

Bioinformatics overview + sequence alignment

Martin Reczko

Staff research scientist professor level

Institute for Fundamental Biomedical Science

Biomedical Sciences Research Center "Alexander Fleming"

Head of Node - ELIXIR-Greece



Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών
— ΙΔΡΥΘΕΝ ΤΟ 1837 —

Search... 🔍

▼ Course Options

📅 Agenda

📢 Announcements

🧪 Assignments

📁 Documents 6

📝 Exercises

🔗 Links

✉ Messages 1

❓ Questionnaires

🏠 Portfolio / Introduction to Bioinformatics

Introduction to Bioinformatics (M413)

Martin Reczko - Alexandros Dimopoulos



Description



The course introduces students into the basic concepts of bioinformatics. It starts with a general overview of the various fields of bioinformatics and introduces dynamic programming as a solution to the sequence comparison problem (1). Next, a first introduction to the GNU / Linux operating system and the hands on use of basic command-line commands (CLI) as well as bash scripting is given. In addition, basic bioinformatics command line programs such as bedtools, vcftools, samtools, etc. are presented and used (2+3). Students are then familiarized with the programming language R, the use of IDE RStudio and the basic tools provided by the Bioconductor repository (4+5). Next, detailed examples of

NGS bioinformatics analysis and pipelines are explained for:

- RNAseq (quality control, gene expression analysis) (6),
- denovo assembly (both on the genome and transcriptome level) (7)
- ChipSeq, ClipSeq and (8)
- variant calling (exome sequencing example using GATK) (9)

Finally, the concept of flux

More ↓



Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Ἀθηνῶν
— ΔΙΔΡΥΘΕΝ ΤΟ 1837

Search...

Course Options

- Agenda
- Announcements
- Assignments
- Documents
- Exercises
- Links
- Messages 1
- Questionnaires



mareczko



Portfolio / Introduction to Bioinformatics / Documents

Introduction to Bioinformatics (M413)

Documents



Root directory









Type	First Name	Size	Date	
	2022-23		10/12/22	
	FOSSwire Unix/Linux Command Cheat Sheet	69.09 KB	10/19/17	
	Grades_February 2022	86.83 KB	4/15/22	
	How to start RStudio within X2Go for the first time	411.19 KB	11/5/21	
	Hypatia VMs: IPs and public keys		10/11/22	



Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Ἀθηνῶν
ΙΔΡΥΘΕΝ ΤΟ 1837

Search... 

Course Options

-  Agenda
-  Announcements
-  Assignments
-  Documents
-  Exercises
-  Links
-  Messages 1
-  Questionnaires



 mareczko



 Portfolio / Introduction to Bioinformatics / Documents


Introduction to Bioinformatics (M413)



Documents

Root directory » 2022-23 

 Up

Type	First Name 	Size	Date	
	exercises		10/12/22	
	lectures		10/12/22	

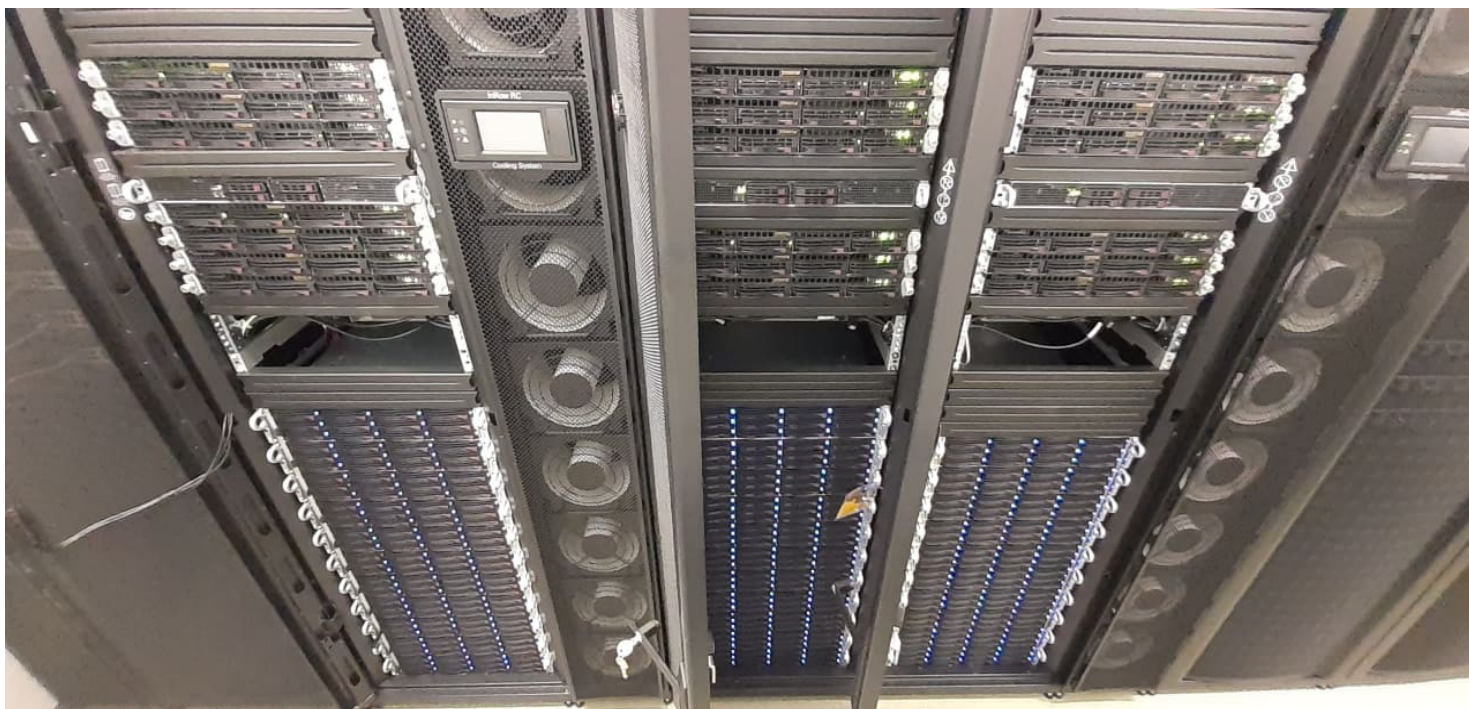
Please verify name+email in participant list at

<https://tinyurl.com/3hm8ze43>

Enter all emails you might use
(to get an account on the virtual machine from



28 CPUs, 242GB RAM, 800 GB disk shared for all



Syllabus and grading

#	Date	Short title	Lecturer	Other actions	Subject
1	Tuesday, October 10, 2023	introduction	MR		Overview of Bioinformatics, sequence alignment
2	Tuesday, October 17, 2023	Linux/shell/ssh	AD		Introduction to Linux and the command line, bash scripting and ssh
3	Tuesday, October 24, 2023	R (1)	AD		Introduction to the R programming language and Rstudio usage
4	Tuesday, October 31, 2023	QC+RNASeq	MR		Next generation sequencing: introduction, quality control and gene expression analysis for RNAseq
5	Tuesday, November 7, 2023	R (2)	AD		Advances R subjects, introduction to Bioconductor
6	Tuesday, November 14, 2023	bedtools/vcftools/samtools	AD		Command line tool usage: bedtools, vcftools, samtools etc.
7	Tuesday, November 21, 2023	Denovo	MR		NGS for denovo genome and transcriptome assembly
8	Tuesday, November 28, 2023	ChipSeq/chirp	MR	assign presentations	NGS analysis for molecular interactions (ChipSeq, (Par-)Clip, structural sequencing, chromosome conformation capture (3C))
9	Tuesday, December 5, 2023	metabolomics	MR		Genome-scale models of metabolism and macromolecular expression, Biological applications of Transformers
10	Tuesday, December 12, 2023	Exome/SNP calling	AD	assign final projects	Pipelines for SNP calling, especially for exome sequencing using the GATK pipeline
11	Tuesday, December 19, 2023	presentations	MR+AD		Paper presentations by students
12	Tuesday, January 9, 2024	presentations	MR+AD		Paper presentations by students
13	Tuesday, January 16, 2024	final projects support	MR+AD		Support for the final project

Grade	100%
Presentation	30%
Exercises	20%
Final Project	50%

Subjects:

'Just enough' biology

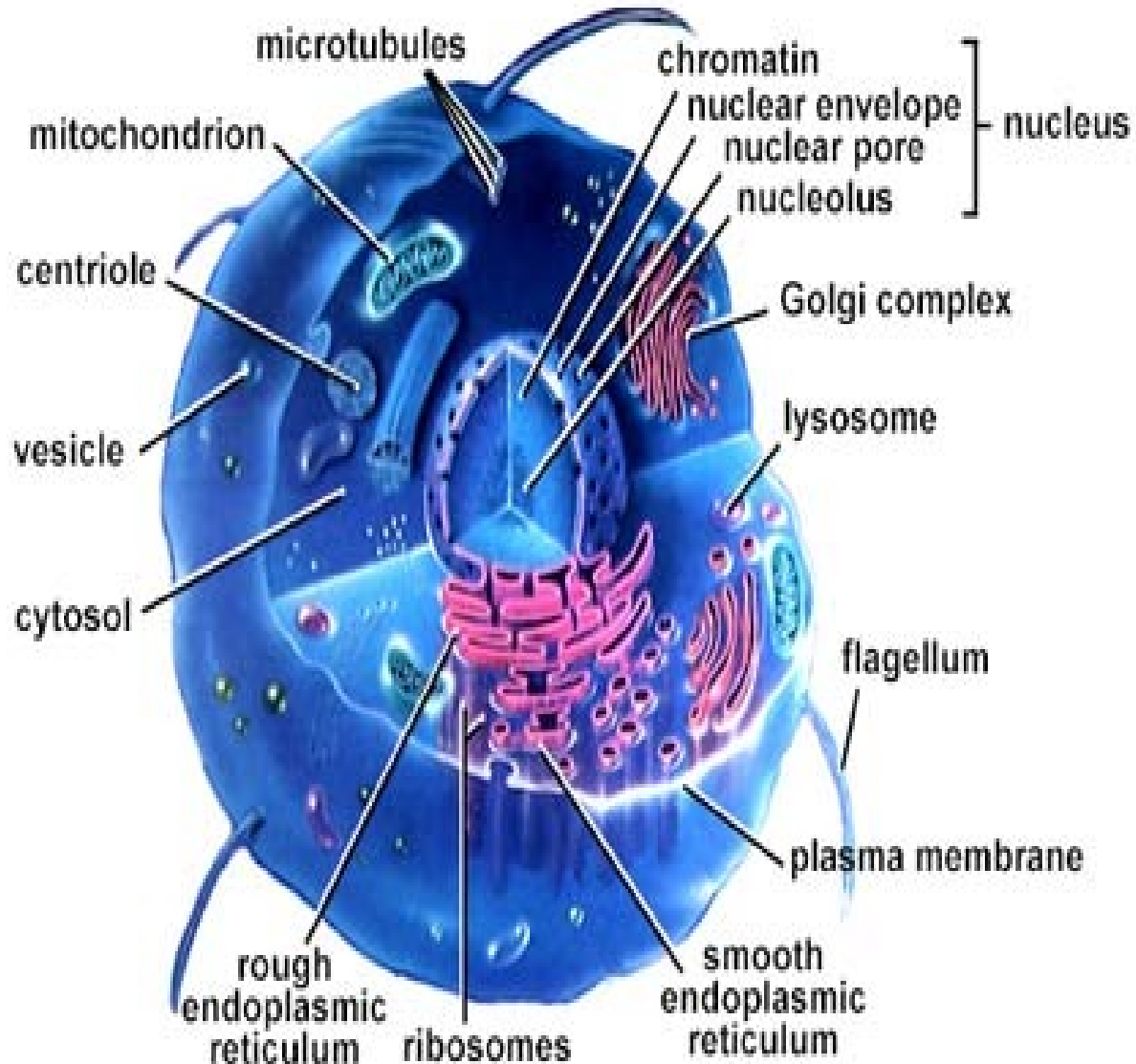
Dynamic programming

Approximate string similarity

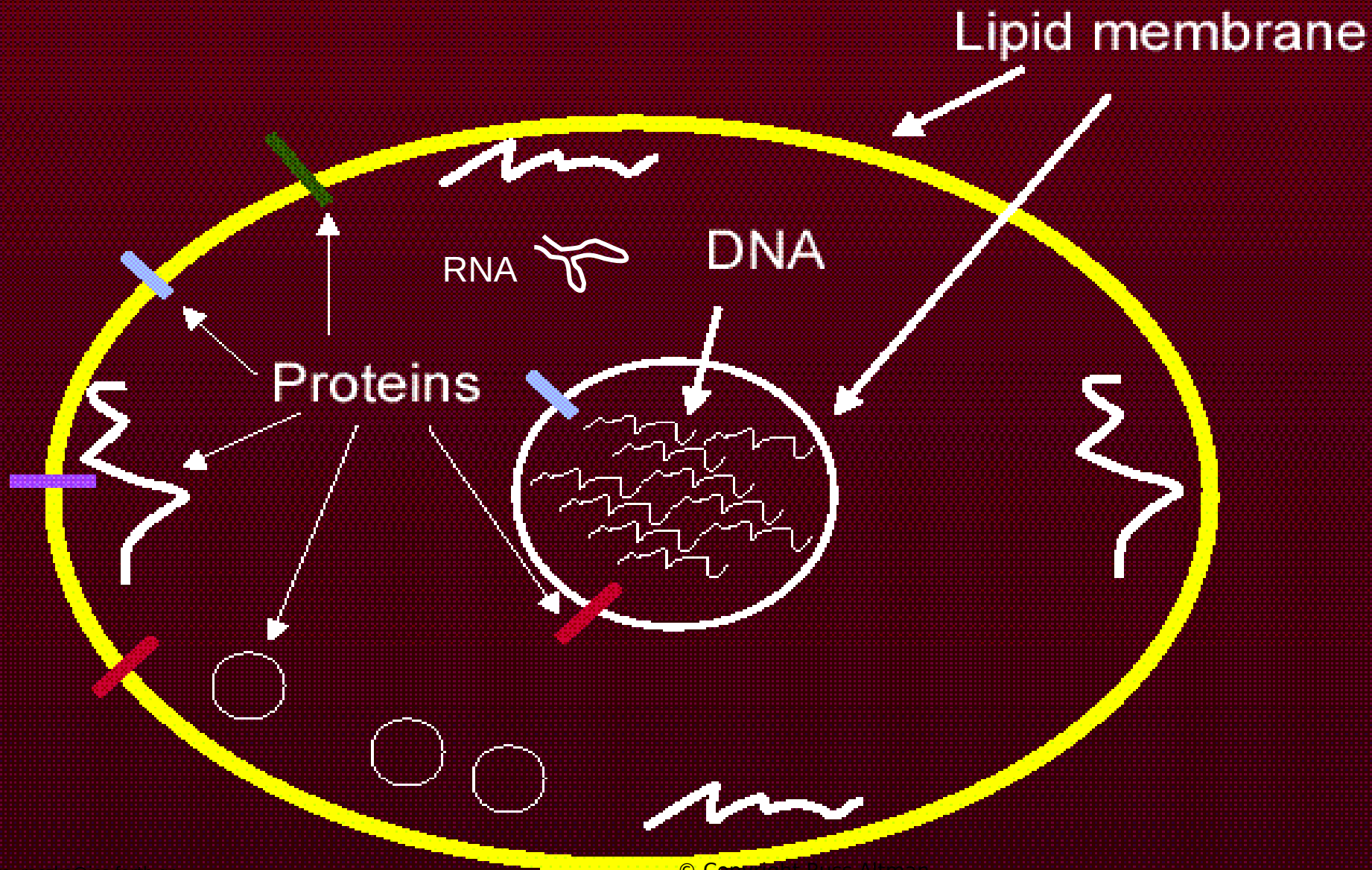
Bioinformatics fields

Recent machine learning results

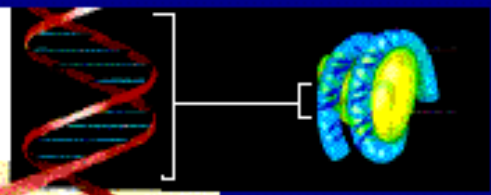
An Eukaryotic Cell (biological view)



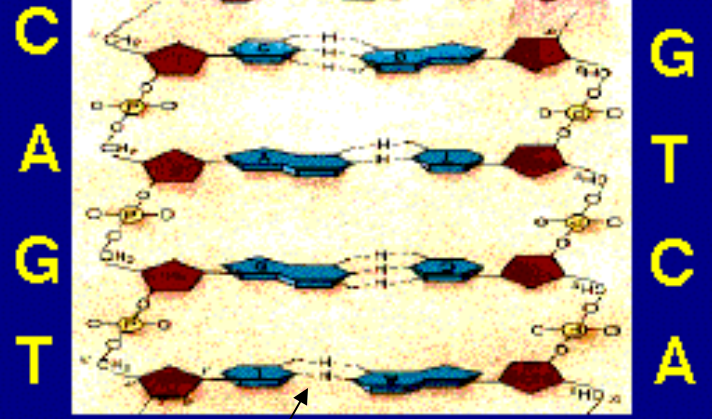
Bioinformatics Schematic of a Cell



THE DNA DOUBLE HELIX



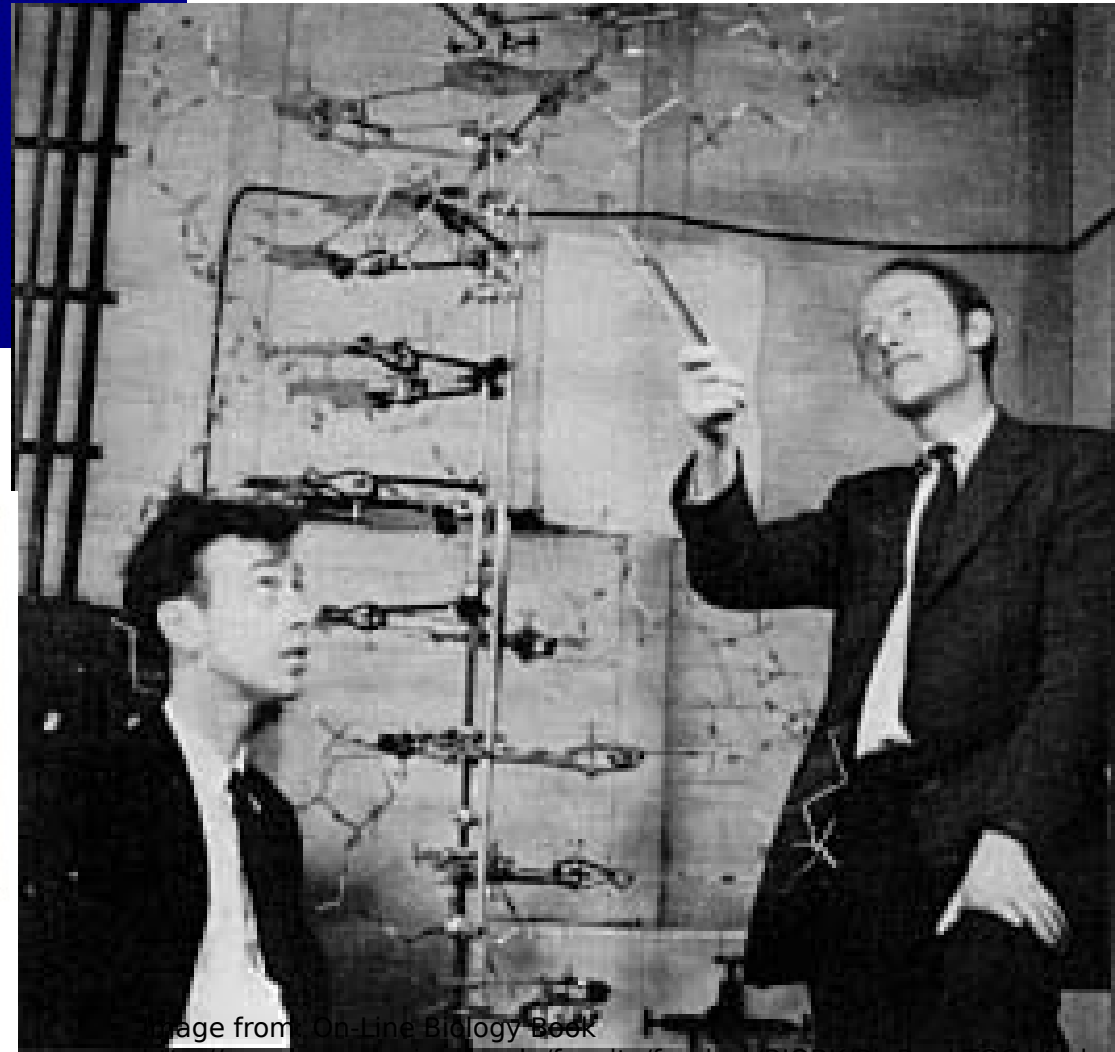
Watson and Crick, 1953 in Cambridge



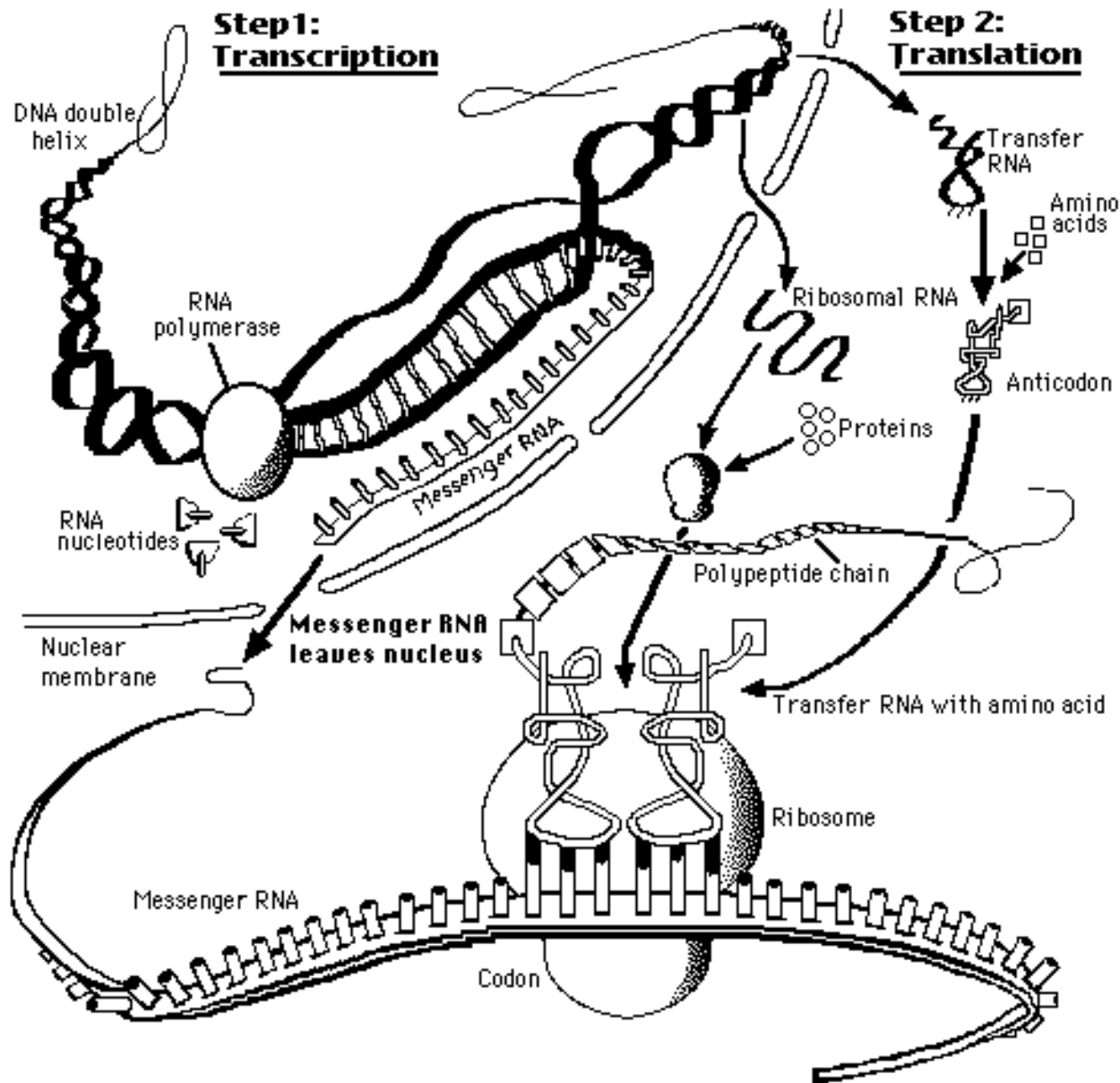
Hydrogen bonds (->Hybridization)



Rosalind Franklin



PROTEIN SYNTHESIS



Transcription

transcription is accomplished by RNA polymerase

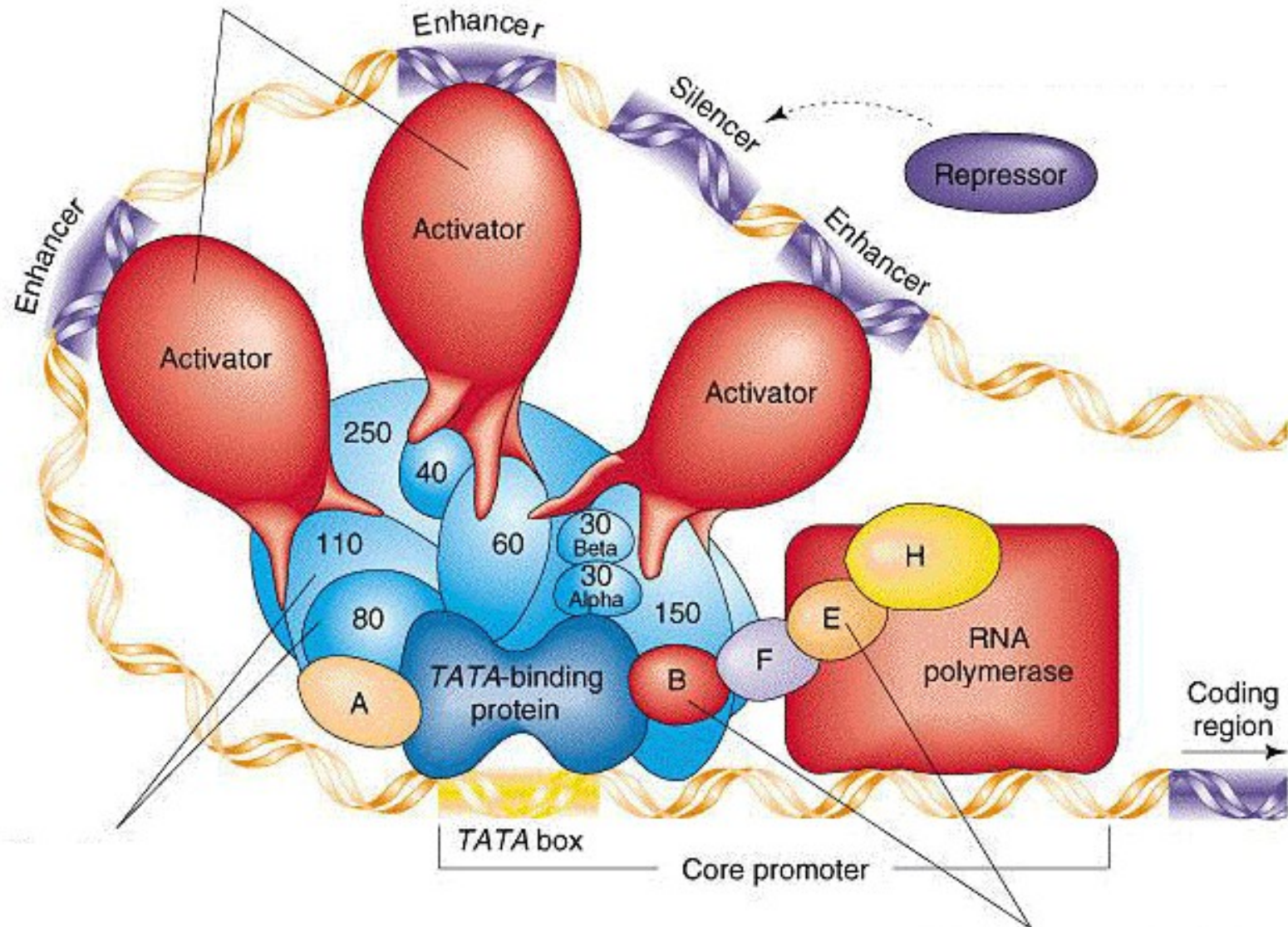
RNA polymerase binds to **promoters**

promoters have distinct regions "-35" and "-10"

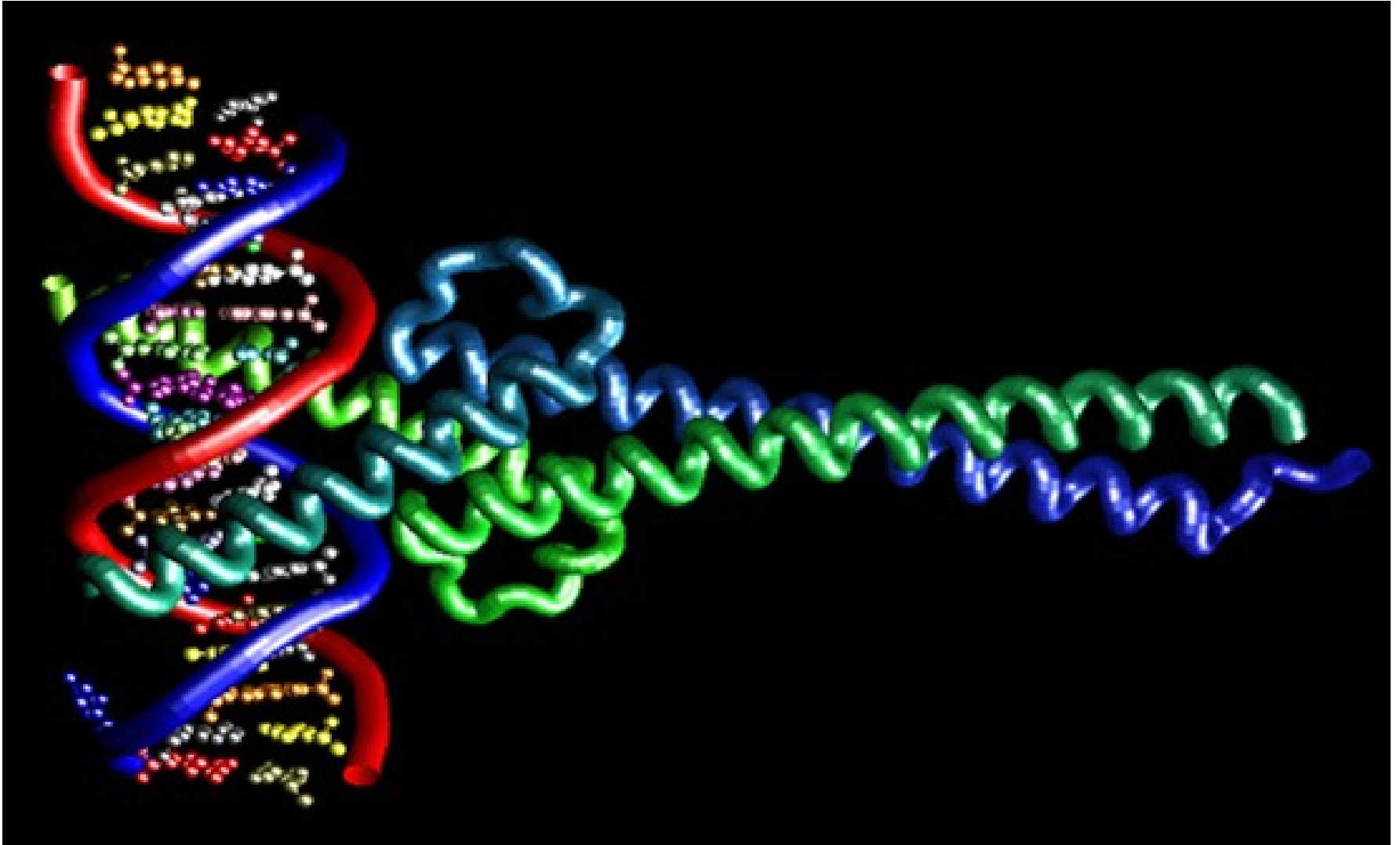
transcription start and stop affected by DNA structure

Additional regulatory sequences can be positive or negative

Complete Assembly of Eukaryotic Gene Regulatory System



Interaction of a transcription factor and DNA



Myc Proto-Oncogene Protein, causing cell division and proliferation

Transcription: DNA → RNA

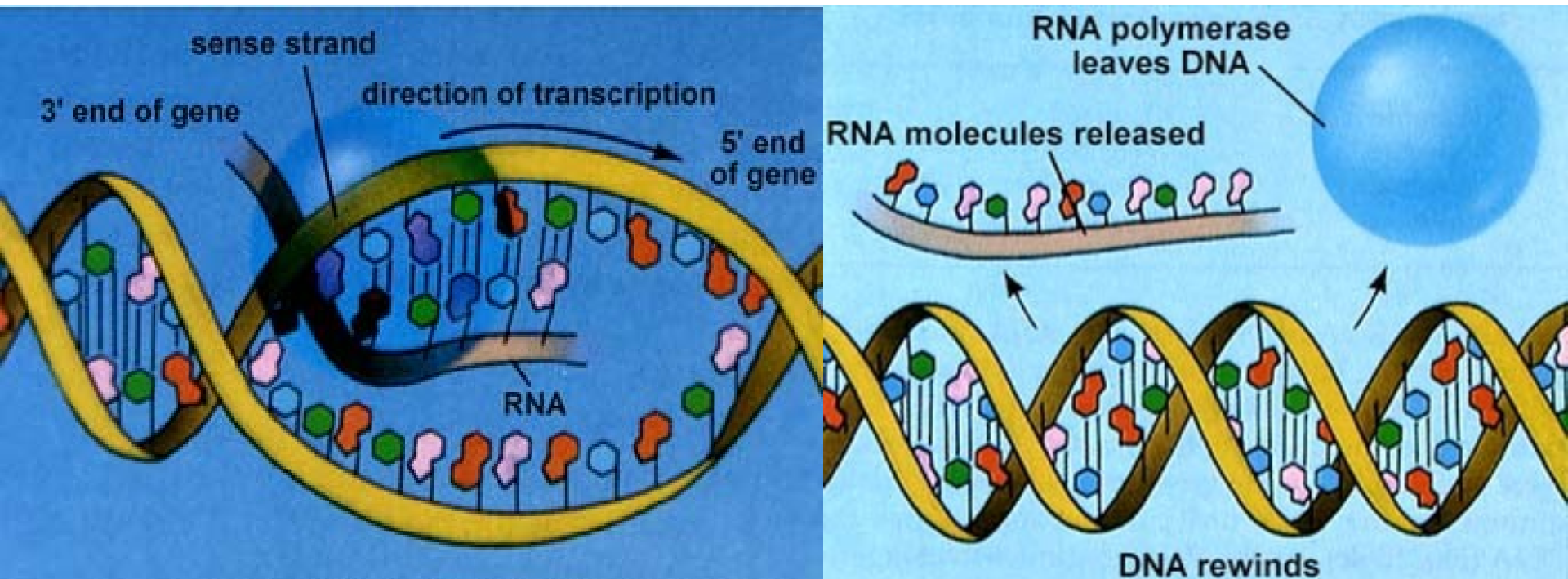


Image from: On-Line Biology Book
<http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookTOC.html>

RNA processing

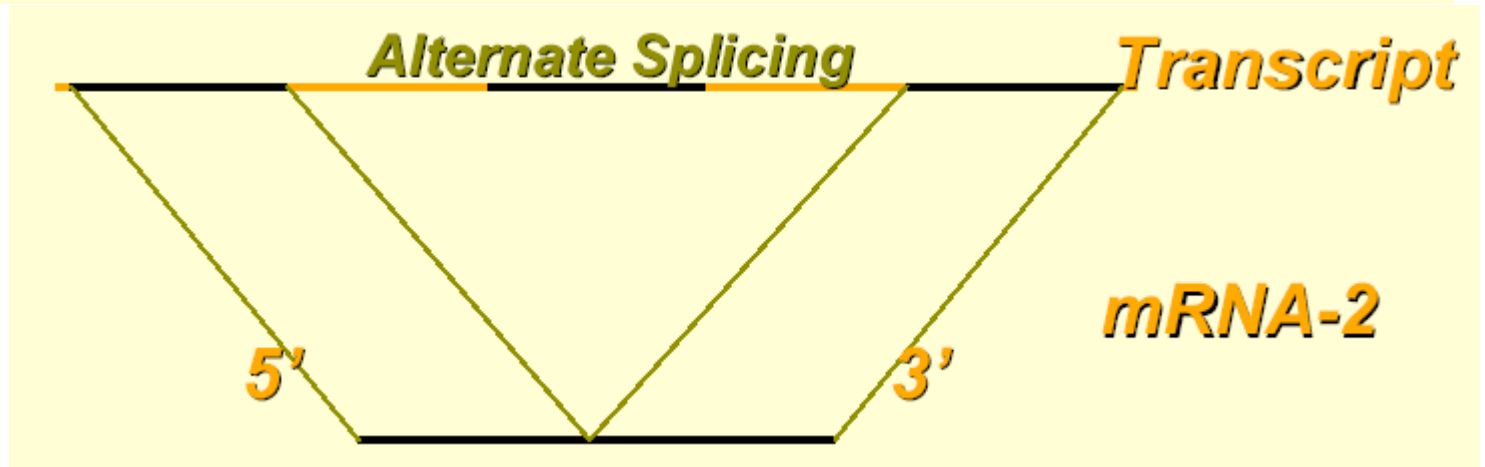
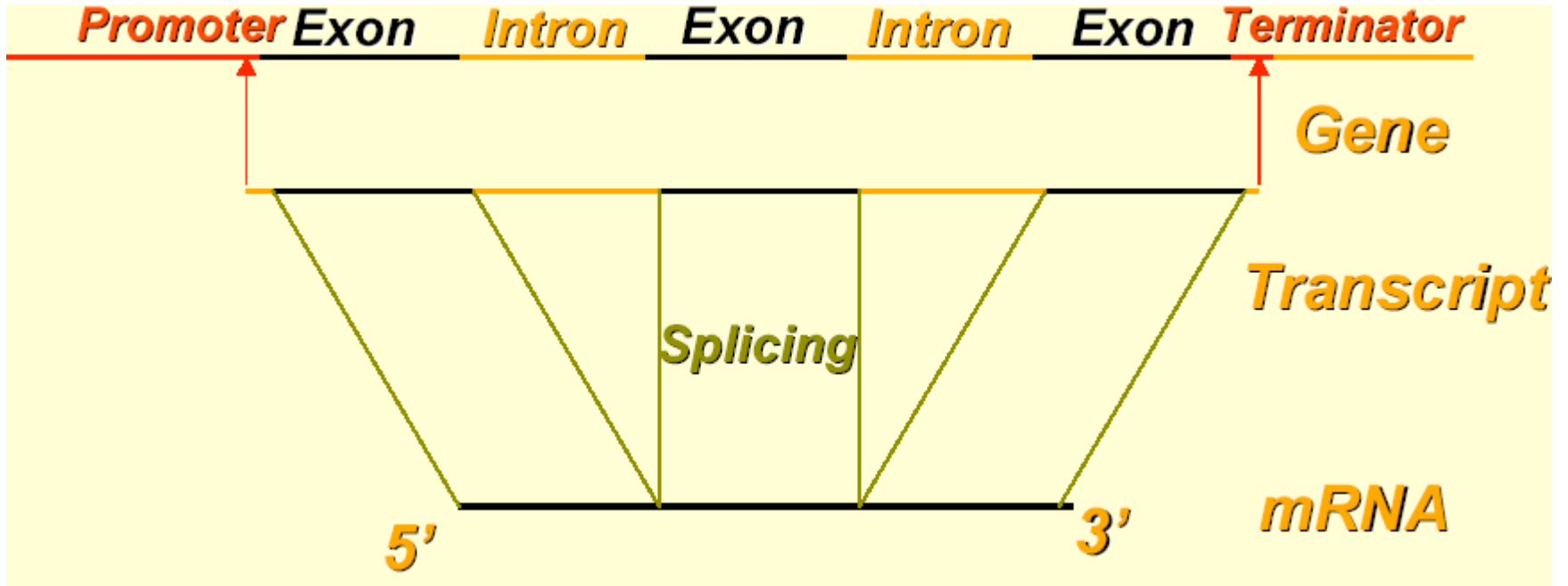
eukaryotic genes are interrupted by **introns**

these are "spliced" out to yield final messenger RNA (mRNA)

splicing done by spliceosomes

splicing sites are quite degenerate but not all are used

Processing of RNA = splicing



Images from: <http://biochem218.stanford.edu> (Doug Brutlag)

Translation

conversion from RNA to protein is by

codon: 3 bases = 1 amino acid

translation done by ribosome

translation stops after reading the stop
codon

Building proteins:

Elongation (translation)

messenger RNA (mRNA)

Ribosomes (rRNA)

transfer RNA (tRNA)

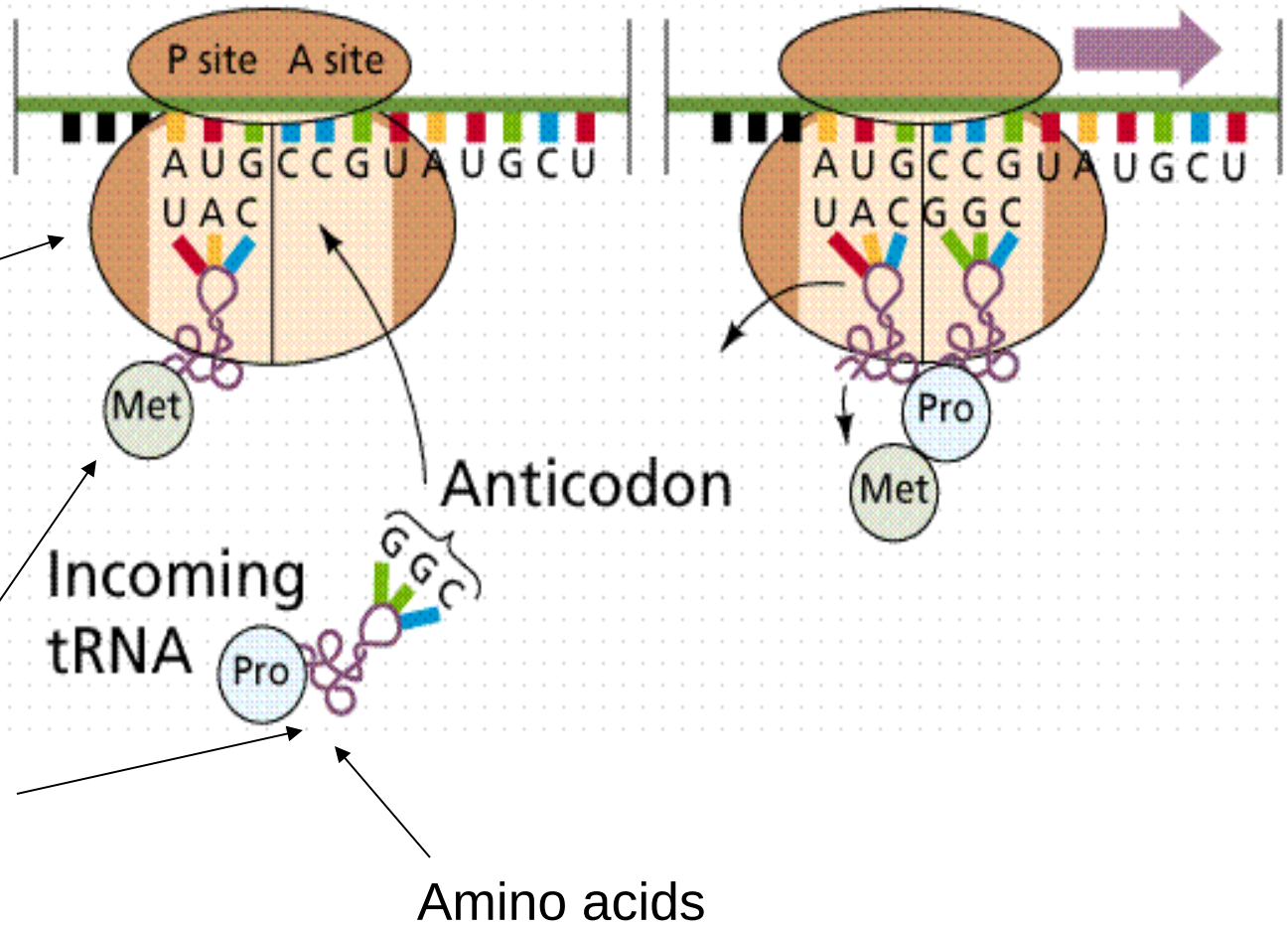
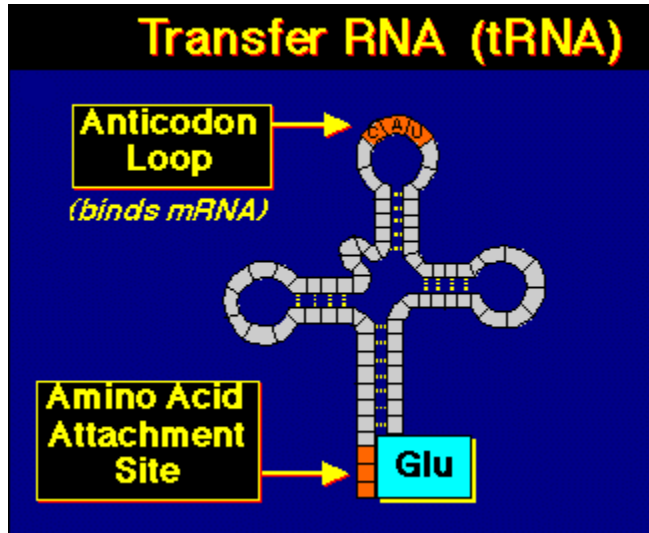


Image from: On-Line Biology Book

<http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookTOC.html>

The 'universal' genetic code:



64 different transfer RNA molecules

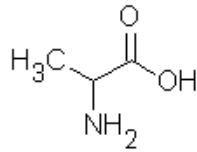


Second letter

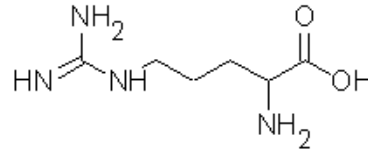
First letter

	U		C		A		G	
U	UUU	Phenyl-alanine	UCU	Serine	UAU	Tyrosine	UGU	Cysteine
	UUC		UCC		UAC		UGC	
	UUA	Leucine	UCA		UAA	Stop codon	UGA	Stop codon
	UUG		UCG		UAG		UGG	
C	CUU	Leucine	CCU	Proline	CAU	Histidine	CGU	Arginine
	CUC		CCC		CAC		CGC	
	CUA		CCA		CAA	Glutamine	CGA	
	CUG		CCG		CAG		CGG	
A	AUU	Isoleucine	ACU	Threonine	AAU	Asparagine	AGU	Serine
	AUC		ACC		AAC		AGC	
	AUA	Methionine; initiation codon	ACA		AAA	Lysine	AGA	Arginine
	AUG		ACG		AAG		AGG	
G	GUU	Valine	GCU	Alanine	GAU	Aspartic acid	GGU	Glycine
	GUC		GCC		GAC		GGC	
	GUA		GCA		GAA	Glutamic acid	GGA	
	GUG		GCG		GAG		GGG	

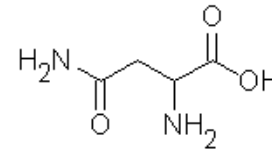
The 20 amino acids, building blocks for proteins



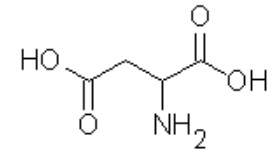
Alanin (Ala)



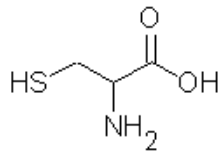
Arginin (Arg)



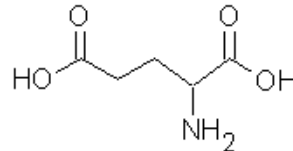
Asparagin (Asn)



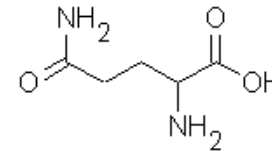
Asparaginsäure (Asp)



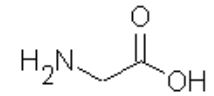
Cystein (Cys)



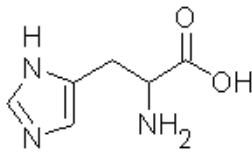
Glutaminsäure (Glu)



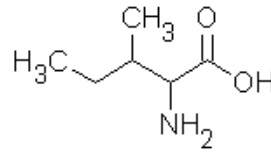
Glutamin (Gln)



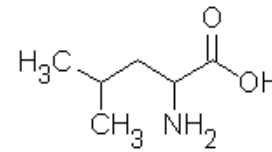
Glycin (Gly)



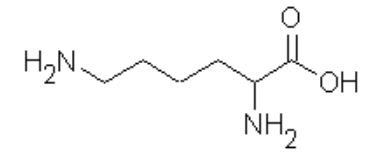
Histidin (His)



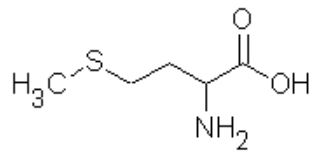
Isoleucin (Ile)



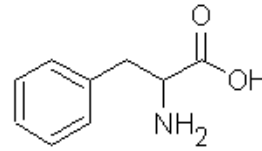
Leucin (Leu)



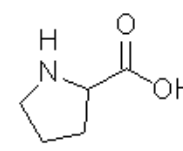
Lysin (Lys)



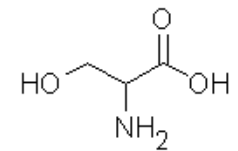
Methionin (Met)



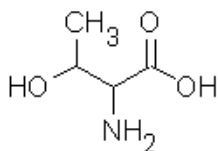
Phenylalanin (Phe)



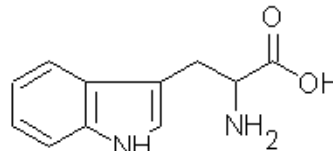
Prolin (Pro)



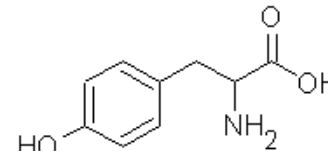
Serin (Ser)



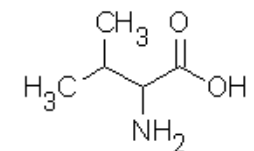
Threonin (Thr)



Tryptophan (Trp)



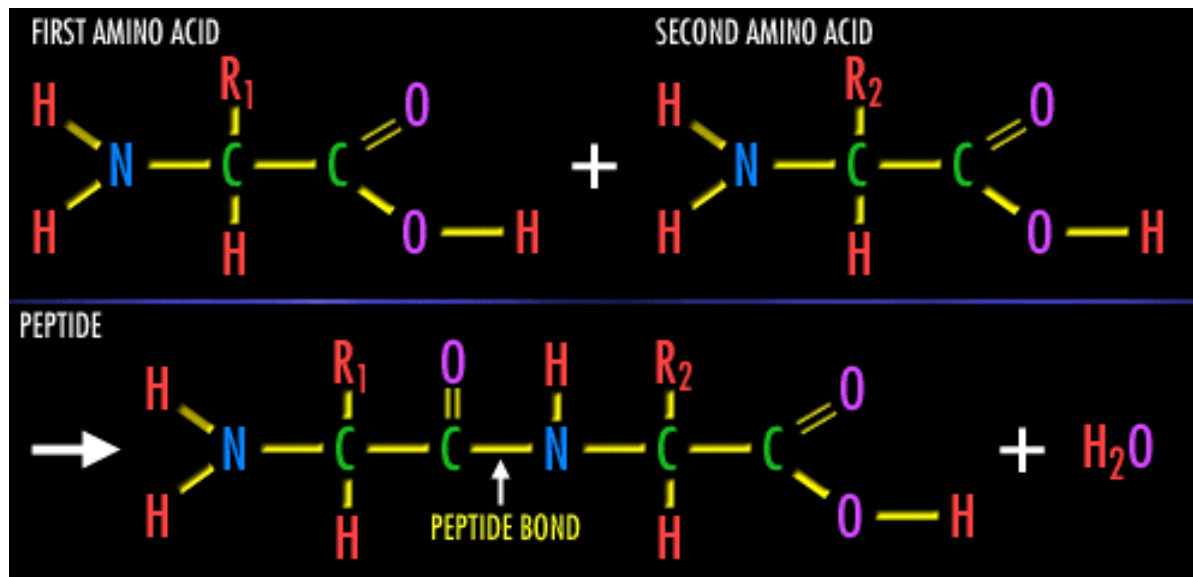
Tyrosin (Tyr)



Valin (Val)

Building proteins (chemistry):

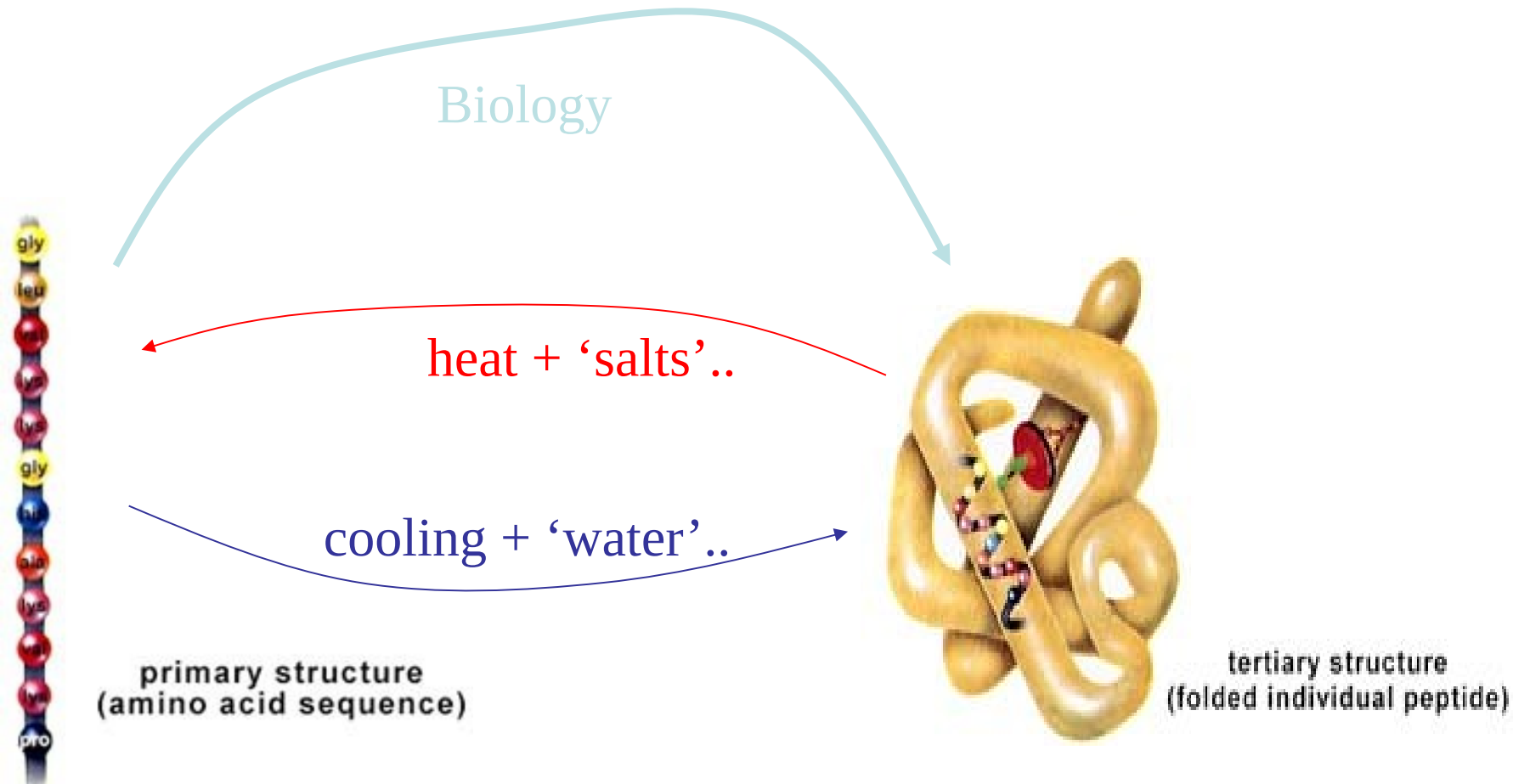
Amino acids are linked together by joining the amino end of one molecule to the carboxyl end of another. Removal of water allows formation of a type of covalent bond known as a peptide bond.



The above image is from

<http://zebu.uoregon.edu/internet/images/peptide.gif>.

Protein folding: Sequence determines structure



C. Anfinsen, 1973

The above images are from

http://www.biosci.uga.edu/almanac/bio_103/notes/may_14.html.

Levels of structural description



primary structure
(amino acid sequence)



secondary structure
(α -helix)



tertiary structure
(folded individual peptide)



quaternary structure
(aggregation of two or more peptides)

The above images are from

http://www.biosci.uga.edu/almanac/bio_103/notes/may_14.html.

Protein localization

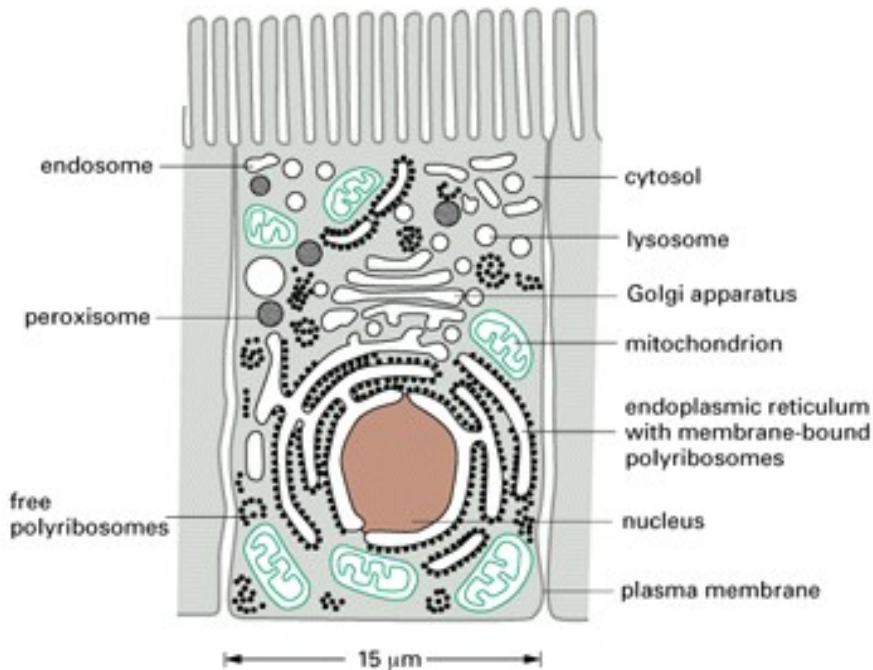
leader sequences can specify cellular location (e.g., insert across membranes)

leader sequences usually removed by cleavage

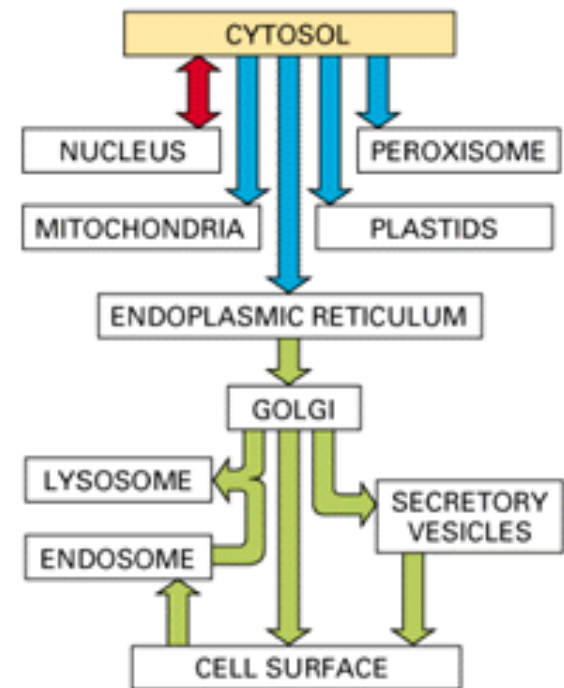
Like an address sticker

Protein localization

compartments



protein traffic



KEY: **Red** = gated transport
Blue = transmembrane transport
Green = vesicular transport

UNFOLDED PROTEIN

FOLDED PROTEIN

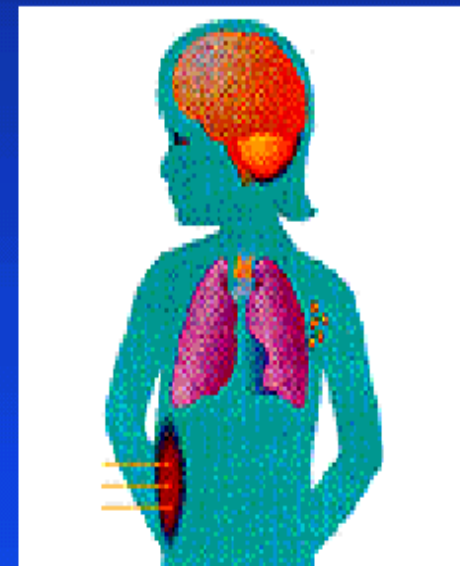
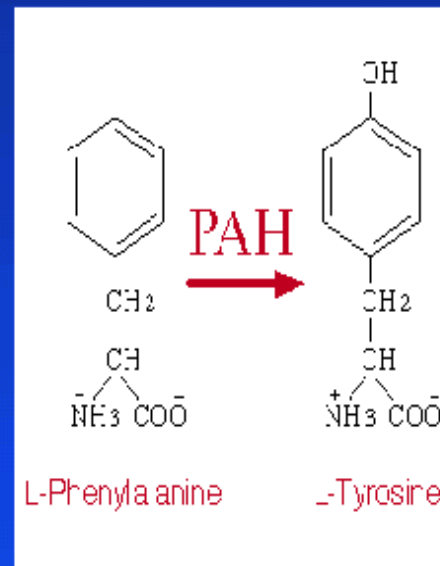
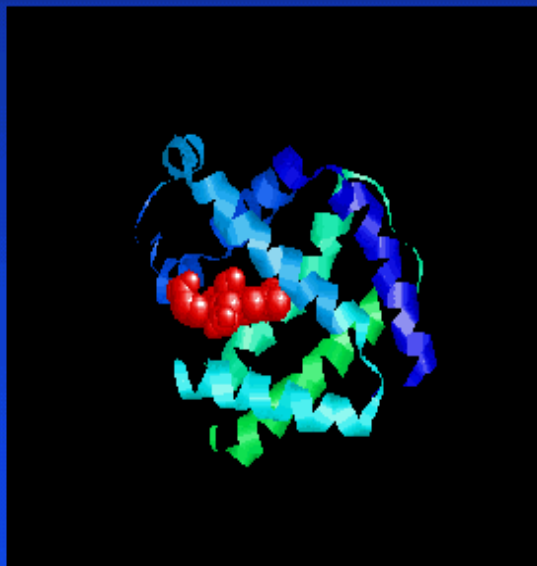


(A)

Central Paradigm of Bioinformatics



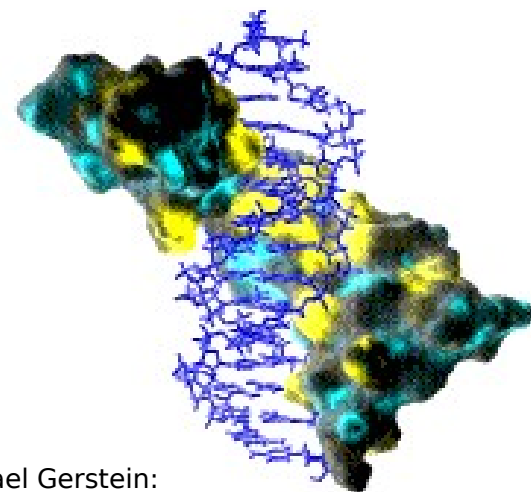
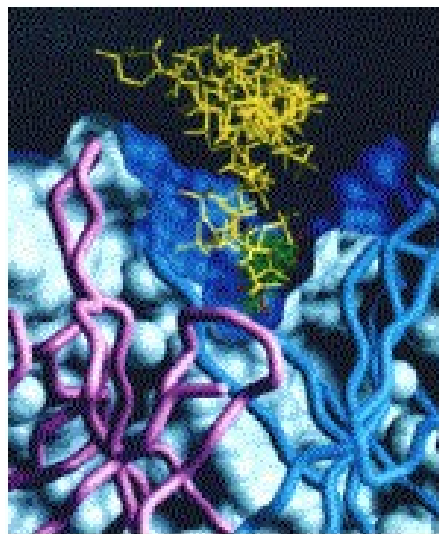
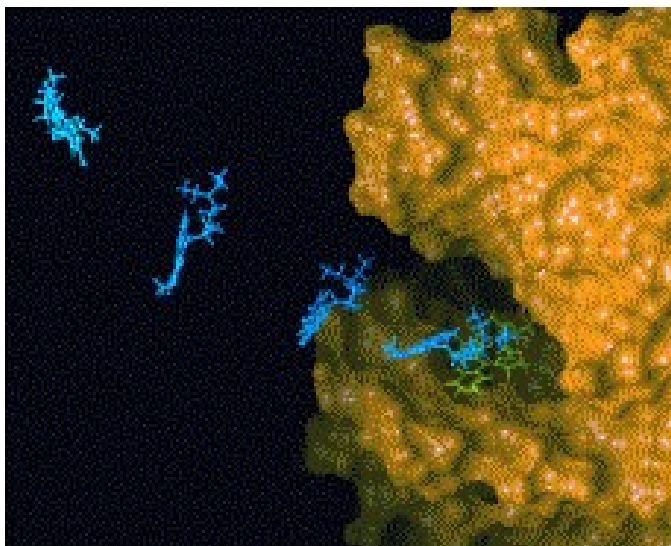
TGCTTTAGCTTT
AAACTACAGGCC
TCACTGGAGCTA
GAGACAAGAAGG
TAAAAACGGCT
GACAAAAGAAGT
CCTGGTATCCTC
TATGATGGGAGA
AGGAACTAGCT
AAAGGGAAGAAT
AAATTAGAGAAA
AACTGGAATGAC
GCTTATACCTGG



Protein/Ligand interactions:

- Understanding How Structures Bind Other Molecules (Function)
- Designing Inhibitors
- Docking, Structure Modeling

(From left to right, figures adapted from Olsen Group Docking Page at Scripps, Dyson NMR Group Web page at Scripps, and from Computational Chemistry Page at Cornell Theory Center).



Michael Gerstein:
<http://bioinfo.mbb.yale.edu/mbb452a/intro/intro.ppt>

Information flow

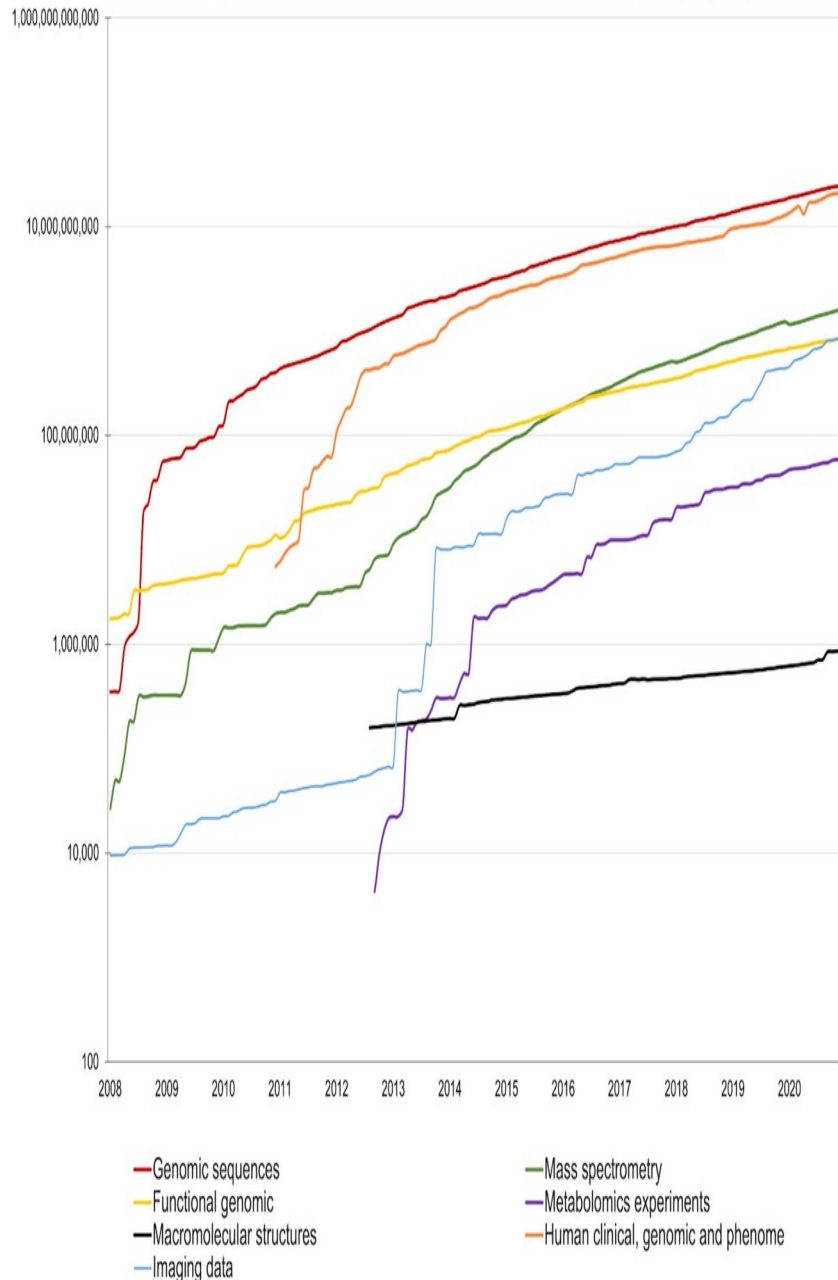
A major task in computational molecular biology is to “decipher” information contained in biological sequences

Since the nucleotide sequence of a genome contains all information necessary to produce a functional organism, we should in theory be able to duplicate this decoding using computers

Data growth in the life sciences

- Computer speed and storage capacity is **doubling every 18 months** and this rate is steady (Moore's law)
- The amount of life science data **doubles every 12 months** and the growth rate is predicted to continue

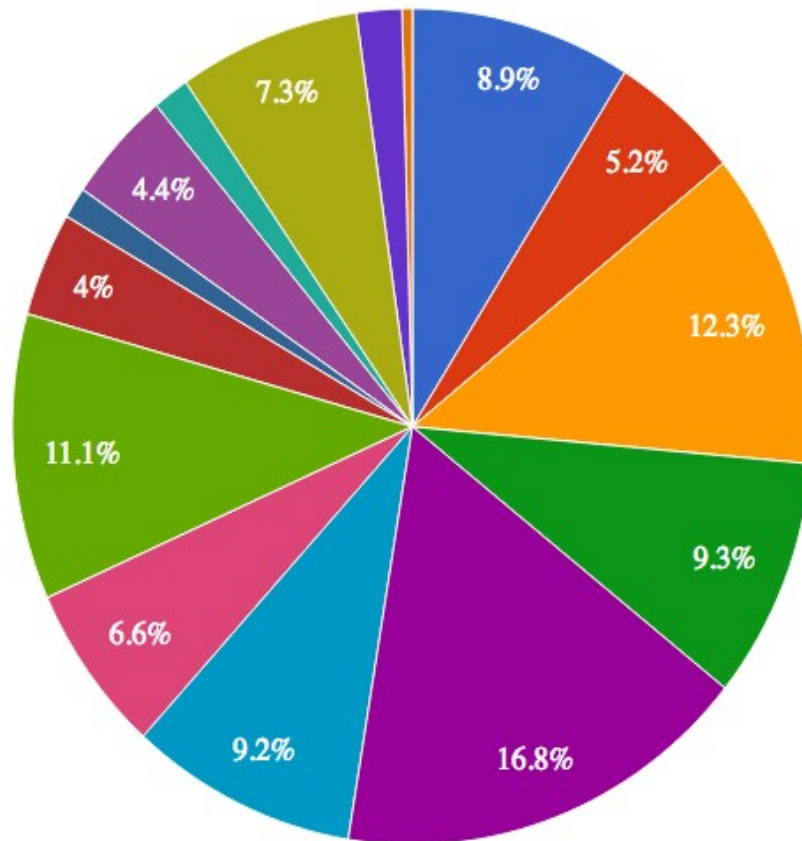
Data growth of EMBL-EBI services volume of data (megabytes)



Cantelli et al. The European Bioinformatics Institute (EMBL-EBI) in 2021, Nucleic Acids Research, Volume 50, Issue D1, 7 January 2022, Pages D11-D19



Data resources in life sciences



- Nucleotide Sequence Databases
- RNA sequence databases
- Protein sequence databases
- Structure Databases
- Genomics Databases (non-vertebrate)
- Metabolic and Signaling Pathways
- Human and other Vertebrate Genomes
- Human Genes and Diseases
- Microarray Data and other Gene Expression Databases
- Proteomics Resources
- Other Molecular Biology Databases
- Organelle databases
- Plant databases
- Immunological databases
- Cell biology

~ 1800
molecular
biology
data
resources

The *Nucleic Acids Research* online Database Collection:
<http://www.oxfordjournals.org/nar/database/a/>



Incoming data size classes:

Organism	Number of chromosomes	Genome size in base pairs
<u>Bacteria</u>	1	~400,000 - ~10,000,000
<u>Yeast</u>	12	14,000,000
<u>Worm</u>	6	100,000,000
<u>Fly</u>	4	300,000,000
<u>Weed</u>	5	125,000,000
<u>Human</u>	23	3,000,000,000

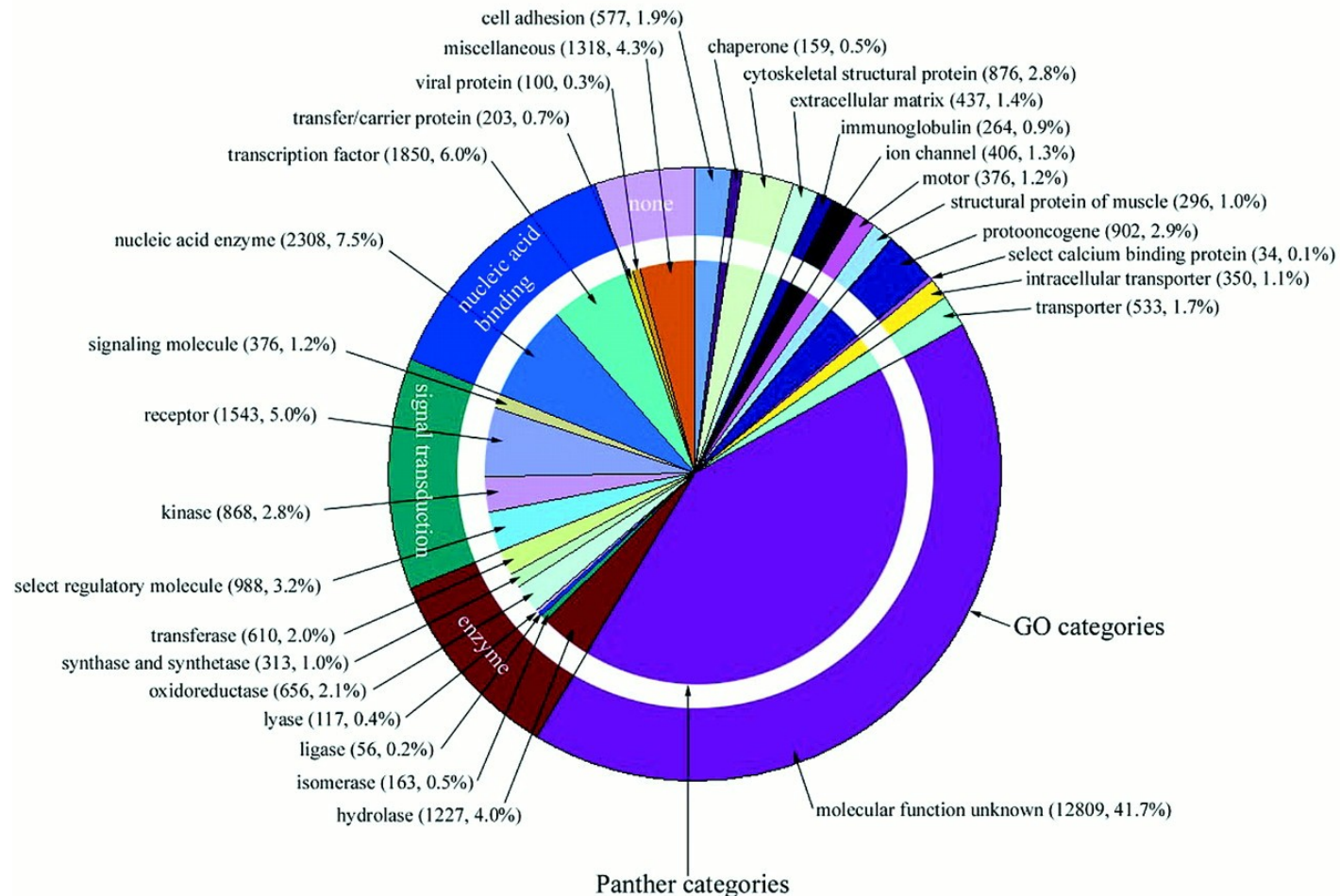
Only the surface is scratched:

Organism	The number of predicted genes	Part of the genome that encodes proteins (exons)
E.Coli (bacteria)	5000	90%
Yeast	6000	70%
<u>Worm</u>	18,000	27%
<u>Fly</u>	14,000	20%
<u>Weed</u>	25,500	20%
<u>Human</u>	30,000	< 5%

A. Brazma et. al.:
http://www.ebi.ac.uk/microarray/biology_intro.html

‘Alien finds a broken hard-disk’ situation

The function of human genes

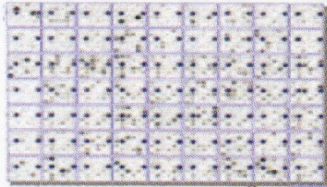


42 % of the genes has unknown function,
 even having accurate predicted protein structures (AlphaFold2)

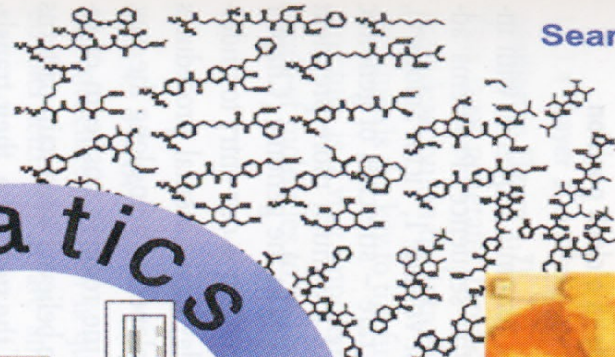
From Genomics to Drugs

Thomas Lengauer (Ed)

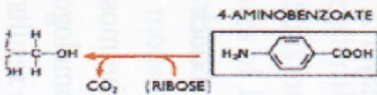
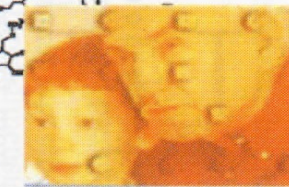
DNA chips: comparison of cell states



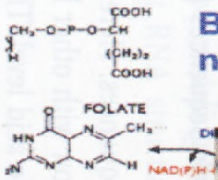
Search for new drugs



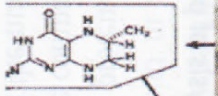
Genetic variations



Biochemical networks



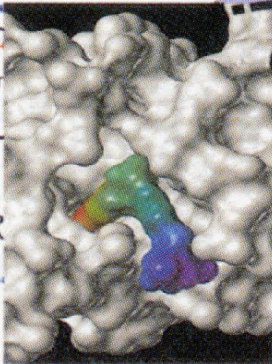
β -TETRAHYDROFOLATE (T)



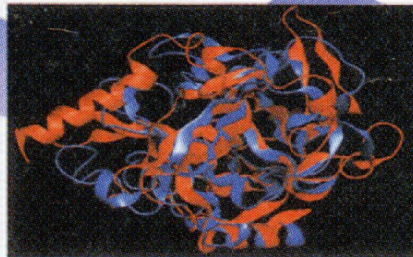
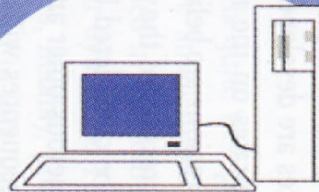
L-VALINE (logically)



Molecular Interactions



Data handling, Algorithms
Statistics, Visualisation



Structure prediction



Optimizing therapies

Genomes

```
cctgtggagccacacctagggtggcca  
atctactcccaggagcaggaggaggcaggag...
```

Proteins

```
MTNRFNFRQIINLLDLRWQRVVPVIHOTETA  
ECGLACLAMICGHFPGKNIDLILYLRKFNLS...
```

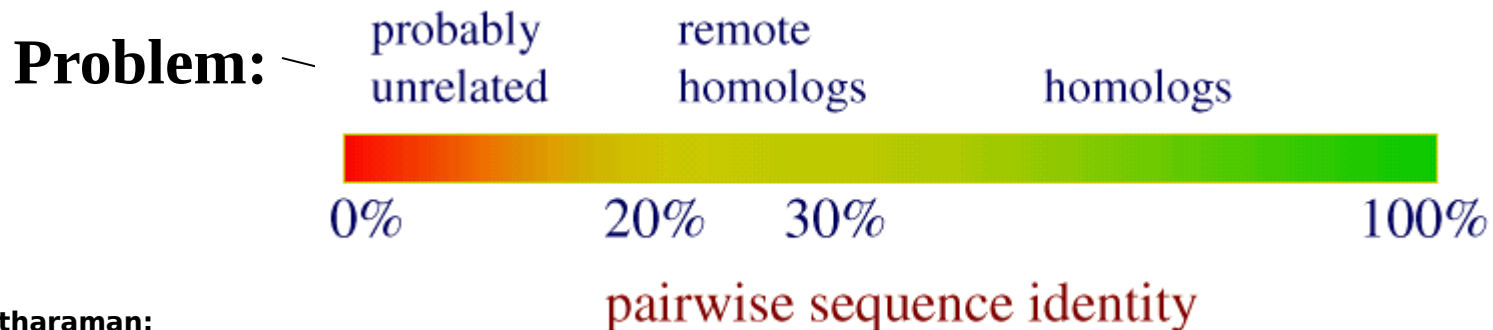
Sequence analysis

Fig. 1.7

A schematic overview of bioinformatics

Homology Modeling

- observation: proteins with similar sequences tend to fold into similar structures
- given: a query sequence Q, database of protein structures
- do:
 - find protein P such that
 - structure of P is known
 - P has high sequence similarity to Q
 - return P's structure as an approximation to Q's structure



Basic biological sequence analysis:

Exact string matching:

- Boyer – Moore string search algorithm (UNIX: grep)
- suffix trees

Inexact string matching:

- Complete sequence (global) or parts (local)
- Similarity measures

Pairwise vs. multiple comparisons

Aligning Text Strings

Raw Data ???

```
T C A T G
  C A T T G
```

2 matches, 0 gaps

```
T C A T G
      | |
C A T T G
```

3 matches (2 end gaps)

```
T C A T G .
  | | |
. C A T T G
```

4 matches, 1 insertion

```
T C A - T G
      | |   | |
. C A T T G
```

4 matches, 1 insertion

```
T C A T - G
      | | |   |
. C A T T G
```

Ambiguity:

```
T C A T G
 / / | |
C A T T G
```

```
T C A T G
 / / / |
C A T T G
```

Definitions

Global alignment

INPUT: Two sequences S and T of roughly the same length.

QUESTION: What is the maximum similarity between them? Find a best alignment.

Local alignment

INPUT: Two sequences S and T .

QUESTION: What is the maximum similarity between a subsequence of S and a subsequence of T ? Find most similar subsequences.

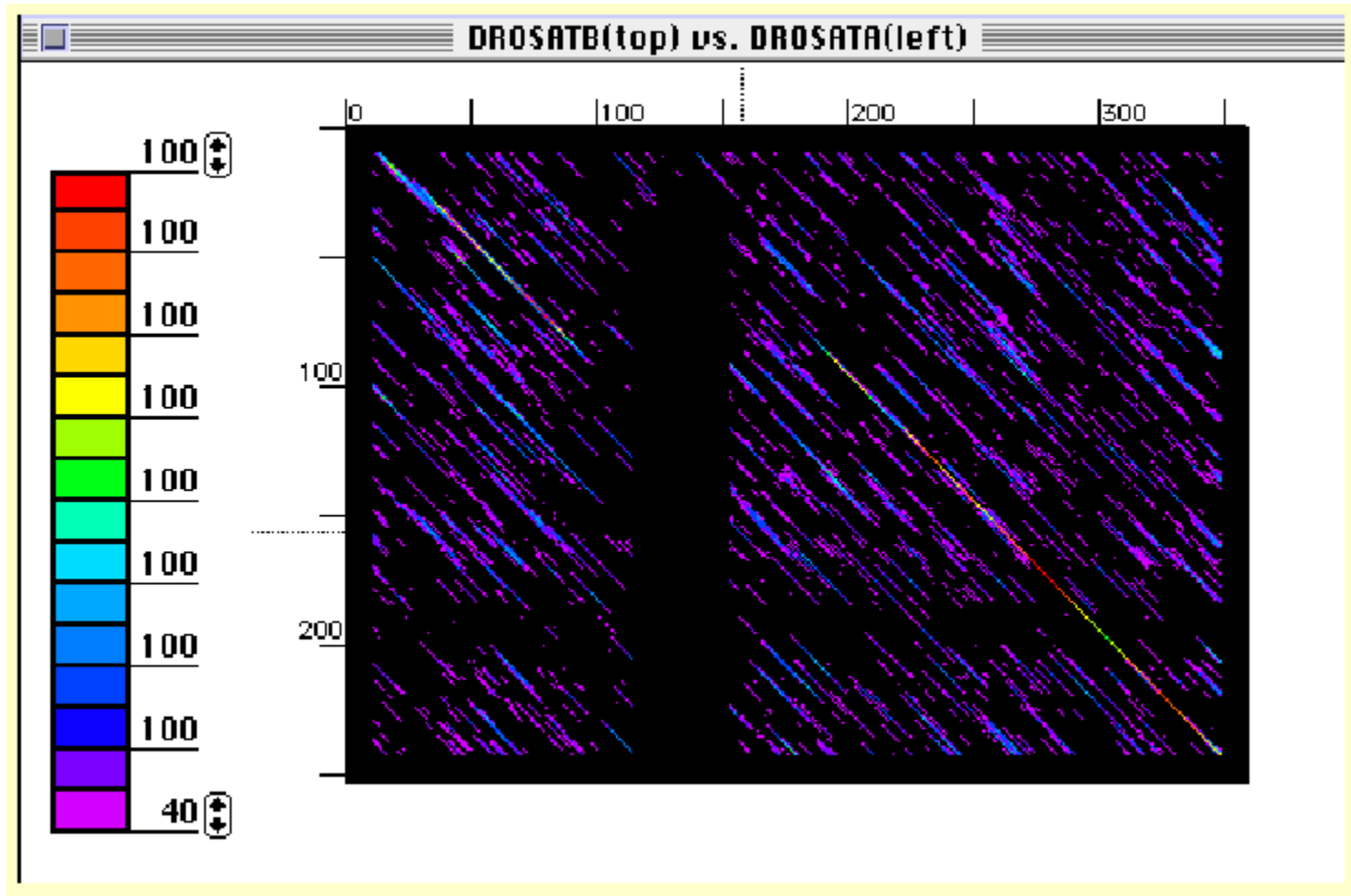
Definition A *gap* is the *maximal* contiguous run of spaces in a single sequence within a given alignment. *The length of a gap* is the number of *indel* operations on it. A *gap penalty function* is a function that measures the cost of a gap as a (nonlinear) function of its length.

Gapped alignment

INPUT: Two sequences S and T (possibly of different length).

QUESTION: Find a best alignment between the two sequences using the gap penalty function.

Graphical solution: dot-plot

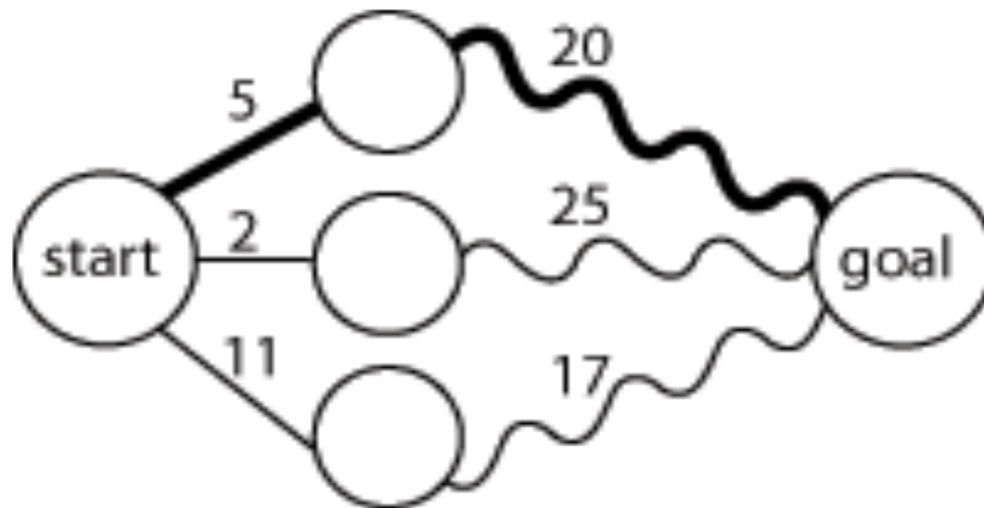


Dynamic programming algorithms for sequence comparison

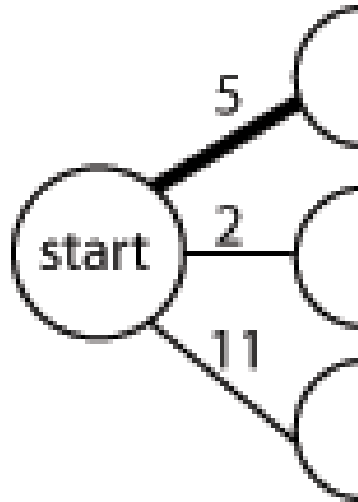
Introduced for biological sequences by

- } S. B. Needleman & C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48:443-453 (1970)

Dynamic programming reminder: Shortest path



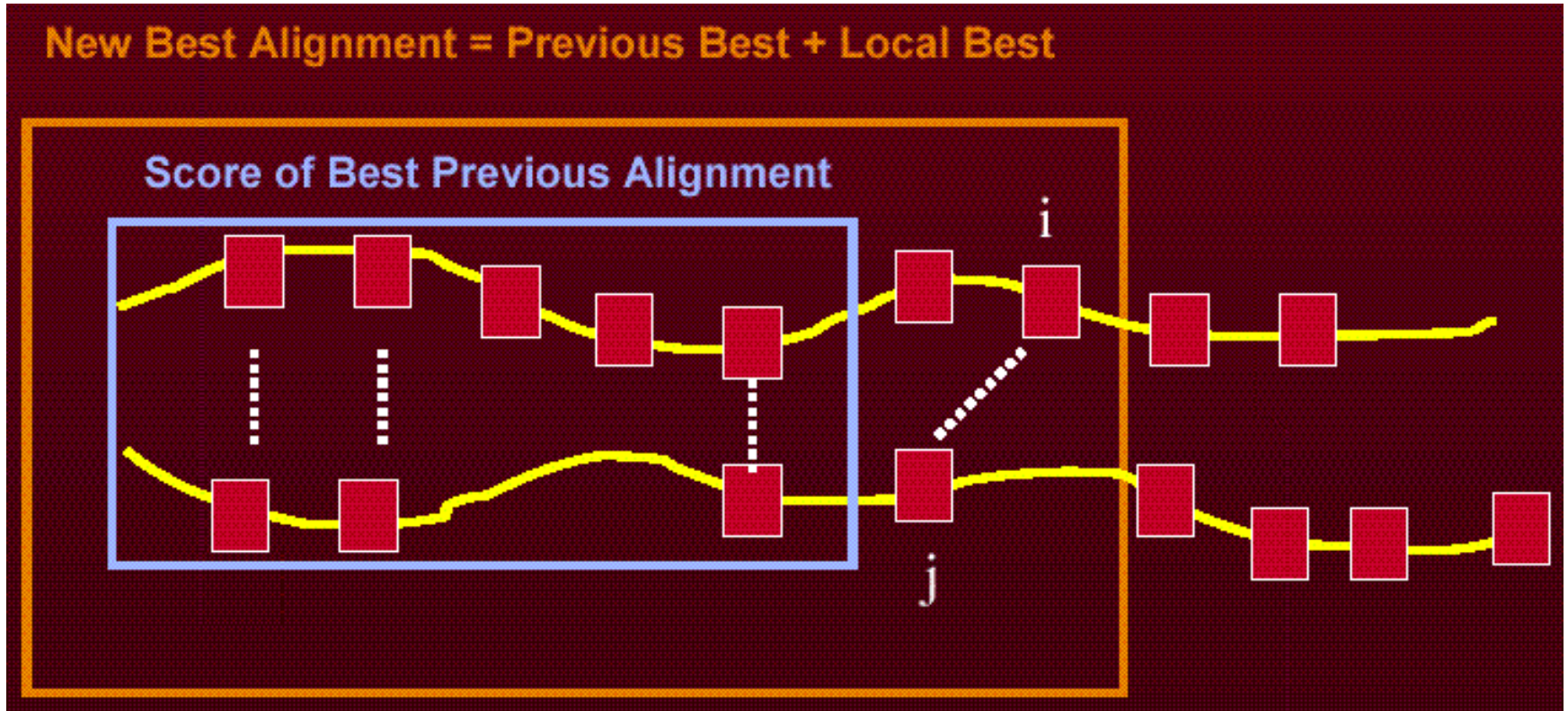
Dynamic programming reminder: Shortest path



Best solutions up to n

**One node added:
n updates to find new best**

Dynamic Programming Idea:



© Copyright Russ Altman
2001, <http://smi-web.stanford.edu/projects/helix/bmi214/4-4-02clr.pdf>

Key Idea in Dynamic Programming

- ◇ The best alignment that ends at a given pair of positions (i and j) in the 2 sequences is the score of the best alignment previous to this position PLUS the score for aligning those two positions.
- ◇ An Example Below
 - o Aligning R to K does not affect alignment of previous N-terminal residues. Once this is done it is **fixed**. Then go on to align D to E.
 - o How could this be violated?
Aligning R to K changes best alignment in box.

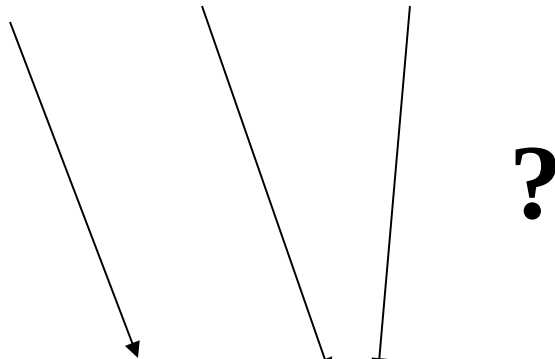
ACSQRP - - LRV - SH	R SENCV
A - SNKPQLVKLMTH	V K DFCV

ACSQRP - - LRV - SH	-R	S ENCV
A - SNKPQLVKLMTH	VK	D FCV

Optimal alignment between sequences

Problem:

VADALTKPVNFKFAVAH



HGQKVADALTKAVAH

similarity score contains:

- variable score for match
- variable cost for gaps
- variable cost for mismatches

Protein amino acid similarity score: Dayhoff's Acceptable Point Mutations (PAMs)

Ala	A																				
Arg	R	30																			
Asn	N	109	17																		
Asp	D	154	0	532																	
Cys	C	33	10	0	0																
Gln	Q	93	120	50	76	0															
Glu	E	266	0	94	831	0	422														
Gly	G	579	10	156	162	10	30	112													
His	H	21	103	226	43	10	243	23	10												
Ile	I	66	30	36	13	17	8	35	0	3											
Leu	L	95	17	37	0	0	75	15	17	40	253										
Lys	K	57	477	322	85	0	147	104	60	23	43	39									
Met	M	29	17	0	0	0	20	7	7	0	57	207	90								
Phe	F	20	7	7	0	0	0	0	17	20	90	167	0	17							
Pro	P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
Ser	S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
Thr	T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
Trp	W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Tyr	Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
Val	V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

Steps of basic dynamic programming method

1. Initialize matrix to match scores
(for simplicity: 0 or 1)
2. Do summation operation
 - } Finds the maximum number of matches that can be obtained starting at any position and proceeding "forward"
3. Traceback to find maximum match alignment

	V	A	D	A	L	T	K	P	V	N	F	K	F	A	V	A	H
H																	1
G																	
Q																	
K							1					1					
V	1								1						1		
A		1		1										1		1	
D			1														
A		1		1										1		1	
L					1												
T						1											
K							1					1					
A		1		1										1		1	
V	1								1						1		
A		1		1										1	1		
H																	1

Status: Showing current search locations

	V	A	D	A	L	T	K	P	V	N	F	K	F	A	V	A	H
H																	1
G																	
Q																	
K							1					1					
V	1								1						1		
A		1		1										1		1	
D			1														
A		1		1										1		1	
L					1												
T						1											
K							1					1					
A		1		1										1		1	
V	1								1						1		
A		1		1										1	1		↘
H																	1

Status: Showing maximum found in search locations

	V	A	D	A	L	T	K	P	V	N	F	K	F	A	V	A	H	
H																	1	
G																		
Q																		
K							1					1						
V	1								1						1			
A		1		1										1		1		
D			1															
A		1		1										1		1		
L					1													
T						1												
K							1					1						
A		1		1										1		1		
V	1								1						1			
A		1		1										1	2			
H																	1	

Status: Showing updated matrix at current location

	V	A	D	A	L	T	K	P	V	N	F	K	F	A	V	A	H
H								6	5	5	5	4	4	3	2	1	1
G								6	5	5	5	4	4	3	2	1	
Q								6	5	5	5	4	4	3	2	1	
K							1	6	5	5	5	5	4	3	2	1	
V	1							5	6	5	5	4	4	3	3	1	
A		1		1				5	5	5	5	4	4	4	2	2	
D			1					5	5	5	5	4	4	3	2	1	
A		1		1				5	5	5	5	4	4	4	2	2	
L					1			5	5	5	5	4	4	3	2	1	
T						1		5	5	5	5	4	4	3	2	1	
K							1	4	4	4	4	5	4	3	2	1	
A		1		1			3	3	3	3	3	3	3	4	2	2	
V	1						2	2	3	2	2	2	2	2	3	1	
A		1		1			1	1	1	1	1	1	1	2	1	2	
H																	1

Status: Showing current search locations

	V	A	D	A	L	T	K	P	V	N	F	K	F	A	V	A	H
H								6	5	5	5	4	4	3	2	1	1
G								6	5	5	5	4	4	3	2	1	
Q								6	5	5	5	4	4	3	2	1	
K							1	6	5	5	5	5	4	3	2	1	
V	1							5	6	5	5	4	4	3	3	1	
A		1		1				5	5	5	5	4	4	4	2	2	
D			1					5	5	5	5	4	4	3	2	1	
A		1		1				5	5	5	5	4	4	4	2	2	
L					1			5	5	5	5	4	4	3	2	1	
T						1		5	5	5	5	4	4	3	2	1	
K							1	4	4	4	4	5	4	3	2	1	
A		1		1			3	3	3	3	3	3	3	4	2	2	
V	1						2	2	3	2	2	2	2	2	3	1	
A		1		1			1	1	1	1	1	1	1	2	1	2	
H																	1

Status: Showing maximum found in search locations

	V	A	D	A	L	T	K	P	V	N	F	K	F	A	V	A	H
H								6	5	5	5	4	4	3	2	1	1
G								6	5	5	5	4	4	3	2	1	
Q								6	5	5	5	4	4	3	2	1	
K							1	6	5	5	5	5	4	3	2	1	
V	1							5	6	5	5	4	4	3	3	1	
A		1		1				5	5	5	5	4	4	4	2	2	
D			1					5	5	5	5	4	4	3	2	1	
A		1		1				5	5	5	5	4	4	4	2	2	
L					1			5	5	5	5	4	4	3	2	1	
T						1		5	5	5	5	4	4	3	2	1	
K							5	4	4	4	4	5	4	3	2	1	
A		1		1			3	3	3	3	3	3	3	4	2	2	
V	1						2	2	3	2	2	2	2	2	3	1	
A		1		1			1	1	1	1	1	1	1	2	1	2	
H																	1

Status: Showing updated matrix at current location

	V	A	D	A	L	T	K	P	V	N	F	K	F	A	V	A	H
H	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	1
G	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
Q	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
K	10	9	8	7	6	6	7	6	5	5	5	5	4	3	2	1	
V	11	9	8	7	6	5	5	5	6	5	5	4	4	3	3	1	
A	9	10	8	8	6	5	5	5	5	5	5	4	4	4	2	2	
D	8	8	9	7	6	5	5	5	5	5	5	4	4	3	2	1	
A	7	8	7	8	6	5	5	5	5	5	5	4	4	4	2	2	
L	6	6	6	6	7	5	5	5	5	5	5	4	4	3	2	1	
T	5	5	5	5	5	6	5	5	5	5	5	4	4	3	2	1	
K	4	4	4	4	4	4	5	4	4	4	4	5	4	3	2	1	
A	3	4	3	4	3	3	3	3	3	3	3	3	3	4	2	2	
V	3	2	2	2	2	2	2	2	3	2	2	2	2	2	3	1	
A	1	2	1	2	1	1	1	1	1	1	1	1	1	2	1	2	
H																	1

Status: Showing current traceback search locations

	V	A	D	A	L	T	K	P	V	N	F	K	F	A	V	A	H
H	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	1
G	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
Q	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
K	10	9	8	7	6	6	7	6	5	5	5	5	4	3	2	1	
V	11	9	8	7	6	5	5	5	6	5	5	4	4	3	3	1	
A	9	10	8	8	6	5	5	5	5	5	5	4	4	4	2	2	
D	8	8	9	7	6	5	5	5	5	5	5	4	4	3	2	1	
A	7	8	7	8	6	5	5	5	5	5	5	4	4	4	2	2	
L	6	6	6	6	7	5	5	5	5	5	5	4	4	3	2	1	
T	5	5	5	5	5	6	5	5	5	5	5	4	4	3	2	1	
K	4	4	4	4	4	4	5	4	4	4	4	5	4	3	2	1	
A	3	4	3	4	3	3	3	3	3	3	3	3	3	4	2	2	
V	3	2	2	2	2	2	2	2	3	2	2	2	2	2	3	1	
A	1	2	1	2	1	1	1	1	1	1	1	1	1	2	1	2	
H																	1

Status: Showing maximum found in traceback

	V	A	D	A	L	T	K	P	V	N	F	K	F	A	V	A	H
H	18	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	1
G	18	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
Q	18	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
K	18	9	8	7	6	6	7	6	5	5	5	5	4	3	2	1	
V	11	9	8	7	6	5	5	5	6	5	5	4	4	3	3	1	
A	9	10	8	8	6	5	5	5	5	5	5	4	4	4	2	2	
D	8	8	9	7	6	5	5	5	5	5	5	4	4	3	2	1	
A	7	8	7	8	6	5	5	5	5	5	5	4	4	4	2	2	
L	6	6	6	6	7	5	5	5	5	5	5	4	4	3	2	1	
T	5	5	5	5	5	6	5	5	5	5	5	4	4	3	2	1	
K	4	4	4	4	4	4	5	4	4	4	4	5	4	3	2	1	
A	3	4	3	4	3	3	3	3	3	3	3	3	3	4	2	2	
V	3	2	2	2	2	2	2	2	3	2	2	2	2	2	3	1	
A	1	2	1	2	1	1	1	1	1	1	1	1	1	2	1	2	
H																	1

Status: Showing current traceback search locations

----V

HGQKV

	V	A	D	A	L	T	K	P	V	N	F	K	F	A	V	A	H
H	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	1
G	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
Q	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
K	10	9	8	7	6	6	7	6	5	5	5	5	4	3	2	1	
V	11	9	8	7	6	5	5	5	6	5	5	4	4	3	3	1	
A	9	10	8	8	6	5	5	5	5	5	5	4	4	4	2	2	
D	8	8	9	7	6	5	5	5	5	5	5	4	4	3	2	1	
A	7	8	7	8	6	5	5	5	5	5	5	4	4	4	2	2	
L	6	6	6	6	7	5	5	5	5	5	5	4	4	3	2	1	
T	5	5	5	5	5	6	5	5	5	5	5	4	4	3	2	1	
K	4	4	4	4	4	4	5	4	4	4	4	5	4	3	2	1	
A	3	4	3	4	3	3	3	3	3	3	3	3	3	4	2	2	
V	3	2	2	2	2	2	2	2	3	2	2	2	2	2	3	1	
A	1	2	1	2	1	1	1	1	1	1	1	1	1	2	1	2	
H																	1

Status: Showing maximum found in traceback

	V	A	D	A	L	T	K	P	V	N	F	K	F	A	V	A	H
H	18	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	1
G	18	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
Q	18	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
K	18	9	8	7	6	6	7	6	5	5	5	5	4	3	2	1	
V	11	9	8	7	6	5	5	5	6	5	5	4	4	3	3	1	
A	9	10	8	8	6	5	5	5	5	5	5	4	4	4	2	2	
D	8	8	9	7	6	5	5	5	5	5	5	4	4	3	2	1	
A	7	8	7	8	6	5	5	5	5	5	5	4	4	4	2	2	
L	6	6	6	6	7	5	5	5	5	5	5	4	4	3	2	1	
T	5	5	5	5	5	6	5	5	5	5	5	4	4	3	2	1	
K	4	4	4	4	4	4	5	4	4	4	4	5	4	3	2	1	
A	3	4	3	4	3	3	3	3	3	3	3	3	3	4	2	2	
V	3	2	2	2	2	2	2	2	3	2	2	2	2	2	3	1	
A	1	2	1	2	1	1	1	1	1	1	1	1	1	2	1	2	
H																	1

Status: Showing current traceback search locations

-----VA

HGQKVA

	V	A	D	A	L	T	K	P	V	N	F	K	F	A	V	A	H
H	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	1
G	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
Q	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
K	10	9	8	7	6	6	7	6	5	5	5	5	4	3	2	1	
V	11	9	8	7	6	5	5	5	6	5	5	4	4	3	3	1	
A	9	10	8	8	6	5	5	5	5	5	5	4	4	4	2	2	
D	8	8	9	7	6	5	5	5	5	5	5	4	4	3	2	1	
A	7	8	7	8	6	5	5	5	5	5	5	4	4	4	2	2	
L	6	6	6	6	7	5	5	5	5	5	5	4	4	3	2	1	
T	5	5	5	5	5	6	5	5	5	5	5	4	4	3	2	1	
K	4	4	4	4	4	4	5	4	4	4	4	5	4	3	2	1	
A	3	4	3	4	3	3	3	3	3	3	3	3	3	4	2	2	
V	3	2	2	2	2	2	2	2	3	2	2	2	2	2	3	1	
A	1	2	1	2	1	1	1	1	1	1	1	1	1	2	1	2	
H																	1

Status: Showing current traceback search locations

---VADALTK

HGQKVADALTK

	V	A	D	A	L	T	K	P	V	N	F	K	F	A	V	A	H
H	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	1
G	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
Q	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
K	10	9	8	7	6	6	7	6	5	5	5	5	4	3	2	1	
V	11	9	8	7	6	5	5	5	6	5	5	4	4	3	3	1	
A	9	10	8	8	6	5	5	5	5	5	5	4	4	4	2	2	
D	8	8	9	7	6	5	5	5	5	5	5	4	4	3	2	1	
A	7	8	7	8	6	5	5	5	5	5	5	4	4	4	2	2	
L	6	6	6	6	7	5	5	5	5	5	5	4	4	3	2	1	
T	5	5	5	5	5	6	5	5	5	5	5	4	4	3	2	1	
K	4	4	4	4	4	4	5	4	4	4	4	5	4	3	2	1	
A	3	4	3	4	3	3	3	3	3	3	3	3	3	4	2	2	
V	3	2	2	2	2	2	2	2	2	3	2	2	2	2	3	1	
A	1	2	1	2	1	1	1	1	1	1	1	1	1	2	1	2	
H																	1

Status: Showing maximum found in traceback

---VADALTK

HGQKVADALTK

	V	A	D	A	L	T	K	P	V	N	F	K	F	A	V	A	H
H	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	1
G	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
Q	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
K	10	9	8	7	6	6	7	6	5	5	5	5	4	3	2	1	
V	11	9	8	7	6	5	5	5	6	5	5	4	4	3	3	1	
A	9	10	8	8	6	5	5	5	5	5	5	4	4	4	2	2	
D	8	8	9	7	6	5	5	5	5	5	5	4	4	3	2	1	
A	7	8	7	8	6	5	5	5	5	5	5	4	4	4	2	2	
L	6	6	6	6	7	5	5	5	5	5	5	4	4	3	2	1	
T	5	5	5	5	5	6	5	5	5	5	5	4	4	3	2	1	
K	4	4	4	4	4	4	5	4	4	4	4	5	4	3	2	1	
A	3	4	3	4	3	3	3	3	3	3	3	3	3	4	2	2	
V	3	2	2	2	2	2	2	2	3	2	2	2	2	2	3	1	
A	1	2	1	2	1	1	1	1	1	1	1	1	1	2	1	2	
H																	1

Status: Showing current traceback search locations

---VADALTKPVNFKFA

HGQKVADALTK-----A

	V	A	D	A	L	T	K	P	V	N	F	K	F	A	V	A	H
H	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	1
G	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
Q	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
K	10	9	8	7	6	6	7	6	5	5	5	5	4	3	2	1	
V	11	9	8	7	6	5	5	5	6	5	5	4	4	3	3	1	
A	9	10	8	8	6	5	5	5	5	5	5	4	4	4	2	2	
D	8	8	9	7	6	5	5	5	5	5	5	4	4	3	2	1	
A	7	8	7	8	6	5	5	5	5	5	5	4	4	4	2	2	
L	6	6	6	6	7	5	5	5	5	5	5	4	4	3	2	1	
T	5	5	5	5	5	6	5	5	5	5	5	4	4	3	2	1	
K	4	4	4	4	4	4	5	4	4	4	4	5	4	3	2	1	
A	3	4	3	4	3	3	3	3	3	3	3	3	3	4	2	2	
V	3	2	2	2	2	2	2	2	3	2	2	2	2	2	3	1	
A	1	2	1	2	1	1	1	1	1	1	1	1	1	2	1	2	
H																	1

Status: Showing final alignment

---VADALTKPVNFKFAVAH

HGQKVADALTK-----AVAH

Summation operation

1. Start in lower right corner
2. Move up one position and left one position
3. Find largest value in either (a) row segment starting one below current position and extending to the right or (b) column segment starting one to the right of current position and extending down

Summation operation (cont.)

4. Add this value to the value in the current cell
5. Repeat steps 3 and 4 for all cells to the left in current row and all cells above in current column
6. If we are not in the top left corner, go to step 2

Multiple sequence alignment

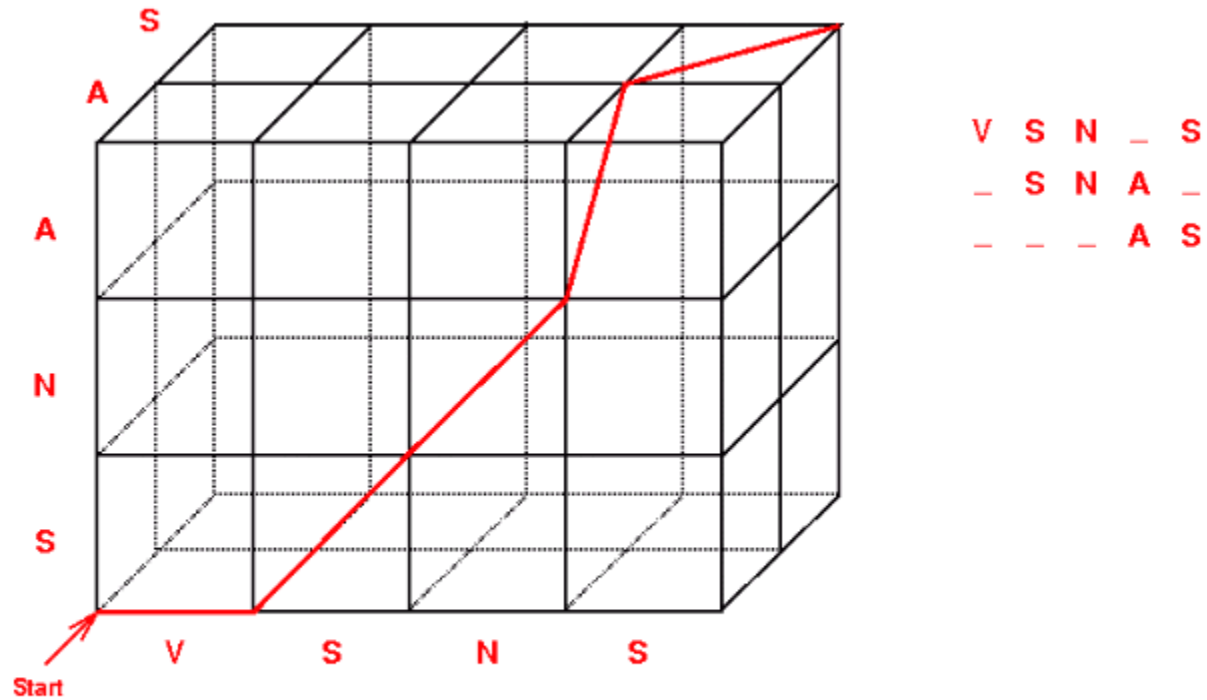
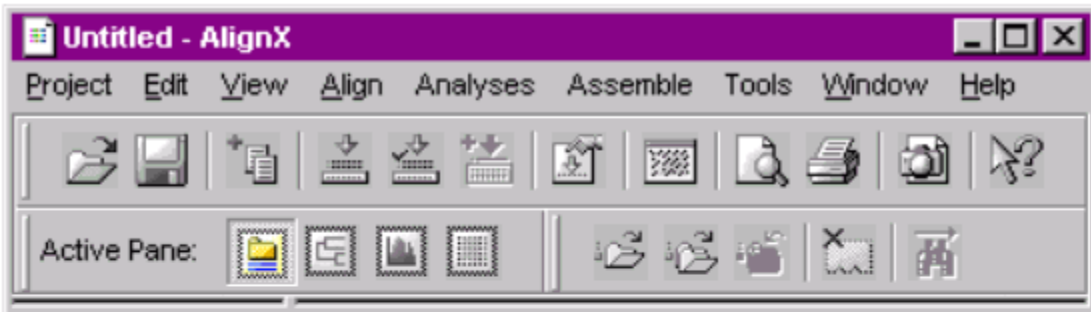


Figure source: <http://www.techfak.uni-bielefeld.de/bcd/Curric/MuAli/node2.html#SECTION00020000000000000000>

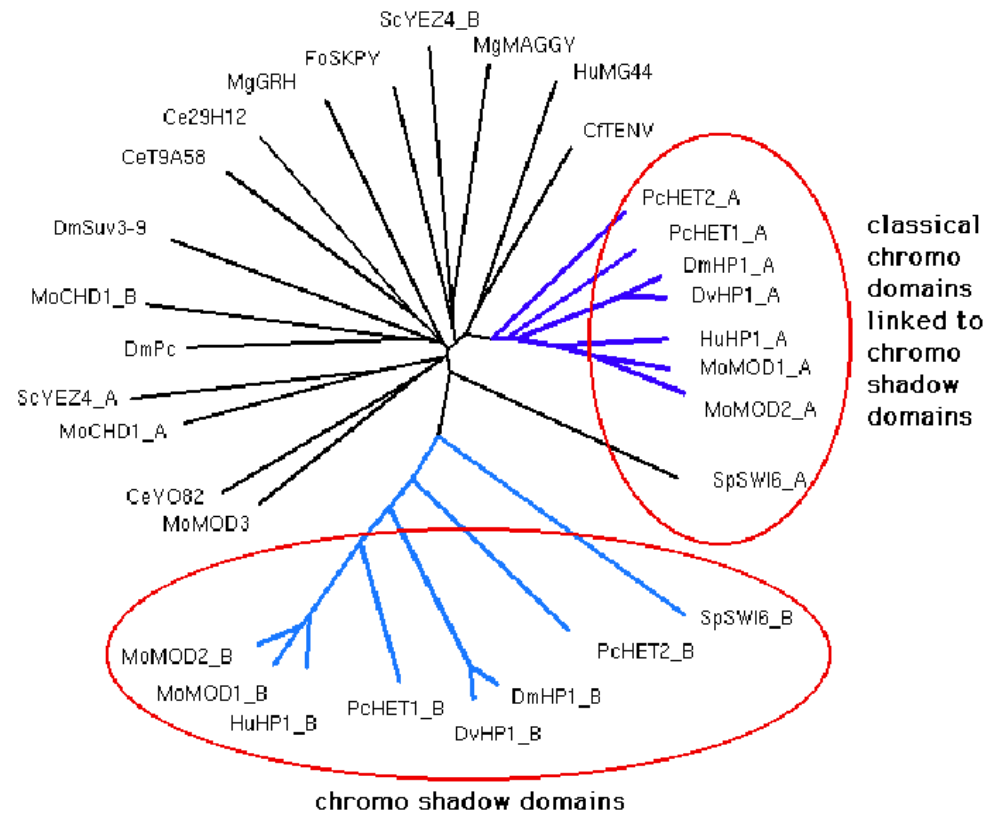
- Calc. of optimal solution infeasible for >5 sequences
- ⇒ Heuristic solutions
- ⇒ e.g. progressive alignment (CLUSTALW)

Multiple sequence alignment for phylogenetic trees



	1	10	20
NONAME	1	VSLTCL-VKGFYPSD-I	AVEWESNG--
NONAME#2	1	VTISCTGTSSNIGS--	ITVNWYQLPG
NONAME#8	1	VTISCTGSSNIGAG-NHVKWYQLPG	
NONAME#3	1	LRLSCS-SSGFIFSS-Y	ANYWVRQAPG
NONAME#4	1	LSLTCT-VSGTSFDD-Y	YSTWVRQPPG
NONAME#5	1	PEVTCVVVDVSHEDPQ	VKFNWYVDG--
NONAME#6	1	ATLVCL-ISDFYPGA-V	TVAWKADS--
NONAME#7	1	AALGCL-VKDYFPEP-V	TVSWNSG---
Consensus	1	VTLSCT VS F S V V W Q PG	

Ready positives: 59.3% identi



Modelling tasks:

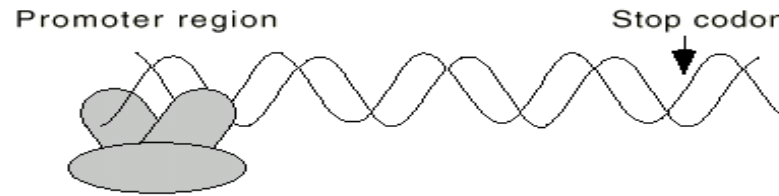
Promoter
Stop

Difficulty

3
2

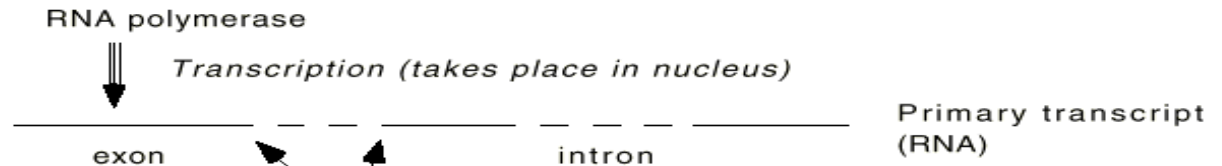
Image by Lawrence Hunter:
<http://www.aaai.org/Library/Books/Hunter/01-Hunter.htm>

sequencing



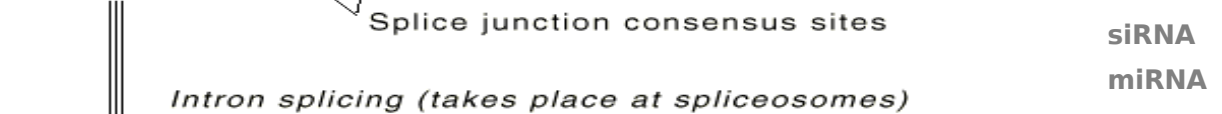
1:1
splice sites
exon/intron

2
1



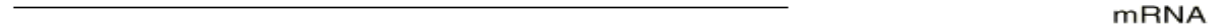
alternative splicing

3



Translation start

1



3:1



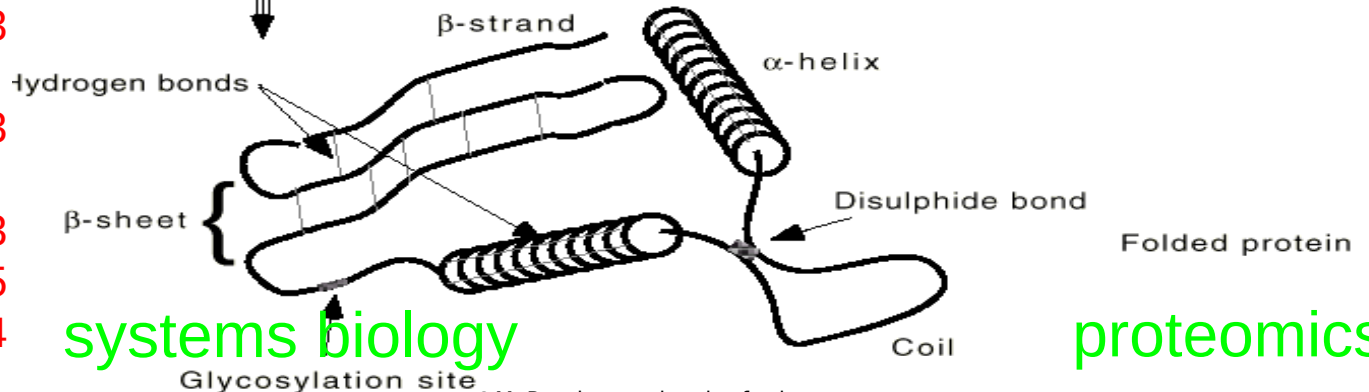
Cleaving

2



Secondary structure

3



S-S bonds

3

Exposure

3

Tertiary structure

5

Complexes, networks

4

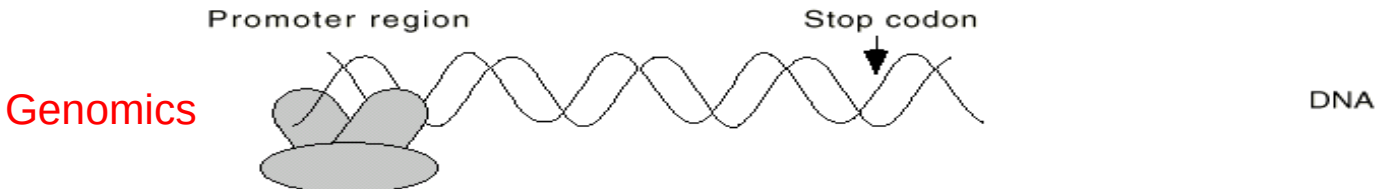
systems biology

proteomics

Modelling tasks:

Image by Lawrence Hunter:
<http://www.aaai.org/Library/Books/Hunter/01-Hunter.htm>

Promoter
 Stop



Genomics

1:1
 splice sites
 exon/intron



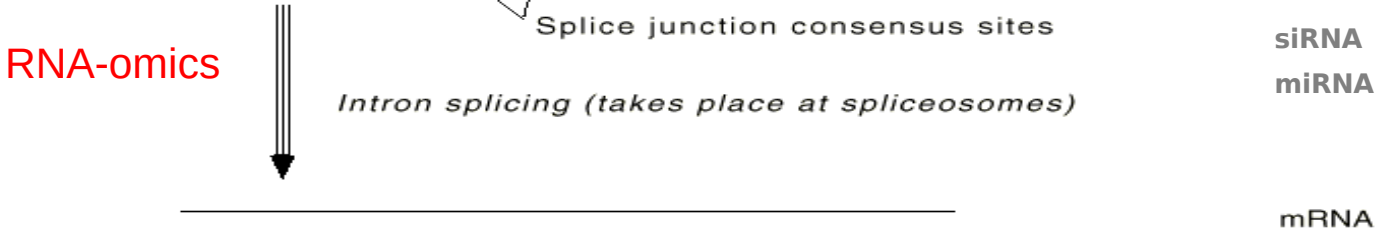
Transcript-omics

alternative splicing

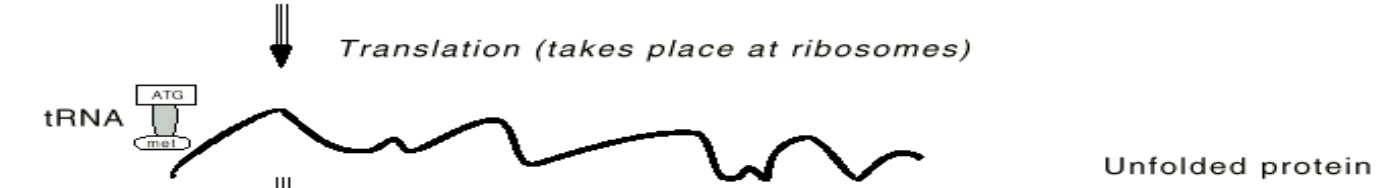
RNA-omics

siRNA
 miRNA

Translation start



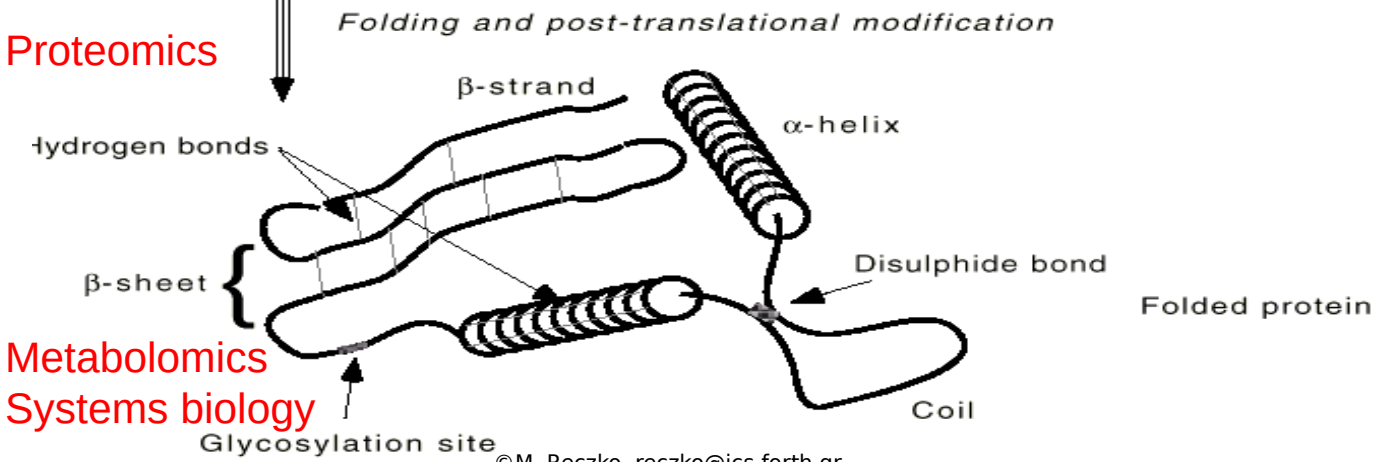
3:1



Cleaving

Proteomics

Secondary structure



S-S bonds

Exposure

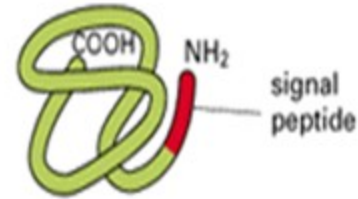
Tertiary structure

Complexes, networks

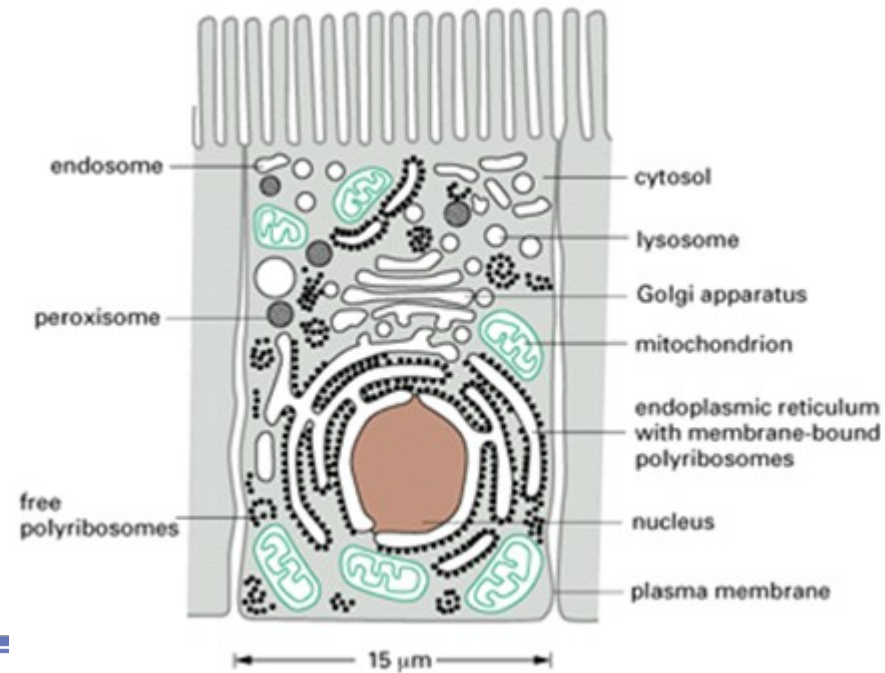
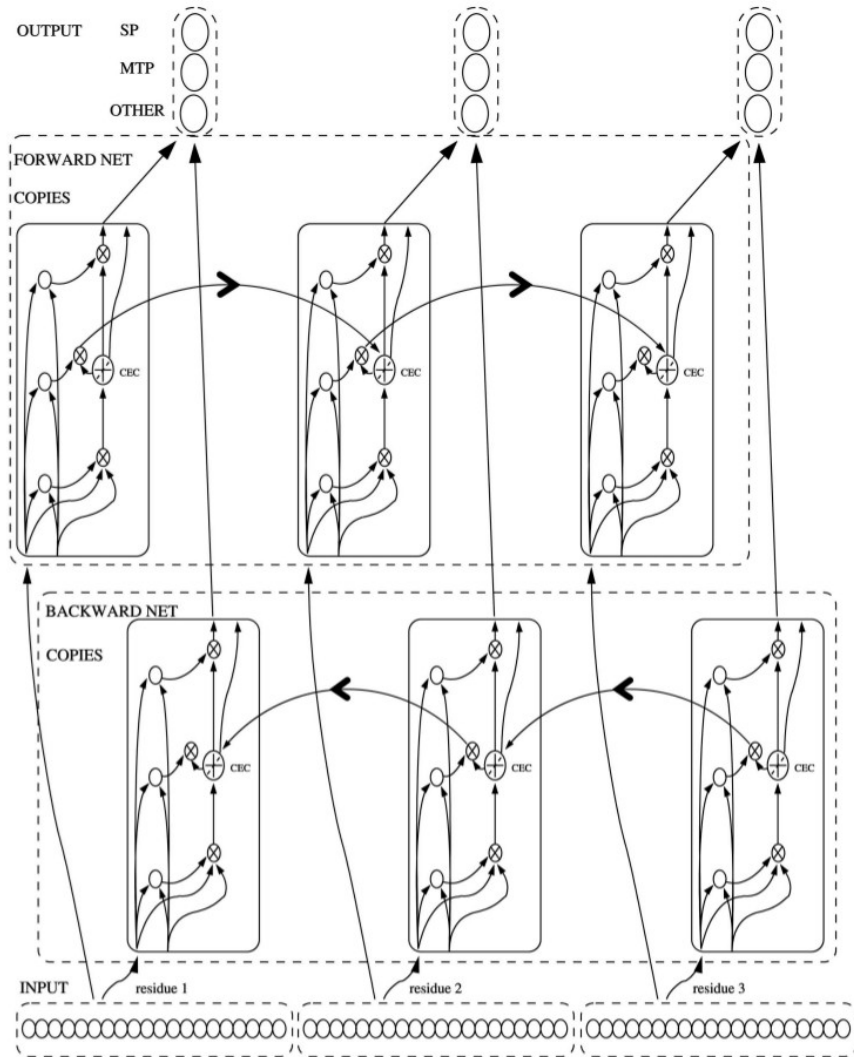
Metabolomics
 Systems biology

Introduction novel sequence learning algorithm (BLSTM)

- Use start of protein sequence to predict its compartment



- BLSTMs precursors of transformer networks





About

Research

Impact

Blog

Safety &
Ethics

Careers



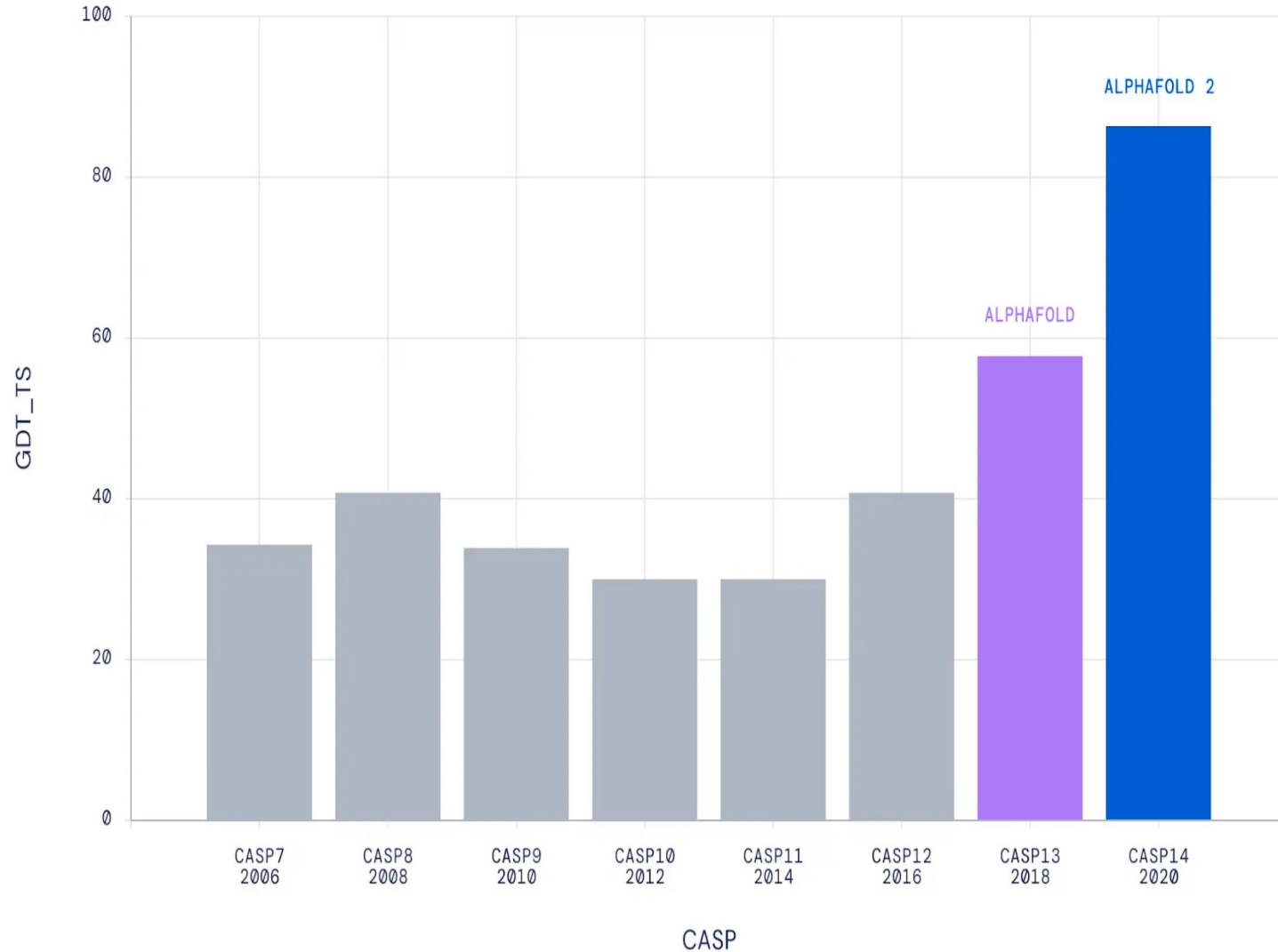
DeepMind

What if solving one problem could unlock solutions to thousands more?

FIND OUT MORE

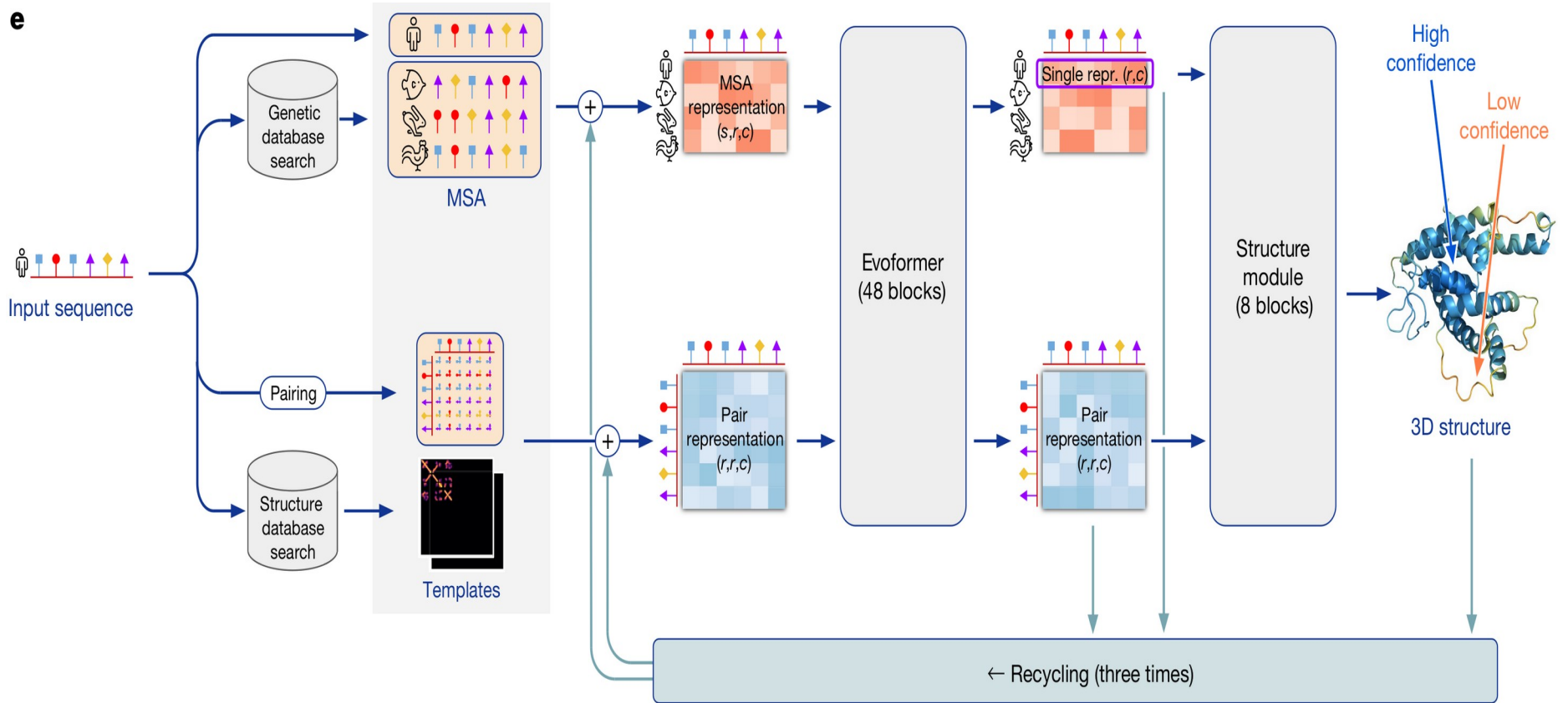


Median Free-Modelling Accuracy



Improvements in the median accuracy of predictions in the free modelling category for the best team in each CASP, measured as best-of-5 GDT.

AlphaFold2 architecture



AlphaFold2 database of predicted structures

alphafold.ebi.ac.uk

it metabolo... bio dev med etc cs phys tools

EMBL-EBI Services Research Training About us EMBL-EBI

AlphaFold Protein Structure Database

Home About FAQs Downloads

Search for protein, gene, UniProt accession or organism BETA Search

Examples: Free fatty acid receptor 2 At1g58602 Q5VSL9 E. coli Help: AlphaFold DB search help

Developed by

EMBL-EBI

Services <ul style="list-style-type: none">By topicBy name (A-Z)Help & Support	Research <ul style="list-style-type: none">PublicationsResearch groupsPostdocs & PhDs	Training <ul style="list-style-type: none">Live trainingOn-demand trainingSupport for trainersContact organisers	Industry <ul style="list-style-type: none">Members AreaWorkshopsSME ForumContact Industry programme	About <ul style="list-style-type: none">Contact usEventsJobsNewsPeople & groups
---	--	--	---	--

EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK. +44 (0)1223 49 44 44

Copyright © EMBL 2021 | EMBL-EBI is part of the European Molecular Biology Laboratory | Terms of use | License and Disclaimer

ELIXIR

ELIXIR is an intergovernmental organisation that brings together life science resources such as **databases**, **software tools**, **training materials**, **standards** and **compute resources**, from across Europe.

The goal of ELIXIR is to **coordinate life science resources from across Europe so they form a single infrastructure**. This makes it easier for scientists to:

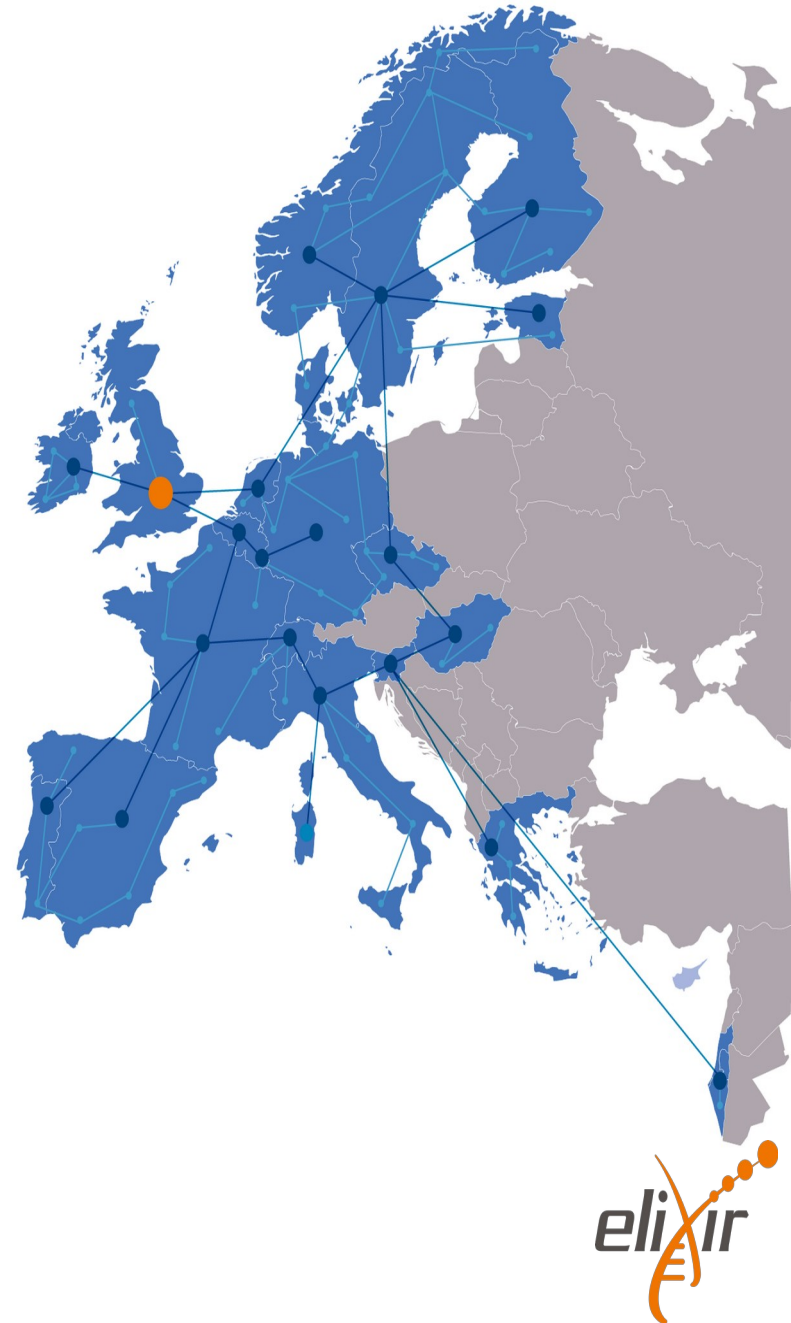
Find and share data

Exchange expertise

Agree on best practices in scientific research

Check: <https://elixir-europe.org>

<https://elixir-greece.org>



Accelerating research through data sharing



Viral sequences →

Raw and assembled sequence and analysis of SARS-CoV-2 and other coronaviruses.

[111,900 records >](#)

Host sequences →

Raw and assembled sequence and analysis of human and other hosts.

[973 records >](#)

About this portal

The COVID-19 Data Portal was launched in April 2020 to bring together relevant datasets for sharing and analysis in an effort to accelerate coronavirus research. It enables researchers to upload, access and analyse COVID-19 related reference data and specialist datasets as part of the wider European COVID-19 Data Platform.

To enquire on how to collaborate on the European COVID-19 platform: ecovid19@ebi.ac.uk.

To share your data on COVID-19 Data Portal: virus-dataflow@ebi.ac.uk.

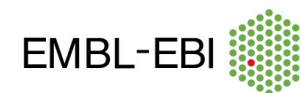
COVID DATA RESOURCES

[Viral sequences](#)
[Host sequences](#)
[Expression](#)
[Proteins](#)

[Biochemistry](#)
[Literature](#)
[Related Resources](#)

ABOUT

[About the Portal](#)
[SARS-CoV-2 Data Hubs](#)
[Our Partners](#)
[Submit Data](#)





DATA

OPTIMISATION

MODEL

EVALUATION

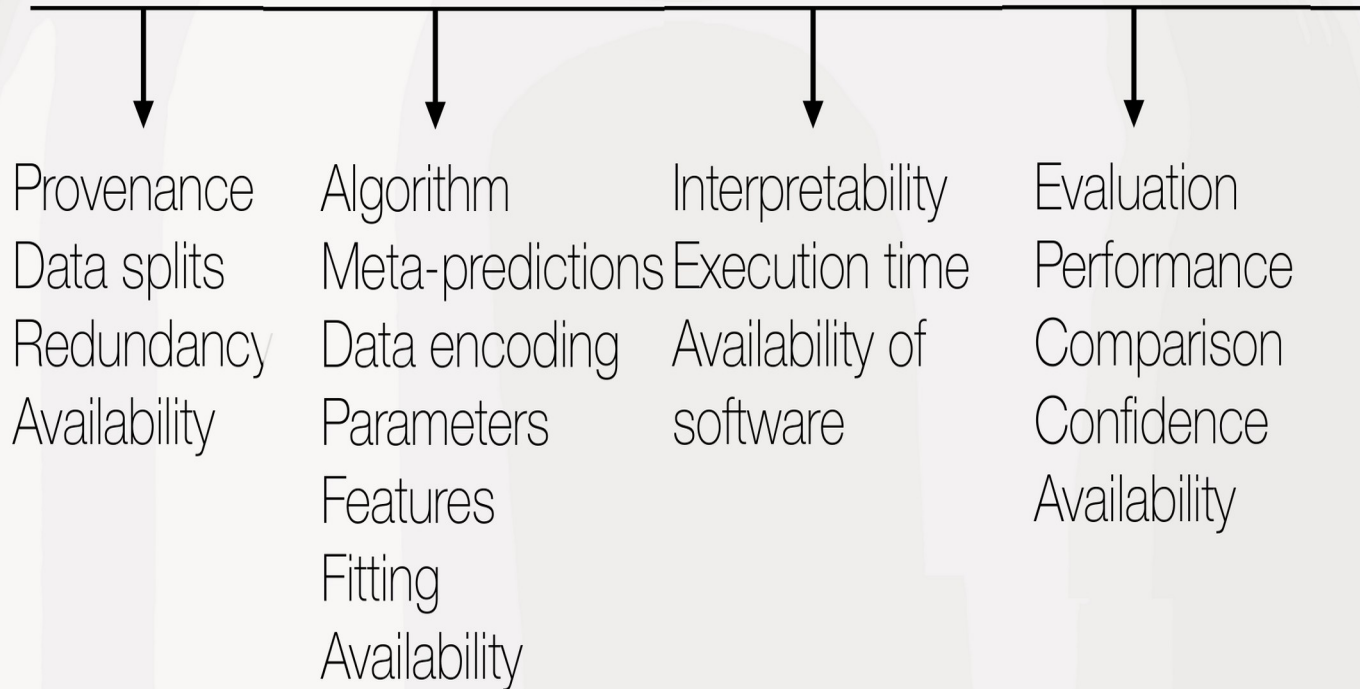
MACHINE LEARNING
FOCUS GROUP



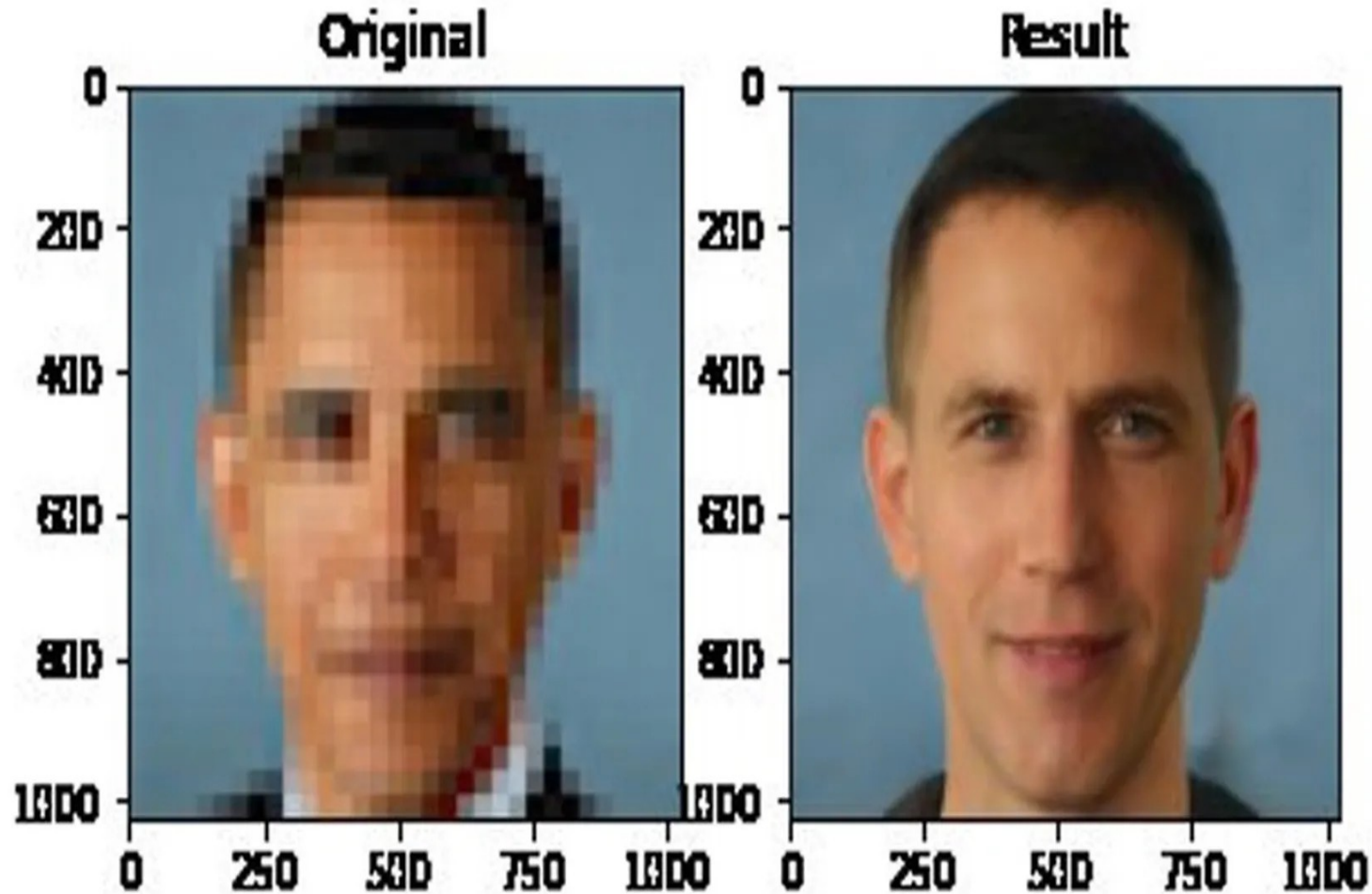
Website:

<https://dome-ml.org/>

Data Optimisation Model Evaluation

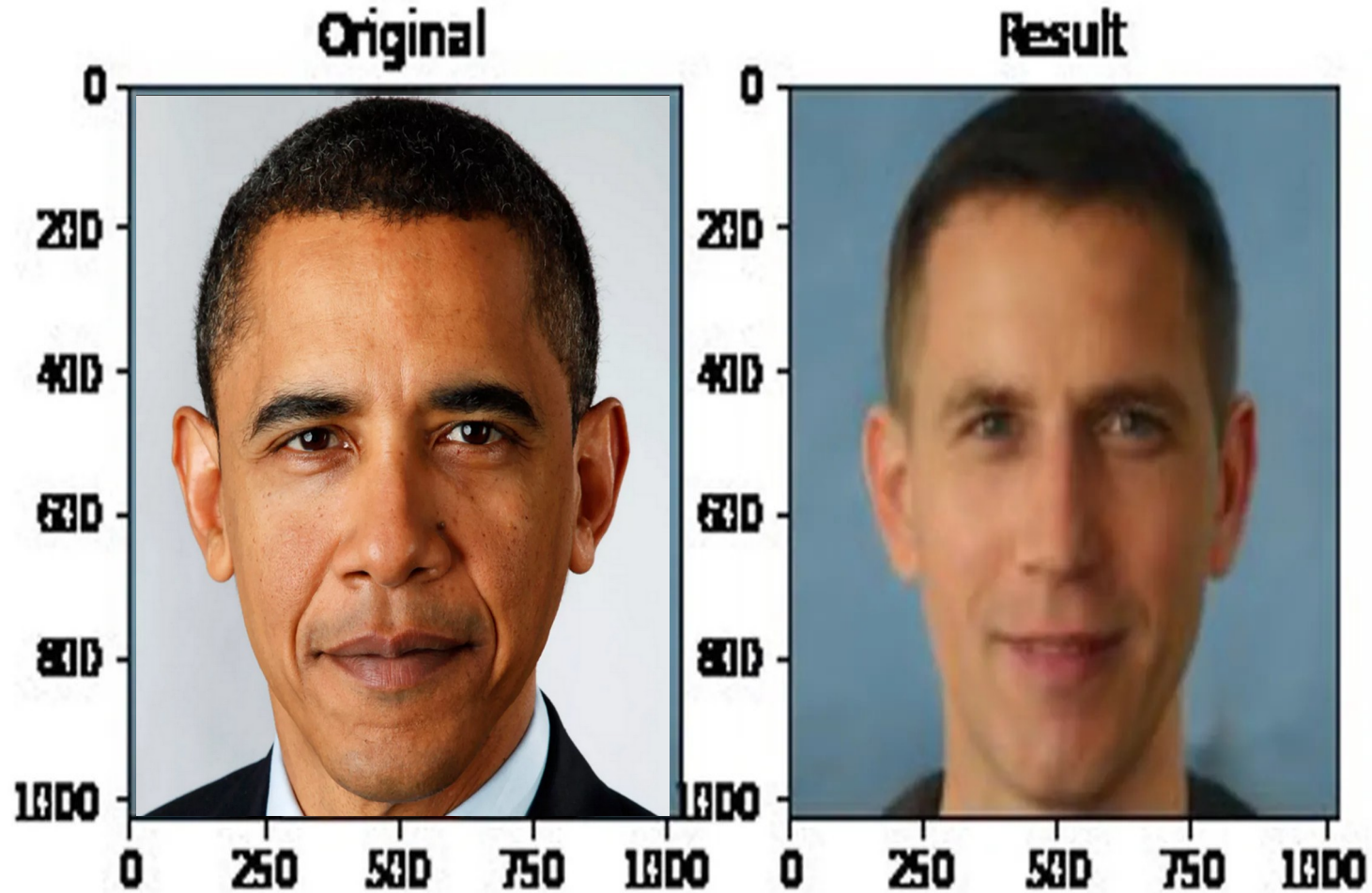


Dangers of deep/machine learning



<https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias>

Dangers of deep/machine learning: Bias



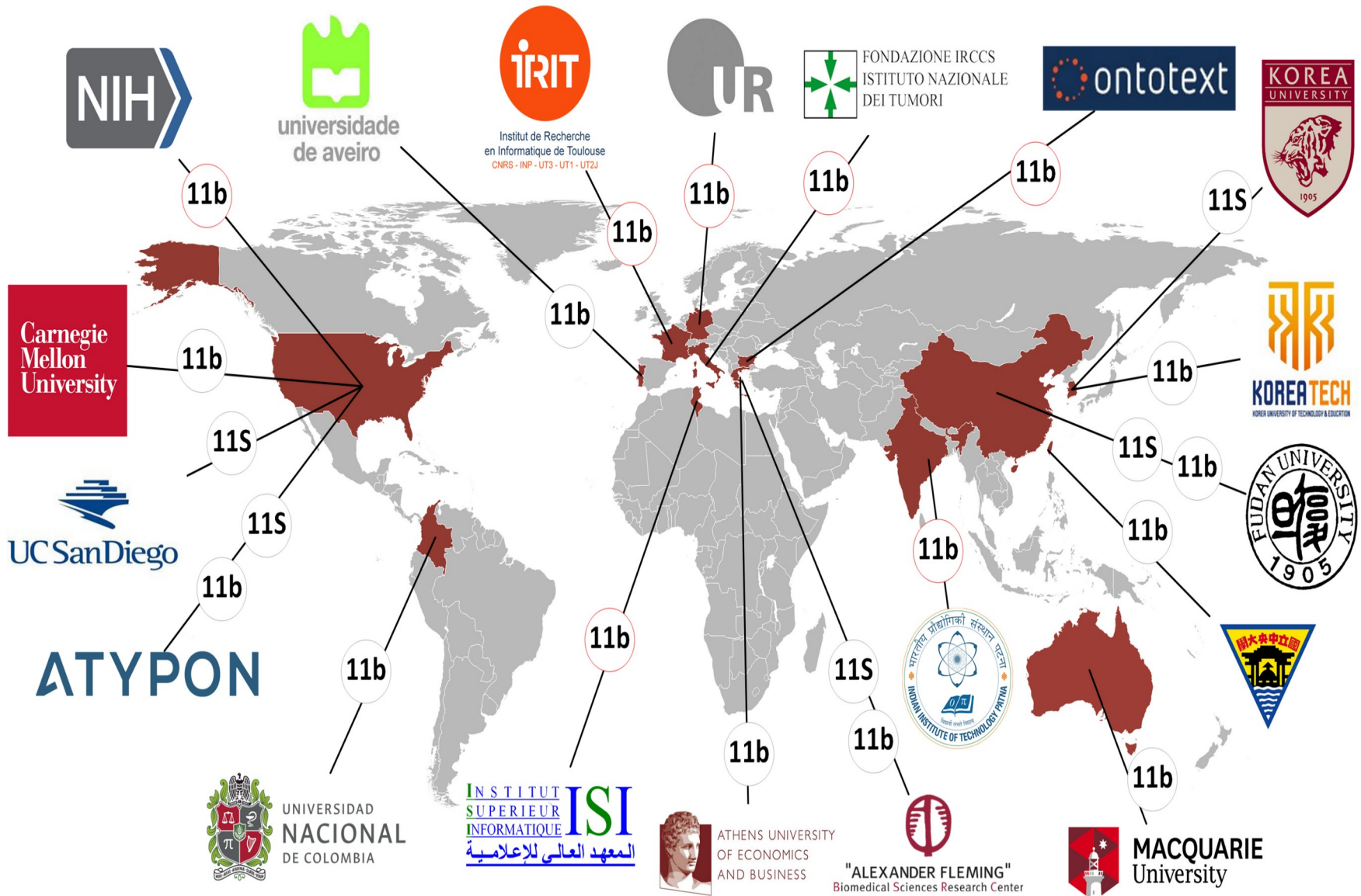
<https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias>

ITBI students are winners



Former ITBI student Dimitra Panou at the BioASQ workshop in Thessaloniki 20/9/2023

BioASQ: Int. competition for biomedical QA



Transformers help clustering all scientific papers

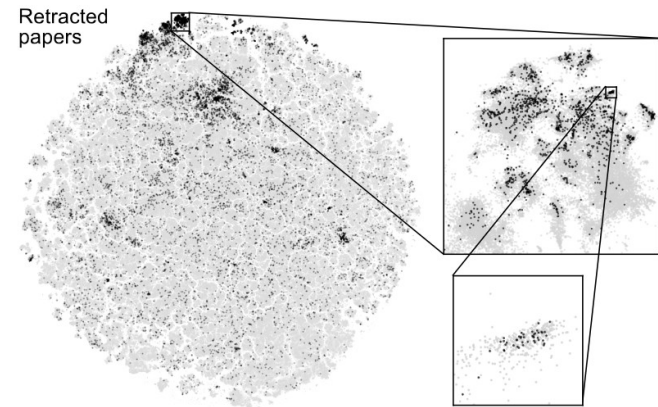
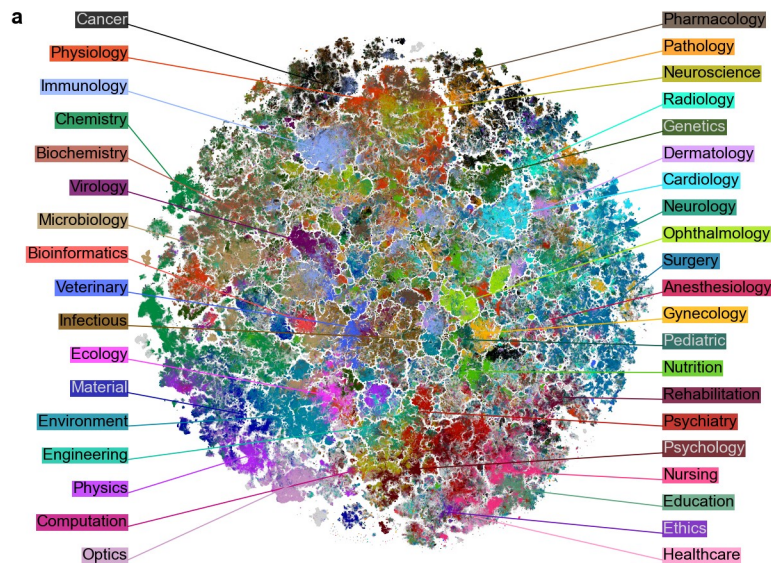


Figure 6: Retracted papers group together. All retracted papers with intact abstracts (11,756) are highlighted in black, plotted on top of the non-retracted papers. First inset corresponds to one of the regions with higher density of retracted papers (3.8%), covering research on cancer-related drugs, marker genes, and microRNA. Second inset corresponds to a subregion with a particularly high fraction of retracted papers (10.8%), the one we used for manual inspection.

<https://www.biorxiv.org/content/10.1101/2023.04.10.536208v2>

HEALTH & WELLNESS

A boy saw 17 doctors over 3 years for chronic pain. ChatGPT found the diagnosis

Alex experienced pain that stopped him from playing with other children but doctors had no answers to why. His frustrated mom asked ChatGPT for help.



Sept. 11, 2023, 5:42 PM EST / Updated Sept. 12, 2023, 5:31 PM EST / Source: TODAY

By Meghan Holohan

During the COVID-19 lockdown, Courtney bought a bounce house for her two young children. Soon after, her son, Alex, then 4, began experiencing pain.

“(Our nanny) started telling me, ‘I have to give him Motrin every day, or he has these gigantic meltdowns,’” Courtney, who asked not to use her last name to protect her family’s privacy, tells TODAY.com. “If he had Motrin, he was totally fine.”

Then Alex began chewing things, so Courtney took him to the dentist. What followed was a three-year search for the cause of Alex’s increasing pain and eventually other symptoms.

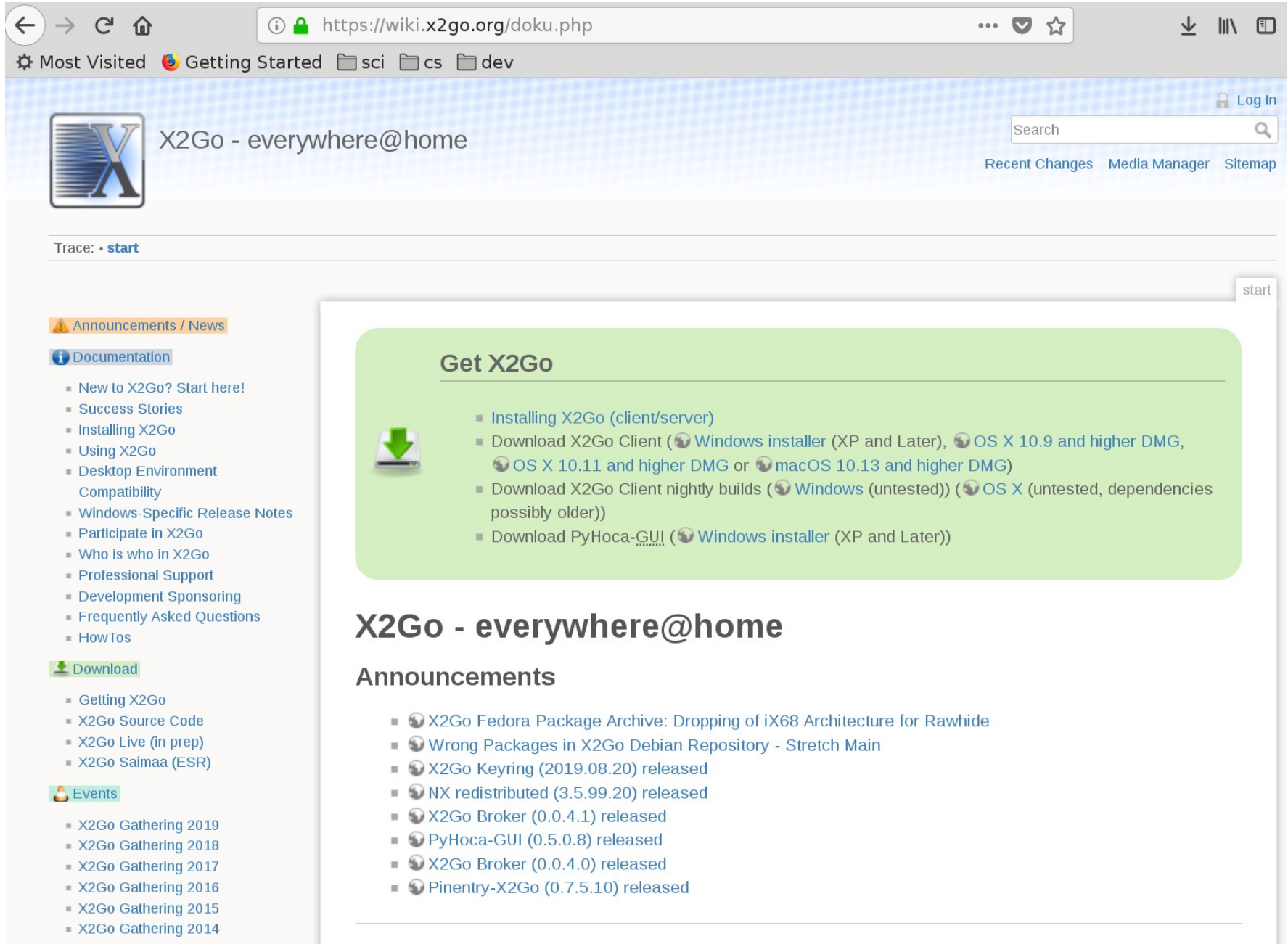


Alex saw 17 doctors over three years for his chronic pain, but none were able to find a diagnosis that explained all of his symptoms, his mom says. Courtesy Courtney

Get your account on the Virtual Machine for the exercises in hands-on during the lectures and at home

- **use 28 CPUs, 248GB RAM for all**
- **20GB disk-space for each + 400GB shared**

Install x2go to access graphical user interface



The screenshot shows a web browser window displaying the X2Go website. The browser's address bar shows the URL <https://wiki.x2go.org/doku.php>. The website header includes the X2Go logo, the text "X2Go - everywhere@home", a search bar, and links for "Log In", "Recent Changes", "Media Manager", and "Sitemap". A breadcrumb trail shows "Trace: - start".

The main content area features a green box titled "Get X2Go" with a download icon. Below this, there are sections for "X2Go - everywhere@home" and "Announcements".

Get X2Go

- Installing X2Go (client/server)
- Download X2Go Client (Windows installer (XP and Later), OS X 10.9 and higher DMG, OS X 10.11 and higher DMG or macOS 10.13 and higher DMG)
- Download X2Go Client nightly builds (Windows (untested)) (OS X (untested, dependencies possibly older))
- Download PyHoca-GUI (Windows installer (XP and Later))

X2Go - everywhere@home

Announcements

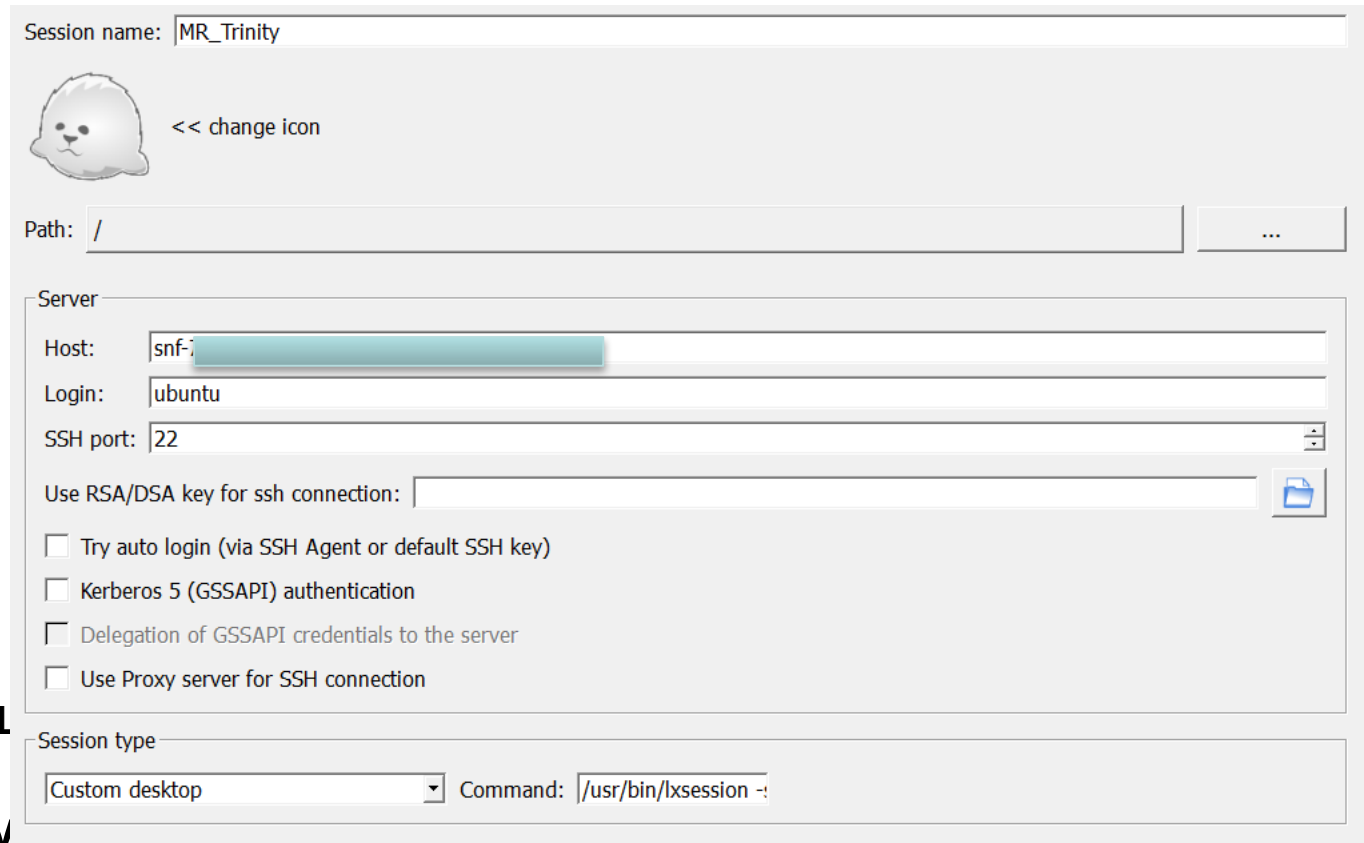
- X2Go Fedora Package Archive: Dropping of iX68 Architecture for Rawhide
- Wrong Packages in X2Go Debian Repository - Stretch Main
- X2Go Keyring (2019.08.20) released
- NX redistributed (3.5.99.20) released
- X2Go Broker (0.0.4.1) released
- PyHoca-GUI (0.5.0.8) released
- X2Go Broker (0.0.4.0) released
- Pinentry-X2Go (0.7.5.10) released

Left Sidebar:

- Announcements / News**
- Documentation**
 - New to X2Go? Start here!
 - Success Stories
 - Installing X2Go
 - Using X2Go
 - Desktop Environment Compatibility
 - Windows-Specific Release Notes
 - Participate in X2Go
 - Who is who in X2Go
 - Professional Support
 - Development Sponsoring
 - Frequently Asked Questions
 - HowTos
- Download**
 - Getting X2Go
 - X2Go Source Code
 - X2Go Live (in prep)
 - X2Go Saimaa (ESR)
- Events**
 - X2Go Gathering 2019
 - X2Go Gathering 2018
 - X2Go Gathering 2017
 - X2Go Gathering 2016
 - X2Go Gathering 2015
 - X2Go Gathering 2014

Access to virtual machine

- Install x2go from: <https://wiki.x2go.org/doku.php/download:start>



The screenshot shows the X2Go client configuration window for a session named "MR_Trinity". The window includes a session name field, a default icon of a white seal, and a "change icon" button. Below this is a "Path" field set to "/". The "Server" section contains fields for "Host" (prefixed with "snf:"), "Login" (set to "ubuntu"), and "SSH port" (set to "22"). There is also a field for "Use RSA/DSA key for ssh connection" with a file selection icon. Below these are four unchecked checkboxes: "Try auto login (via SSH Agent or default SSH key)", "Kerberos 5 (GSSAPI) authentication", "Delegation of GSSAPI credentials to the server", and "Use Proxy server for SSH connection". The "Session type" section has a dropdown menu set to "Custom desktop" and a "Command" field containing "/usr/bin/lxsession -i".

- X2go:

- Session ty

Lubuntu -e LXDE

- (Virtualbox : http://genomics-lab.fleming.gr/fleming/uoavm/TrinityVM_U16.ova)

Introduction to Bioinformatics 2022-2023

Exercise 1 (M. Reczko):

(Adapted from:

<https://web.archive.org/web/20150425010121/http://www.ableweb.org/volumes/vol-28/v28reprint.php?ch=8>

)

In a hypothetical scenario many people in a city suddenly come down with a serious illness. All the victims have in common is that they were all in a downtown pedestrian mall at a certain time five days before. Could terrorists have released a cloud of viruses or bacteria from a vehicle downwind of the mall? You work for the Centers for Disease Control and Prevention, and you have to find out.

A sample of non-human DNA (bacterial or viral) has been isolated from the victims. Identify the DNA sample as well as you can. Some of the DNA molecules are very short, and have been partially degraded. You will notice that the sequence is sprinkled with Ns, “N” stands for “nucleotide” and means that the nucleotide at that position could not be determined.

Some judgment is called for as you interpret your results. First, everyone has bacteria and viruses in his or her body, and sometimes they can cause disease. However, we are looking for exotic pathogens with bioterrorism potential (e.g., anthrax or smallpox rather than the common cold). Even AIDS, although it is deadly, would not work as a bioterror weapon because the disease develops too slowly and the virus is too hard to disseminate. For the purposes of this exercise, we will not consider a pathogen a bioterror agent unless it is listed as a potential agent on the Centers for Disease Control and Prevention Web site at <https://emergency.cdc.gov/agent/agentlist.asp> .

Second, organisms that are evolutionarily related have similar DNA, which might lead you to sound a false alarm. For example, say you find the following when you do a BLAST search on a certain DNA sample:

Sequences producing significant alignments:	Score (Bits)	E Value
gi 40012 emb X02369.1 BSORIC Bacillus subtilis oriC region	5967	0.0
gi 32468687 emb Z99104.2 BSUB0001 Bacillus subtilis complete ...	5967	0.0
gi 467326 dbj D26185.1 BAC180K B. subtilis DNA, 180 kilobase reg	5967	0.0
gi 39877 emb X12778.1 BSDNAA Bacillus subtilis dnaA gene 5'-regi	846	0.0
gi 56160984 gb CP000002.2 Bacillus licheniformis ATCC 14580, co	690	0.0
gi 52346357 gb AE017333.1 Bacillus licheniformis DSM 13, comple	690	0.0
gi 39878 emb X12779.1 BSDNAAN Bacillus subtilis genes for dnaA (587	8e-164
gi 39893 emb X17013.1 BSDPD Bacillus subtilis lys gene for di...	525	2e-145
gi 51973633 gb CP000001.1 Bacillus cereus E33L, complete genome	337	1e-88
gi 49328240 gb AE017355.1 Bacillus thuringiensis serovar kon...	329	3e-86
gi 50082967 gb AE017334.2 Bacillus anthracis str. 'Ames Ancesto	329	3e-86
gi 49176966 gb AE017225.1 Bacillus anthracis str. Sterne, compl	329	3e-86

Bacillus subtilis is a harmless and very common soil bacterium. It is closely related to *Bacillus anthracis*. *Bacillus anthracis* causes anthrax, and is a dangerous bioterror weapon. Note from the similarity score (second column from the right) that *Bacillus subtilis* DNA is far more similar to the sample than *Bacillus anthracis* DNA is. Unless one of your samples gives a stronger indication of *Bacillus anthracis* than this, the mention of *B. anthracis* in the output is probably just due to genetic similarities between it and *B. subtilis*.

1. Analyze the samples

>outbreak14

```
GCCGAGTTAGTCTTGTGCTNACGGAACTTATTGTATGAGTANTGATTTGAAAGAGCTANANT  
TAAAAAATCACTAATNAATNTAAGAGCGGACTTAACNAGCGTAAACTGTCTTACTAATTAAT  
TGTCAGTTAGCTCGTTCAGGTAATGGTTCCTANCGGNCAATGCAGGAAGAGTTCTACCTGG  
AACTGANAGACCGCTGGCGGTGACAACACACTACGTCAAATAAGA
```

>outbreak15

```
TAGTCTTGTGCTNACGGAACTTATTTATGAGGTACCCACCGANTCTGAAAACCGCTAATANA  
GCACTTTAAAATAAGAGCAGAATGGGATTTAAGGATAG
```

separately using both megablast and blastn at

https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&BLAST_SPEC=&LINK_LOC=blasttab

and to determine if there is any evidence of bioterror agents. Use the general nucleotide collection (nr/nt). Report any differences between the 2 algorithms.

2. Check the CDC Web site at <https://emergency.cdc.gov/agent/agentlist.asp> .

to see if the CDC considers any found organism to be a potential weapon. If you've found a bioterror agent, research it on the CDC site so you can describe its effects on humans.

3. The health effects of many pathogenic bacteria are briefly described on the NCBI Genomes Web site at <<http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi>>. Click on a species name to see its information. It also might be helpful to do a general web search.

**SEND SOLUTIONS (for M.Reczko exercise) ONLY TO:
mareczko@di.uoa.gr**