Gene Structure & Gene Finding

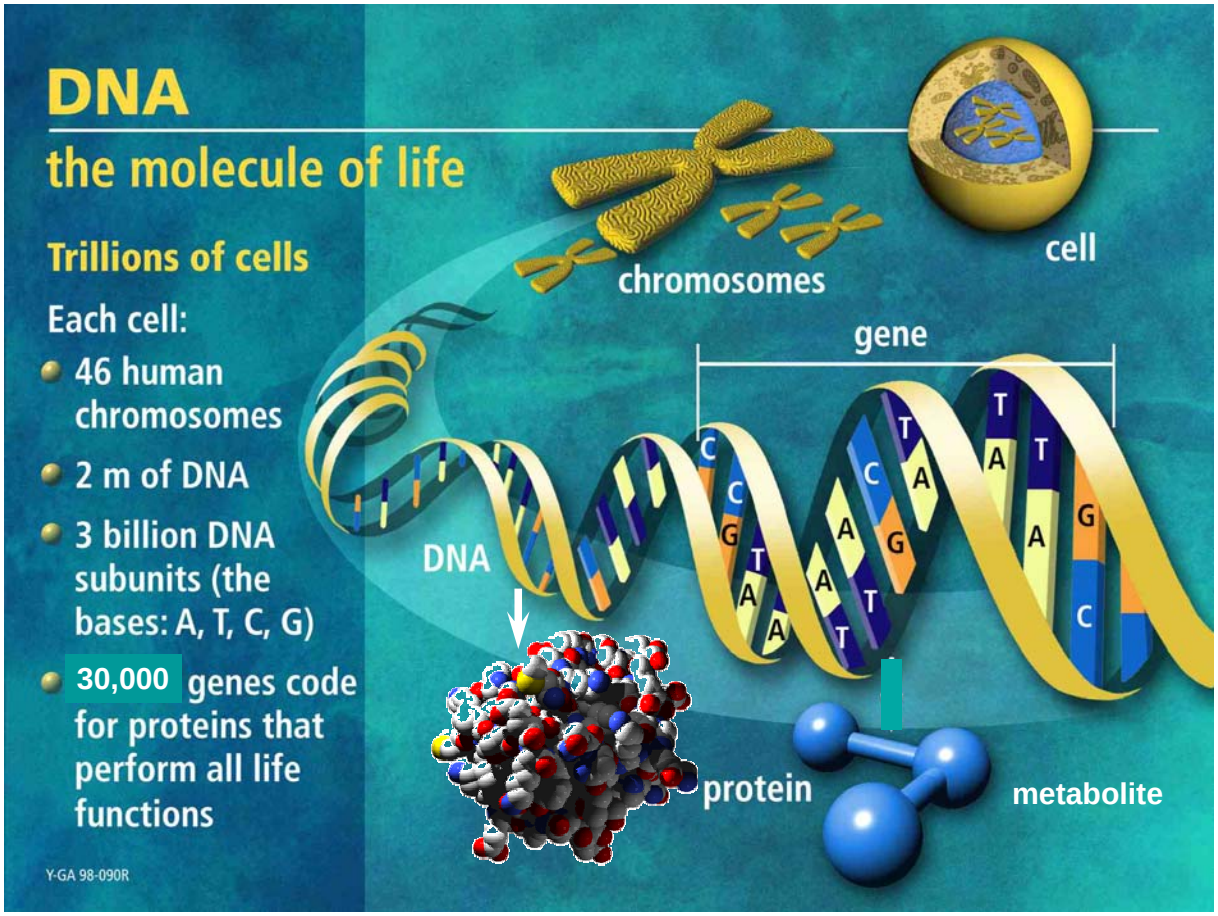
David Wishart

Rm. 3-41 Athabasca Hall

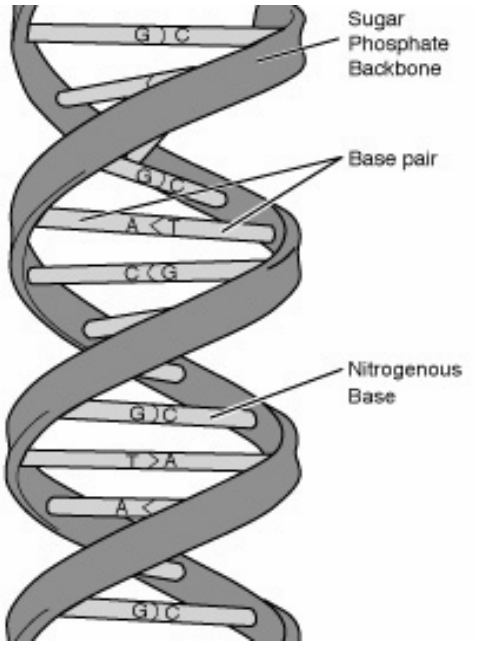
david.wishart@ualberta.ca

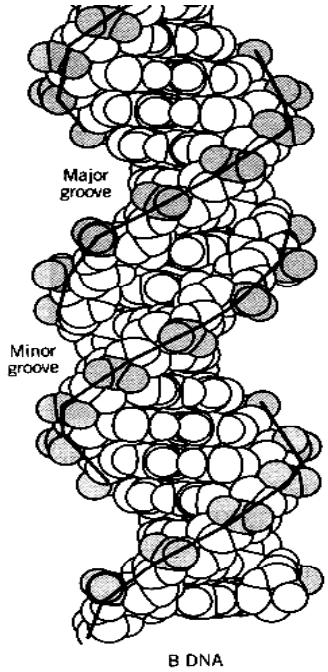
Outline for Next 3 Weeks

- **Genes and Gene Finding (Prokaryotes)**
- **Genes and Gene Finding (Eukaryotes)**
- **Genome and Proteome Annotation**
- **Fundamentals of Transcript Measurement**
- **Microarrays**
- **More Microarrays**



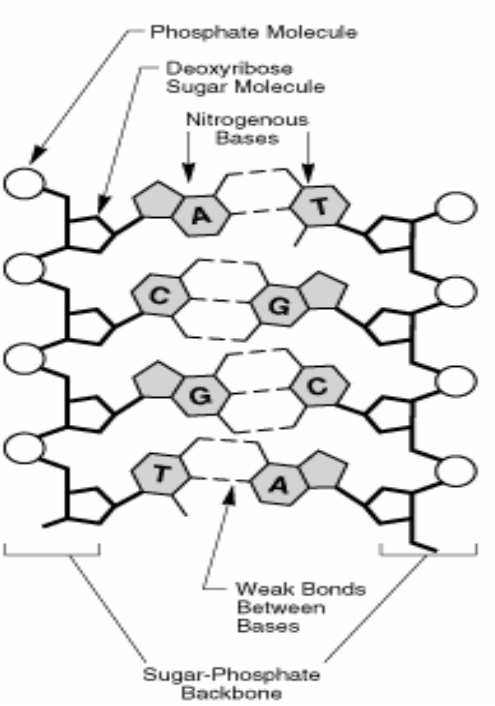
DNA Structure





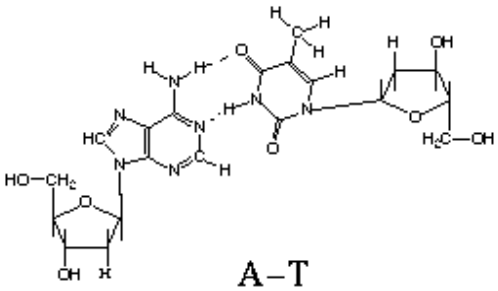
DNA - base pairing

- **Hydrogen Bonds**
- **Base Stacking**
- **Hydrophobic Effect**



Base-pairing (Details)

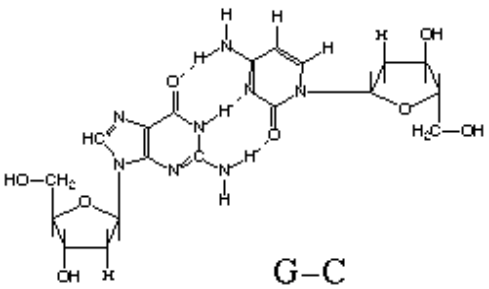
DNA Basepairs



A–T

Adenosine–Thymidine
(Adenine–Thymine)

2 H-bonds



G–C

Guanosine–Cytidine
(Guanine–Cytosine)

3 H-bonds

DNA Sequences

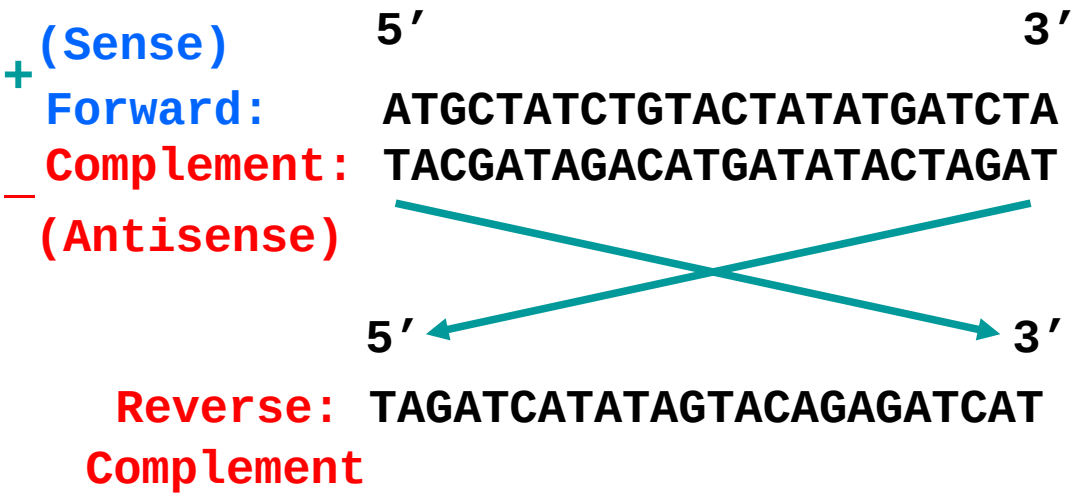
```
          5′                                3′
Single:   ATGCTATCTGTACTATATGATCTA

          5′                                3′
Paired:   ATGCTATCTGTACTATATGATCTA
          TACGATAGACATGATATACTAGAT
```

```
Read this way----->
5′                                  3′
ATGATCGATAGACTGATCGATCGATCGATTAGATCC

TACTAGCTATCTGACTAGCTAGCTAGCTAATCTAGG
3′                                  5′
                                <---Read this way
```

DNA Sequence Nomenclature

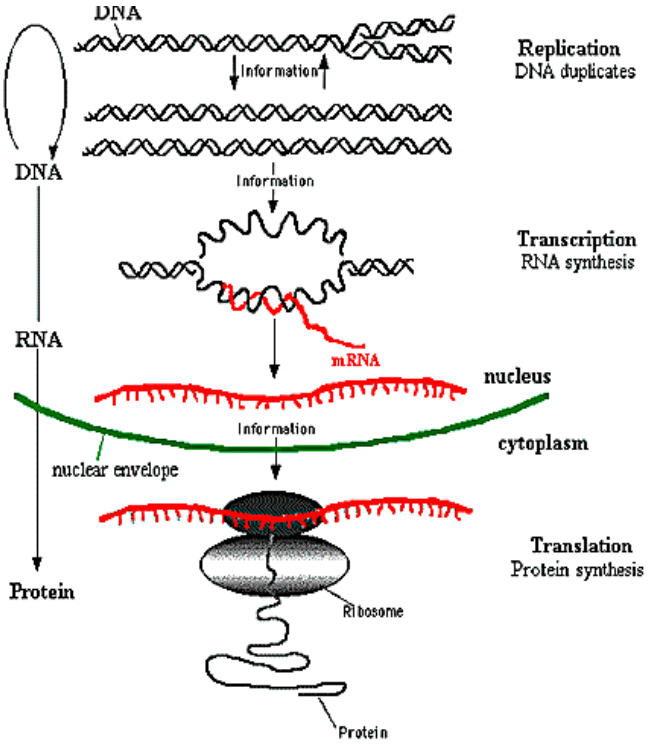


The Fundamental Paradigm

DNA

RNA

Protein



RNA Polymerase

5′ 3′

Forward: ATGCTATCTGTACTATATGATCTA

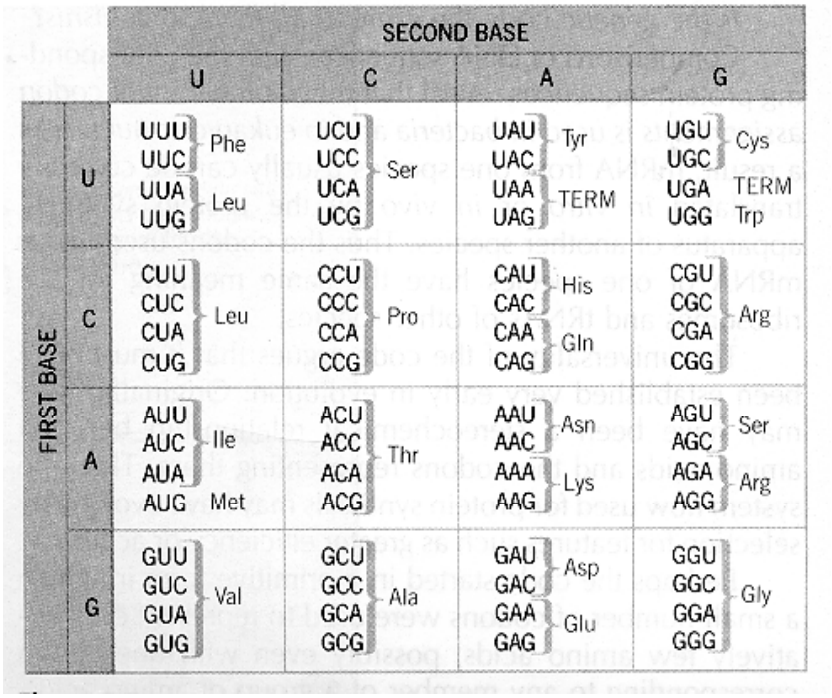
Complement: TACGATAGACATGATATACTAGAT

AUGCUAU

Forward: CTGTACTATATGATCTA

Complement: TACGATAGACATGATATACTAGAT

The Genetic Code



FIRST BASE	SECOND BASE: U	SECOND BASE: C	SECOND BASE: A	SECOND BASE: G
U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr UAA TERM UAG TERM	UGU Cys UGC Cys UGA TERM UGG Trp
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg
A	AUU Ile AUC Ile AUA Ile AUG Met	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly

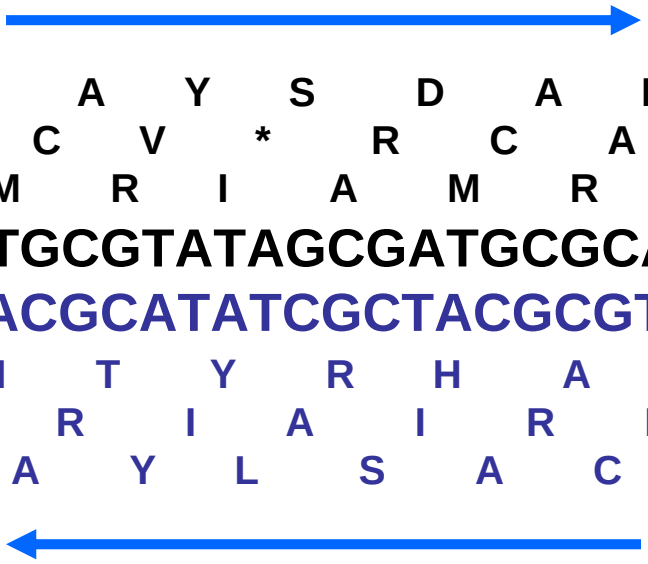
The Genetic Code

TABLE 19-3 Alterations in the Standard Genetic Code in Mitochondria

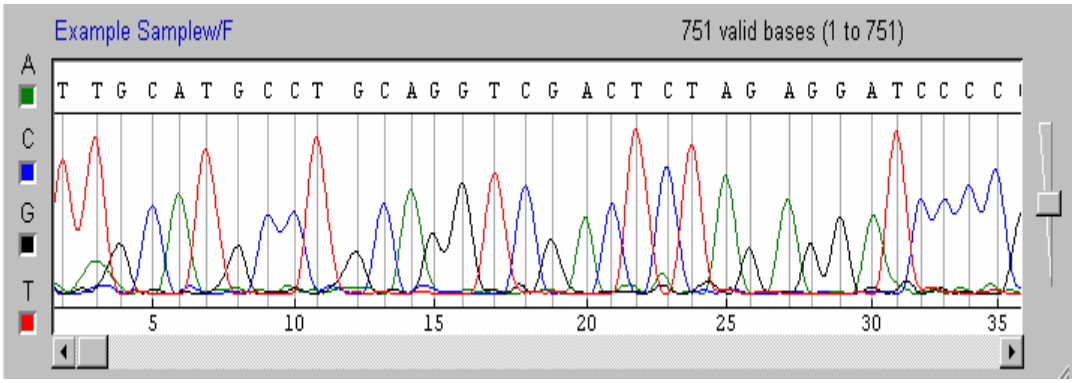
Codon	Standard Code: Nuclear-Encoded Proteins	Mitochondria				
		Mammals	*Drosophila*	*Neurospora*	Yeasts	Plants
UGA	Stop	Trp	Trp	Trp	Trp	Stop
AGA, AGG	Arg	Stop	Ser	Arg	Arg	Arg
AUA	Ile	Met	Met	Ile	Met	Ile
AUU	Ile	Met	Met	Met	Met	Ile
CUU, CUC, CUA, CUG	Leu	Leu	Leu	Leu	Thr	Leu

SOURCE: S. Anderson et al., 1981, *Nature* 290:457; P. Borst, in *International Cell Biology 1980–1981*, H. G. Schweiger, ed., Springer-Verlag, p. 239; C. Breitenberger and U. L. Raj Bhandary, 1985, *Trends Biochem. Sci.* 10:478–483; V. K. Eckenrode and C. S. Levings, 1986, *In Vitro Cell Dev. Biol.* 22:169–176; J. M. Gualber et al., 1989, *Nature* 341:660–662; and P. S. Covello and M. W. Gray, 1989, *Nature* 341:662–666.

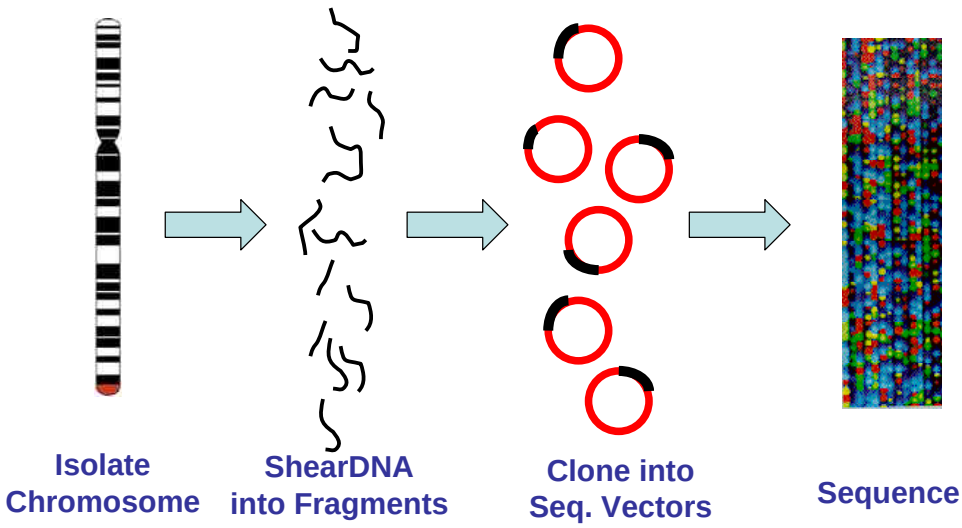
Translating DNA/RNA



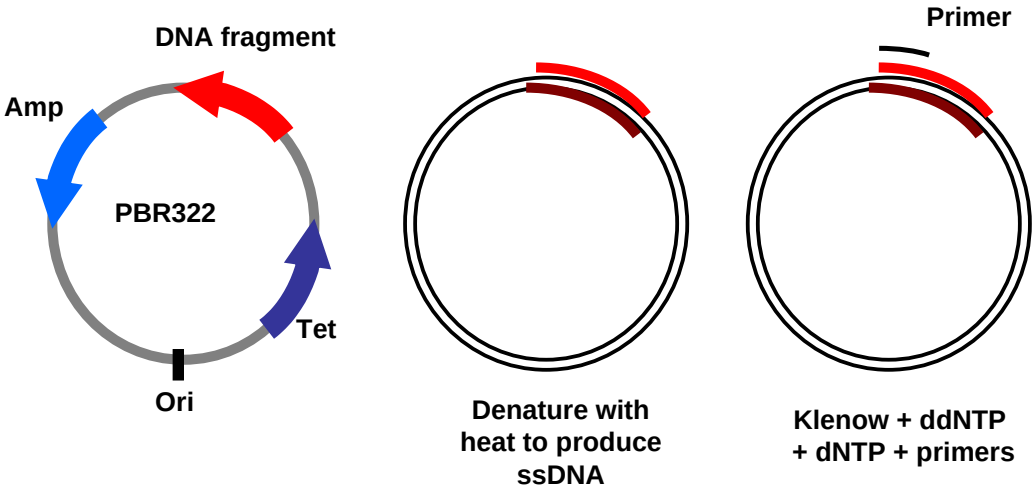
DNA Sequencing



Shotgun Sequencing

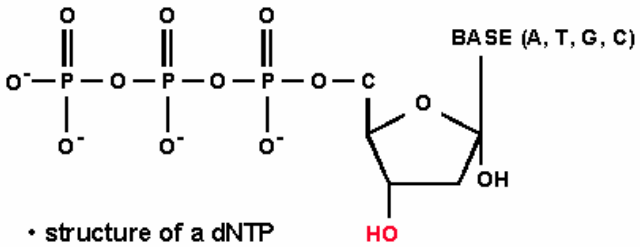


Principles of DNA Sequencing



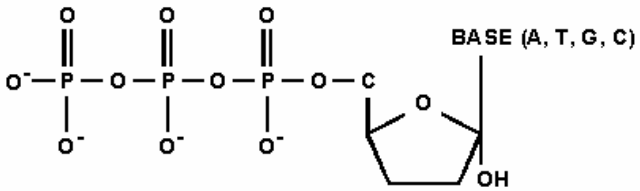
The Secret to Sanger Sequencing

• Structure of the dideoxynucleotide



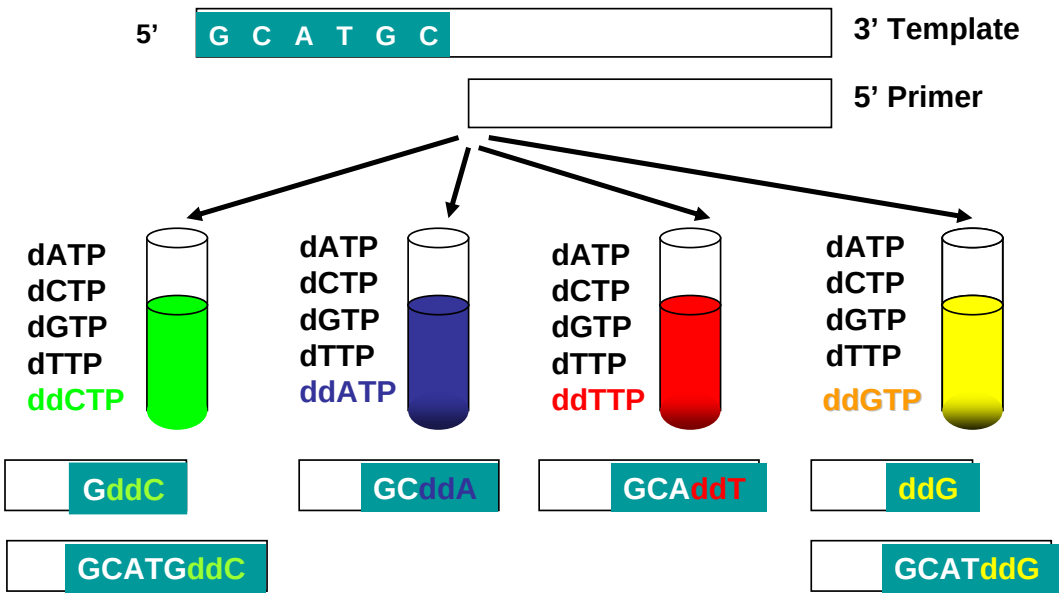
• structure of a dNTP

HO

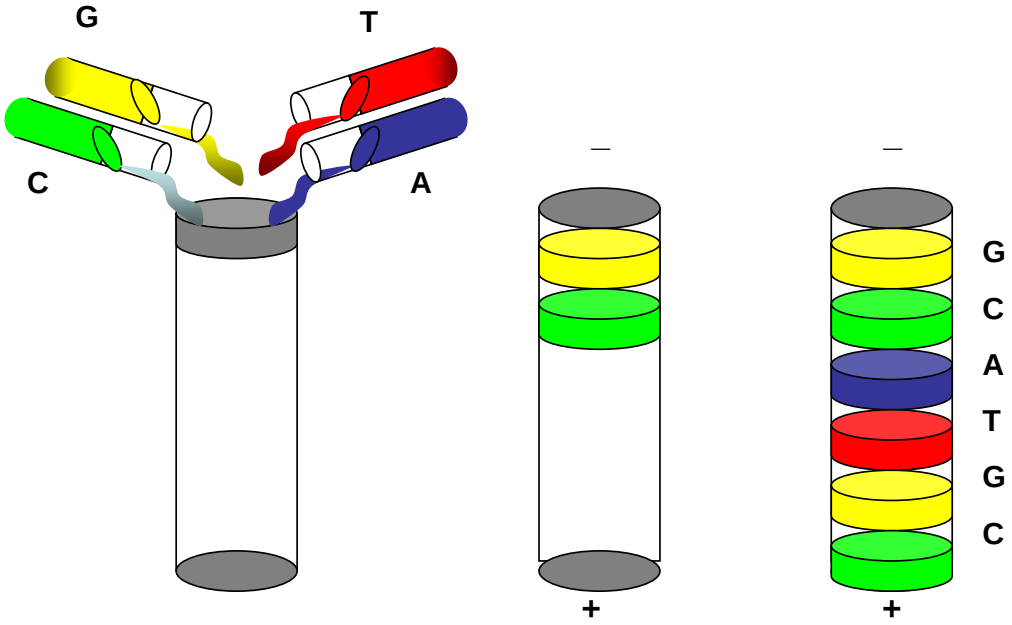


• structure of a ddNTP

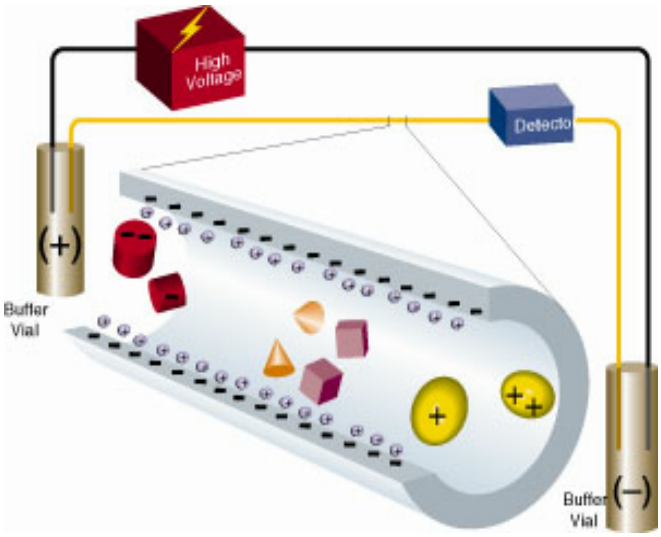
Principles of DNA Sequencing



Principles of DNA Sequencing



Capillary Electrophoresis

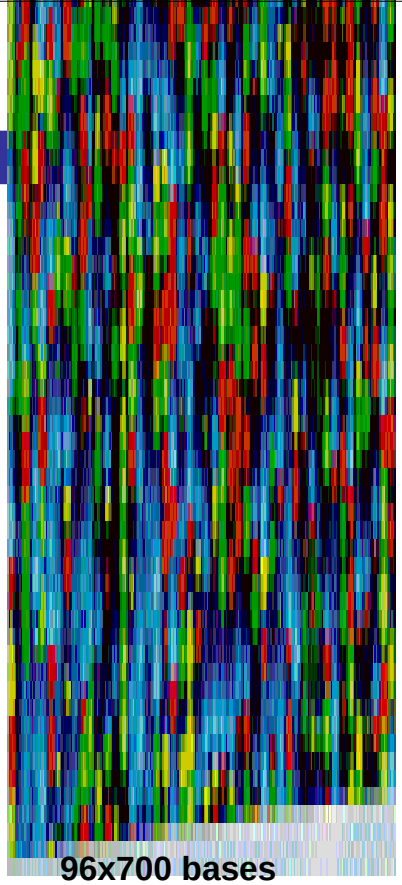


Separation by Electro-osmotic Flow

Multiplexed Fluorescent

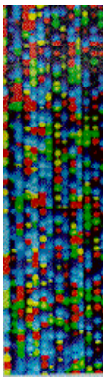


ABI 3700



96x700 bases

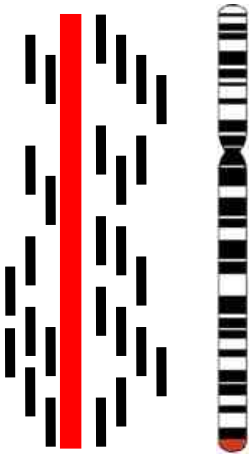
Shotgun Sequencing



Sequence Chromatogram



Send to Computer



Assembled Sequence

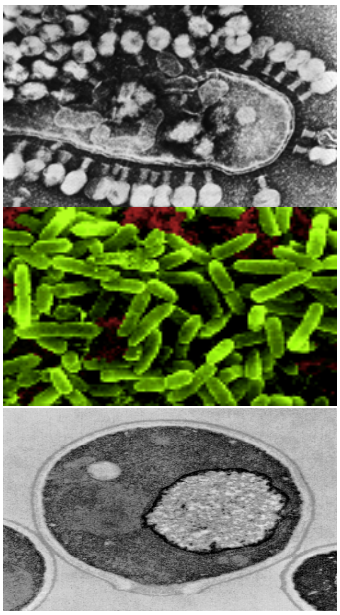
Shotgun Sequencing

- Very efficient process for small-scale (~10 kb) sequencing (preferred method)
- First applied to whole genome sequencing in 1995 (*H. influenzae*)
- Now standard for all prokaryotic genome sequencing projects
- Successfully applied to *D. melanogaster*
- Moderately successful for *H. sapiens*

The Finished Product

GATTACAGATTACAGATTACAGATTACAGATTACAG
ATTACAGATTACAGATTACAGATTACAGATTACAGA
TTACAGATTACAGATTACAGATTACAGATTACAGAT
TACAGATTAGAGATTACAGATTACAGATTACAGATT
ACAGATTACAGATTACAGATTACAGATTACAGATTA
CAGATTACAGATTACAGATTACAGATTACAGATTAC
AGATTACAGATTACAGATTACAGATTACAGATTACA
GATTACAGATTACAGATTACAGATTACAGATTACAG
ATTACAGATTACAGATTACAGATTACAGATTACAGA
TTACAGATTACAGATTACAGATTACAGATTACAGAT

Sequencing Successes

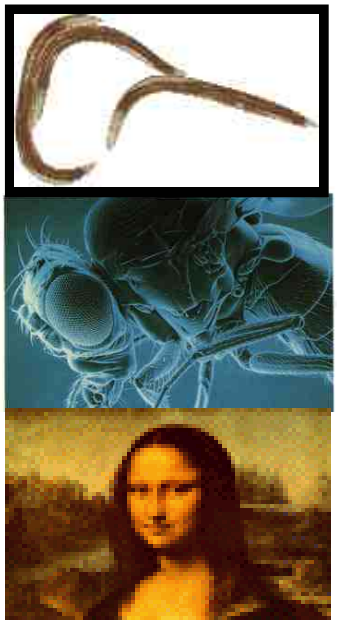


T7 bacteriophage
completed in 1983
39,937 bp, 59 coded proteins

Escherichia coli
completed in 1998
4,639,221 bp, 4293 ORFs

Sacchoromyces cerevisae
completed in 1996
12,069,252 bp, 5800 genes

Sequencing Successes



Caenorhabditis elegans
completed in 1998
95,078,296 bp, 19,099 genes

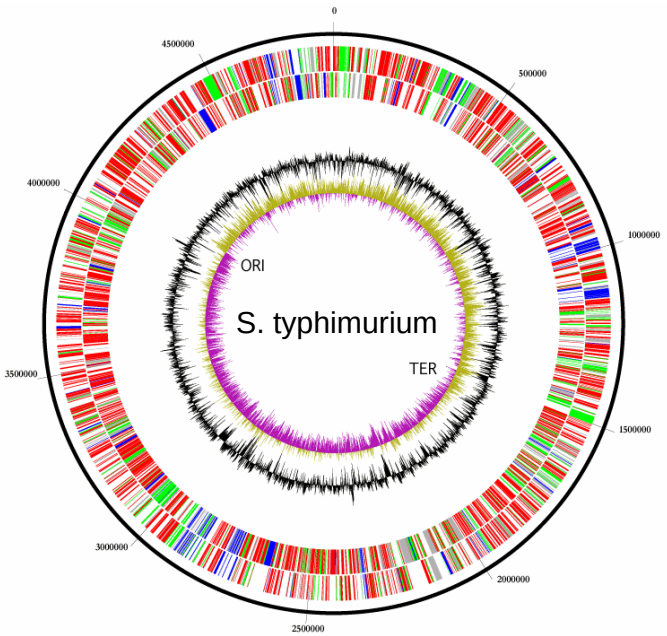
Drosophila melanogaster
completed in 2000
116,117,226 bp, 13,601 genes

Homo sapiens
completed in 2003
3,201,762,515 bp, 31,780 genes

Genomes to Date

- **5 vertebrates** (human, mouse, rat, fugu, zebrafish)
- **2 plants (arabadopsis, rice)**
- **2 insects (fruit fly, mosquito)**
- **2 nematodes (C. elegans, C. briggsae)**
- **1 sea squirt**
- **4 parasites (plasmodium, guillardia)**
- **4 fungi (S. cerevisae, S. pombe)**
- **140 bacteria and archebacteria**
- **1000+ viruses**

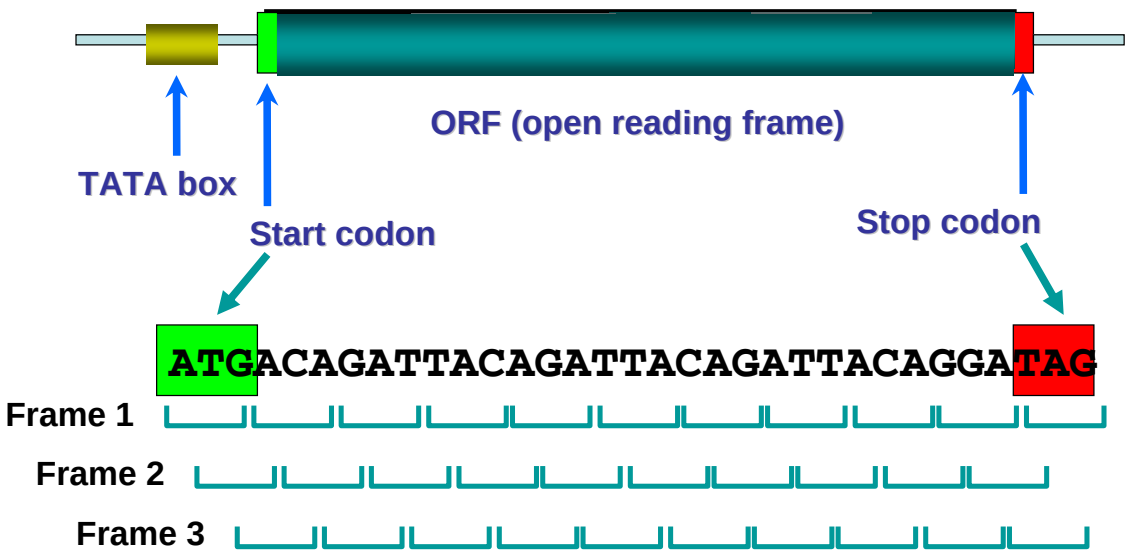
Gene Finding in Prokaryotes



Prokaryotes

- Simple gene structure
- Small genomes (0.5 to 10 million bp)
- No introns (uninterrupted)
- Genes are called Open Reading Frames of “ORFs” (include start & stop codon)
- High coding density (>90%)
- Some genes overlap (nested)
- Some genes are quite short (<60 bp)

Prokaryotic Gene Structure



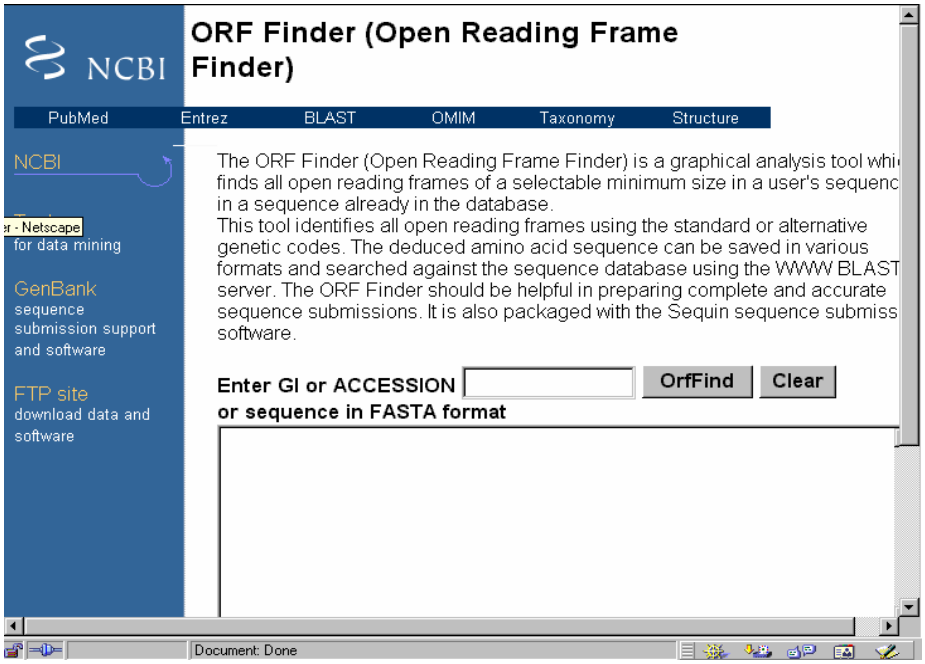
Gene Finding In Prokaryotes

- Scan forward strand until a start codon is found
- Staying in same frame scan in groups of three until a stop codon is found
- If # of codons between start and end is greater than 50, identify as gene and go to last start codon and proceed with step 1
- If # codons between start and end is less than 50, go back to last start codon and go to step 1
- At end of chromosome, repeat process for reverse complement

ORF Finding Tools

- http://www.ncbi.nlm.nih.gov/gorf/gorf.html
- http://alfa.ist.utl.pt/~pedromc/SMS/orf_find.html
- http://www.cbc.umn.edu/diogenes/diogenes.html
- http://www.nih.go.jp/~jun/cgi-bin/frameplot.pl

NCBI ORF Finder



But...

- **Prokaryotic genes are not always so simple to find**
- **When applied to whole genomes, simple ORF finding programs tend to overlook small genes and tend to overpredict the number of long genes**
- **Can we include other genome signals?**
- **Can we account for alternative signals?**

Key Prokaryotic Gene Signals

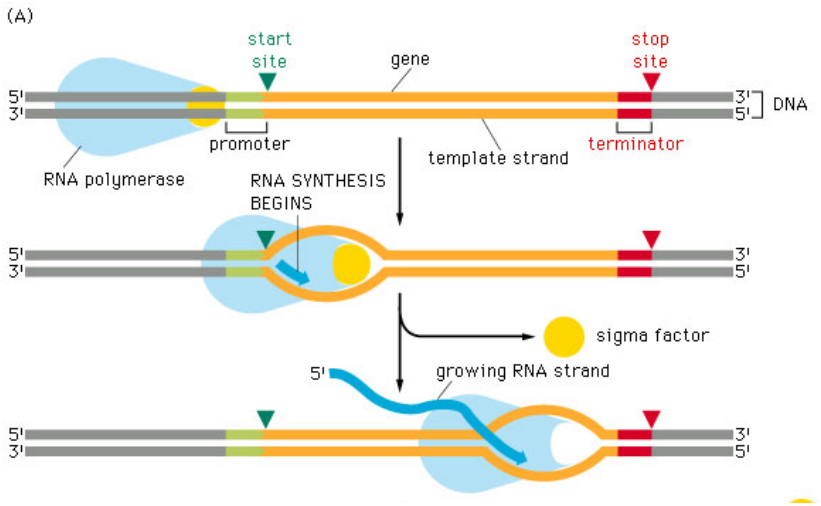
- Alternate start codons
- RNA polymerase promoter site (-10, -35 site or TATA box)
- Shine-Dalgarno sequence (Ribosome binding site-RBS)
- Stem-loop (rho-independent) terminators
- High GC content (CpG islands)

Alternate Start Codons (E. coli)

Class I	ATG	Met
	GTG	Val
	TTG	Leu
Class IIa	CTG	Met
	ATT	Val
	ATA	Leu
	ACG	Thr

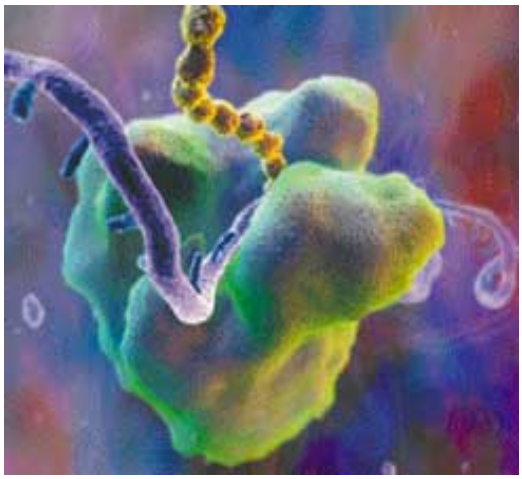
-10, -35 Site (RNA pol Promoter)

-36	-35	-34	-33	-32	….	-13	-12	-11	-10	-9	-8
T	T	G	A	C		T	A	t	A	A	T

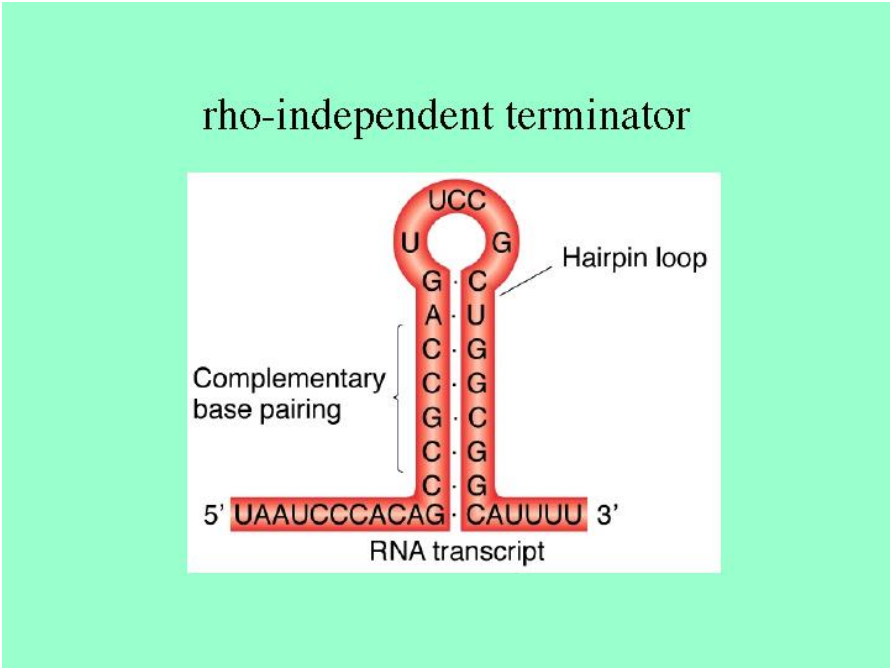


RBS (Shine Dalgarno Seq)

-13	-12	-11	-10	-9	-8	..	-1	0	1	2	3	4
G	G	G	G	G	G		n	A	T	G	n	C



Terminator Stem-loops



Simple Methods to Gene Site Identification



A PSSM

- Use a consensus sequence (CNNTGA)
- Use a regular expression (C[TG]A*)
- Use a custom scoring matrix called a position specific scoring matrix (PSSM) built from multiple sequence alignments

Building a PSSM - Step 1

```
A T T T A G T A T C
G T T C T G T A A C
A T T T T G T A G C
A A G C T G T A A C
C A T T T G T A C A
```

Multiple Alignment



A	3	2	0	0	1	0	0	5	2	1
C	1	0	0	2	0	0	0	0	1	4
G	1	0	1	0	0	5	0	0	1	0
T	0	3	4	3	4	0	5	0	1	0

Table of Occurrences

Building a PSSM - Step 2

A	3	2	0	0	1	0	0	5	2	1
C	1	0	0	2	0	0	0	0	1	4
G	1	0	1	0	0	5	0	0	1	0
T	0	3	4	3	4	0	5	0	1	0

Table of Occurrences



A	.6	.4	0	0	.2	0	0	1	.4	.2
C	.2	0	0	.4	0	0	0	0	.2	.8
G	.2	0	.2	0	0	1	0	0	.2	0
T	0	.6	.8	.6	.8	0	1	0	.2	0

PSSM with no pseudocounts

Pseudocounts

- Method to account for small sample size of multi-sequence alignment
- Gets around problem of having "0" score in PSSM or profile
- Defined by a correction factor "B" which reflects overall composition of sequences under consideration
- $B = \sqrt{N}$ or B = 0.1 which falls off with N where N = # sequences

Pseudocounts

- $Score(X_i) = (q_x + p_x)/(N + B)$

- q = observed counts of residue X at pos. i
- p = pseudocounts of X = B*frequency(X)
- N = total number of sequences in MSA
- B = number of pseudocounts (assume $\sqrt{N}$)

$$Score(A_1) = (3 + \sqrt{5}(0.32))/(5 + \sqrt{5}) = 0.51$$

Including Pseudocounts - Step 2

A	3	2	0	0	1	0	0	5	2	1
C	1	0	0	2	0	0	0	0	1	4
G	1	0	1	0	0	5	0	0	1	0
T	0	3	4	3	4	0	5	0	1	0

Table of Occurrences



A	.51	.38	.09	.09	.24	.09	.09	.79	.38	.24
C	.19	.06	.06	.33	.06	.06	.06	.06	.19	.61
G	.19	.06	.19	.06	.06	.75	.06	.06	.19	.06
T	.09	.51	.65	.51	.65	.09	.79	.09	.24	.09

PSSM with pseudocounts

Calculating Log-odds - Step 3

A	.51	.38	.09	.09	.24	.09	.09	.79	.38	.24
C	.19	.06	.06	.33	.06	.06	.06	.06	.19	.61
G	.19	.06	.19	.06	.06	.75	.06	.06	.19	.06
T	.09	.51	.65	.51	.65	.09	.79	.09	.24	.09

PSSM with pseudocounts



$-\text{Log}_{10}$

A	0.2	0.4	1.1	1.1	0.7	1.1	1.1	0.1	0.4	0.7
C	0.7	1.2	1.2	0.4	1.2	1.2	1.2	1.2	0.7	0.1
G	0.7	1.2	0.7	1.2	1.2	0.1	1.2	1.2	0.7	1.2
T	1.1	0.2	0.1	0.2	0.1	1.1	0.1	1.1	0.7	1.1

Log-odds PSSM

Scoring a Sequence - Step 4

A	0.2	0.4	1.1	1.1	0.7	1.1	1.1	0.1	0.4	0.7
C	0.7	1.2	1.2	0.4	1.2	1.2	1.2	1.2	0.7	0.1
G	0.7	1.2	0.7	1.2	1.2	0.1	1.2	1.2	0.7	1.2
T	1.1	0.2	0.1	0.2	0.1	1.1	0.1	1.1	0.7	1.1

Log-odds PSSM

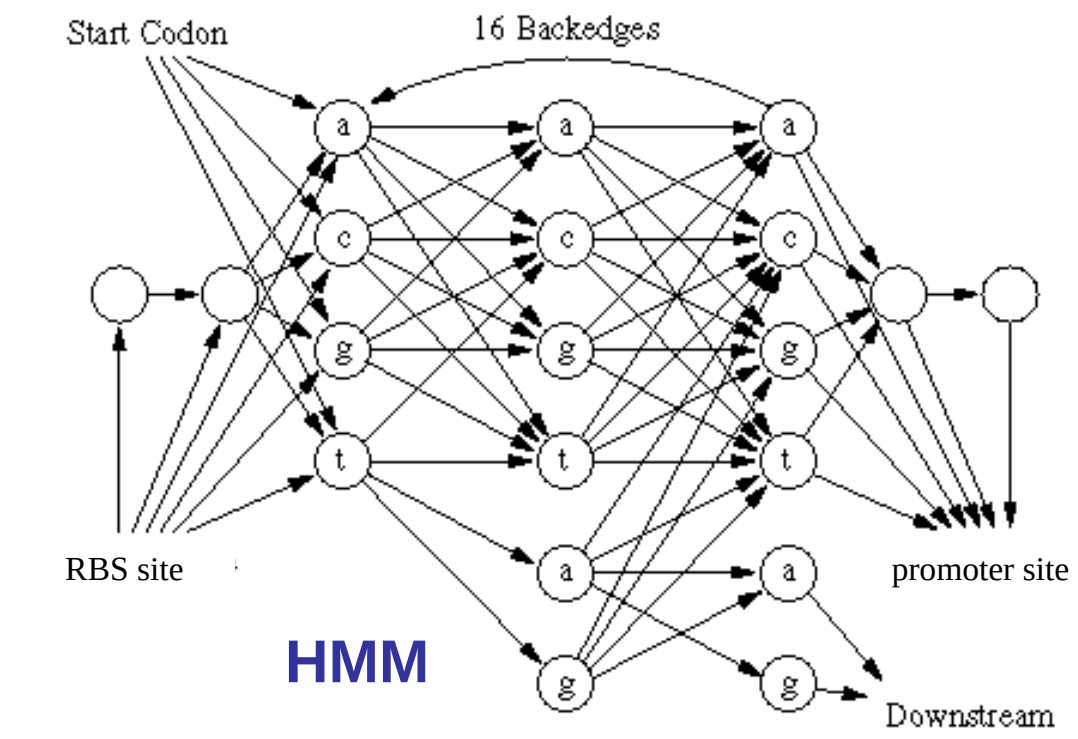
A T T T A G T A T C

Score = 2.5

(Lowest score wins)

A	**0.2**	0.4	1.1	1.1	**0.7**	1.1	1.1	**0.1**	0.4	0.7
C	0.7	1.2	1.2	0.4	1.2	1.2	1.2	1.2	0.7	**0.1**
G	0.7	1.2	0.7	1.2	1.2	**0.1**	1.2	1.2	0.7	1.2
T	1.1	**0.2**	**0.1**	**0.2**	0.1	1.1	**0.1**	1.1	**0.7**	1.1

More Sophisticated Methods



HMM

More Sophisticated Methods

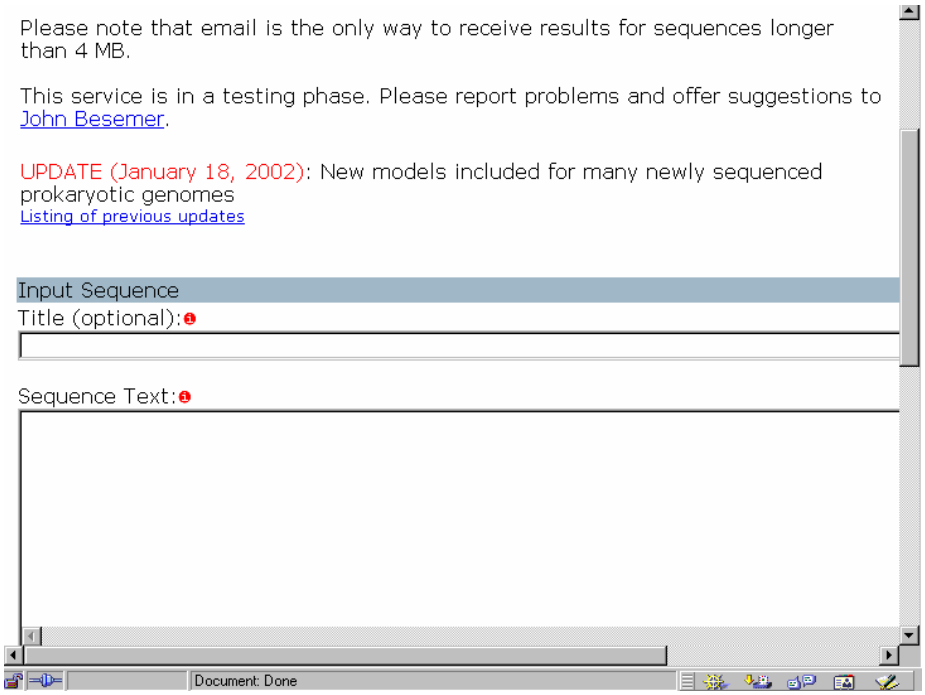
- **GLIMMER**
  - **http://www.tigr.org/software/glimmer/**
  - **Uses interpolated markov models (IMM)**
  - **Requires training of sample genes**
  - **Takes about 1 minute/genome**
- **GeneMark.hmm**
  - **http://opal.biology.gatech.edu/GeneMark/gmhmm2_prok.cgi**
  - **Available as a web server**
  - **Uses hidden markov models (HMM)**

Glimmer Performance

Glimmer 2.0's Accuracy

Organism	Genes annotated	Annotated genes found	% found
H. influenzae	1738	1720	99.0
M. genitalium	483	480	99.4
M. jannaschii	1727	1721	99.7
H. pylori	1590	1550	97.5
E. coli	4269	4158	97.4
B. subtilis	4100	4030	98.3
A. fulgidis	2437	2404	98.6
B. burgdorferi	853	843	99.3
T. pallidum	1039	1014	97.6
T. maritima	1877	1854	98.8

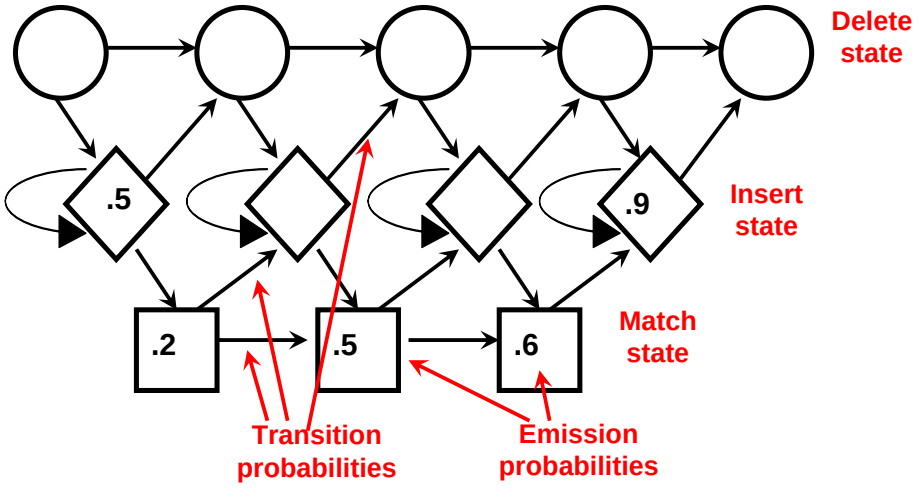
Genemark.hmm



Hidden Markov Models

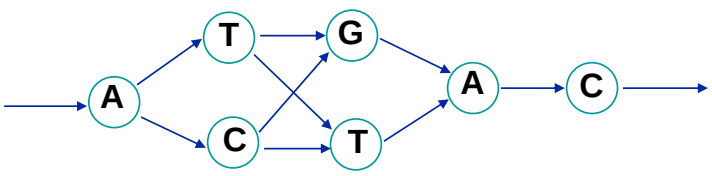
- **Markov Model is a chain of events or states**
- **Each state has a set of emission probabilities for occupying that state**
- **MSA creates a Markov model of emission and transition probabilities**
- **Typically have a "Topology" which assumes a sequence of events is a multiplicative product of individual probabilities (independent, 1st order)**

Hidden Markov Topology



Hidden Markov Models

States -- well defined conditions
Edges -- transitions between the states



ATGAC
ATTAC
ACGAC
ACTAC

Each transition is assigned a probability.

Probability of the sequence:
single path with the highest probability --- *Viterbi* path
sum of the probabilities over all paths -- *Baum-Welch* method

Making a Markov Model

```
A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A - - - A T C
A C C G - - A T C
```

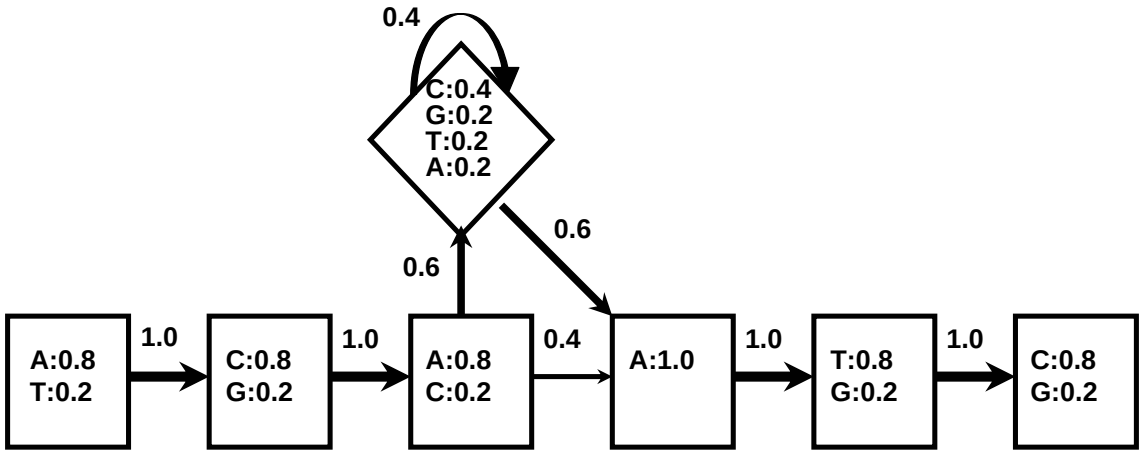
[AT][CG][AC][ACGT-](3)A[TG][GC]

~3600 possible valid sequences

Making a Markov Model

			Δ=.4	Δ=.6	Δ=.6			
	p(C)=.8 p(G)=.2		p(A)=.2 p(C)=.4 p(T)=.2 p(G)=.2			p(T)=.8 p(G)=.2		
A	C	A	-	-	-	A	T	G
T	C	A	A	C	T	A	T	C
A	C	A	C	-	-	A	G	C
A	G	A	-	-	-	A	T	C
A	C	C	G	-	-	A	T	C
p(A)=.8 p(T)=.2		p(A)=.8 p(C)=.2				p(A)=1	p(C)=.8 p(G)=.2	

Making a Markov Model



P(ACAC--ATC)=0.8x1.0x0.8x1.0x0.8x1.0x0.6x0.4
x0.6x1.0x1.0x0.8x1.0x0.8 = 0.0047

Log-Odds (LOD)

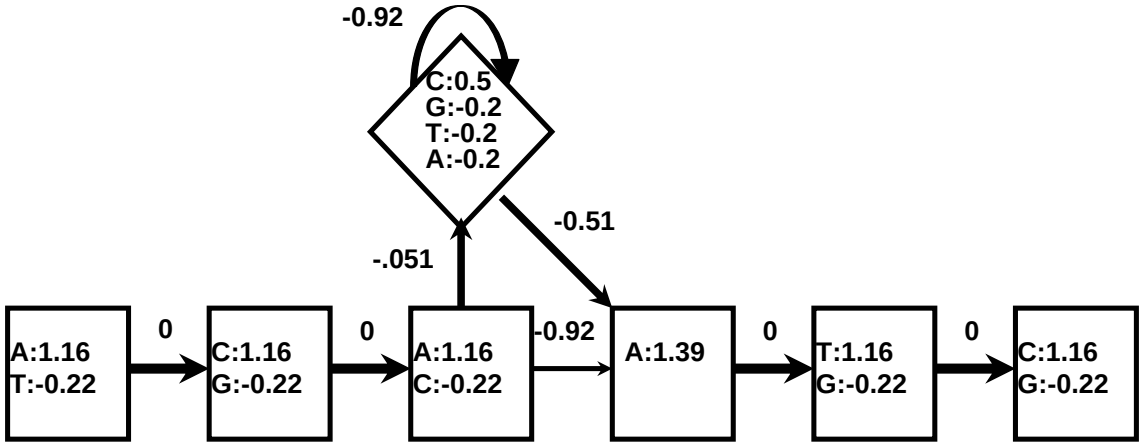
Def'n - LOD is the logarithm of the probability of an event divided by the probability of a null model

For DNA: $LOD(S) = \log \frac{P(S)}{0.25^L} = \log P(S) - L\log 0.25$

For protein: $LOD(S) = \log \frac{P(S)}{0.05^L} = \log P(S) - L\log 0.05$

S = sequence, L = length

Making a LOD Markov Model



LOD(ACAC--ATC)=1.16+0+1.16+0+1.16-0.51+
0.5-0.51+1.39+0+1.16+0+1.16 = 6.64

Other Sequences...

- P(ACA---ATG) = 0.0033 (LOD = 4.9)
- P(TCAACTATC) = 0.000075 (LOD = 3.0)
- P(ACAC--AGC) = 0.0012 (LOD = 5.3)
- P(AGA---ATC) = 0.0033 (LOD = 4.9)
- P(ACCG--ATC) = 0.00059 (LOD = 4.6)
- P(TGCT--AGG) = 0.000023 (LOD = -0.97) Worst
- P(ACAC--ATG) = 0.0047 (LOD = 6.7) Best

HMM Issues

- How to find the “optimal sequence” or score a new sequence?
- Answer: Use Dynamic Programming (called the Viterbi algorithm) to find the optimal path
- How to deal with sparse data?
- Answer: Use Pseudocounts (i.e. add fake data that reflects natural substitution patterns or known frequencies)

HMM’s in Gene Prediction

- Can be used to make a 1st order position specific profile or weight matrix for splice sites, start sites or coding regions
- Mostly used in creating “higher order” Markov Models where dinucleotide (2nd order), trinucleotide (3rd order) or pentanucleotide (5th order) probabilities are used to recognize coding regions

HMM Order & Conditional Probability

Order

1st P(ACTGTC) = p(A) x p(C) x p(T) x p(G) x p(T) ...

2nd P(ACTGTC) = p(A) x p(C|A) x p(T|C) x p(G|T) …

3rd P(ACTGCG) = p(A) x p(C|A) x p(T|AC) x p(G|CT)...



P(T|AC) = #(ACT)/#ACT+#ACA+#ACG+#ACC

Probability of T given AC

Bottom Line...

- Gene finding in prokaryotes is now a "solved" problem
- Accuracy of the best methods approaches 99%
- Gene predictions should always be compared against a BLAST search to ensure accuracy and to catch possible sequencing errors