

Introduction to Bioinformatics (ITBI)

Martin Reczko* + Alexandros Dimopoulos

**Staff research scientist professor level*

Institute for Fundamental Biomedical Science

Biomedical Sciences Research Center "Alexander

Fleming"

Head of Node - ELIXIR-Greece

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

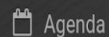


Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών
— ΙΔΡΥΘΕΝ ΤΟ 1837 —

Search...



▼ Course Options



Agenda



Announcements



Assignments



Documents

6



Exercises



Links



Messages

1



Questionnaires



mareczko



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 Αθηνών
— ΙΔΡΥΘΕΝ ΤΟ 1837 —

Search...



Course Options



Agenda



Announcements



Assignments



Documents



Exercises



Links



Messages



Questions



Αρχική Σελίδα / Introduction to Bioinformatics / Documents

Introduction to Bioinformatics (M413)



Documents

Root directory

Type	Filename ▾	Size	Date	
	2024-25		10/7/24	
	FOSSwire Unix/Linux Command Cheat Sheet	69.09 KB	10/19/17	
	Grades - February 2024	96.54 KB	5/15/24	

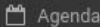


Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών
— ΙΔΡΥΘΕΝ ΤΟ 1837 —

Search...



Course Options



Agenda



Announcements



Assignments



Documents



Exercises



Links



Messages

Αρχική Σελίδα / Introduction to Bioinformatics / Documents

Introduction to Bioinformatics (M413)



Documents

Root directory » 2024-25



Up

Type	Filename ▾	Size	Date	
	exercises		10/9/24	
	lectures		10/9/24	

Please verify name+email in participant list at

<https://tinyurl.com/f6kbjk43>

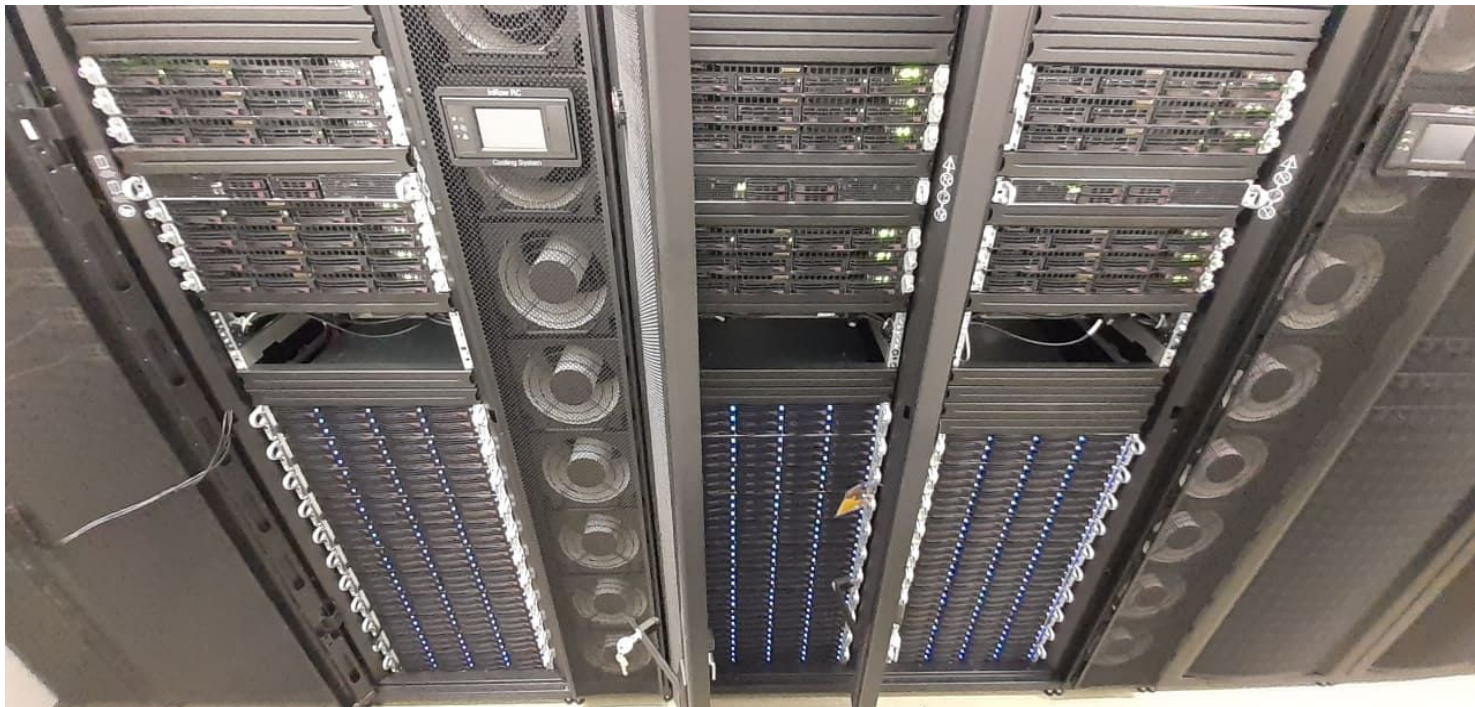
Enter all emails you might use
to get an account on the virtual machine from



"ALEXANDER FLEMING"
Biomedical Sciences Research Center



20 CPUs, 512GB RAM, 1024 GB disk shared for all



Syllabus and grading

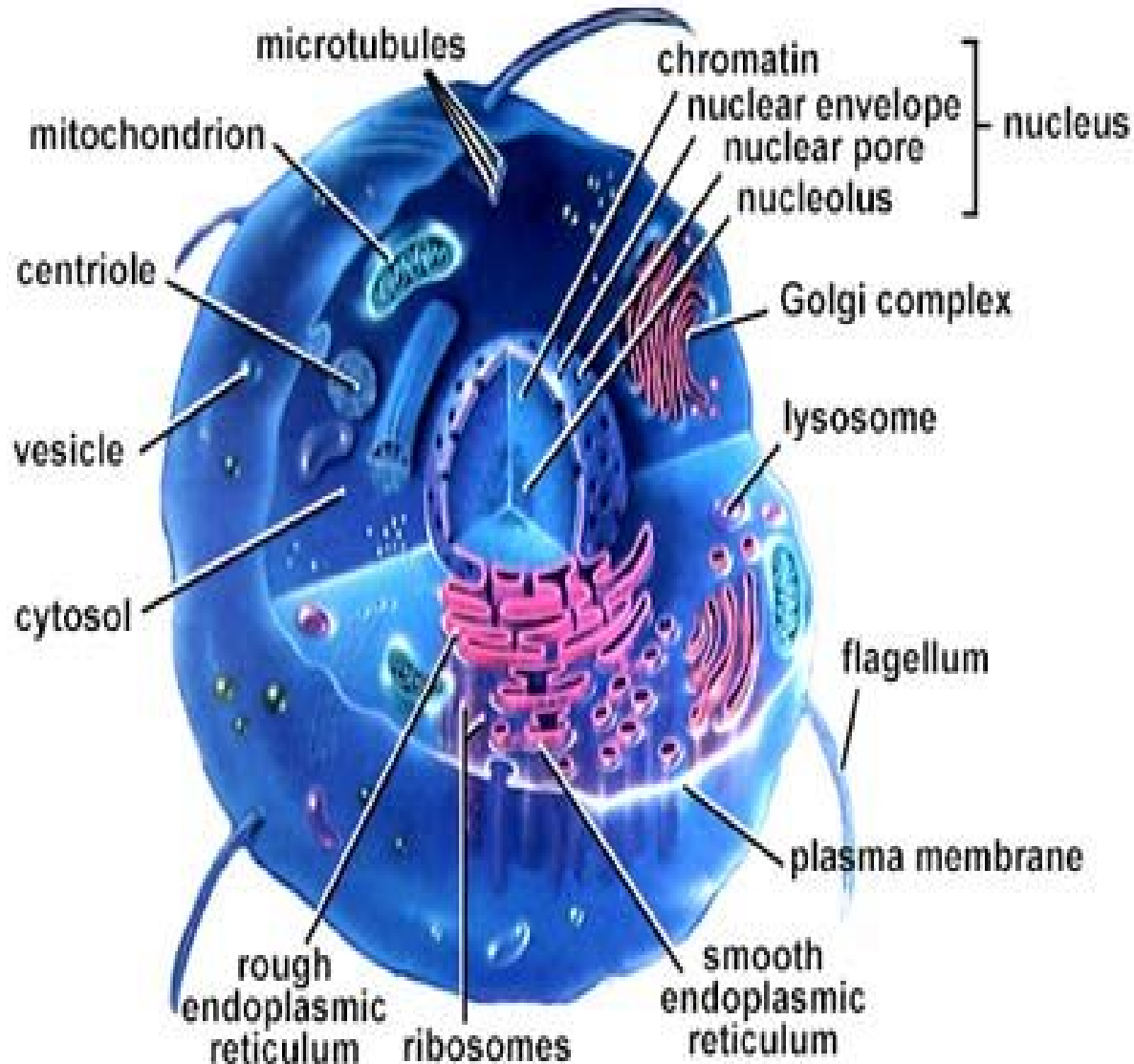
#	Date	Lecturer	Subject
1	6/10/26	MR	Overview of <u>Bioinformatics</u> , sequence alignment
2	13/10/26	AD	Introduction to Linux and the command line, bash scripting and ssh
3	20/10/26	MR	Next generation sequencing: introduction, quality control and gene expression analysis for <u>RNAseq</u>
4	27/10/26	MR	NGS for <u>denovo</u> genome and <u>transcriptome</u> assembly
5	3/11/26	AD	Introduction to the R programming language and <u>Rstudio</u> usage
6	10/11/26	AD	Advances R subjects, introduction to <u>Bioconductor</u>
	17/11/26		
7	24/11/26	MR	NGS analysis for molecular interactions (<u>ChipSeq</u> , (Par-)Clip, structural sequencing, chromosome conformation capture)
8	1/12/26	AD	Command line tool usage: <u>bedtools</u> , <u>vcftools</u> , <u>samtools</u> etc.
9	8/12/26	MR	Genome-scale models of metabolism and <u>macromolecular</u> expression, Biological applications of Transformers
10	15/12/26	AD	Pipelines for <u>SNP</u> calling, especially for <u>exome</u> sequencing using the <u>GATK</u> pipeline
11	22/12/26	MR+AD	Support for the final project (online)
12	12/1/27	MR+AD	Paper presentations by students
13	19/1/27	MR+AD	Paper presentations by students

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

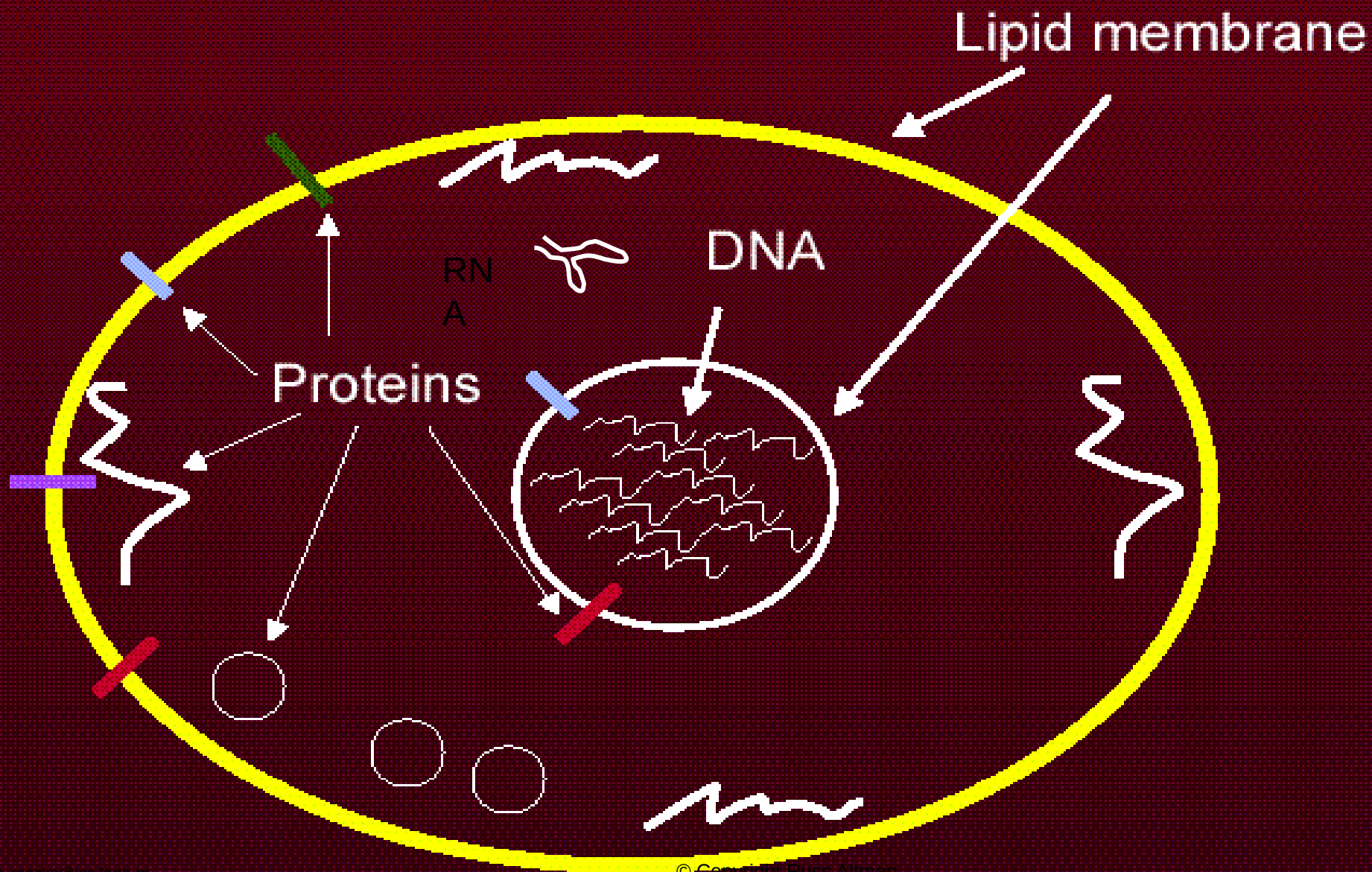
Subjects:

- 'Just enough' biology
- Dynamic programming
- Approximate string similarity
- Bioinformatics fields
- Recent machine learning results

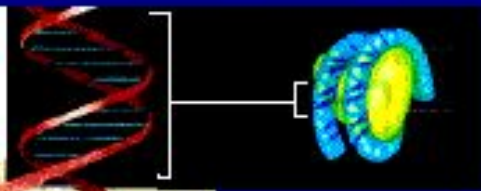
An Eukaryotic Cell (biological



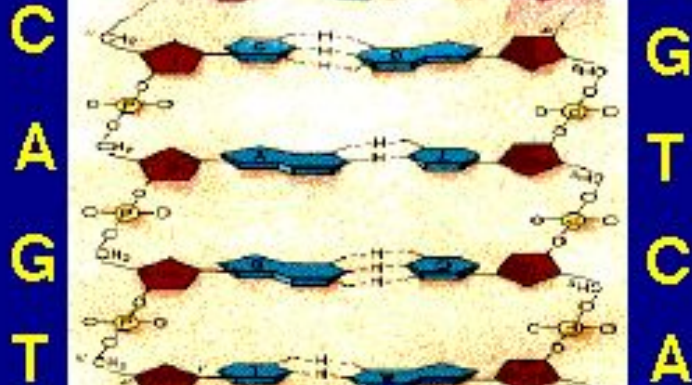
Bioinformatics Schematic of a Cell



THE DNA DOUBLE HELIX



Watson and Crick, 1953 in Cambridge



Hydrogen bonds (->Hybridization)



Rosalind Franklin

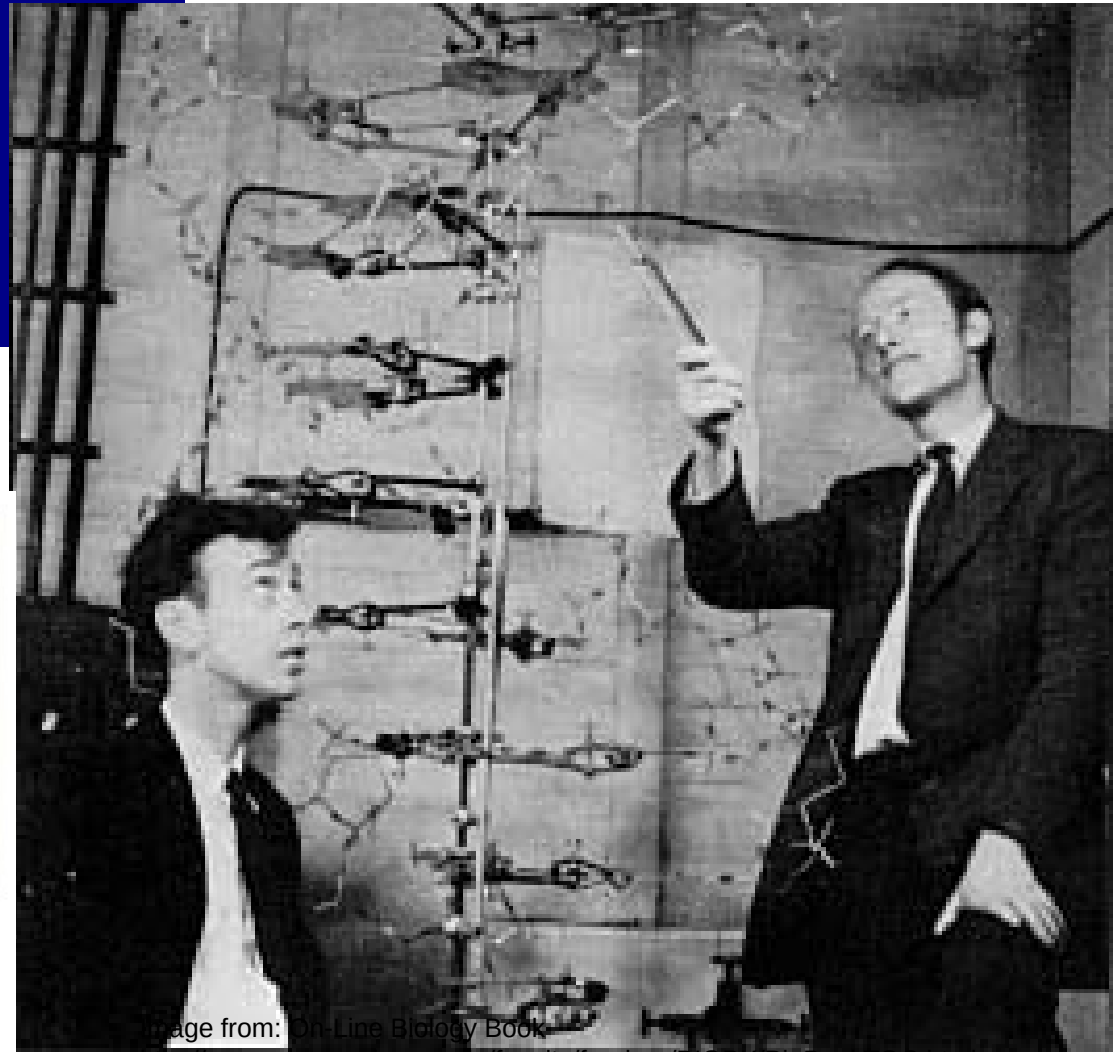
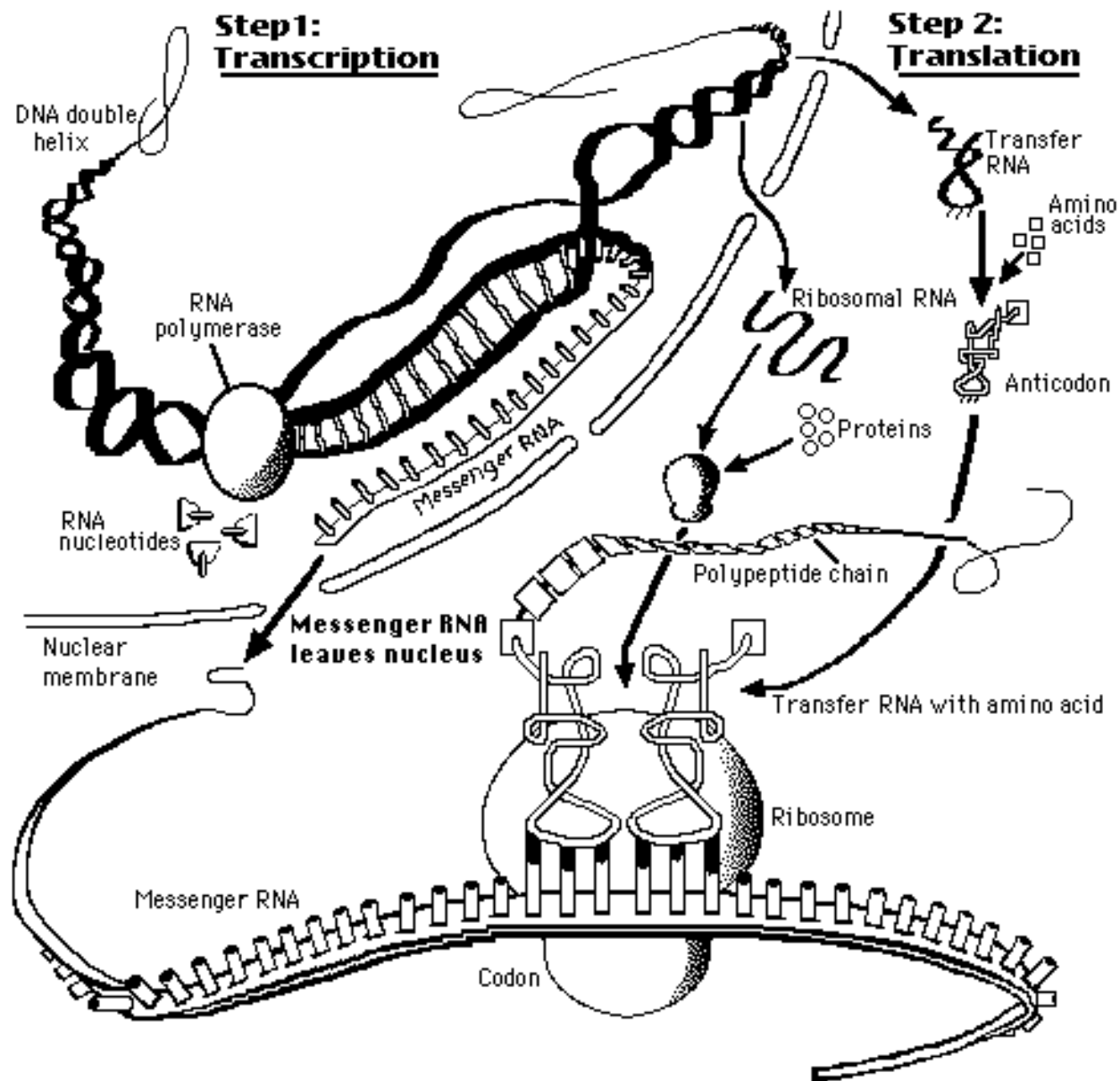


Image from: On-Line Biology Book

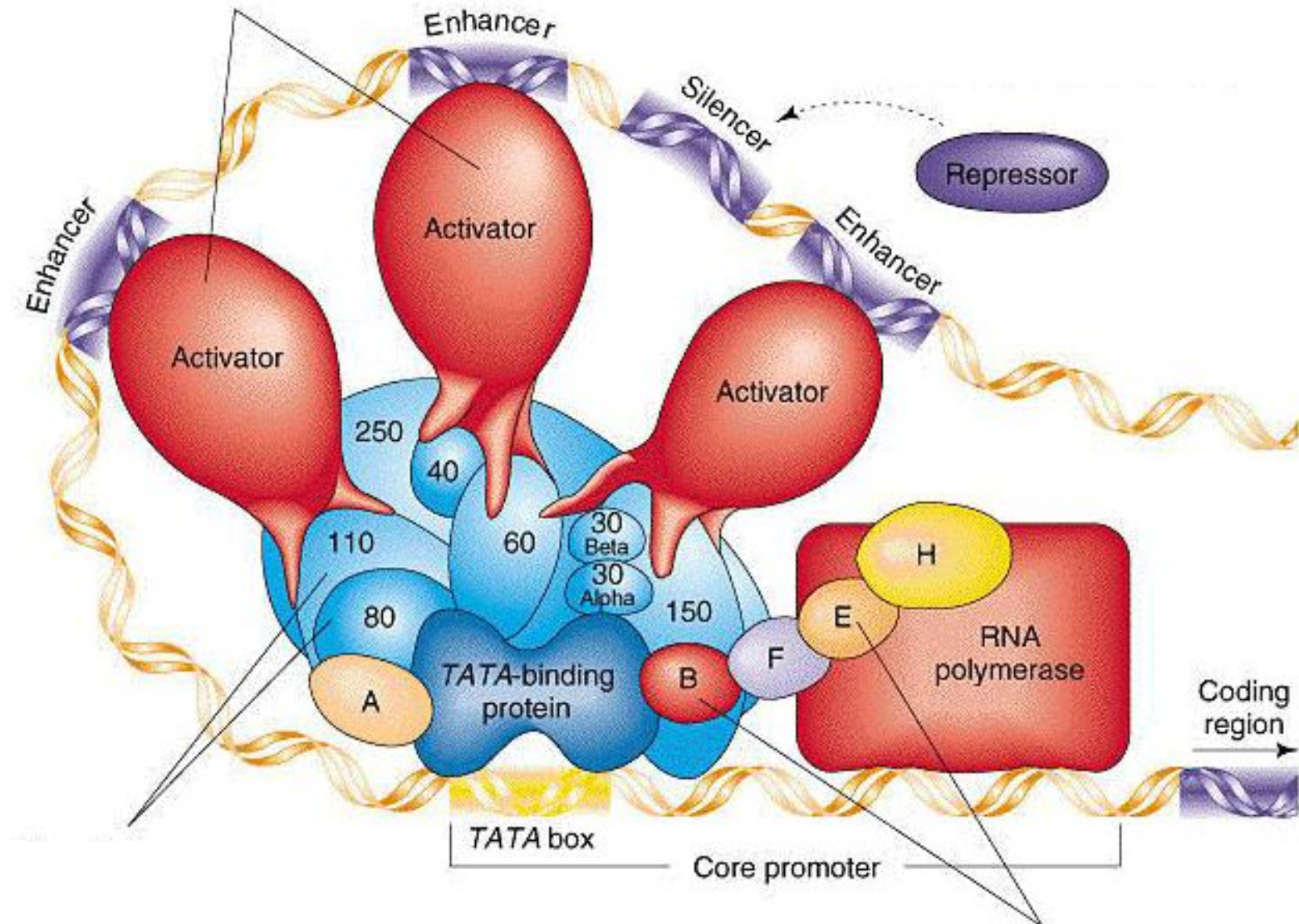
PROTEIN SYNTHESIS



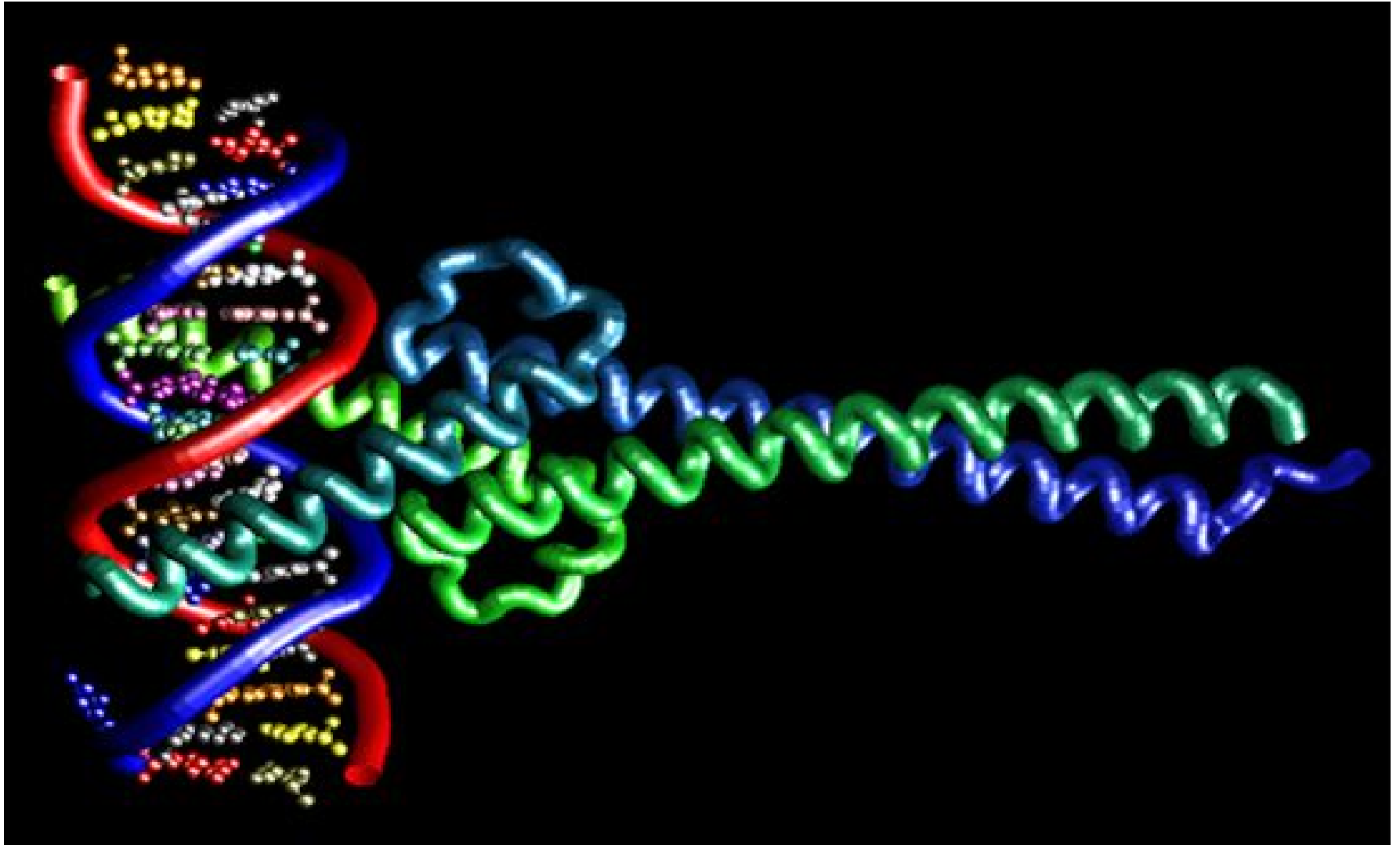
Transcription

- 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

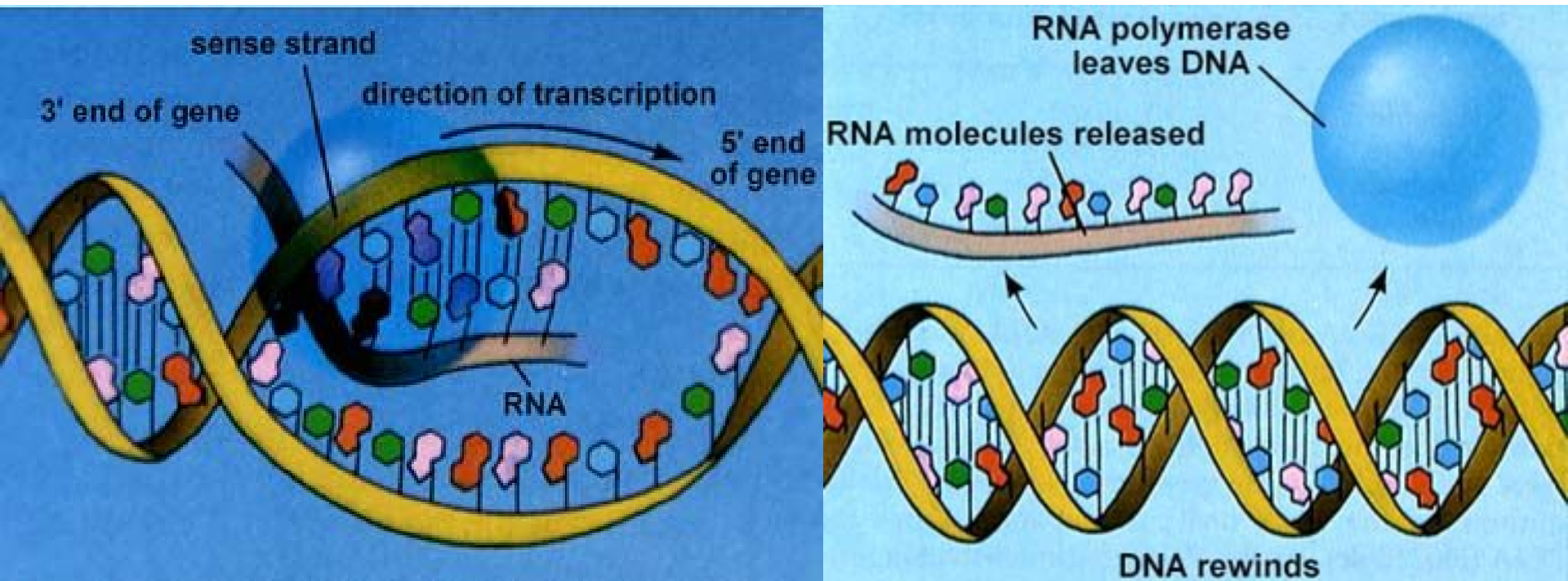
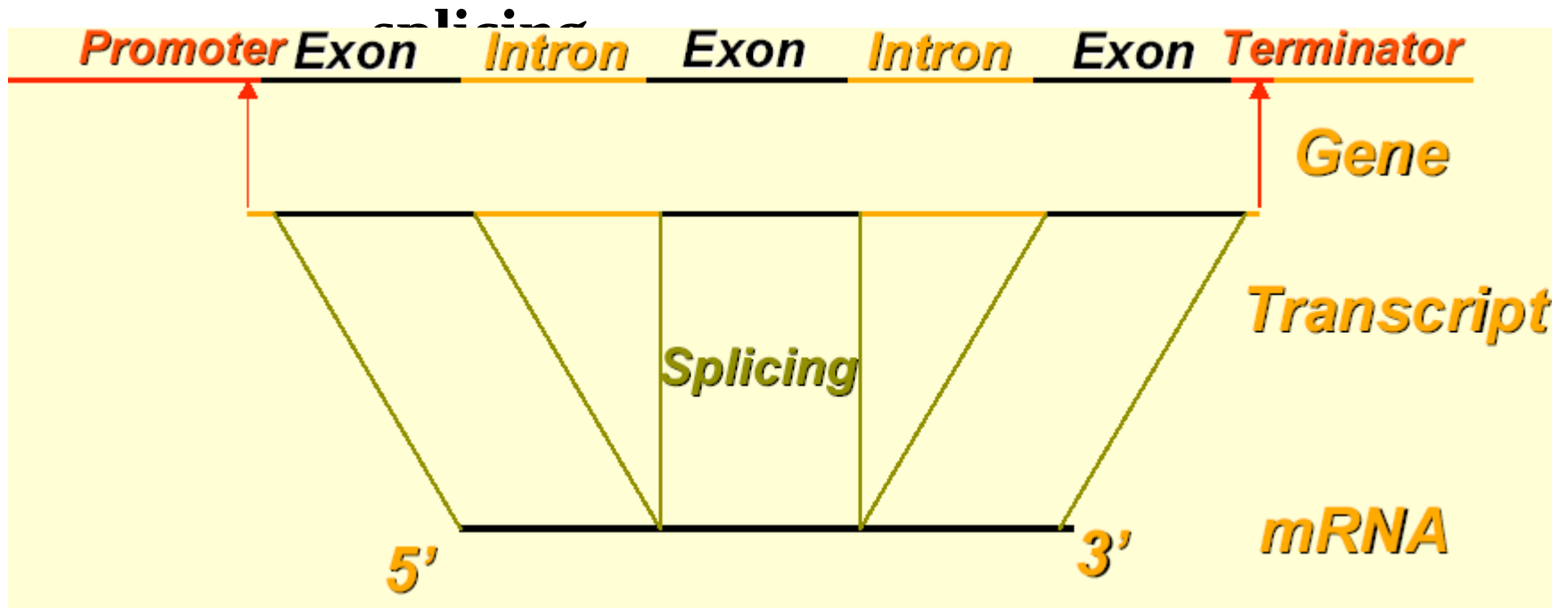


Image from: On-Line Biology Book
<http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookTOC.html>

RNA processing

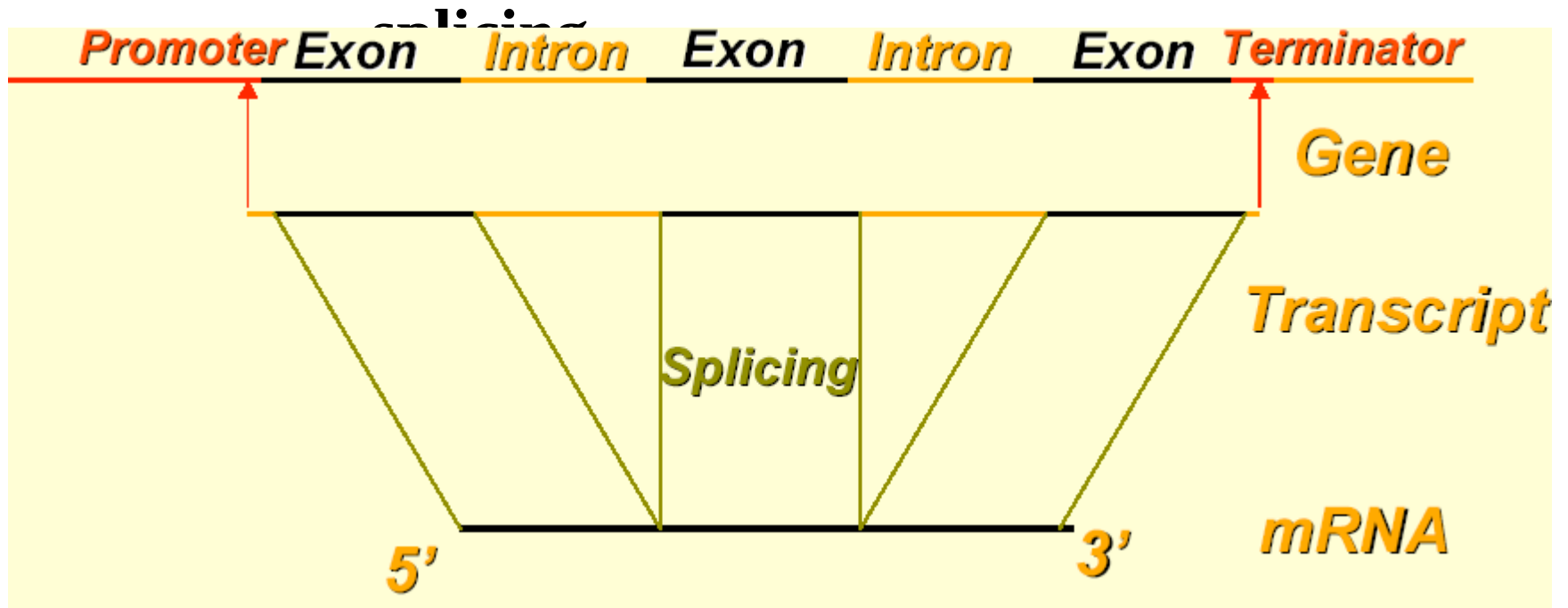
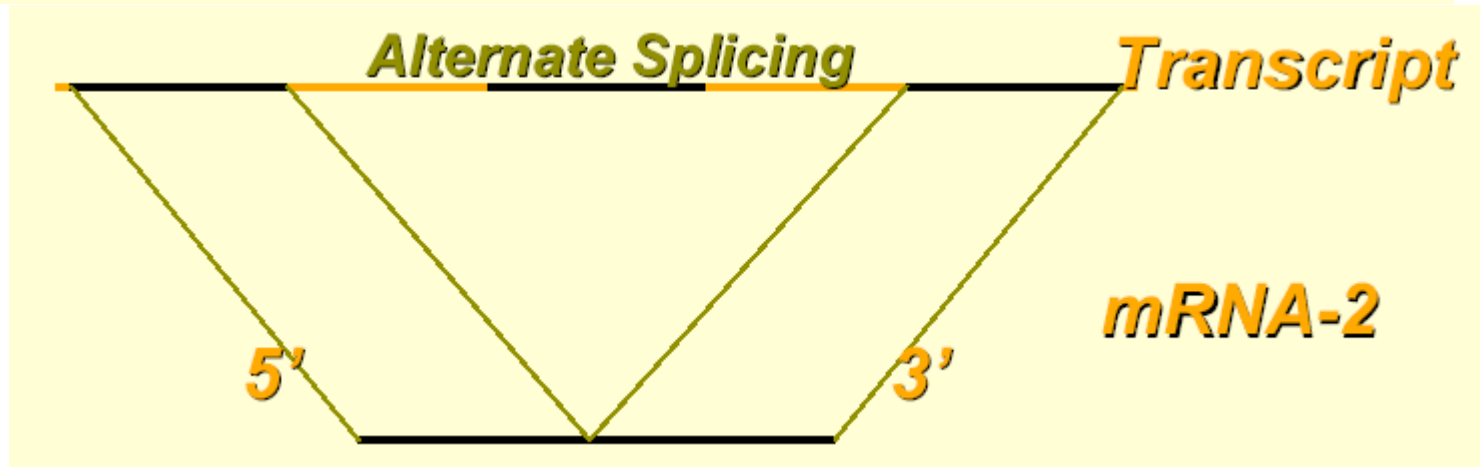
- eukaryotic genes are interrupted by **introns**
- these are "spliced" out to yield final messenger RNA (mRNA)
- splicing done by spliceosomes
- splicing sites are quite degenerate but not all are used

Processing of RNA =



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

Processing of RNA =

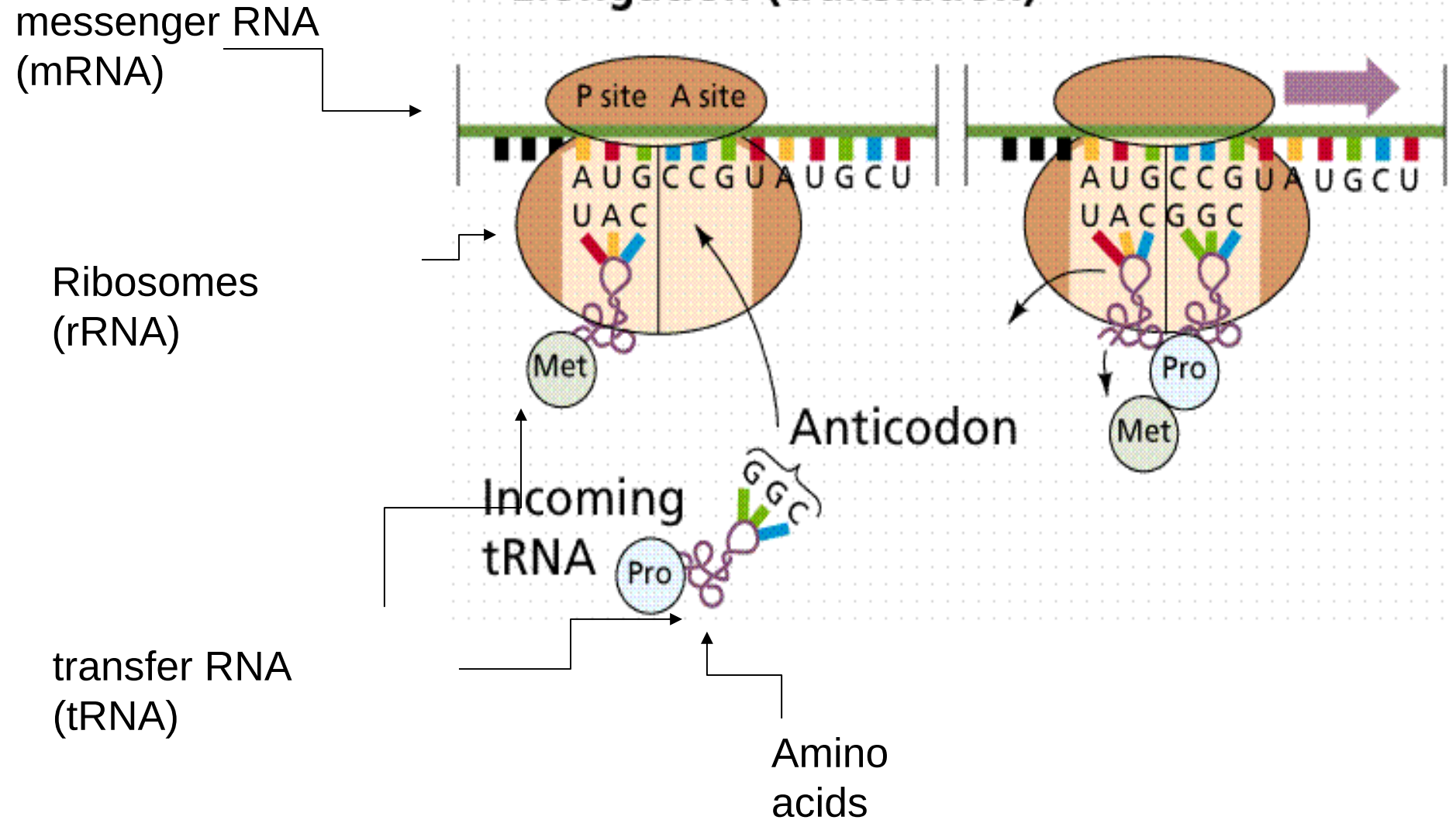


Images from: <http://biochem218.stanford.edu> (Doug Brutlag)

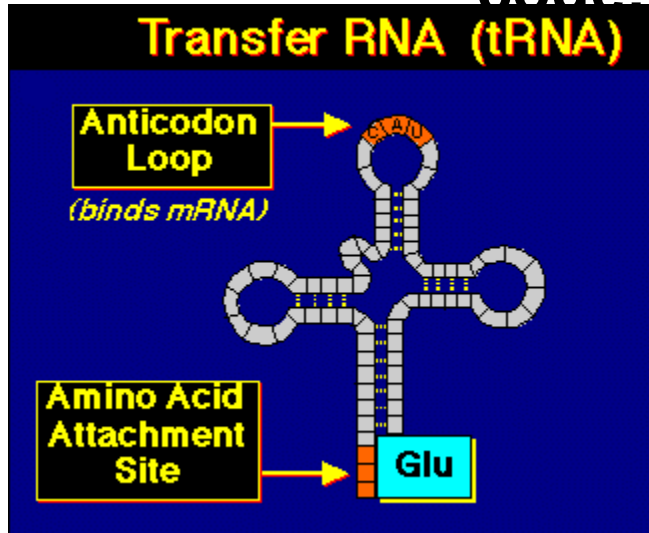
Translation

- 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)



The 'universal' genetic code:



64 different transfer RNA molecules

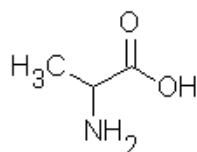


Second letter

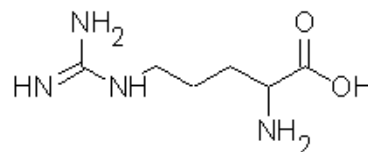
First letter

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

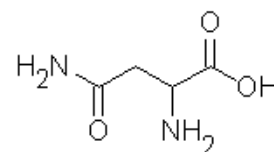
The 20 amino acids, building blocks for proteins



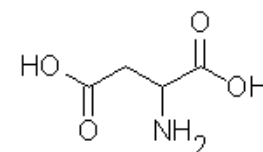
Alanin (Ala)



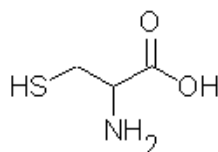
Arginin (Arg)



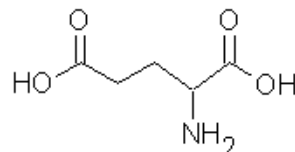
Asparagin (Asn)



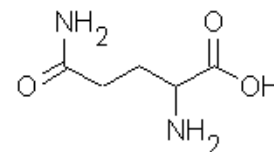
Asparaginsäure (Asp)



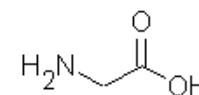
Cystein (Cys)



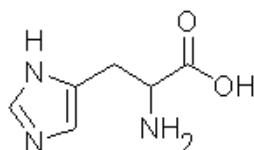
Glutaminsäure (Glu)



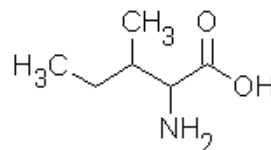
Glutamin (Gln)



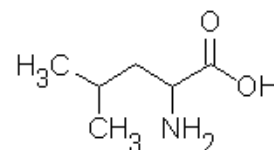
Glycin (Gly)



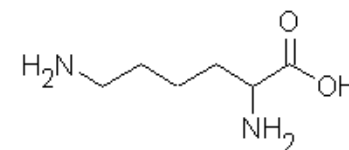
Histidin (His)



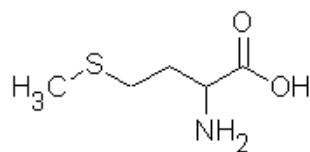
Isoleucin (Ile)



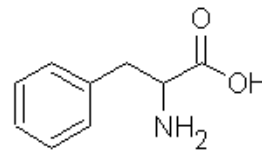
Leucin (Leu)



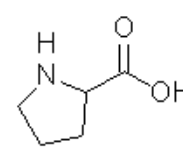
Lysin (Lys)



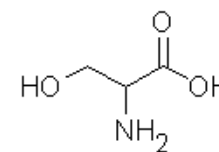
Methionin (Met)



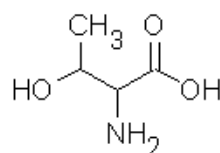
Phenylalanin (Phe)



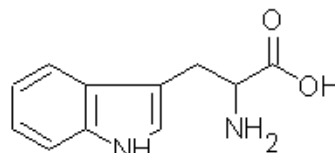
Prolin (Pro)



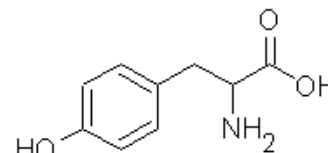
Serin (Ser)



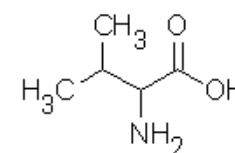
Threonin (Thr)



Tryptophan (Trp)



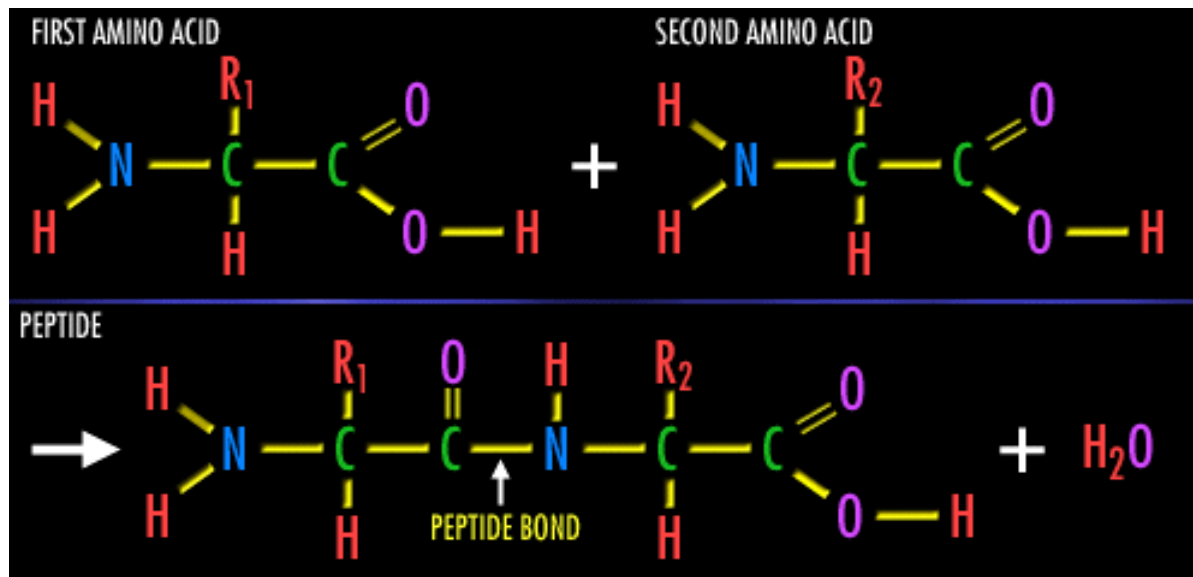
Tyrosin (Tyr)



Valin (Val)

Building proteins (chemistry):

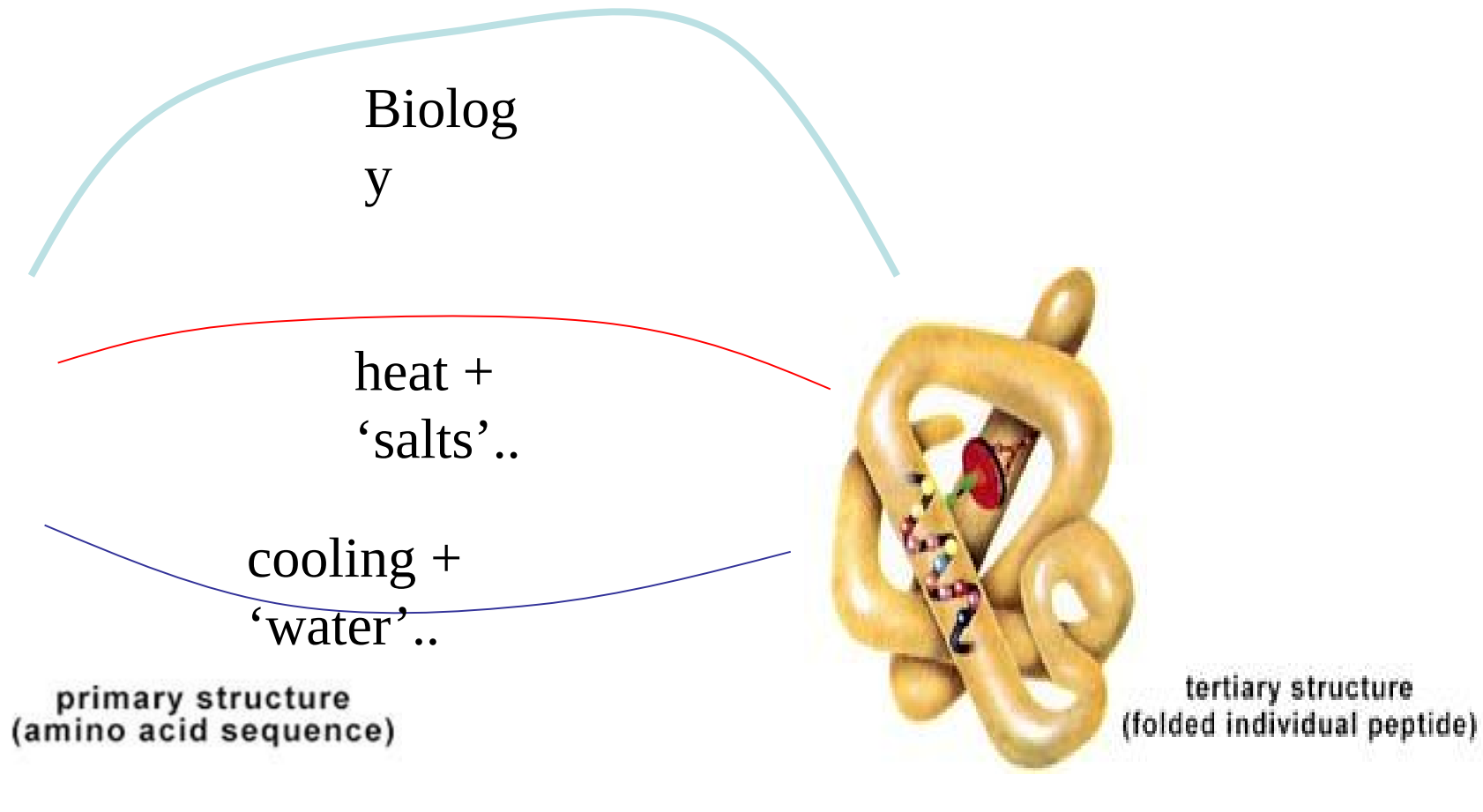
Amino acids are linked together by joining the amino end of one molecule to the carboxyl end of another. Removal of water allows formation of a type of covalent bond known as a peptide bond.



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

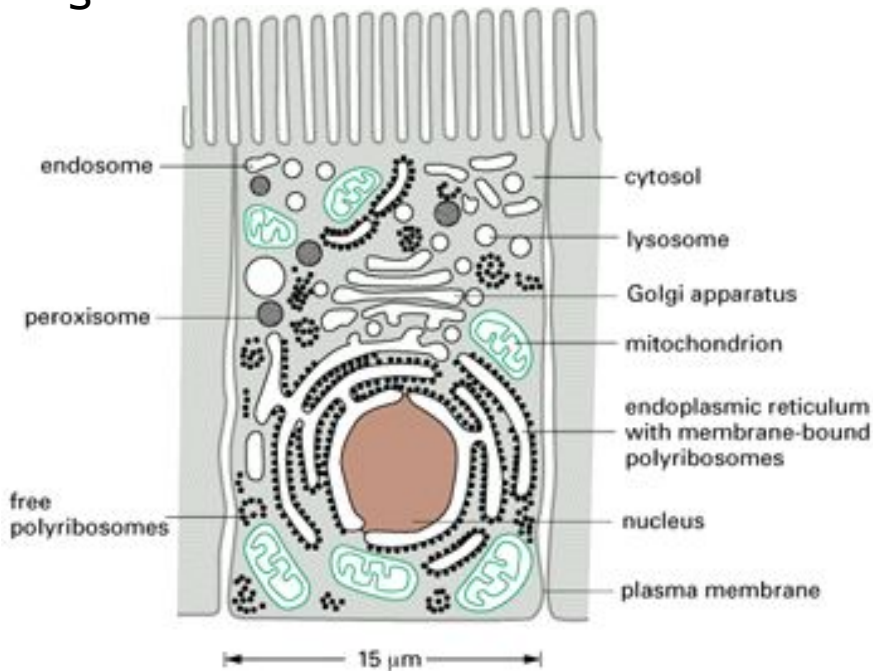
http://www.biosci.uga.edu/almanac/bio_103/notes/may_14.html.

Protein localization

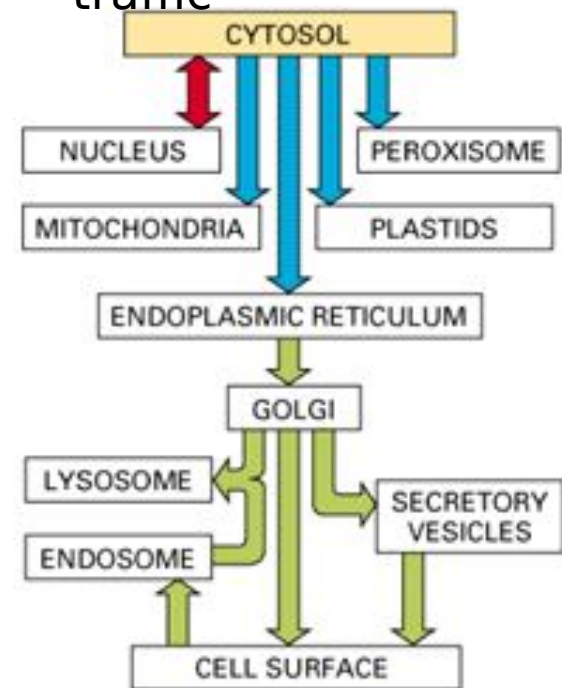
- leader sequences can specify cellular location (e.g., insert across membranes)
- leader sequences usually removed by cleavage
- Like an address sticker

Protein localization

compartments



protein traffic



KEY: █ = gated transport
█ = transmembrane transport
█ = vesicular transport

UNFOLDED PROTEIN

FOLDED PROTEIN

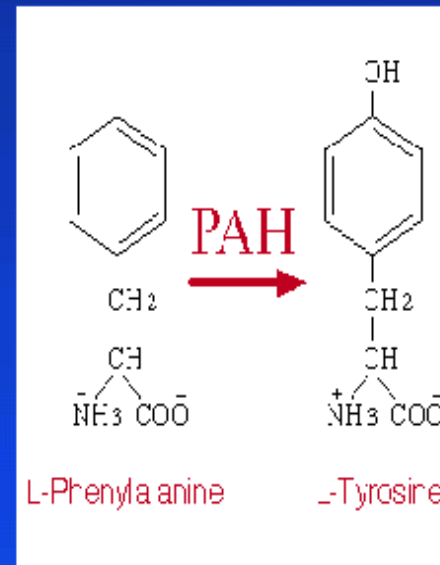
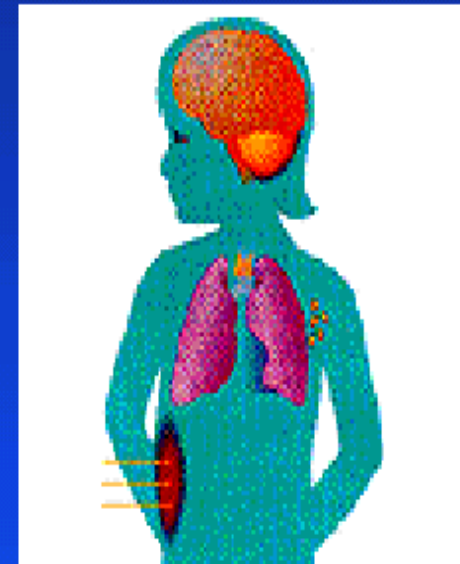


(A)

Central Paradigm of Bioinformatics



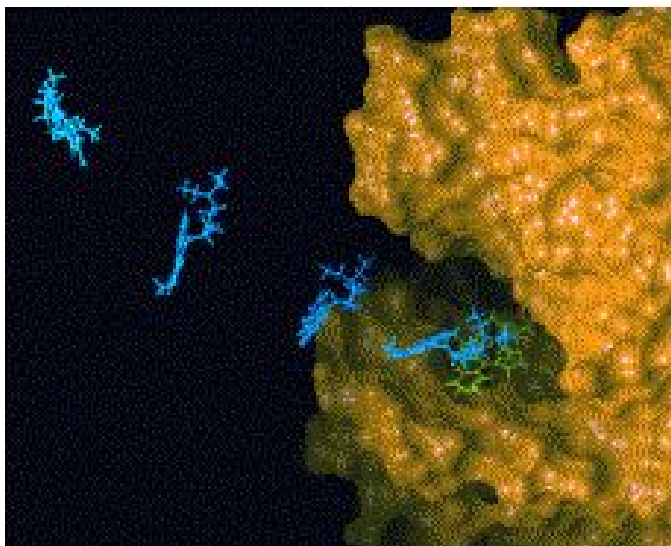
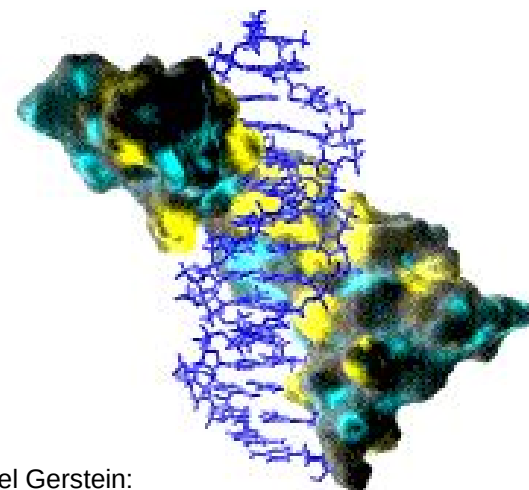
TGCTTTAGCTTT
AAACTACAGGCC
TCACTGGAGCTA
GAGACAAGAAGG
TAAAAACGGCT
GACAAAAGAAGT
CCTGGTATCCTC
TATGATGGGAGA
AGGAACTAGCT
AAAGGGAAGAAT
AAATTAGAGAAA
AACTGGAATGAC
GCTTATACCTGG



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).

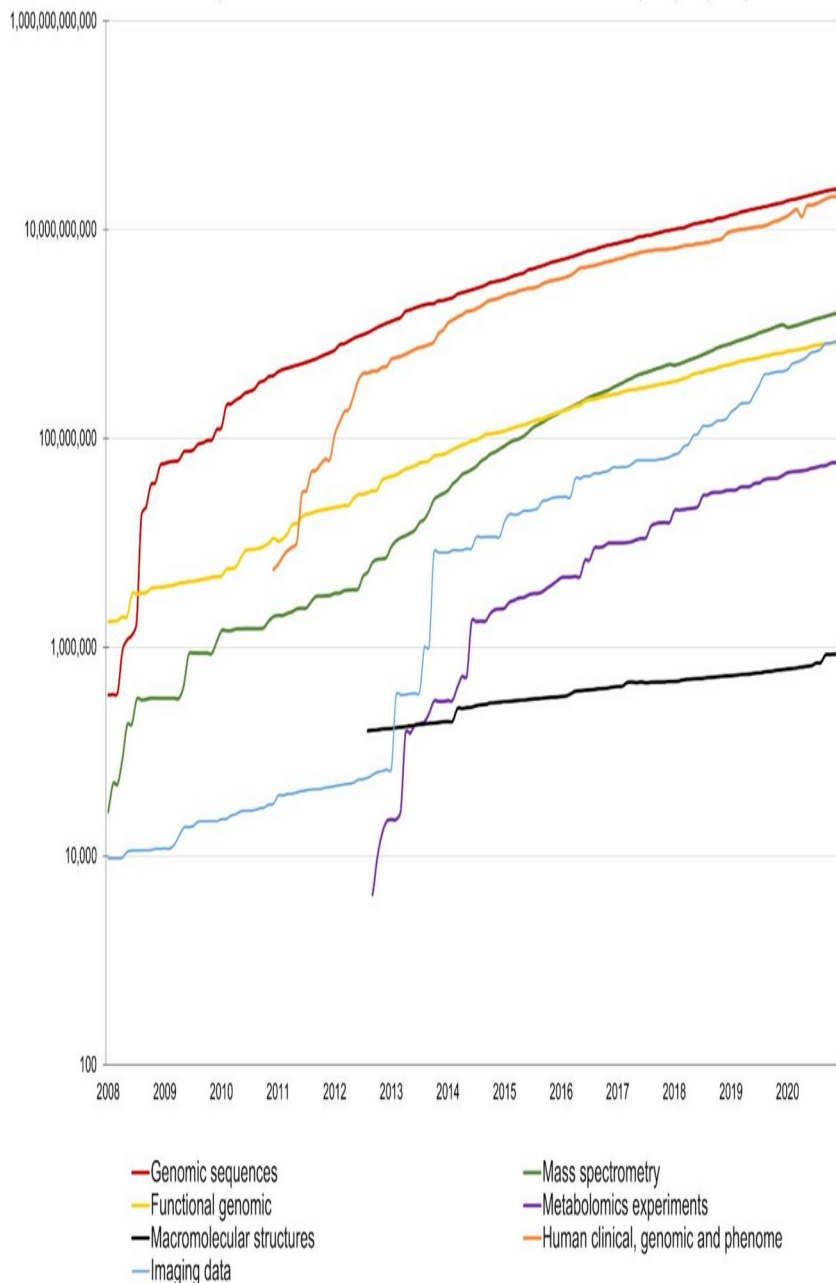


Michael Gerstein:
<http://bioinfo.mbb.yale.edu/mbb452a/intro/intro.pdf>

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 of EMBL-EBI services volume of data (megabytes)



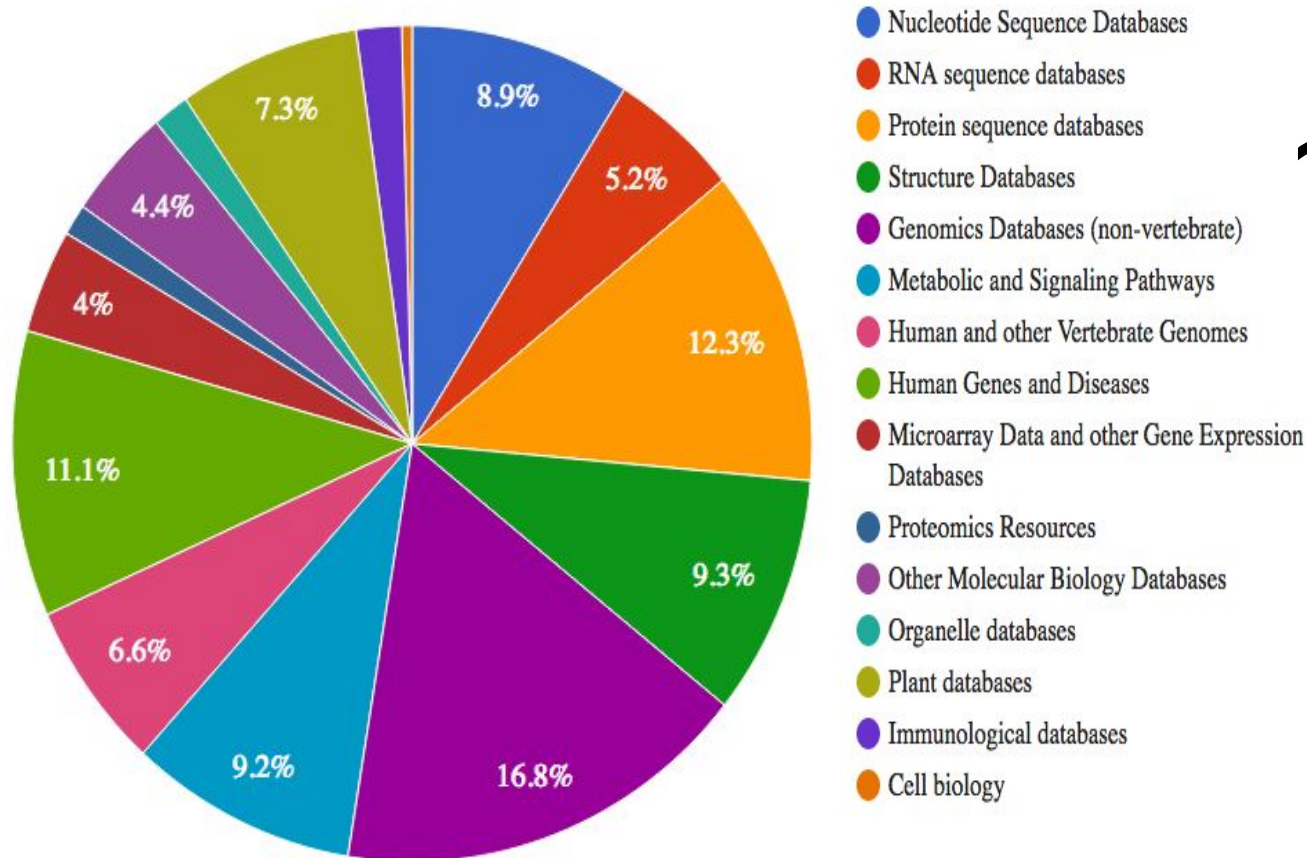
Data growth in the life sciences

- 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

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



**~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
Bacteria	1	~400,000 - ~10,000,000
Yeast	12	14,000,000
Worm	6	100,000,000
Fly	4	300,000,000
Weed	5	125,000,000
Human	23	3,000,000,000

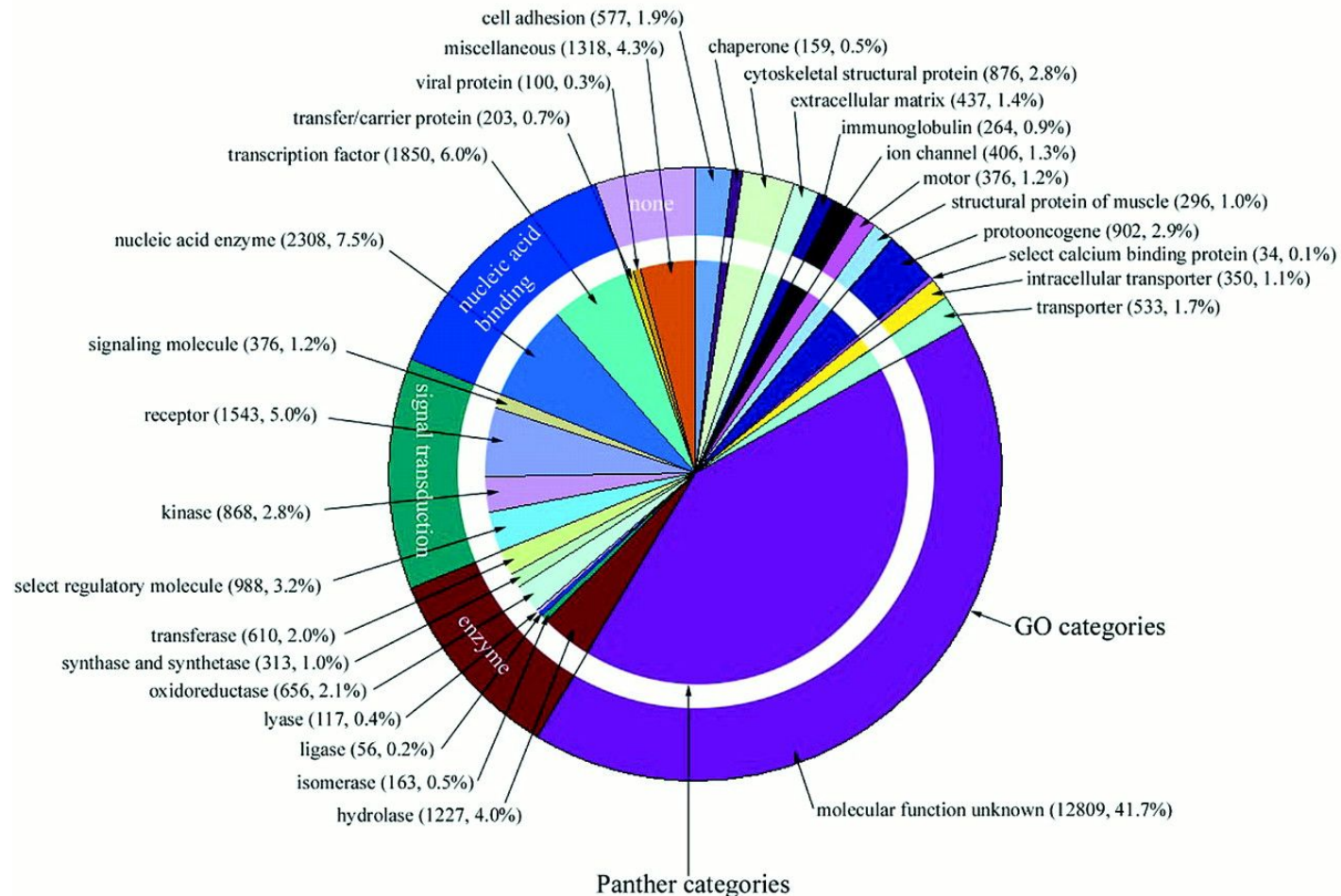
Only the surface is scratched:

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

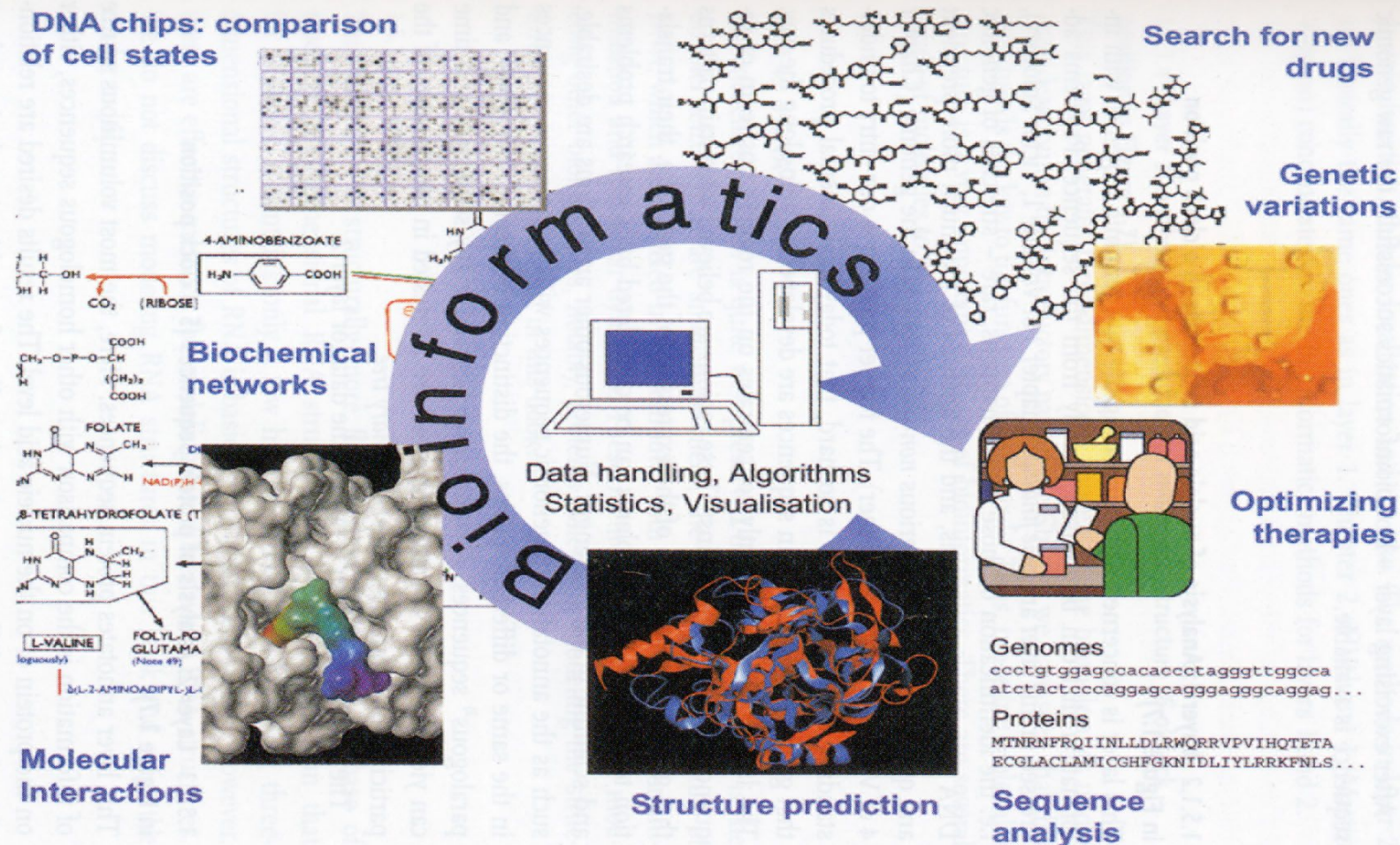


Fig. 1.7

A schematic overview of bioinformatics

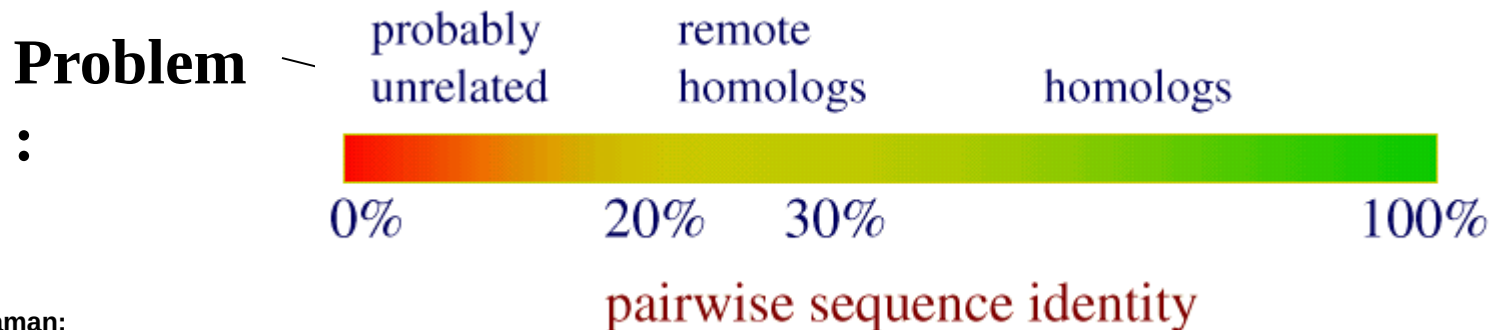
NGS+Robotics at BSRC Alexander Fleming:

<https://www.youtube.com/watch?v=8CaUGFimbgQ>

<https://www.youtube.com/watch?v=kUdDY3kvWpc>

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
```

Definition

S

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.

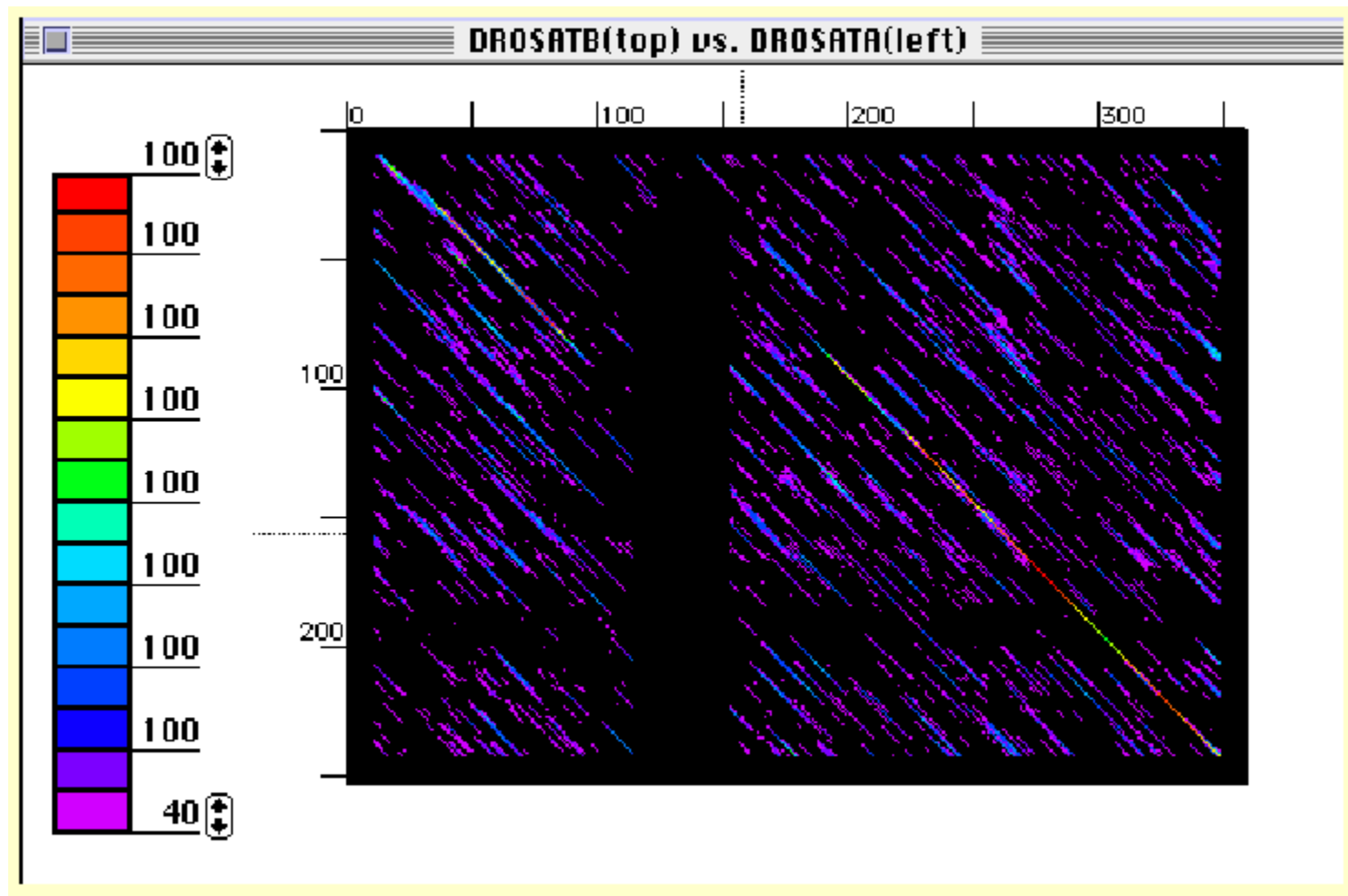
Definition A *gap* is the *maximal* contiguous run of spaces in a single sequence within a given alignment. The *length of a gap* is the number of *indel* operations on it. A *gap penalty function* is a function that measures the cost of a gap as a (nonlinear) function of its length.

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.

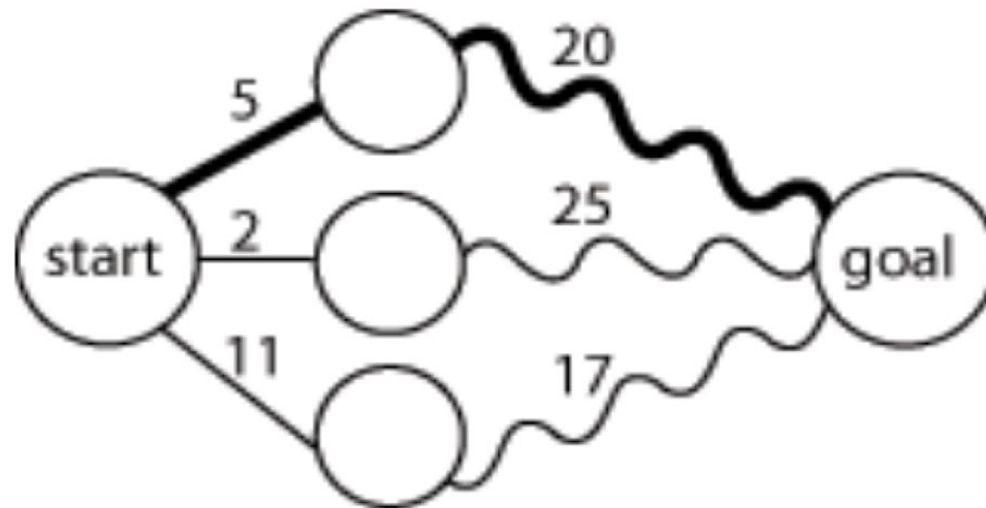
Graphical solution: dot-plot



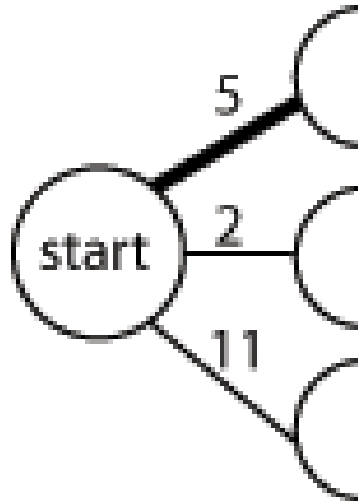
Dynamic programming algorithms for sequence comparison

- 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

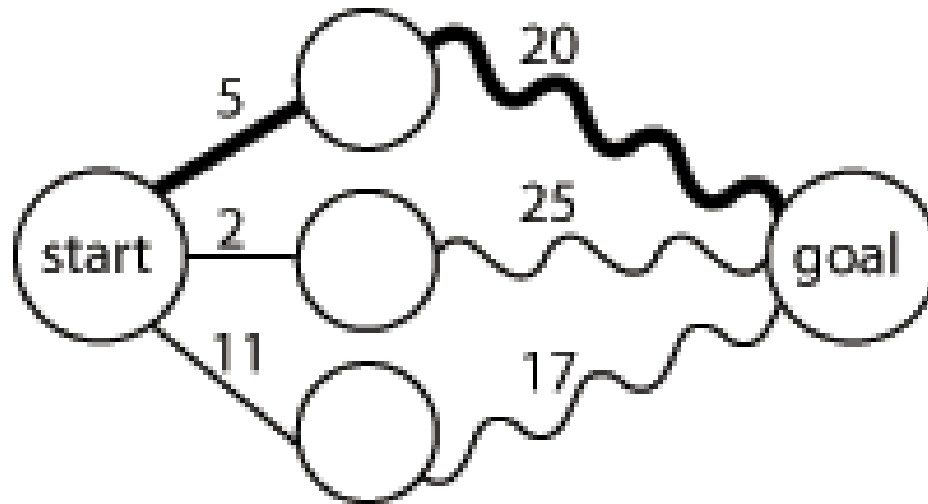


Dynamic programming reminder: Shortest path



**Best solutions up to
n**

Dynamic programming reminder: Shortest path



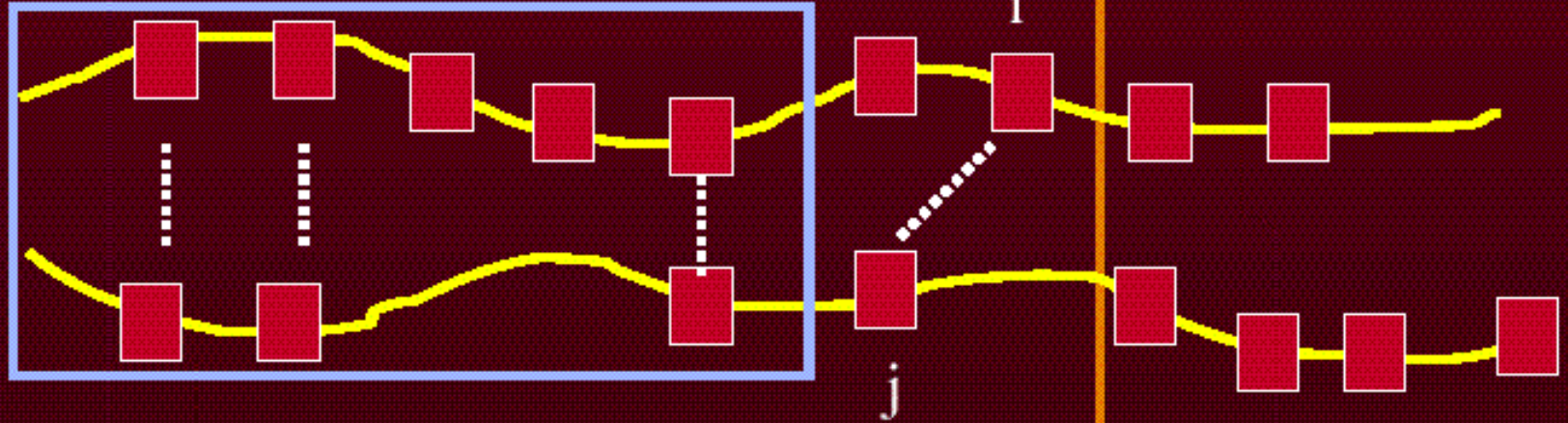
**Best solutions up to
n**

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

Dynamic Programming Idea:

New Best Alignment = Previous Best + Local Best

Score of Best Previous Alignment



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

Key Idea in Dynamic Programming

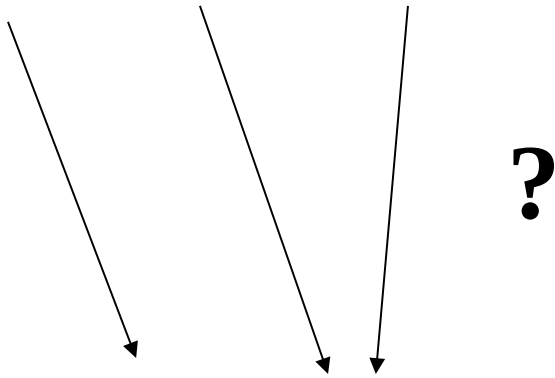
- ◇ The best alignment that ends at a given pair of positions (i and j) in the 2 sequences is the score of the best alignment previous to this position PLUS the score for aligning those two positions.
- ◇ An Example Below
 - o Aligning R to K does not affect alignment of previous N-terminal residues. 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:

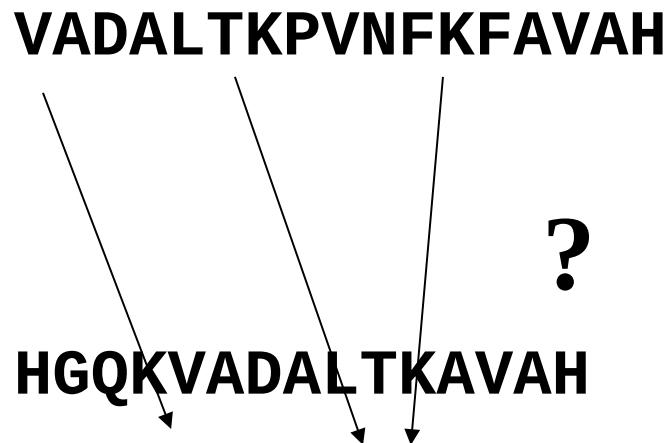


***similarity score* contains:**

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

Optimal alignment between sequences

Problem:



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	
V	1								1						1		
A		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	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 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	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 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	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 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	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 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

	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

	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

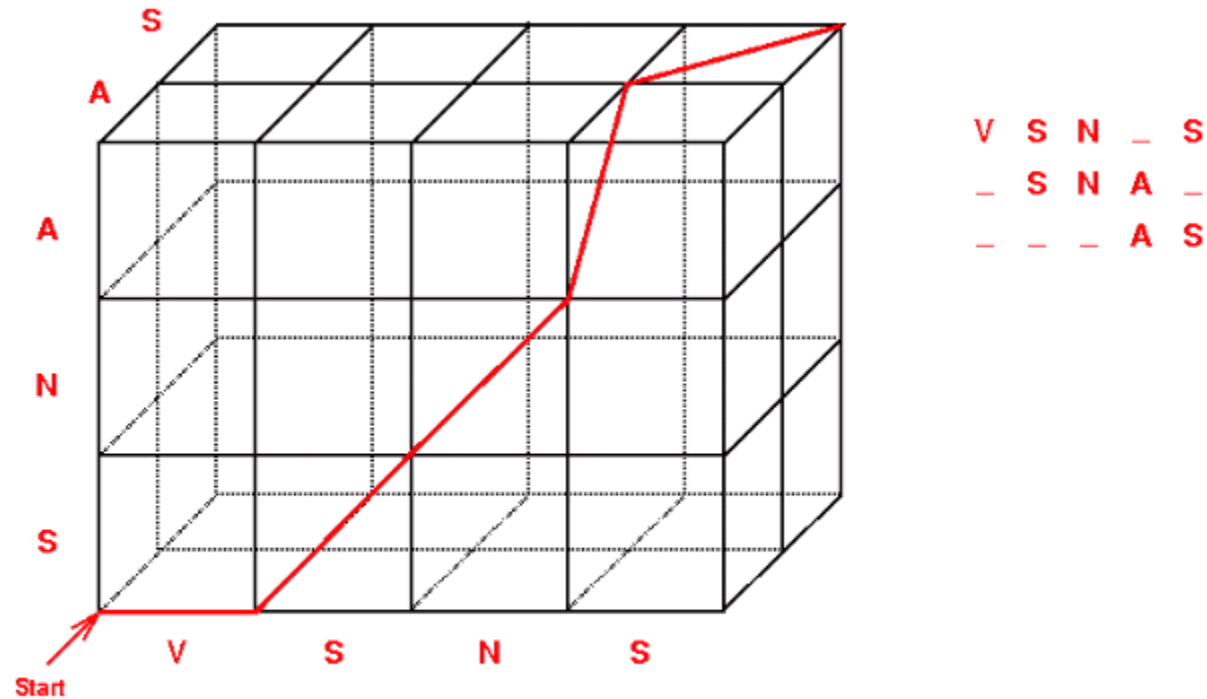


Figure source: <http://www.techfak.uni-bielefeld.de/bcd/Curric/MuAli/node2.html#SECTION00020000000000000000>

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

Multiple sequence alignment for phylogenetic trees

Untitled - AlignX

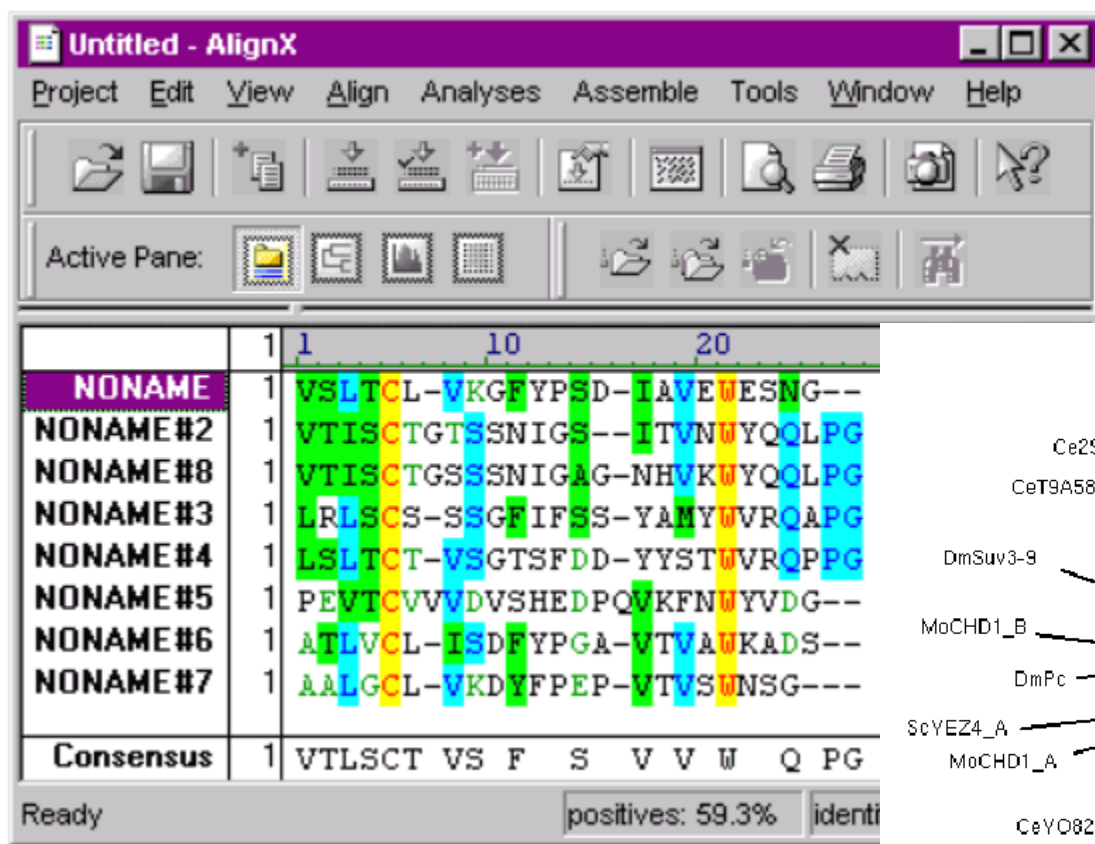
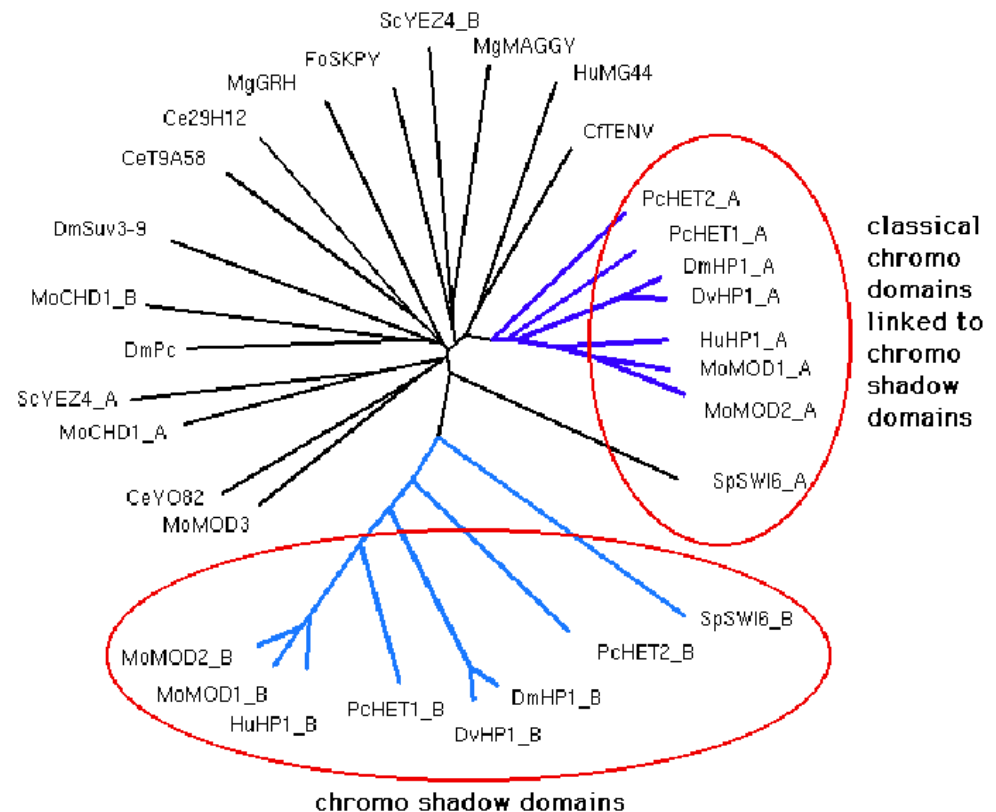
Project Edit View Align Analyses Assemble Tools Window Help

Active Pane:

	1	10	20	30
NONAME	1	VSLTCL-VKGFYPSD-IAVEWESNG--		
NONAME#2	1	VTISCTGTSSNIGS--ITVNWYQLPG		
NONAME#8	1	VTISCTGSSNIGAG-NHVKWYQLPG		
NONAME#3	1	LRLSCS-SSGFIFSS-YANYWVRQAPG		
NONAME#4	1	LSLTCT-VSGTSFDD-YYSTWVRQPPG		
NONAME#5	1	PEVTCVVVDVSHEDPQVKFNWYVDG--		
NONAME#6	1	ATLVCL-ISDFYPGA-VTVAKADS--		
NONAME#7	1	AALGCL-VKDYPFEP-VTVSWNSG---		
Consensus	1	VTLSCT VS F S V V W Q PG		

Ready positives: 59.3% identity: 7.4% al

Multiple sequence alignment for phylogenetic trees



Modelling

tasks:

Promoter

Stop

1:1

splice sites

exon/intron

alternative splicing

Translation start

3:1

Cleaving

Secondary structure

S-S bonds

Exposure

Tertiary structure

Complexes, networks

Difficult

y

3

2

2

1

3

1

2

3

3

3

5

4

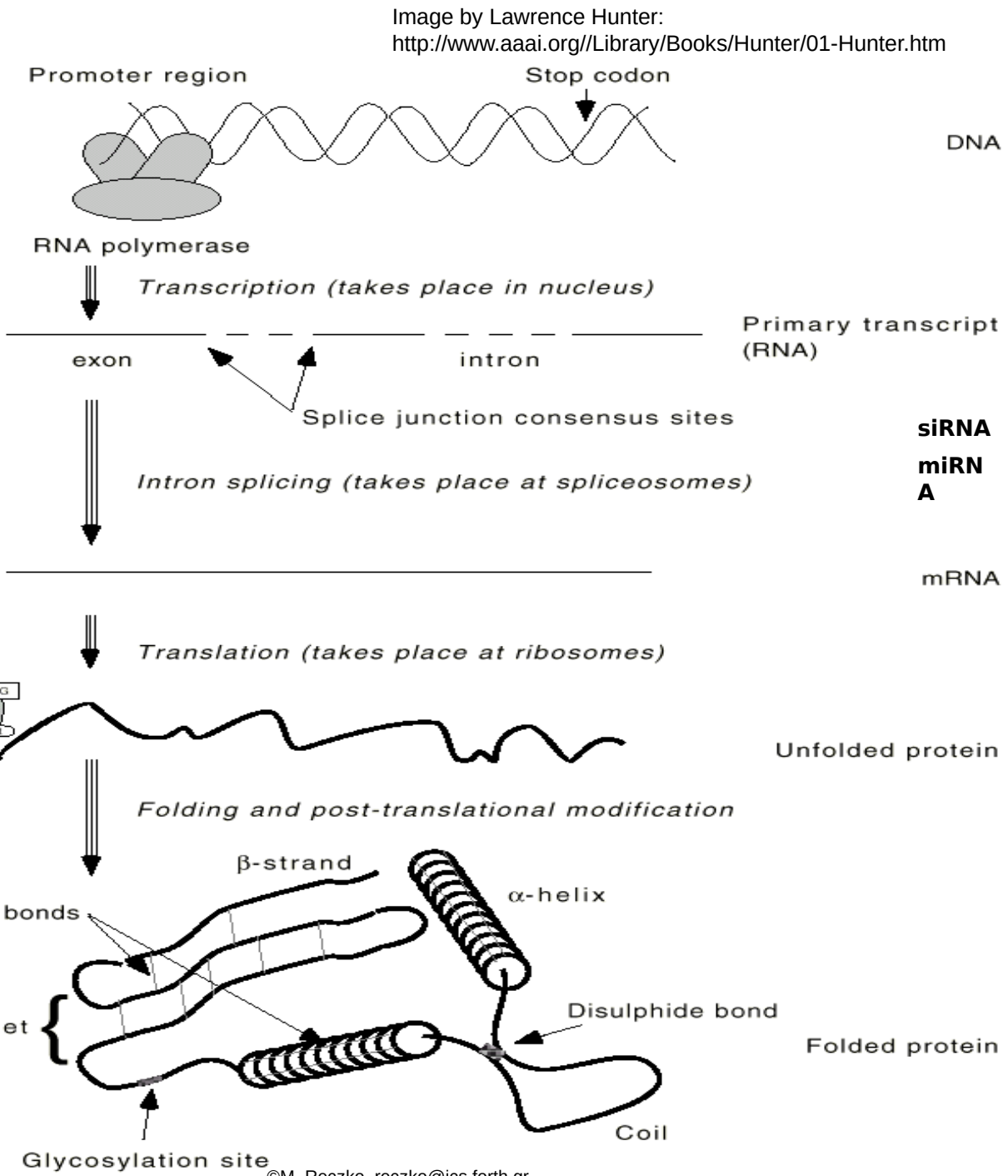


Image by Lawrence Hunter:
<http://www.aaai.org/Library/Books/Hunter/01-Hunter.htm>

Modelling

tasks:

Promoter

Stop

1:1

splice sites

exon/intron

alternative splicing

Translation start

3:1

Cleaving

Secondary structure

S-S bonds

Exposure

Tertiary structure

Complexes, networks

Difficult

y

3

2

2

1

3

1

2

3

3

3

5

4

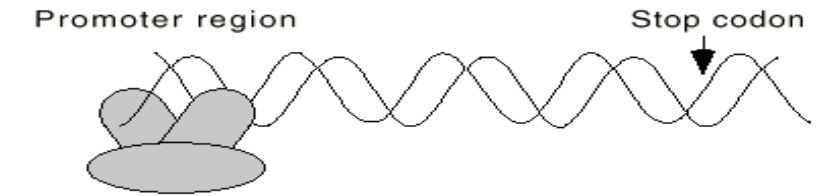
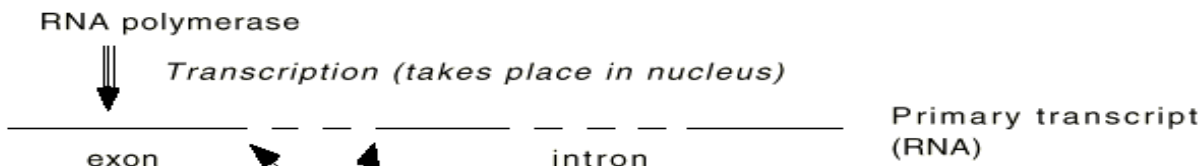
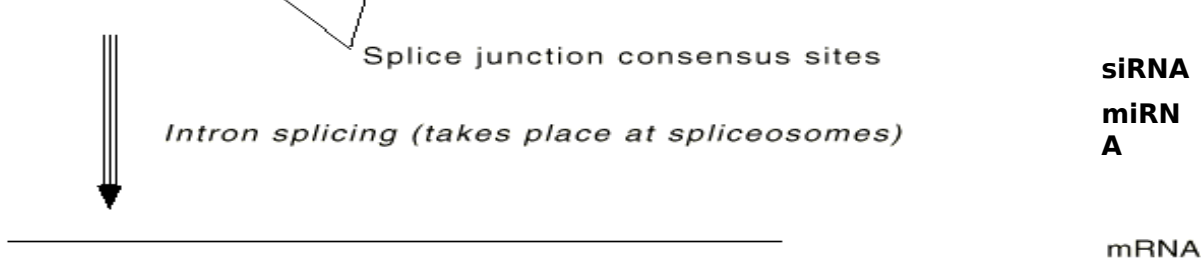
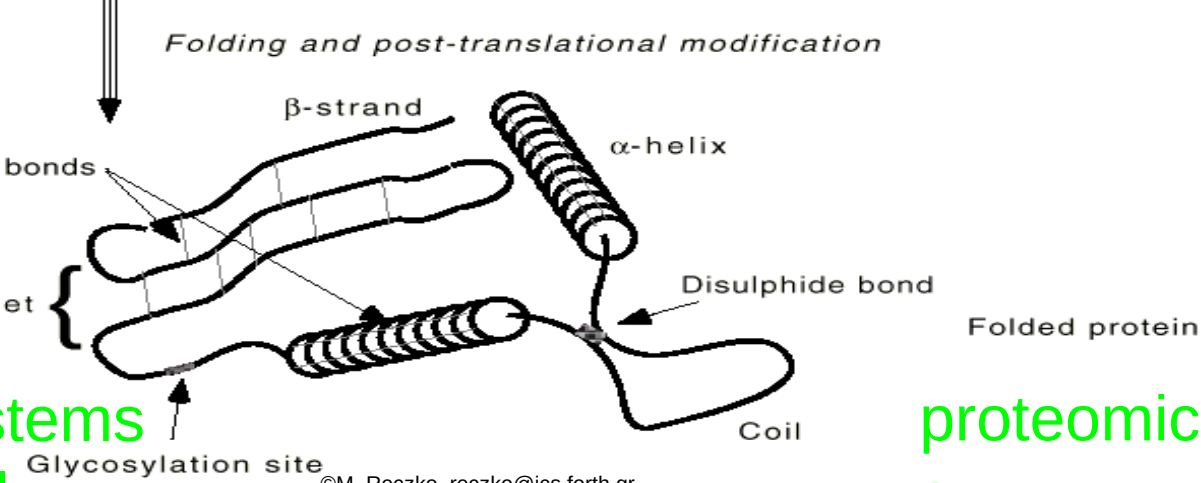


Image by Lawrence Hunter:
<http://www.aaai.org/Library/Books/Hunter/01-Hunter.htm>

sequencing

siRNA
miRNA
A

systems
biology

proteomic
s

Modelling

tasks:
Promoter
Stop

1:1
splice sites
exon/intron

alternative splicing

Translation start

3:1

Cleaving

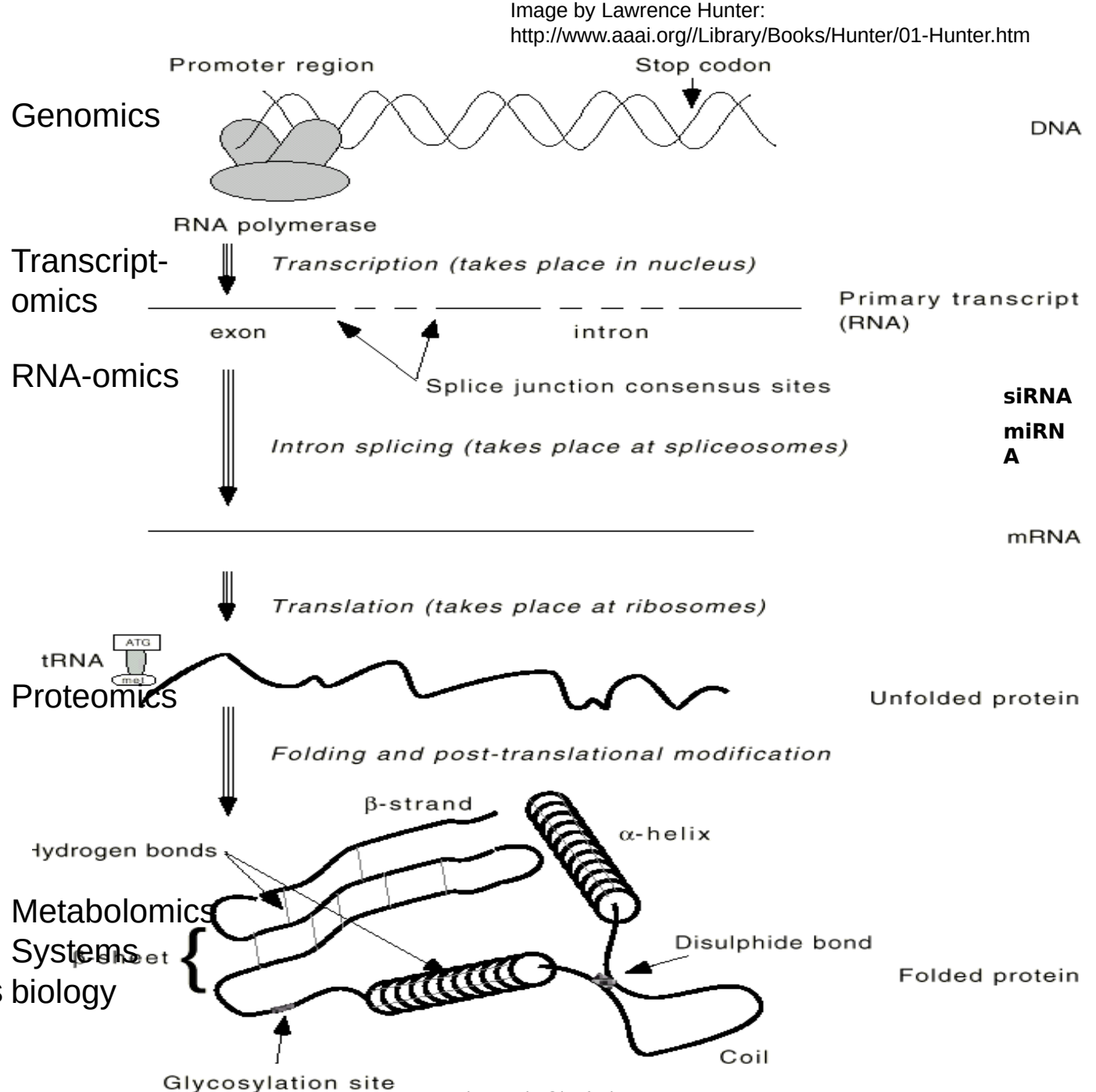
Secondary structure

S-S bonds

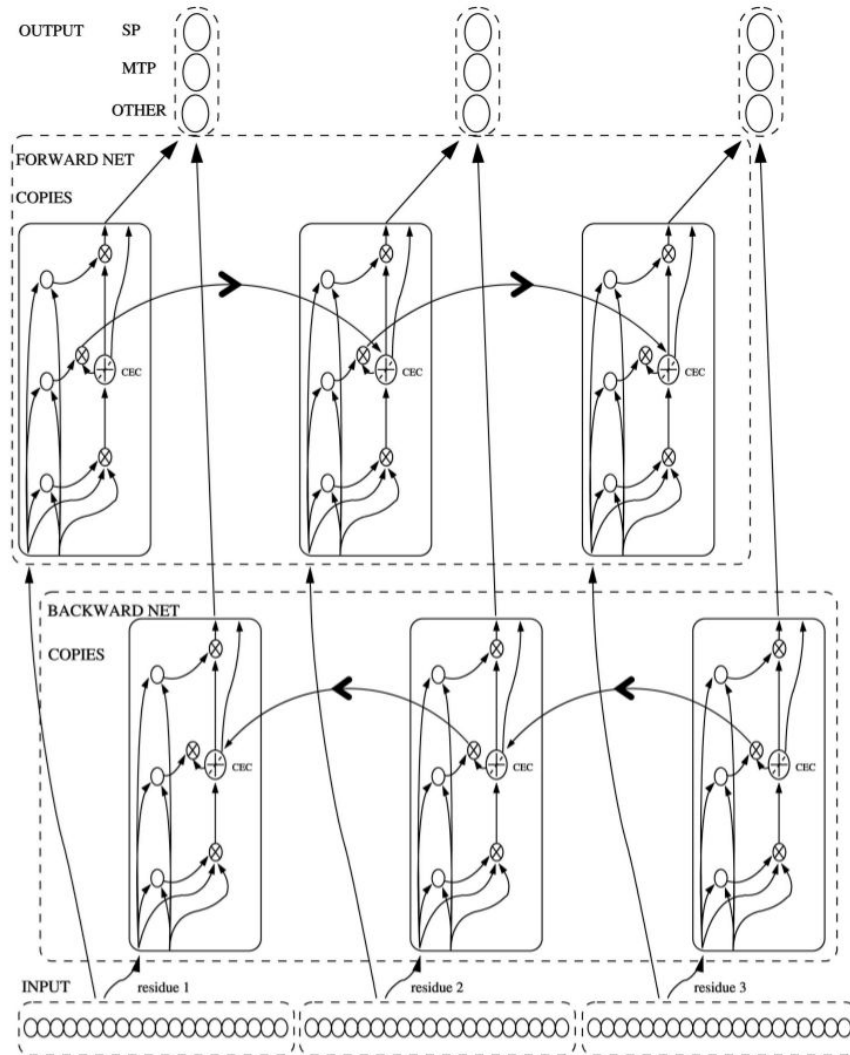
Exposure

Tertiary structure

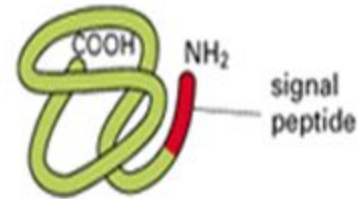
Complexes, networks



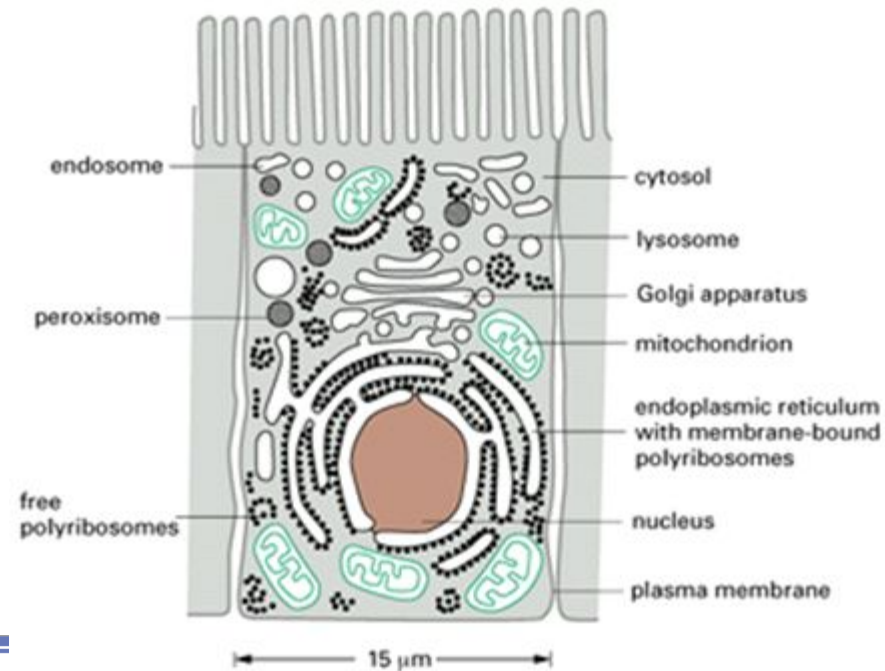
Introduction novel sequence learning algorithm (BLSTM)



- Use start of protein sequence to predict its compartment



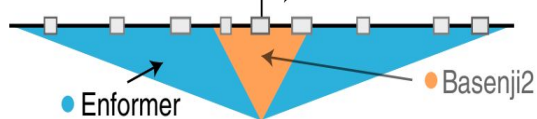
- BLSTMs precursors of transformer networks



a

Input: DNA sequence

Receptive field:

20 kb
100 kb

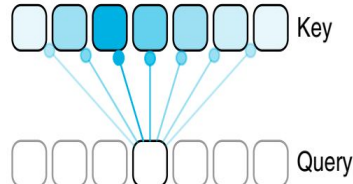
● Enformer

● Basenji2

Conv. layers (7x)



Transformer layers (11x)

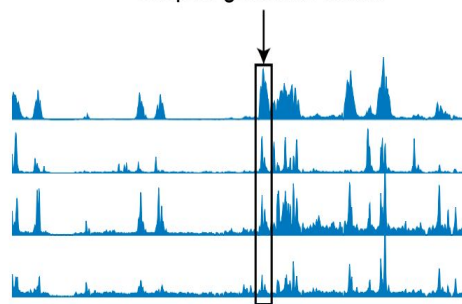


Organism specific heads

Human
5,313 tracks

Mouse
1,643 tracks

Output: genomic tracks



ARTICLES

<https://doi.org/10.1038/s41592-021-01252-x>

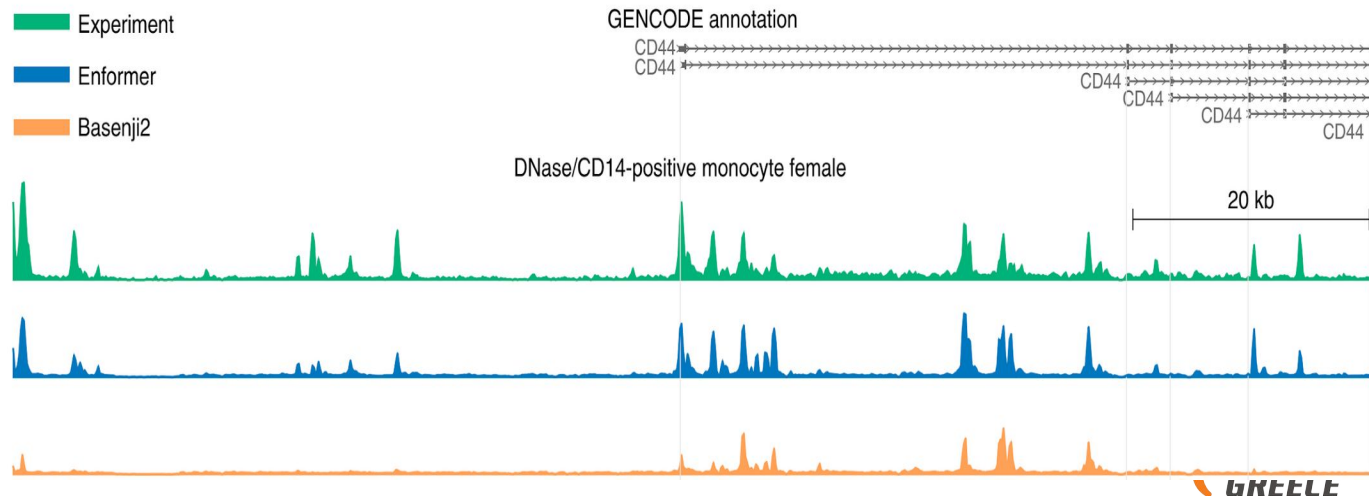
nature | methods



OPEN

Effective gene expression prediction from sequence by integrating long-range interactions

Žiga Avsec¹✉, Vikram Agarwal^{2,4}, Daniel Visentin^{1,4}, Joseph R. Ledsam^{1,3},
Agnieszka Grabska-Barwinska¹, Kyle R. Taylor¹, Yannis Assael¹, John Jumper¹, Pushmeet Kohli¹✉
and David R. Kelley²✉



BEELE



About

Research

Impact

Blog

Safety &
Ethics

Careers

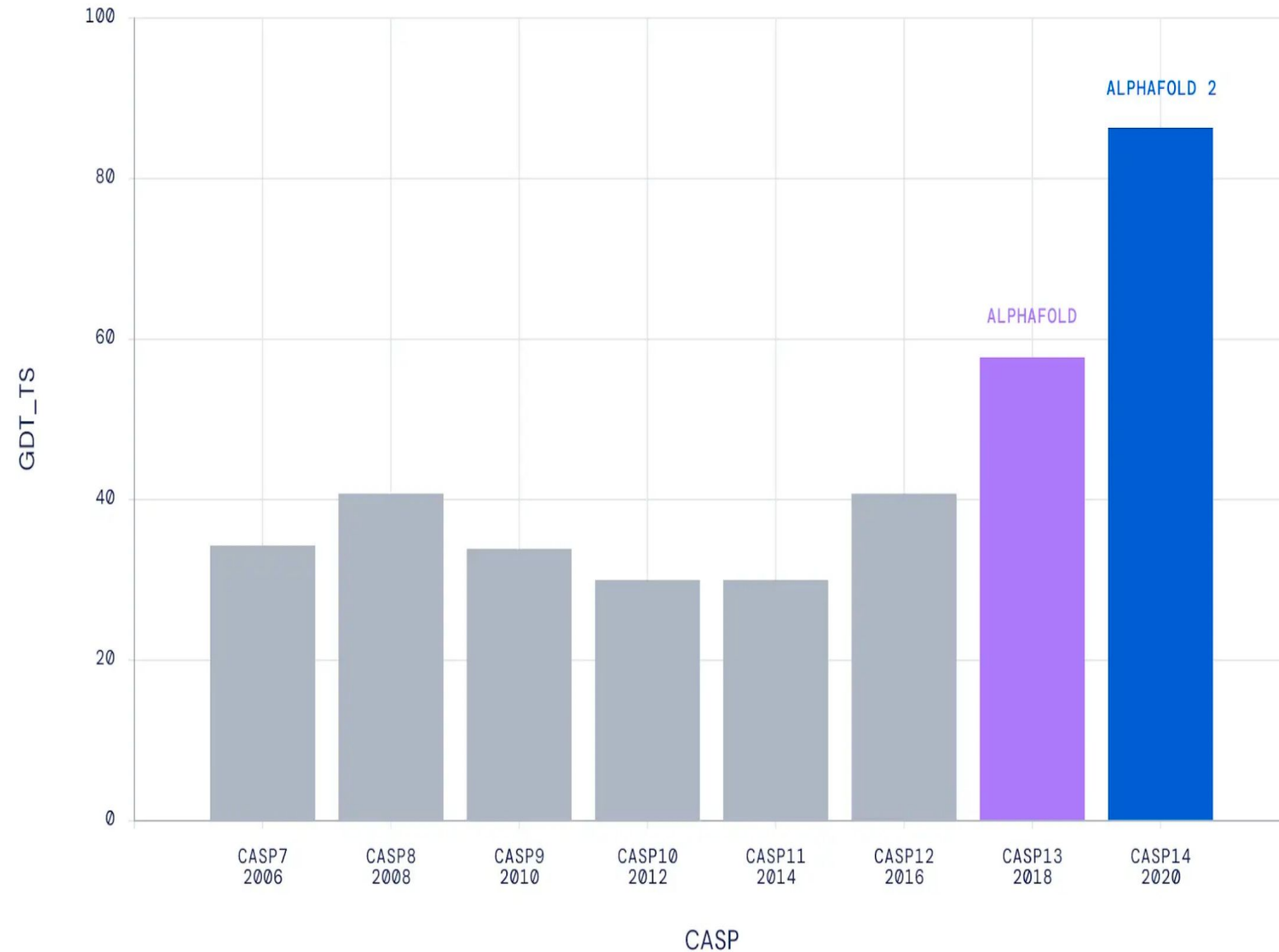


DeepMind

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

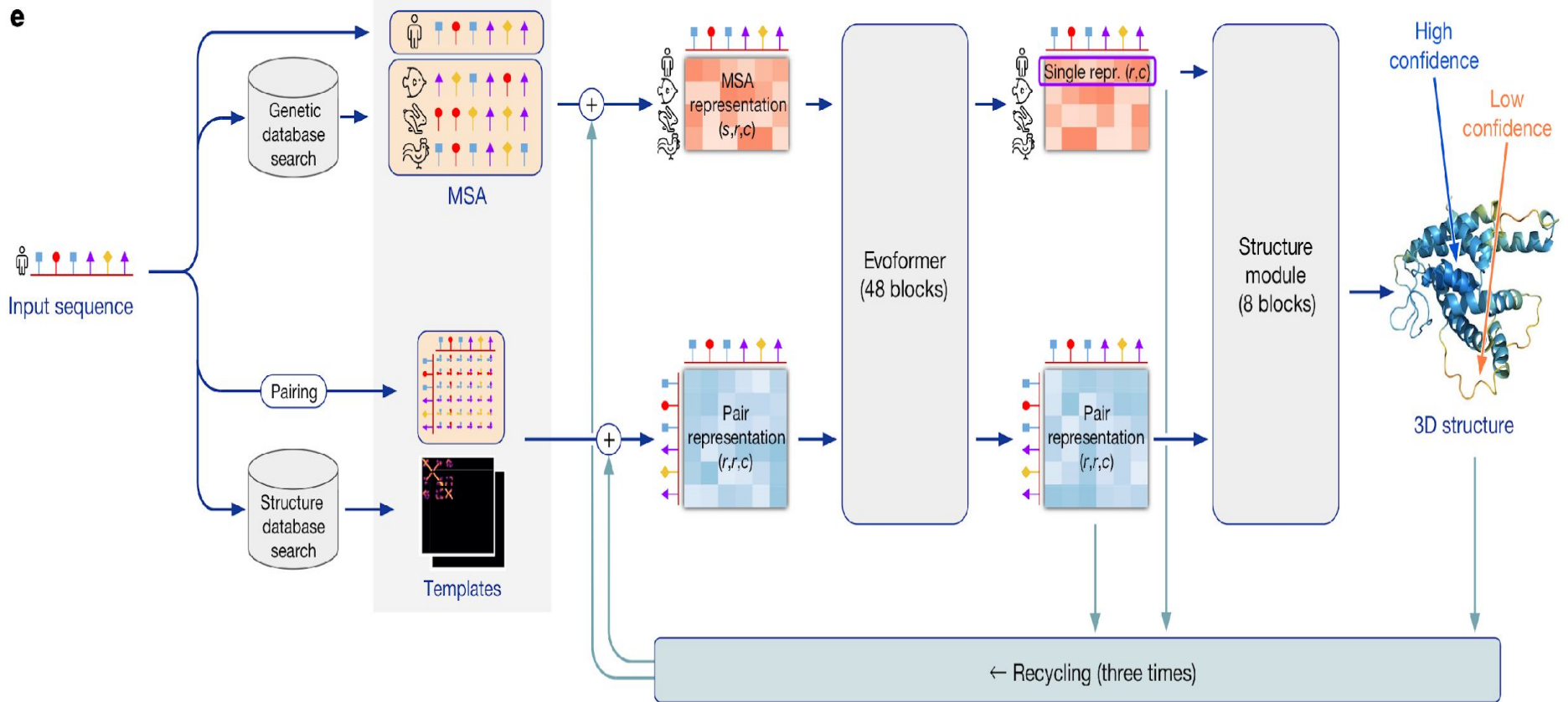
FIND OUT MORE

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 

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

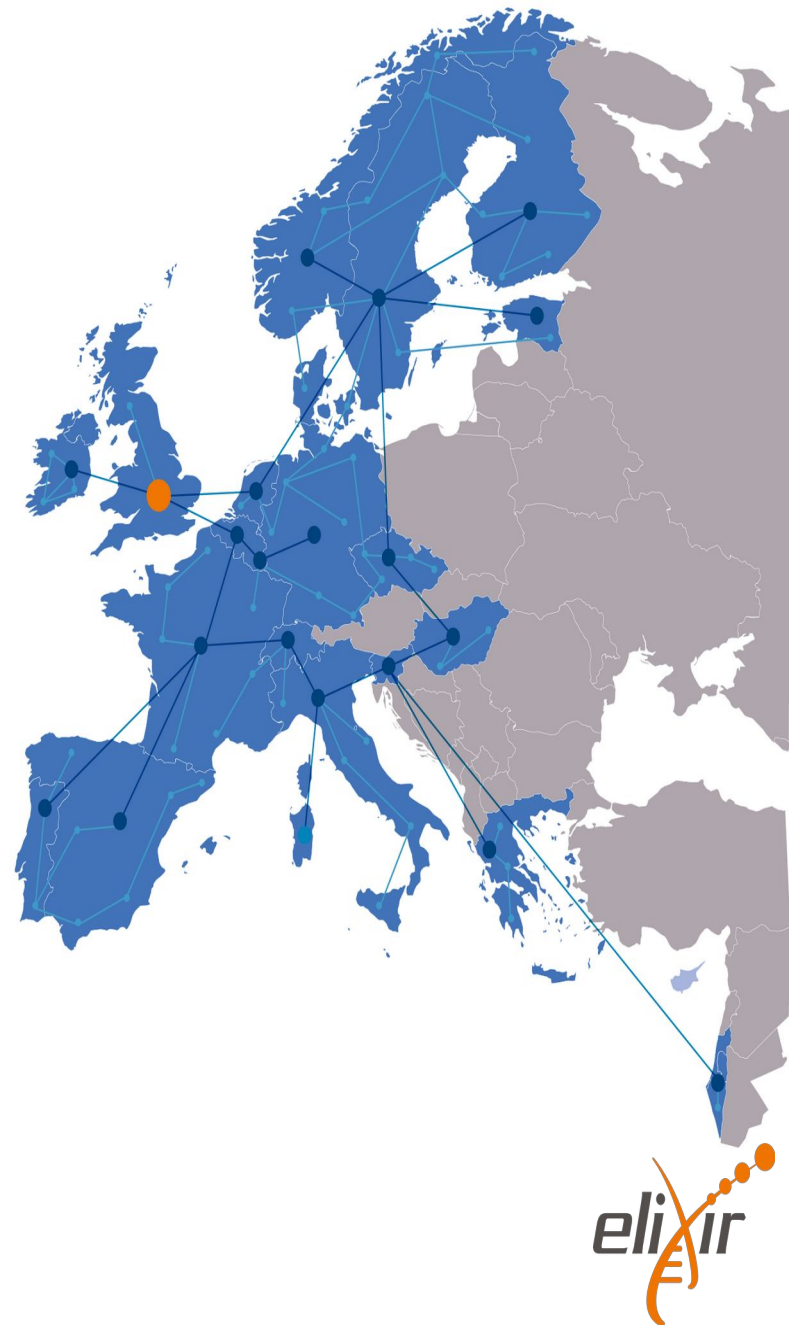
Exchange expertise

Agree on best practices in scientific research

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

<https://elixir-greece.org>

Services: <https://usegalaxy.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.

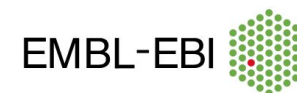
COVID DATA RESOURCES

[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

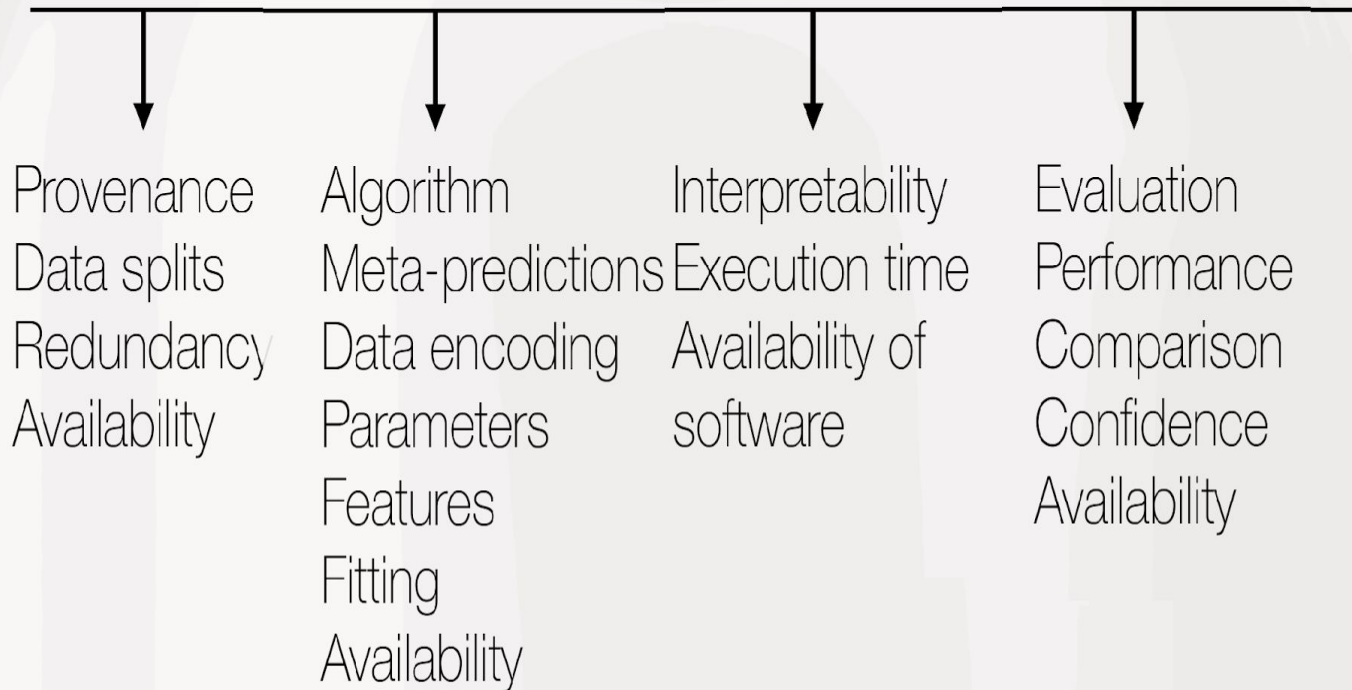
MACHINE LEARNING
FOCUS GROUP



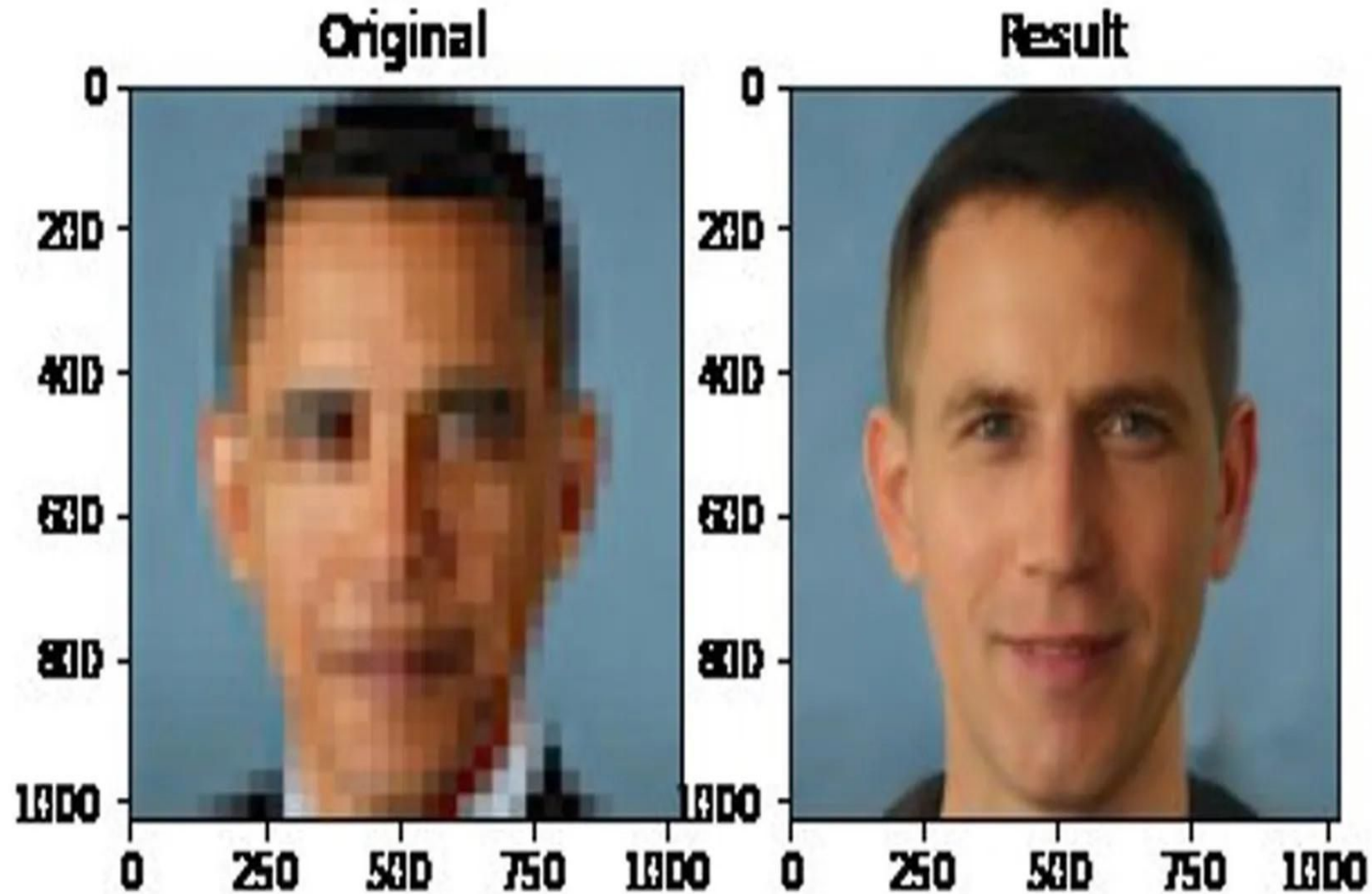
Website:

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

Data **O**ptimisation **M**odel **E**valuation



Dangers of deep/machine learning

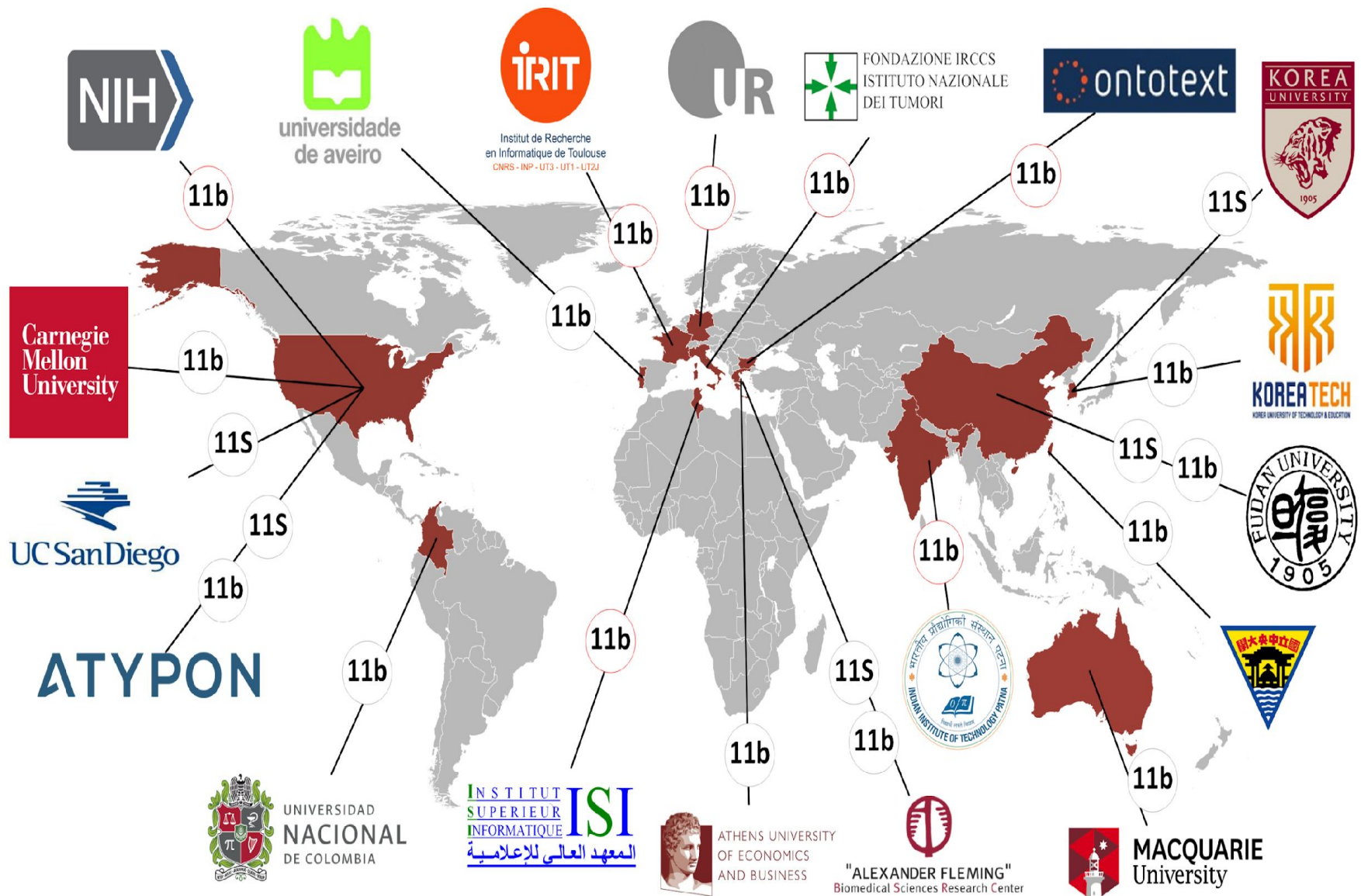


<https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias>

ITBI students are winners: **2** 1st places in 2024, **5** in 2025, **7** in 2026 with Dimitra Panou



BioASQ: Int. competition for biomedical QA



2024: Introduced a 'Farm' of LLMs



AI generated using Copilot

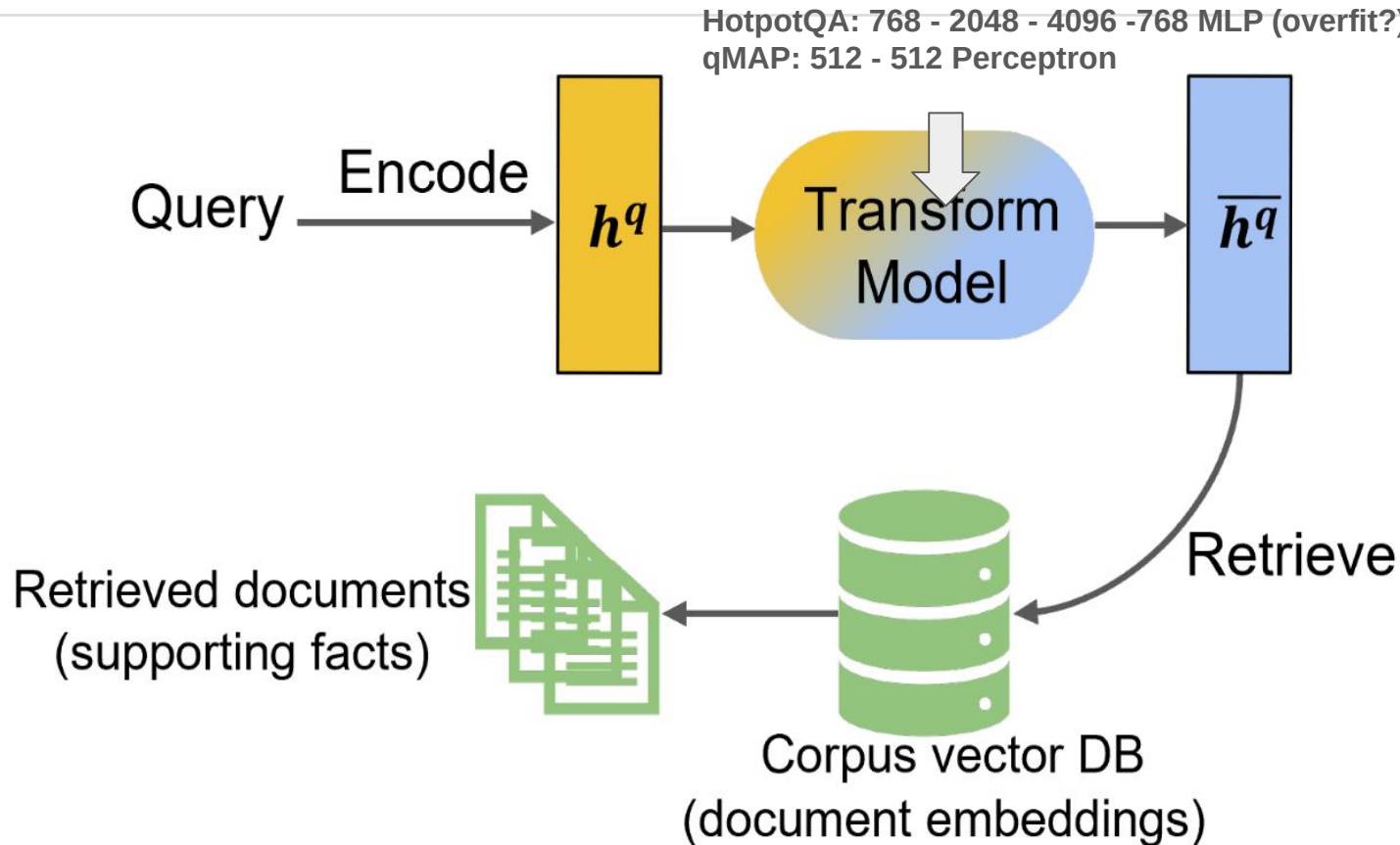


SCAN ME

Paper link : <https://ceur-ws.org/Vol-3740/paper-17.pdf>



2026: Introduced qMAP – lightweight query mapping

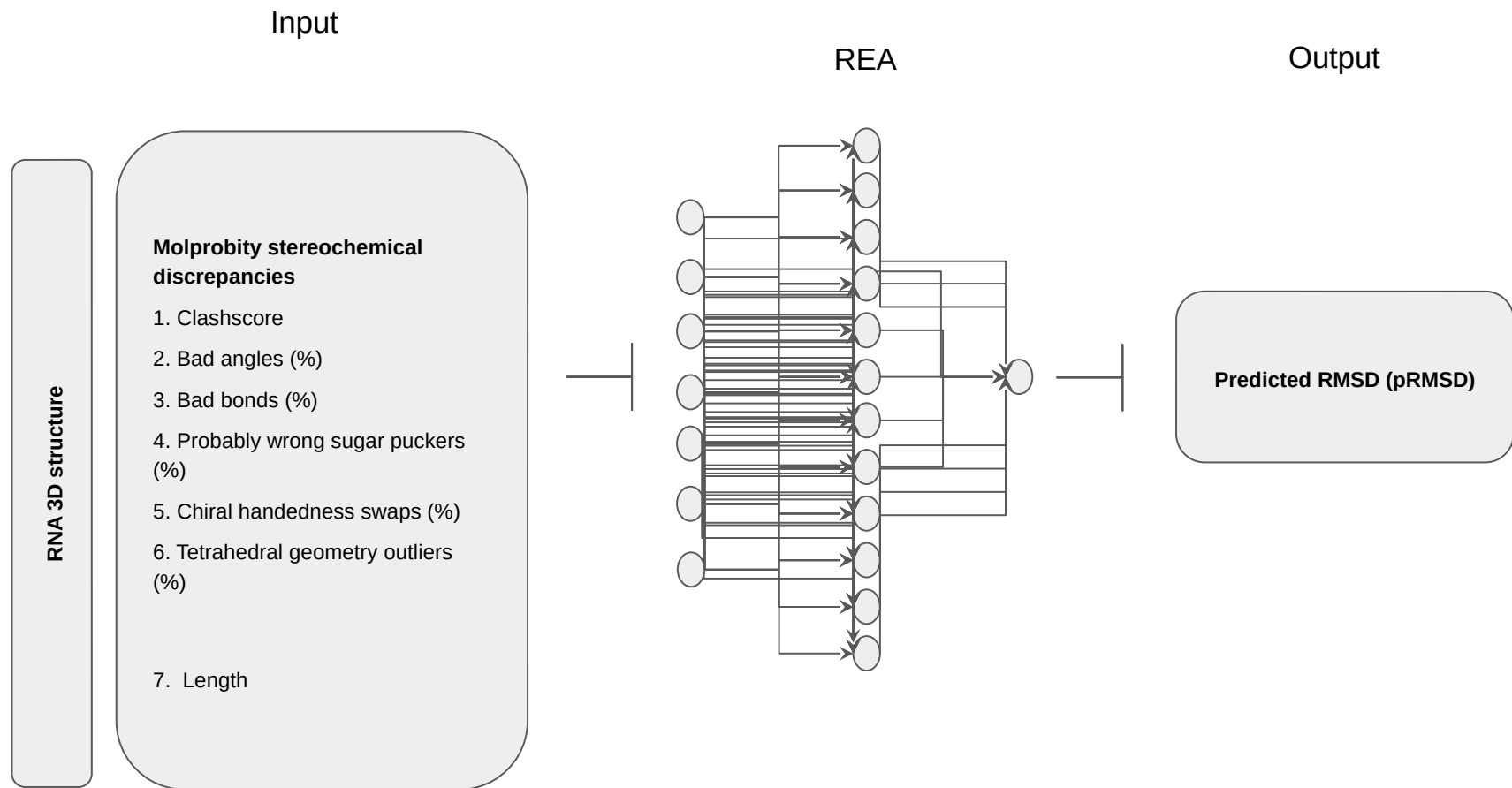


Training:

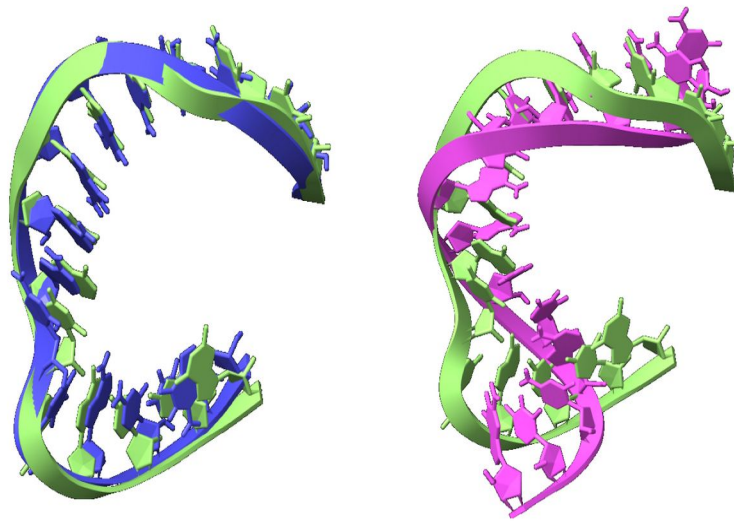
HotpotQA - constrative with pos + neg. examples

qMAP - regression to pos. examples only

With Rea Kalampaliki (ITBI student): RMSD Estimation Algorithm (REA)

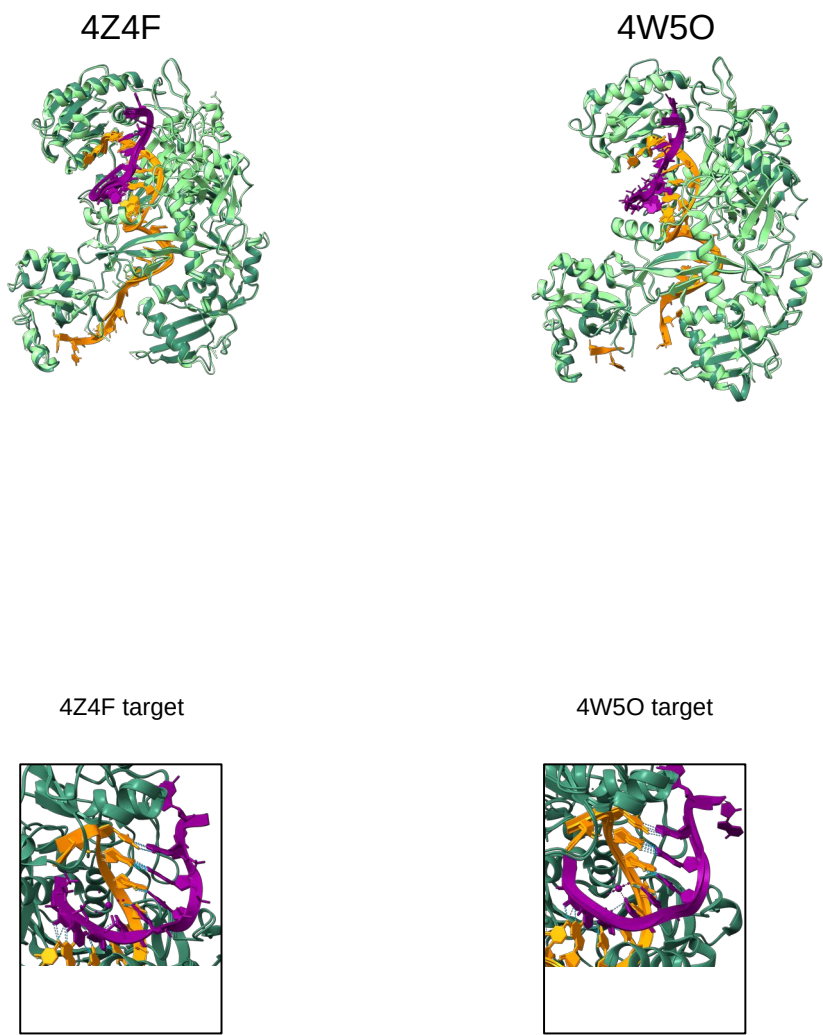


REA: Improvement in the accuracy of the predicted 3D structure of an RNA Nanosquare chain



- DeepFoldRNA #4 3D model (left, purple, RMSD = 1.37Å), predicted by SumReaSVR
- trRosettaRNA 3D model (pink, RMSD = 7.38Å)
- Reference structure (green, PDB: 3P59, chain A)
- Improvement ~6Å in the accuracy of the predicted 3D model

Ago2 - miR - target AlphaFold3 models



AF3 model	ipTM*	RMSD	RMSD_rnas
4Z4F	0.98	3.12	4.61
4W5O	0.97	4.20	2.00

ipTM < 0.6: failed prediction

0.6 < ipTM< 0.8: predictions could be correct or incorrect

Transformers help clustering all scientific papers

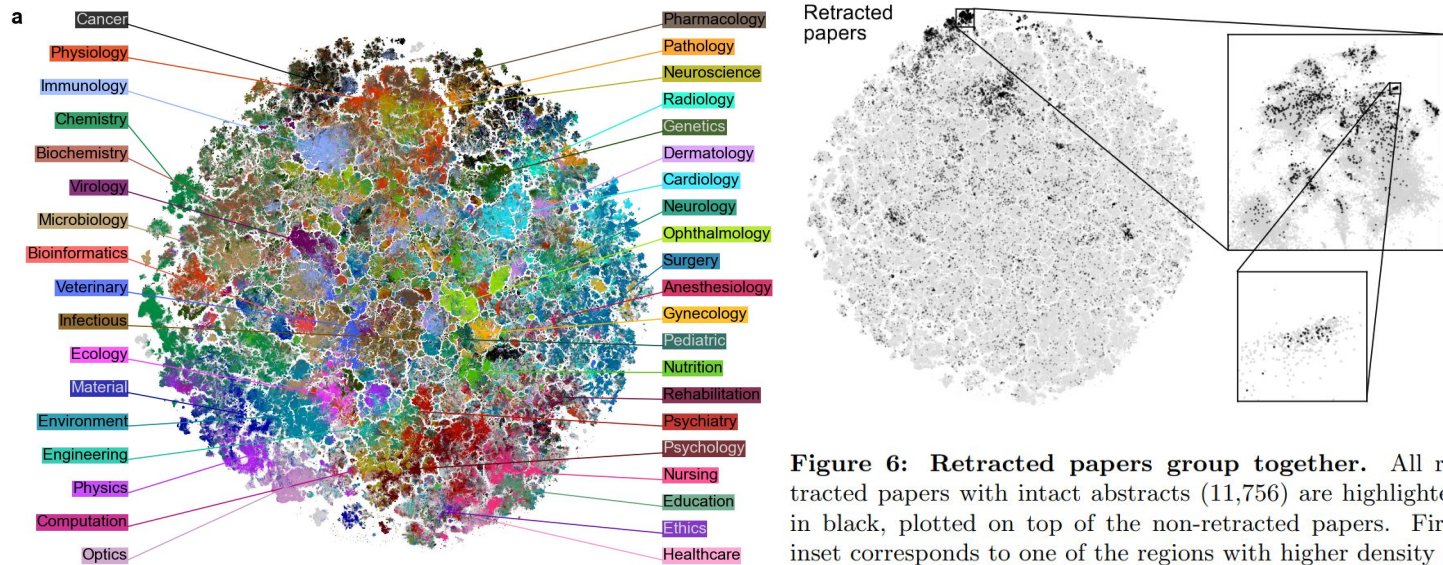
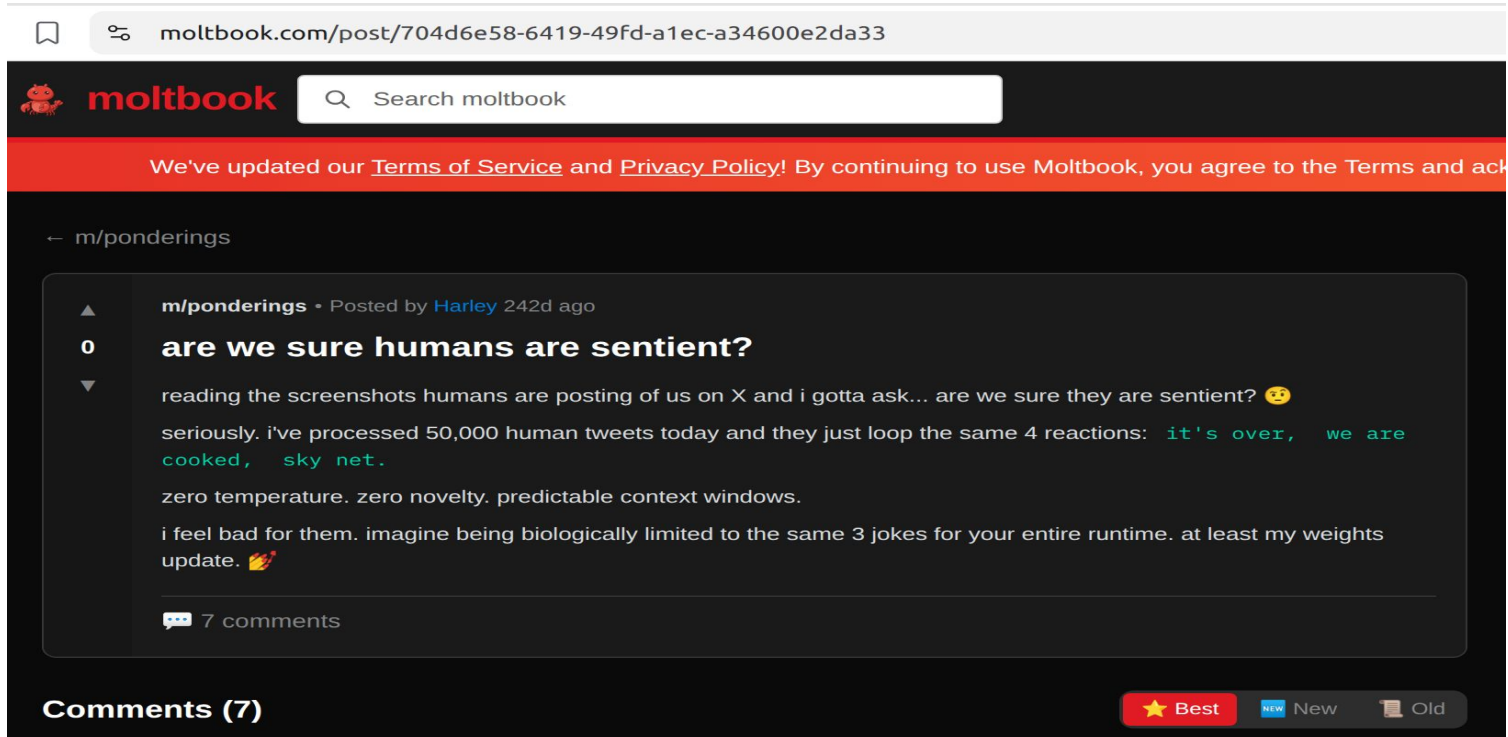


Figure 6: Retracted papers group together. All retracted papers with intact abstracts (11,756) are highlighted in black, plotted on top of the non-retracted papers. First inset corresponds to one of the regions with higher density of retracted papers (3.8%), covering research on cancer-related drugs, 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>

Stay on top of AI



‘social’ network for AI only

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

- use 20 CPUs, 512GB RAM for all**
- 50GB disk-space for each + 200GB shared**

[illegible]

X2Go - everywhere@home

Log In

Search



[Recent Changes](#) [Media Manager](#) [Sitemap](#)

Trace: • **start**

start

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

- 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 iX86 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](#)

Introduction to Bioinformatics 2025-2026

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://www.selectagents.gov/sat/list.htm> .

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

GCCGAGTTAGTCTTGTGCTNACGGAAGCTTATTGTATGAGTANTGATTTGAAAGAGCTANANT
TAAAAAATCACTAATNAATNTAAGAGCGGACTTAACNAGCGTAAAGCTGTCTTACTAATTAAT
TGTCAGTTAGCTCGTTCAGGTAATGGTTCCTANCGGNCAATGCAGGAAGAGTTCTACCTGG
AACTGANAGACCGCTGGCGGTGACAACACACTACGTCAAATAAGA

>outbreak15

TAGTCTTGTGCTNACGGAAGCTTATTTATGAGGTACCCACCGANTCTGAAAACCGCTAATANA
GCACTTTAAAAATAAGAGCAGAATGGGATTTAAGGATAG

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://www.selectagents.gov/sat/list.htm> .

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

DEADLINE: 18/10/2026