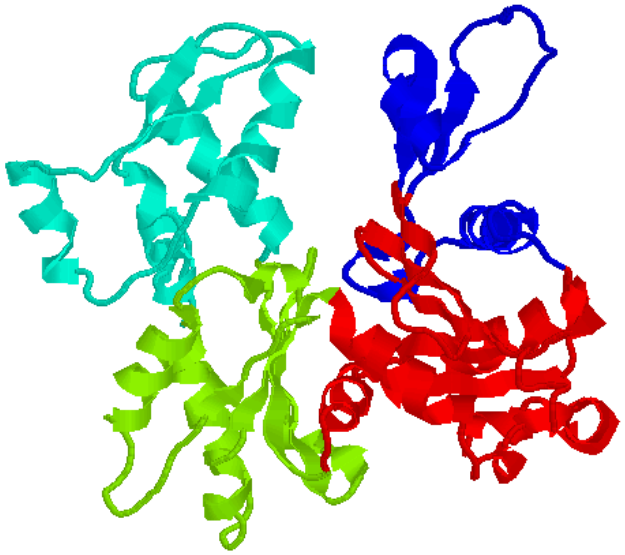
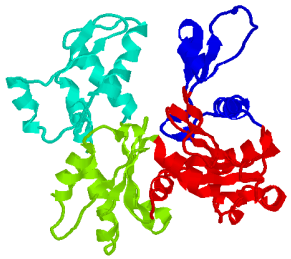
Computational Gene Finding



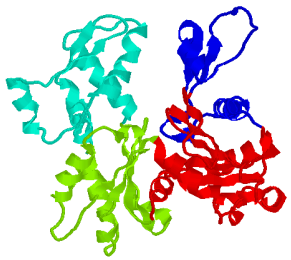
Dong Xu

Digital Biology Laboratory
Computer Science Department
Christopher S. Life Sciences Center
University of Missouri, Columbia
E-mail: xudong@missouri.edu
http://digbio.missouri.edu



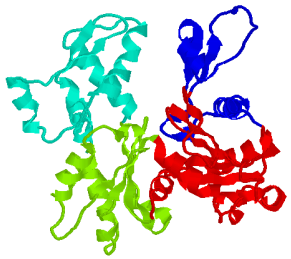
Lecture Outline

- Protein-encoding genes and gene structures
- Computational models for coding regions
- Computational models for coding-region boundaries
- Markov chain model for coding regions



What Is a Gene?

Definition: A gene is the nucleotide sequence that stores the information which specifies the order of the monomers in a final **functional polypeptide** or RNA molecule, or set of closely related isoforms (Epp CD, Nature, 389: 537).



Gene and Disease

Monogenic Diseases

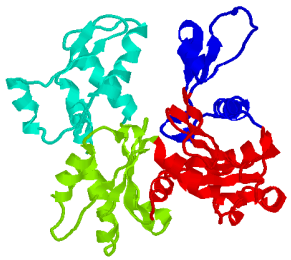
- Cystic fibrosis
- Huntington's disease
- Haemophilia
- Phenylketonuria

Common Diseases

- Alzheimer disease
- Adult onset diabetes
- Cancer
- Cardiovascular disease
- Depression

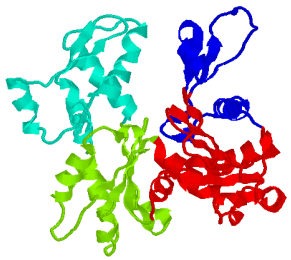
Infections

- Influenza
- Hepatitis
- AIDS



Genetic Code

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



Reading Frame

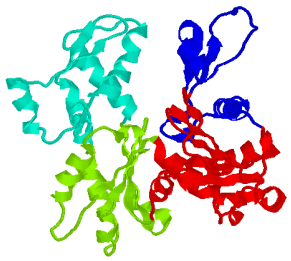
- Reading (or translation) frame: each DNA segment has six possible reading frames

Forward strand: ATGGCTTACGCTTGA

Reading frame #1	Reading frame #2	Reading frame #3
ATG	TGG	GGC
GCT	CTT	TTA
TAC	ACG	CGC
GCT	CTT	TTG
TGC	GA.	A..

Reverse strand: TCAAGCGTAAGCCAT

Reading frame #4	Reading frame #5	Reading frame #6
TCA	CAA	AAG
AGC	GCG	CGT
GTA	TAA	AAG
AGC	GCC	CCA
CAT	AT.	T..



Prokaryotic Gene Structure

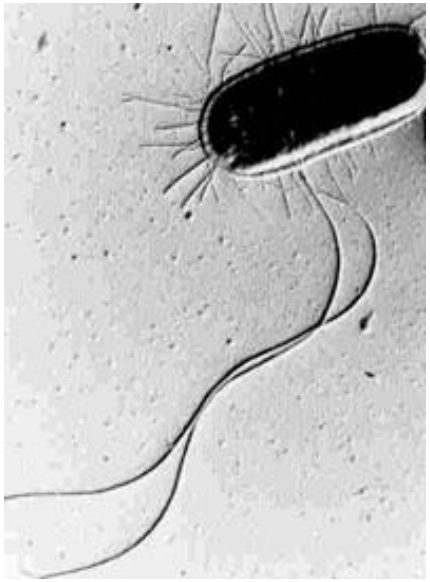
Coding region of Open Reading Frame

Promoter region (maybe)

Ribosome binding site (maybe)

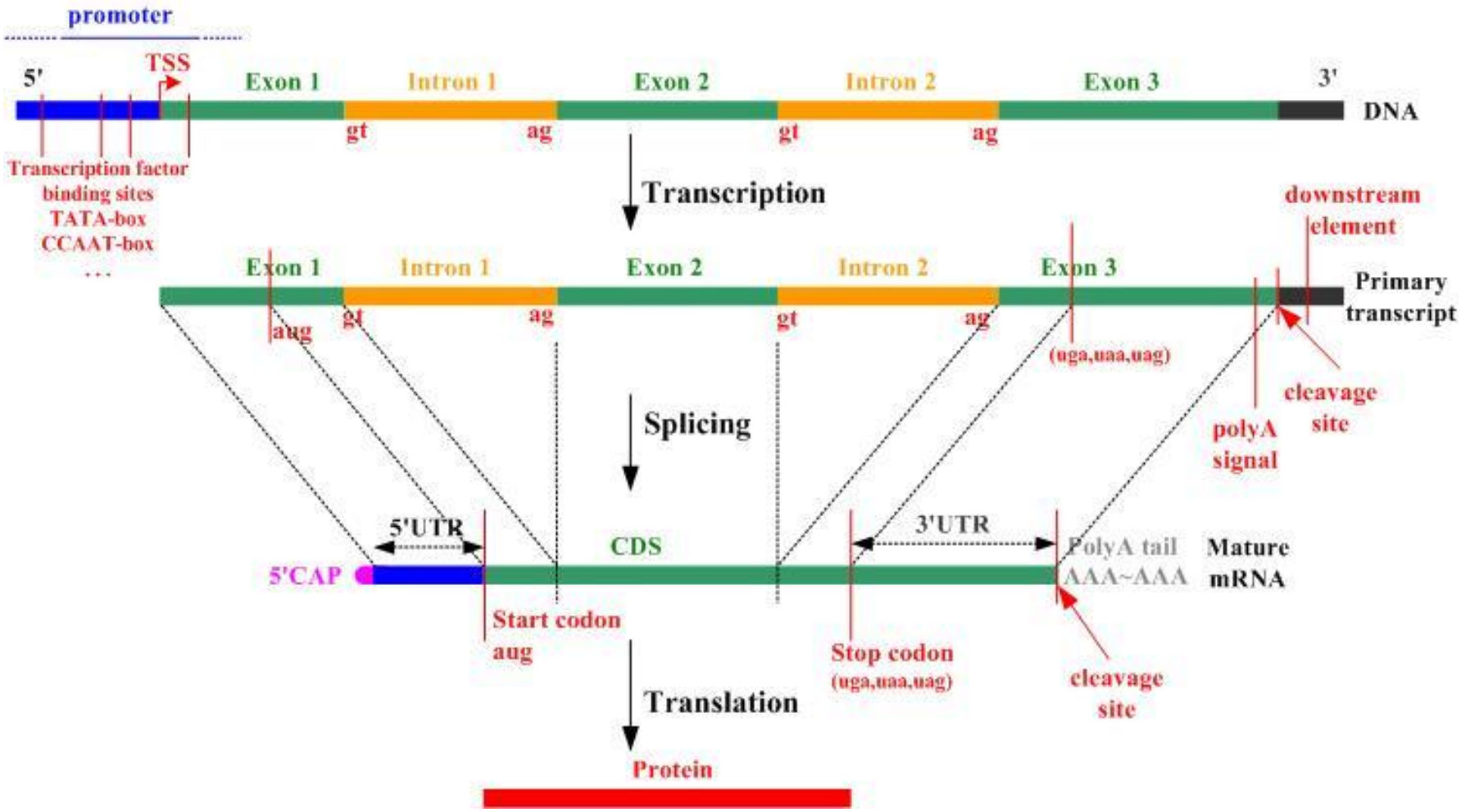
Termination sequence (maybe)

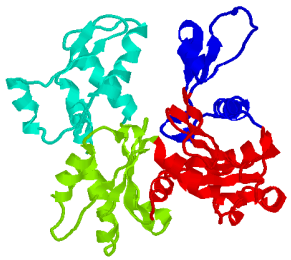
Start codon / Stop Codon



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

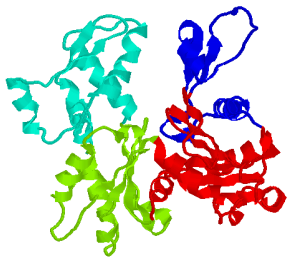
Eukaryotic Gene Structure





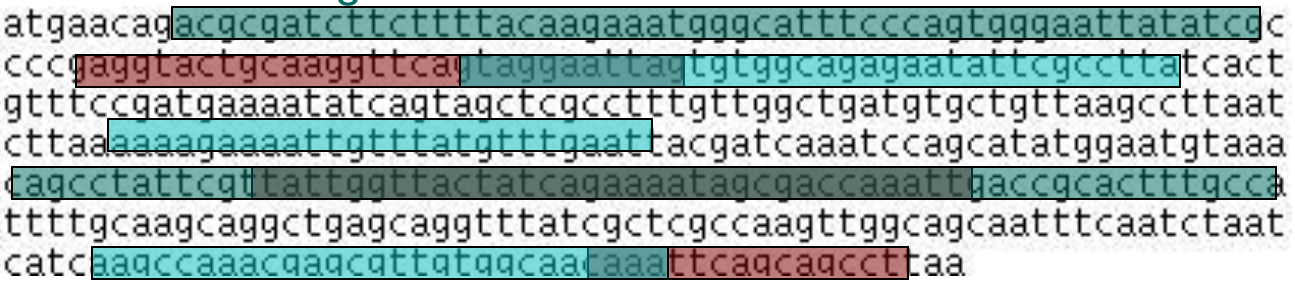
Gene Structure Rules

- Each coding region (exon) has a fixed translation frame (no gaps allowed)
- All exons of a gene are on the same strand
- Neighboring exons of a gene can have different reading frames

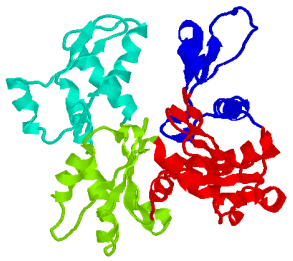


Computational Gene Finding

- The Problem: Given a stretch of DNA sequence, find all coding regions and construct gene structures from identified exons if needed

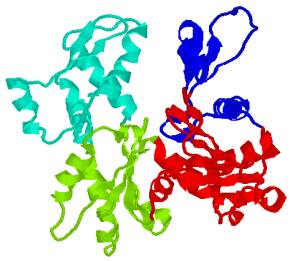


- A gene finding problem can be decomposed into two problems:
  - identification of coding potential of a region in a particular frame
  - identification of boundaries between coding and non-coding regions



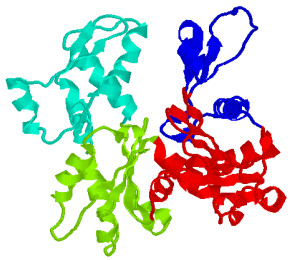
Repetitive Sequence

- Definition
  - DNA sequences that made up of copies of the same or nearly the same nucleotide sequence
  - Present in many copies per chromosome set



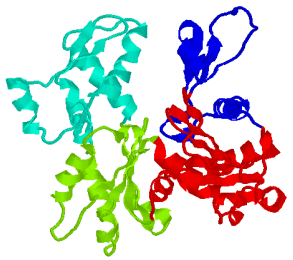
Repeat Filtering

- RepeatMasker
  - Uses precompiled representative sequence libraries to find homologous copies of known repeat families
  - Use Blast
  - http://www.repeatmasker.org/



Gene Finding Tools

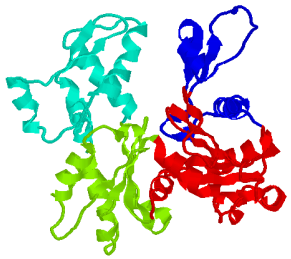
- Genscan (http://genes.mit.edu/GENSCAN.html)
- GeneMarkHMM (http://opal.biology.gatech.edu/GeneMark/)
- GRAIL (http://compbio.ornl.gov/Grail-1.3/)
- Genie (http://www.fruitfly.org/seq_tools/genie.html)
- Glimmer (http://www.tigr.org/softlab/glimmer)



Testing Finding Tools

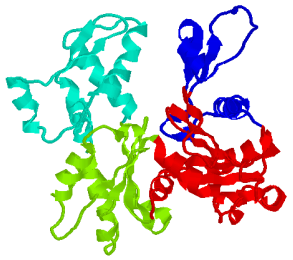
- Access Genscan (http://genes.mit.edu/GENSCAN.html)
- Use a sequence at

http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=8077108



Lecture Outline

- Protein-encoding genes and gene structures
- Computational models for coding regions
- Computational models for coding-region boundaries
- Markov chain model for coding regions



Coding Signal Detection (1)

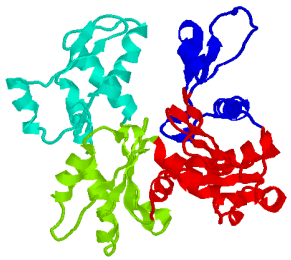
- Frequency distribution of dimers in protein sequence (shewanella)

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

The average frequency is 5%

Some amino acids prefer to be next to each other

Some other amino acids prefer to be not next to each other



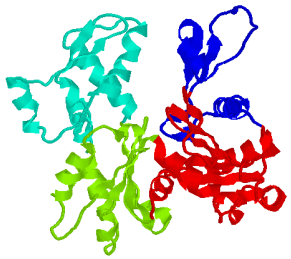
Coding Signal Detection (2)

- Dimer bias (or preference) could imply di-codon (6-mers like AAA TTT) bias in coding versus non-coding regions

- 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 dicocon occurrences
  - frequency of dicodon X (e.g, AAAAAA) in noncoding region, total number of occurrences of X divided by total number of dicodon 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?



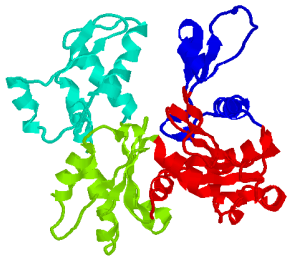
Coding Signal Detection (3)

- 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



Coding Signal Detection (4)

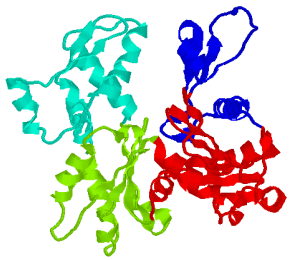
- Dicodon frequencies in coding *versus* non-coding are genome-dependent

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

shewanella

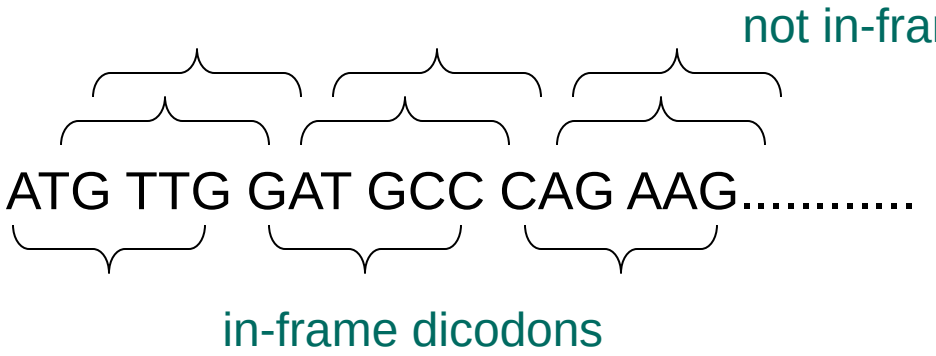
bovine

Name	ala	arg	asn	asp	cys	glu	gln	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
ala	11.4	5.9	3.1	4.5	1.9	5.8	3.6	7.7	1.9	4.3	9.7	4.3	2.1	3.7	6.4	6.4	5.6	1.1	2.6	6.8
arg	8.5	7.7	4	4.6	2.3	5.9	3.8	7.6	2.5	4.4	9.2	5	1.7	4	5.3	6.3	5	1.5	3.4	6.5
asn	6.3	4.9	4.9	4.4	2.1	5.3	4.1	6.9	2.2	5.6	9.7	5.4	2.1	4.1	5.9	7.3	5.3	1.9	4.6	6.2
asp	7.4	4.9	3.5	5.4	2.4	6.6	3.4	7.4	2.1	5.4	9.5	4.7	2	4.4	5.4	6.8	5.7	1.6	4	6.4
cys	6.9	5.9	4	5.4	2.7	5.6	4.9	7.1	3	4.4	8.8	5.4	1.6	3.5	6.8	7.4	5.7	1.4	2.7	5.7
glu	7.8	5.3	4.3	6.4	1.9	9.7	3.7	6.8	2	5.1	8.2	6.2	2.2	3.3	4.8	5.3	5.4	1.2	3.2	6.2
gln	7.9	5.6	4.2	5	2	6.6	5.1	6.9	2.1	4.7	9.3	5.7	2	3.3	5.9	5.7	6.1	1.6	3.3	6.2
gly	7.9	5.8	3.9	5	1.9	6.2	3.5	8	1.8	4.7	8.7	5.2	1.7	3.7	6.9	7.4	5.8	1.4	3.2	6.2
his	6	5.8	4.3	3.5	2.9	5.1	4.1	6.3	3.2	4.5	10.6	4.8	1.6	4.5	6.7	6.6	6.1	1.7	3.9	6.9
ile	6.2	4.9	4.9	4.7	2.4	5.3	4.6	5.8	2.2	6	9.9	5.3	2.1	4.1	5.3	7.7	6.9	1.2	3.7	6
leu	7.7	5.6	4.1	4.7	2.1	5.8	4.5	6.8	2.1	4.6	11	5.4	1.9	3.7	5.7	7	5.5	1.2	3.1	6.4
lys	6.3	5.2	4.8	5.2	2.1	7.2	3.7	6.7	2.2	6	8.5	7.5	2	3.5	4.8	6.1	5.8	1.6	3.5	6.3
met	9.3	5.3	4.1	5.9	1.6	6.1	3.5	6.4	1.6	4.1	9.6	6.6	2.6	4	5.1	6.9	5.5	1	3.2	6.6
phe	6	5.4	4.5	5.2	2.5	5.5	4.1	6.5	2.3	5.3	10.2	5.2	1.8	4.1	5.3	7.8	5.8	1.4	3.9	6.2
pro	8.5	5.4	3.1	5.1	1.9	6.7	3.9	9.5	1.9	4.3	7.7	4.3	1.7	3.3	8.7	6.9	5.7	1.4	2.8	6.4
ser	6.7	5.4	3.8	4.9	2.3	5.4	4	7.9	2.1	4.5	9.5	5.2	1.8	4	5.7	8.6	6.2	1.4	3	6.4
thr	7.5	4.6	3.7	5	2.6	5.7	3.8	6.8	2	5.2	9.7	4.4	1.8	3.9	6	7.2	7.3	1.5	3.5	6.9
trp	7.1	5.2	4.9	5.5	2.3	5.4	4.3	5.8	2.2	5.6	9.5	6.6	2.1	3.8	4.1	6.4	5.9	1.7	3.7	6.8
tyr	5.8	5.7	5	5.1	2.3	5.7	4.1	6.2	2.4	5	8.6	5.6	1.9	5	4.8	6.7	6.3	1.5	4.8	6.5
val	7.6	5	4.4	5.2	2.4	5.7	3.7	6.3	1.9	5	9.3	5.1	2.1	4.1	5.5	6.9	6.6	1.1	3.6	7.4



Coding Signal Detection (5)

- in-frame versus any-frame dicodons

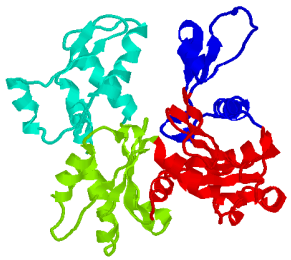


In-frame:
ATG TTG
GAT GCC
CAG AAG

Not in-frame:
TGTTGG, ATGCCC
AGAAG ., GTTGGA
AGCCCA, AGAAG ..

more sensitive

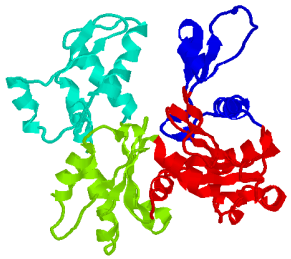
any-frame



Computational Model (1)

- Preference model:
  - for each dicodon X (e.g., AAA AAA), calculate its frequencies in coding and non-coding regions, FC(X), FN(X)
  - calculate X's preference value P(X) = log (FC(X)/FN(X))

- Properties:
  - P(X) is 0 if X has the same frequencies in coding and non-coding regions
  - P(X) has positive score if X has higher frequency in coding than in non-coding region; the larger the difference the more positive the score is
  - P(X) has negative score if X has higher frequency in non-coding than in coding region; the larger the difference the more negative the score is



Computational Model (2)

- Example

AAA ATT, AAA GAC, AAA TAG have the following frequencies

FC(AAA ATT) = 1.4%, FN(AAA ATT) = 5.2%
FC(AAA GAC) = 1.9%, FN(AAA GAC) = 4.8%
FC(AAA TAG) = 0.0%, FN(AAA TAG) = 6.3%

We have

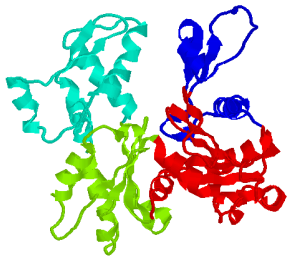
P(AAA ATT) = log (1.4/5.2) = -0.57
P(AAA GAC) = log (1.9/4.8) = -0.40
P(AAA TAG) = - infinity (treating STOP codons differently)

A region consisting of only these dicodons is probably a non-coding region

- Coding preference of a region (an any-frame model)

Calculate the preference scores of all dicodons of the region and sum them up;

If the total score is positive, predict the region to be a coding region; otherwise a non-coding region.



Computational Model (3)

- In-frame preference model (most commonly used in prediction programs)

Data collection step:

For each known coding region,

calculate in-frame preference score, $P_0(X)$, of each dicodon X; e.g., ATG TGC CGC GCT .....

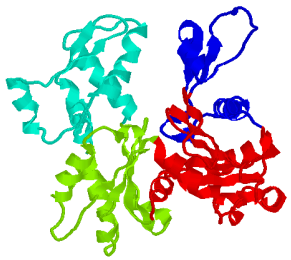
calculate (in-frame + 1) preference score, $P_1(X)$, of each dicodon X; e.g., ATG TGC CGC GCT .....

calculate (in-frame + 2) preference score, $P_2(X)$, of each dicodon X; e.g., ATG TGC CGC GCT .....

Application step:

For each possible reading frame of a region, calculate the total in-frame preference score $\Sigma P_0(X)$, the total (in-frame + 1) preference score $\Sigma P_1(X)$, the total (in-frame + 2) preference score $\Sigma P_2(X)$, and sum them up

If the score is positive, predict it to be a coding region; otherwise non-coding



Computational Model (4)

- Prediction procedure of coding region

Procedure:

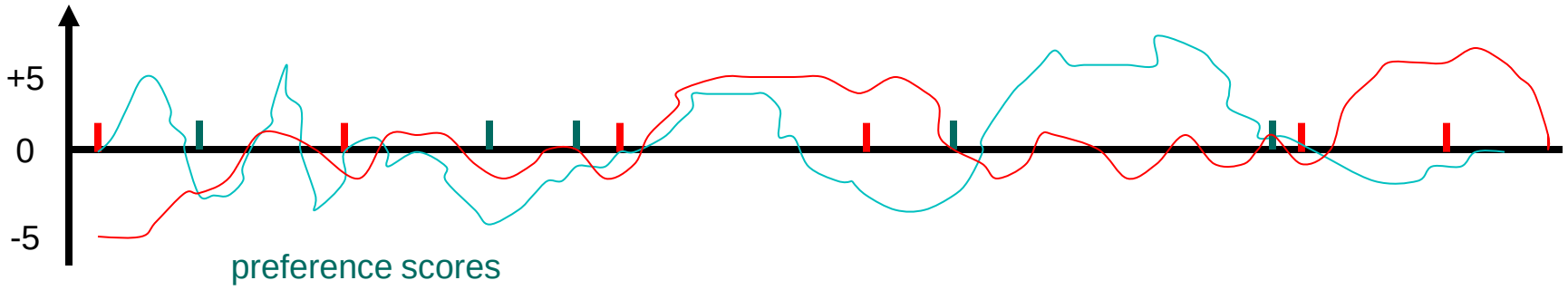
Calculate all ORFs of a DNA segment;

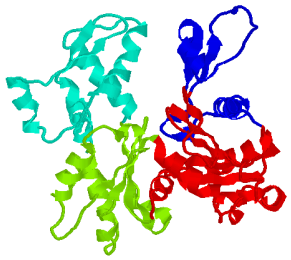
For each ORF, do the following

slide through the ORF with an increment of 10 base-pairs

calculate the preference score, in same frame of ORF, within a window of 60 base-pairs; and assign the score to the center of the window

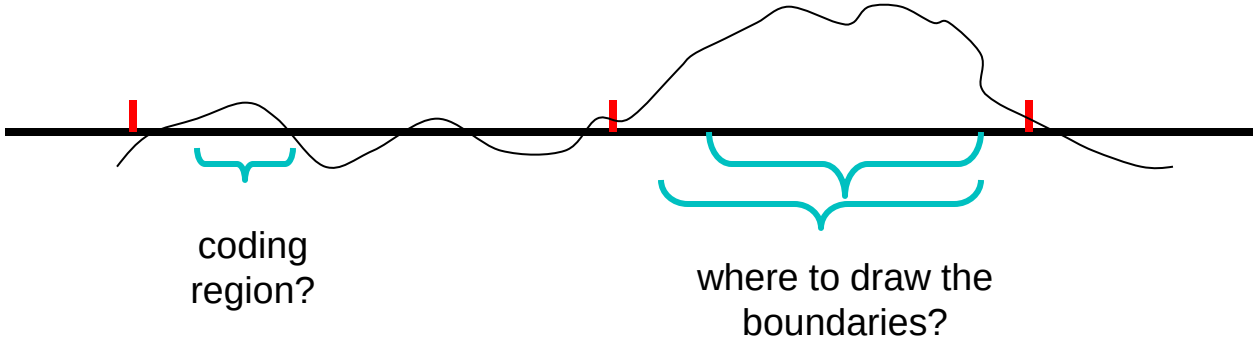
Example (forward strand in one particular frame)





Computational Model (5)

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



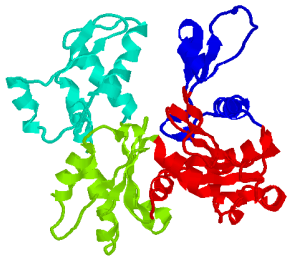
- Need a training set with known coding and non-coding regions
  - ↙ select threshold(s) to include as many known coding regions as possible, and in the same time to exclude as many known non-coding regions as possible

where to draw the line?

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



Computational Model (6)

- Why dicodon (6mer)?

Codon (3mer) -based models are not nearly as information rich as dicodon-based models

Tricodon (9mers)-based models need too many data points for it to be practical

People have used 7-mer or 8-mer based models; they could provide better prediction methods 6-mer based models

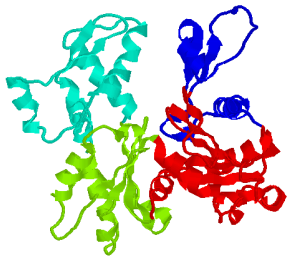
There are

4*4*4 = 64 codons

4*4*4*4*4*4 = 4,096 di-codons

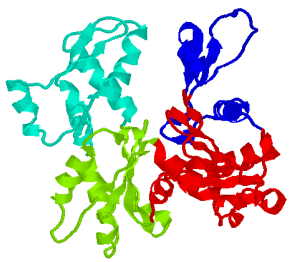
4*4*4*4*4*4*4*4*4= 262,144 tricodons

To make our statistics reliable, we would need at least ~15 occurrences of each X-mer; so for tricodon-based models, we need at least 15*262144 = 3932160 coding bases in our training data, which is probably not going to be available for most of the genomes



Lecture Outline

- Protein-encoding genes and gene structures
- Computational models for coding regions
- Computational models for coding-region boundaries
- Markov chain model for coding regions



Signals for Coding-Region Boundaries (1)

- Possible boundaries of an exon

{ translation start, acceptor site }

{ translation stop, donor site }

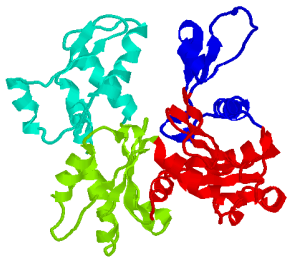
- Splice junctions:

```
EXON |    INTRON                    |   EXON
                |                 |
               \_/               \_/
A  G  G    T    A......C    A    G
64 73 100 100 62      65 100 100
(Percent occurrence)
```

- Translation start
  - ↙ in-frame ATG

donor site: coding region | GT

acceptor: AG | coding region



Signals for Coding-Region Boundaries (2)

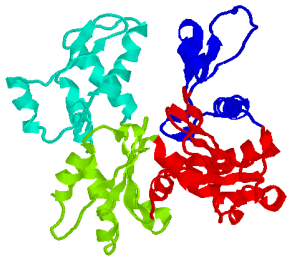
- Both splice junction sites and translation starts have certain distribution profiles
- Acceptor site (human genome)
  - if we align all known acceptor sites (with their splice junction site aligned), we have the following nucleotide distribution

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

	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1
A	11.1	12.7	3.2	4.8	12.7	8.7	16.7	16.7	12.7	9.5	26.2	6.3	100	0.0	21.4
C	36.5	30.9	19.1	23.0	34.9	39.7	34.9	40.5	40.5	36.5	33.3	68.2	0.0	0.0	7.9
G	9.5	10.3	15.1	12.7	8.7	9.5	16.7	4.8	2.4	6.3	13.5	0.0	0.0	100	62.7
U	38.9	41.3	58.7	55.6	42.1	40.5	30.9	37.3	44.4	47.6	27.0	25.4	0.0	0.0	7.9

- Donor site (human genome)

	-3	-2	-1	1	2	3	4	5	6
A	34.0	60.4	9.2	0.0	0.0	52.6	71.3	7.1	16.0
C	36.3	12.9	3.3	0.0	0.0	2.8	7.6	5.5	16.5
G	18.3	12.5	80.3	100	0.0	41.9	11.8	81.4	20.9
U	11.4	14.2	7.3	0.0	100	2.5	9.3	5.9	46.2



Model for Splice Sites (1)

- Information content
  - for a weight matrix, the information content of each column is calculated as

-F(A)*log (F(A)/.25) - F(C)*log (F(C)/.25) - F(G)*log (F(G)/.25) - F(T) *log (F(T)/.25)

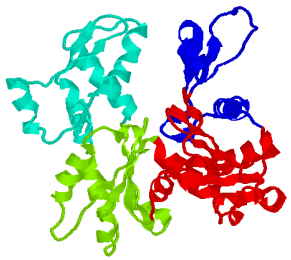
  - when a column has evenly distributed nucleotides, the information content is lowest

	-3	-2	-1	1	2	3	4	5	6
A	34.0	60.4	9.2	0.0	0.0	52.6	71.3	7.1	16.0
C	36.3	12.9	3.3	0.0	0.0	2.8	7.6	5.5	16.5
G	18.3	12.5	80.3	100	0.0	41.9	11.8	81.4	20.9
U	11.4	14.2	7.3	0.0	100	2.5	9.3	5.9	46.2

column -3: -0.34*log (.34/.25) - 0.363*log (.363/.25) -0.183* log (.183/.25) - 0.114* log (.114/.25) = 0.04

column -1: -0.092*log (.092/.25) - 0.03*log (.033/.25) -0.803* log (.803/.25) - 0.073* log (.073/.25) = 0.30

Only need to consider positions with “high” information content



Model for Splice Sites (2)

- Weight matrix model
  - build a weight matrix for donor, acceptor, translation start site, respectively
  - using positions with high information

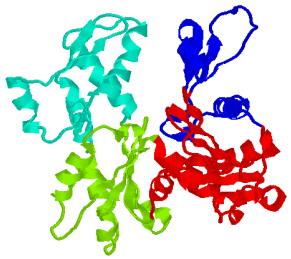
- Application of weight matrix model

	-3	-2	-1	1	2	3	4	5	6
A	34.0	60.4	9.2	0.0	0.0	52.6	71.3	7.1	16.0
C	36.3	12.9	3.3	0.0	0.0	2.8	7.6	5.5	16.5
G	18.3	12.5	80.3	100	0.0	41.9	11.8	81.4	20.9
U	11.4	14.2	7.3	0.0	100	2.5	9.3	5.9	46.2

  - add up frequencies of corresponding letter in corresponding positions

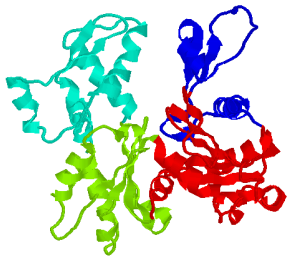
AAGGTAAGT: 0.34+0.60+0.80+1.0+1.0+0.52+0.71+0.81+0.46 = 6.24

TGTGTCTCA: 0.11+0.12+0.03+1.0+1.0+0.02+0.07+0.05+0.16 = 2.56



Lecture Outline

- Protein-encoding genes and gene structures
- Computational models for coding regions
- Computational models for coding-region boundaries
- Markov chain model for coding regions



Why Markov Chain?

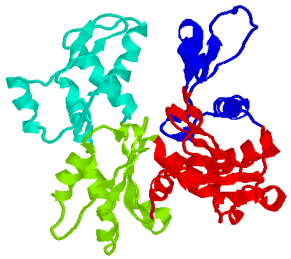
- Preference model cannot capture all the dependence relationship among adjacent dicodons

- Markov chain model has been a popular model for modeling dependence in a linear sequence (a chain of events)

- Basic assumption of the model for a chain of events:

  The "occurrence" of each event depends only on the most recent events right before this event

- Example: the weather of today is a function of only the weather of past seven days (i.e., it is independent of the weather of eight days ago)

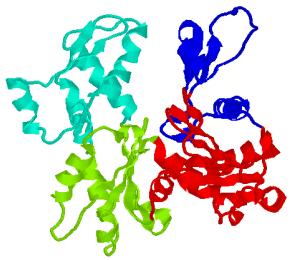


Markov Chain Model (1)

- Basics of probabilities:
  - P(A) represents the probability of A being true
  - P(A, B) represents the event of having both A and B being true
  - if A and B are independent, P(A, B) = P(A) * P(B)
  - P(A | B), conditional probability, of A being true under the condition B is true (this applies only when B is true)

- Zero-th order Markov chain is equivalent to "all events are independent"
- First order Markov chain: the occurrence of an event depends only on the event right before it

$$P(A_1 A_2 A_3 A_4 A_5 A_6) = P(A_1)\, P(A_2 \mid A_1)\, P(A_3 \mid A_2)\, P(A_4 \mid A_3)\, P(A_5 \mid A_4)\, P(A_6 \mid A_5)$$



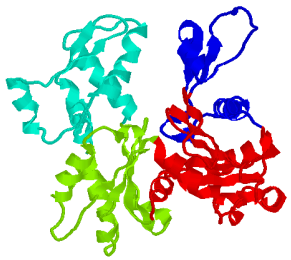
Markov Chain Model (2)

- K-th order Markov chain model:

Example of 5th order Markov chain:

$P(A_1 A_2 A_3 A_4 A_5 A_6 A_7 A_8 A_9 A_{10} A_{11}) = P(A_1 A_2 A_3 A_4 A_5) * P(A_6 \mid A_1 A_2 A_3 A_4 A_5) * P(A_7 \mid A_2 A_3 A_4 A_5 A_6) * P(A_8 \mid A_3 A_4 A_5 A_6 A_7) * P(A_9 \mid A_4 A_5 A_6 A_7 A_8) * P(A_{10} \mid A_5 A_6 A_7 A_8 A_9) * P(A_{11} \mid A_6 A_7 A_8 A_9 A_{10})$

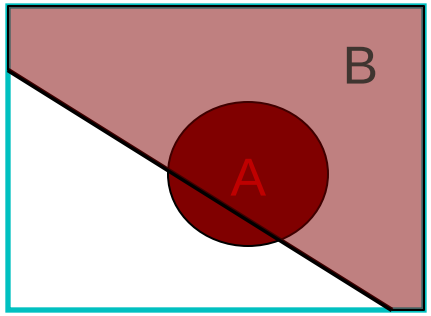
- Markov chain model allows us to “decompose” a large problem into a collection of smaller problems



Markov Chain Model (3)

- Definition of conditional probability

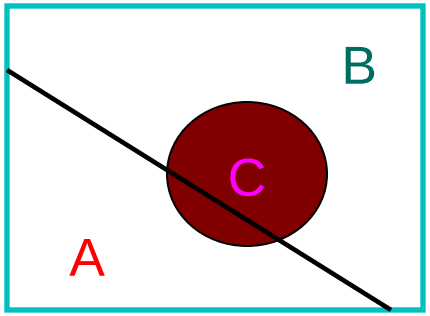
$$P(A \mid B) = P (A, B) / P(B)$$

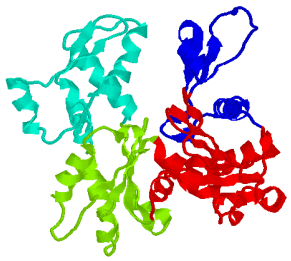


- Decomposition rule

$$P(C) = P(C \mid A)\, P(A) + P(C \mid B)\, P(B)$$

as long A and B do not overlap and A plus B completely covers C





Markov Chain Model for Coding Region (1)

- Any-frame Markov chain model

Bayesian formula for coding:

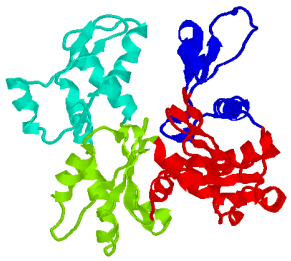
$$P(\text{coding} \mid A_1 \ldots A_n) = P(\text{coding}, A_1 \ldots A_n) / P(A_1 \ldots A_n)$$

$$= \frac{P(A_1 \ldots A_n \mid \text{coding}) * P(\text{coding})}{P(A_1 \ldots A_n \mid \text{coding})\,P(\text{coding}) + P(A_1 \ldots A_n \mid \text{noncoding})\,P(\text{noncoding})}$$

Bayesian formula for non-coding:

$$P(\text{non-coding} \mid A_1 \ldots A_n)$$

$$= \frac{P(A_1 \ldots A_n \mid \text{noncoding}) * P(\text{noncoding})}{P(A_1 \ldots A_n \mid \text{coding})\,P(\text{coding}) + P(A_1 \ldots A_n \mid \text{noncoding})\,P(\text{noncoding})}$$

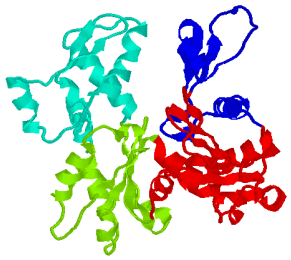


Markov Chain Model for Coding Region (2)

This formula decomposes a problem of "predicting a region $A_1 \ldots A_n$ being a (non) coding region" to the following four problems

1. Estimating probability of seeing $A_1 \ldots A_n$ in noncoding regions
2. Estimating probability of coding bases in a whole genome
3. Estimating probability of noncoding bases in a whole genome
4. Estimating probability of seeing $A_1 \ldots A_n$ in coding regions

All these can be estimated using known coding and noncoding sequence data



Markov Chain Model for Coding Region (3)

- Any-frame Markov chain model

Markov chain model (5^{th} order) :

a priori probability

conditional probability

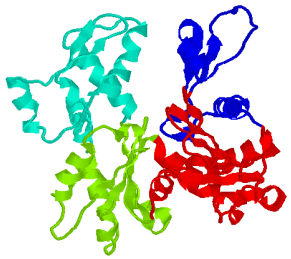
$$P(A_1 \ldots A_n \mid \text{coding}) = P(A_1 A_2 A_3 A_4 A_5 \mid \text{coding}) * P(A_6 \mid A_1 A_2 A_3 A_4 A_5, \text{coding}) * P(A_7 \mid A_2 A_3 A_4 A_5 A_6, \text{coding}) * \ldots * P(A_n \mid A_{n-5} A_{n-4} A_{n-3} A_{n-2} A_{n-1}, \text{coding})$$

Markov chain model (5^{th} order) :

$$P(A_1 \ldots A_n \mid \text{noncoding}) = P(A_1 A_2 A_3 A_4 A_5 \mid \text{noncoding}) * P(A_6 \mid A_1 A_2 A_3 A_4 A_5, \text{noncoding}) * P(A_7 \mid A_2 A_3 A_4 A_5 A_6, \text{noncoding}) * \ldots * P(A_n \mid A_{n-5} A_{n-4} A_{n-3} A_{n-2} A_{n-1}, \text{noncoding})$$

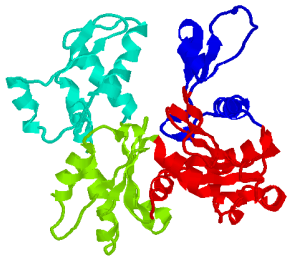
P(coding): total # coding bases/total # all bases

P(noncoding): total # noncoding bases/total # all bases



Build Markov Tables (1)

- *a priori* probability tables (5th order): P(5mer |coding) and P(5mer | noncoding)
  - 5mer frequency table for coding regions
  - 5mer frequency table for noncoding regions
- Conditional probability tables (5th order): P(X | 5mer, coding) and P(X | 5mer, noncoding) (X could be A, C, G, T)
  - For a fixed 5mer (e.g., ATT GT), what is the probability to have A, C, G or T following it in coding region
  - For a fixed 5mer (e.g., ATT GT), what is the probability to have A, C, G or T following it in noncoding region
- P(coding) = ~0.02 and P(noncoding) = ~0.98



Build Markov Tables (2)

a priori probabilities for coding PAC

AAA AA: 0.000012
AAA AC: 0.000001
AAA AG: 0.000101
…….

a priori probabilities for noncoding PAN

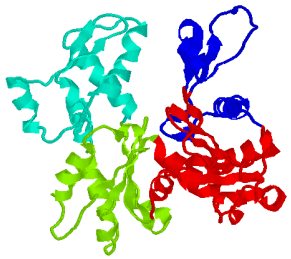
AAA AA: 0.000329
AAA AC: 0.000201
AAA AG: 0.000982
…….

conditional probabilities for coding PC

	A	C	G	T
AAA AA:	0.17	0.39	0.01	0.43
AAA AC:	0.12	0.44	0.02	0.42
AAA AG:	0.01	0.69	0.10	0.20
…….				

conditional probabilities for noncoding PN

	A	C	G	T
AAA AA:	0.71	0.09	0.00	0.20
AAA AC:	0.61	0.19	0.02	0.18
AAA AG:	0.01	0.69	0.10	0.20
…….				



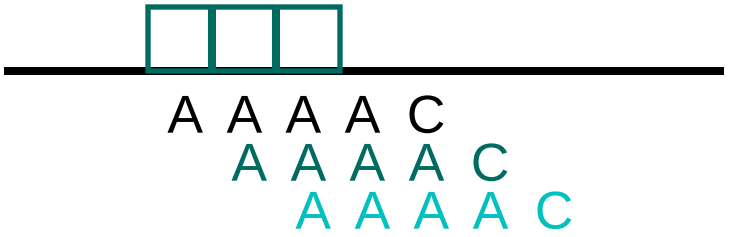
In-Frame Markov Chain Model (1)

- In-frame Markov tables

a priori probabilities for coding $PAC0_{0, 1, 2}$

AAA AA:	0.000012	0.000230 0.000009
AAA AC:	0.000001	0.000182 0.000011
AAA AG:	0.000101	0.000301 0.000101
…….		

translation frame



conditional probabilities for coding PC_0

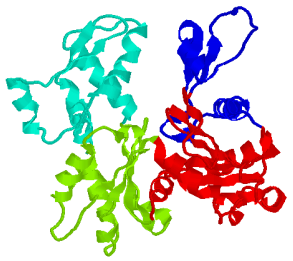
AAA AA:	0.17	0.39	0.01	0.43
AAA AC:	0.12	0.44	0.02	0.42
AAA AG:	0.01	0.69	0.10	0.20
…….				

conditional probabilities for coding PC_1

AAA AA:	0.33	0.12	0.10	0.35
AAA AC:	0.02	0.49	0.12	0.37
AAA AG:	0.10	0.60	0.15	0.15
…….				

conditional probabilities for coding PC_2

AAA AA:	0.17	0.39	0.01	0.43
AAA AC:	0.12	0.44	0.02	0.42
AAA AG:	0.01	0.69	0.10	0.20
…….				



In-Frame Markov Chain Model (2)

- In-frame Markov chain (5th order) calculation

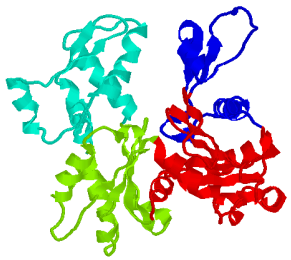
$P_0(A_1 \ldots A_n \mid \text{coding}) = P_0(A_1 A_2 A_3 A_4 A_5 \mid \text{coding}) * P_0(A_6 \mid A_1 A_2 A_3 A_4 A_5, \text{coding}) *$
$P_1(A_7 \mid A_2 A_3 A_4 A_5 A_6, \text{coding}) * P_2(A_8 \mid A_3 A_4 A_5 A_6 A_7, \text{coding}) * .....$

0	1	2	0	1	2	0	1	2

A A A A C T G C .....

$P(A_1 \ldots A_n \mid \text{noncoding}) = P(A_1 A_2 A_3 A_4 A_5 \mid \text{noncoding}) * P(A_6 \mid A_1 A_2 A_3 A_4 A_5, \text{noncoding}) *$
$P(A_7 \mid A_2 A_3 A_4 A_5 A_6, \text{noncoding}) * \ldots. * P(A_n \mid A_{n-5} A_{n-4} A_{n-3} A_{n-2} A_{n-1}, \text{noncoding})$

Calculation for non-coding regions stays the same



In-Frame Markov Chain Model (3)

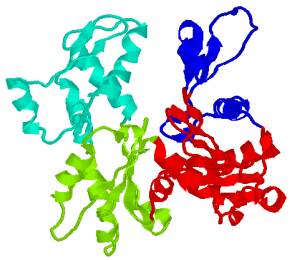
- Markov tables for Human genome

a priori table

0.001801	0.000959	0.000854
0.000949	0.001041	0.000674
0.000979	0.001836	0.001349
0.001240	0.001109	0.000854
0.001031	0.000689	0.000937
0.000785	0.000824	0.000846
0.000523	0.000277	0.000255
0.001390	0.000817	0.000614
0.002405	0.000727	0.001970
0.001666	0.000742	0.000997
0.001816	0.000712	0.001491
0.001382	0.000367	0.000682
0.000747	0.000562	0.000667
0.000792	0.000884	0.000367
0.000635	0.000929	0.001348
0.001218	0.001086	0.000419
0.002286	0.000749	0.000539
0.001405	0.000570	0.000382
0.001449	0.001574	0.000899
0.001785	0.000322	0.000300
0.001845	0.000637	0.000899
0.001696	0.000569	0.000569
0.000874	0.000172	0.000247
0.002129	0.000517	0.000757
0.001240	0.000262	0.000172
0.000657	0.000435	0.000150
0.000949	0.000352	0.000225
0.000814	0.000315	0.000075
0.000740	0.000412	0.000330
0.001121	0.000697	0.000510
0.000941	0.001378	0.000839
0.001546	0.000607	0.000285
0.005282	0.002652	0.000622
0.001942	0.001581	0.000360
0.002331	0.003154	0.000614

in-frame conditional probabilities

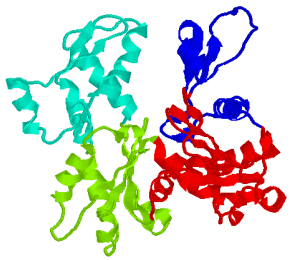
0.186712	0.170170	0.427390	0.215729
0.259907	0.338652	0.110139	0.291302
0.282435	0.229042	0.305389	0.183134
0.228937	0.222835	0.271063	0.277165
0.260952	0.246270	0.384087	0.108691
0.295271	0.342875	0.047604	0.314250
0.214186	0.328511	0.228651	0.228651
0.145116	0.247365	0.430077	0.177442
0.301228	0.211205	0.285700	0.201867
0.291471	0.349795	0.076202	0.282532
0.374479	0.197526	0.251042	0.176953
0.124338	0.302720	0.410809	0.162134
0.000001	0.589869	0.000001	0.410131
0.226334	0.349059	0.075547	0.349059
0.000001	0.364719	0.270561	0.364719
0.153399	0.460197	0.116503	0.269902
0.166720	0.290826	0.372530	0.169924
0.324470	0.404241	0.085158	0.186131
0.164926	0.448450	0.185647	0.200977
0.050185	0.414194	0.343088	0.192534
0.198359	0.161969	0.542543	0.097128
0.237835	0.422977	0.105672	0.233516
0.111173	0.512849	0.273464	0.102514
0.059626	0.291251	0.540305	0.108818
0.048228	0.355315	0.469882	0.126575
0.204605	0.431860	0.227256	0.136279
0.094442	0.378024	0.417396	0.110139
0.036582	0.256972	0.596402	0.110045
0.000001	0.737294	0.000001	0.262706



In-Frame Markov Chain Model (4)

- Coding score procedure
  - for a DNA segment [i, j], calculate Markov coding scores scoreC[0], scoreC[1], scoreC[2], representing three frames (one strand), and non-coding score scoreN
  - if MAX { scoreC[0], scoreC[1], scoreC[2] } > scoreN, the region is predicted to coding; otherwise non-coding

calculated reading frame in reference to the starting point of the first base



Application of Markov Chain Model

- Prediction procedure

Procedure:

Calculate all ORFs of a DNA segment;

For each ORF, do the following

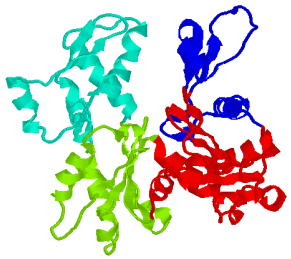
slide through the ORF with an increment of 10 base-pairs

calculate the preference score, in same frame of ORF, within a window of 60 base-pairs; and assign the score to the center of the window

- A computing issue

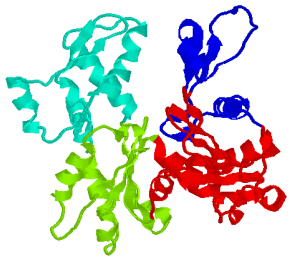
Multiplication of many small numbers (probabilities) is generally problematic in computer

Converting a * b * c * d ..... *z to log (a) + log (b) + log (c) + log (d) ... log (z)



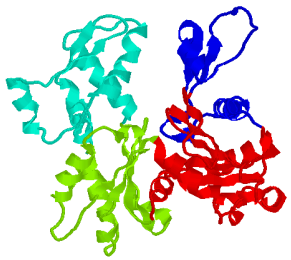
References

- Chapter 9 in “Current Topics in Computational Molecular Biology, edited by Tao Jiang, Ying Xu, and Michael Zhang. MIT Press. 2002.”
- Chapter 9 in “Pavel Pevzner: Computational Molecular Biology - An Algorithmic Approach. MIT Press, 2000.”



Selected Reading

- http://www.ncbi.nlm.nih.gov/pubmed/20221925
- http://www.ncbi.nlm.nih.gov/pubmed/12364589
- http://www.ncbi.nlm.nih.gov/pubmed/16728949
- http://www.ncbi.nlm.nih.gov/pubmed/21653517
- http://www.ncbi.nlm.nih.gov/pubmed/19564452
- http://www.ncbi.nlm.nih.gov/pubmed/19494180
- http://www.ncbi.nlm.nih.gov/pubmed/10779491



Acknowledgments

This file is for the educational purpose only. Some materials (including pictures and text) were taken from the Internet at the public domain.