



Gene Finding Project

Charles Yan



Gene Finding

- **Summary of the project**
 - Download the genome of E. Coli K12
 - Gene-finding using k^{th} -order Markov chains, where $k = 1, 2, 3$
 - Gene-finding using inhomogeneous Markov chains



The Genome of E. Coli K12

- **Go to**
<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed>
- **1.2 Search Genome** for NC_000913 (which is the access number for E. coli K12)

The Genome of E. Coli K12

The screenshot shows a web browser window with the address `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=genome`. The search results page for `NC_000913` is displayed. The search bar shows "Genome" selected and "NC_000913" entered. The results list shows one entry: "1: NC_000913" with a red arrow pointing to it and the text "Click to access the sequence". The entry details are: "Escherichia coli K12, complete genome", "DNA; circular; Length: 4,639,675 nt", "Replicon Type: chromosome", and "Created: 2001/10/15".

Address `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=genome` Start Search Go

Adobe Y! Search Web Mail My Yahoo! Ansv

Entrez Genome + Add Tab

NCBI Entrez Genome My N Sign

All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy

Search Genome for NC_000913 Go Clear Save Search

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Send to

All: 1

1: [NC_000913](#) Click to access the sequence

Escherichia coli K12, complete genome
DNA; circular; Length: 4,639,675 nt
Replicon Type: chromosome
Created: 2001/10/15

About Entrez
Entrez Genome Help
Submitting Genome Project Genome sequence
Microbial Genome Projects PDB neighbors

The Genome of E. Coli K12

The screenshot shows the NCBI Entrez Genome search results page. At the top, there is a navigation bar with links for All Databases, PubMed, Nucleotide, Protein, Genome, Structure, PMC, Taxonomy, and Books. Below this is a search bar with 'Genome' selected and a search button. The main content area shows 'All: 1' result. The 'Genome' link is highlighted in blue.

[Genome](#) > [Bacteria](#) > ***Escherichia coli K12, complete genome***

Lineage: [Bacteria](#) ; [Proteobacteria](#) ; [Gammaproteobacteria](#) ; [Enterobacteriales](#) ; [Enterobacteriaceae](#) ; [Escherichia](#) ; [Esche](#)

Click to access the sequence

Genome Info:	Features:	BLAST homologs:	Links:	Review Info:
Refseq: NC_000913	Genes: 4437	COG	Genome Project	Publications: [2]
GenBank: U00096	Protein coding: 4243	3D Structure	Refseq FTP	Refseq Status: Provisio
Length: 4,639,675 nt	Structural RNAs: 157	TaxMap	GenBank FTP	Seq. Status: Completec
GC Content: 50%	Pseudo genes: 31	TaxPlot	BLAST	Sequencing center: Univ Wisconsin
% Coding: 86%	Others: 93	GenePlot	TraceAssembly	Completed: 2001/10/15
Topology: circular	Contigs: 1	gMap	CDD	Organism Group

The Genome of E. Coli K12

NCBI Nucleotide

Search Nucleotide for [] Go Clear

Display GenBank Show 5 File select "file" to save the entry to a file

Range: from begin to end Reverse complemented strand Features: + Refresh

1: [NC_000913](#) Reports Escherichia coli K12

[Comment](#) [Features](#) [Sequence](#)

LOCUS NC_000913
DEFINITION Escherichia coli K12
ACCESSION NC_000913
VERSION NC_000913.2 GI:4917157
PROJECT GenomeProject:225
KEYWORDS .
SOURCE Escherichia coli K12
ORGANISM [Escherichia coli K12](#)
Bacteria; Proteobacteria; Enterobacteriaceae;
REFERENCE 1 (bases 1 to 46396)
AUTHORS Riley,M., Abe,T., Anand,B., Chaudhuri,R.R., Glasner,J., Mori,H., Perna,N.T., Plunkett,G., Rudd,K.E., Serres,M.H., Thomas,G.H., Thomson,N.R., Wishart,D. and Wanner,B.L.
TITLE Escherichia coli K-12: a cooperatively developed annotation snapshot--2005
JOURNAL Nucleic Acids Res. 34 (1), 1-9 (2006)
PIRMEID [16397293](#)

File Download

Do you want to open or save this file?

Name: sequences.gb
Type: Unknown File Type
From: www.ncbi.nlm.nih.gov

Open Save Cancel

While files from the Internet can be useful, some files can potentially harm your computer. If you do not trust the source, do not open or save this file. [What's the risk?](#)



The Genome of E. Coli K12

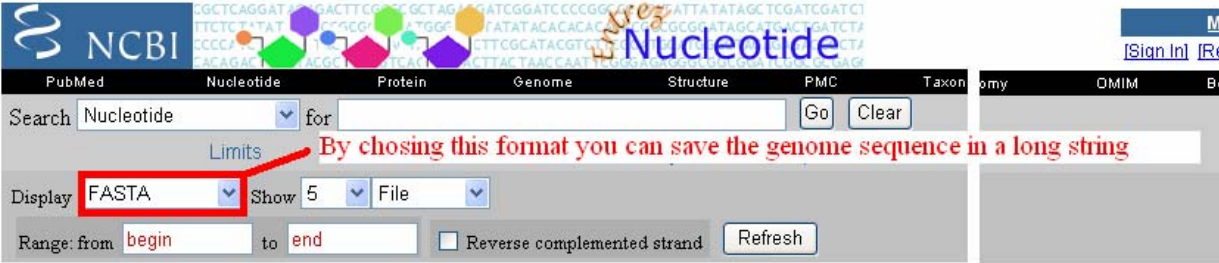
- **Genome sequence.** Close to the end of the file, you will find some thing like this:

```
ORIGIN
1 agcttttcat tctgactgca acgggcaata tgtctctgtg tggattaaaa aaagagtgtc
61 tgatagcagc ttctgaactg gttacctgcc gtgagtaaat taaaatttta ttgacttagg
121 tcaactaaata ctttaaccaa tataggcata gcgcacagac agataaaaaa tacagagtac
181 acaacatcca tgaaacgcat tagcaccacc attaccacca ccatcaccat taccacaggt
241 aacggtgccg gctgacgcgt acaggaaaca cagaaaaaag cccgcacctg acagtgcggg
301 cttttttttt cgaccaaaagg taacgaggta acaacctatgc gagtgttgaa gttcggcggg
361 acatcagtg gcaaatgcaga acgttttctg cgtgttgccg atattctgga aagcaatgcc
421 aggcaggggg aggtggccac cgtcctctct gcccccgcca aaatcaccaa ccacctgggtg
481 gcgatgattg aaaaaacat tagcggccag gatgctttac ccaatatcag cgatgccgaa
541 cgtatttttg ccgaactttt gacgggactc gccgccgccc agccgggggt cccgctggcg
601 caattgaaaa ctttcgtcga tcaggaattt gcccaataa aacatgtcct gcatggcatt
661 agtttggttg ggcagtgcgc ggatagcatc aacgctgcgc tgatttgccg tggcgagaaa
721 atgtcgatcg ccattatggc cggcgtatta gaagcgcgcg gtcacaacgt tactgttatc
781 gatccggtcg aaaaactgct ggcagtgggg cattacctcg aatctaccgt cgatattgct
841 gagtccacce gccgtattgc ggcaagccgc attccggctg atcacatggt gctgatggca
901 ggtttcaccg ccggtaatga aaaaggcgaa ctggtggtgc ttggacgcaa cggttccgac
961 tcaatgatac cgtatgtagg tcaatgattc cgggggactt cttggggact tgggggggac
```

- This is the real sequence of the genome. The number at each line show the index of the starting letter. In this format, the sequence is shown in 6 columns with each column having 10 letters.

The Genome of E. Coli K12

- You can also get the genome sequence in a long string by selecting the format to save the file



The screenshot shows the NCBI Nucleotide search interface. The search criteria are set to 'Nucleotide'. The 'Display' dropdown menu is set to 'FASTA', which is highlighted with a red box. A red arrow points from a text box to this dropdown. The text box contains the instruction: 'By choosing this format you can save the genome sequence in a long string'. Below the search bar, there are options for 'Limits', 'Show' (set to 5), and 'File' format. The 'Range' is set from 'begin' to 'end'. A 'Refresh' button is also visible.

By choosing this format you can save the genome sequence in a long string

```
1: U00096 Reports Escherichia coli... [gi:48994873]
>gi|48994873|gb|U00096.2| Escherichia coli K12 MG1655, complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTTACCTGCGCGTGAAGTAAATTTAAATTTTATTGACTTAGGTCACTAAATACTTTAAACCAA
TATAGGCATAGCGCACAGACAGATAAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACC
ATTACCACCACCATCACCATTACCACAGGTAACGGTGGCGGGTGCAGCGGTACAGGAAACACAGAAAAAAG
CCCGCACCTGACAGTGGCGGCTTTTTTTTCGACCAAAGGTAACGAGGTAACAAACCATGCGAGTGTGAA
GTTCCGGCGGTACATCAGTGGCAAAATGCAGAACGTTTTCTCGGTGTTGCCGATATTCTGGAAGCAATGCC
AGGCAGGGCAGGTGGCCACCGTCCTCTCTGCCCCGCCAAAATCACCAACCCTGGTGGCGATGATTG
AAAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGATTTTTTGCCGAACTTTT
GACGGGACTCGCCCGCCCGCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACCTTTCGTGATCAGGAATTT
GCCCAAATAAAACATGCTCTGCATGGCATTAGTTTGTGGGGCAGTGCCCGGATAGCATCAACGCTGGCC
TGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAAGCCGCGGTCACAACGCT
TACTGTTATCGATCCGGTCAAAAAATGCTGGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCT
```

Genes

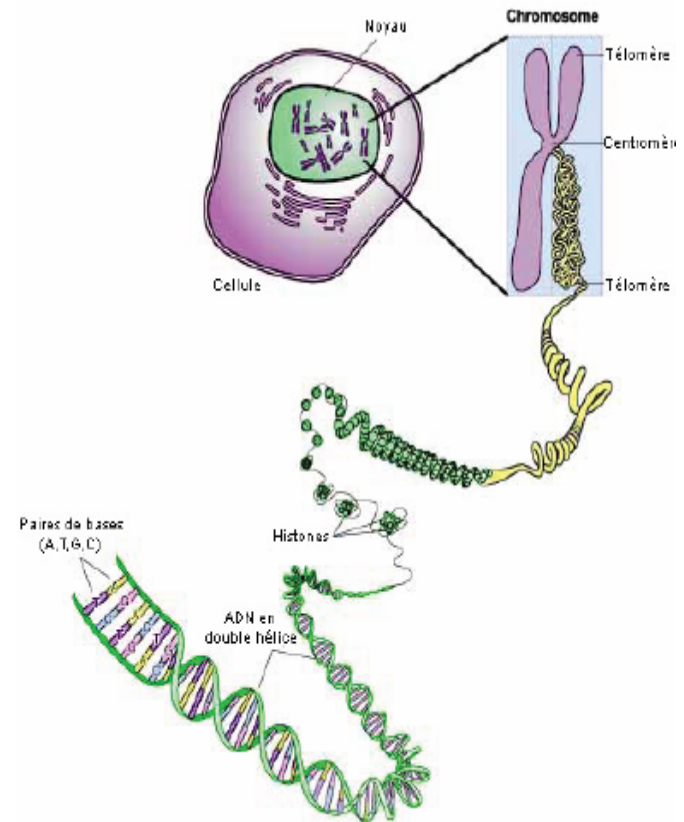
- In the middle of the file, you will see something like

```
LPMTK...GLG...YHAKPKVNEKA...VTIRHADLMGVFCILSGSLNQR"
gene 16864..17508 → The starting point and ending point of a gene
/gene="ytjB"
/locus_tag="b4387"
/note="synonyms: ECK4379, JW4350, smp"
/db_xref="ECOCYC:EG10951"
/db_xref="GeneID:946089"
CDS 16864..17508
/gene="ytjB"
/locus_tag="b4387"
/note="orf, hypothetical protein"
/codon_start=1
/transl_table=11
/product="hypothetical protein"
/protein_id="NP_418804.1"
/db_xref="GI:16132204"
/db_xref="ASAP:ABE-0014387"
/db_xref="ECOCYC:EG10951"
/db_xref="GeneID:946089"
/translation="MARTKLFRLHRAVIVLFLCLALLVALMQGASWFSQNHQQRNPQ
LEELARTLARQVTLNVAPLMRTDSPDEKRIQAILDQLTDESRILDAGVYDEQGDLIAR
SGESVEVRDRLALDGGKAGGYFNQQIVEPIAGKNGPLGYLRLTLDTHTLATEAQQVDN
TTNILRMLLLSLAIGVVLTRTLQGRTRWQQSPFLLTASKPVPPEEESEKKE"
```

- This example shows that the DNA sequence from 16864 to 17508 is a gene.

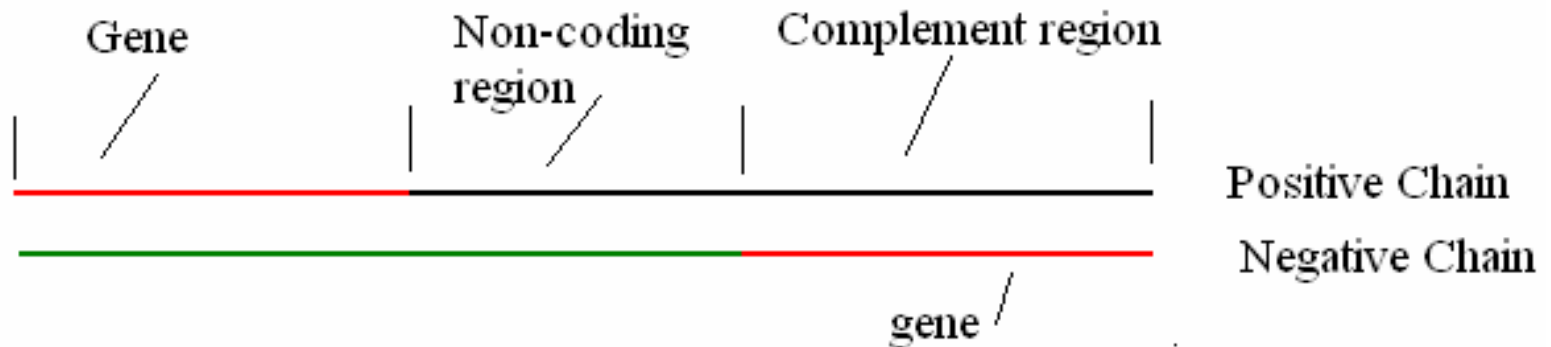
Complement Regions

- DNA is a double helix molecule. It has two complementary chains. The sequence we see in this file is only of them. This chain is often referred as ***positive chain***. The other one is ***negative chain***.



Complement Regions

- There can be genes in both chains. If the gene is on the negative chain. The corresponding region on the positive chain is called *complement region*.



Complement Regions

- This is an example of a complement region.

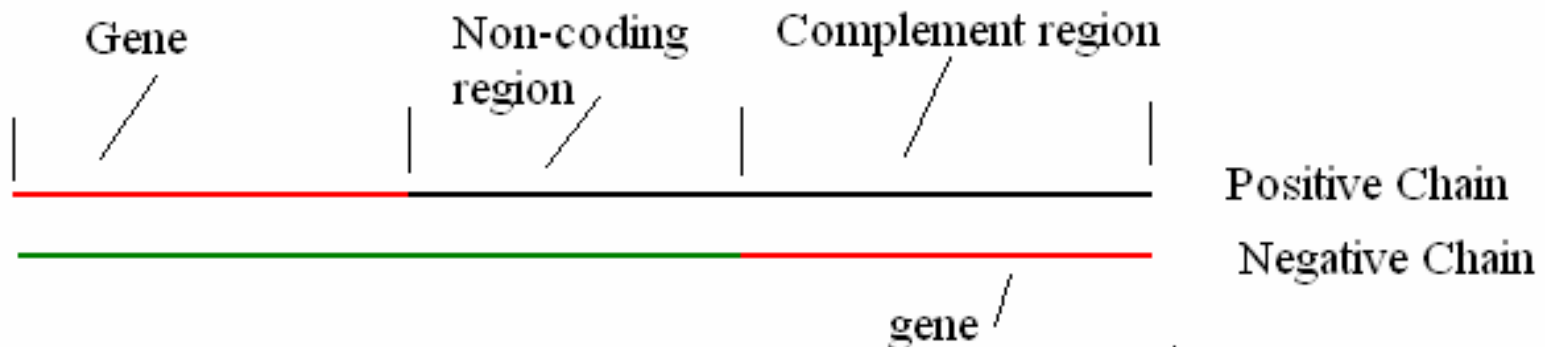
```
gene      GAVR"
          complement(18553..19884)
          /gene="yjjj"
          /locus_tag="b4385"
          /note="synonyms: ECK4377, JW4348"
          /db_xref="ECOCYC:EG12342"
          /db_xref="GeneID:944883"
CDS       complement(18553..19884)
          /gene="yjjj"
          /locus_tag="b4385"
          /note="orf, hypothetical protein"
          /codon_start=1
          /transl_table=11
          /product="predicted DNA-binding transcriptional regulator"
          /protein_id="NP_418802.1"
          /db_xref="GI:16132202"
          /db_xref="ASAP:ABE-0014383"
          /db_xref="ECOCYC:EG12342"
          /db_xref="GeneID:944883"
          /translation="MSELTDLLLQGPRSAPELRQRLAISQATFSRLVAREDRVIRFGK
ARATRYALLRPYRGIERIPVWRVDDTGKAHKFADIRLCWPQGSCLVTGADGDEQWFDG
LPWYLTDLRPQGFLGRAWGRKLAQAQLNLTDDIRLWQEEDVLYALTTFNGEYTGGLVVG
EGNYQRWITAQHPAEIPLDQKLTHYEQLASDALAGEIVGSSAGGEQPKFTYYAQTSPG
NKHVLVKFTVPQQTAVSQRWGDLLIAESIAAQILRDGGIHAIESTVLVTSNRQVFLEA
ERFDCKGNDGRLPIVSLEAVQSEF ISSPGSWPQAMRRLCEQQLVTHQSVAQTEVIWAF
GRLIANSMDMHAGNLSFYLSPEPPFALTPVYDMLPMVYAPNSAGMLRDAAEVKFDLNVS
KSAWLTAIPLAQQFWQTVARDPRISEAFRHIAQEMPEKIRQIEEKVARMGG"
```

This key word shows that the negative chain is a gene. The corresponding region on the positive chain is 18553-19884

Non-Coding Regions

The rest of the genome that are not labeled as gene or complement does not encode genetic information. These regions are *non-coding regions*.

The following figure shows that the positive chain is divided into three types of region: gene, non-coding region and complement region.



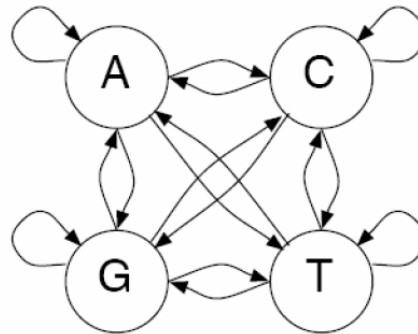


1st-order Markov chain

- Since there are three types of regions on the sequence we have, we will develop three models corresponding to them: *gene model*, *non-coding model* and *complement model*.

1st-order Markov chain

- For these models, we use the same structure as we shown in the example of identifying CpG island.



States: A,C,G,T

Emissions: corresponding letter

Transitions: $a_{st} = P(x_i = t \mid x_{i-1} = s)$

The structure of the 1st-order Markov chain model.

1st-order Markov chain

■ Then, each model is reduced to a transition probability table. Here is an example for the gene model (1st-order Markov chain). **We will need to estimate the probabilities for each model.**

This is the model for gene

End state

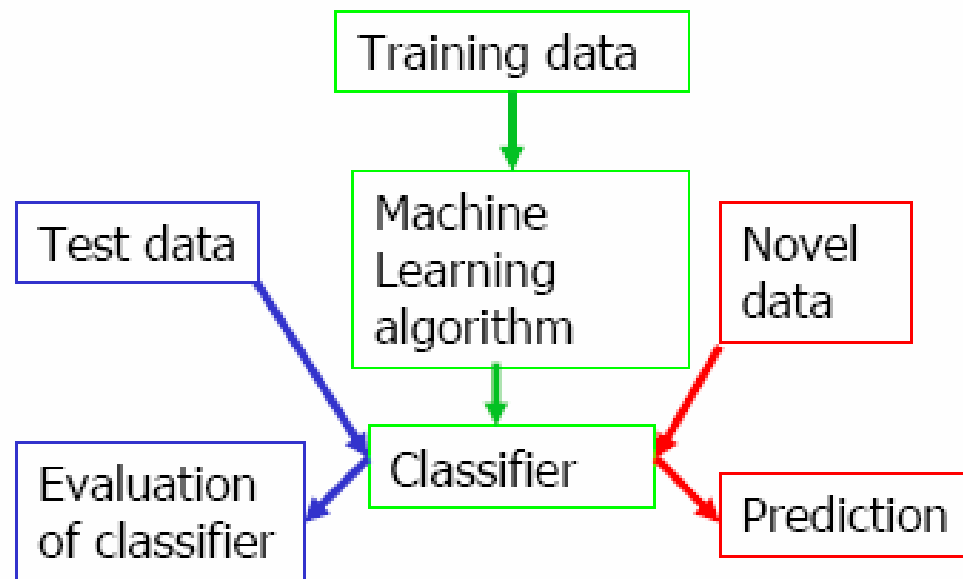
Gene	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

Starting state

The probability that the C state transit into G state. In another word, it the probability that you will see a C followed by a G on the genome sequence.

Machine-Learning Approach

- Split data into a **training set** and a **test set**
- Use the training set to train a classifier
- Test the classifier on test set
- The classifier then can be applied to novel data



Three-Fold Cross-Validation

- The genome sequence will be divided into three parts. In the first round of experiment, part 1 and 2 are used to estimate the probabilities. Then the models are used to make predictions on part 3. Then we rotate through the three parts. Until predictions are made on each part.

3-fold cross-validation

E.Coli K12 Genome
4,639,675

	Training Set	Test Set
Round 1	————— ————	—————
Round 2	————— ————	—————
Round 3	————— ————	—————



Estimation of the Transition Probabilities

- We will use maximum likelihood approach to estimate the probability. Let $a(s, t)$ be the probability that state s transit into state t . The formula to calculate $a(s, t)$ is:

$$a(s, t) = \frac{C_{st}}{\sum_m C_{sm}}$$

- When we estimate the probabilities for the gene model, C_{st} is the number of times that t follows s on gene sequences. In another word, it the number of times that ts appears on gene sequences. C_{sm} is the number of times that s is follow by any letter, that is the number of times s appear on gene sequences.

Training

- We will have three transition probability tables.

This is the model for gene

End state

Gene	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	<u>0.274</u>	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

Starting state

The probability that the C state transit into G state. In another word, it the probability that you will see a C followed by a G on the genome sequence.



Prediction

Now we have three models, i.e. three probability tables. Then, how are we going to make predictions using them? For example, we have a sequence $X=x_1x_2x_3\dots x_n$, where $x_i \in \{A, T, C, G\}$. Then, **the probability that this sequence belongs to gene is given by**

$$\begin{aligned}P_{gene}(X) &= P_{gene}(x_1) \cdot P_{gene}(x_2 | x_1) \cdot P_{gene}(x_3 | x_2x_1) \cdot \dots \cdot P_{gene}(x_n | x_{n-1}x_{n-2}\dots x_1) \\ &= P_{gene}(x_1) \cdot P_{gene}(x_2 | x_1) \cdot P_{gene}(x_3 | x_2) \cdot \dots \cdot P_{gene}(x_n | x_{n-1}) \\ &= P_{gene}(x_1) \cdot a_{gene}(x_1, x_2) \cdot a_{gene}(x_2, x_3) \cdot \dots \cdot a_{gene}(x_{n-1}, x_n) \\ &= P_{gene}(x_1) \cdot \prod_{i=2}^n a_{gene}(x_{i-1}, x_i)\end{aligned}$$

Where $a_{gene}(x_{i-1}, x_i)$ is the transition probability from x_{i-1} to x_i in the gene model, and $P_{gene}(x_1)$ is the prior probability that x_1 appears in genes, that is, the fraction of the letters in genes that are x_1 .



Prediction

The probabilities that X belongs to non-coding region or complement region are calculated similarly

$$P_{non-coding}(X) = P_{non-coding}(x_1) \cdot \prod_{i=2}^n a_{non-coding}(x_{i-1}, x_i)$$

$$P_{complement}(X) = P_{complement}(x_1) \cdot \prod_{i=2}^n a_{complement}(x_{i-1}, x_i)$$

Then we compare $P_{gene}(X)$, $P_{complement}(X)$ and $P_{non-coding}(X)$ and assign X to the class that give the largest probability. For example, if $P_{gene}(X)$ is the largest, then X is predicted to be a gene.



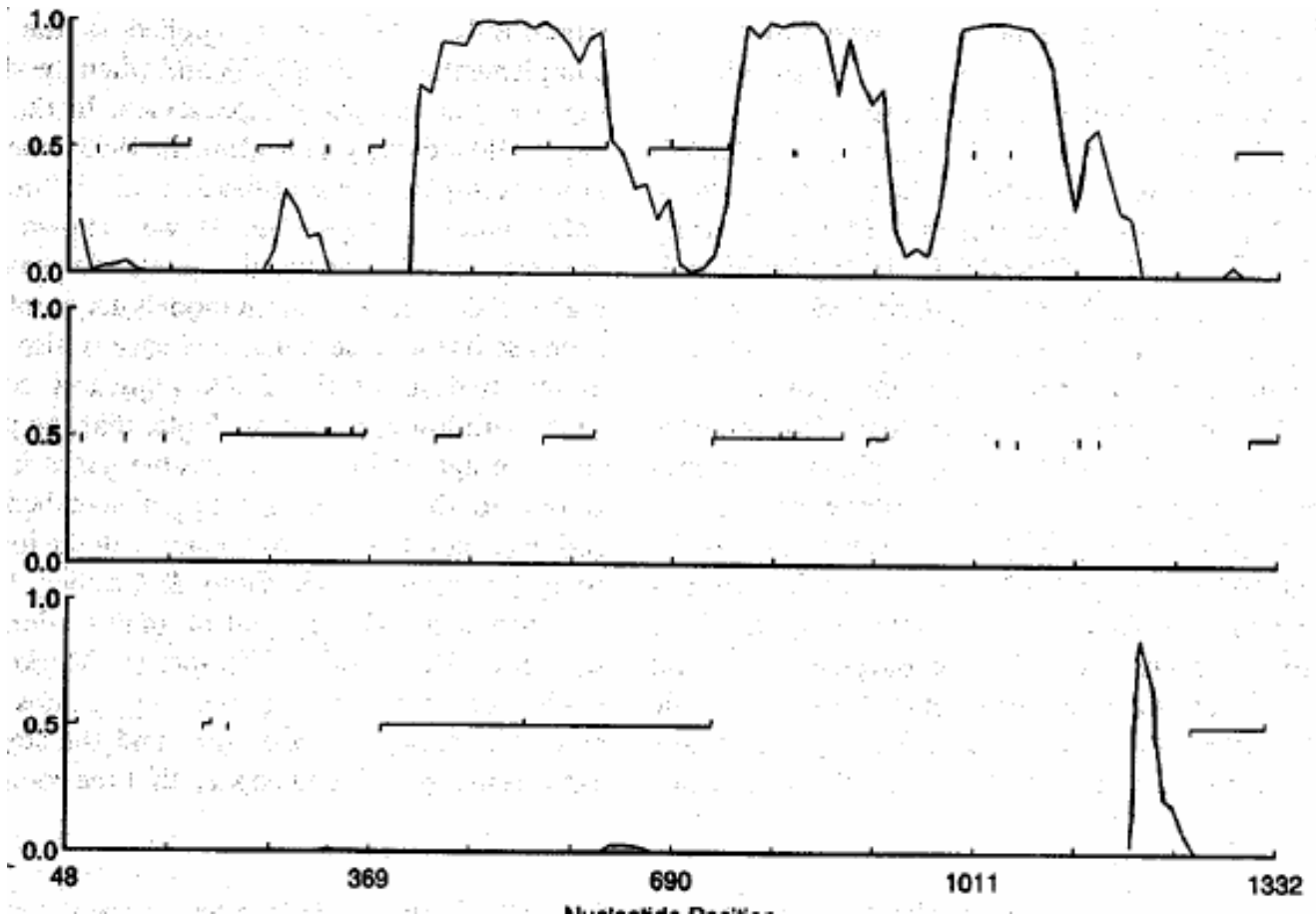
Prediction

NOTE 1: Since $a_{gene}(x_{i-1}, x_i)$ is less than one, a long product of $\prod_{i=2}^n a_{gene}(x_{i-1}, x_i)$ may cause under flow problem. So when we implement the program, we calculate and compare the probabilities in the log space. For example

$$\log(P_{gene}(X)) = \log(P_{gene}(x_1)) + \sum_{i=2}^{i=n} \log(a_{gene}(x_{i-1}, x_i))$$

NOTE 2: The test set contains one long sequence. We will divide the sequence into 100-letter fragments. Predictions are to be made for each fragment separately.

Prediction





K^{th} -Order Markov Chain

- When $K=2$ is used, the changes in the method include:
 - (1) The size of the transition probability table for each model will become 16×4 .

	A	C	T	G
AA				
AT				
AC				
AG				
TA				
TT				
TC				
TG				
...				



Kth-Order Markov Chain

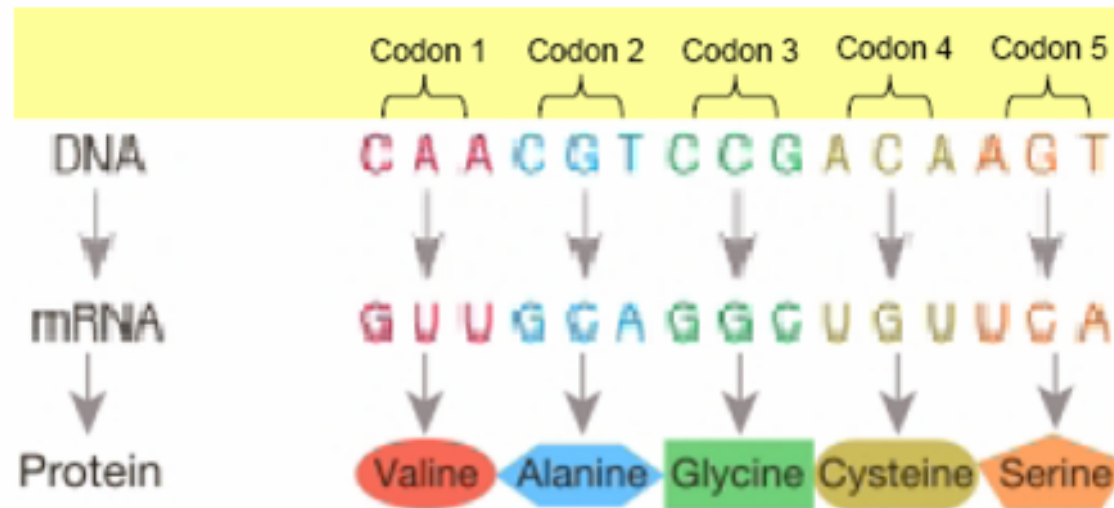
(2) The formula used in calculating the probabilities change to

$$\begin{aligned} P_{gene}(X) &= P_{gene}(x_1) \cdot P_{gene}(x_2 | x_1) \cdot P_{gene}(x_3 | x_2 x_1) \cdot \dots \cdot P_{gene}(x_n | x_{n-1} x_{n-2} \dots x_1) \\ &= P_{gene}(x_1) \cdot P_{gene}(x_2 | x_1) \cdot P_{gene}(x_3 | x_2 x_1) \cdot \dots \cdot P_{gene}(x_n | x_{n-1} x_{n-2}) \\ &\dots \\ &= P_{gene}(x_1 x_2) \cdot \prod_{i=3}^n a_{gene}(x_{i-2} x_{i-1}, x_i) \end{aligned}$$

When other value of K is used, similar changes apply.

Inhomogeneous Markov Chains

- When DNA is translated into proteins, three bases (the letters for DNA, which are A, T, G, and C) make up a codon and encode one amino acid residue (the letter for Proteins).





Inhomogeneous Markov Chains

- Each codon has three positions. In the previous models (referred as *homogeneous models*), we do not distinguish between the three positions. In this section, we will build different models (referred as *inhomogeneous models*) for different positions.

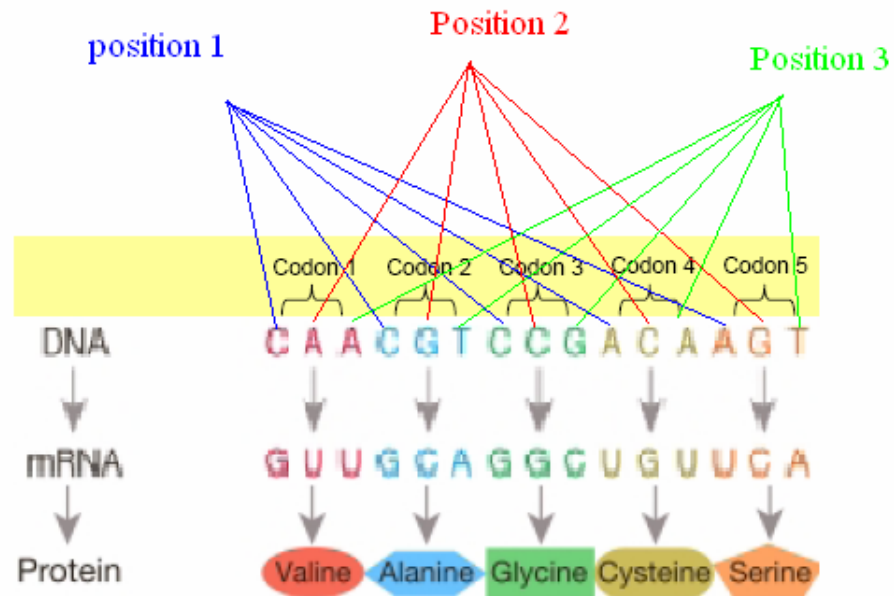
Gene Models

- The gene model will be split into three models, each for one codon position:

a_{gene_code1}

a_{gene_code2}

a_{gene_code3}





Gene Models

α_{gene_code1}

Gene Position 1	A	C	G	T
A				
C				
G				
T				

α_{gene_code2}

Gene Position 2	A	C	G	T
A				
C				
G				
T				

α_{gene_code3}

Gene Position 3	A	C	G	T
A				
C				
G				
T				



Gene Models

- In the prediction stage, for a given sequence $X = x_1x_2x_3 \dots x_n$, we will have to calculate three probabilities.

(1) The probability that X is a gene sequence starting with position 1.

$$P_{\text{gene_code1}}(X) = P_{\text{gene_code1}}(x_1) \cdot a_{\text{gene_code1}}(x_1, x_2) \cdot a_{\text{gene_code2}}(x_2, x_3) \cdot a_{\text{gene_code3}}(x_3, x_4) \cdot a_{\text{gene_code1}}(x_4, x_5) \cdot a_{\text{gene_code2}}(x_5, x_6) \cdot a_{\text{gene_code3}}(x_6, x_7) \dots$$

(2) The probability that X is a gene sequence starting with position 2.

$$P_{\text{gene_code2}}(X) = P_{\text{gene_code2}}(x_1) \cdot a_{\text{gene_code2}}(x_1, x_2) \cdot a_{\text{gene_code3}}(x_2, x_3) \cdot a_{\text{gene_code1}}(x_3, x_4) \cdot a_{\text{gene_code3}}(x_4, x_5) \cdot a_{\text{gene_code1}}(x_5, x_6) \cdot a_{\text{gene_code2}}(x_6, x_7) \dots$$

(3) The probability that X is a gene sequence starting with position 3.

$$P_{\text{gene_code3}}(X) = P_{\text{gene_code3}}(x_1) \cdot a_{\text{gene_code3}}(x_1, x_2) \cdot a_{\text{gene_code3}}(x_2, x_3) \cdot a_{\text{gene_code1}}(x_3, x_4) \cdot a_{\text{gene_code2}}(x_4, x_5) \cdot a_{\text{gene_code3}}(x_5, x_6) \cdot a_{\text{gene_code1}}(x_6, x_7) \dots$$



Complement Region Models

- We treat complement region the same way as we do genes. Three models will be built for three positions.
- Three probabilities will be calculated when prediction is to be made for a sequence $X = x_1x_2x_3 \dots x_n$,

(1) $P_{\text{complement_code1}}(X)$: The probability that X is a complement region starting with position 1.

(2) $P_{\text{complement_code2}}(X)$: The probability that X is a complement region starting with position 2.

(2) $P_{\text{complement_code3}}(X)$: The probability that X is a complement region starting with position 3.



Non-Coding Region Model

- Since the non-coding region does not contain codons, every position will be considered the same. There is no change to the non-coding region model. will be calculated as described in the homogeneous models.



Inhomogeneous Markov Chains

Then for each input sequence, we need to calculate and compare 7 probabilities:

$$P_{non-coding}(X)$$

$$P_{complement_code3}(X)$$

$$P_{complement_code2}(X)$$

$$P_{complement_code1}(X)$$

$$P_{gene_code1}(X)$$

$$P_{gene_code2}(X)$$

$$P_{gene_code3}(X)$$

