MEET AND CODE

techsoup BALKANS

BioinfoRmatika

October 6th 2018

Content

- Sequences – string
  Library BioStrings
- Central dogma of molecular biology
- Transcription, translation, open reading frame (ORF)

___

- Mutations
- Distance/difference between two DNA sequences

___

- Genomic ranges

___

- Break

___

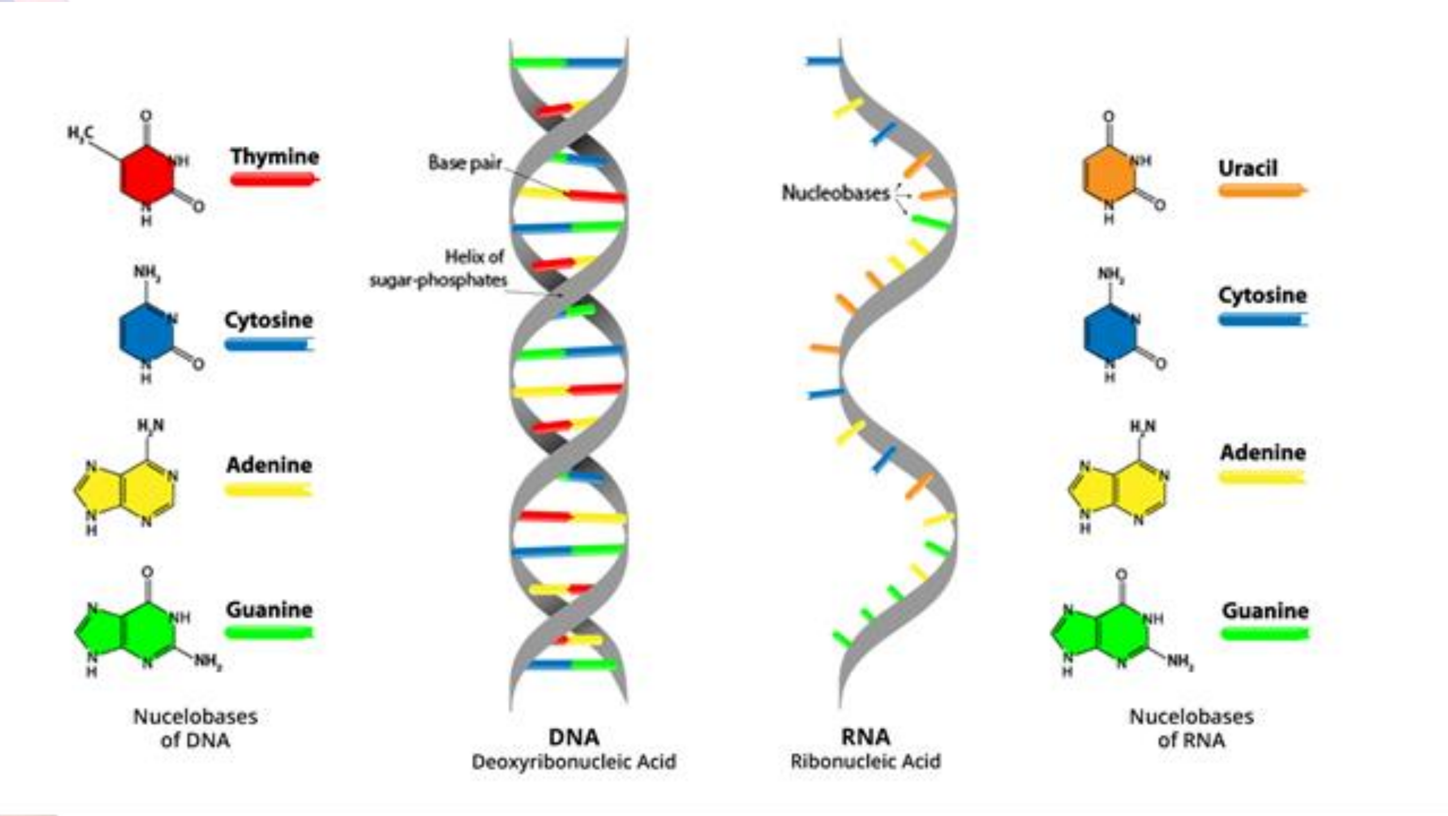
- Final assignment: finding a motif in a bacterial genome

String

- A sequence of characters (char)
- s = “some text”

- substring = substr(s, start, end)

DNA/RNA/proteins



Source: www.technologynetworks.com

BioStrings

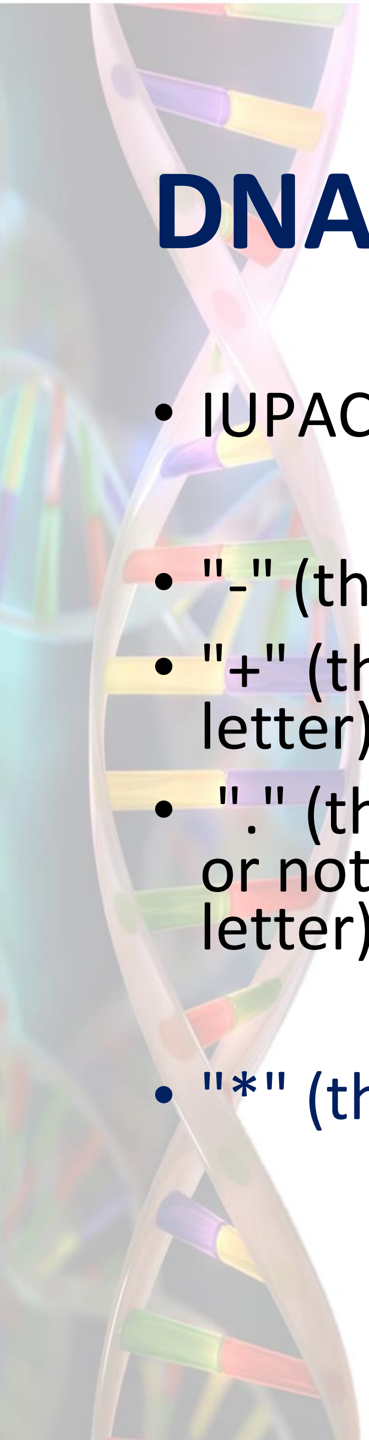
- **string = sequence**

- **DNAString: DNA sequence**
- **RNAString: RNA sequence**
- **AAString: aminoacid sequence (proteins)**

- s = “ACGT”
- DNAseq = DNAString(s)

Difference?

DNAString is a special data structure with its own methods



DNA String

- IUPAC (extended)

- "-" (the gap letter)
- "+" (the hard masking letter)
- "." (the not a letter or not available letter)

- "*" (the stop letter)

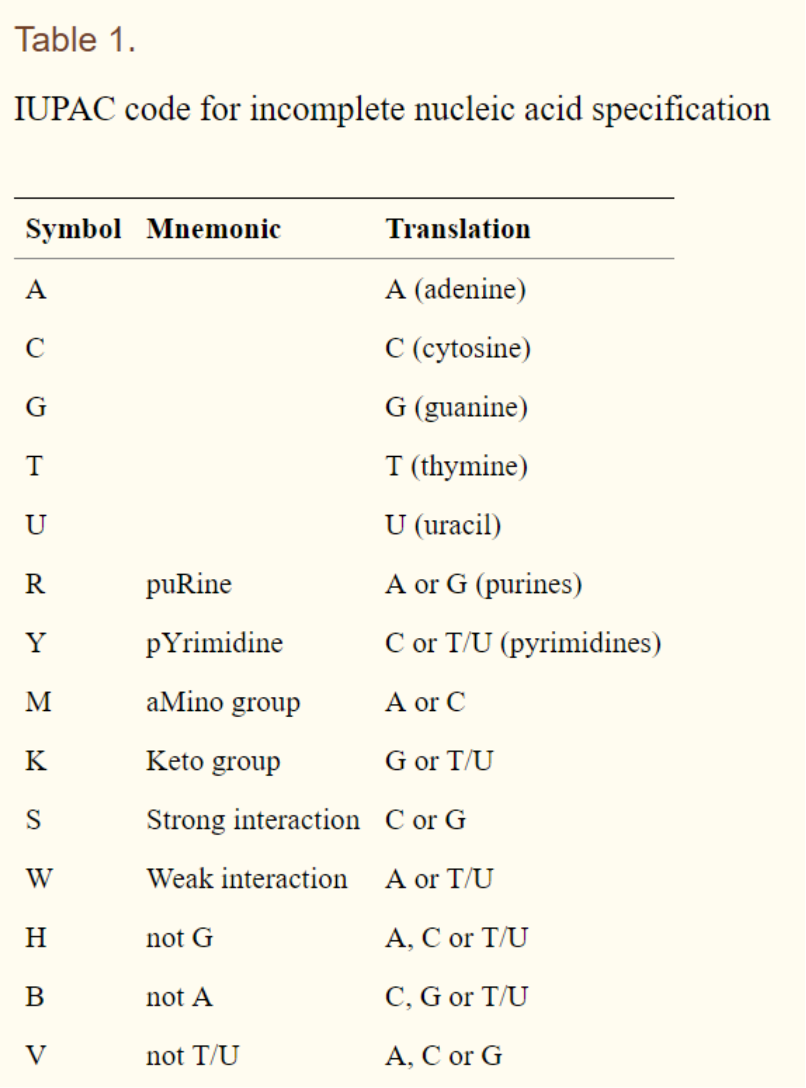


Table 1.

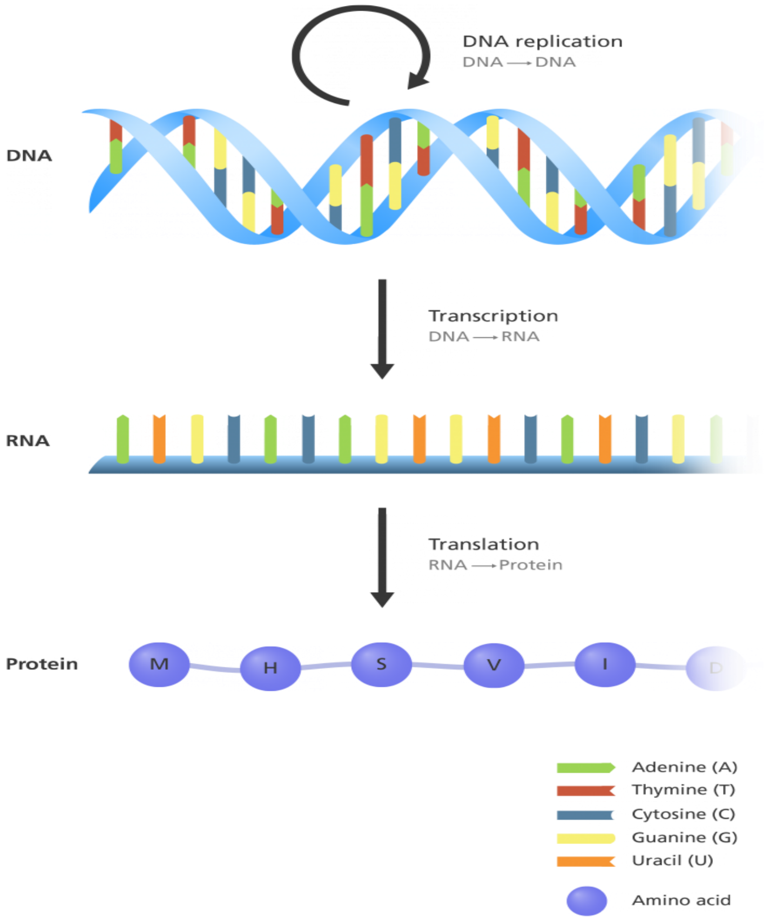
IUPAC code for incomplete nucleic acid specification

Symbol	Mnemonic	Translation
A		A (adenine)
C		C (cytosine)
G		G (guanine)
T		T (thymine)
U		U (uracil)
R	puRine	A or G (purines)
Y	pYrimidine	C or T/U (pyrimidines)
M	aMino group	A or C
K	Keto group	G or T/U
S	Strong interaction	C or G
W	Weak interaction	A or T/U
H	not G	A, C or T/U
B	not A	C, G or T/U
V	not T/U	A, C or G

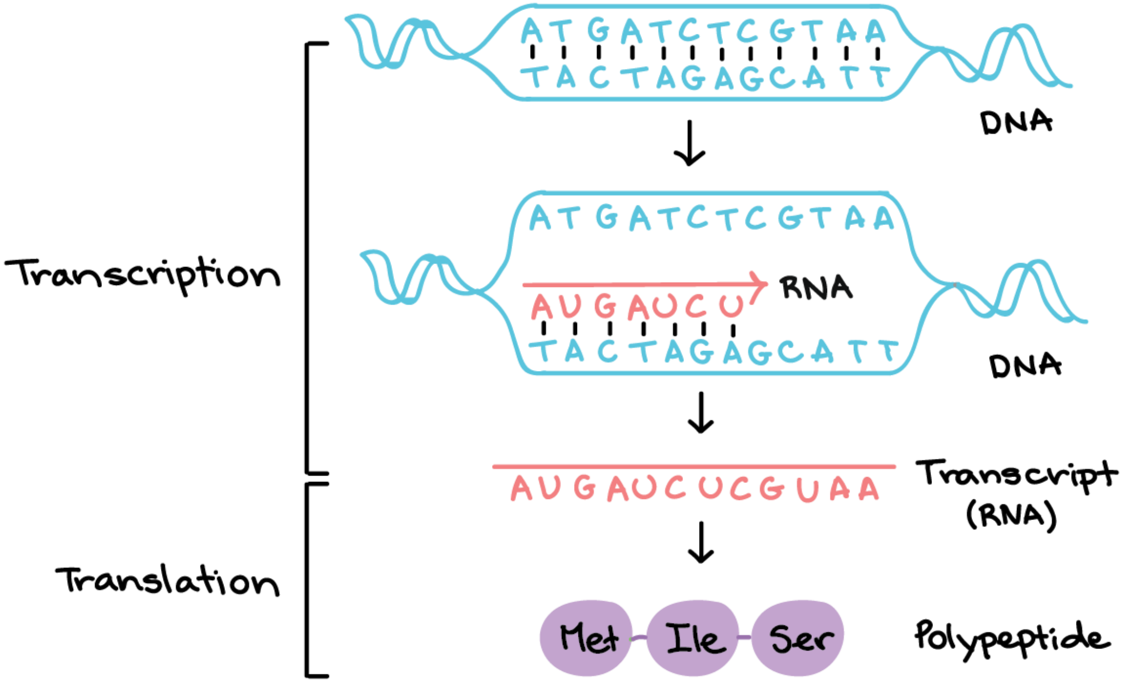
The central dogma of molecular biology

Useful functions:

- reverse
- complement
- reverseComplement
- translate



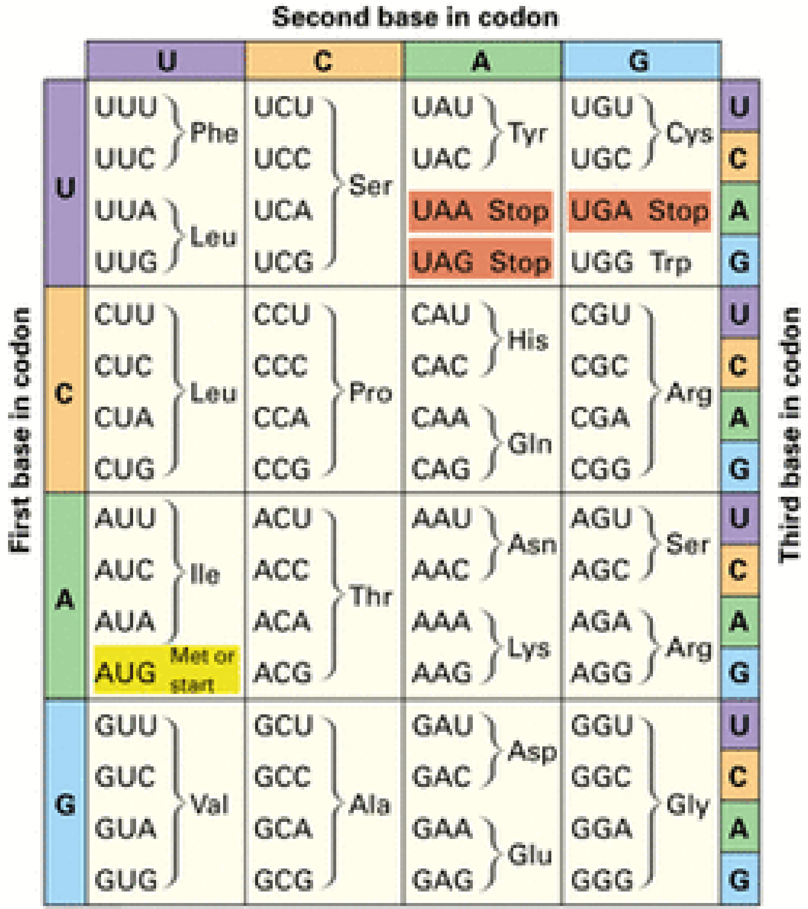
Transcription and translation



source: khanacademy

Genetic code

- 20 amino acids
- 64 codons
- 3 stop codons



Second base in codon

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

source: geneticcontrolandproteinfunction.wordpress.com

Motifs

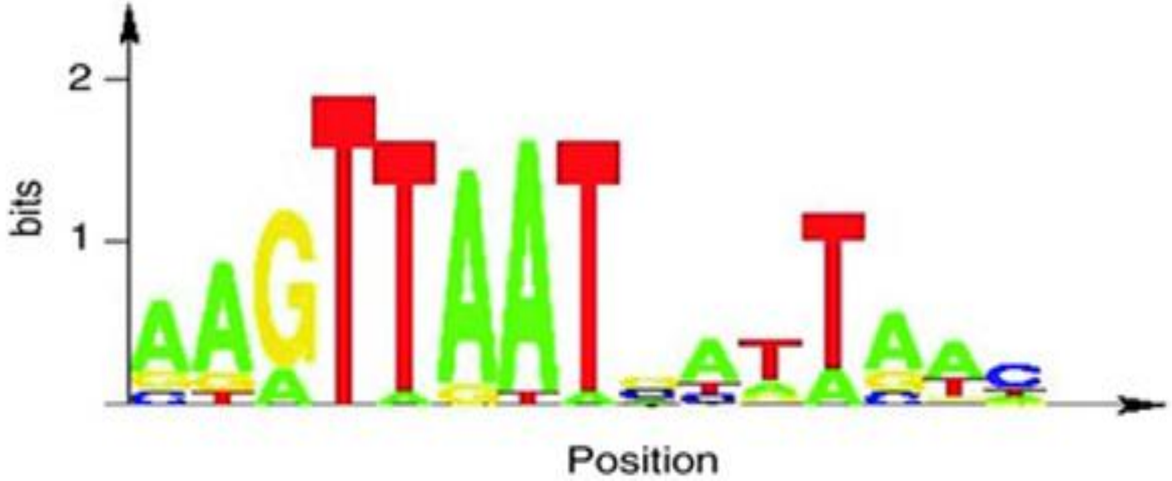
The consensus sequence, position weight matrix and sequence logo of a motif

(a) “AAGTTAATGGATCTG”

(b)

$$\begin{pmatrix} 14 & 16 & 4 & 0 & 1 & 19 & 20 & 1 & 4 & 13 & 4 & 4 & 13 & 12 & 3 \\ 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7 & 3 & 1 & 0 & 3 & 1 & 12 \\ 4 & 3 & 17 & 0 & 0 & 2 & 0 & 0 & 9 & 1 & 3 & 0 & 5 & 2 & 2 \\ 0 & 2 & 0 & 21 & 20 & 0 & 1 & 20 & 1 & 4 & 13 & 17 & 0 & 6 & 4 \end{pmatrix}$$

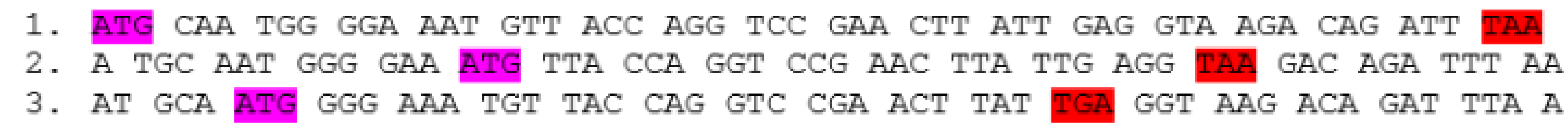
(c)



ORF

= open reading frame (a sequence of codons uninterrupted by a STOP codon)

ATGCAATGGGGAAATGTTACCAGGTCCGAACTTATTGAGGTAAGACAGATTTAA



```
1.  ATG CAA TGG GGA AAT GTT ACC AGG TCC GAA CTT ATT GAG GTA AGA CAG ATT TAA
2.  A TGC AAT GGG GAA ATG TTA CCA GGT CCG AAC TTA TTG AGG TAA GAC AGA TTT AA
3.  AT GCA ATG GGG AAA TGT TAC CAG GTC CGA ACT TAT TGA GGT AAG ACA GAT TTA A
```

There are 3 reading frames

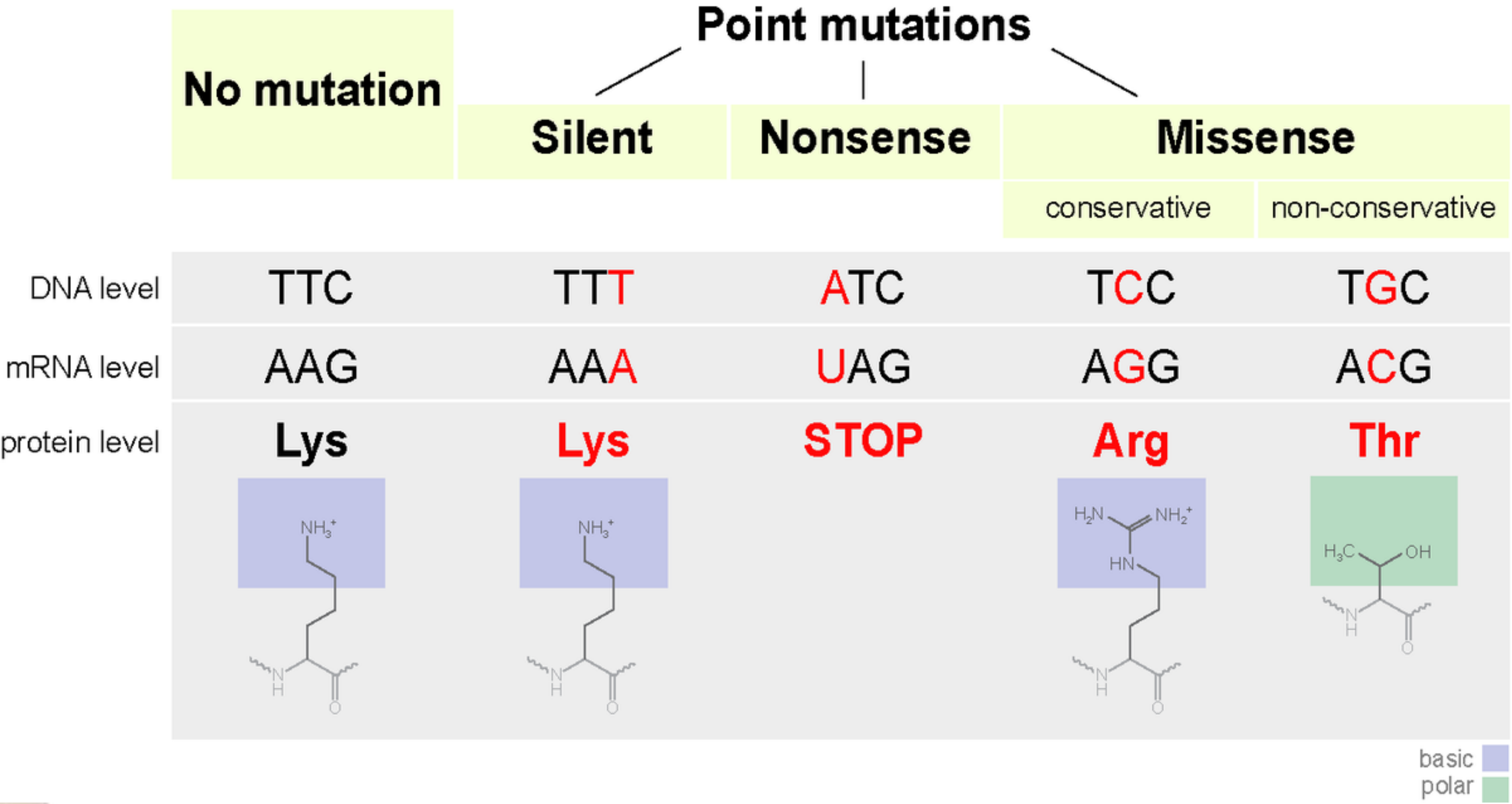
The distance between two sequences

Hamming distance

For two sequences of equal length the number of diferring positions

```
GAGCCTACTAACGGGAT
CATCGTAATGACGGCCT
```

Point mutations



		Point mutations			
	No mutation	Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	Lys	Lys	STOP	Arg	Thr

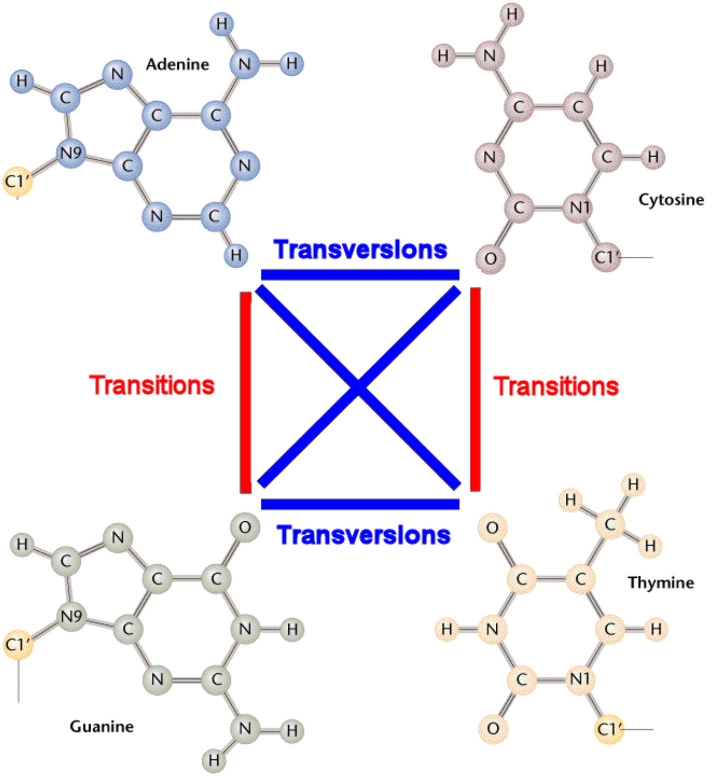
NH_3^+
NH_3^+
H_2N
NH_2^+
HN
H_3C
OH
N
H
O

basic 🟦
polar 🟩

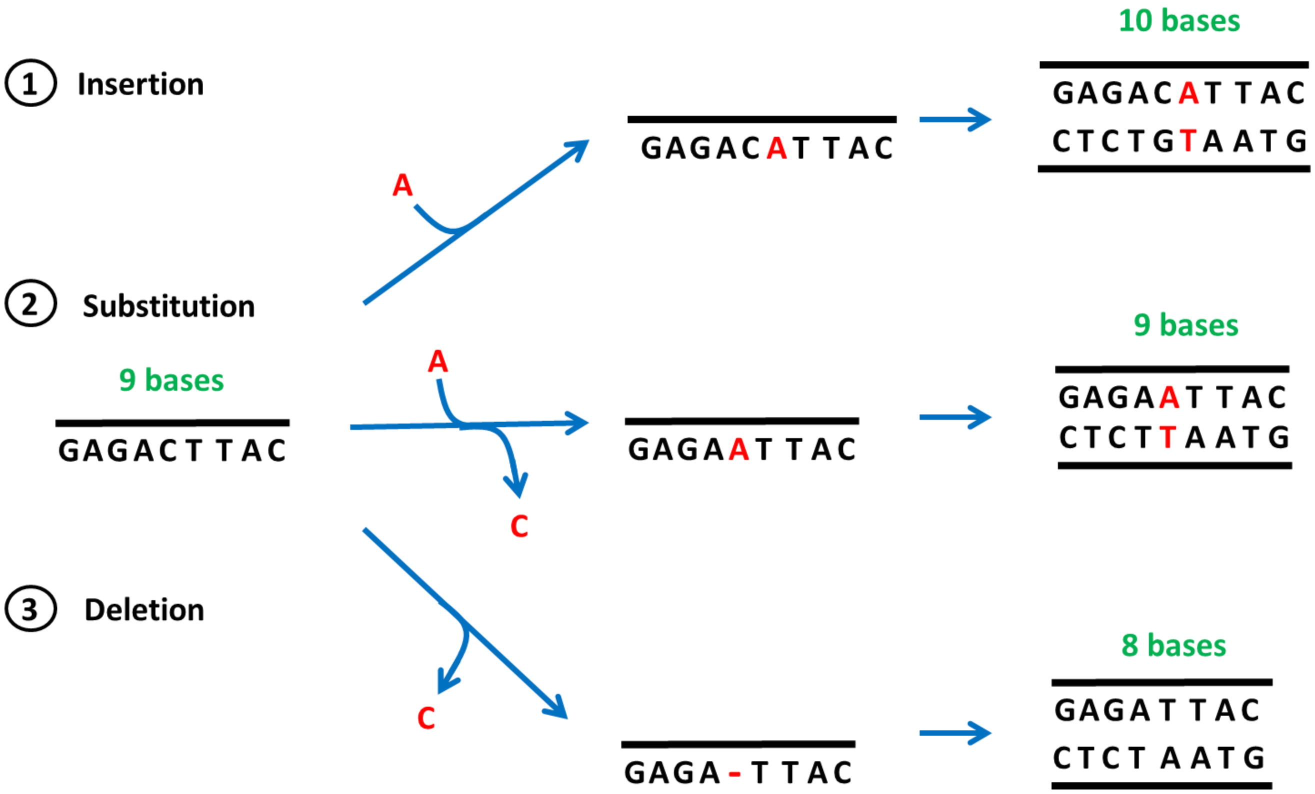
Transitions/transversions

- *Transition:* between A and G, or. C and T
- G → A or A → G
- C → T or T → C

- *Transversions*
- G → T or C
- A → T or C
- C → A or G
- T → A or G



Mutations



1 Insertion

2 Substitution

3 Deletion

9 bases

GAGACT TAC

A

GAGACAT TAC

10 bases

GAGACAT TAC

CTCTGTAATG

A

C

GAGAAT TAC

9 bases

GAGAAT TAC

CTCTTAATG

C

GAGA - T TAC

8 bases

GAGAT TAC

CTCT AATG

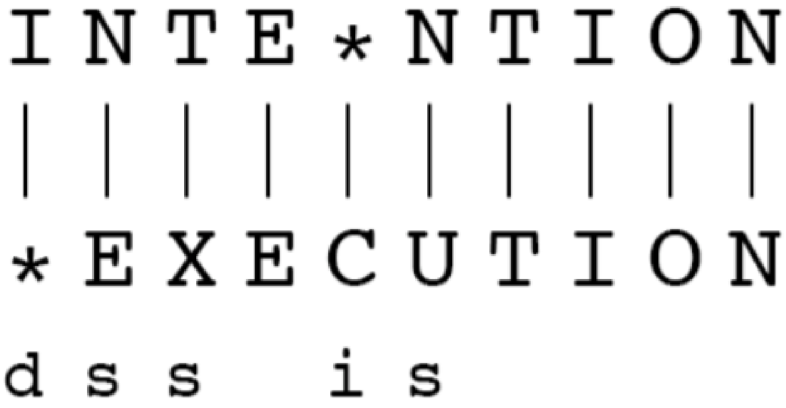
Source: www2.le.ac.uk

Edit distance

- transforming sequences
- transformation with the minimal number of changes
- operations: **insertions**

  deletions

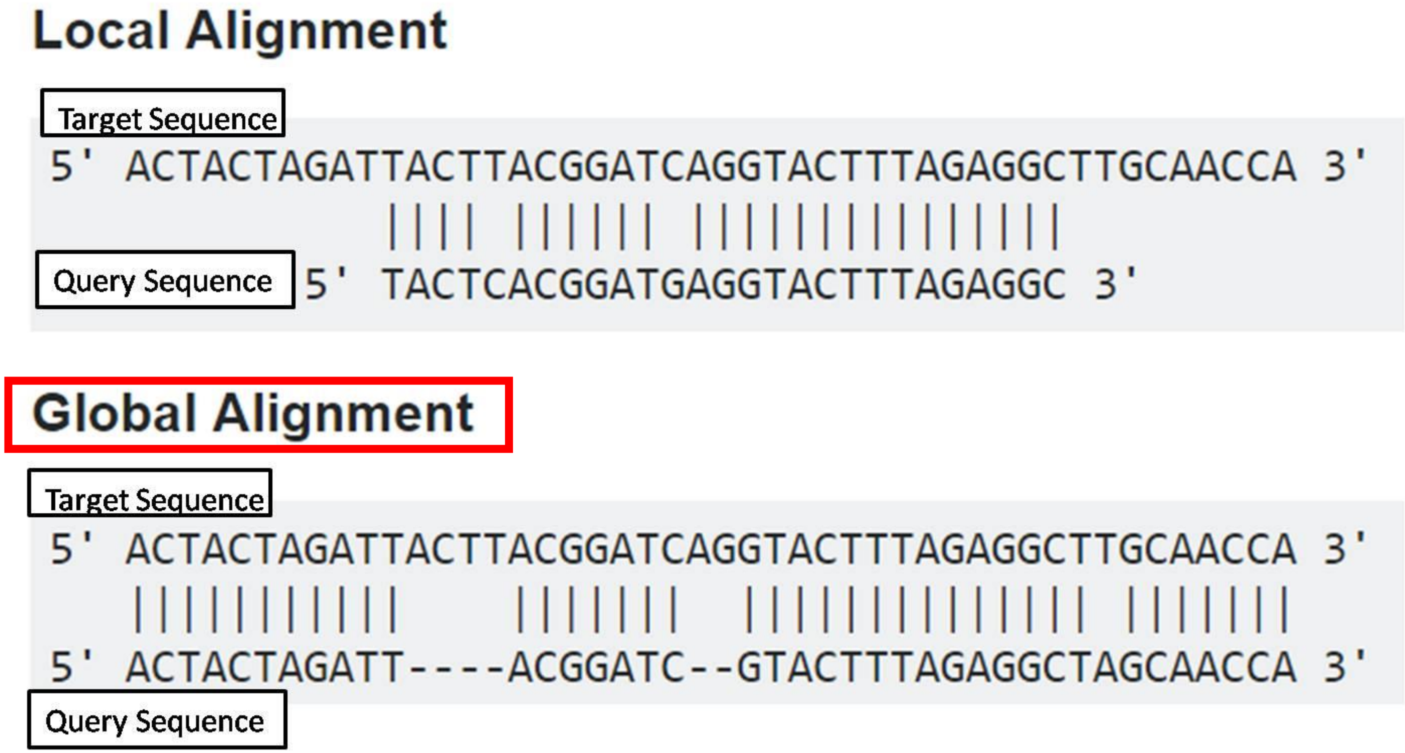
  supstitutions
- operations get penalties



```
I N T E * N T I O N
| | | | | | | | | |
* E X E C U T I O N
d s s   i s
```

Levenshtein edit distance

Used to align sequences



Local Alignment

Target Sequence

```
5'  ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA  3'
              |||| |||||| |||||||||||||||
          5'  TACTCACGGATGAGGTACTTTAGAGGC  3'
```

Query Sequence

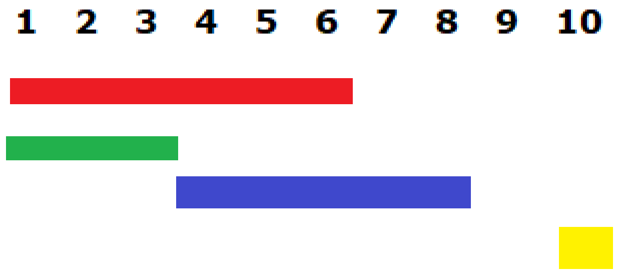
Global Alignment

Target Sequence

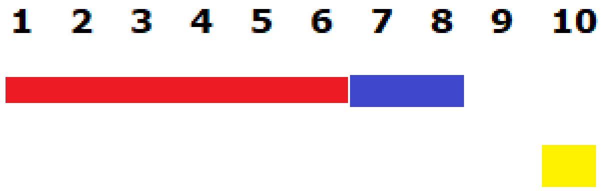
```
5'  ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA  3'
    |||||||||||    |||||||  |||||||||||||| |||||||
5'  ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA  3'
```

Query Sequence

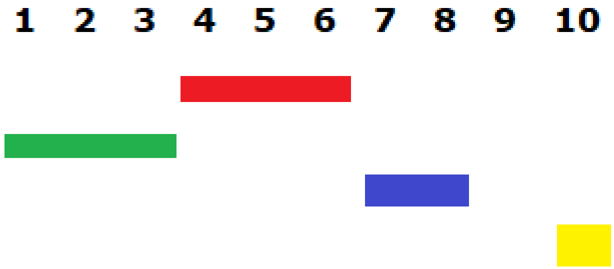
IRanges



reduce

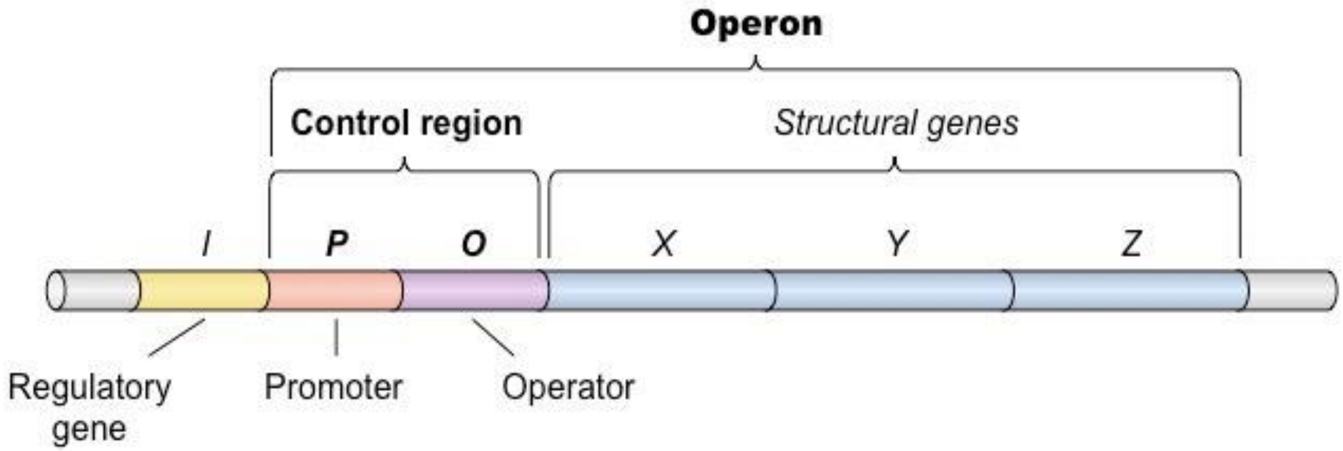


disjoin



The *E. coli* genome

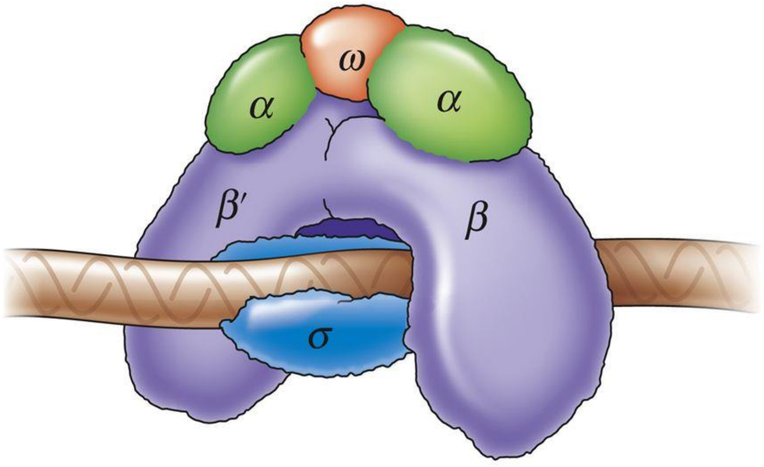
- 4.6 million nucleotides
- 4288 protein-coding genes
- 2584 operons



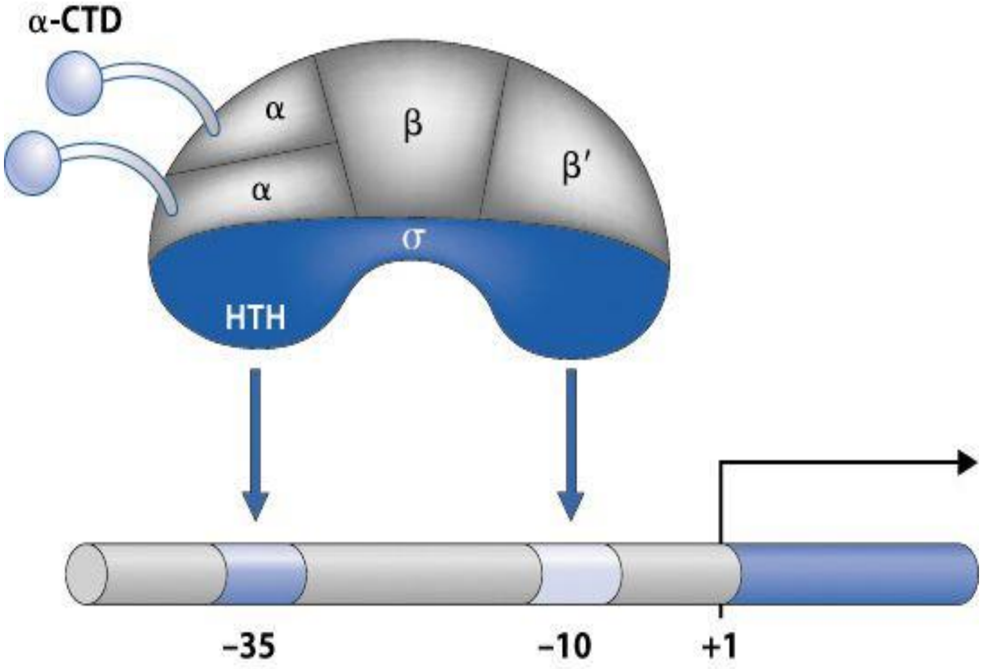
Transcription

Binding of the transcription initiation complex to the promoter

Figure 8.1 *E. coli* RNA polymerase

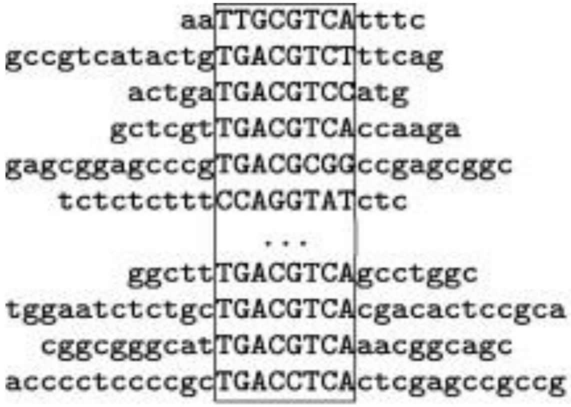


THE CELL: A MOLECULAR APPROACH 7e, Figure 8.1
© 2016 Sinauer Associates, Inc.

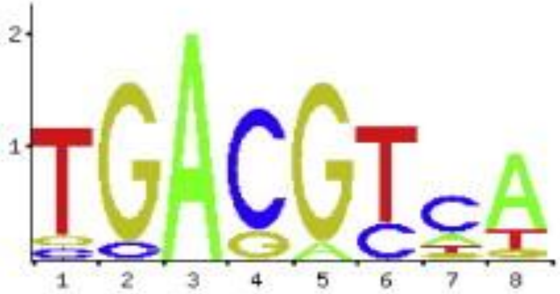


Final assignment

Discovering a motif:
left - a multiple sequence alignment
right - a motif logo
down - logarithmic position weight matrix



```
          aaTTGCGTCAtttc
gccgtcatactgTGACGTCTttcag
       actgaTGACGTCCatg
      gctcgtTGACGTCAccaaga
gagcggagcccgTGACGCGGccgagcggc
   tctctctttCCAGGTATctc
                ...
       ggcttTGACGTCAgcctggc
tggaatctctgcTGACGTCAcgacactccgca
  cggcgggcatTGACGTCAaacggcagc
acccctccccgcTGACCTCActcgagccgccg
```

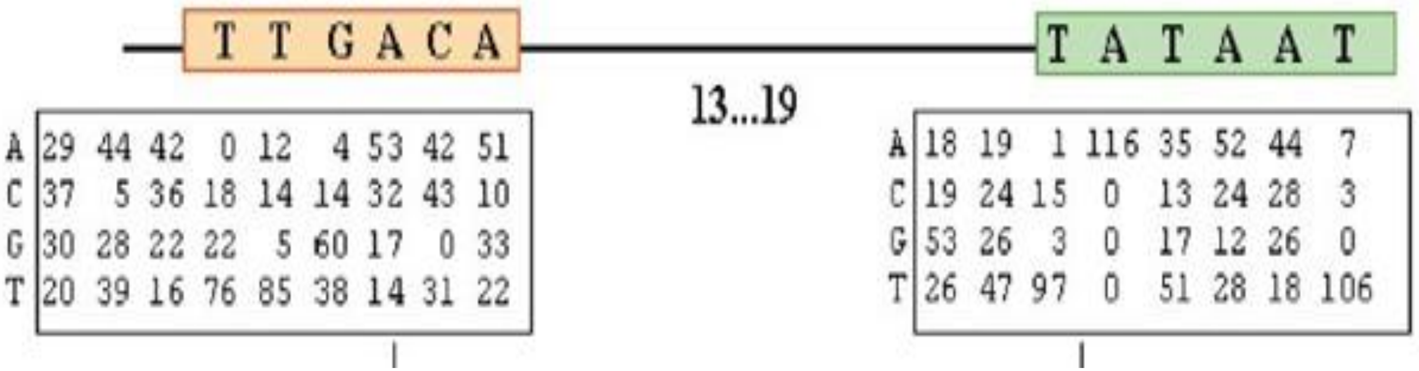


$$M(i,x) = \log_2 \frac{\text{frequency of letter } x \text{ at position } i}{\text{background frequency of letter } x} \tag{1}$$

$$\begin{array}{cc}
A & [\,-3.219 \quad -3.219 \quad 3.785 \quad -3.219 \quad 1.396 \quad -3.219 \quad 2.084 \quad 3.467\,] \\
C & [\,1.396 \quad 1.396 \quad -3.219 \quad 3.585 \quad -3.219 \quad 2.488 \quad 3.334 \quad -3.219\,] \\
G & [\,1.396 \quad 3.690 \quad -3.219 \quad 2.084 \quad 3.690 \quad -3.219 \quad 1.396 \quad 1.396\,] \\
T & [\,3.585 \quad -3.219 \quad -3.219 \quad -3.219 \quad -3.219 \quad 3.467 \quad 1.396 \quad 2.084\,]
\end{array}$$

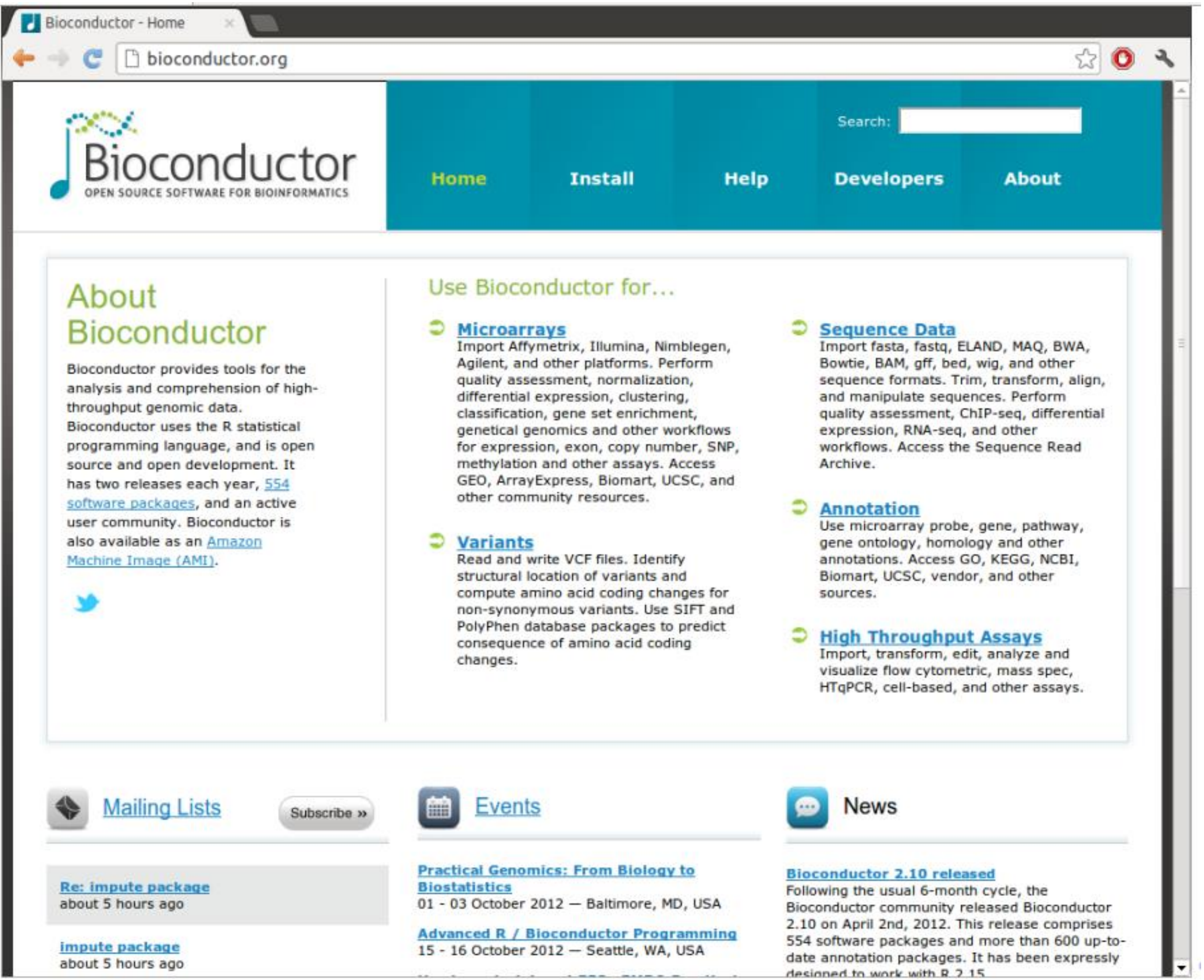
Final assignment

The consensus sequences (TTAGACA and TATAAT) and the position weight matrices for the two motifs in the sigma70 binding site



A	29	44	42	0	12	4	53	42	51
C	37	5	36	18	14	14	32	43	10
G	30	28	22	22	5	60	17	0	33
T	20	39	16	76	85	38	14	31	22

A	18	19	1	116	35	52	44	7
C	19	24	15	0	13	24	28	3
G	53	26	3	0	17	12	26	0
T	26	47	97	0	51	28	18	106



About Bioconductor

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, 554 software packages, and an active user community. Bioconductor is also available as an Amazon Machine Image (AMI).

Use Bioconductor for...

Microarrays
Import Affymetrix, Illumina, Nimblegen, Agilent, and other platforms. Perform quality assessment, normalization, differential expression, clustering, classification, gene set enrichment, genetical genomics and other workflows for expression, exon, copy number, SNP, methylation and other assays. Access GEO, ArrayExpress, Biomart, UCSC, and other community resources.

Variants
Read and write VCF files. Identify structural location of variants and compute amino acid coding changes for non-synonymous variants. Use SIFT and PolyPhen database packages to predict consequence of amino acid coding changes.

Sequence Data
Import fasta, fastq, ELAND, MAQ, BWA, Bowtie, BAM, gff, bed, wig, and other sequence formats. Trim, transform, align, and manipulate sequences. Perform quality assessment, ChIP-seq, differential expression, RNA-seq, and other workflows. Access the Sequence Read Archive.

Annotation
Use microarray probe, gene, pathway, gene ontology, homology and other annotations. Access GO, KEGG, NCBI, Biomart, UCSC, vendor, and other sources.

High Throughput Assays
Import, transform, edit, analyze and visualize flow cytometric, mass spec, HTqPCR, cell-based, and other assays.

Mailing Lists

Re: impute package
about 5 hours ago

impute package
about 5 hours ago

Events

Practical Genomics: From Biology to Biostatistics
01 - 03 October 2012 — Baltimore, MD, USA

Advanced R / Bioconductor Programming
15 - 16 October 2012 — Seattle, WA, USA

News

Bioconductor 2.10 released
Following the usual 6-month cycle, the Bioconductor community released Bioconductor 2.10 on April 2nd, 2012. This release comprises 554 software packages and more than 600 up-to-date annotation packages. It has been expressly designed to work with R 2.15