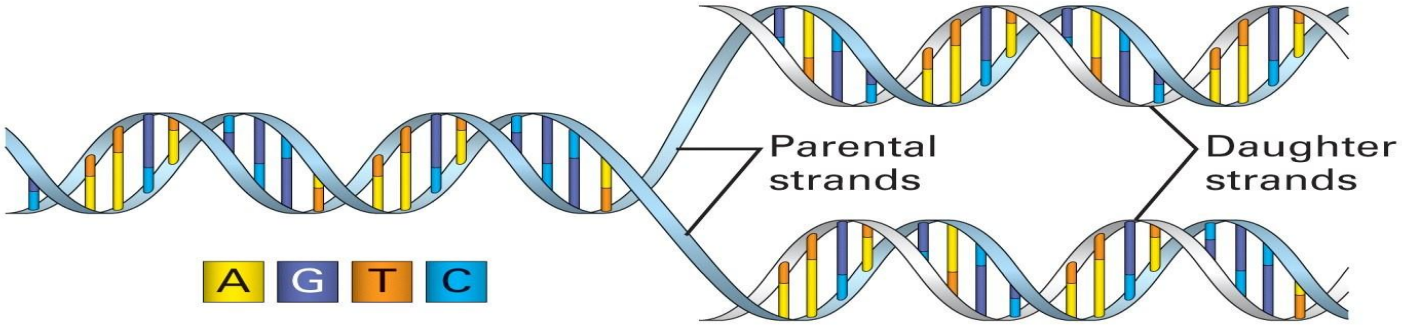
BASICS ON BASES:

A-G-T-C AS WORDS



■ The bases Adenine, Guanine, Thymine, and Cytosine form chemical pairs A-T and C-G → DNA double helix

- This lecture approaches the DNA-world by considering *words,* short strings of letters drawn from an alphabet, which in the case DNA is the set of letters A-G-T-C forming *k-words* or *k- tuples* (*k* is the word length).
- DNA sequences from different regions of a genome differ by their *k-tuple* content and different organisms differ as well.
- We take a look at computational issues on words, how to count words and how words can be located along a string.
- Word distribution description includes probabilistic modelling.
- Some statistics used to describe word frequencies.

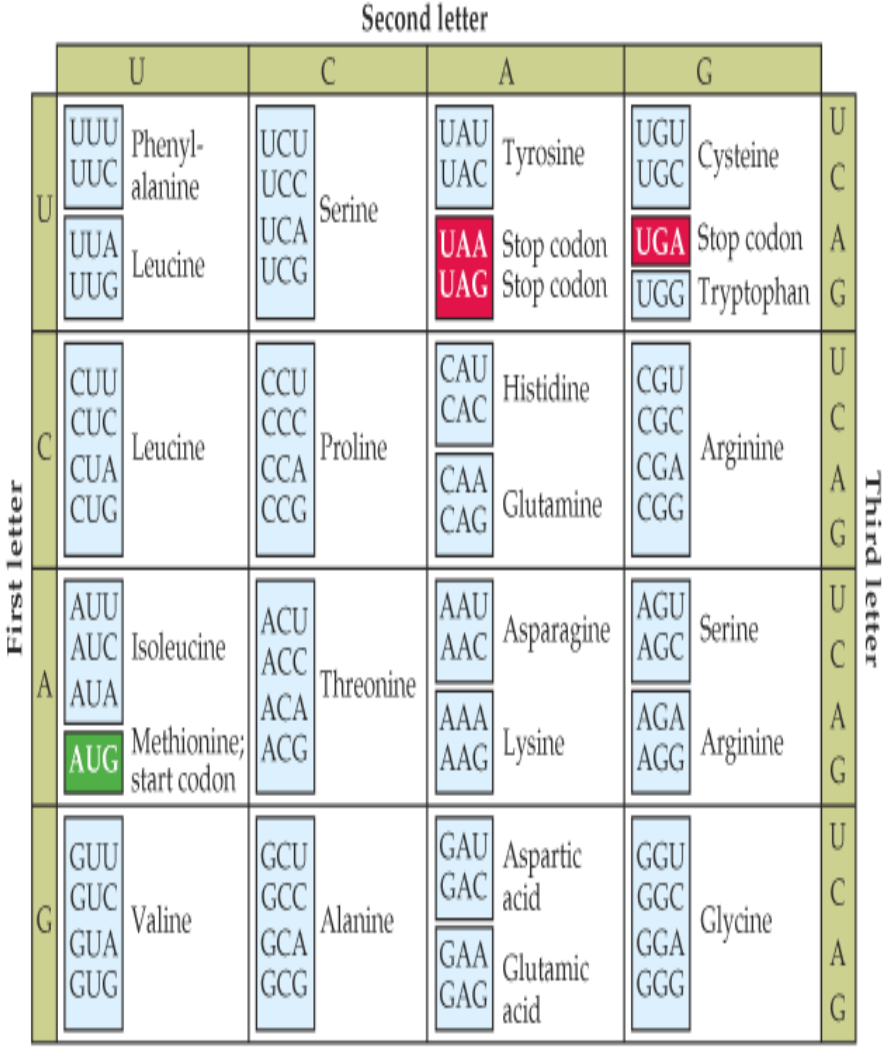
- Next week lectures:
  - The biological perspective on DNA-world and A-G-T-C.
  - Flow of biological information, DNA, RNA, proteins
- Next week also Biology for methodological scientists: The reading group in Meilahti campus starts (Wednesday, see the calendar and course list).
- In the the *wet-lab* biology course, Measurement techniques, you extract DNA from yourselves in Wednesday 23. September.

A cell of an organism contains DNA-molecules, organized into chromosomes

Organism	#base pairs	#chromosomes
Escherichia coli (bacterium)	$4x10^6$	1
Saccharomyces cerevisiae (yeast)	$1.35x10^7$	17
Drosophila melanogaster (insect)	$1.65x10^8$	4
Homo sapiens (human)	$2.9x10^9$	23
Zea mays (corn / maize)	$5.0x10^9$	10

DNA codes for proteins

- The DNA-code A-G-T-C through RNA-code, A-G-U-C, codes for 20 different amino acids.
- Trinucleotides (triplets) allow 4^3 = 64 possible trinucleotides.
- Triplets are also called *codons.*



DNA makes new copies of itself, replicates

- In this process, mistakes can occur.
- The cell repair machinery may, or may not, correct the mistakes.
- Mistakes can be moved on as mutations.
- This in one (simple) mechanism that generates differences to DNA-differences between organisms.
- This is can be considered as *a string manipulation issue*

Biological string manipulation

- One type of a mutation is *deletion*: removal of one or more contiguous bases (substring)
  - …TT**GA**TCA… => …TTTCA…
- Another type is and *insertion*: insertion of a substring
  - …GGCTAG… => …GG**TCAAC**TAG…
- Point mutation: substitution of a base
  - …ACG**G**CT… => …ACG**C**CT…

From: Esa Pitkänen

Given a DNA sequence, we might ask a number of questions

What sort of statistics should be used to describe the sequence?

What sort of organism did this sequence come from?

Does the description of this sequence differ from the description of other DNA in the organism?

What sort of sequence is this? What does it do?

Biological words

- We can try to answer questions like these by considering the *words* in a sequence
- A *k*-word (or a *k-tuple*) is a string of length *k* drawn from some alphabet
- A DNA k-word is a string of length k that consists of letters A, C, G, T
  - 1-words: individual nucleotides (bases)
  - 2-words: dinucleotides (AA, AC, AG, AT, CA, ...)
  - 3-words: codons (AAA, AAC, …)
  - 4-words and beyond

1-words: base composition

- Typically DNA exists as *duplex* molecule (two complementary strands)

```
5′-GGATCGAAGCTAAGGGCT-3′
3′-CCTAGCTTCGATTCCCGA-5′
```

These are something we can determine experimentally.

Top strand: 7 G, 3 C, 5 A, 3 T
Bottom strand: 3 G, 7 C, 3 A, 5 T
Duplex molecule: 10 G, 10 C, 8 A, 8 T
Base frequencies: 10/36 10/36 8/36 8/36

fr(G + C) = 20/36, fr(A + T) = 1 – fr(G + C) = 16/36

G+C content

- fr(G + C), or *G+C content* is a simple statistics for describing genomes
- Notice that one value is enough characterise fr(A), fr(C), fr(G) and fr(T) for duplex DNA
- Is G+C content (= base composition) able to tell the difference between genomes of different organisms?
  - Simple computational experiment, if we have the genome sequences under study (-> exercises)

G+C content for various organisms

Bacteria

- *Mycoplasma genitalium* 31.6%
- *Escherichia coli* K-12 50.7%
- *Pseudomonas aeruginosa* PAO1 66.4%
- *Pyrococcus abyssi* 44.6%
- *Thermoplasma volcanium* 39.9%

worm

- *Caenorhabditis elegans* 36%

plant

- *Arabidopsis thaliana* 35%

human

- *Homo sapiens* 41%

Base frequencies in duplex molecules

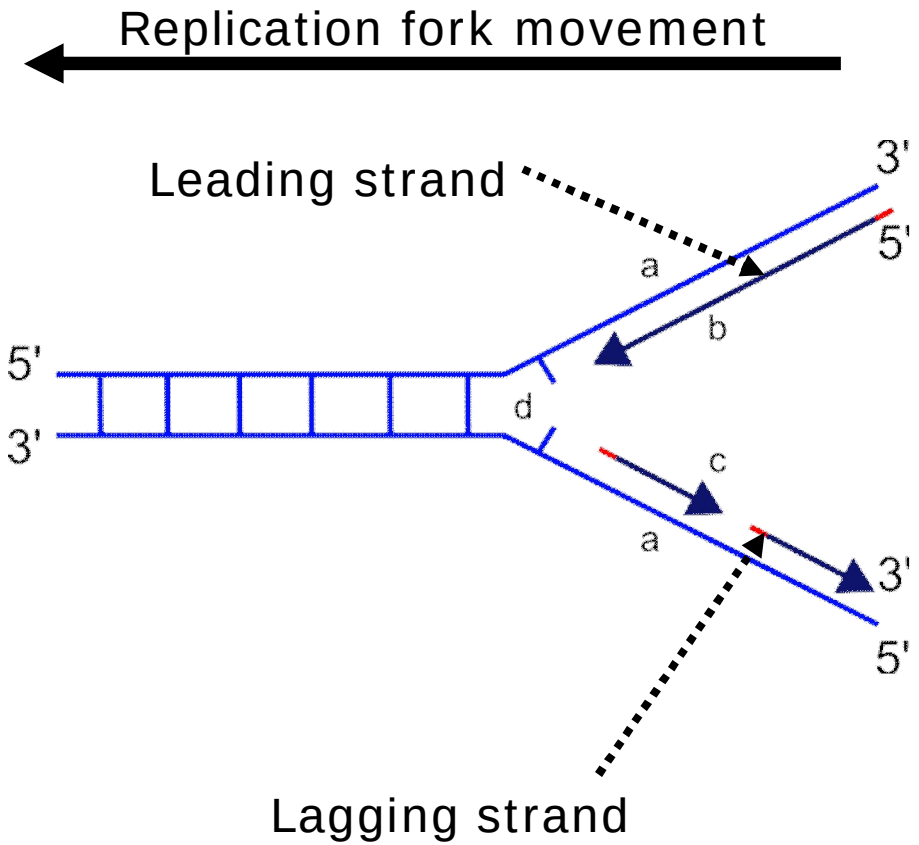
- Consider a DNA sequence generated randomly, with probability of each letter being independent of position in sequence
- You could expect to find a uniform distribution of bases in genomes…

```
5′-...GGATCGAAGCTAAGGGCT...-3′
3′-...CCTAGCTTCGATTCCCGA...-5′
```

- This is not, however, the case in genomes, especially in bacteria
  - This phenomenon is called *GC skew*

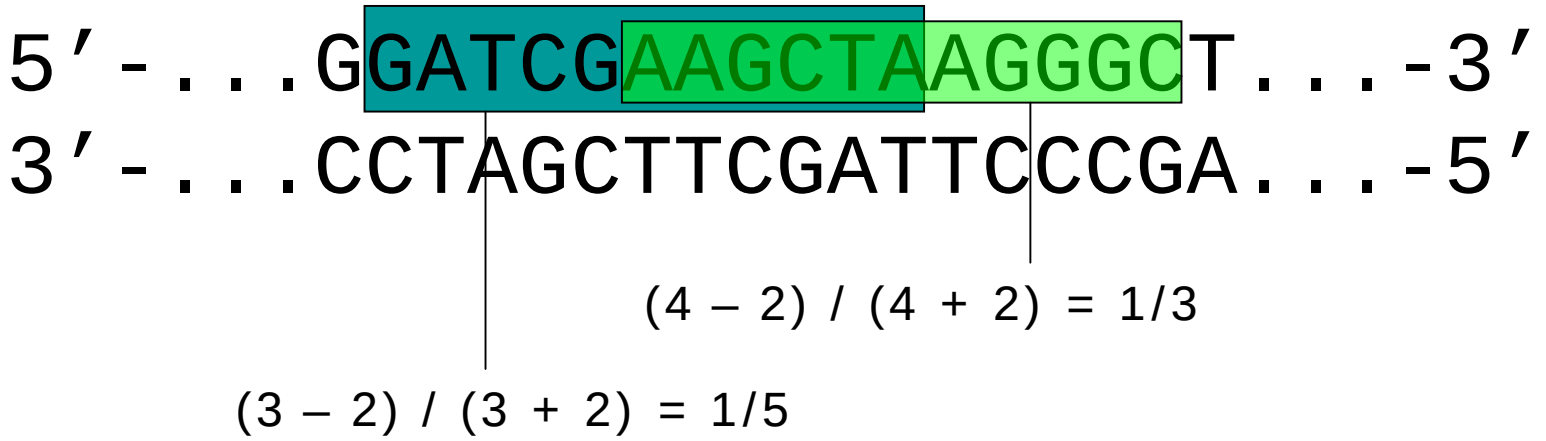
DNA replication fork

- When DNA is replicated, the molecule takes the *replication fork* form
- New complementary DNA is synthesised at both strands of the ”fork”
- New strand in 5’-3’ direction corresponding to replication fork movement is called *leading strand* and the other *lagging strand*
- This process has specific starting points in genome (*origins of replication)*
- Observation: Leading strands have an excess of G over C
- This can be described by *GC skew* statistics



GC skew

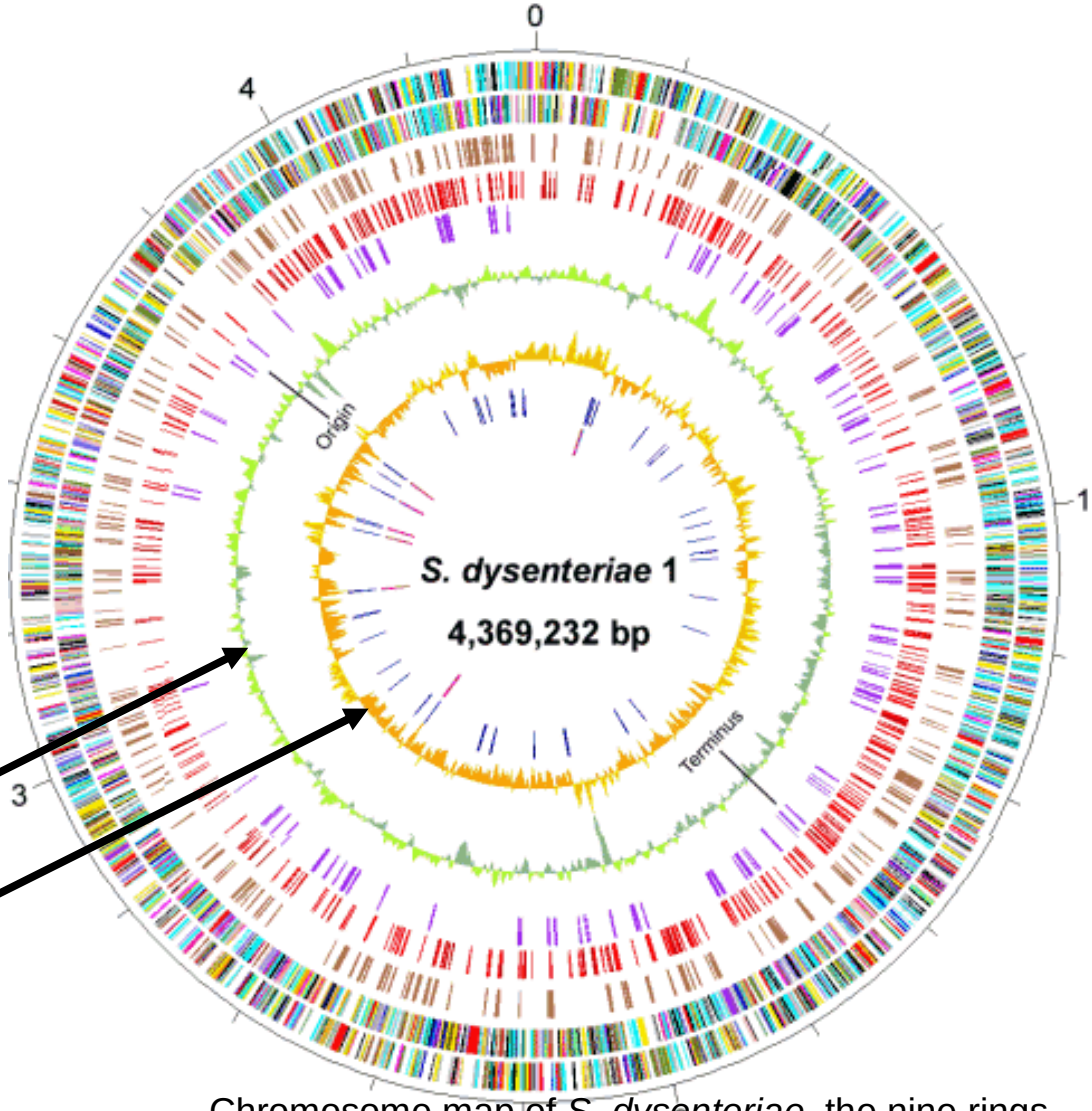
- GC skew is defined as (#G - #C) / (#G + #C)
- It is calculated at successive positions in intervals (windows) of specific width



From: Esa Pitkänen

G-C content & GC skew

- G-C content & GC skew statistics can be displayed with a *circular genome map*



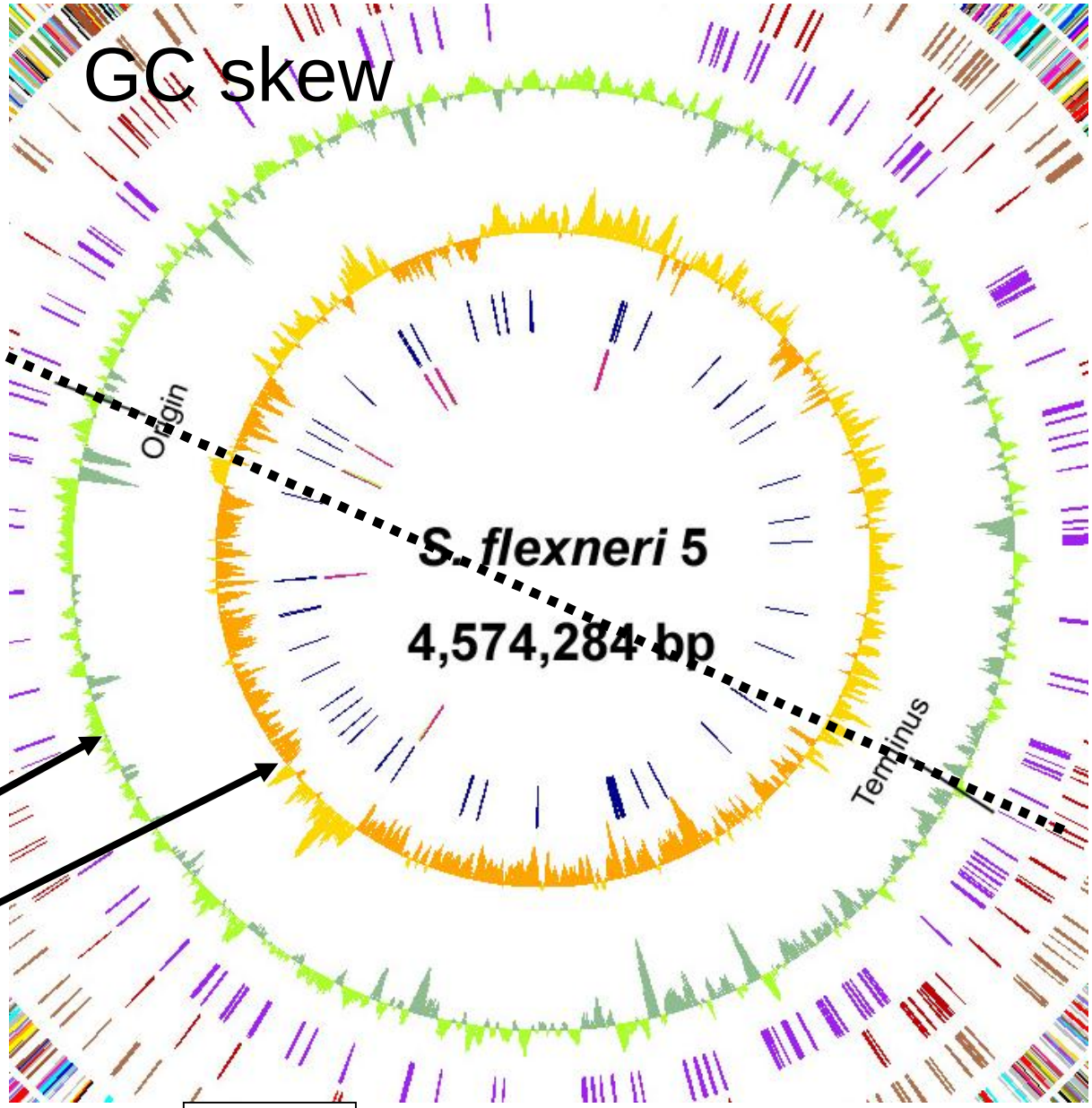
G+C content

GC skew

(10kb window size)

Chromosome map of *S. dysenteriae*, the nine rings describe different properties of the genome
http://www.mgc.ac.cn/ShiBASE/circular_Sd197.htm

- GC skew often changes sign at origins and termini of replication



G+C content

GC skew

(10kb window size)

Nie et al., BMC Genomics, 2006

2-words: dinucleotides

- Let's consider a sequence $L_1,L_2,...,L_n$ where each letter L_i is drawn from the DNA alphabet {A, C, G, T}

- We have 16 possible dinucleotides $l_i l_{i+1}$: AA, AC, AG, ..., TG, TT.

i.i.d. model for nucleotides

- Assume that bases
  - occur **i**ndependently of each other
  - bases at each position are **i**dentically **d**istributed

- Probability of the base A, C, G, T occuring is p_A, p_C, p_G, p_T, respectively
  - For example, we could use $p_A=p_C=p_G=p_T=0.25$ or estimate the values from known genome data

- Probability of $l_i l_{i+1}$ is then $P_{li} P_{li+1}$
  - For example, $P(TG) = p_T\, p_G$

What is i.i.d ?

In probability theory and statistics a sequence or other collection of random variables is

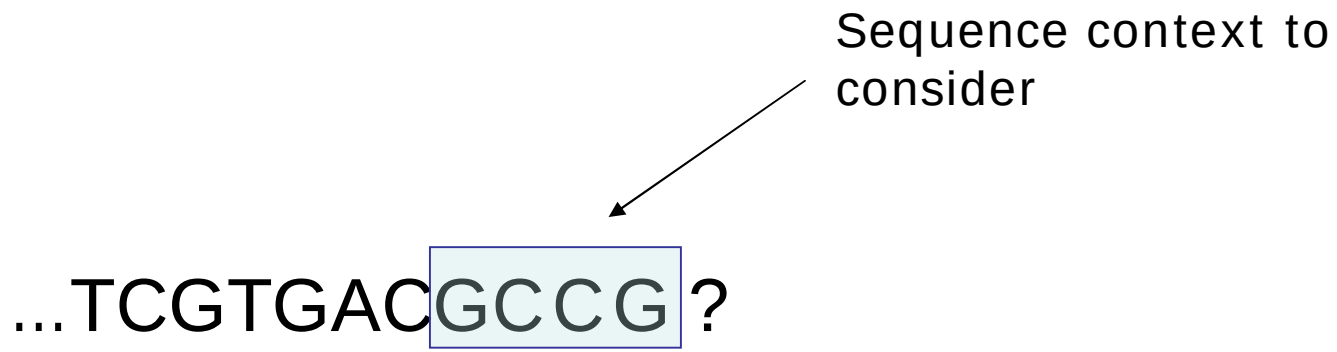
independent and identically distributed (**i.i.d.**) if each random variable has the same probability distribution as the others and all are mutually independent.

2-words: is what we see surprising?

- We can test whether a sequence is ”unexpected”, for example, with a χ^2 test
- Test statistic for a particular dinucleotide r_1r_2 is $\chi^2 = (O - E)^2 / E$ where
  - O is the observed number of dinucleotide r_1r_2
  - E is the expected number of dinucleotide r_1r_2
  - $E = (n - 1)p_{r1}p_{r2}$ under i.i.d. model
- Basic idea: high values of χ^2 indicate deviation from the model
  - Actual procedure is more detailed -> basic statistics courses

Refining the i.i.d. model

- i.i.d. model describes some organisms well but fails to characterise many others
- We can refine the model by having the DNA letter at some position depend on letters at preceding positions

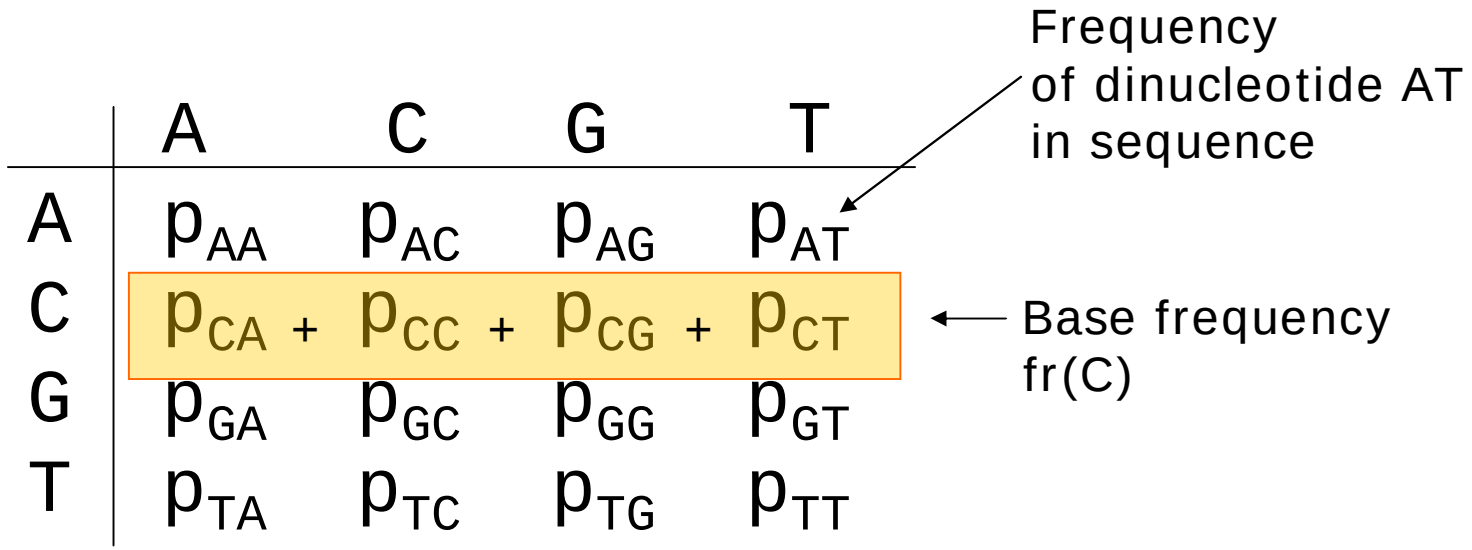


First-order Markov chains

- Lets assume that in sequence X the letter at position t, X_t, depends only on the previous letter X_{t-1} (*first-order markov chain*)
- Probability of letter j occuring at position t given $X_{t-1} = i$: $p_{ij} = P(X_t = j \mid X_{t-1} = i)$
- We consider *homogeneous* Markov chains: probability p_{ij} is independent of position t

Estimating p_{ij}

- We can estimate probabilities p_{ij} (”the probability that j follows i”) from observed dinucleotide frequencies



…the values p_{AA}, p_{AC}, …, p_{TG}, p_{TT} sum to 1

Estimating p_{ij}

Dinucleotide frequency

- $p_{ij} = P(X_t = j \mid X_{t-1} = i) = \frac{P(X_t = j, X_{t-1} = i)}{P(X_{t-1} = i)}$

Probability of transition i -> j

Base frequency of nucleotide i, fr(i)

0.052 / 0.345 ≈ 0.151

	A	C	G	T
A	0.146	0.052	0.058	0.089
C	0.063	0.029	0.010	0.056
G	0.050	0.030	0.028	0.051
T	0.086	0.047	0.063	0.140

$P(X_t = j, X_{t-1} = i)$

	A	C	G	T
A	0.423	0.151	0.168	0.258
C	0.399	0.184	0.063	0.354
G	0.314	0.189	0.176	0.321
T	0.258	0.138	0.187	0.415

$P(X_t = j \mid X_{t-1} = i)$

Simulating a DNA sequence

- From a transition matrix, it is easy to generate a DNA sequence of length n:
  - First, choose the starting base randomly according to the base frequency distribution
  - Then, choose next base according to the distribution $P(x_t \mid x_{t-1})$ until n bases have been chosen

T T C T T C A A

	A	C	G	T
A	0.423	0.151	0.168	0.258
C	0.399	0.184	0.063	0.354
G	0.314	0.189	0.176	0.321
T	0.258	0.138	0.187	0.415

$P(X_t = j \mid X_{t-1} = i)$

Now we can quickly generate sequences of arbitrary length...

.

ttcttcaaaataaggatagtgattcttattggcttaagggataacaatttagatctttttcatgaatcatgtatgtcaacgttaaaagttgaactgcaataagttc
ttacacacgattgtttatctgcgtgcgaagcatttcactacatttgccgatgcagccaaaagtatttaacatttggtaaacaaattgacttaaatcgcgcacttaga
gtttgacgtttcatagttgatgcgtgtctaacaattacttttagtttttaaatgcgtttgtctacaatcattaatcagctctggaaaaacattaatgcatttaaac
cacaatggataattagttacttattttaaaattcacaaagtaattattcgaatagtgccctaagagagtactggggttaatggcaaagaaaattactgtagtgaaga
ttaagcctgttattatcacctgggtactctggtgaatgcacataagcaaatgctacttcagtgtcaaagcaaaaaaatttactgataggactaaaaacccttttattt
ttagaatttgtaaaaatgtgacctcttgcttataacatcatatttattgggtcgttctaggacactgtgattgccttctaactcttatttagcaaaaaattgtcata
gctttgaggtcagacaaacaagtgaatggaagacagaaaaagctcagcctagaattagcatgttttgagtggggaattacttggttaactaaagtgttcatgactgt
tcagcatatgattgttggtgagcactacaaagatagaagagttaaactaggtagtggtgatttcgctaacacagttttcatacaagttctattttctcaatggtttt
ggataagaaaacagcaaacaaatttagtattattttcctagtaaaaagcaaacatcaaggagaaattggaagctgcttgttcagtttgcattaaattaaaaatttat
ttgaagtattcgagcaatgttgacagtctgcgttcttcaaataagcagcaaatcccctcaaaattgggcaaaaacctaccctggcttcttttaaaaaaccaagaaa
agtcctatataagcaacaaatttcaaacctttgttaaaaattctgctgctgaataaataggcattacagcaatgcaattaggtgcaaaaaaggccatcctctttct
ttttttgtacaattgttcaagcaactttgaatttgcagattttaacccactgtctatatgggacttcgaattaaattgactggtctgcatcacaaatttcaactgcc
caatgtaatcatattctagagtattaaaaatacaaaaagtacaattagttatgcccattggcctggcaatttatttactccactttccacgttttggggatatttta
acttgaatagttcacaatcaaaacataggaaggatctactgctaaaagcaaaagcgtattggaatgataaaaaactttgatgtttaaaaaactacaaccttaatgaa
ttaaagttgaaaaaatattcaaaaaaagaaattcagttcttggcgagtaatatttttgatgtttgagatcagggttacaaaataagtgcatgagattaactcttcaa
atataaactgatttaagtgtatttgctaataacattttcgaaaaggaatattatggtaagaattcataaaaatgtttaatactgatacaactttcttttatatcctc
catttggccagaatactgttgcacacaactaattggaaaaaaaatagaacgggtcaatctcagtgggaggagaagaaaaaagttggtgcaggaaatagtttctacta
acctggtataaaaacatcaagtaacattcaaattgcaaatgaaaactaaccgatctaagcattgattgatttttctcatgcctttcgcctagttttaataaacgcgc
cccaactctcatcttcggttcaaatgatctattgtatttatgcactaacgtgcttttatgttagcatttttcaccctgaagttccgagtcattggcgtcactcacaa
atgacattacaatttttctatgttttgttctgttgagtcaaagtgcatgcctacaattctttcttatatagaactagacaaaatagaaaaaggcacttttggagtct
gaatgtcccttagtttcaaaaaggaaattgttgaatttttgtggttagttaaattttgaacaaactagtatagtggtgacaaacgatcaccttgagtcggtgacta
taaaagaaaaaggagattaaaaatacctgcggtgccacatttttgttacgggcatttaaggtttgcatgtgttgagcaattgaaacctacaactcaataagtcatg
ttaagtcacttctttgaaaaaaaaaaagacccttaagcaagctc

25

Results from simulating a DNA sequence

Dinucleotide frequencies

	Simulated	Observed
aa	0.145	0.146
ac	0.050	0.052
ag	0.055	0.058
at	0.092	0.089
ca	0.065	0.063
cc	0.028	0.029
cg	0.011	0.010
ct	0.058	0.056
ga	0.048	0.050
gc	0.032	0.030
gg	0.029	0.028
gt	0.050	0.051
ta	0.084	0.086
tc	0.052	0.047
tg	0.064	0.063
tt	0.138	0.0140

n = 10000

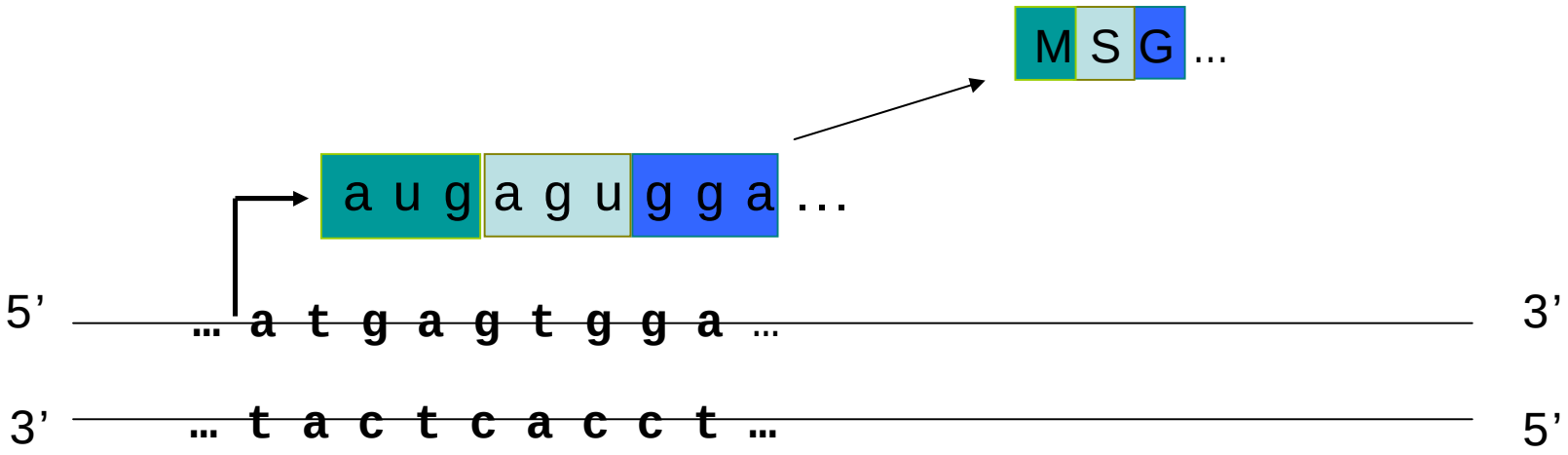
Simulating a DNA sequence

- The model is able to generate correct proportions of 1- and 2-words in genomes...
- ...but fails with k=3 and beyond.

```
ttcttcaaaataaggatagtgattcttattggcttaagggataacaatttagatctttttcatgaatcatgtatgtcaacgttaaaagttgaactgcaataagttc
ttacacacgattgtttatctgcgtgcgaagcatttcactacatttgccgatgcagccaaaagtatttaacatttggtaaacaaattgacttaaatcgcgcacttaga
gtttgacgtttcatagttgatgcgtgtctaacaattacttttagtttttaaatgcgtttgtctacaatcattaatcagctctggaaaaacattaatgcatttaaac
cacaatggataattagttacttattttaaaattcacaaagtaattattcgaatagtgccctaagagagtactggggttaatggcaaagaaaattactgtagtgaaga
ttaagcctgttattatcacctgggtactctggtgaatgcacataagcaaatgctacttcagtgtcaaagcaaaaaaatttactgataggactaaaaacccttttattt
ttagaatttgtaaaaatgtgacctcttgcttataacatcatatttattgggtcgttctaggacactgtgattgccttctaactcttatttagcaaaaaattgtcata
gctttgaggtcagacaaacaagtgaatggaagacagaaaaagctcagcctagaattagcatgttttgagtggggaattacttggttaactaaagtgttcatgactgt
tcagcatatgattgttggtgagcactacaaagatagaagagttaaactaggtagtggtgatttcgctaacacagttttcatacaagttctattttctcaatggtttt
ggataagaaaacagcaaacaaatttagtattatttcctagtaaaaagcaaacatcaaggagaaattggaagctgcttgttcagtttgcattaaattaaaaatttat
ttgaagtattcgagcaatgttgacagtctgcgttcttcaaataagcagcaaatcccctcaaaattgggcaaaaacctaccctggcttctttttaaaaaaccaagaaa
agtcctatataagcaacaaatttcaaacctttgttaaaaattctgctgctgaataaataggcattacagcaatgcaattaggtgcaaaaaaggccatcctctttct
ttttttgtacaattgttcaagcaactttgaatttgcagattttaacccactgtctatatgggacttcgaattaaattgactggtctgcatcacaaatttcaactgcc
caatgtaatcatattctagagtattaaaaatacaaaaagtacaattagttatgcccattggcctggcaatttatttactccactttccacgttttggggatatttta
acttgaatagttcacaatcaaaacataggaaggatctactgctaaaagcaaaagcgtattggaatgataaaaaactttgatgtttaaaaaactacaaccttaatgaa
ttaaagttgaaaaaatattcaaaaaaagaaattcagttcttggcgagtaatatttttgatgtttgagatcagggttacaaaataagtgcatgagattaactcttcaa
atataaactgatttaagtgtatttgctaataacattttcgaaaaggaatattatggtaagaattcataaaaatgtttaatactgatacaactttcttttatatcctc
catttggccagaatactgttgcacacaactaattggaaaaaaaatagaacgggtcaatctcagtgggaggagaagaaaaaagttggtgcaggaaatagtttctacta
acctggtataaaaacatcaagtaacattcaaattgcaaatgaaaactaaccgatctaagcattgattgatttttctcatgcctttcgcctagttttaataaacgcgc
cccaactctcatcttcggttcaaatgatctattgtatttatgcactaacgtgcttttatgttagcatttttcaccctgaagttccgagtcattggcgtcactcacaa
atgacattacaatttttctatgttttgttctgttgagtcaaagtgcatgcctacaattctttcttatatagaactagacaaaatagaaaaaggcacttttggagtct
gaatgtcccttagtttcaaaaaggaaattgttgaatttttgtggttagttaaattttgaacaaactagtatagtggtgacaaacgatcaccttgagtcggtgacta
taaaagaaaaaggagattaaaaatacctgcggtgccacatttttgttacgggcatttaaggtttgcatgtgttgagcaattgaaacctacaactcaataagtcatg
ttaagtcacttctttgaaaaaaaaaaagacccttaagcaagctc
```

3-words: codons

- We can extend the previous method to 3-words
- k=3 is an important case in study of DNA sequences because of genetic code



3-word probabilities

- Let's again assume a sequence L of independent bases
- Probability of 3-word $r_1r_2r_3$ at position i,i+1,i+2 in sequence L is

$$P(L_i = r_1, L_{i+1} = r_2, L_{i+2} = r_3) =$$

$$P(L_i = r_1)P(L_{i+1} = r_2)P(L_{i+2} = r_3)$$

3-words in Escherichia coli genome

Word	Count	Observed	Expected	Word	Count	Observed	Expected
AAA	108924	**0.02348**	**0.01492**	CAA	76614	0.01651	0.01541
AAC	82582	0.01780	0.01541	CAC	66751	0.01439	0.01591
AAG	63369	0.01366	0.01537	CAG	104799	0.02259	0.01588
AAT	82995	0.01789	0.01490	CAT	76985	0.01659	0.01539
ACA	58637	0.01264	0.01541	CCA	86436	0.01863	0.01591
ACC	74897	0.01614	0.01591	CCC	47775	0.01030	0.01643
ACG	73263	0.01579	0.01588	CCG	87036	0.01876	0.01640
ACT	49865	0.01075	0.01539	CCT	50426	0.01087	0.01589
AGA	56621	0.01220	0.01537	CGA	70938	0.01529	0.01588
AGC	80860	0.01743	0.01588	CGC	115695	**0.02494**	**0.01640**
AGG	50624	0.01091	0.01584	CGG	86877	0.01872	0.01636
AGT	49772	0.01073	0.01536	CGT	73160	0.01577	0.01586
ATA	63697	0.01373	0.01490	CTA	26764	**0.00577**	**0.01539**
ATC	86486	0.01864	0.01539	CTC	42733	0.00921	0.01589
ATG	76238	0.01643	0.01536	CTG	102909	0.02218	0.01586
ATT	83398	0.01797	0.01489	CTT	63655	0.01372	0.01537

2nd order Markov Chains

- Markov chains readily generalise to higher orders
- In 2nd order markov chain, position t depends on positions t-1 and t-2
- Transition matrix:

	A	C	G	T
AA				
AC				
AG				
AT				
CA				
…				

Codon Adaptation Index (CAI)

- Observation: cells prefer certain codons over others in highly expressed genes
  - Gene expression: DNA is transcribed into RNA (and possibly translated into protein)
- CAI is a statistic used to compare the distribution of codons ***observed*** with the ***preferred*** codons for highly expressed genes

Amino acid	Codon	Predicted	Gene class I	Gene class II
Phe	TTT	0.493	0.551	0.291
	TTC	0.507	0.449	0.709
Ala	GCT	0.246	0.145	0.275
	GCC	0.254	0.276	0.164
	GCA	0.246	0.196	0.240
	GCG	0.254	0.382	0.323
Asn	AAT	0.493	0.409	0.172
	AAC	0.507	0.591	0.828

Moderately expressed

Highly expressed

Codon frequencies for some genes in E. coli

Codon Adaptation Index (CAI)

- Consider an amino acid sequence $X = x_1x_2...x_n$
- Let p_k be the probability that codon k is used in highly expressed genes
- Let q_k be the highest probability that a codon coding for the same amino acid as codon k has
  - For example, if codon k is ”GCC”, the corresponding amino acid is Alanine (see genetic code table; also GCT, GCA, GCG code for Alanine)
  - Assume that $p_{GCC} = 0.164$, $p_{GCT} = 0.275$, $p_{GCA} = 0.240$, $p_{GCG} = \mathbf{0.323}$
  - Now $q_{GCC} = q_{GCT} = q_{GCA} = q_{GCG} = \mathbf{0.323}$

Codon Adaptation Index (CAI)

- CAI is defined as

$$CAI = \left(\prod_{k=1}^{n} p_k / q_k \right)^{1/n}$$

- CAI can be given also in *log-odds* form:

$$\log(CAI) = (1/n) \sum_{k=1}^{n} \log(p_k / q_k)$$

CAI: example with an E. coli gene

q_k
p_k

M	A	L	T	K	A	E	M	S	E	Y	L	…
ATG	GCG	CTT	ACA	AAA	GCT	GAA	ATG	TCA	GAA	TAT	CTG	
1.00	**0.47**	0.02	0.45	**0.80**	**0.47**	**0.79**	**1.00**	**0.43**	**0.79**	0.19	0.02	
	0.06	0.02	**0.47**	0.20	0.06	0.21		0.32	0.21	**0.81**	0.02	
	0.28	0.04	0.04		0.28			0.03			0.04	
	0.20	0.03	0.05		0.20			0.01			0.03	
		0.01						0.04			0.01	
		0.89						0.18			**0.89**	
ATG	**GCT**	TTA	ACT	**AAA**	**GCT**	**GAA**	**ATG**	**TCT**	**GAA**	TAT	TTA	
	GCC	TTG	**ACC**	AAG	GCC	GAG		TCC	GAG	**TAC**	TTG	
	GCA	CTT	ACA		GCA			TCA			CTT	
	GCG	CTC	ACG		GCG			TCG			CTC	
		CTA						AGT			CTA	
		CTG						AGC			**CTG**	
1.00	0.20	0.04	0.04	0.80	0.47	0.79	1.00	0.03	0.79	0.19	0.89	…
1.00	0.47	0.89	0.47	0.80	0.47	0.79	1.00	0.43	0.79	0.81	0.89	

$1/n$

Biological words: summary

- Simple 1-, 2- and 3-word models can describe interesting properties of DNA sequences
  - GC skew can identify DNA replication origins
  - It can also reveal *genome rearrangement* events and *lateral transfer* of DNA
  - GC content can be used to locate genes: human genes are comparably GC-rich
  - CAI predicts high gene expression levels

  - k=3 models can help to identify correct *reading frames :*
    - Reading frame starts from a start codon and stops in a stop codon
    - Consider what happens when a single extra base is introduced in a reading frame

Note on programming languages

- Working with probability distributions is straightforward with R.
- You can use R in Computer science classrooms Linux systems
- Python works too!

Example Python code for generating DNA sequences with first-order Markov chains.

```
#!/usr/bin/env python

import sys, random

n = int(sys.argv[1])
```

Initialisation: use packages 'sys' and 'random', read sequence length from input.

```
tm = {'a' : {'a' : 0.423, 'c' : 0.151, 'g' : 0.168, 't' : 0.258},
      'c' : {'a' : 0.399, 'c' : 0.184, 'g' : 0.063, 't' : 0.354},
      'g' : {'a' : 0.314, 'c' : 0.189, 'g' : 0.176, 't' : 0.321},
      't' : {'a' : 0.258, 'c' : 0.138, 'g' : 0.187, 't' : 0.415}}

pi = {'a' : 0.345, 'c' : 0.158, 'g' : 0.159, 't' : 0.337}
```

Transition matrix tm and initial distribution pi.

```
def choose(dist):
    r = random.random()
    sum = 0.0
    keys = dist.keys()
    for k in keys:
        sum += dist[k]
        if sum > r:
            return k
    return keys[-1]
```

Function choose(), returns a key (here 'a', 'c', 'g' or 't') of the dictionary 'dist' chosen randomly according to probabilities in dictionary values.

```
c = choose(pi)
for i in range(n - 1):
    sys.stdout.write(c)
    c = choose(tm[c])
sys.stdout.write(c)
sys.stdout.write("\n")
```

Choose the first letter, then choose next letter according to $P(x_t \mid x_{t-1})$.

BASICS ON BIOLOGICAL DATABASES

- Storage of information
- Sources of data
- Go to: http://www.ncbi.nlm.nih.gov/
  - Have a look, what kind of databases
  - Familiarize yourself, at least, with PubMed, visit also OMIM

FASTA format
the basic format – and an important practical concept

Header line,
begins with >

```
>Hepatitis delta virus, complete genome
atgagccaagttccgaacaaggattcgcggggaggatagatcagcgcccgagaggggtga
gtcggtaaagagcattggaacgtcggagatacaactcccaagaaggaaaaaagagaaagc
aagaagcggatgaatttccccataacgccagtgaaactctaggaaggggaaagagggaag
gtggaagagaaggaggcgggcctcccgatccgaggggcccggcggccaagtttggaggac
actccggcccgaagggttgagagtaccccagagggaggaagccacacggagtagaacaga
gaaatcacctccagaggacccccttcagcgaacagagagcgcatcgcgagagggagtagac
catagcgataggaggggatgctaggagttgggggagaccgaagcgaggaggaaagcaaag
agagcagcggggctagcaggtgggtgttccgcccccccgagaggggacgagtgaggcttat
cccggggaactcgacttatcgtccccacatagcagactcccggacccccctttcaaagtga
…
```