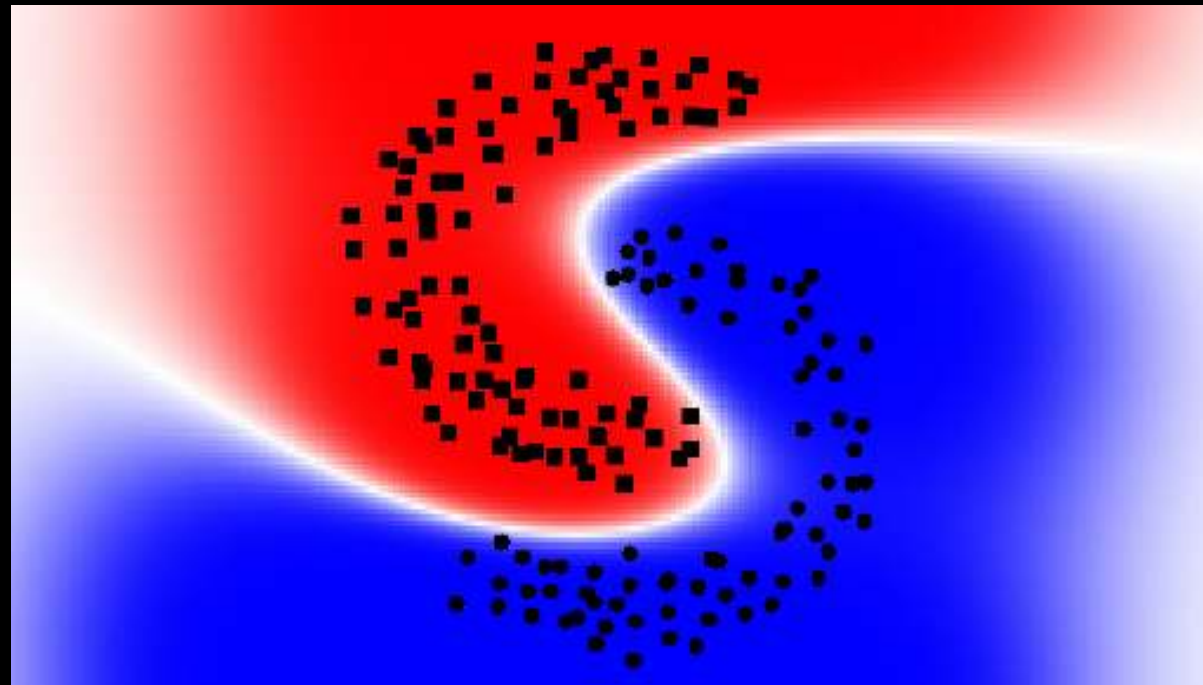


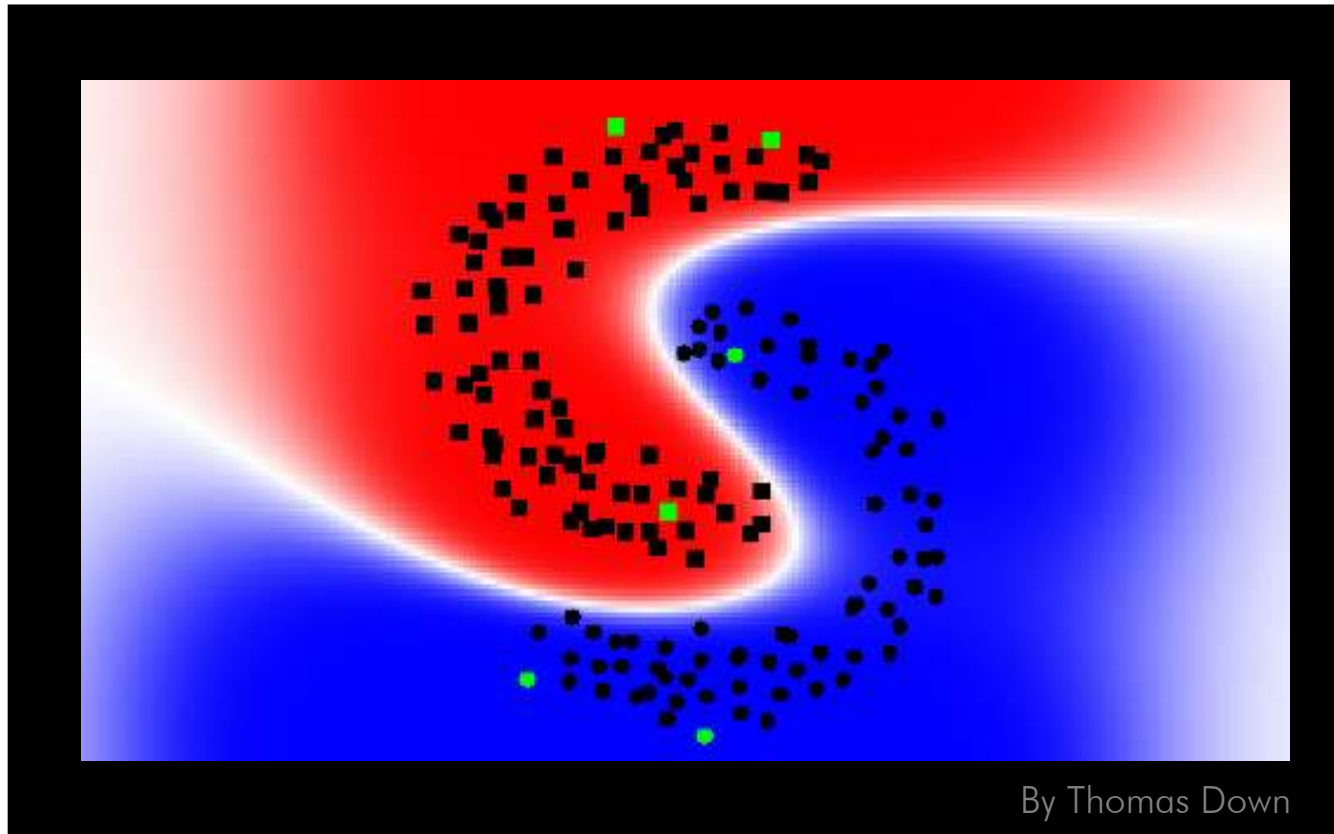
Αλγόριθμοι στη Μοριακή Βιολογία



<http://eclass.di.uoa.gr/courses/D461/>

By Thomas Down

Αλγόριθμοι στη Μοριακή Βιολογία



... και εσύ ποιός είσαι?

... και εσύ ποιός είσαι?

Γιώργος Σ. Βερνίκος, PhD
Comparative Genomics

... και εσύ ποιός είσαι?



<http://www.sanger.ac.uk/>



**ερν
e Ge**

<http://bioinformatics.biol.uoa.gr/>



<http://www.cam.ac.uk/>

... και εσύ ποιός είσαι?

GeneVito
File Edit View Tools Annotation Go To Colors Links Help

GENOME : Methanococcus jannaschii.

Gene Position on Circular Map :

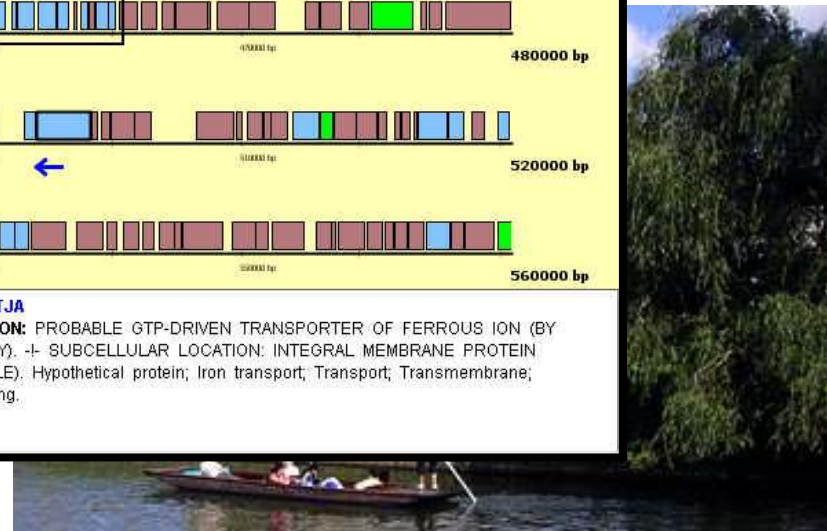
Protein ID: FEOB_METJA
SubCellular Location: INTEGRAL-MEMBRANE
PREDclass category: MEMBRANE
Enzyme class: NOT_AVAILABLE
Number of tm's (PRED-TMR2): 8
Number of tm's (SwissProt): 8

FEOB_METJA
-I- FUNCTION: PROBABLE GTP-DRIVEN TRANSPORTER OF FERROUS ION (BY SIMILARITY). -+ SUBCELLULAR LOCATION: INTEGRAL MEMBRANE PROTEIN (PROBABLE). Hypothetical protein; Iron transport; Transport; Transmembrane; GTP-binding.

```
GGAGGGCTTGCCTCTGCTTTGGTGTGTTGGGATAAATAGCTAAGGAGGTAGTTGTTGG  
AAGTTTGGCAATGTTATATGGGACTGGAGAGGAAAATCTCTCATCTGTTATTGGCT  
1815 CATGCATTCTCTCCAGTATCTGCCTATGCATTTATGGCATTTCCTTTAATTTACC  
1870 TCCCATGTATTGCAACATTAGCAGTTATAAAGCAAGAAATGGGGTGGAAATGGGC  
1925 GTTATTGTCAGTAACTTATGAGATGATATTAGCTTATGTTGTAGCTTTGGTAAATC  
1980 TCCGTTATTGGAATCTATTATTTAA  
2035
```

<http://www>

<http://www.cam.ac.uk/>



... και εσύ ποιός είσαι?

The screenshot displays the GeneVito software interface. On the left, a sidebar shows a tree view of annotations for the protein SYG_METJA, including SubCellular Location (CYTOPLASM), PREDclass category (GLOBULAR), and Enzyme class (Ligases(6)). The main window shows a gene map with exons represented by blue boxes and introns by lines. A red arrow points to a specific exon. A 'Motif Search Results' window is open, listing 91 proteins that contain the motif 'M...[VP]...[ST]'. The protein SYG_METJA is highlighted in the list. Below the gene map, the DNA sequence is shown, with the start codon ATG highlighted in red. The sequence is: ATCAAATTGCCCTATAAAAAGCTTATGTTCTGCGGTTGGTTAATAAAGATGATATG... (lines 18-20). The protein description for SYG_METJA is: -!- CATALYTIC ACTIVITY: ATP + L-GLYCINE + TRNA(GLY) = AMP + PYROPHOSPHATE + L-GLYCYL-TRNA(GLY) (BY SIMILARITY). -!- SUBCELLULAR LOCATION: CYTOPLASMIC.-!- SIMILARITY: BELONGS TO CLASS-II AMINOACYL-TRNA SYNTHETASE FAMILY. Aminoacyl-tRNA synthetase; Protein biosynthesis; Ligase; ATP-binding.

<http://www>

<http://www.cam.ac.uk/>

... και εσύ ποιός είσαι?

GeneVito

File Edit View Tools Annotation Go To Colors Links Help

GENOME : *Chlamydia trachomatis*.

Gene Position on Circular Map :

1 bp 40000 bp 80000 bp 120000 bp 160000 bp 200000 bp 240000 bp 280000 bp

Cut Position: 91603
Recognition name: a/agctt

SubCellular Location: NOT_AVAILABLE

PREDclass category: GLOBULAR

Enzyme class: Transferases(2.)

Number of tm's (PRED-TMR2): NOT_AVAILABLE

Number of tm's (SwissProt): NOT_AVAILABLE

GTGTTTAAGCACGCGAGTGGCGACGAAATGGAGTTTGGCTGCCAAGACTTGCATTACAG
1540 715 AAGCTGGGCTGCAAGAAAAAGATATAGATTGGTTAGTTCCCTCATCAGGCAAATGA
1595 770 GCGTATTATCGATGCTATTGCAAAAACGTTTGGCTGTTAAAGACTCTCGGGTATTT
1850 825 AAAACTCTTGGCTAAGTATGGTAACACAGCAGCCTCTTCTGTGGGGGATTGCTTTAG
1705 880 ACGAACTTTTACGTACACATGATATCCATGTTGCGGAGCGGTTGTTGTTAGTAGC
1760 935 TTTTGGGGGAGGCTTATCTTGGGGAGCAGTGATTTTACAGCAAGTGTAA
990

IN GOVERNING THE TOTAL RATE OF FATTY ACID PRODUCTION, POSSESSES BOTH ACETOACETYL-ACP SYNTHASE AND ACETYL TRANSACYLASE ACTIVITIES (BY SIMILARITY). -I- **CATALYTIC ACTIVITY:** ACYL-[ACYL-CARRIER PROTEIN] + MALONYL-[ACYL-CARRIER PROTEIN] = 3-OXOACYL-[ACYL-CARRIER PROTEIN] + CO(2) + [ACYL-CARRIER PROTEIN]. -I- **PATHWAY:** FATTY ACID BIOSYNTHESIS.-I- **SIMILARITY:** BELONGS TO THE FABH FAMILY. Fatty acid biosynthesis; Transferase; Acyltransferase; Multifunctional enzyme.

<http://www>

... και εσύ ποιός είσαι?

The screenshot displays the GeneVito software interface. The main window shows a genome map for *Chlamydia trachomatis* with a scale from 0 to 40,000 bp. A detailed view of a gene is shown below, with a scale from 160,000 to 280,000 bp. The gene's enzyme class is listed as Transferases(2.). The number of transmembrane domains (tm's) is noted as NOT_AVAILABLE. The amino acid sequence is shown with a red stop codon (TAA) at position 935. The gene's function is described as governing the total rate of fatty acid production, possessing both acetoacetyl-ACP synthase and acetyl transacylase activities.

email: gsv@sanger.ac.uk
gvernikos@gmail.com
e-class: <http://eclass.di.uoa.gr/courses/D461/>

<http://www>

Bioinformatics Godfathers

Sanger



NCBI



EBI



Copenhagen Uni

Howard Hughes Medical Institute



Sanger



EBI



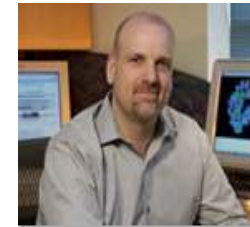
Stockholm Uni



Memorial Sloan-Kettering cancer centre



CNIO Madrid



National Human Genome Research Institute

Society: <http://www.iscb.org/>

Bioinformatics Godfathers

Sanger



NCBI



EBI



Howard Hughes Medical Institute



Copenhagen Uni

Sanger



EBI



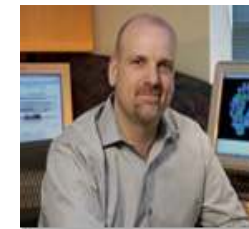
CNIO Madrid



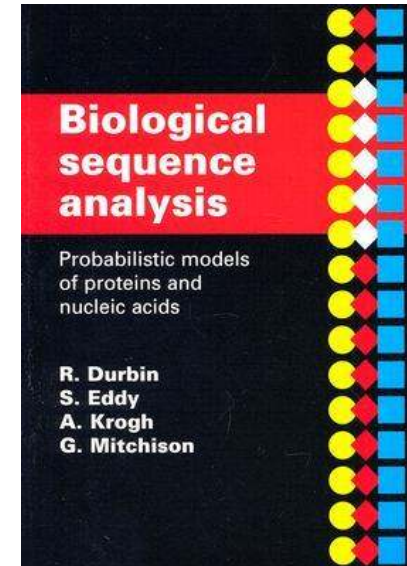
Stockholm Uni



Memorial Sloan-Kettering cancer centre



National Human Genome Research Institute



Society: <http://www.iscb.org/>

Bioinformatics Godfathers

Sanger



NCBI



EBI



Howard Hughes Medical Institute



Copenhagen Uni

Sanger



EBI



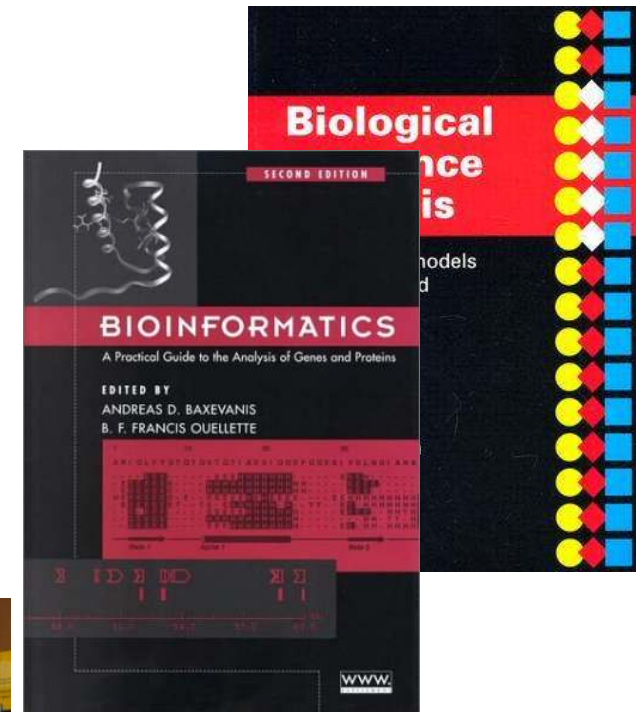
Stockholm Uni



Memorial Sloan-Kettering cancer centre



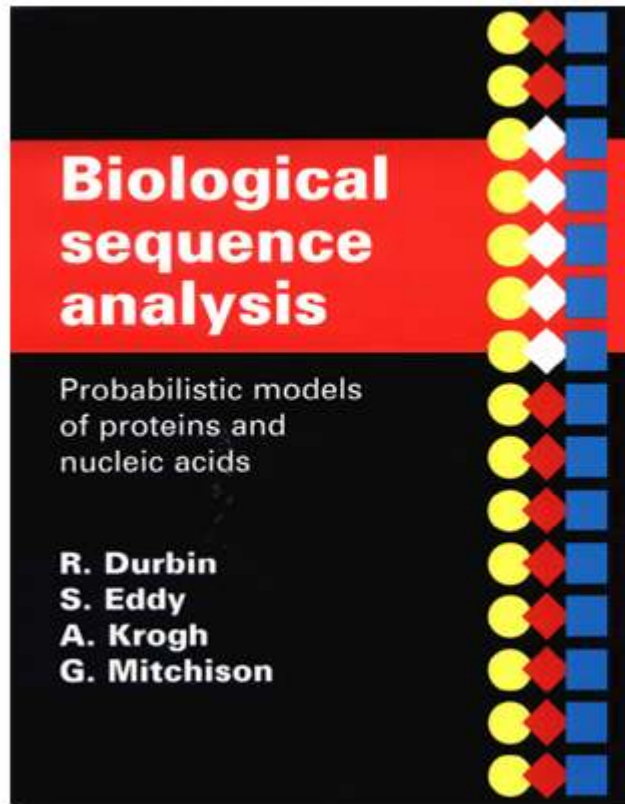
CNIO Madrid



National Human Genome Research Institute

Society: <http://www.iscb.org/>

Biological Sequence Analysis



1. Sequence Alignment
2. Markov Chains & Hidden Markov Models
3. Phylogenetic Trees

4. Sparse Bayesian Learning & The Relevance Vector Machine



<http://eclass.di.uoa.gr/courses/D461/>
Υλικό Μαθήματος/Βιβλιογραφία

Bioinformatics ... at the high school level

A First Attempt to Bring Computational Biology into Advanced **High School** Biology Classrooms

(<http://dx.doi.org/doi:10.1371/journal.pcbi.1002244>)

Suzanne Gallagher, William Coon, Kristin Donley, Abby Scott, Debra Goldberg. PLoS Comput Biol, Vol. 7, No. 10. (27 October 2011), e1002244.

Ten Simple Rules for Teaching Bioinformatics at the **High School** Level

(<http://dx.doi.org/doi:10.1371/journal.pcbi.1002243>)

David Form, Fran Lewitter. PLoS Comput Biol, Vol. 7, No. 10. (27 October 2011), e1002243.

Teaching Bioinformatics at the **Secondary School** Level

(<http://dx.doi.org/doi:10.1371/journal.pcbi.1002242>)

Fran Lewitter, Philip Bourne. PLoS Comput Biol, Vol. 7, No. 10. (27 October 2011), e1002242.



ctacgaatccagcccggttacgatcatagctatttaaggaatatttattgctccttcccttccggcggcgaacgccagcgcgtcatgatagccatggcgctga
ttgaggaccatctgcgctttcatgcagcgcgtcatgtatcgattgccacaaatgtctttaccggctcgtacatcgcagcaatgatctcatcgagcgaataca
gggttcgcttacgcgactcatcgccatcgtcgccgctttatcgctttcacggcagagatggcatttctctctatgcaggaatttgtaaccgaccagttcagc
gtatgcaggccgcgataaagcatccgtagctcagctggatagagtactcggctacgaaccgagcggcggaggttcgaatcctcccggatgcaccagc
tgcatacgtcccataatttgcgccagataattctgcaacgtctgtggcgagtggcggcacagctcattttcatcatcagcgcgccacggacaggccgta
cgccgacacagcgcagcaactgcgcggcatgactaaattcaaagggcatgctgcatcttctgcggtcgtcgcgaagaggtgtgccgaacggaga
tcgacattctccggctggcaaccagcagcccaagatacagggcgggtatccacattgtggcacttgcggccttaacgataacgcgccatacagctcgattt
gatgggctgcaggccgagcgtgagcagggcattactattgacgtcgcgggtcgtctaaaagcgcgggtatcaactggattatgccttttcgaatacgggtga
aatacgcgataacggtctgacggataacattacctgcccgattatgaacgccgcaaaaaaacagacgctgggtgacctggctgcccggttaaatatt
gccacggagaatattattgcctgcgggtgatggcgccaacgatctgattcccgtggcgttaactgtagataaaatgactgaactgcgtgccgatgttagaaca
tgctggcaccggtattgcctggaaagcgaagccggtgtacgggaaaaatccaccatcagattaattatcacggttcgaattgcttcttttctattgaaga
ggcgtattatcttcggtcgtgcgaccccggtagagctgtaccattaaagatcgtcccggtagcttaggttgcccaacgtggggcagtggcgtaacgacc
cgacggaaactggcccaggtaaggccagcccggtaacctgagcttctggcagatTTTTGTGAAATATCTGACCAATCCACTGGTATGGATCATTATT
ATCTTTATCCTGGTGAAGGGGACTCCGCGGGCGGCCACGTTCA

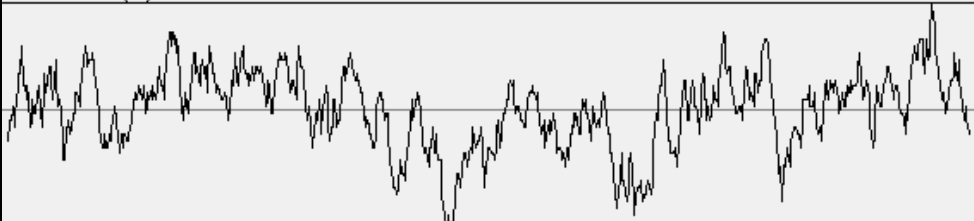
ctacgaatccagcccggttacgatcatagctatttaaggaatthttattgctccttcctttccggcggcgaacgccagcgcgtcatgatagccatggcgcgtga
ttgaggaccatctgcgctttcatgcagcgcctcatgatcgcattgccacaaatgctttaccggtctcgtacatcgcagcaatgatctcatcgagcgaataca
gggttcgcttacgcgactcatcgccatcgtcgccgctttatcgctttcacggcagagatggcatttctctctatgcaggaatttgtaaccgaccagttcagc
gtatgcaggccgcgataaagcatccgtagctcagctggatagagtactcggctacgaaccgagcggctcggaggttcgaatcctcccggatgcaccagc
tgcatacgtcccataattgcgccagataattctgcaacgtctgtggcgagtggcggcacagctcattttcatcatcagcgcgccacggacaggccggtta
cgccgacacagcgcagcaactgcgcgcatgactaaattcaaaggcatgcgtgcatcttctcggctcgtcgaagaggtgtgcccgaacggaga
tcgacattctccggctggcaaccagcagcccaagatacagggcgggtatccacattgtggcactgctggcgttaacgataacgcgccatacagctcgatttt
gatgggctgcaggccgagcgtgagcagggcattactattgacgtcgcggtgctgctaaaagcgcgggtatcaactggattatgccttttcgaatacgggtga
aatacgcgataacggtctgacggataacattacctgcccgattatgaacgccgcaaaaaaacagacgctggtgacctggctgcccgggttaaatatt
gccacggagaatattattgcctgcggtgatggcgccaacgatctgattcccgtggcgttaactgtagataaaaatgactgaactgcgtgccgatgttagaaca
tgctggcaccggtattgcctggaaagcgaagccggtgtacgggaaaaaatccaccatcagattaattatcacggtttcgaattgcttcttttcttattgaaga
ggcgtattatcttcggctcgtcgcaccccggtagagctgtaccattaaagatcgtcccggtagcttaggttgcccaacgtggggcagtggcgtaacgacc
cgacggaactggcccagggtcaaggccagcccgggtcaacctgagcttctggcagattttgtgaaatatactgaccaatccactggtatggatcattattat
atctttatcctggtggaaggggactccgccccggcgcccacggtcag

RBS

ABC transporter

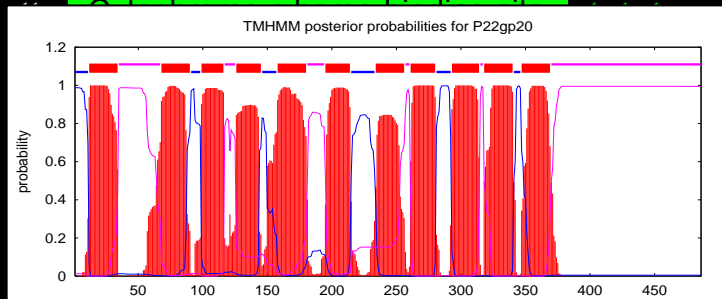
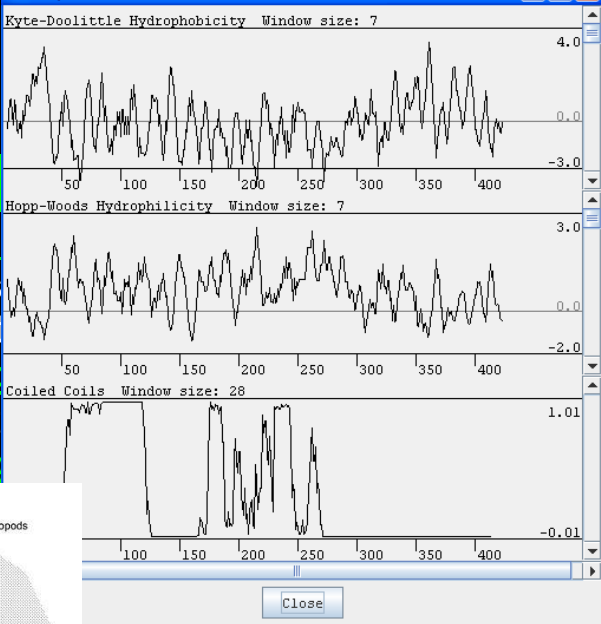
ctaccaatccaacccaattacaatcataactatttaaggaatthattgcgctcttcctttccggcggcgaacgccagcgcgcatgatagccatggcgctga
 ttg: **Cytochrome c heme-binding site** atgtatcgattgccacaaatgtctttaccctctctgtacatcgcagcaatgatctcatcgagcgaataca
 gggttcgcttacgcgactcatcgccatcgtcgccgctttatcgctttcacggcagag **tRNA** ttctctatgcaggaattgtaccgaccagttcagc
 gtatgcaggccgcgataaa **gcatccgtagctcagctggatagagtactcggctacgaaccgagcggctcggaggttcgaatcctccggatgcaccagc**
 tgcatacgtcccatatttcccccagataattctcccacctctctcccccagctccggcacagctcattttcatcatcagcgcgccacggacaggccgta
 cgccgacacagcgc **4Fe-4S ferredoxins, iron-sulfur binding region** gtgcatcttctcggctcgtcgaagaggtgtgcccaacggaga
 tcgacattctccggctggcaaccagcagcccaagatacagggcgggtatccacattctccacttccacttaacnataacgcgccatacagctcgatttt
gatgggctgcaggccgagcgtgagcagggcattactattgacgtcgcggtc **GTP-binding elongation factor** ttatgccttttcgaatacgggtga
 aatacgcgataacottctgacccgataacattacctccgattatgaacgcgcgcaaaaaaacagacgctggtgacctggctgccgggtaaatatt
 gccacggagaata **Ribosomal protein L10 signature** ctgattcccgtggcgtaactgtagataaaatgactgaactcgtgccgatgttagaaca
 tgctggcaccggtattgctggaaagcgaagccggtgtacggcaaaaatccaccatcagattaattatcacggtttcgaattgctcttttcttattgaaga
 ggcgctatt **atcttcggctcgtcgcgaccccggtagagctgtac** **Transcription termination factor signature** tggggcagtggcgtaacgacc
 cgacggaactggcccaggtcaaggccagcccggtaacctaactctcacaattttttataaatatctctgaccaatccactggtatggatcattattat
 atctttatc **ctgggtggaaggggactccgcccggcgccc** **DNA topoisomerase II signature**

GC Content (%) Window size: 71



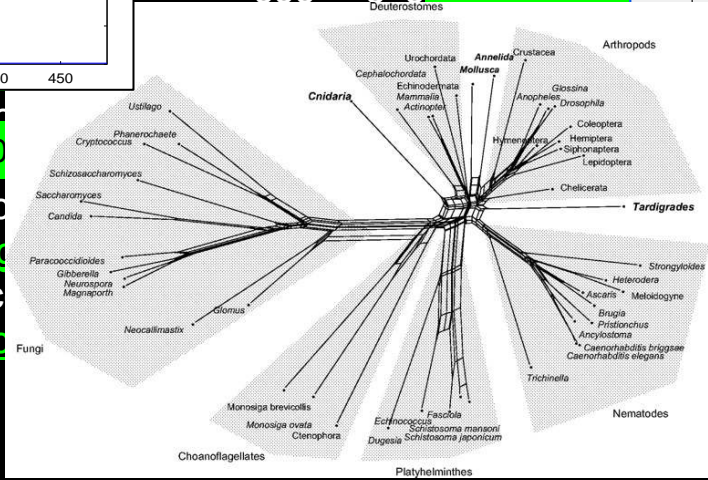
ctacacaaacacaccccccacacacacacacacacatattt**aaagga**atttattgtgcgctctt**ctttccggcgggcga**

Graphs for: STY4090

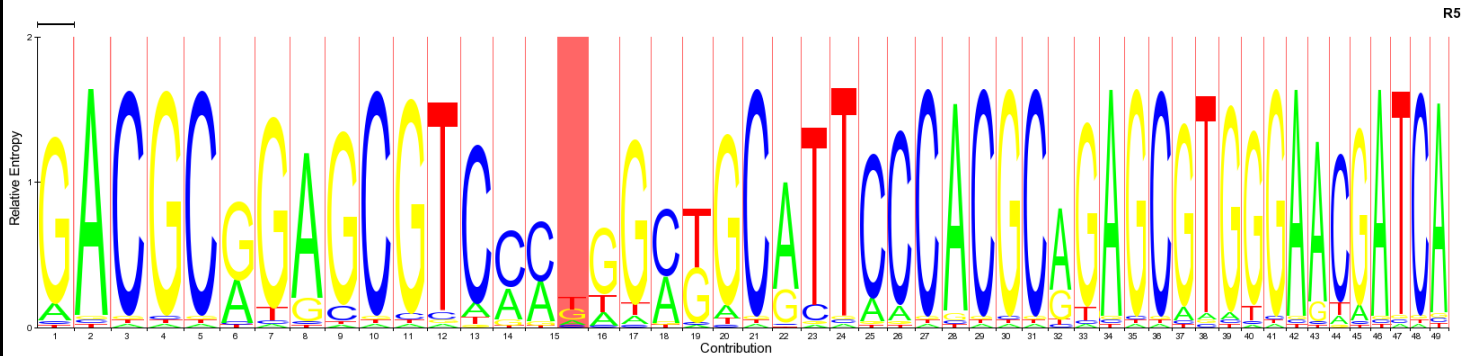


tgccacaaatgtcttaccnctctgtact
 tcgctttcacggcagag **tRNA** ttctct
tagagtactcggctacgaaccgagcga
 ctatagccacttgcgcacagctcatttt
fur binding region gttgcattctgccc

aatacgggataacottctgaccgataacattar
 gccacggagaata **Ribosomal protein L10**
 tgctggcaccggtattgacctggaagcgaagc
 ggcgctatt**atcttcgggtcgtcgacccccggta**
 cgacggaactggcccaggtcaaggccagccc
 atctttatc**ctgggtggaaggggactccgctgg**



tgactgaactcgtgccgatgtagaaca
 acnctttcgaattgctctttttctattgaaga
 ire **gtggggcagtggcgtaacgacc**
tgaccaatccactggtatggatcattattat



R5

Εισαγωγή – Πιθανότητες

What is a probabilistic model?

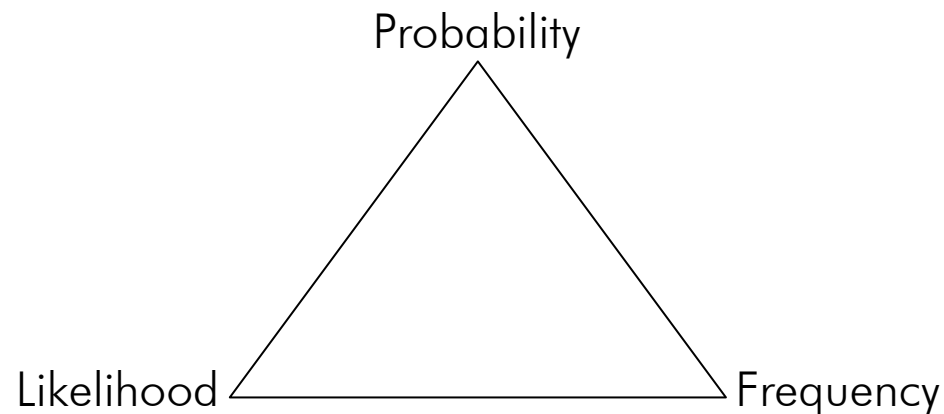
- A system that **simulates** the **object** under consideration.
- Produces **different outcomes** with **different probabilities**.
- A probabilistic **model** can therefore simulate a whole class of **objects**.
- In this context, the **objects** will be **sequences**, and a **model** might describe a **family** of related sequences.

Εισαγωγή – Πιθανότητες

A *probability* gives the *odds of an event, given any parameters*: Given that the mean is zero and the variance one, what are the odds that the draw will be between 1.1 and 1.2?

A *likelihood* gives *the odds of parameters given data*: We drew a 1.3 from the distribution; what are the odds that the mean is zero?

Observed *frequencies* are *estimates* of *probabilities*.



6-αρεις ...!

A familiar probabilistic **system** with a set of **discrete** outcomes is the roll of a six-sided die. A model of a roll of a (possibly loaded) die would have six parameters:

$$p_1 \dots p_6$$

and the probability of rolling i is p_i .

The parameters p_i must satisfy the conditions:

$$p_i \geq 0 \quad \sum_{i=1}^6 p_i = 1$$

A model of a sequence of **three** consecutive **rolls** of a die might be that they were all **independent**, so that the probability of sequence **[1,6,3]** would be the product of the individual probabilities:

$$p_1 p_6 p_3$$

Maximum Likelihood Estimation

What are biological sequences?

Strings from a finite alphabet of residues, generally either four nucleotides (DNA) or twenty amino acids (Proteins).

Assuming that a residue a occurs at random with probability q_a independent of all other residues in the sequence, and the (protein or DNA) sequence is denoted:

$x_1 \dots x_n$,

the probability of the whole sequence is :

$$\prod_{i=1}^n q_{x_i}$$

Maximum Likelihood Estimation

What are biological sequences?

Strings from a finite alphabet of residues, generally either four nucleotides (DNA) or twenty amino acids (Proteins).



Assuming that a residue a occurs at random with probability q_a independent of all other residues in the sequence, and the (protein or DNA) sequence is denoted:

$x_1 \dots x_n$,

the probability of the whole sequence is :

$$\prod_{i=1}^n q_{x_i}$$



Residues occur at random?

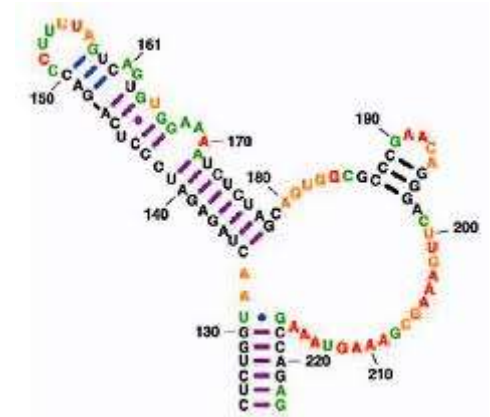
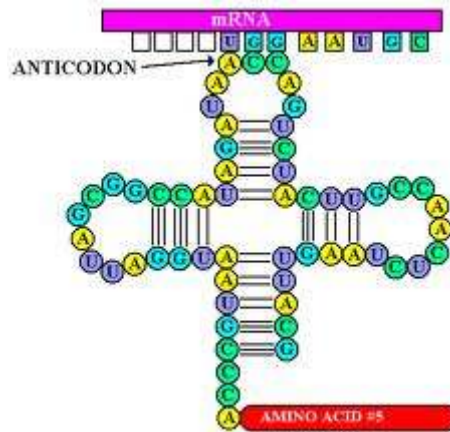
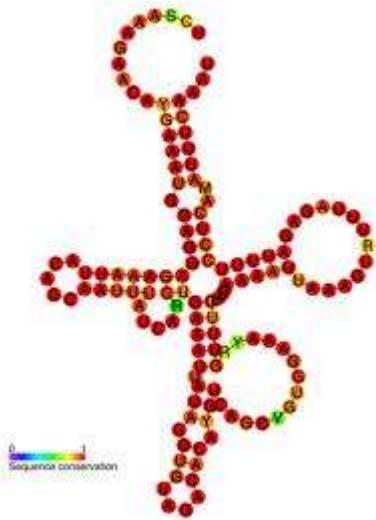
Salmonella enterica subsp. enterica serovar Typhi str. CT18 [gbbct]: 368 CDS's (89135 codons)

fields: [triplet] [amino acid] [fraction] [frequency: per thousand] ([number])

UUU	F	0.51	20.3	(1811)	UCU	S	0.17	11.7	(1043)	UAU	Y	0.51	16.2	(1440)	UGU	C	0.43	5.5	(491)
UUC	F	0.49	19.7	(1759)	UCC	S	0.15	10.8	(964)	UAC	Y	0.49	15.7	(1399)	UGC	C	0.57	7.3	(648)
UUA	L	0.11	10.7	(954)	UCA	S	0.19	13.3	(1186)	UAA	*	0.50	2.1	(184)	UGA	*	0.38	1.6	(140)
UUG	L	0.13	12.5	(1114)	UCG	S	0.13	9.5	(844)	UAG	*	0.12	0.5	(44)	UGG	W	1.00	12.4	(1105)
CUU	L	0.16	15.2	(1356)	CCU	P	0.24	9.5	(849)	CAU	H	0.53	11.2	(994)	CGU	R	0.26	14.7	(1310)
CUC	L	0.14	12.8	(1144)	CCC	P	0.15	6.0	(533)	CAC	H	0.47	9.7	(865)	CGC	R	0.31	17.1	(1523)
CUA	L	0.06	5.4	(483)	CCA	P	0.25	9.9	(880)	CAA	Q	0.33	12.4	(1105)	CGA	R	0.12	6.7	(598)
CUS	L	0.39	36.6	(3263)	CCG	P	0.37	14.8	(1316)	CAG	Q	0.67	25.0	(2230)	CGG	R	0.14	7.8	(696)
AUU	I	0.43	24.8	(2212)	ACU	T	0.23	13.1	(1171)	AAU	N	0.48	21.4	(1910)	AGU	S	0.15	10.7	(951)
AUC	I	0.43	24.5	(2182)	ACC	T	0.31	17.8	(1591)	AAC	N	0.52	23.1	(2060)	AGC	S	0.21	15.0	(1333)
AUA	I	0.14	8.0	(713)	ACA	T	0.21	12.1	(1078)	AAA	K	0.59	34.5	(3074)	AGA	R	0.10	5.6	(503)
AUG	M	1.00	27.4	(2438)	ACG	T	0.26	15.2	(1351)	AAG	K	0.41	23.7	(2116)	AGG	R	0.07	4.1	(363)
GUU	V	0.31	20.7	(1847)	GCU	A	0.22	18.1	(1613)	GAU	D	0.56	31.1	(2772)	GGU	G	0.29	18.2	(1625)
GUC	V	0.24	16.2	(1447)	GCC	A	0.28	23.1	(2062)	GAC	D	0.44	24.8	(2214)	GGC	G	0.35	22.6	(2012)
GUA	V	0.17	11.6	(1033)	GCA	A	0.25	20.3	(1813)	GAA	E	0.57	37.7	(3356)	GGA	G	0.17	11.0	(982)
GUG	V	0.28	18.6	(1657)	GCG	A	0.25	20.5	(1828)	GAG	E	0.43	28.0	(2497)	GGG	G	0.19	11.9	(1060)

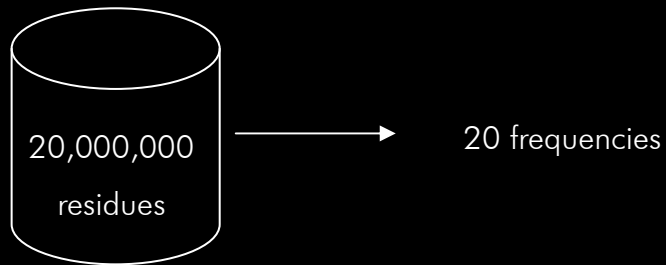


Residues occur at random?



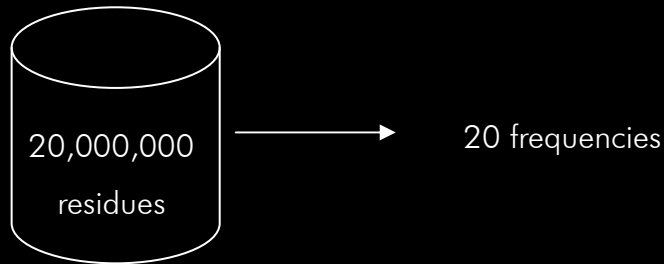
Maximum Likelihood Estimation

The **parameters** for a **probabilistic** model are typically **estimated** from large sets of trusted examples, often called a **training set**. For instance, the probability q_a for amino acid a can be estimated as the **observed** frequency of residues in a database of known protein sequences, such as UNI-PROT:



Maximum Likelihood Estimation

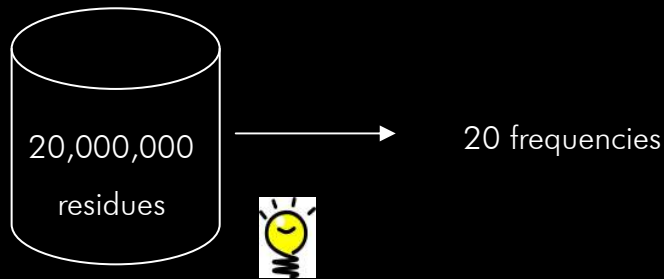
The **parameters** for a **probabilistic** model are typically **estimated** from large sets of trusted examples, often called a **training set**. For instance, the probability q_a for amino acid a can be estimated as the **observed** frequency of residues in a database of known protein sequences, such as UNI-PROT:



Having so much data that, we expect the **frequencies** to be reasonable **estimates** of the underlying **probabilities** of our model. This way of estimating models is called **maximum likelihood estimation (MLE)**.

Maximum Likelihood Estimation

The **parameters** for a **probabilistic** model are typically **estimated** from large sets of trusted examples, often called a **training set**. For instance, the probability q_a for amino acid a can be estimated as the **observed** frequency of residues in a database of known protein sequences, such as UNI-PROT:



3 flips [tail, tail, tail]

$MLE_{\text{heads}} = ?$

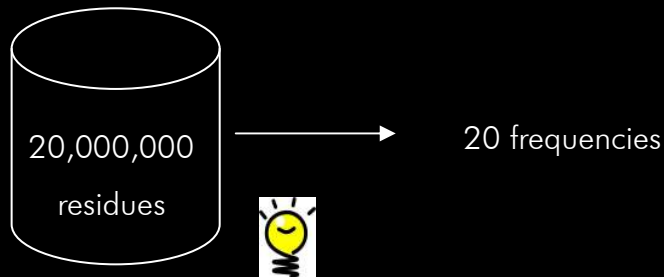


$MLE_{\text{tail}} = ?$

Having so **much** data that, we expect the **frequencies** to be reasonable **estimates** of the underlying **probabilities** of our model. This way of estimating models is called **maximum likelihood estimation (MLE)**.

Maximum Likelihood Estimation

The **parameters** for a **probabilistic** model are typically **estimated** from large sets of trusted examples, often called a **training set**. For instance, the probability q_a for amino acid a can be estimated as the **observed** frequency of residues in a database of known protein sequences, such as UNI-PROT:



3 flips [tail, tail, tail]

$MLE_{\text{heads}} = ?$



$MLE_{\text{tail}} = ?$

Having so **much** data that, we expect the **frequencies** to be reasonable **estimates** of the underlying **probabilities** of our model. This way of estimating models is called **maximum likelihood estimation (MLE)**.

It can be shown that using the **frequencies** with which the amino acids occur in the database as the **probabilities** q_a maximizes the total **probability of all the sequences given the model** (the likelihood).

Given a model with parameters θ and a set of data D , the maximum likelihood estimate for θ is that value which maximizes $P(D | \theta)$.

Conditional, joint and marginal probabilities



D_1



D_2

The probability of rolling i with die D_1 is called:

$P(i \mid D_1)$, the *conditional* probability of rolling i given die D_1 .

Conditional, joint and marginal probabilities



D_1



D_2

The probability of rolling i with die D_1 is called:

$P(i | D_1)$, the *conditional* probability of rolling i given die D_1 .

Picking a die at random with probability $P(D_j)$, the probability of picking die j and rolling an i is:

$P(i, D_j) = P(D_j) P(i | D_j)$, the *joint* probability.

Conditional, joint and marginal probabilities



D_1



D_2

The probability of rolling i with die D_1 is called:

$P(i | D_1)$, the *conditional* probability of rolling i given die D_1 .

Picking a die at random with probability $P(D_j)$, the probability of picking die j and rolling an i is:

$P(i, D_j) = P(D_j) P(i | D_j)$, the *joint* probability.

More generally, we can write: $P(X, Y) = P(X | Y) P(Y)$

If we know the *conditional* and *joint* probabilities, we can calculate a *marginal* probability:

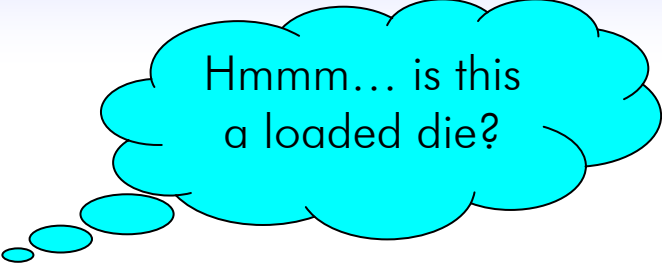
$$P(X) = \sum_Y P(X, Y) = \sum_Y P(X | Y) P(Y)$$

Bayes' theorem



"Consider an occasionally dishonest casino that uses two kinds of dice. Of the dice **99%** are **fair** but **1%** are **loaded** so that a **six** comes up **50%** of the time. We pick up a die from a table at random."

Roll	Outcome
1	Six
2	Six
3	Six



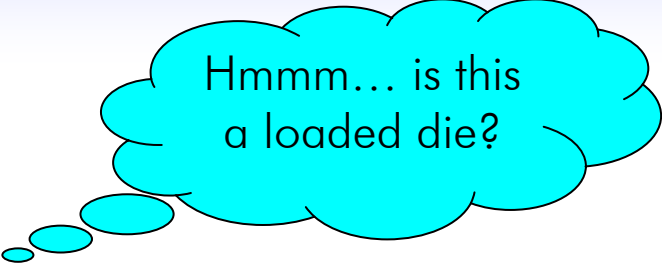
Hmmm... is this a loaded die?

Bayes' theorem



“Consider an occasionally dishonest casino that uses two kinds of dice. Of the dice **99%** are **fair** but **1%** are **loaded** so that a **six** comes up **50%** of the time. We pick up a die from a table at random.”

Roll	Outcome
1	Six
2	Six
3	Six



Hmmm... is this a loaded die?

The probability that we are after, is:

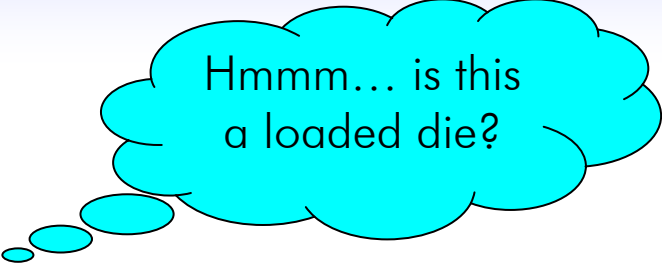
$P(D_{\text{loaded}} \mid 3 \text{ sixes})$, the *posterior probability* of the hypothesis that the die is loaded, given the observed data.

Bayes' theorem



“Consider an occasionally dishonest casino that uses two kinds of dice. Of the dice **99%** are **fair** but **1%** are **loaded** so that a **six** comes up **50%** of the time. We pick up a die from a table at random.”

Roll	Outcome
1	Six
2	Six
3	Six



Hmmm... is this a loaded die?

The probability that we are after, is:

$P(D_{\text{loaded}} \mid 3 \text{ sixes})$, the *posterior probability* of the hypothesis that the die is loaded, given the observed data.

However, what we can directly calculate is the probability of the data given the hypothesis: $P(3 \text{ sixes} \mid D_{\text{loaded}})$, the *likelihood* of the hypothesis.

Bayes' theorem



"Consider an occasionally dishonest casino that uses two kinds of dice. Of the dice **99%** are **fair** but **1%** are **loaded** so that a **six** comes up **50%** of the time. We pick up a die from a table at random."

Roll	Outcome
1	Six
2	Six
3	Six

Hmmm... is this a loaded die?

The probability that we are after, is:

$P(D_{\text{loaded}} \mid 3 \text{ sixes})$, the *posterior probability* of the hypothesis that the die is loaded, given the observed data.

However, what we can directly calculate is the probability of the data given the hypothesis: $P(3 \text{ sixes} \mid D_{\text{loaded}})$, the *likelihood* of the hypothesis.

Knowing the *likelihood* we can calculate the *posterior probabilities*, using the *Baye's theorem*:

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

Bayes' theorem



In the case of our die, Baye's theorem can be written:

$$P(D_{loaded} | 3sixes) = \frac{P(3sixes | D_{loaded})P(D_{loaded})}{P(3sixes)}$$

Bayes' theorem



"Consider an occasionally dishonest casino that uses two kinds of dice. Of the dice 99% are fair but 1% are loaded so that a six comes up 50% of the time. We pick up a die from a table at random."

In the case of our die, Baye's theorem can be written:

$$P(D_{loaded} | 3sixes) = \frac{P(3sixes | D_{loaded})P(D_{loaded})}{P(3sixes)}$$

We have been told, that $P(D_{loaded}) = 0.01$

and that $P(3sixes | D_{loaded}) = 0.5^3 = 0.125$

Bayes' theorem



"Consider an occasionally dishonest casino that uses two kinds of dice. Of the dice 99% are fair but 1% are loaded so that a six comes up 50% of the time. We pick up a die from a table at random."

In the case of our die, Baye's theorem can be written:

$$P(D_{loaded} | 3sixes) = \frac{P(3sixes | D_{loaded})P(D_{loaded})}{P(3sixes)}$$

We have been told, that $P(D_{loaded}) = 0.01$

and that $P(3sixes | D_{loaded}) = 0.5^3 = 0.125$

Thus:

$$P(D_{loaded} | 3sixes) = \frac{(0.5^3)(0.01)}{(0.5^3)(0.01) + (\frac{1}{6})^3(0.99)} = 0.21$$

Bayes' theorem



"Consider an occasionally dishonest casino that uses a lot of dice. Of the dice 99% are fair but 1% are loaded. If a loaded die is used, a six comes up 50% of the time. We pick up a die from the casino at random."

Hmmm... is this a loaded die?

... most probably (79%) this is a fair die!

In the case of our die, Bayes' theorem can be

$$P(D_{loaded} | 3sixes) = \frac{P(3sixes | D_{loaded})P(D_{loaded})}{P(3sixes)}$$

We have been told, that $P(D_{loaded}) = 0.01$

and that $P(3sixes | D_{loaded}) = 0.5^3 = 0.125$

Thus:

$$P(D_{loaded} | 3sixes) = \frac{(0.5^3)(0.01)}{(0.5^3(0.01) + (\frac{1}{6})^3(0.99))} = 0.21$$

Bayes' theorem

“Assuming that, on average, *extracellular* proteins have a slightly different amino acid composition than *intracellular* proteins (e.g. *cysteine* is more common in extracellular than intracellular proteins), lets try to use this information to judge whether a new protein sequence $x = x_1 \dots x_n$ is intracellular or extracellular.”

Bayes' theorem

“Assuming that, on average, *extracellular* proteins have a slightly different amino acid composition than *intracellular* proteins (e.g. *cysteine* is more common in extracellular than intracellular proteins), lets try to use this information to judge whether a new protein sequence $x = x_1 \dots x_n$ is intracellular or extracellular.”

1. We split our training examples from UNI-PROT into intracellular and extracellular proteins.

Bayes' theorem

“Assuming that, on average, *extracellular* proteins have a slightly different amino acid composition than *intracellular* proteins (e.g. *cysteine* is more common in extracellular than intracellular proteins), lets try to use this information to judge whether a new protein sequence $x = x_1 \dots x_n$ is intracellular or extracellular.”

1. We split our training examples from UNI-PROT into intracellular and extracellular proteins.
2. We estimate a set of frequencies q_a^{int} for intracellular proteins, and a corresponding set of extracellular frequencies q_a^{ext} .

Bayes' theorem

“Assuming that, on average, *extracellular* proteins have a slightly different amino acid composition than *intracellular* proteins (e.g. *cysteine* is more common in extracellular than intracellular proteins), lets try to use this information to judge whether a new protein sequence $x = x_1 \dots x_n$ is intracellular or extracellular.”

1. We split our training examples from UNI-PROT into intracellular and extracellular proteins.
2. We estimate a set of frequencies q_a^{int} for intracellular proteins, and a corresponding set of extracellular frequencies q_a^{ext} .
3. We estimate the probability that any new sequence is extracellular, p^{ext} , and the corresponding probability of being intracellular, p^{int} . Assuming that every sequence must be either entirely intracellular or entirely extracellular i.e. $p^{int} = 1 - p^{ext}$, we can write Bayes' theorem:

$$P(ext | x) = \frac{p^{ext} \prod_i q_{xi}^{ext}}{p^{ext} \prod_i q_{xi}^{ext} + p^{int} \prod_i q_{xi}^{int}}$$

Bayes' theorem

"Assuming that, on average, *extracellular* proteins have a slightly different amino acid composition than *intracellular* proteins (e.g. cysteine is more common in extracellular than intracellular proteins), lets try to use this information to judge whether a new protein sequence $x = x_1 \dots x_n$ is intracellular or extracellular."

1. We split our training examples from UNI-PROT into intracellular and extracellular proteins.
2. We estimate a set of frequencies q_a^{int} for intracellular proteins, and a corresponding set of extracellular frequencies q_a^{ext} .
3. We estimate the probability that any new sequence is extracellular, p^{ext} , and the corresponding probability of being intracellular, p^{int} . Assuming that every sequence must be either entirely intracellular or entirely extracellular i.e. $p^{int} = 1 - p^{ext}$, we can write Bayes' theorem:

$$P(ext | x) = \frac{p^{ext} \prod_i q_{xi}^{ext}}{p^{ext} \prod_i q_{xi}^{ext} + p^{int} \prod_i q_{xi}^{int}}$$



p^{ext}, p^{int} : prior probs

$P(ext | x)$: posterior probs

Bayesian parameter estimation

Bayes' theorem can also be used to estimate the parameters θ of a model:

$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)}$$

Bayesian parameter estimation

Bayes' theorem can also be used to estimate the parameters θ of a model:

$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)}$$

One way of using Bayes' theorem to estimate good parameters, is to choose the parameter values for θ that maximize $P(\theta | D)$, a process known as **maximum a posteriori (MAP)** estimation.

Bayesian parameter estimation

Bayes' theorem can also be used to estimate the parameters θ of a model:

$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)}$$

One way of using Bayes' theorem to estimate good parameters, is to choose the parameter values for θ that maximize $P(\theta | D)$, a process known as *maximum a posteriori (MAP)* estimation.

“We are given a die that we expect will be *loaded*, but we *don't know in what way*. We are allowed to *roll* it *ten times*, and we have to give our *best estimates* for the *parameters* p_i . We roll 1, 3, 4, 2, 4, 6, 2, 1, 2, 2. ”

Bayesian parameter estimation

Bayes' theorem can also be used to estimate the parameters θ of a model:

$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)}$$

One way of using Bayes' theorem to estimate good parameters, is to choose the parameter values for θ that maximize $P(\theta | D)$, a process known as **maximum a posteriori (MAP)** estimation.

“We are given a die that we expect will be **loaded**, but we **don't know in what way**. We are allowed to **roll** it **ten times**, and we have to give our **best estimates** for the **parameters** p_i . We roll 1, 3, 4, 2, 4, 6, 2, 1, 2, 2. ”

The **ML** estimate for p_5 , based on the observed frequency, is **0**. Remember though that we **have not seen enough** data to be sure that this die never rolls a five.

Bayesian parameter estimation

Bayes' theorem can also be used to estimate the parameters θ of a model:

$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)}$$

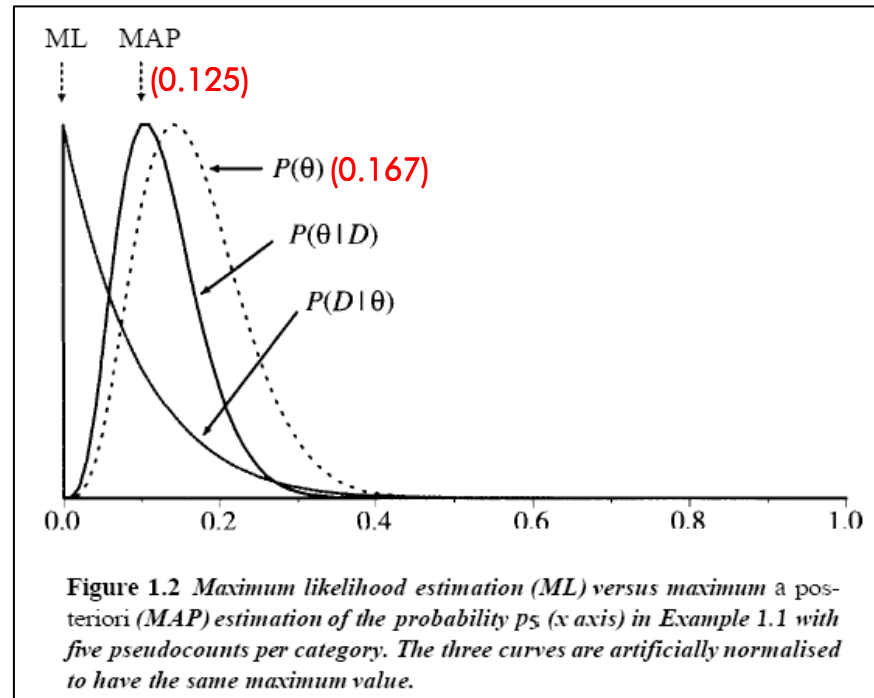
One way of using Bayes' theorem to estimate good parameters, is to choose the parameter values for θ that maximize $P(\theta | D)$, a process known as **maximum a posteriori (MAP)** estimation.

“We are given a die that we expect will be **loaded**, but we **don't know in what way**. We are allowed to **roll** it **ten times**, and we have to give our **best estimates** for the **parameters** p_i . We roll 1, 3, 4, 2, 4, 6, 2, 1, 2, 2. ”

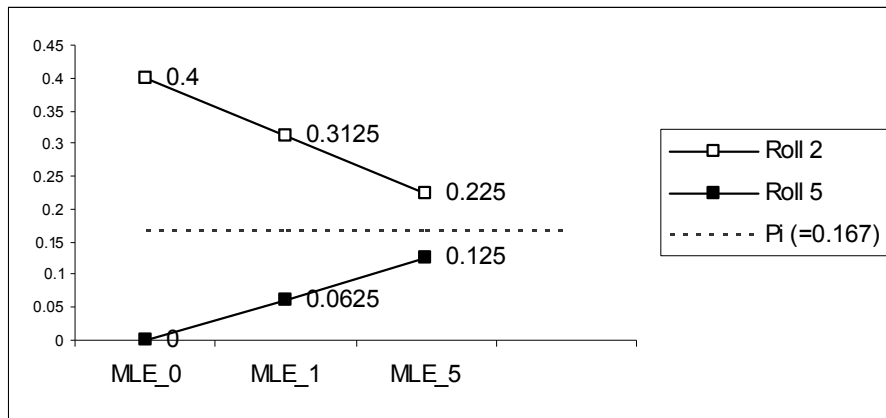
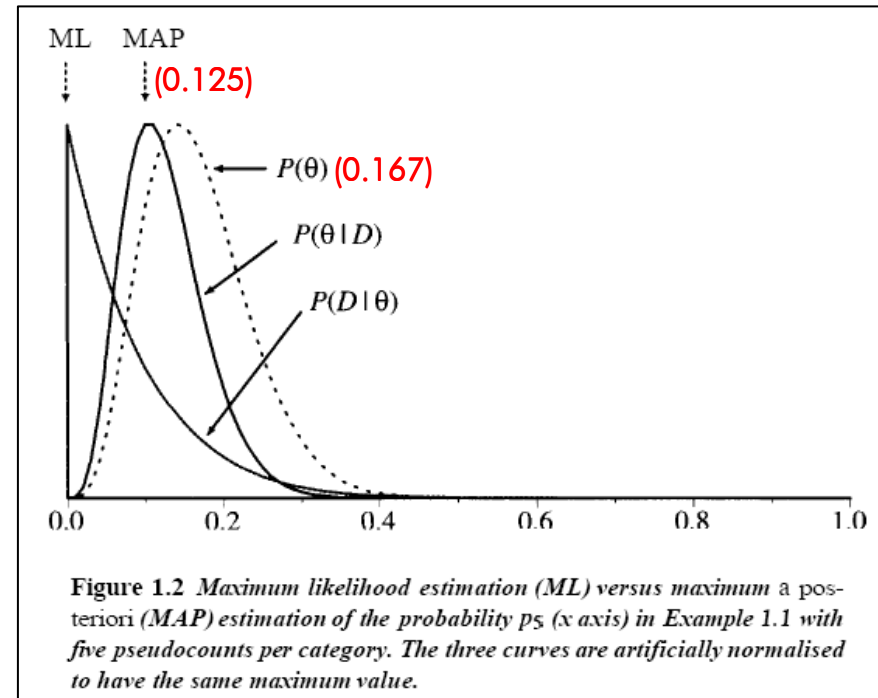
The **ML** estimate for p_5 , based on the observed frequency, is **0**. Remember though that we **have not seen enough** data to be sure that this die never rolls a five.

One well-known approach to this problem is to adjust the observed frequencies used to derive the probabilities by adding some fake extra **pseudocounts**.

MAP vs ML



MAP vs ML

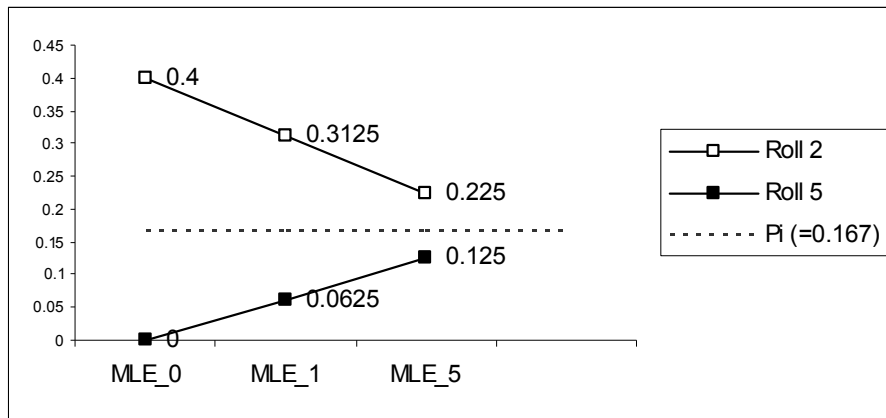
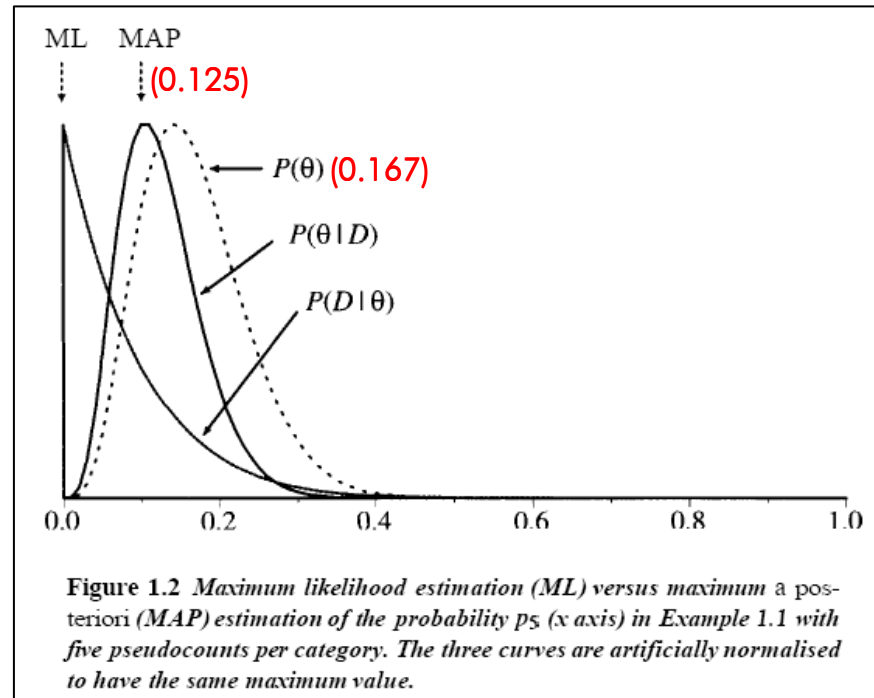


MAP vs ML

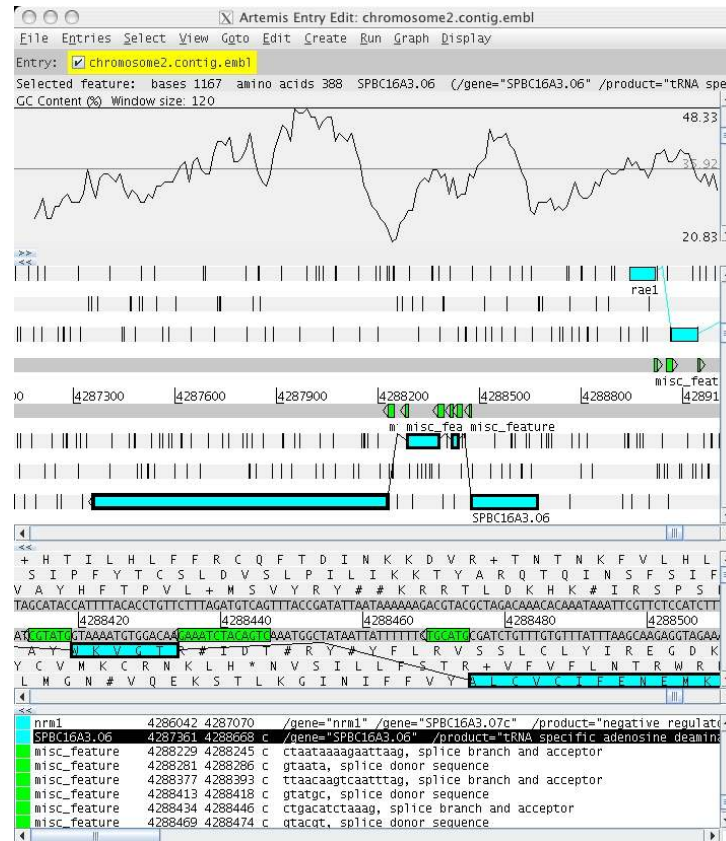
In this example what is our maximum likelihood estimate for p_3 , the probability of rolling a three?

What is the Bayesian estimate if we add one pseudocount per category?

What if we add five pseudocounts per category?

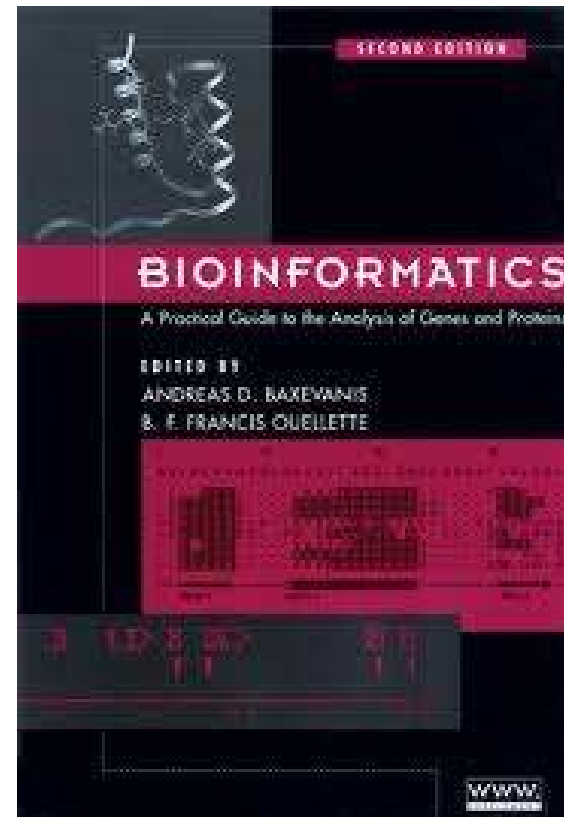
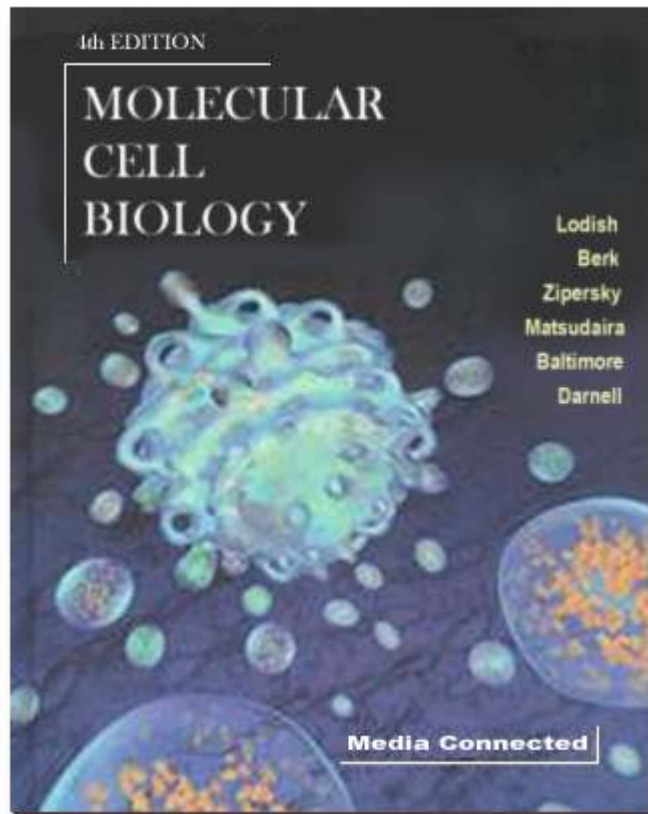


Artemis Demo



Source: <http://www.sanger.ac.uk/resources/software/artemis/>

Further (Biological-Oriented) Reading



<http://eclass.di.uoa.gr/courses/D461/> Υλικό Μαθήματος/Βιβλιογραφία