta, tt occur more frequently than expected in coding regions

tgccgcctggcgtcgcggtttcttttt catctctcttcatctg
acggcggaccgcagcgccaaagaaaaagtagagagaagtagac

CodingStatistics

Fickett and Tung,1992
Guigo and Fickett,1995
(Electronicreserves)

- Codon usage — Species specific
- Codon pair preference — Species specific
- Amino acid usage — Species specific
- Amino acid pair preference — Species specific
- Third position — Any organism
- CG content — Species specific

Look for variations in these measures in coding and non-coding regions (*inter*genic and *intra*genic).

Prokaryotic Gene Finding

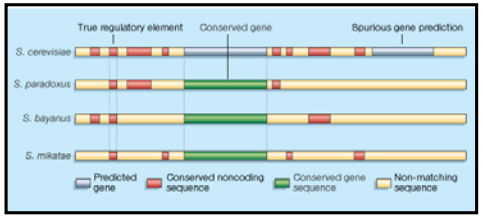
- Identify Open Reading Frames (ORFs)
- Coding Statistics
- Identify individual gene architecture features
- Assemble an integrated gene description
- Homology

SIGNALS IN THE *E.coli* lexA GENE

Prokaryotic Gene Finding

- Identify Open Reading Frames (ORFs)
- Coding Statistics
- Identify individual gene architecture features
- Assemble an integrated gene description
- Homology

Homology



Salzberg, Nature 2003

Gene Finding Questions

- Identify protein coding region
- Identify Open Reading Frame
- Predict mRNA (including UTR's)
- Predict intron/exon structure
  Eukaryotes only
- Regulatory signals
- Protein sequence

Prokaryotic Gene Finding

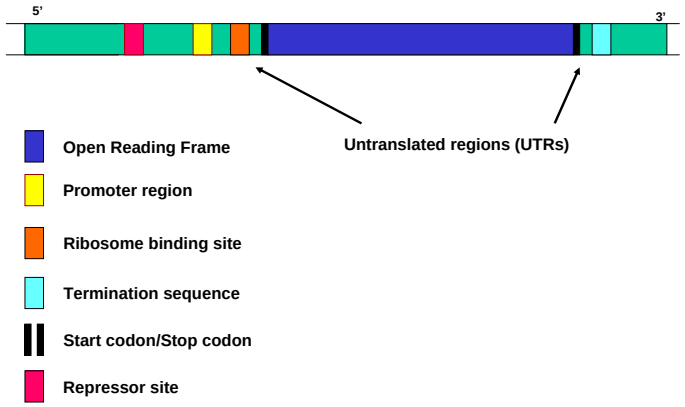
- Genome length: 0.5M bp – 10Mbp
- Coding density: ~90%
- Long ORFs are usually real genes

Early approaches

- Identify ORFs
- Score windows with coding statistics
- Identify gene structure elements

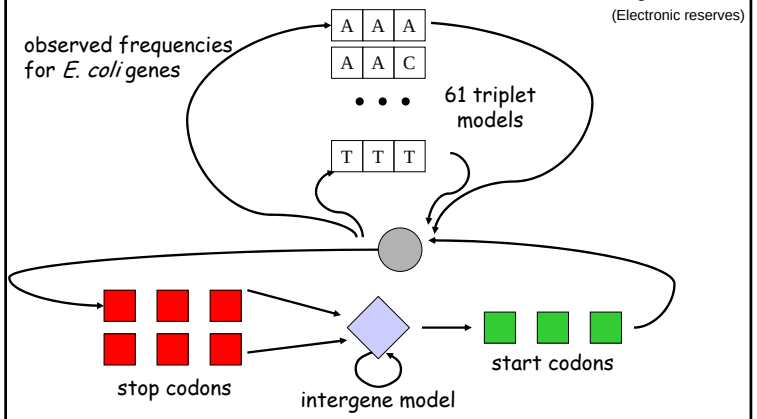
- Parse into a coherent gene model surrounded by intergenic DNA.

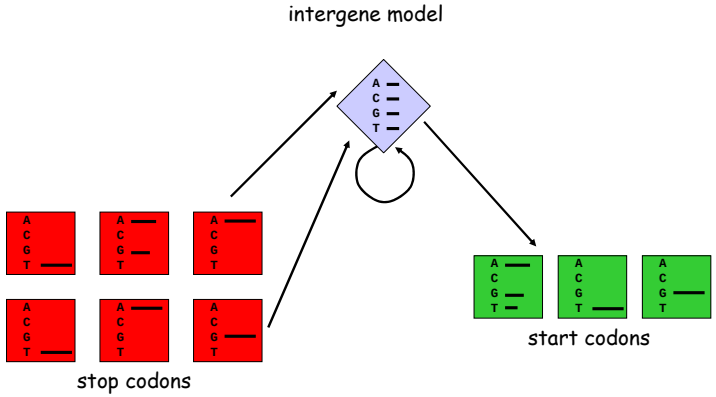
Prokaryotic gene model



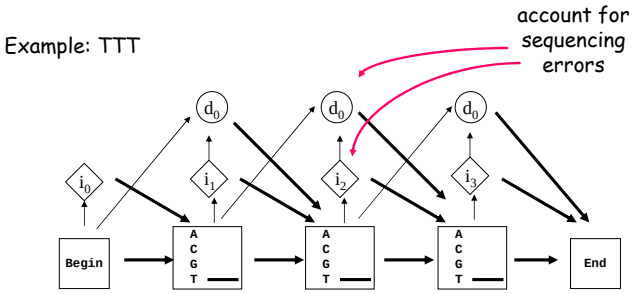
An HMM that finds genes in *E. coli*

Krogh *et al*,1995
(Electronic reserves)

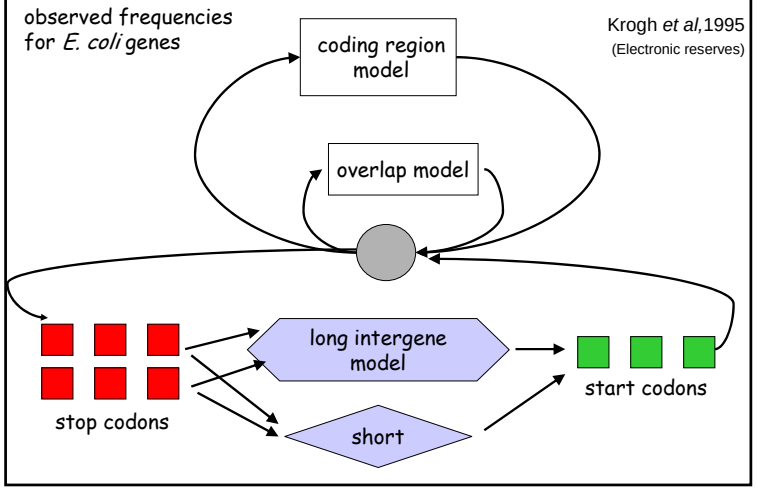




Codon models



Refinements



Parameter estimation

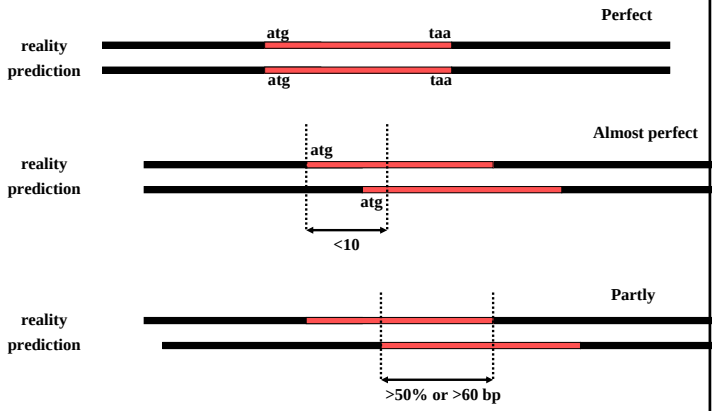
- Data: 429 *E. coli* contigs
- Trained intergenic models with non-coding DNA
- Transitions into codon models set to observed codon frequencies in coding regions

	Training	Test
Contigs	300	129
Base pairs	1,271,528	324,684
Genes	1007	251
Av length	1008	1015

Prediction results on test set

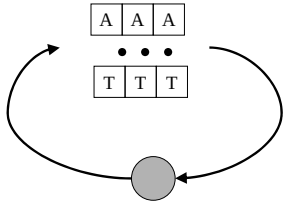
- Exact locations of ~85% of known genes
- Approximate locations of ~10% of known genes
- About half of the false negatives were genes with unusual codon usage.
- Predicted genes: 286
  About 150 were similar to known genes

Performance measures



Outstanding Problems

- Model cannot account for drift in CG content
- Does not take position dependencies into account

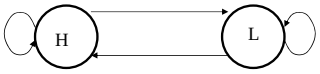


Outstanding Problems

- Model cannot account for drift in CG content
- Does not take position dependencies into account
- Solution:
  - *k*th order Markov chain
  - looks back *k* positions

First-order Markov chain

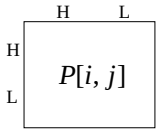
Example: transmembrane region model



H: *hydrophobic*

L: *hydrophilic*

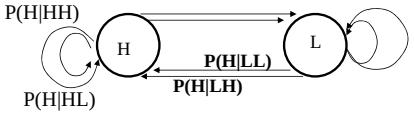
Transition matrix:



$$P(x_t = i \mid x_{t-1} = j)$$

Second-order Markov chain

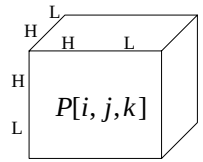
Example: transmembrane region model



H: *hydrophobic*

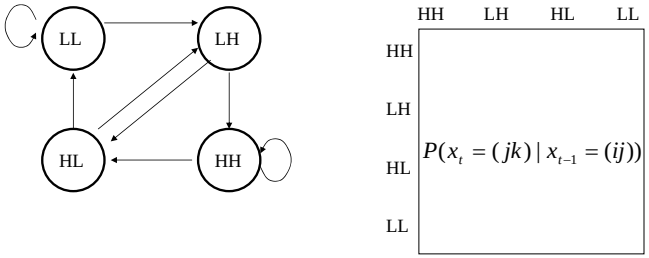
L: *hydrophilic*

Transition matrix:



$$P(x_t = i \mid x_{t-1} = j, x_{t-2} = k)$$

A second-order Markov chain can be expressed as a first order Markov chain with more states and transitions



Glimmer

Salzberg *et al*,1998

- Prokaryotic gene finder
- Finds 98% of all genes in a bacterial genome
- Genome independent
  - Uses all large, non-overlapping ORFs as training data
- *k*th order Markov chain
  - (looks back *k* positions)
- Higher order Markov models require *more* training data