Introduction to Bioinformatics (ITBI)

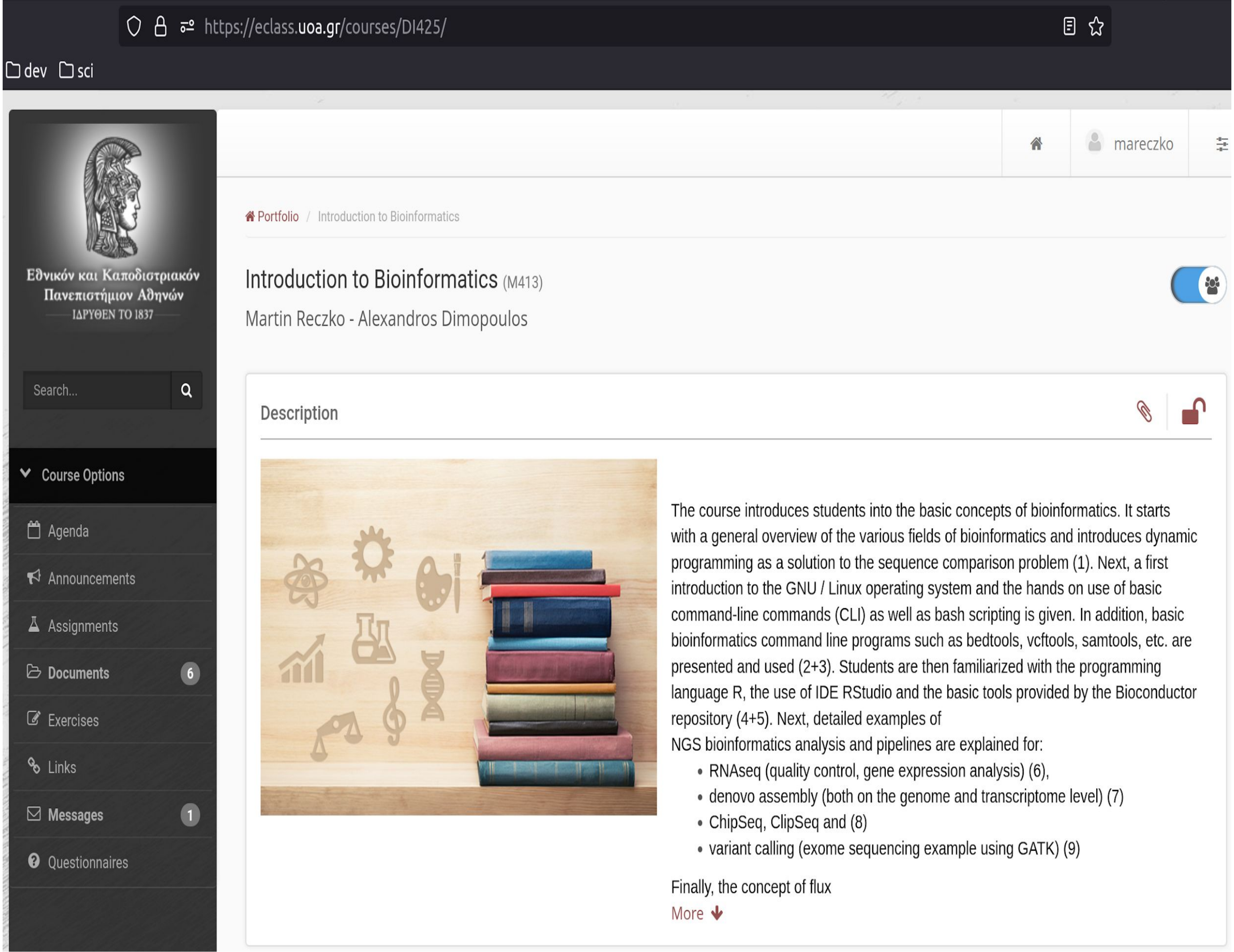
Martin Reczko[*] ***+ Alexandros Dimopoulos***

[*]*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



Introduction to Bioinformatics (M413)

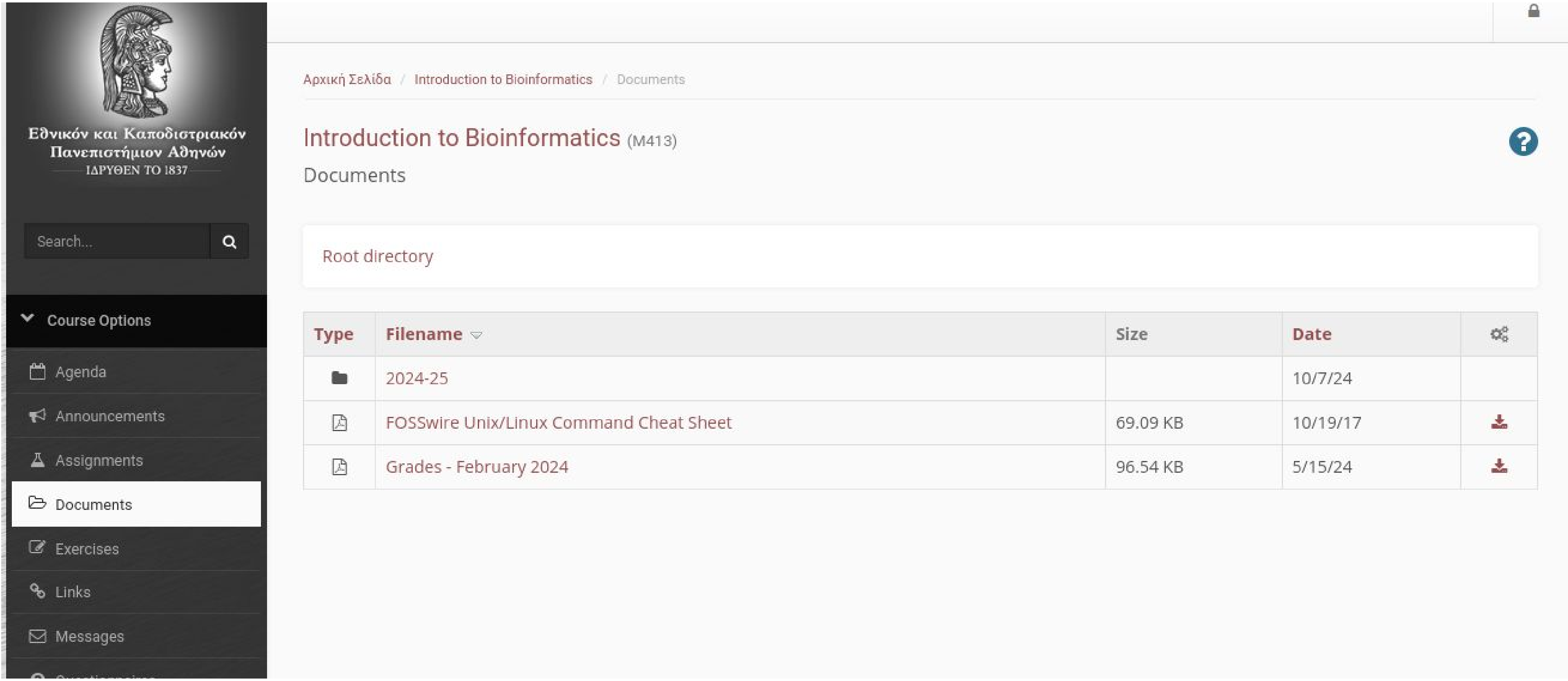
Martin Reczko - Alexandros Dimopoulos

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



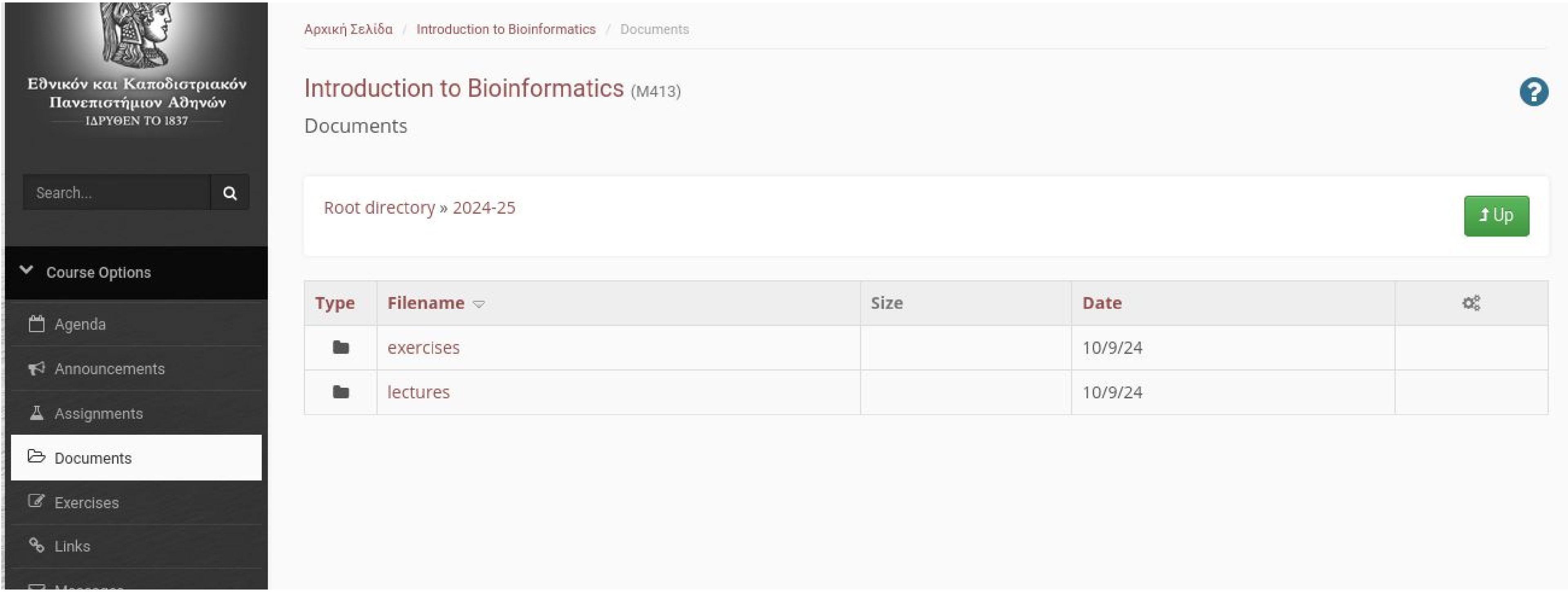
Αρχική Σελίδα / Introduction to Bioinformatics / Documents

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	⬇



Αρχική Σελίδα / Introduction to Bioinformatics / Documents

Introduction to Bioinformatics (M413)

Documents

Root directory » 2024-25

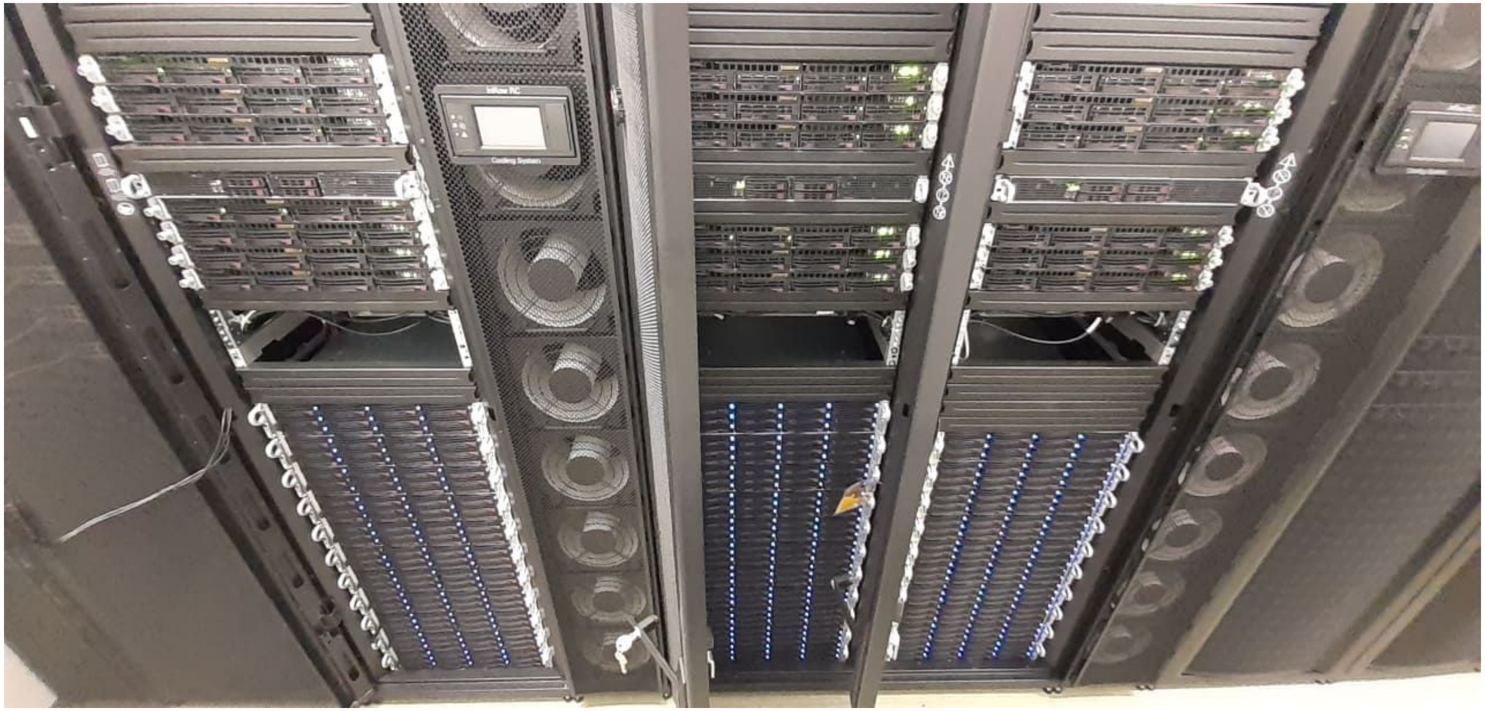
Type	Filename	Size	Date	⚙
📁	exercises		10/9/24	
📁	lectures		10/9/24	

Please verify name+email in participant list at

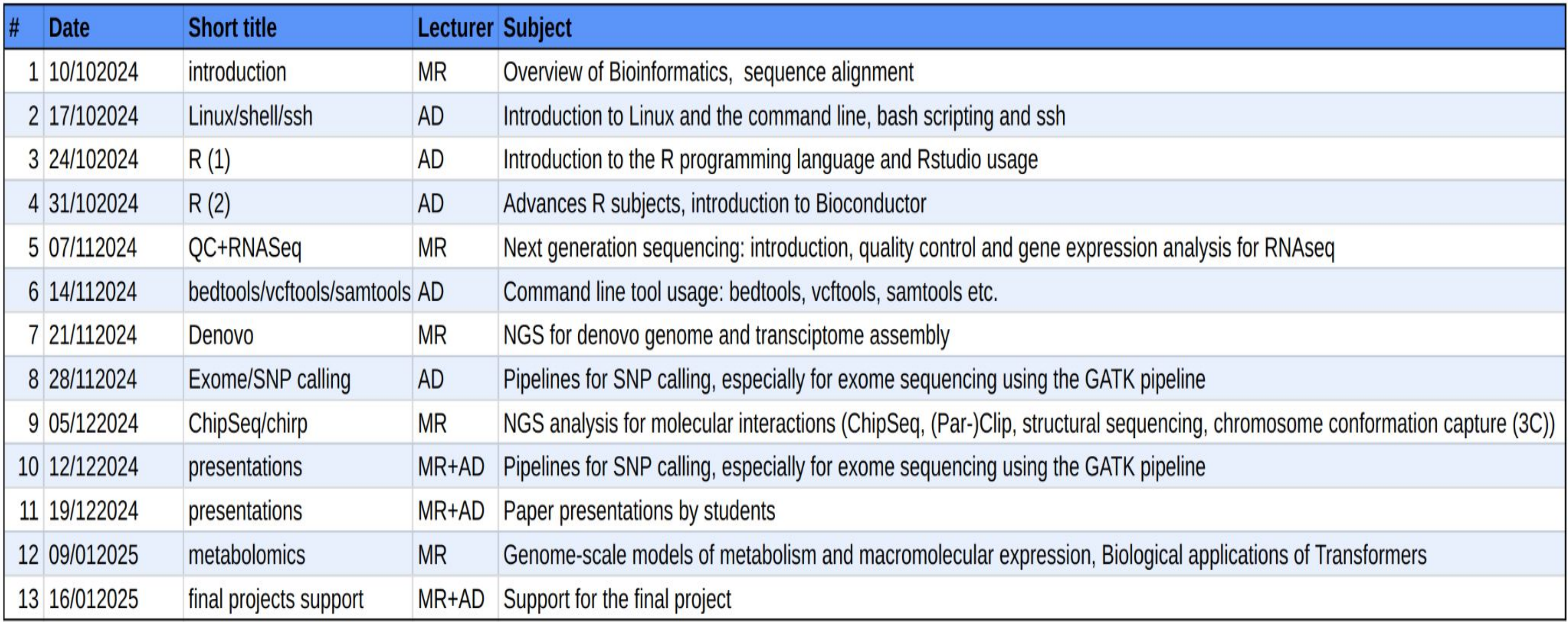
https://tinyurl.com/suzfj6y4

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 f



Syllabus and grading



#	Date	Short title	Lecturer	Subject
1	10/102024	introduction	MR	Overview of Bioinformatics, sequence alignment
2	17/102024	Linux/shell/ssh	AD	Introduction to Linux and the command line, bash scripting and ssh
3	24/102024	R (1)	AD	Introduction to the R programming language and Rstudio usage
4	31/102024	R (2)	AD	Advances R subjects, introduction to Bioconductor
5	07/112024	QC+RNASeq	MR	Next generation sequencing: introduction, quality control and gene expression analysis for RNAseq
6	14/112024	bedtools/vcftools/samtools	AD	Command line tool usage: bedtools, vcftools, samtools etc.
7	21/112024	Denovo	MR	NGS for denovo genome and transciptome assembly
8	28/112024	Exome/SNP calling	AD	Pipelines for SNP calling, especially for exome sequencing using the GATK pipeline
9	05/122024	ChipSeq/chirp	MR	NGS analysis for molecular interactions (ChipSeq, (Par-)Clip, structural sequencing, chromosome conformation capture (3C))
10	12/122024	presentations	MR+AD	Pipelines for SNP calling, especially for exome sequencing using the GATK pipeline
11	19/122024	presentations	MR+AD	Paper presentations by students
12	09/012025	metabolomics	MR	Genome-scale models of metabolism and macromolecular expression, Biological applications of Transformers
13	16/012025	final projects support	MR+AD	Support for the final project

Grade	**100%**	
Presentation		30%
Exercises		20%

Subjects:

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

An Eukaryotic Cell (biological

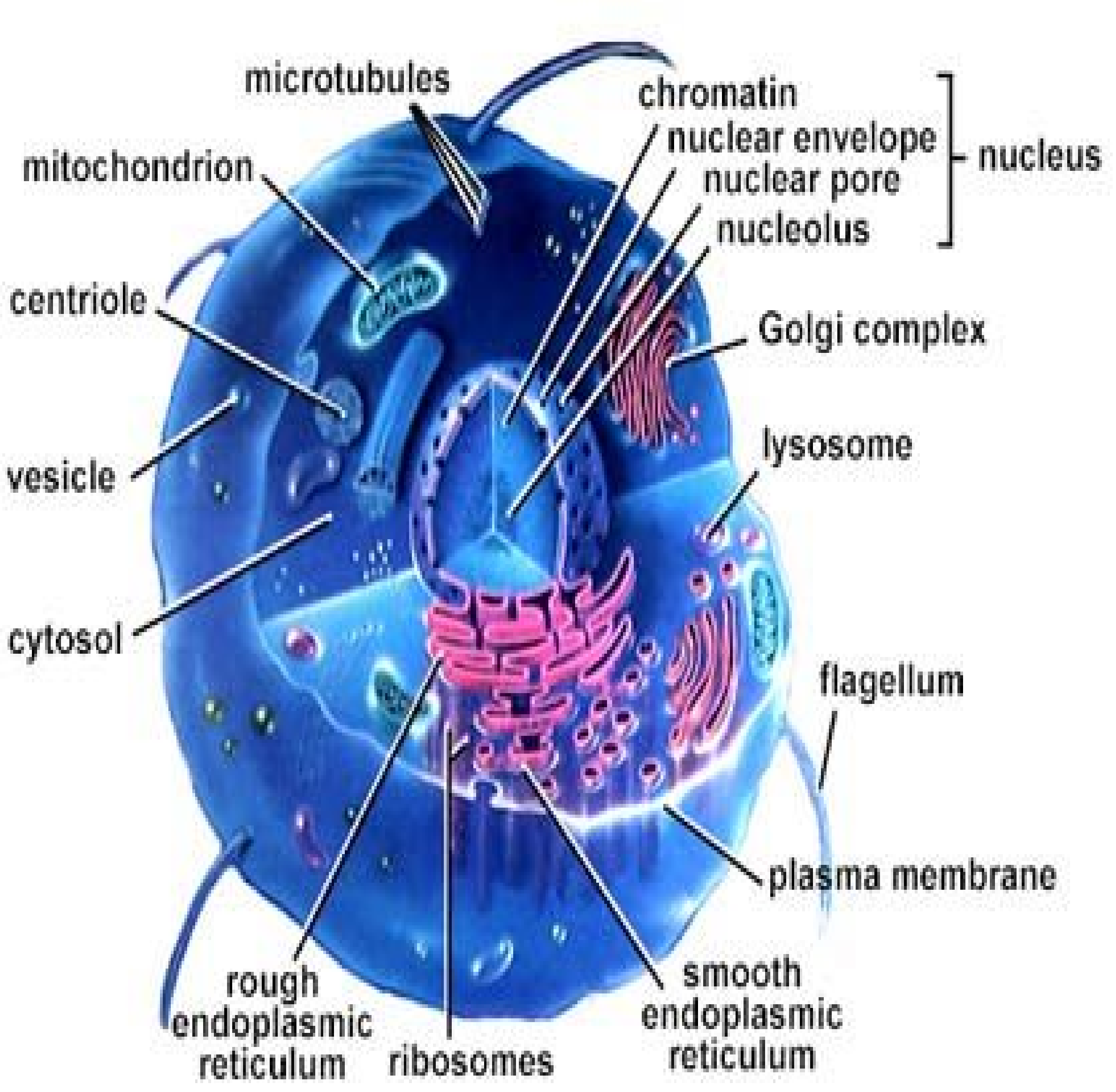
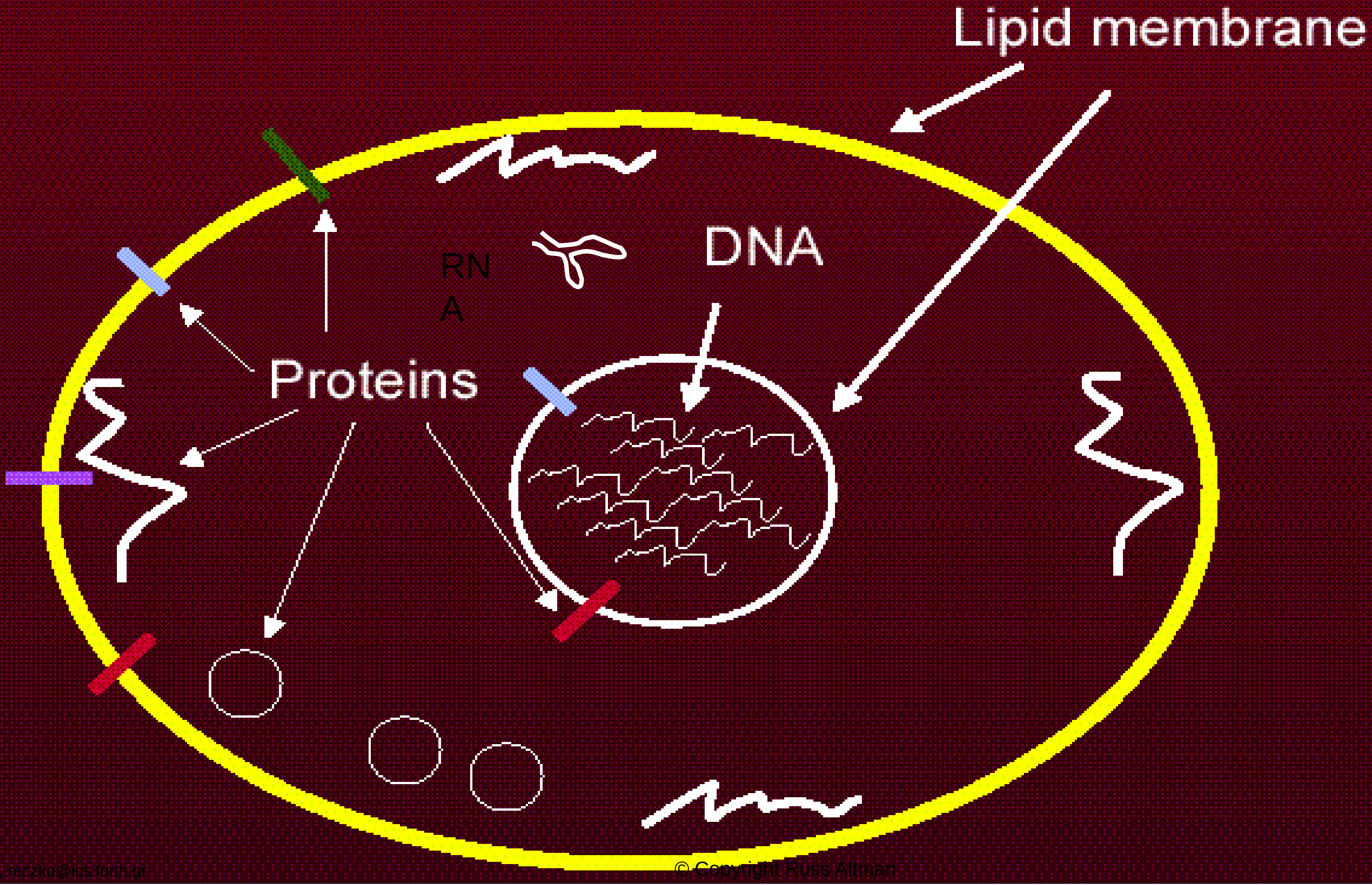
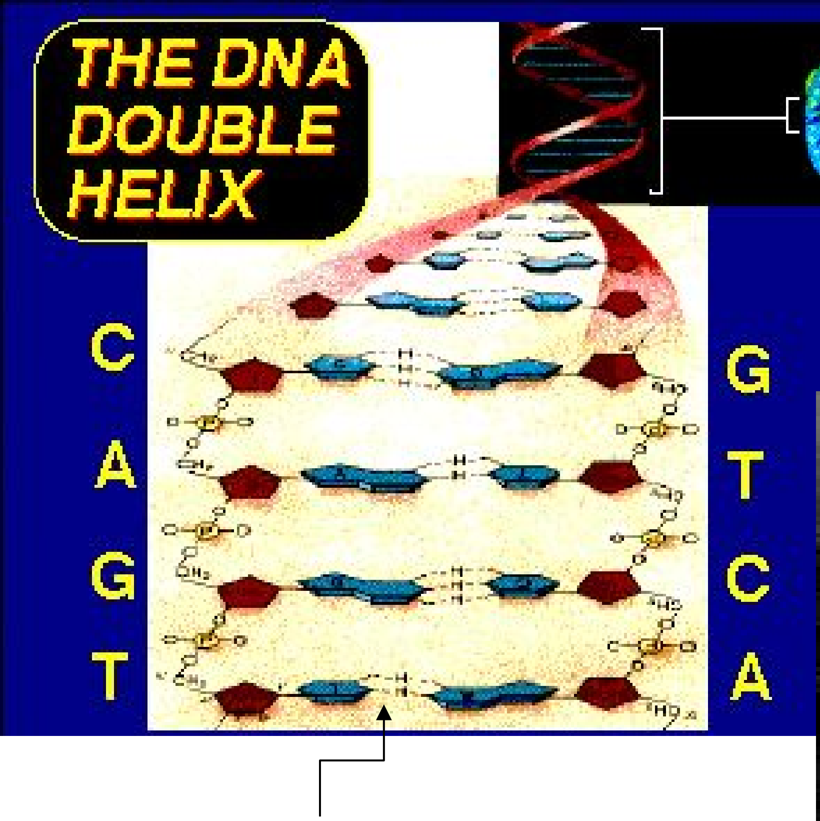


Image from: On-Line Biology Book
http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookTOC.html

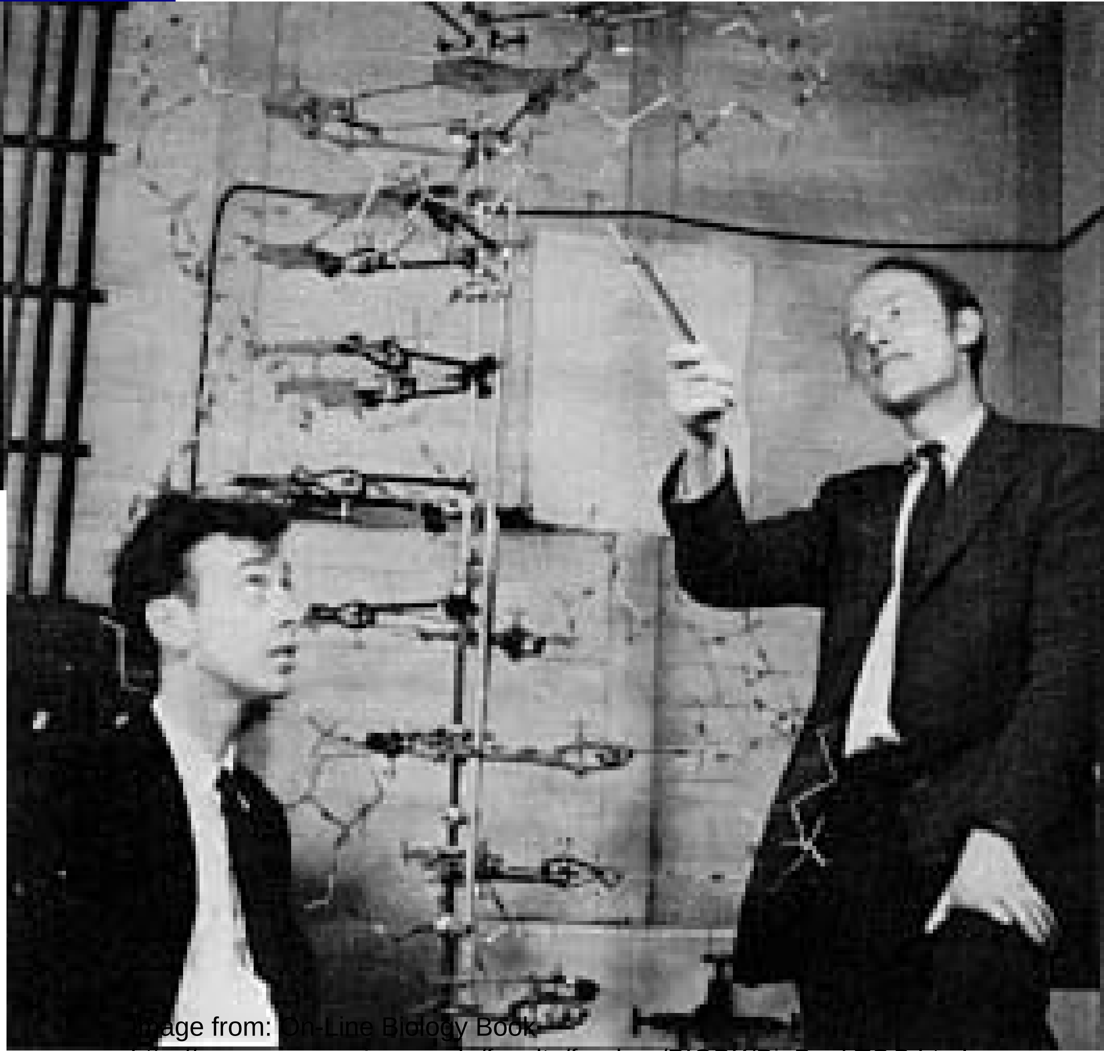
Bioinformatics Schematic of a Cell





Watson and Crick, 1953 in Cambridge

Hydrogen bonds (->*Hybrydization*)



Rosalind Franklin

Image from: On-Line Biology Book

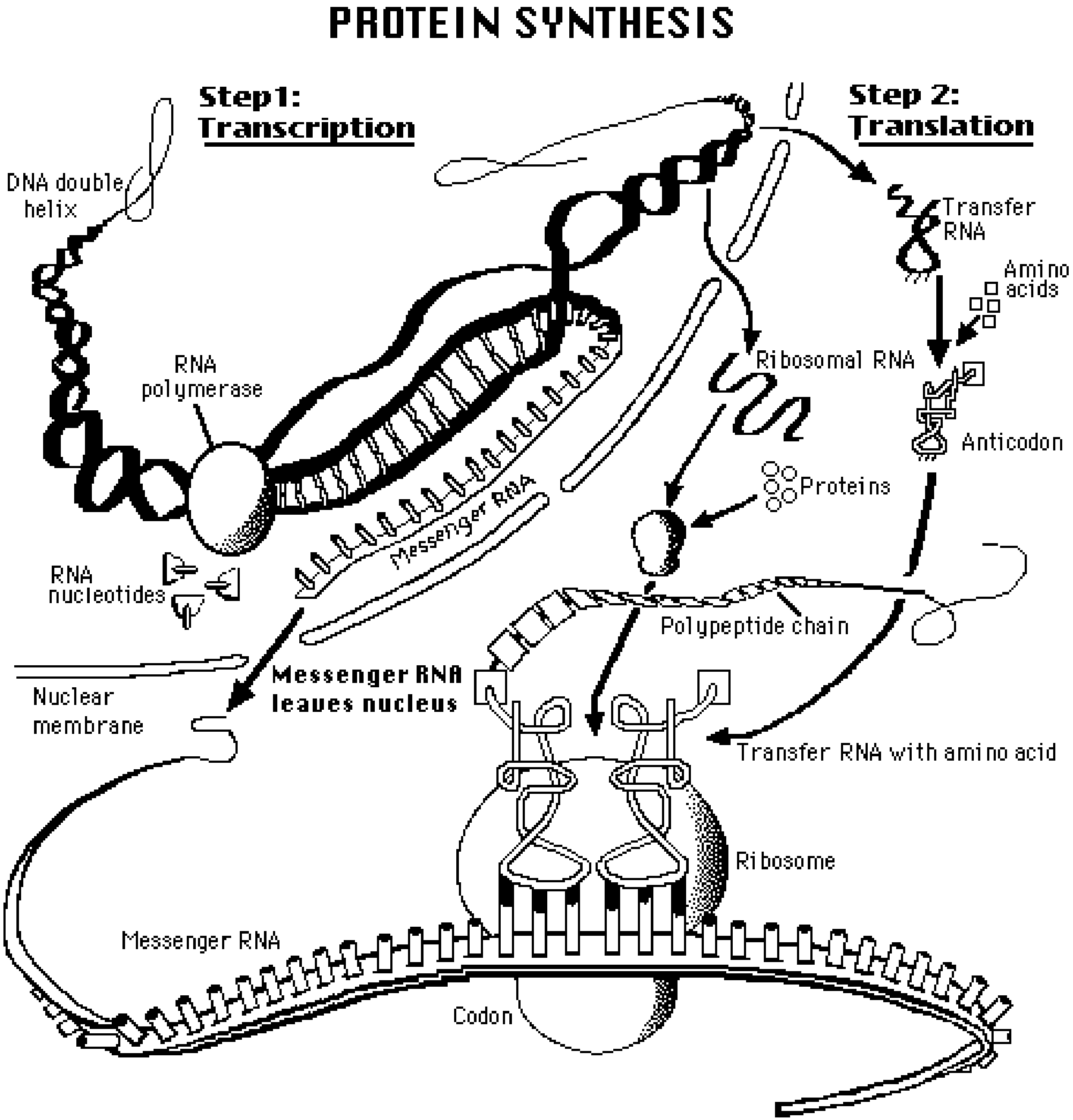
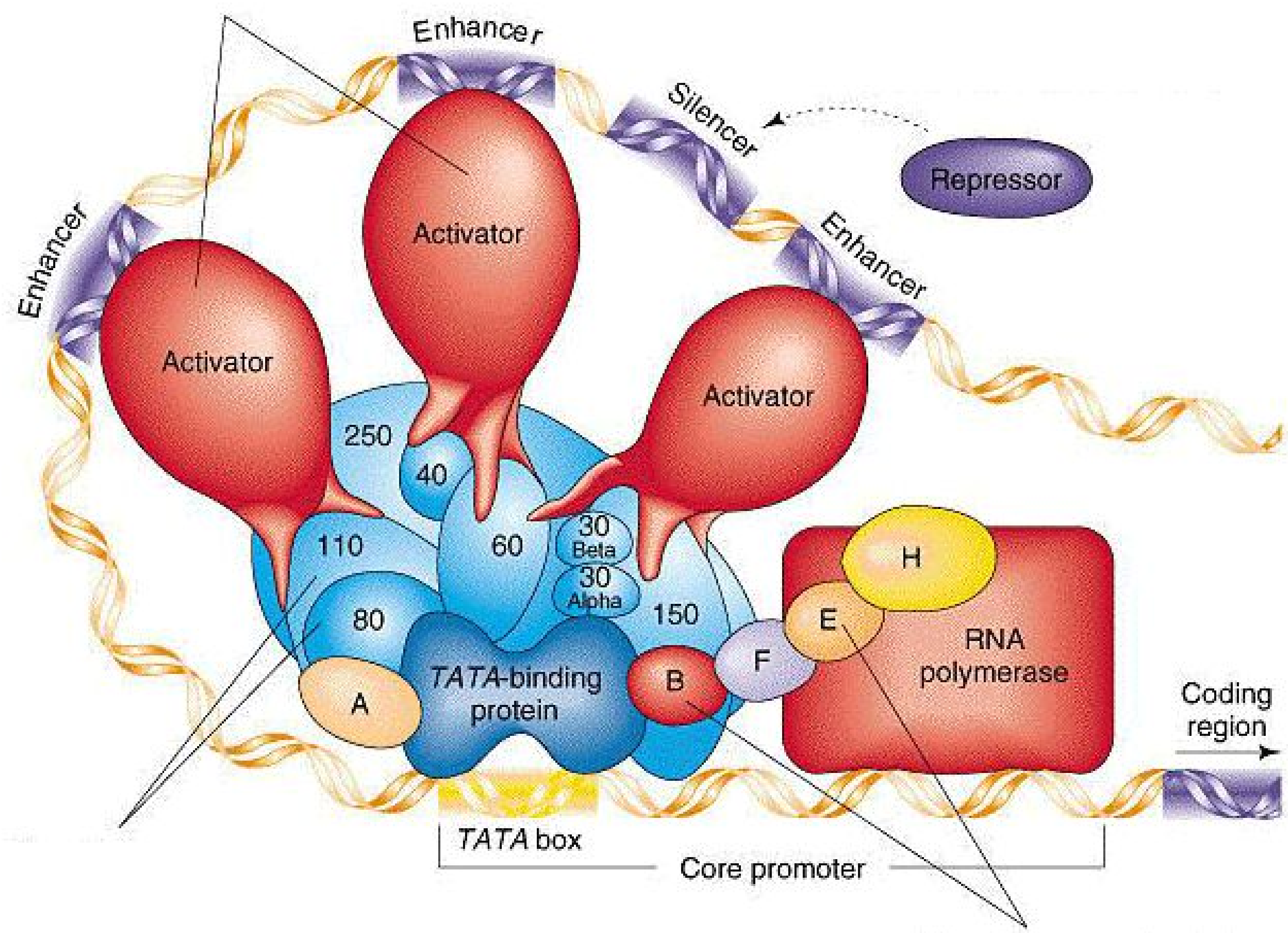


Image from: On-Line Biology Book
http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookTOC.html

Transcription

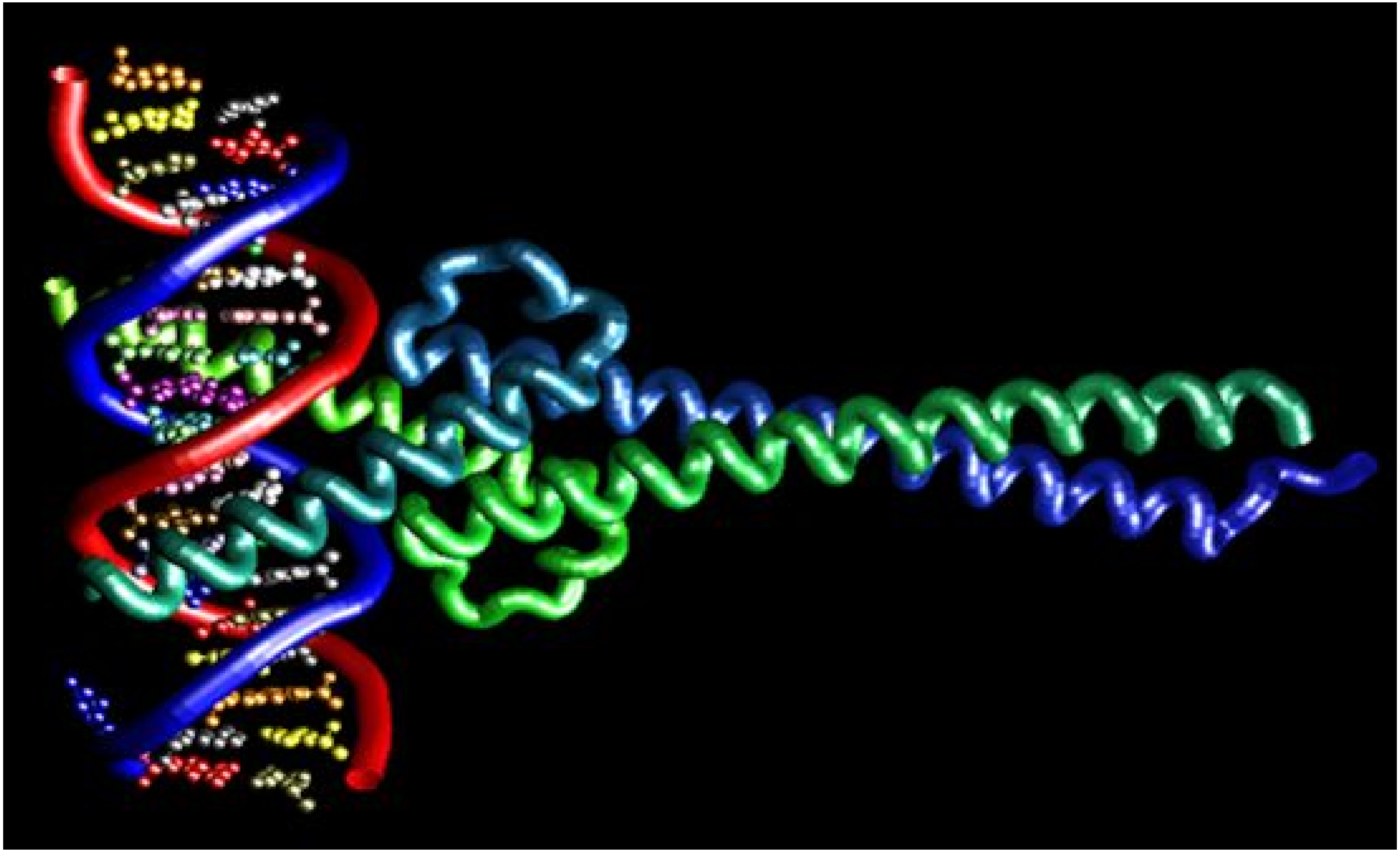
- transcription is accomplished by RNA polymerase
- RNA polymerase binds to **promoters**
- promoters have distinct regions "-35" and "-10"
- transcription start and stop affected by DNA structure
- Additional regulatory sequences can be positive or negative

Complete Assembly of Eukaryotic Gene Regulatory System



http://cats.med.uvm.edu/cats_teachingmod/microbiology/courses/gene_regulation/euk_regulation/eukaryotic_regulation_frameset.htm

Interaction of a transcription factor and DNA



Myc Proto-Oncogene Protein, causing cell division and proliferation

Transcription: DNA -> RNA

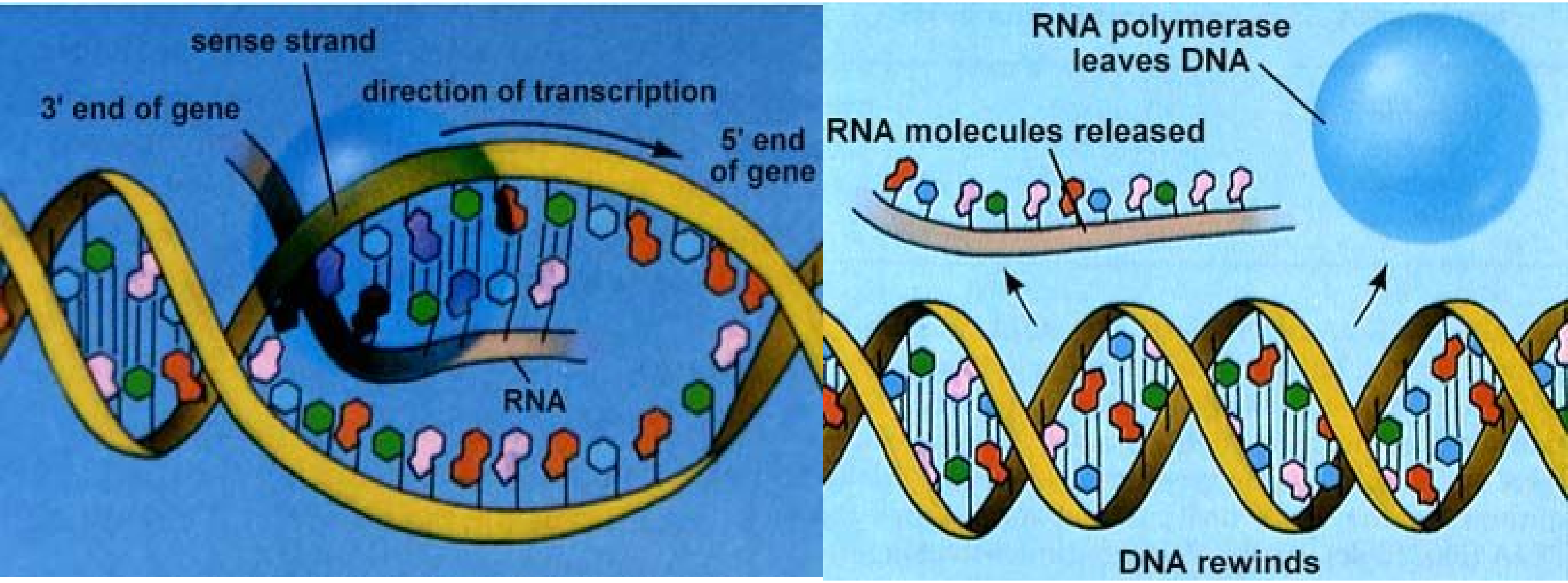
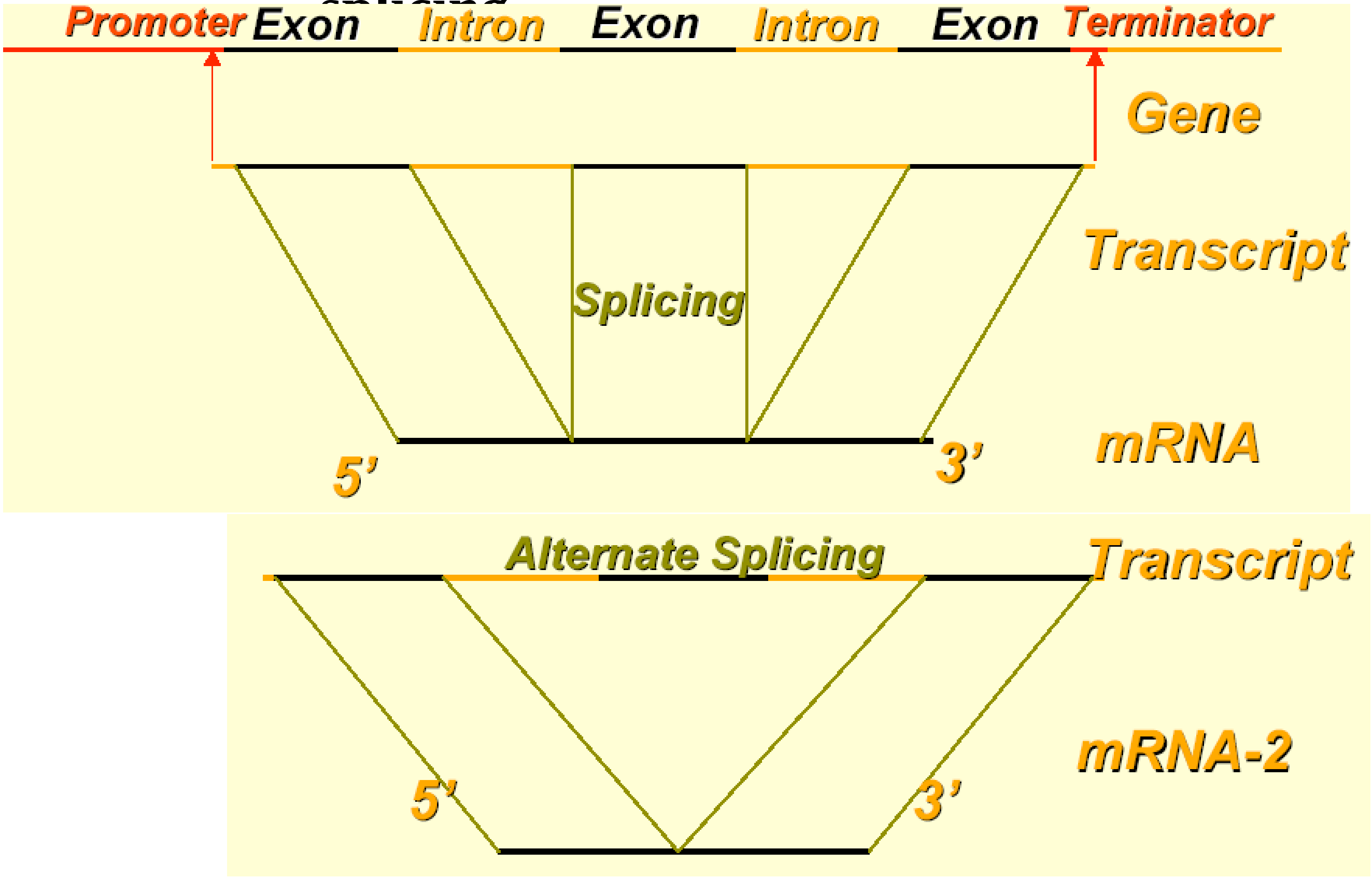


Image from: On-Line Biology Book
http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookTOC.html

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

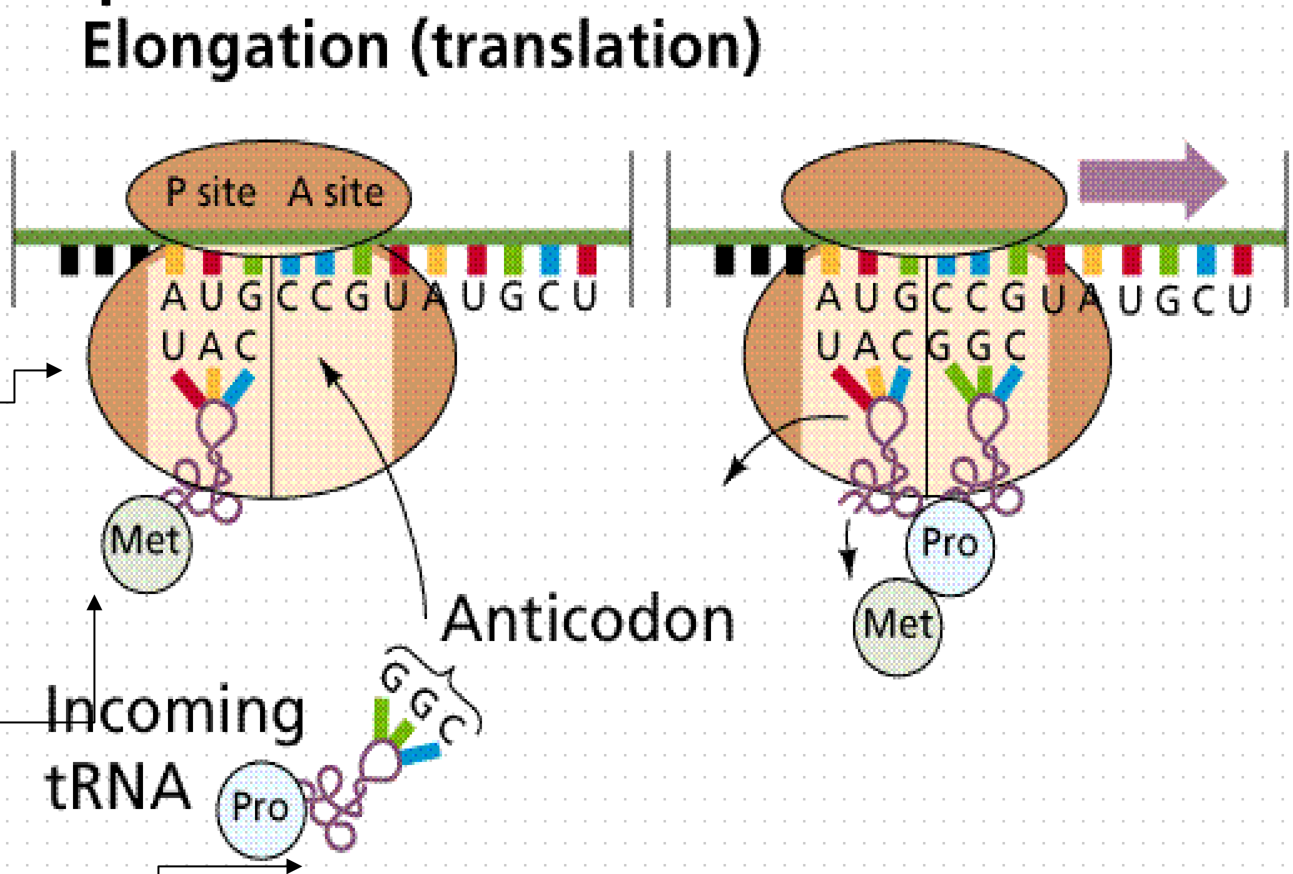


Images from: http://biochem218.stanford.edu (Doug Brutlag)

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:



messenger RNA (mRNA)

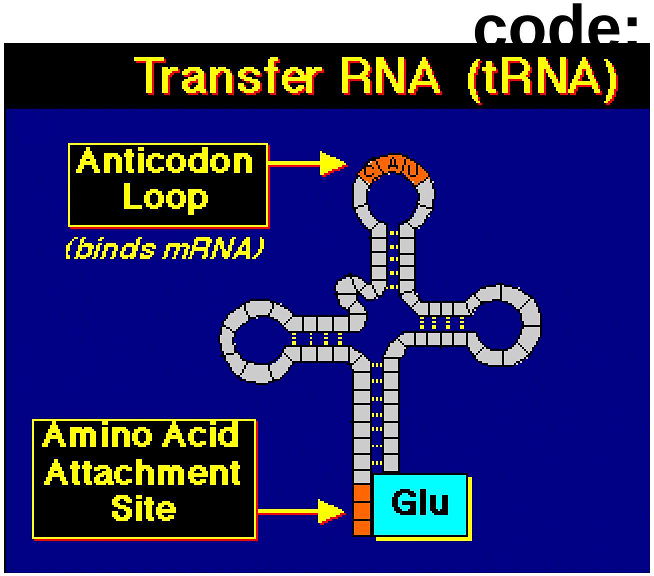
Ribosomes (rRNA)

transfer RNA (tRNA)

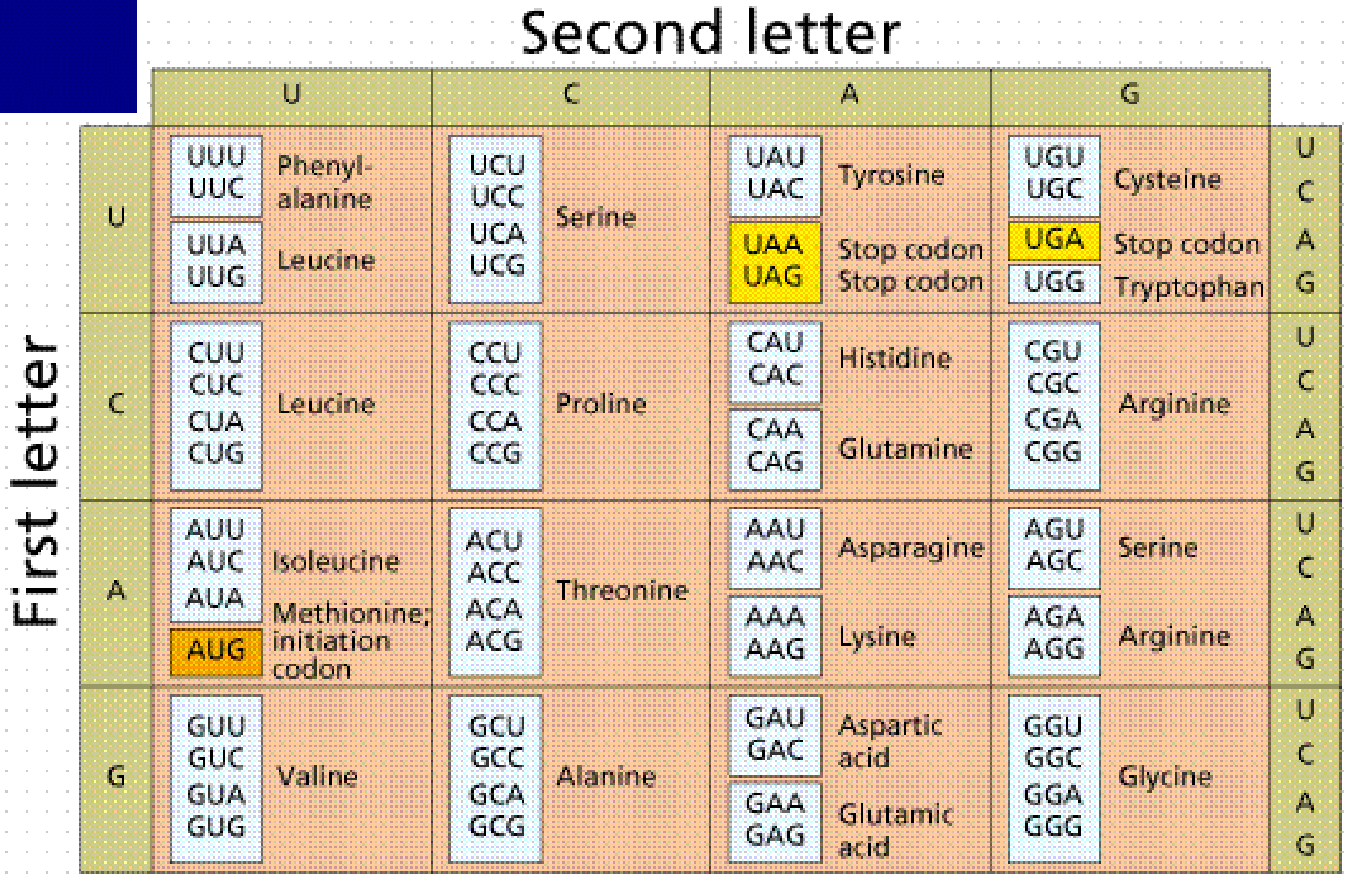
Amino acids

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

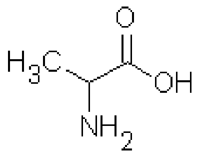


Second letter

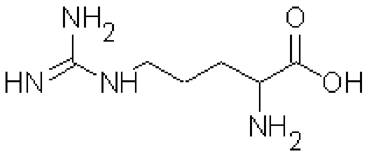
First letter	U	C	A	G	
U	UUU UUC Phenyl-alanine UUA UUG Leucine	UCU UCC UCA UCG Serine	UAU UAC Tyrosine UAA UAG Stop codon Stop codon	UGU UGC Cysteine UGA Stop codon UGG Tryptophan	U C A G
C	CUU CUC CUA CUG Leucine	CCU CCC CCA CCG Proline	CAU CAC Histidine CAA CAG Glutamine	CGU CGC CGA CGG Arginine	U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

Pictures taken from On-Line Biology Book

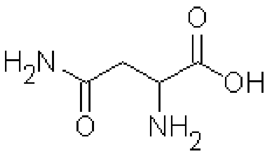
The 20 amino acids, building blocks for proteins



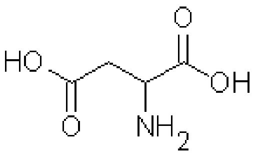
Alanin (Ala)



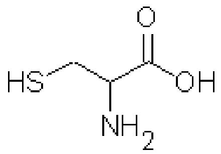
Arginin (Arg)



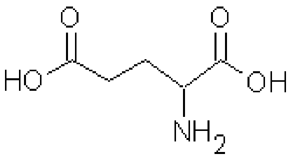
Asparagin (Asn)



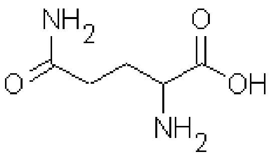
Asparaginsäure (Asp)



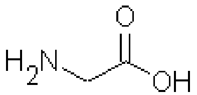
Cystein (Cys)



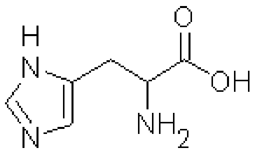
Glutaminsäure (Glu)



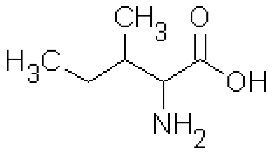
Glutamin (Gln)



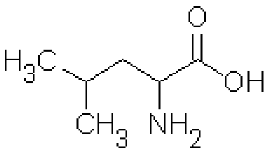
Glycin (Gly)



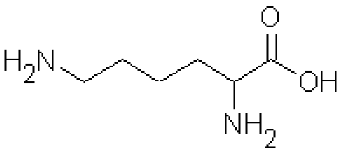
Histidin (His)



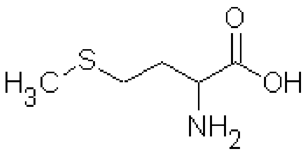
Isoleucin (Ile)



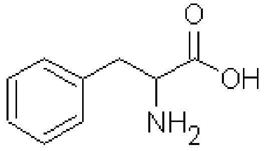
Leucin (Leu)



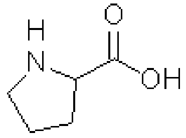
Lysin (Lys)



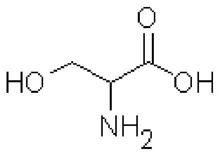
Methionin (Met)



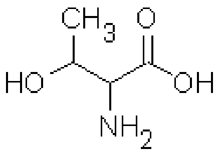
Phenylalanin (Phe)



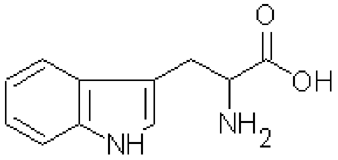
Prolin (Pro)



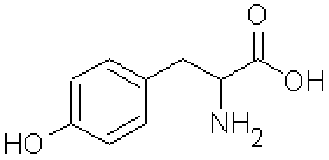
Serin (Ser)



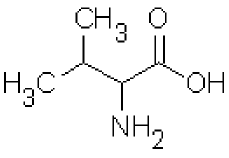
Threonin (Thr)



Tryptophan (Trp)



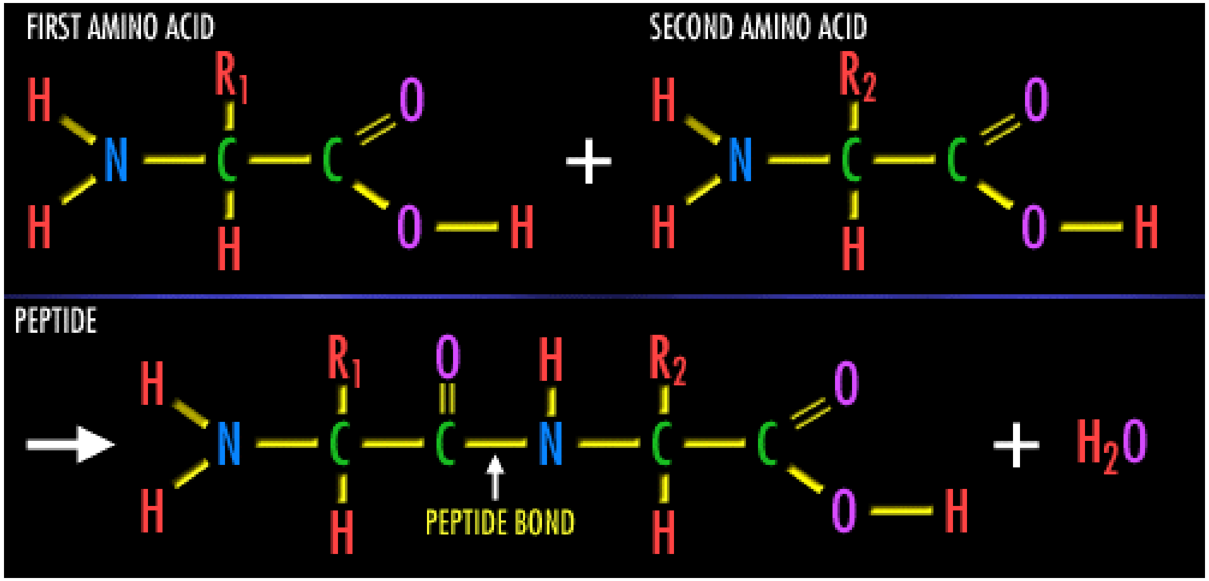
Tyrosin (Tyr)



Valin (Val)

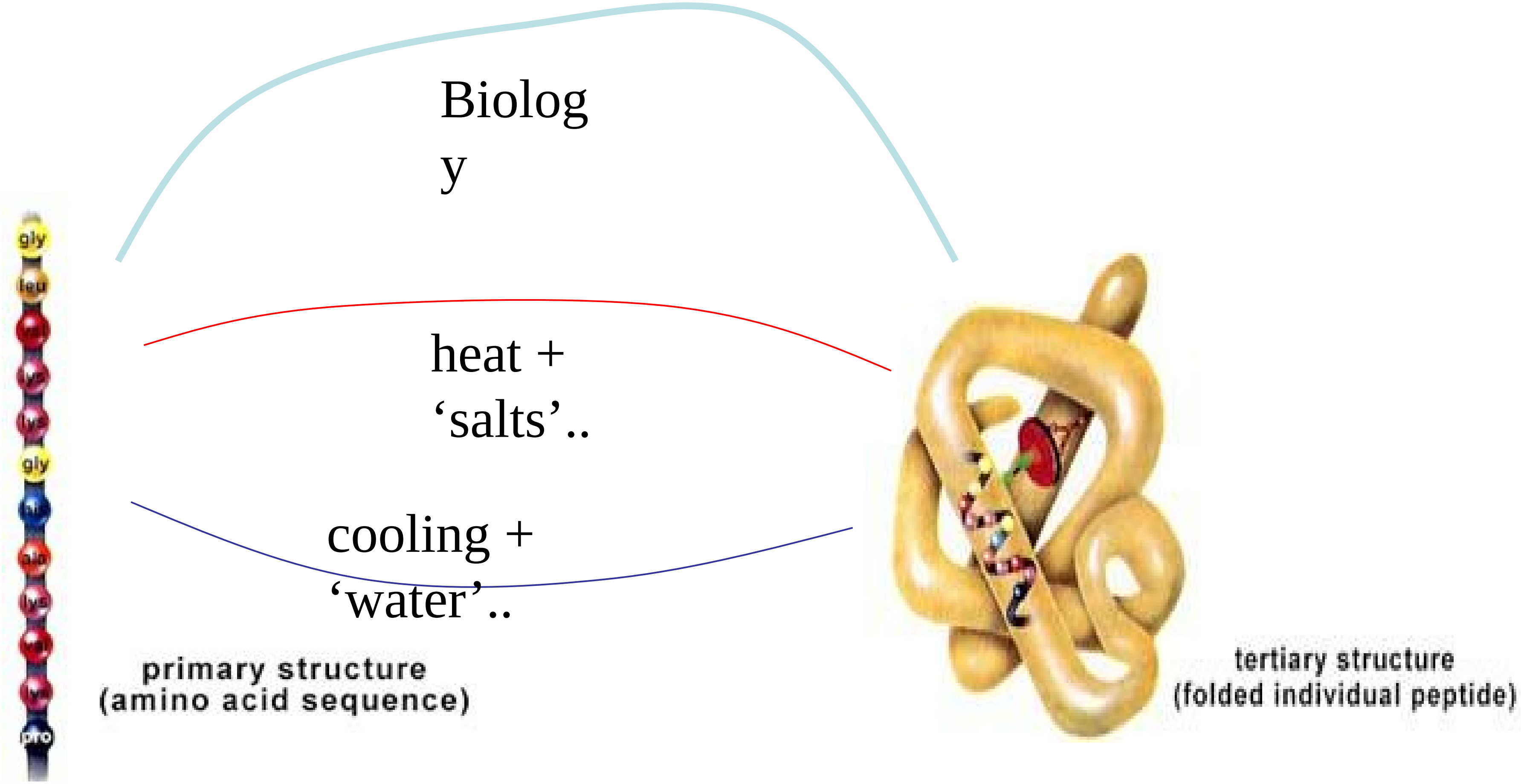
Building proteins (chemistry):

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



Biology

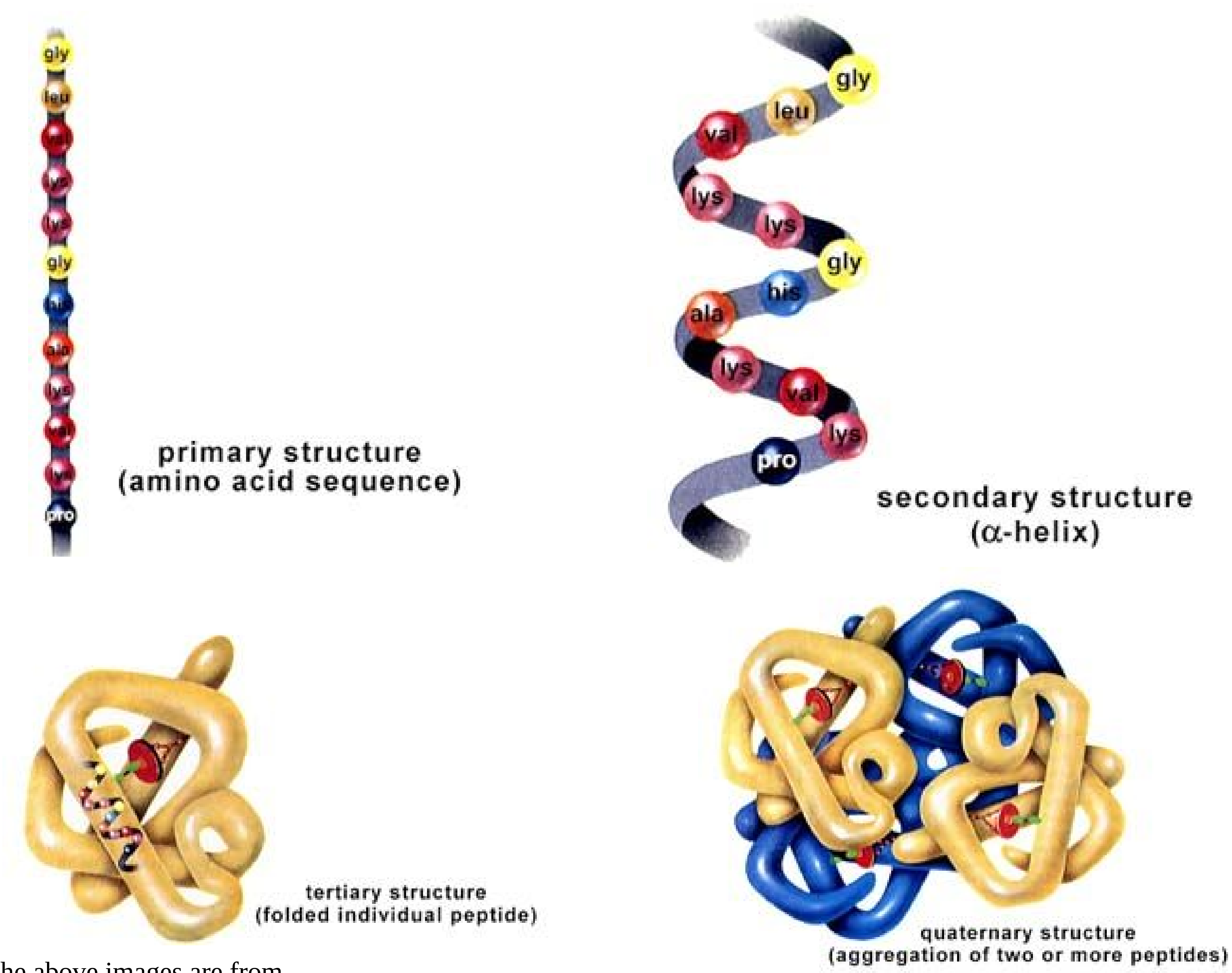
heat +
‘salts’..

cooling +
‘water’..

C. Anfinsen,
1973

The above images are from
http://www.biosci.uga.edu/almanac/bio_103/notes/may_14.html.

Levels of structural description



The above images are from

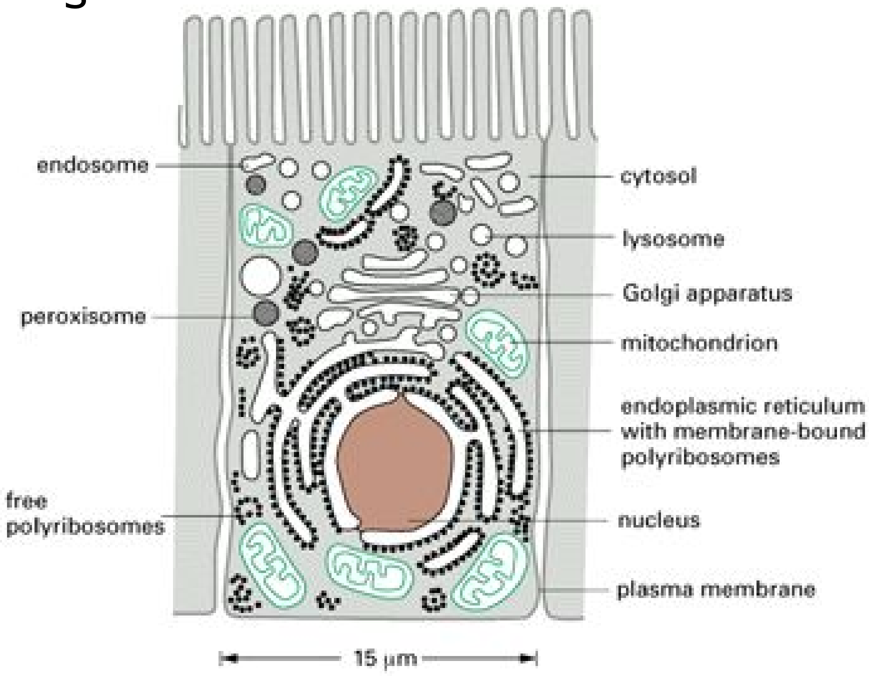
http://www.biosci.uga.edu/almanac/bio_103/notes/may_14.html.

Protein localization

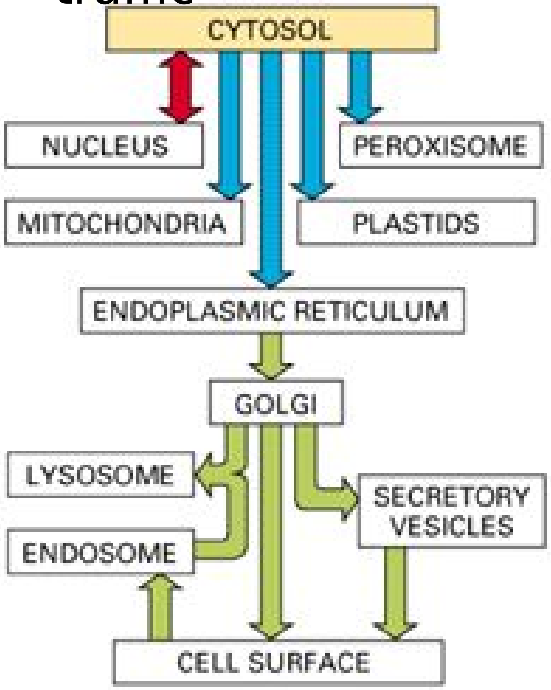
- leader sequences can specify cellular location (e.g., insert across membranes)
- leader sequences usually removed by cleavage
- Like an address sticker

Protein localization

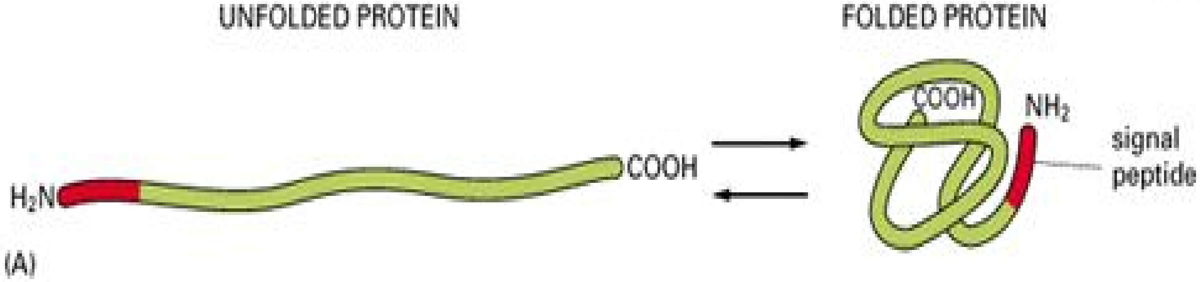
compartment
s



protein
traffic



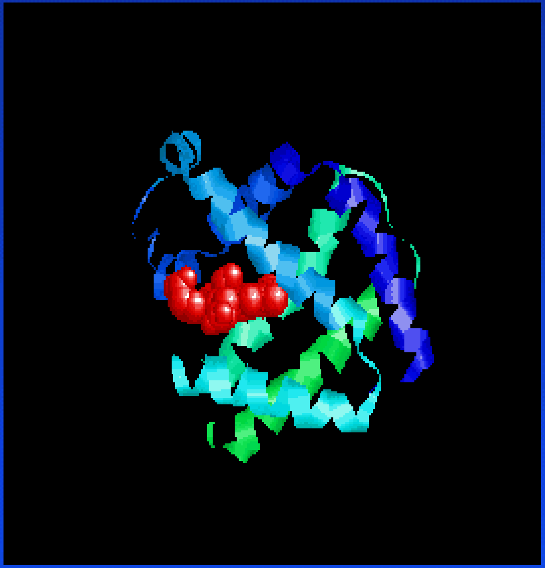
KEY: 🟥 = gated transport
🟦 = transmembrane transport
🟩 = vesicular transport

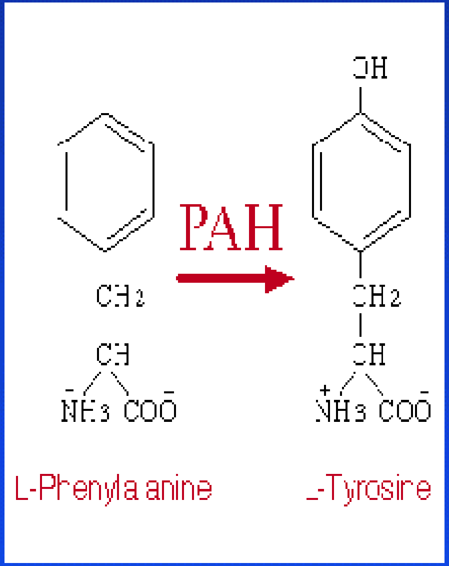


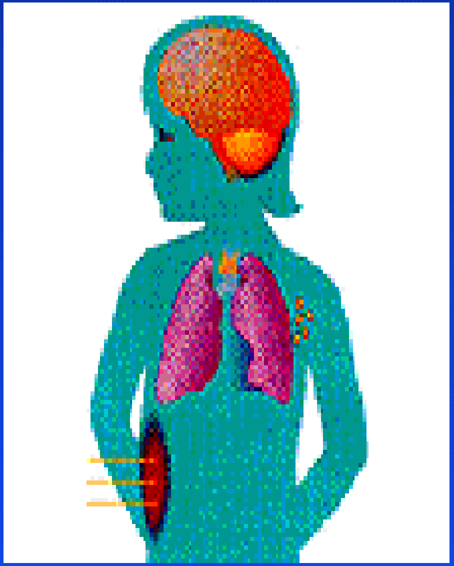
Central Paradigm of Bioinformatics

Genetic Information → Molecular Structure → Biochemical Function → Phenotype (Symptoms)

TGCTTTAGCTTT
AAACTACAGGCC
TCACTGGAGCTA
GAGACAAGAAGG
TAAAAAACGGCT
GACAAAAGAAGT
CCTGGTATCCTC
TATGATGGGAGA
AGGAAACTAGCT
AAAGGGAAGAAT
AAATTAGAGAAA
AACTGGAATGAC
GCTTATACCTGG





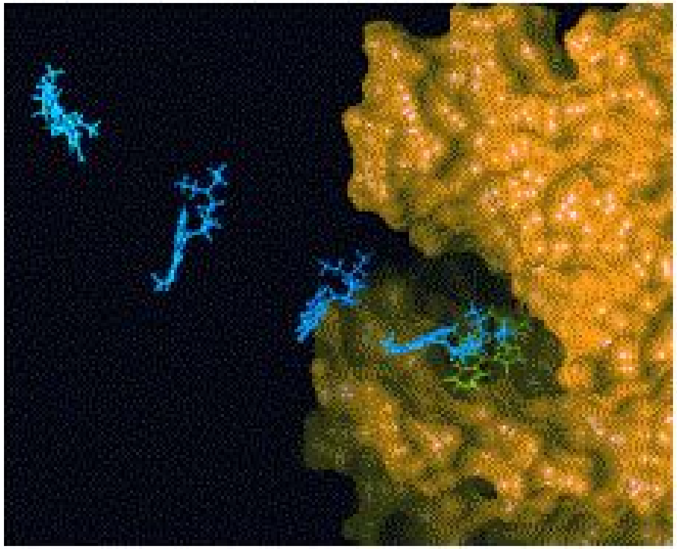


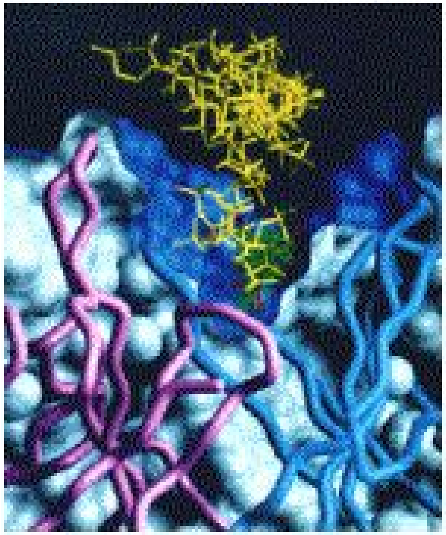
Doug Brutlag:
http://cmgm.stanford.edu/biochem218/01Representation.pdf

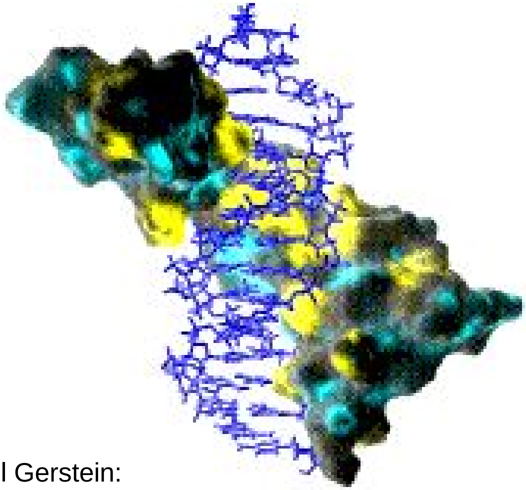
Protein/Ligand interactions:

- 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).



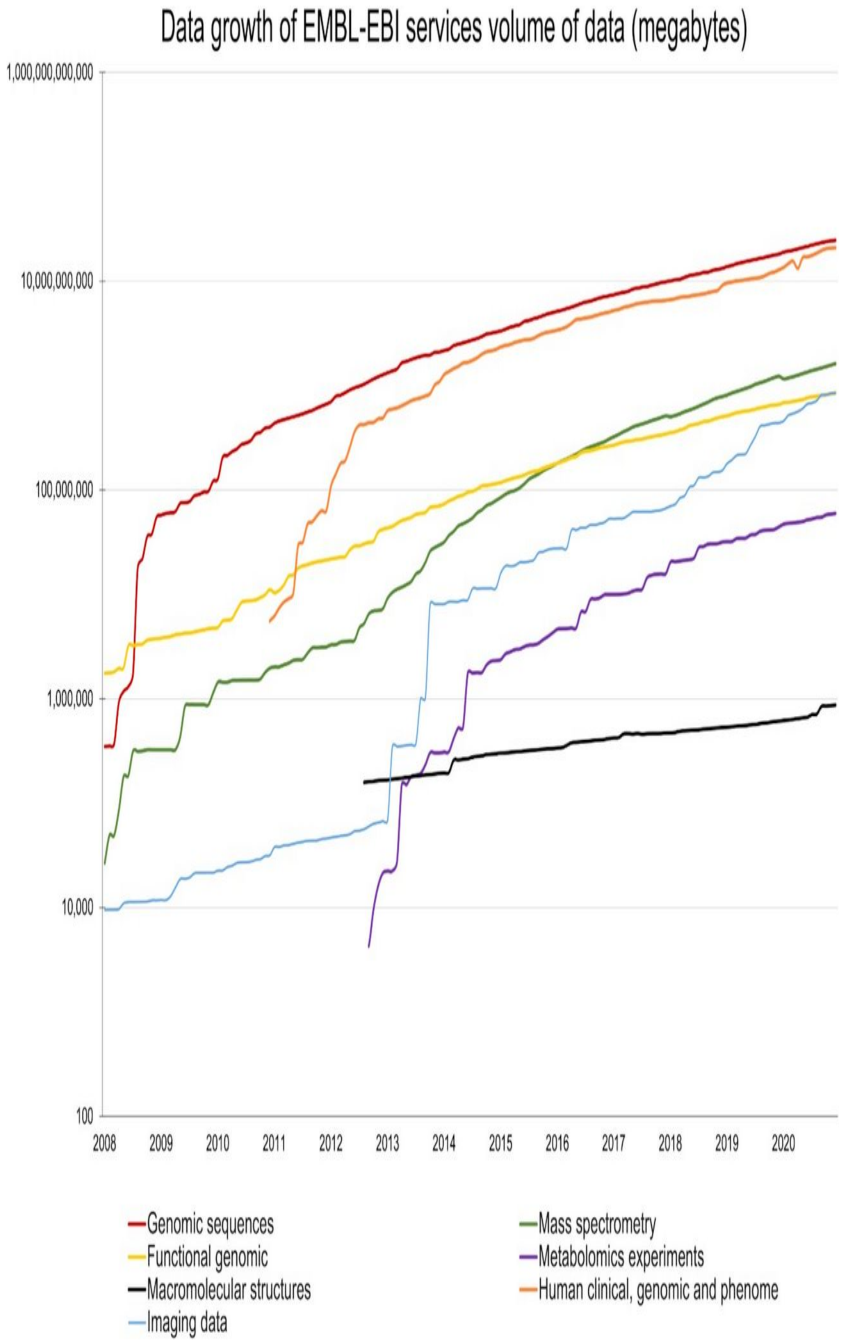




Michael Gerstein:
http://bioinfo.mbb.yale.edu/mbb452a/intro/intro.pdf

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

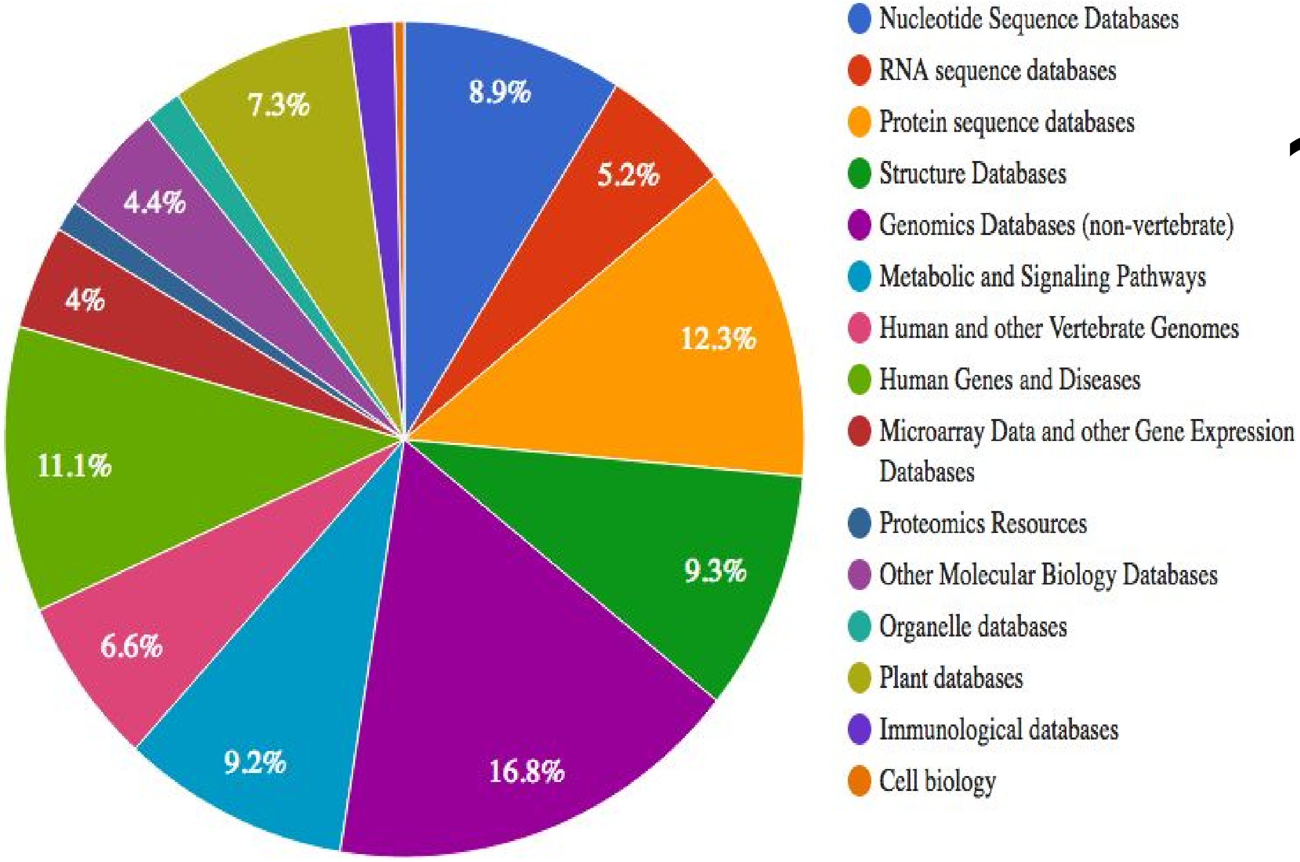


Data growth in the life sciences

- 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

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

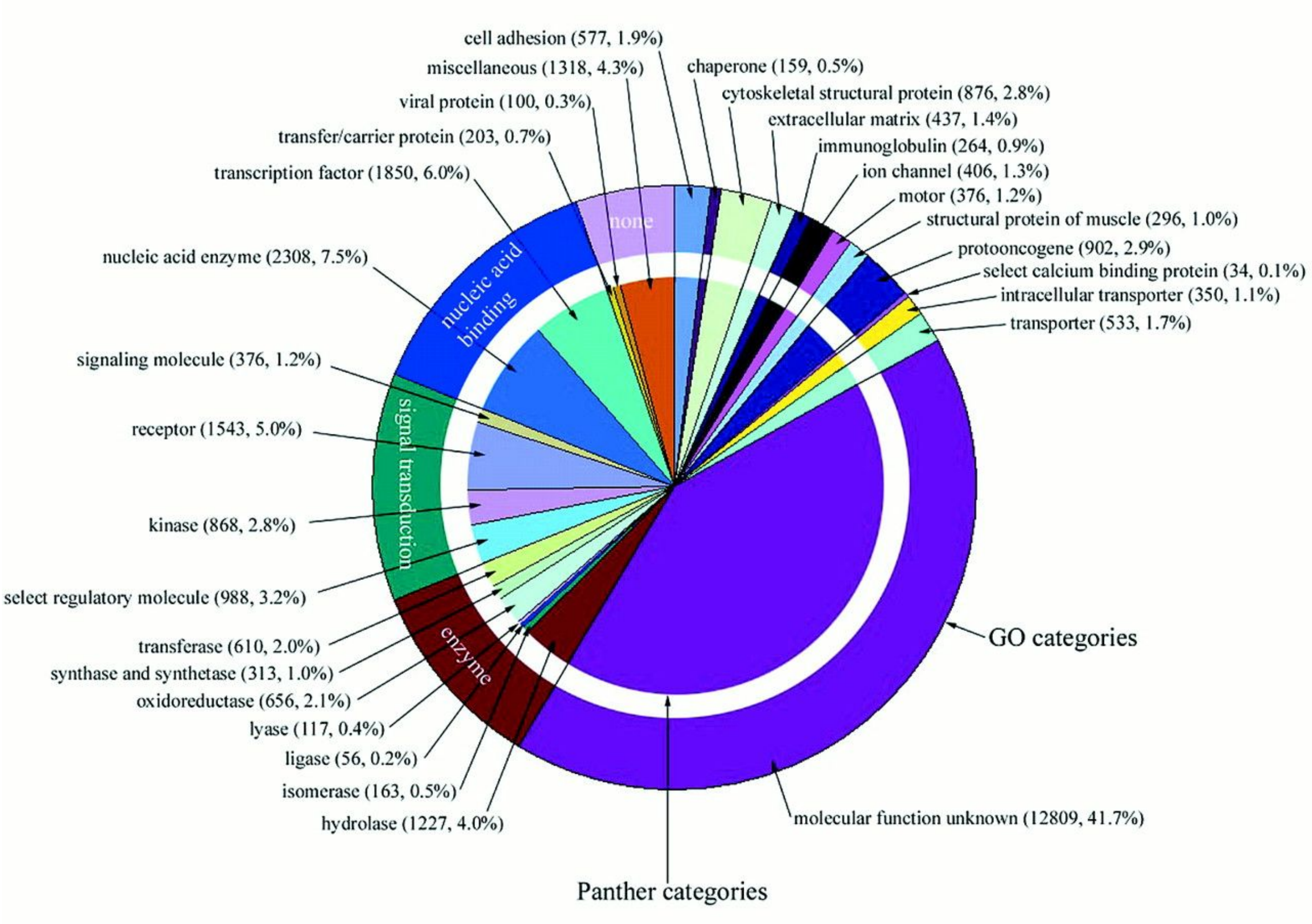
Only the surface is scratched:

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.htmlml

‘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)

Graphics from Dimitris Kafetzopoulos, IMBB

From Genomics to Drugs

Thomas Lengauer (Ed.)

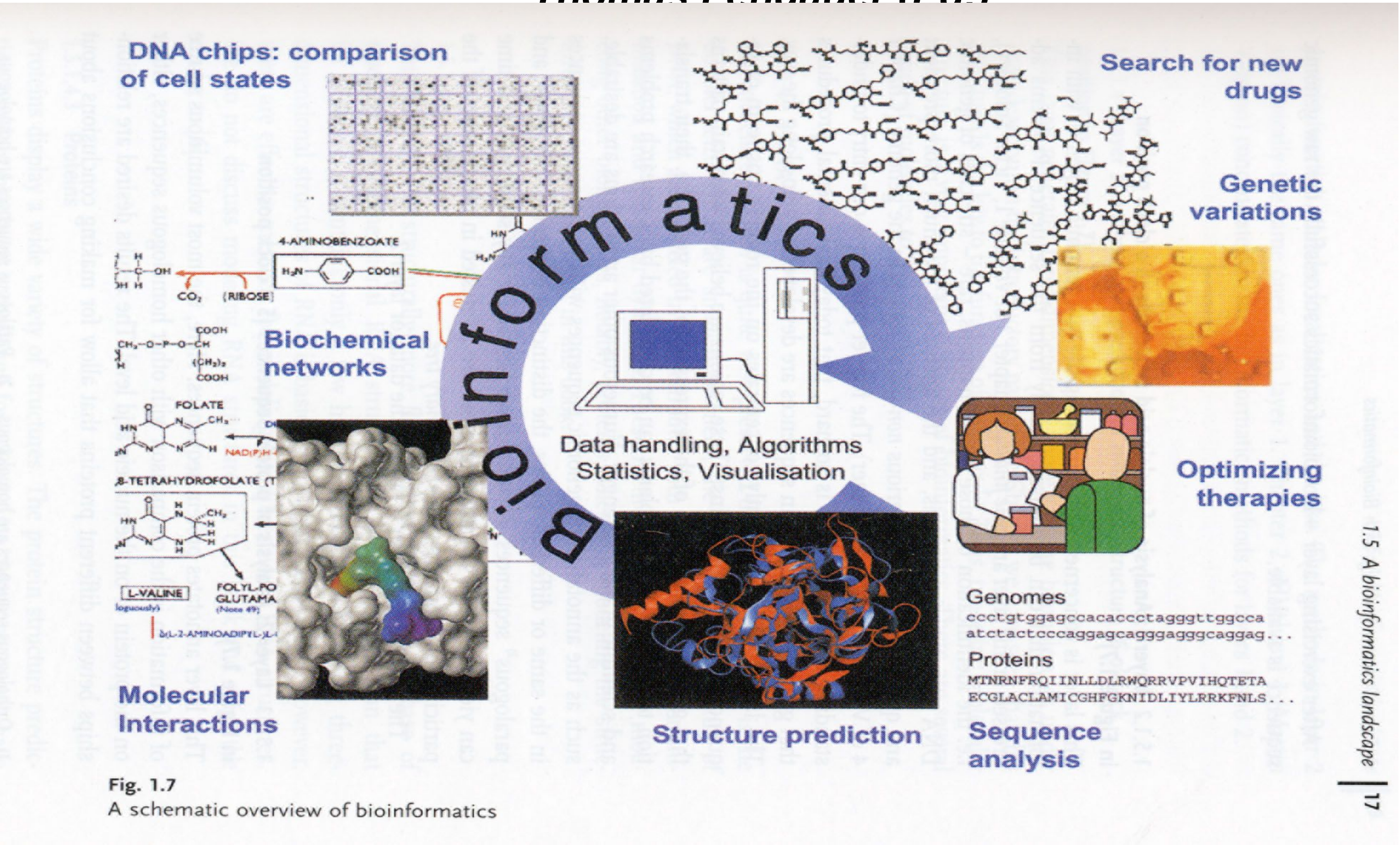


Fig. 1.7
A schematic overview of bioinformatics

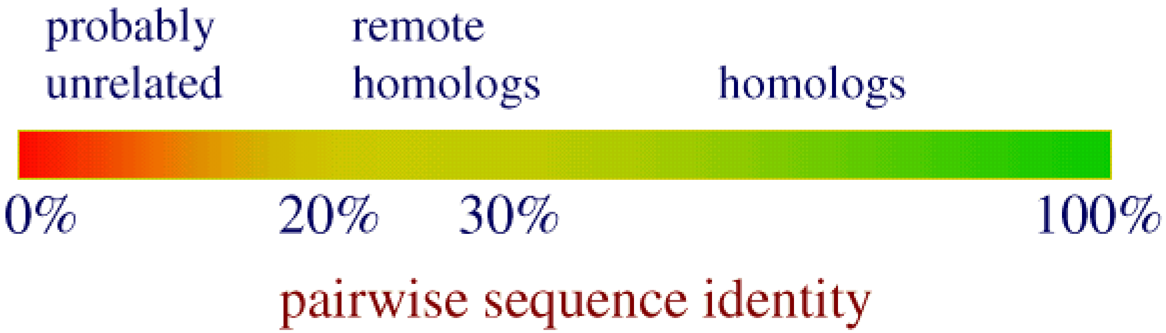
NGS+Robotics at BSRC Alexander Fleming:
https://www.youtube.com/watch?v=8CaUGFimbgQ
https://www.youtube.com/watch?v=kUdDY3kvWpc

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

Problem ¬
:



Mark Craven/Thomas Anantharaman:

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
```

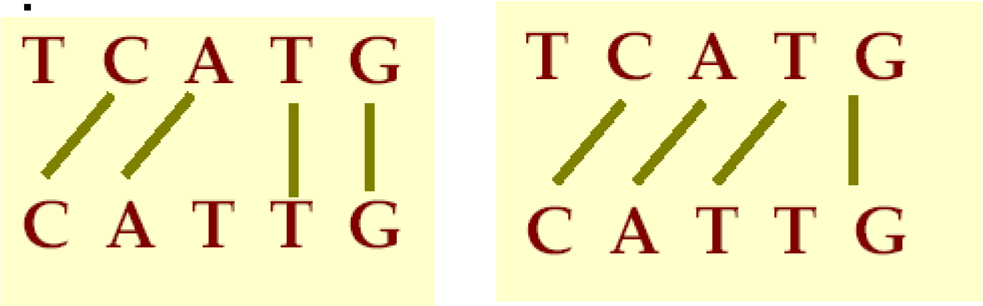
```
2 matches, 0 gaps
  T   C   A   T   G
              |   |
  C   A   T   T   G
```

```
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
```

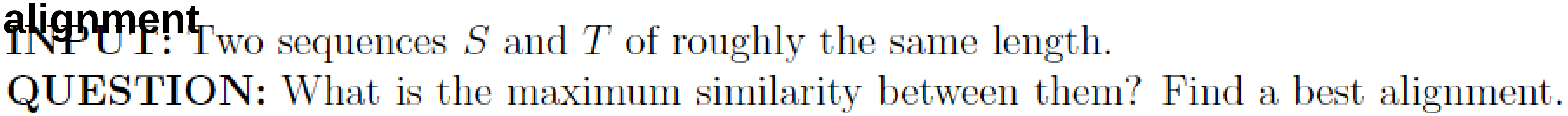
```
4 matches, 1 insertion
  T   C   A   T   -   G
      |   |   |       |
  .   C   A   T   T   G
```

Ambiguity:



Definitions

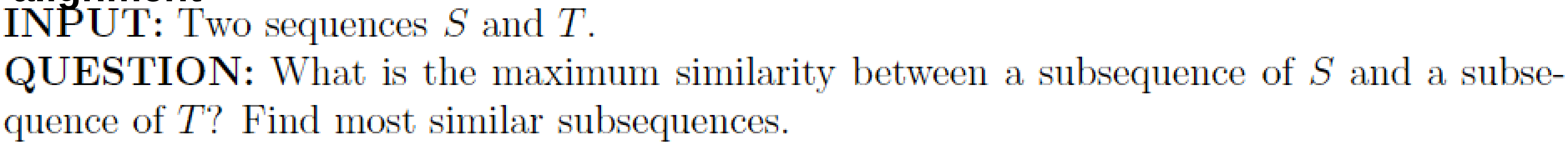
Global alignment



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

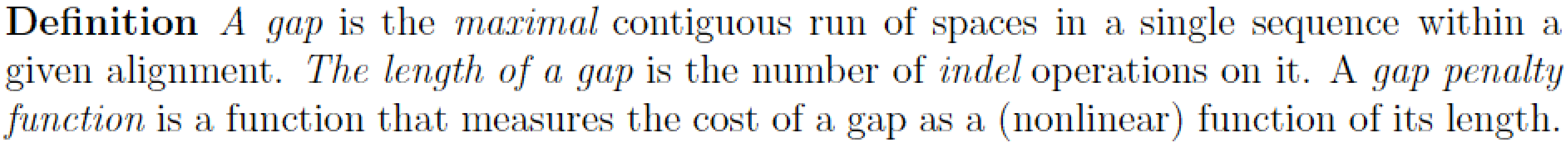
QUESTION: What is the maximum similarity between them? Find a best alignment.

Local alignment



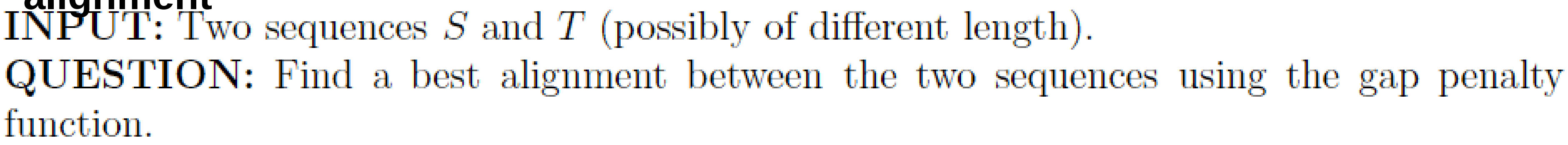
INPUT: Two sequences S and T.

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.

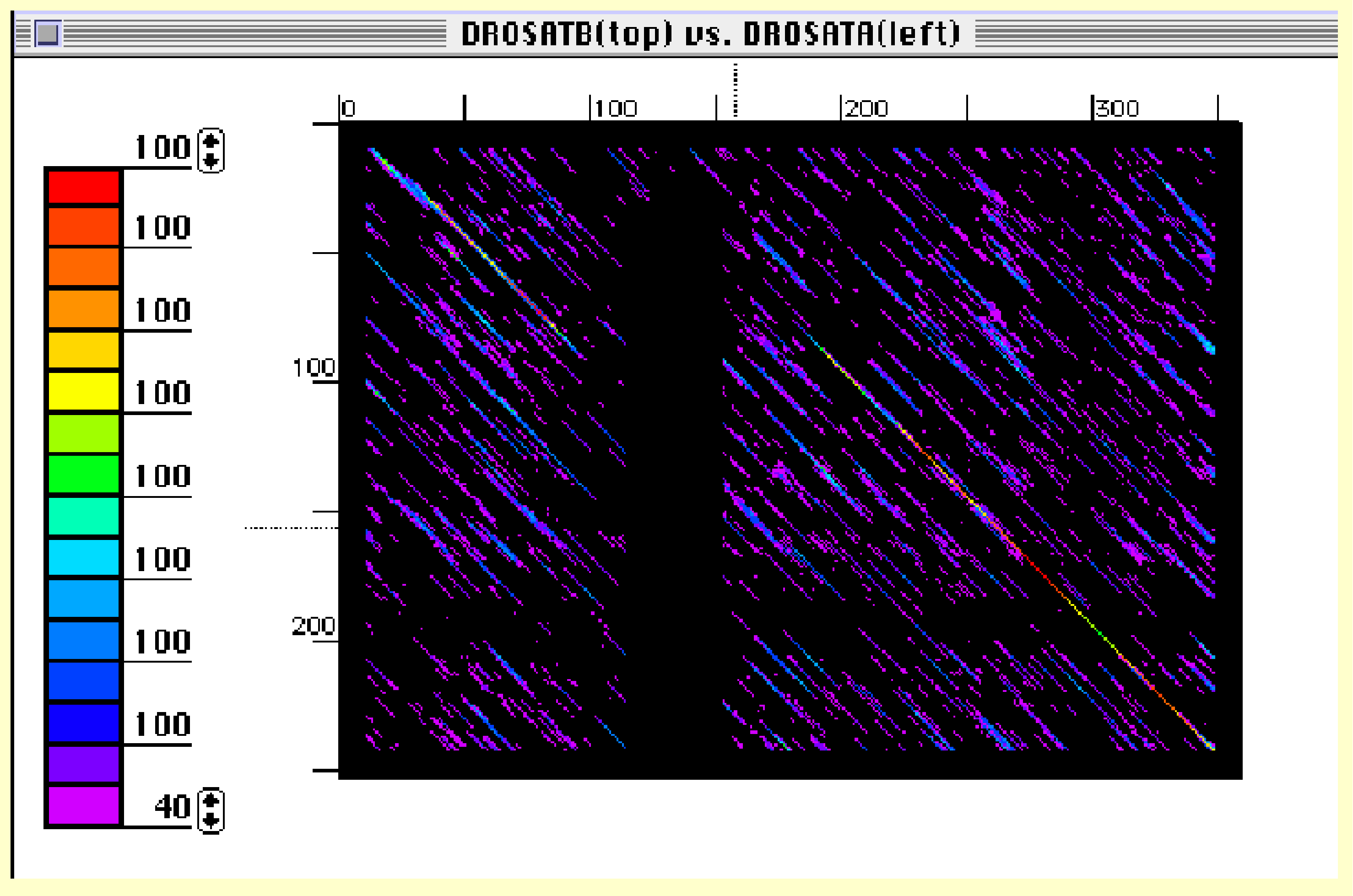
Gapped alignment



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

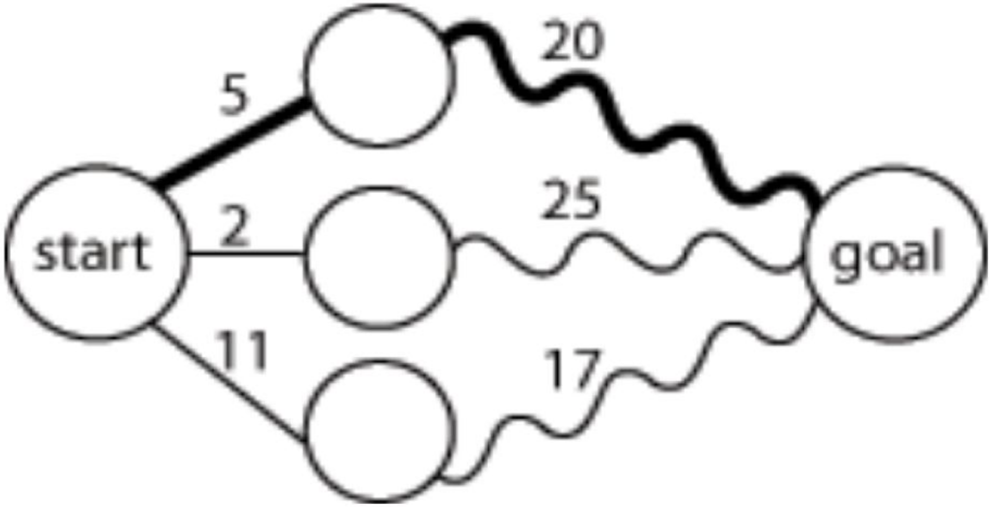


Dynamic programming algorithms for sequence comparison

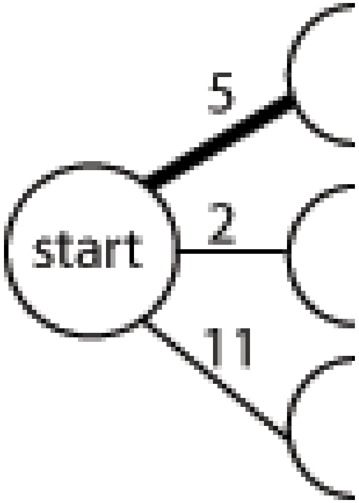
- Introduced for biological sequences by
  - S. B. Needleman & C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol. 48:*443-453 (1970)

Robert F. Murphy: http://www.cmu.edu/bio/education/courses/03310/LectureNotes/LecturesPart07.pptf

Dynamic programming reminder: Shortest path



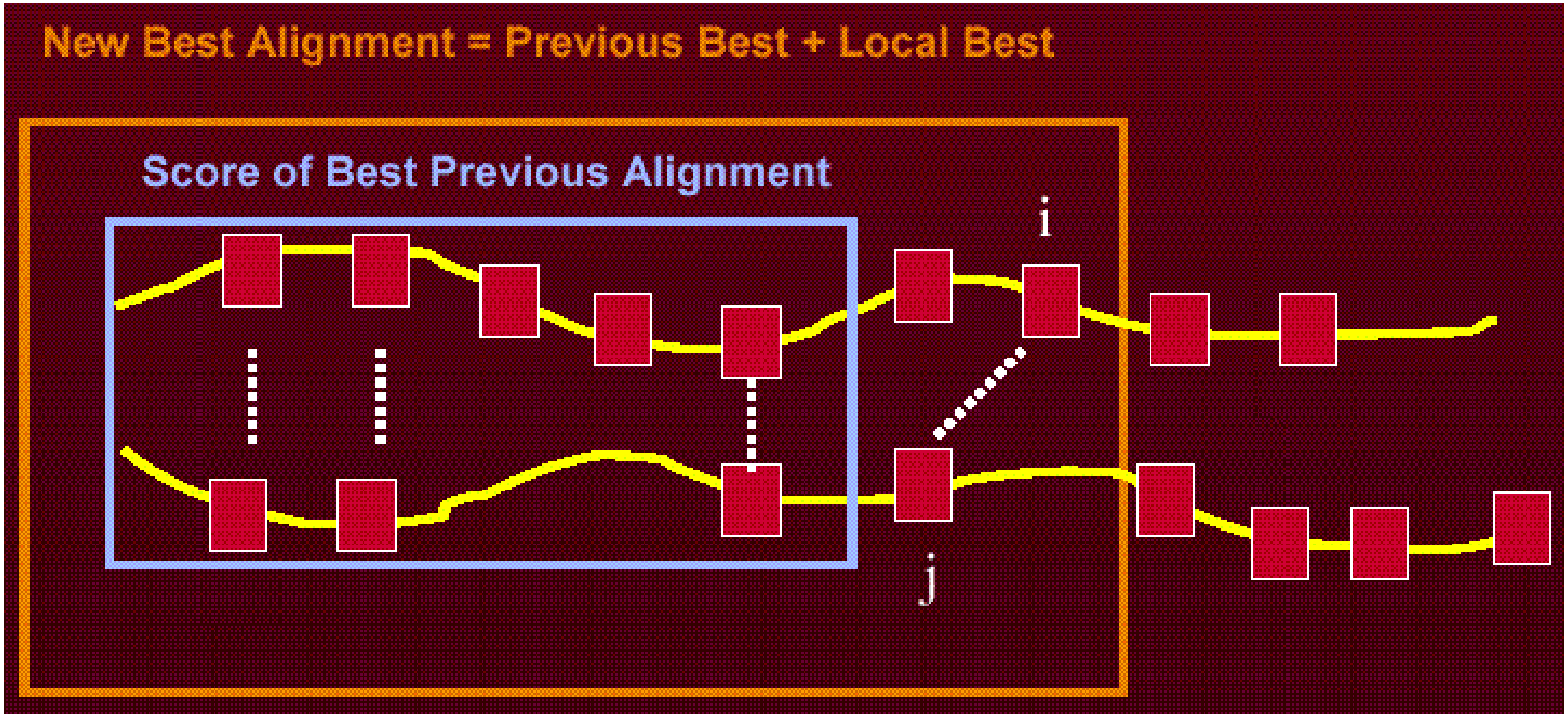
Dynamic programming reminder: Shortest path



Best solutions up to n

One node added:
n updates to find new best

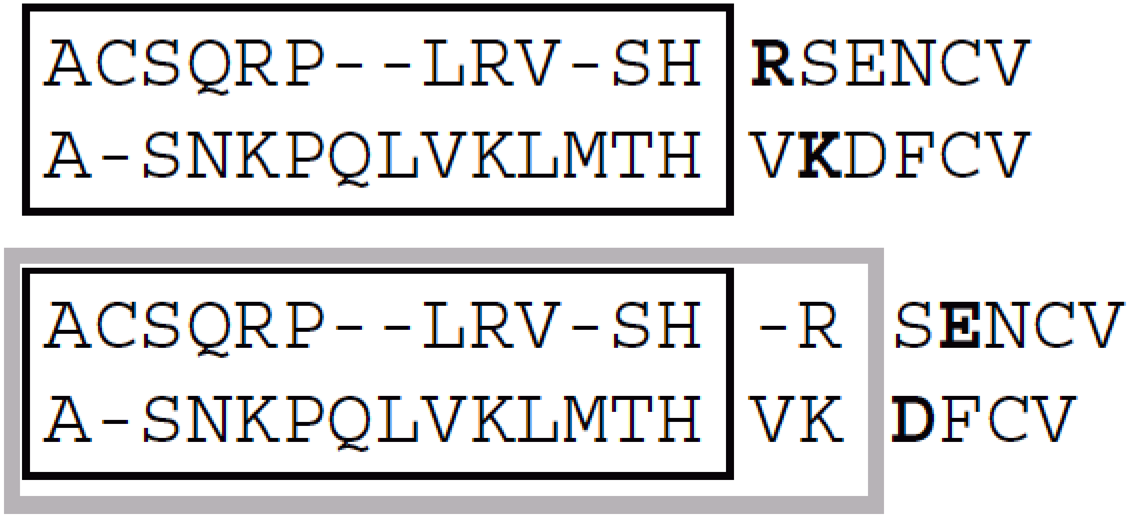
Dynamic Programming Idea:


New Best Alignment = Previous Best + Local Best
Score of Best Previous Alignment
i
j

© 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
    - 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.
    - How could this be violated?
      Aligning R to K changes best alignment in box.

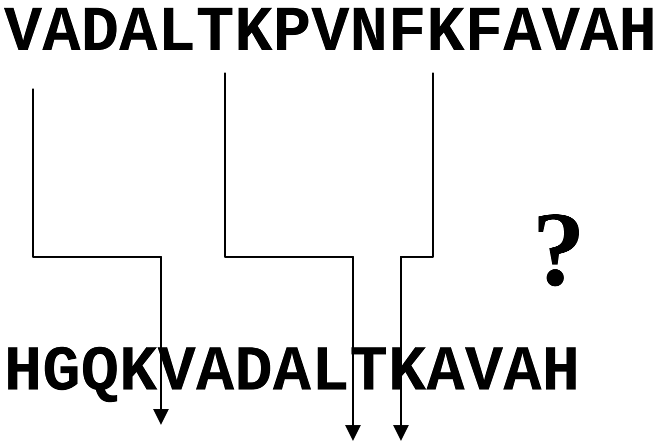


ACSQRP--LRV-SH **R**SENCV
A-SNKPQLVKLMTH V**K**DFCV

ACSQRP--LRV-SH -R S**E**NCV
A-SNKPQLVKLMTH VK **D**FCV

Optimal alignment between sequences

Problem:



VADALTKPVNFKFAVAH

?

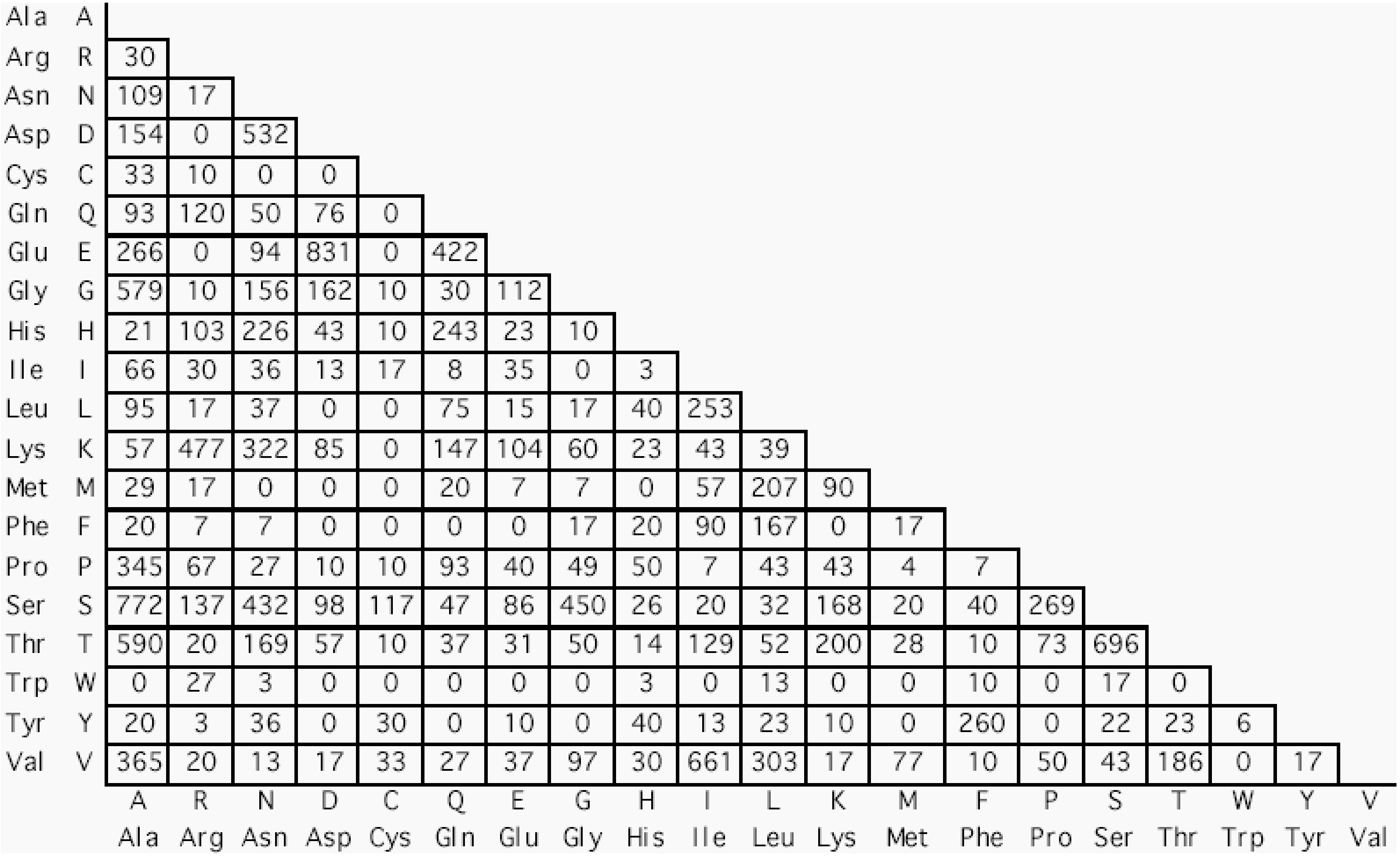
HGQKVADALTKAVAH

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)

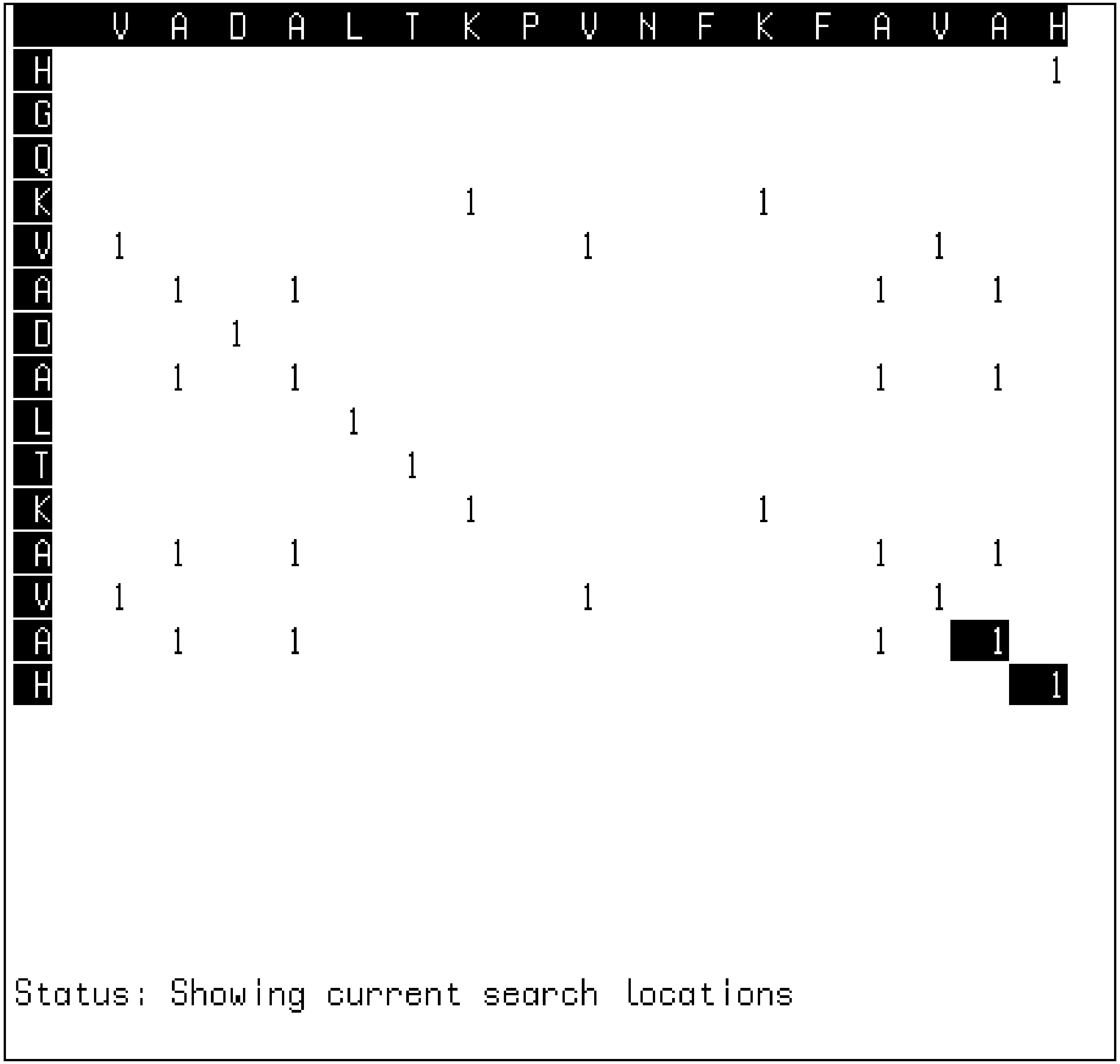


Ala	A																				
Arg	R	30																			
Asn	N	109	17																		
Asp	D	154	0	532																	
Cys	C	33	10	0	0																
Gln	Q	93	120	50	76	0															
Glu	E	266	0	94	831	0	422														
Gly	G	579	10	156	162	10	30	112													
His	H	21	103	226	43	10	243	23	10												
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

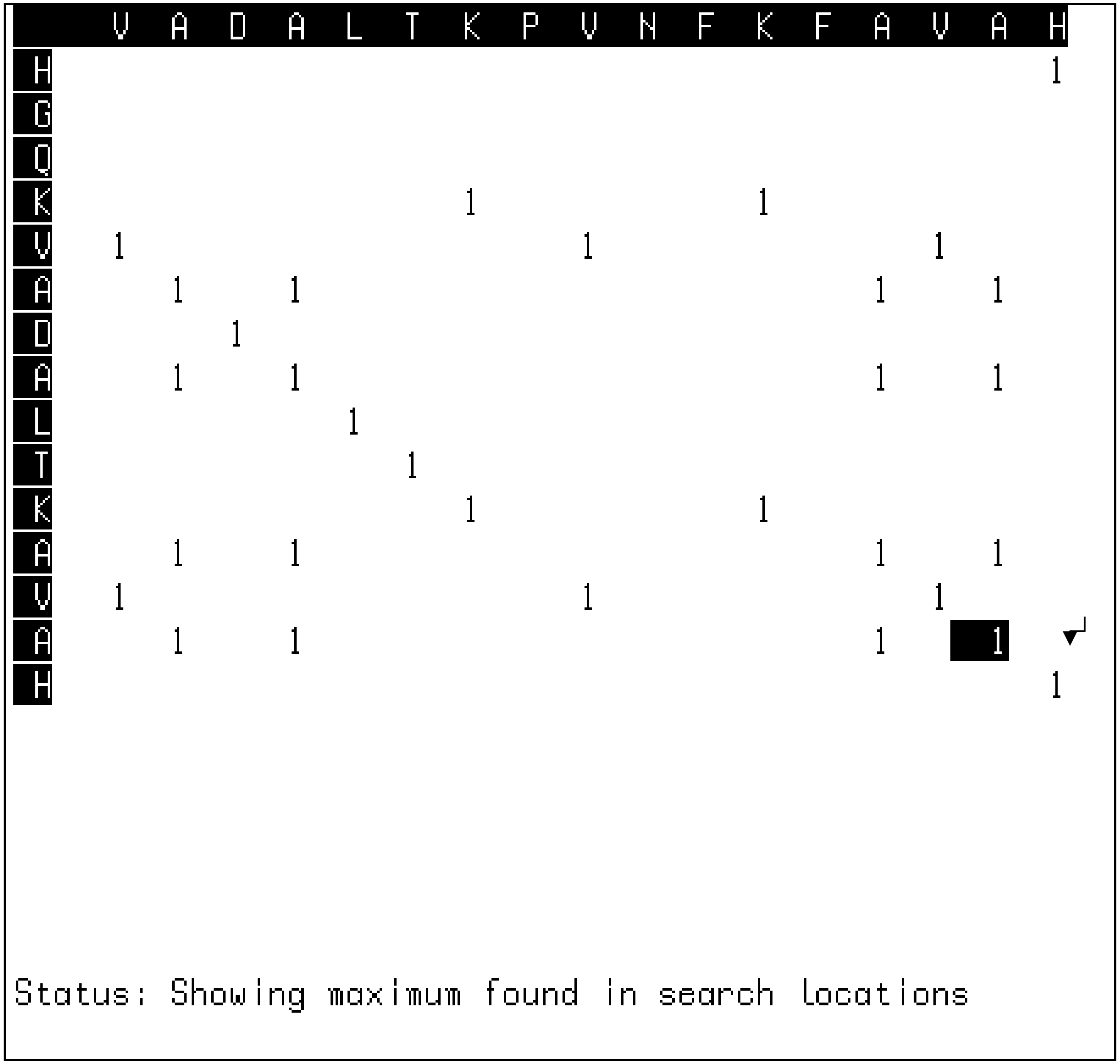
Robert F. Murphy:



	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

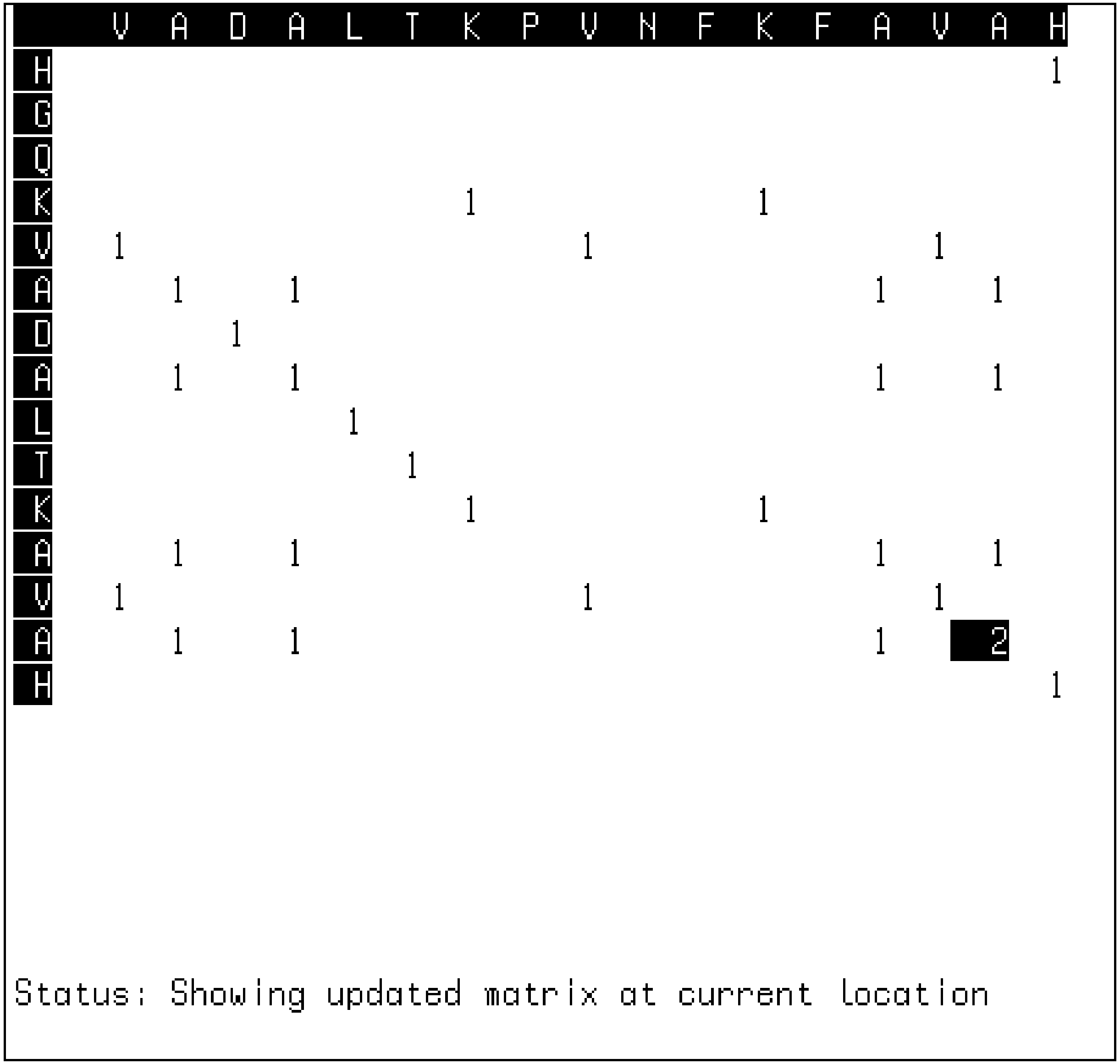
Status: Showing current search locations

Robert F. Murphy:
http://www.cmu.edu/bio/education/courses/03310/LectureNotes/LecturesPart07.pptf



	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

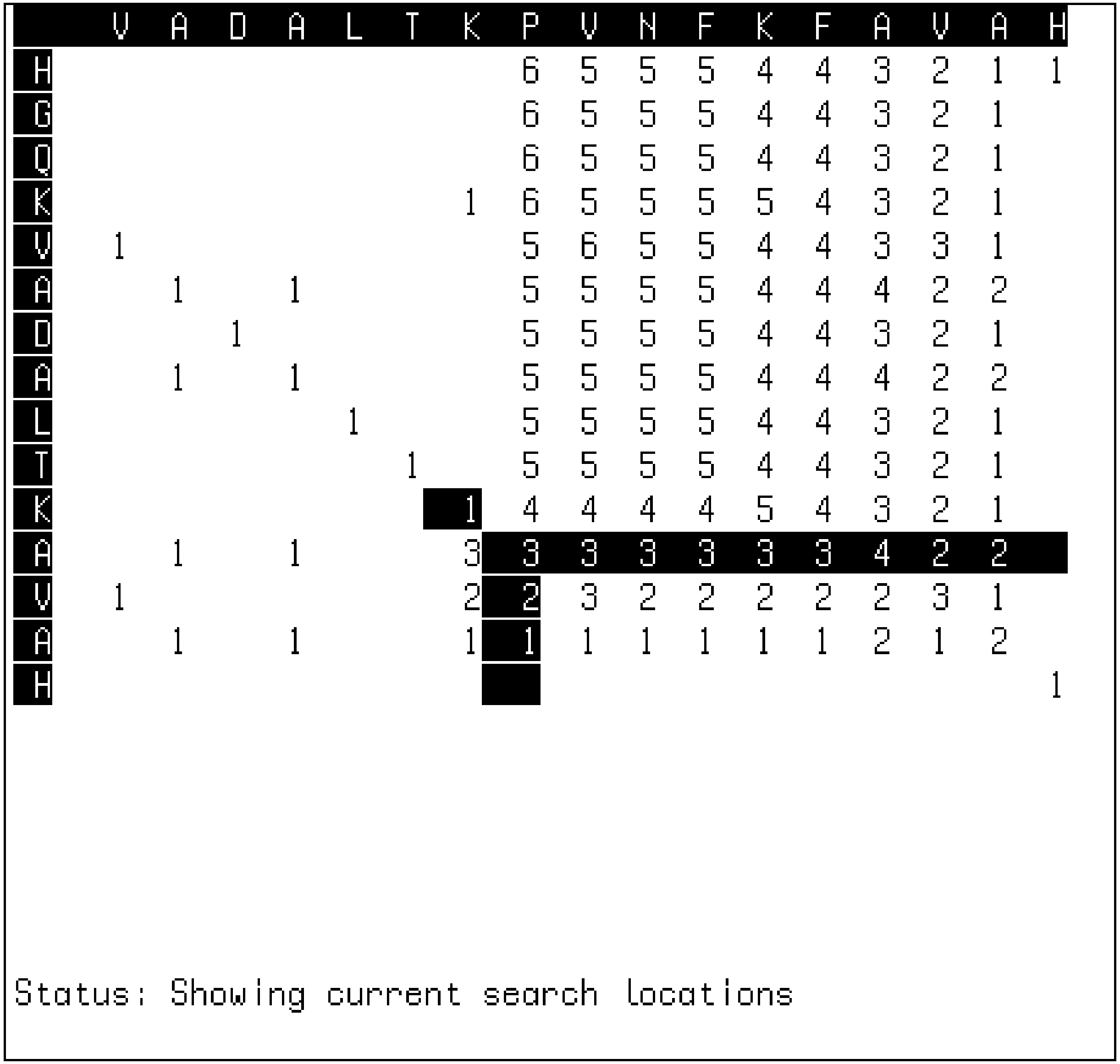
Status: Showing maximum found in search locations



	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

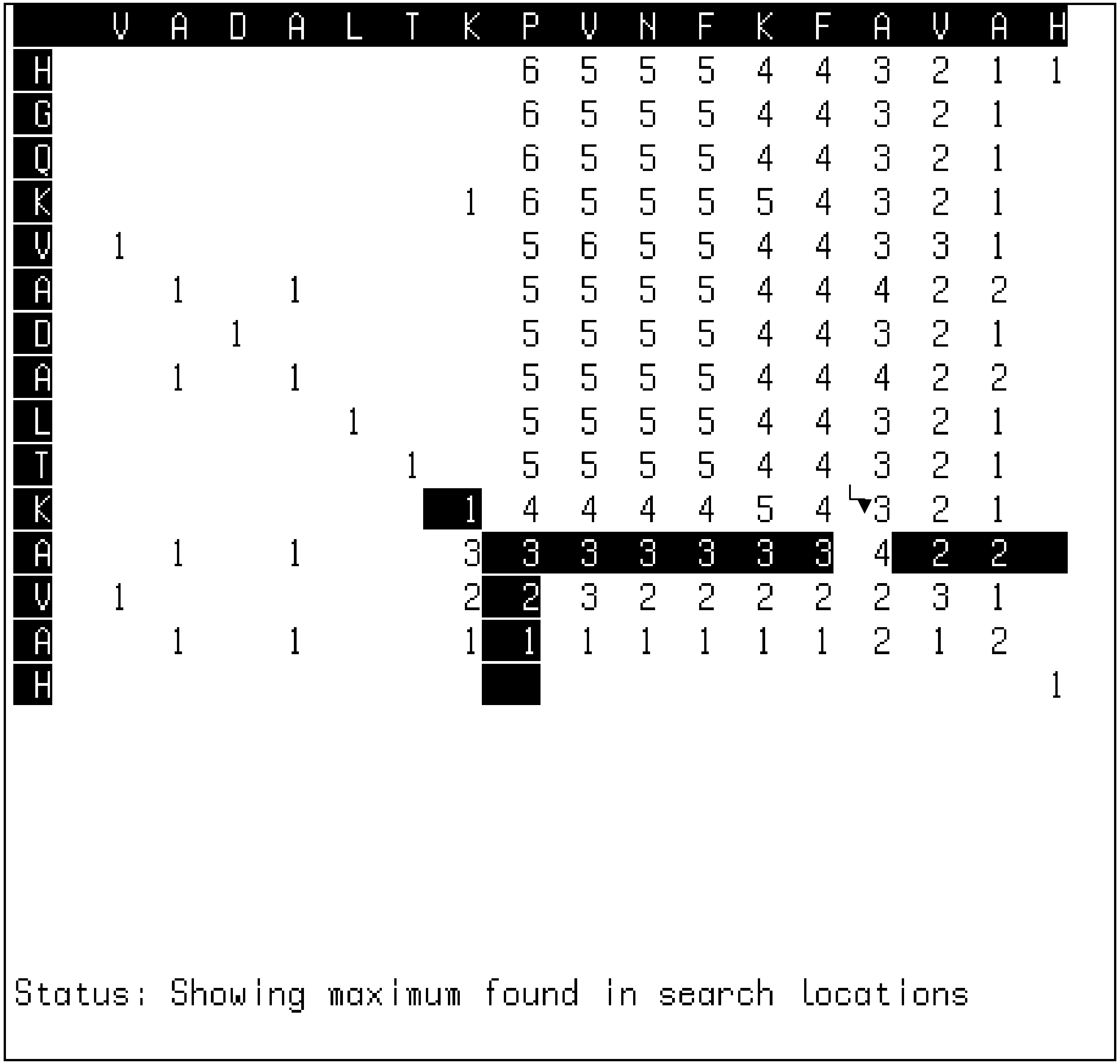
Robert F. Murphy:



	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

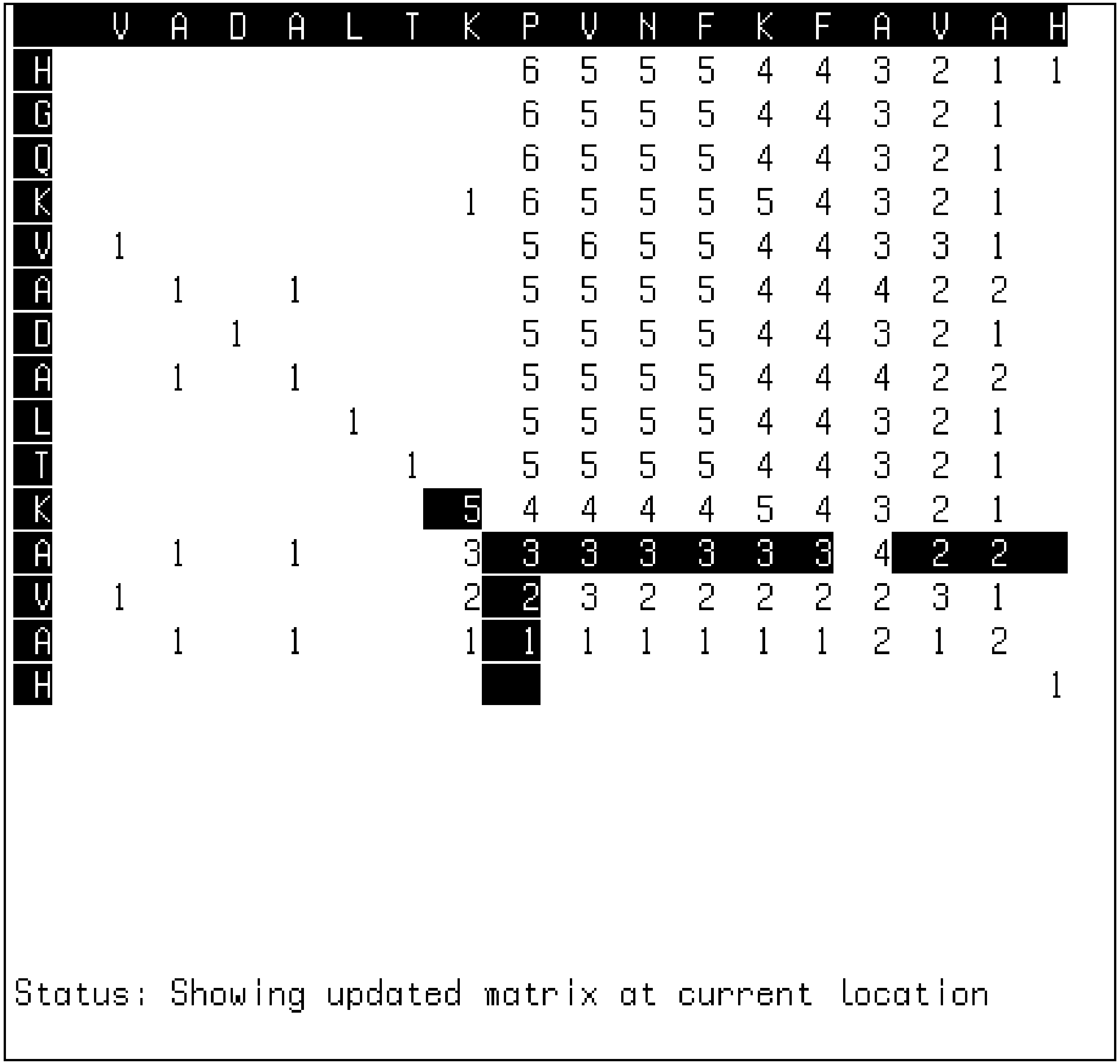
Status: Showing current search locations

Robert F. Murphy:



	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

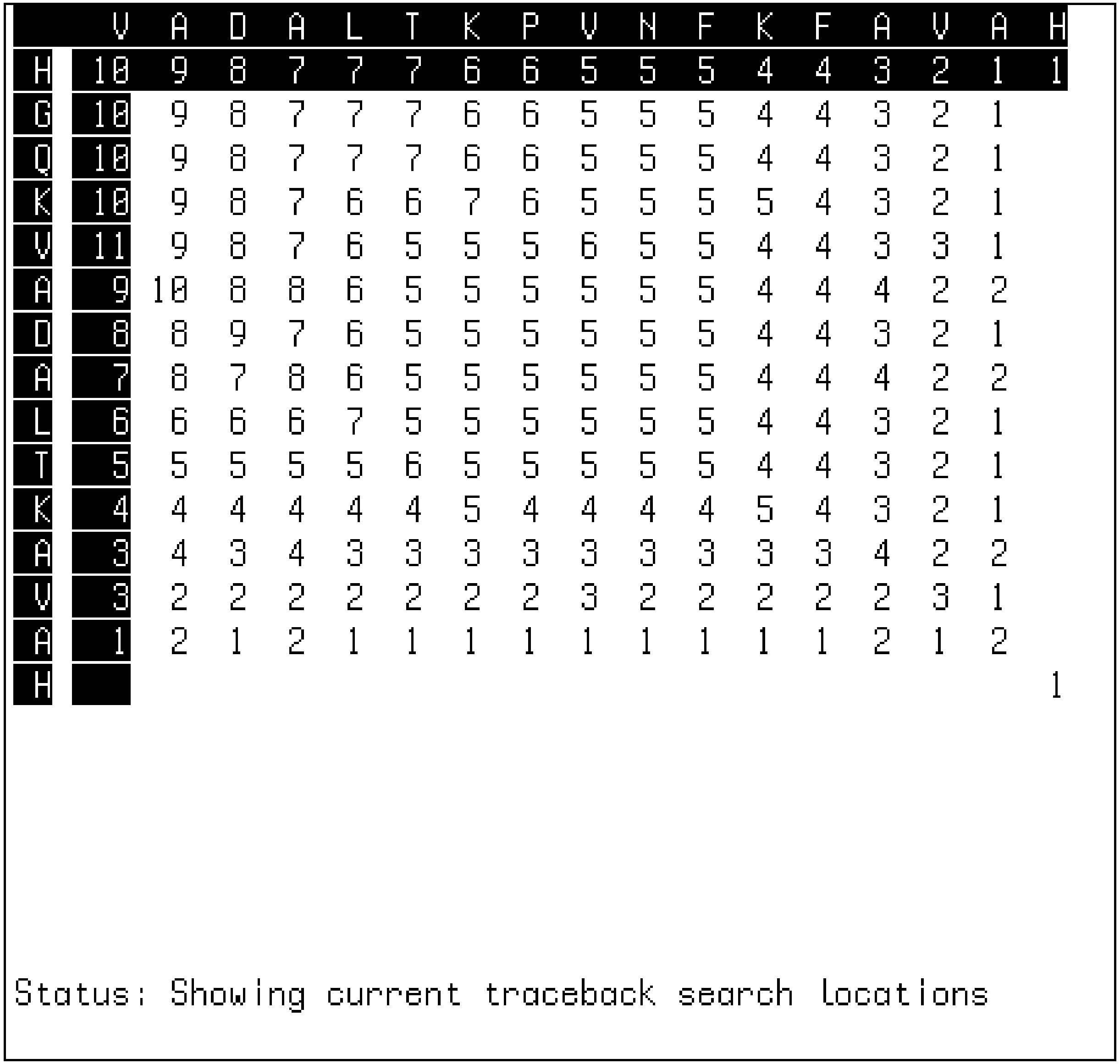
Status: Showing maximum found in search locations



	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

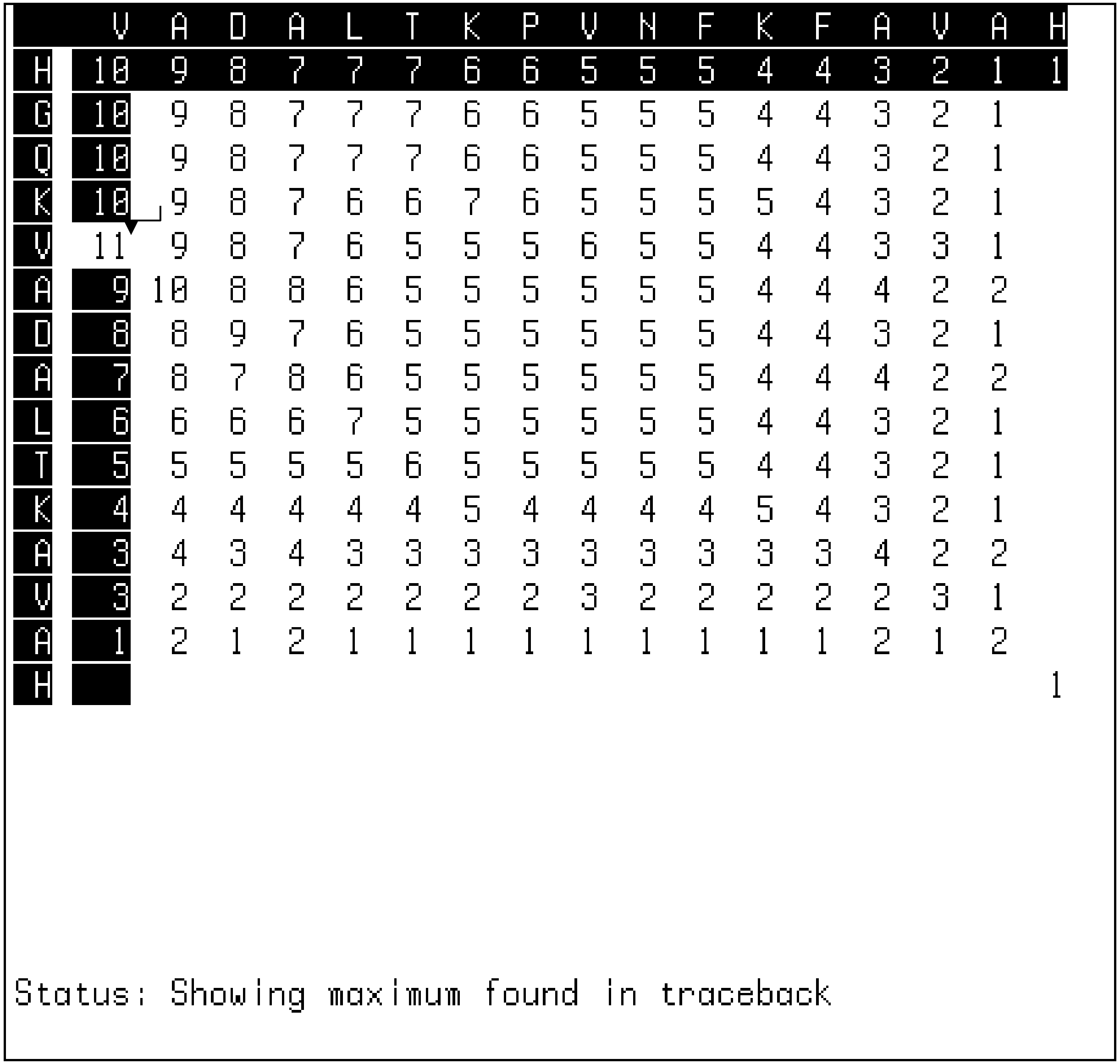
Robert F. Murphy:



	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

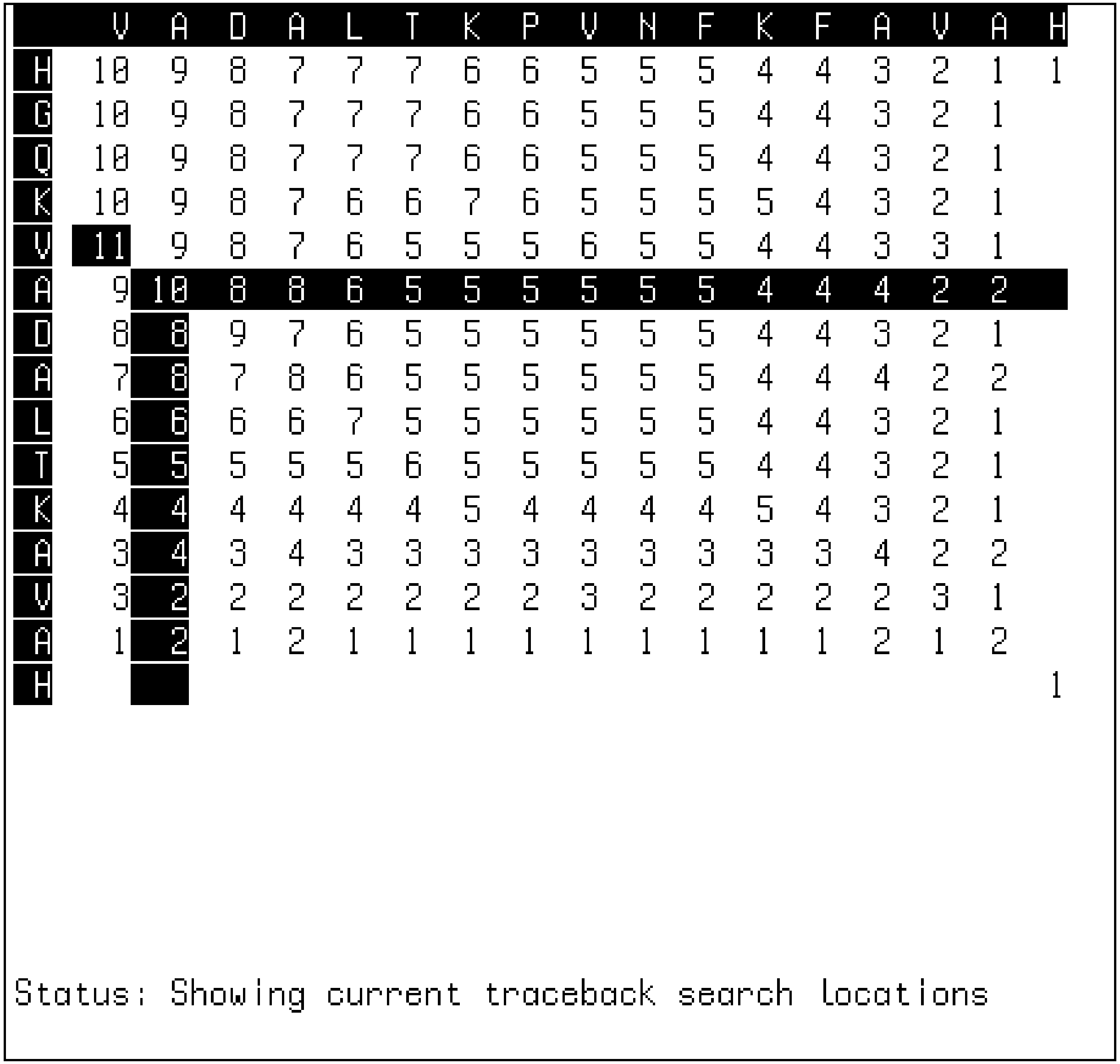
Status: Showing current traceback search locations

Robert F. Murphy:



	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

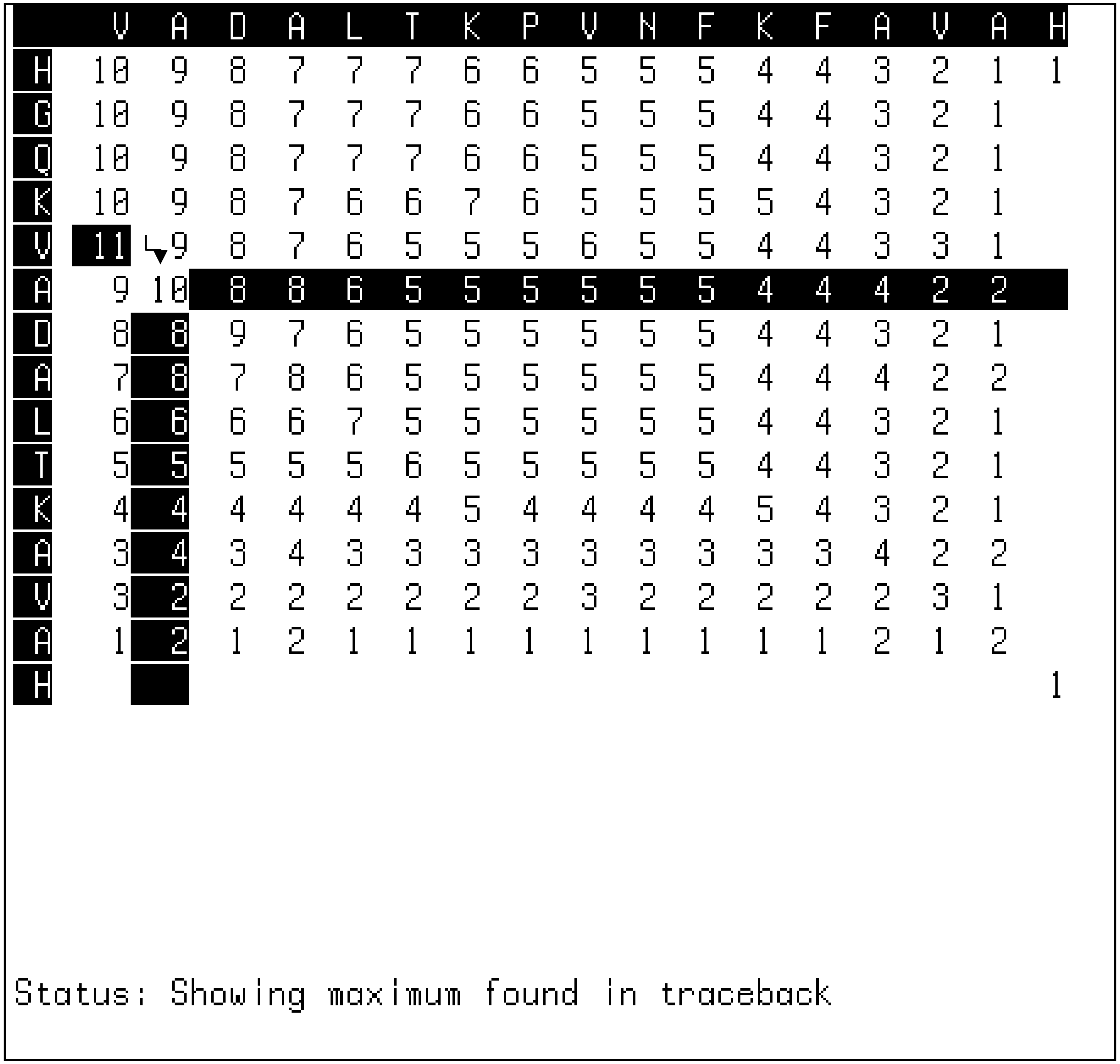


	V	A	D	A	L	T	K	P	V	N	F	K	F	A	V	A	H
H	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	1
G	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
Q	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
K	10	9	8	7	6	6	7	6	5	5	5	5	4	3	2	1	
V	11	9	8	7	6	5	5	5	6	5	5	4	4	3	3	1	
A	9	10	8	8	6	5	5	5	5	5	5	4	4	4	2	2	
D	8	8	9	7	6	5	5	5	5	5	5	4	4	3	2	1	
A	7	8	7	8	6	5	5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

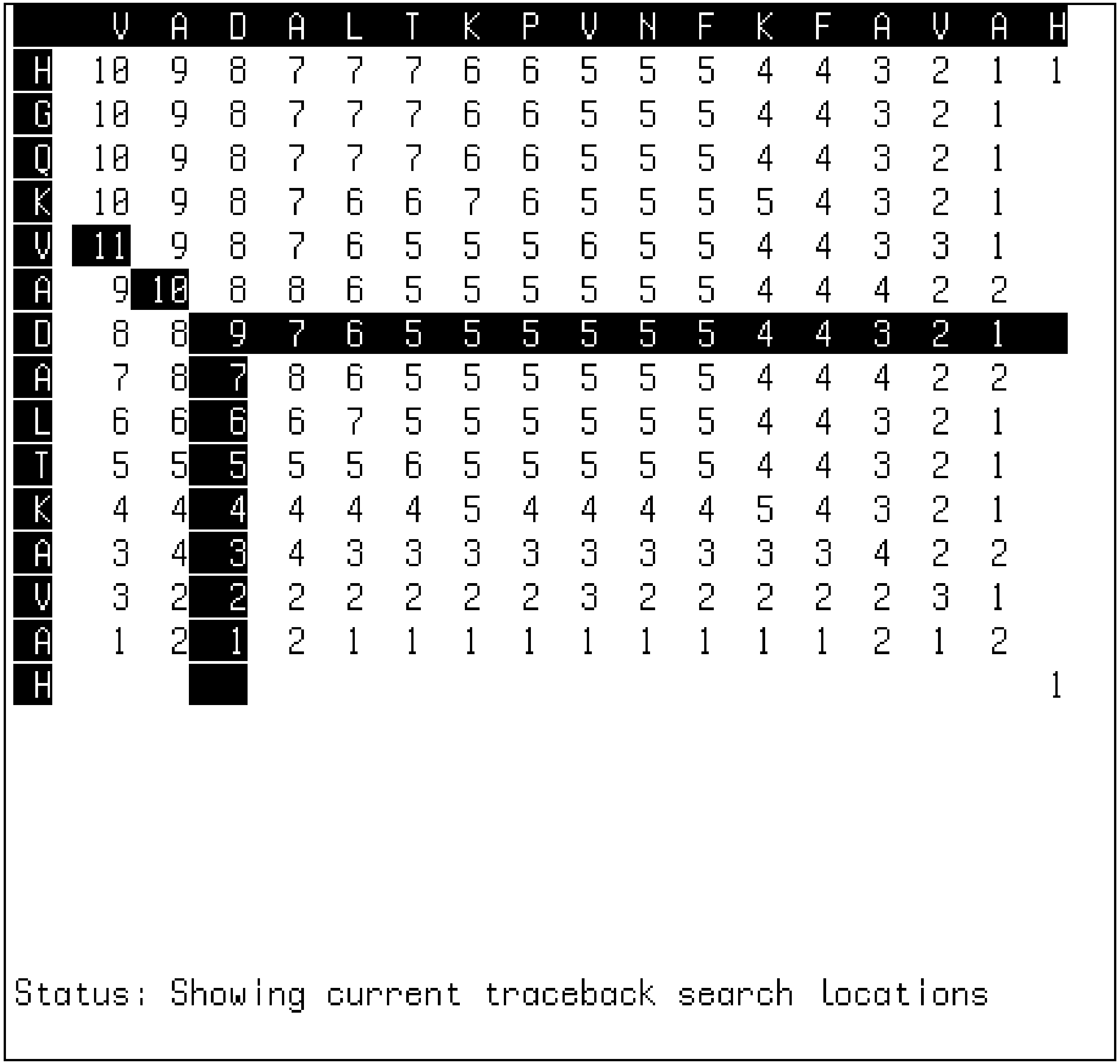
```
----V

HGQKV
```

Robert F. Murphy:



	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

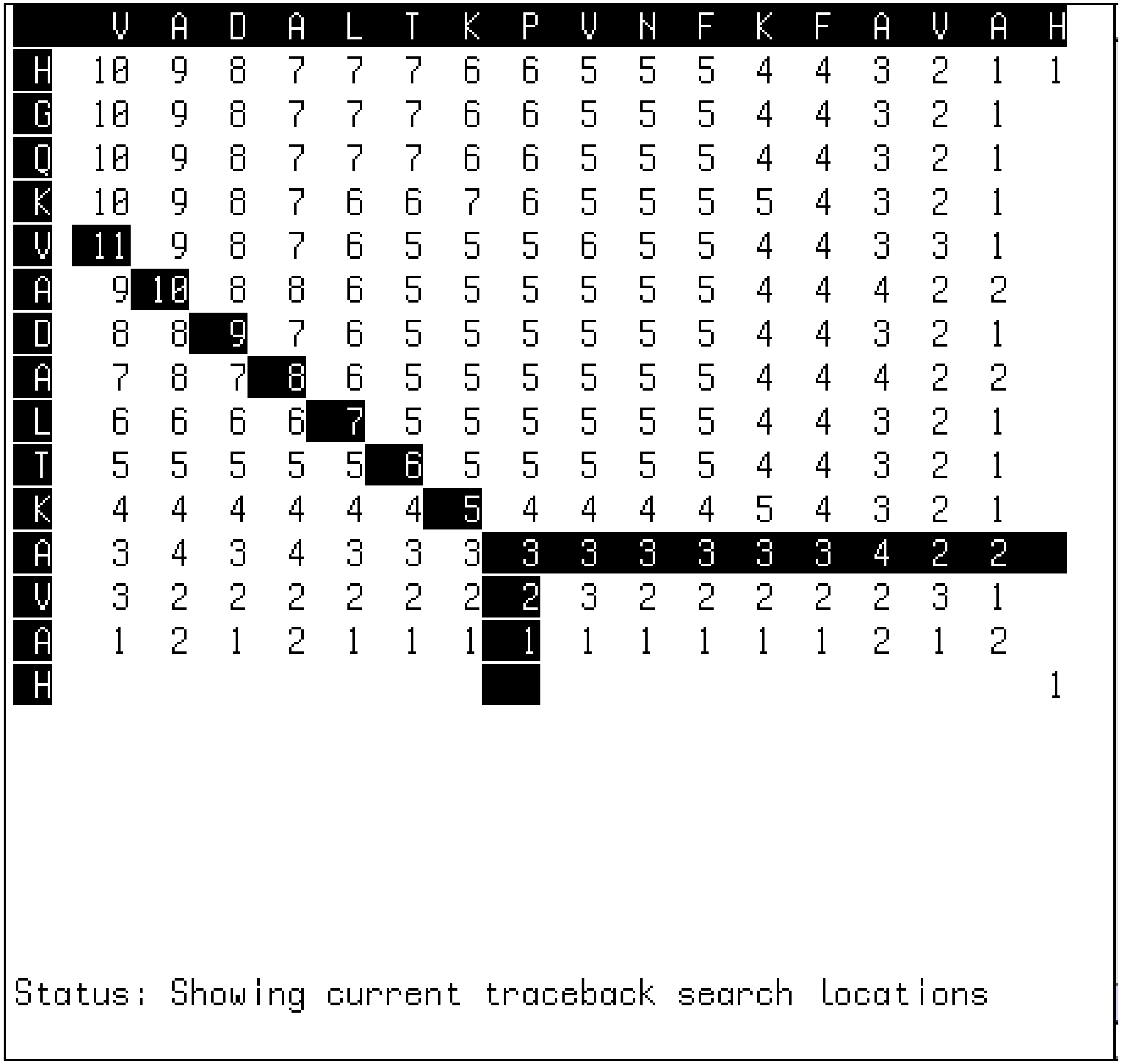


	V	A	D	A	L	T	K	P	V	N	F	K	F	A	V	A	H
H	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	1
G	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
Q	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
K	10	9	8	7	6	6	7	6	5	5	5	5	4	3	2	1	
V	11	9	8	7	6	5	5	5	6	5	5	4	4	3	3	1	
A	9	10	8	8	6	5	5	5	5	5	5	4	4	4	2	2	
D	8	8	9	7	6	5	5	5	5	5	5	4	4	3	2	1	
A	7	8	7	8	6	5	5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

```
----VA

HGQKVA
```

Robert F. Murphy:

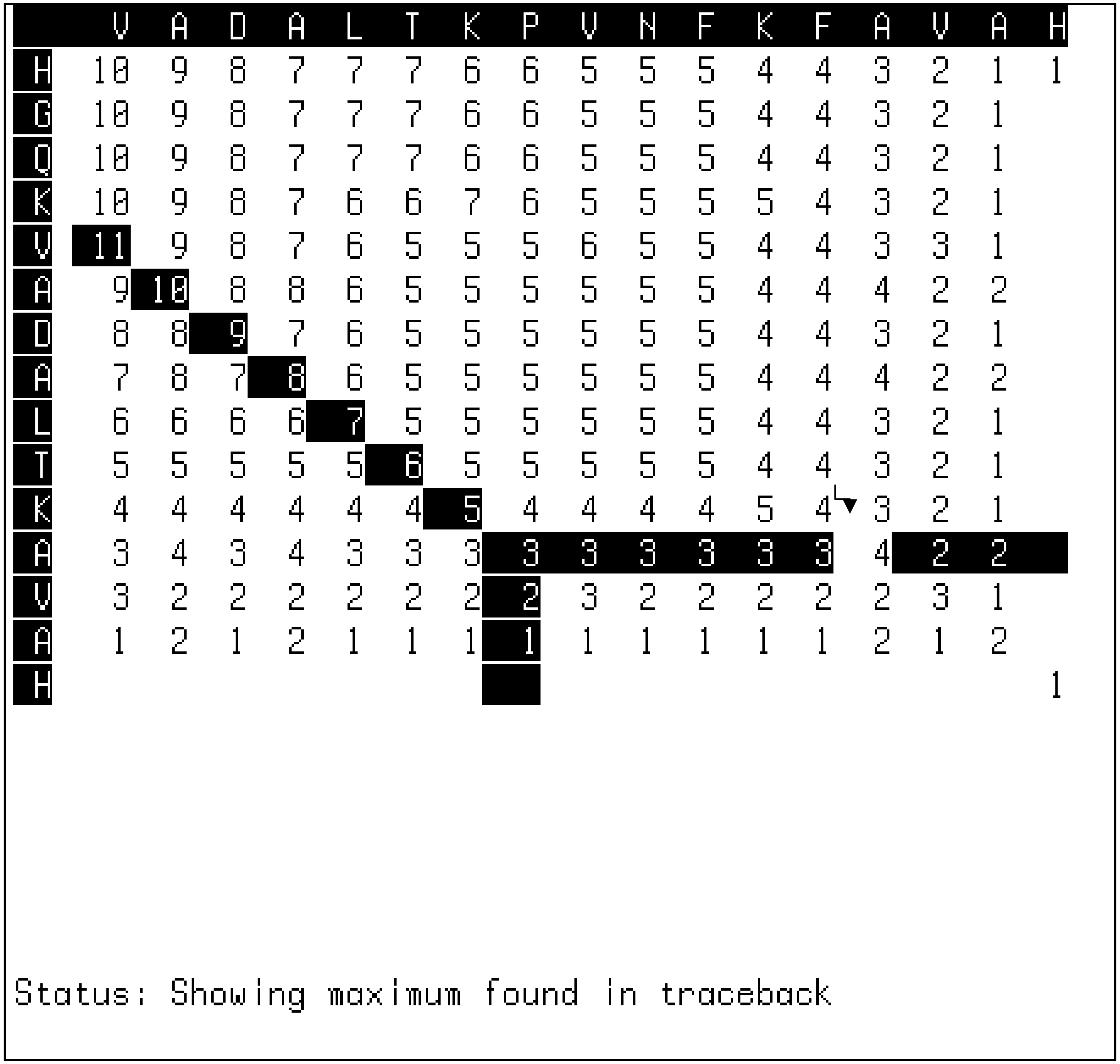


	V	A	D	A	L	T	K	P	V	N	F	K	F	A	V	A	H
H	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	1
G	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
Q	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
K	10	9	8	7	6	6	7	6	5	5	5	5	4	3	2	1	
V	11	9	8	7	6	5	5	5	6	5	5	4	4	3	3	1	
A	9	10	8	8	6	5	5	5	5	5	5	4	4	4	2	2	
D	8	8	9	7	6	5	5	5	5	5	5	4	4	3	2	1	
A	7	8	7	8	6	5	5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

Robert F. Murphy:



	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

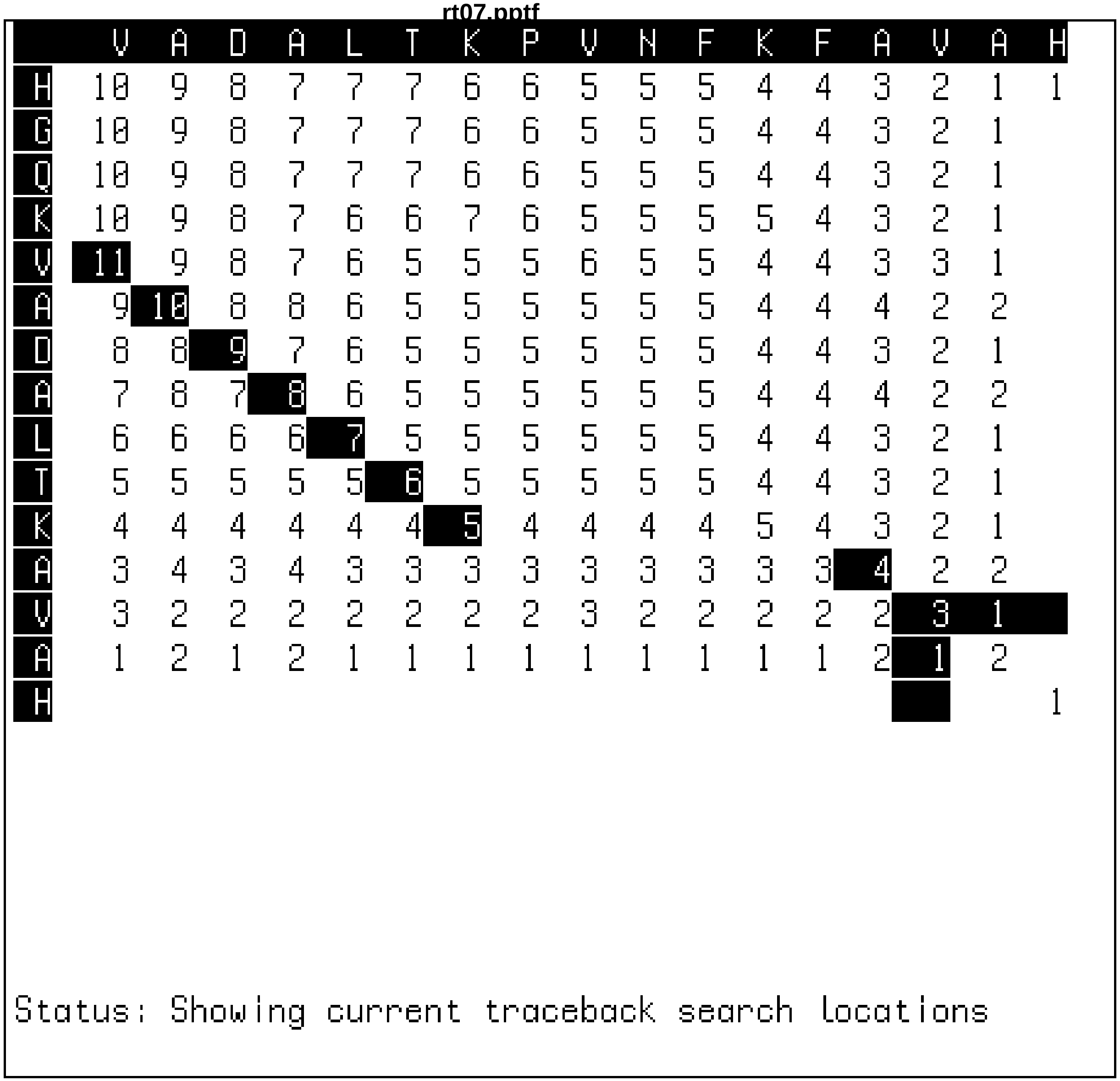
Status: Showing maximum found in traceback

----VADALTK

HGQKVADALTK

Robert F. Murphy:

Robert F. Murphy: http://www.cmu.edu/bio/education/courses/03310/LectureNotes/LecturesPart07.pptf



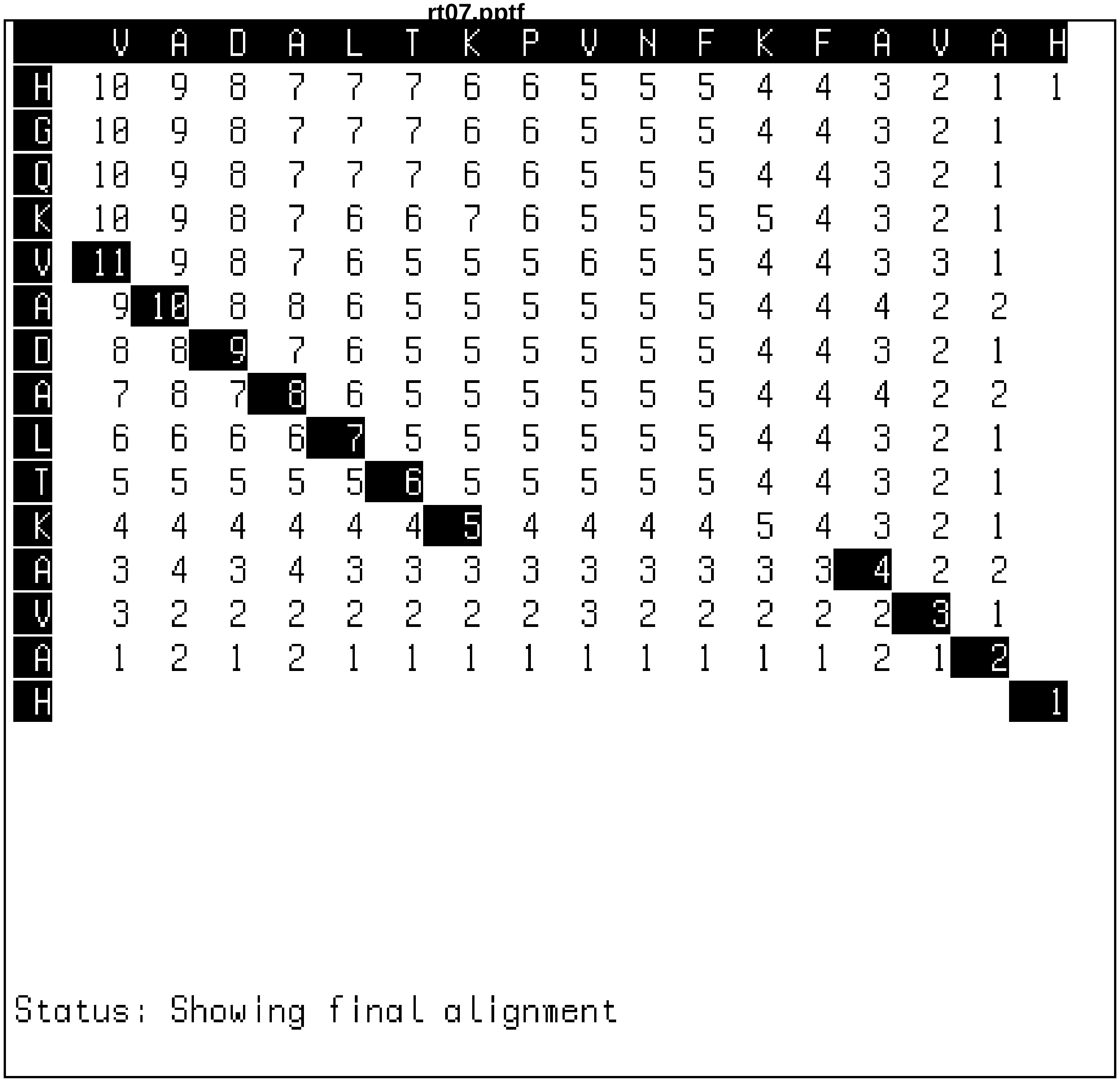
	V	A	D	A	L	T	K	P	V	N	F	K	F	A	V	A	H
H	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	1
G	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
Q	10	9	8	7	7	7	6	6	5	5	5	4	4	3	2	1	
K	10	9	8	7	6	6	7	6	5	5	5	5	4	3	2	1	
V	11	9	8	7	6	5	5	5	6	5	5	4	4	3	3	1	
A	9	10	8	8	6	5	5	5	5	5	5	4	4	4	2	2	
D	8	8	9	7	6	5	5	5	5	5	5	4	4	3	2	1	
A	7	8	7	8	6	5	5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

Robert F. Murphy: http://www.cmu.edu/bio/education/courses/03310/LectureNotes/LecturesPart07.pptf



	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

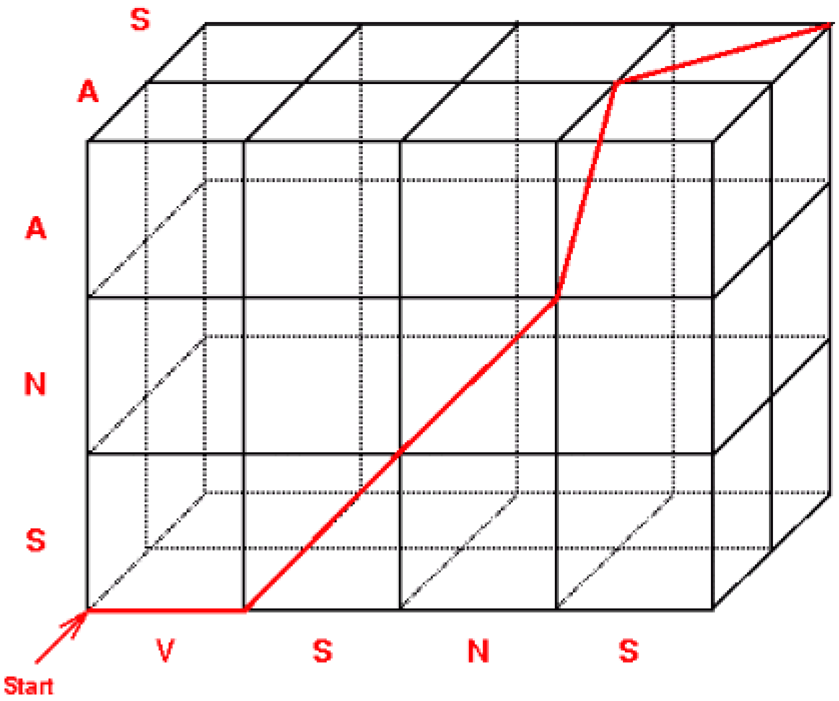
Robert F. Murphy:

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

Robert F. Murphy:

Multiple sequence alignment

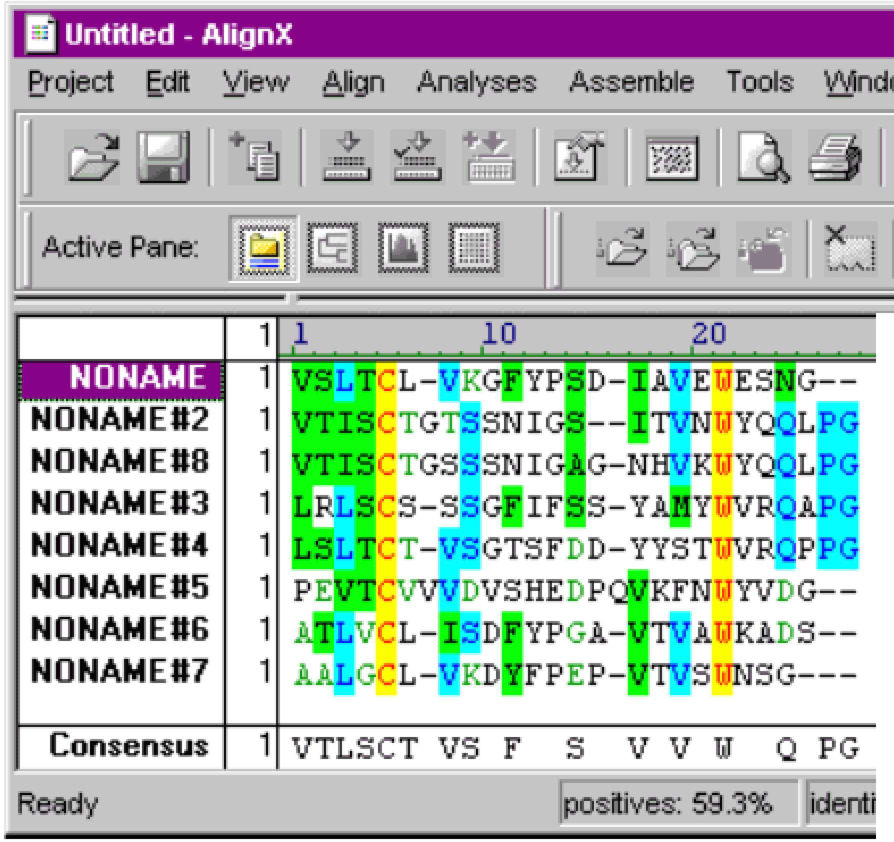


```
V S N _ S
_ S N A _
_ _ _ A S
```

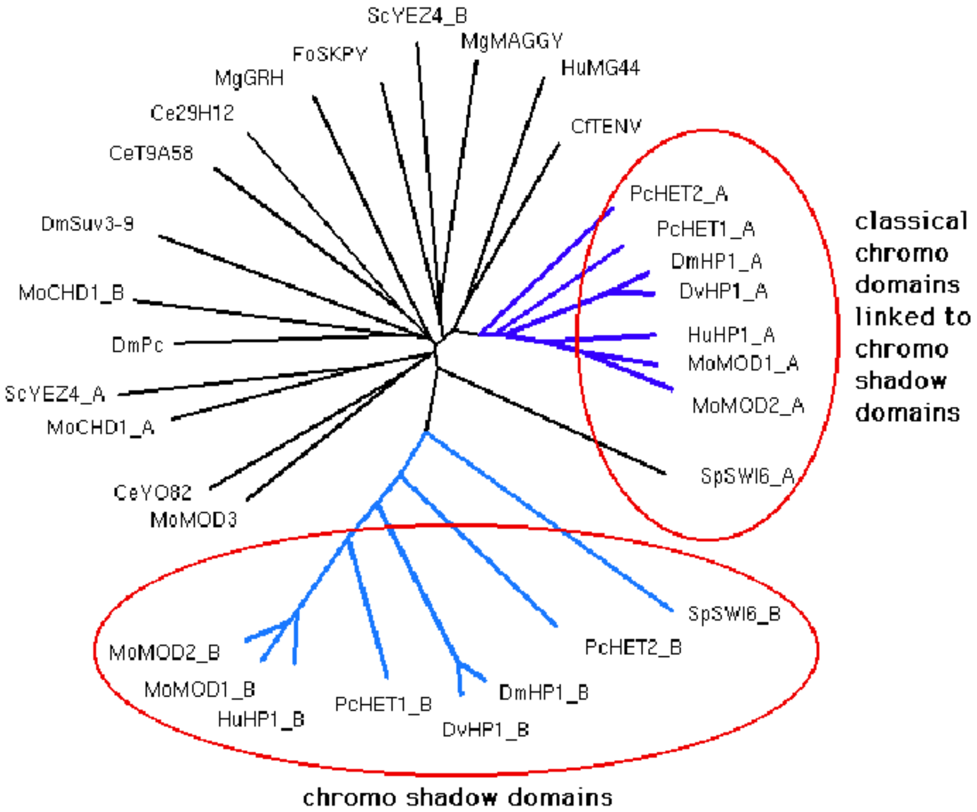
Figure source: http://www.techfak.uni-bielefeld.de/bcd/Curric/MulAli/node2.html#SECTION00020000000000000000

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

Multiple sequence alignment for phylogenetic trees

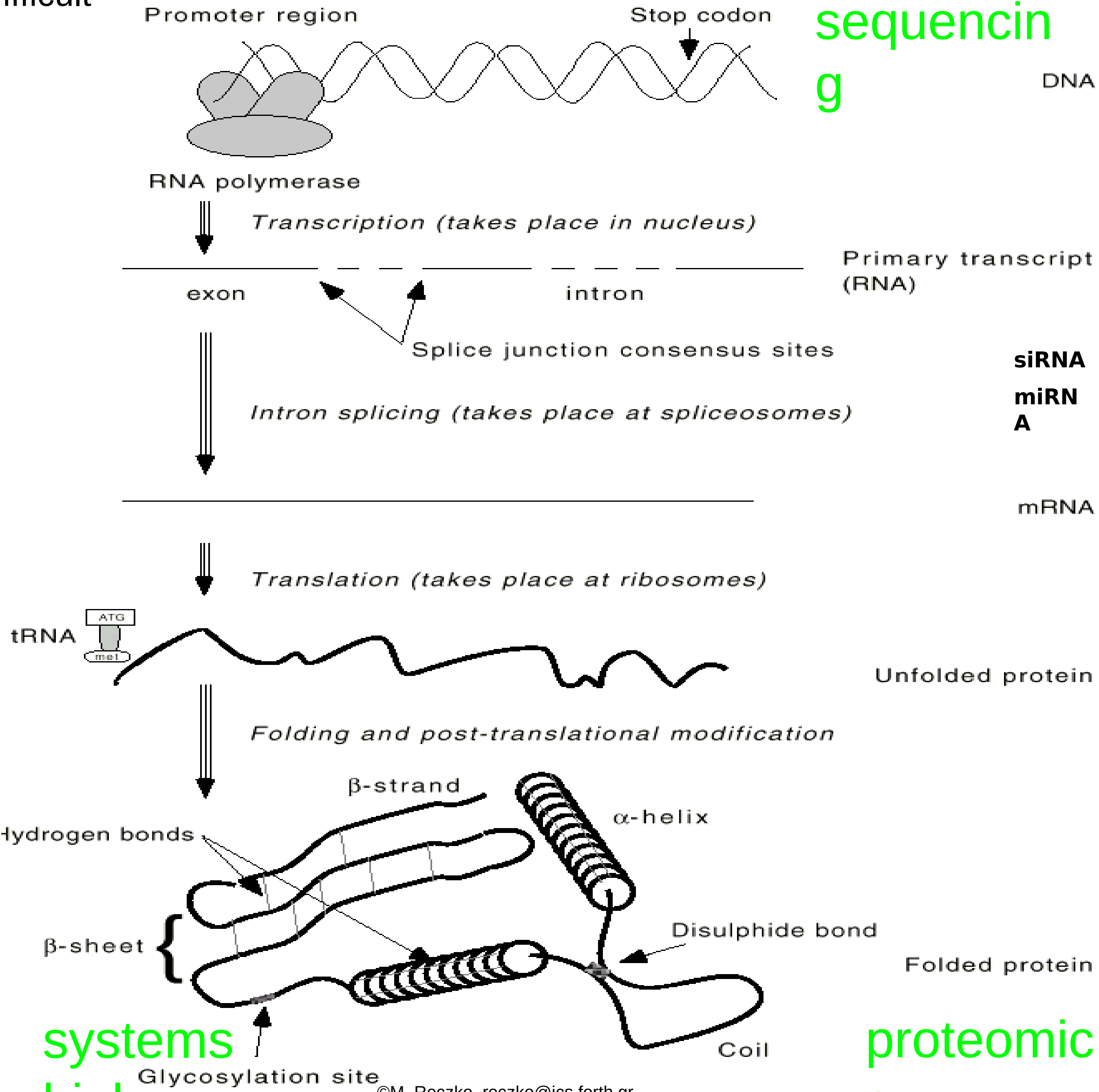


```
                1 1        10        20
NONAME          1 VSLTCL-VKGFYPSD-IAVEWESNG--
NONAME#2        1 VTISCTGTSSNIGS--ITVNWYQQLPG
NONAME#8        1 VTISCTGSSSNIGAG-NHVKWYQQLPG
NONAME#3        1 LRLSCS-SSGFIFSS-YAMYWVRQAPG
NONAME#4        1 LSLTCT-VSGTSFDD-YYSTWVRQPPG
NONAME#5        1 PEVTCVVVDVSHEDPQVKFNWYVDG--
NONAME#6        1 ATLVCL-ISDFYPGA-VTVAWKADS--
NONAME#7        1 AALGCL-VKDYFPEP-VTVSWNSG---

Consensus       1 VTLSCT VS F  S  V V W  Q PG

Ready                          positives: 59.3%   identi
```



Modelling tasks:	Difficulty
Promoter	3
Stop	2
1:1	
splice sites	2
exon/intron	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

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


Promoter region
Stop codon
DNA
RNA polymerase
Transcription (takes place in nucleus)
Primary transcript (RNA)
exon
intron
Splice junction consensus sites
Intron splicing (takes place at spliceosomes)
mRNA
Translation (takes place at ribosomes)
tRNA
ATG
met
Unfolded protein
Folding and post-translational modification
β-strand
α-helix
Hydrogen bonds
Disulphide bond
β-sheet
Folded protein
Coil
Glycosylation site

sequencing

siRNA

miRNA

systems biology

proteomics

Modelling tasks:

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

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

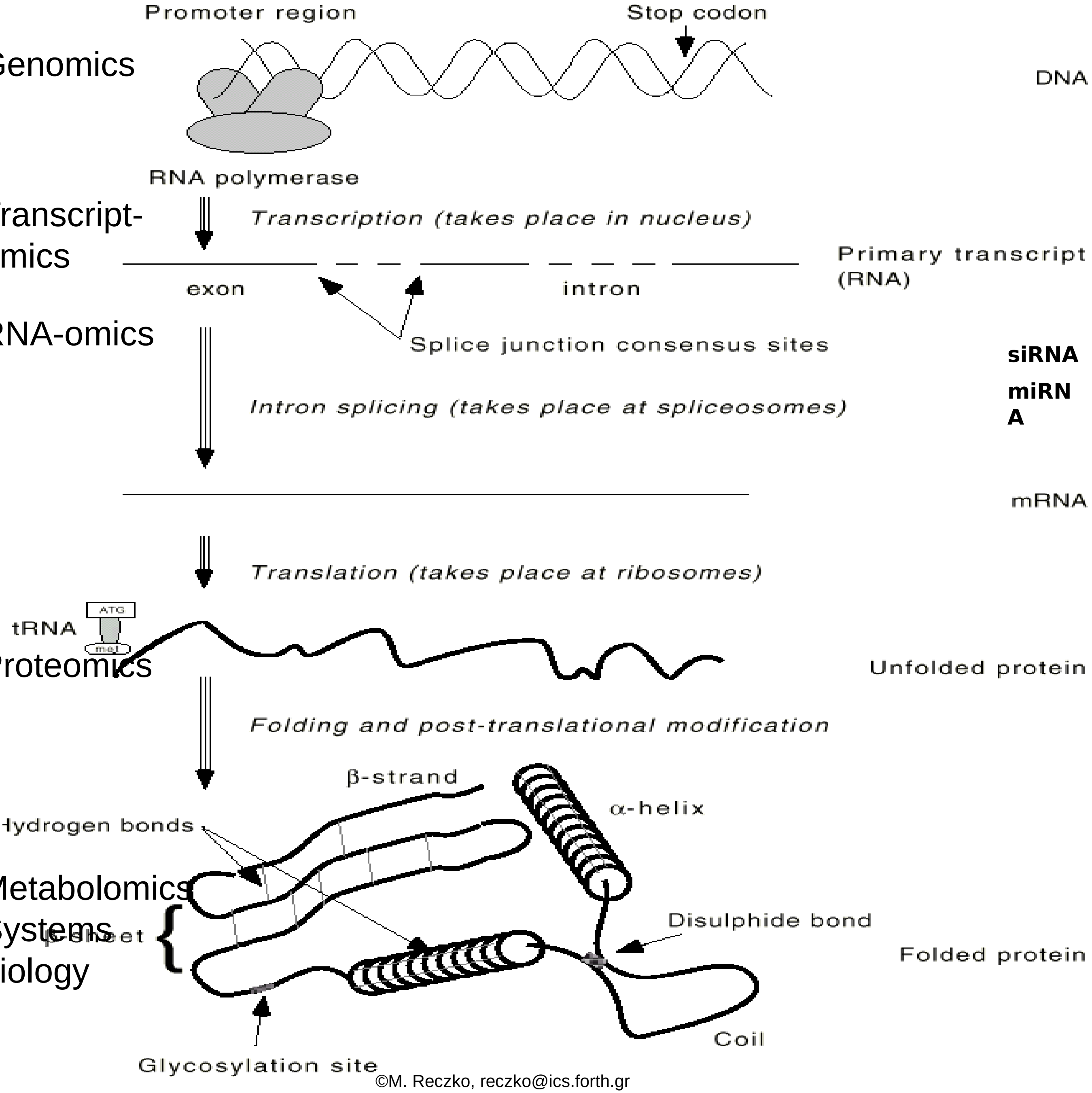
Genomics

Transcript-
omics

RNA-omics

Proteomics

Metabolomics
Systems
biology

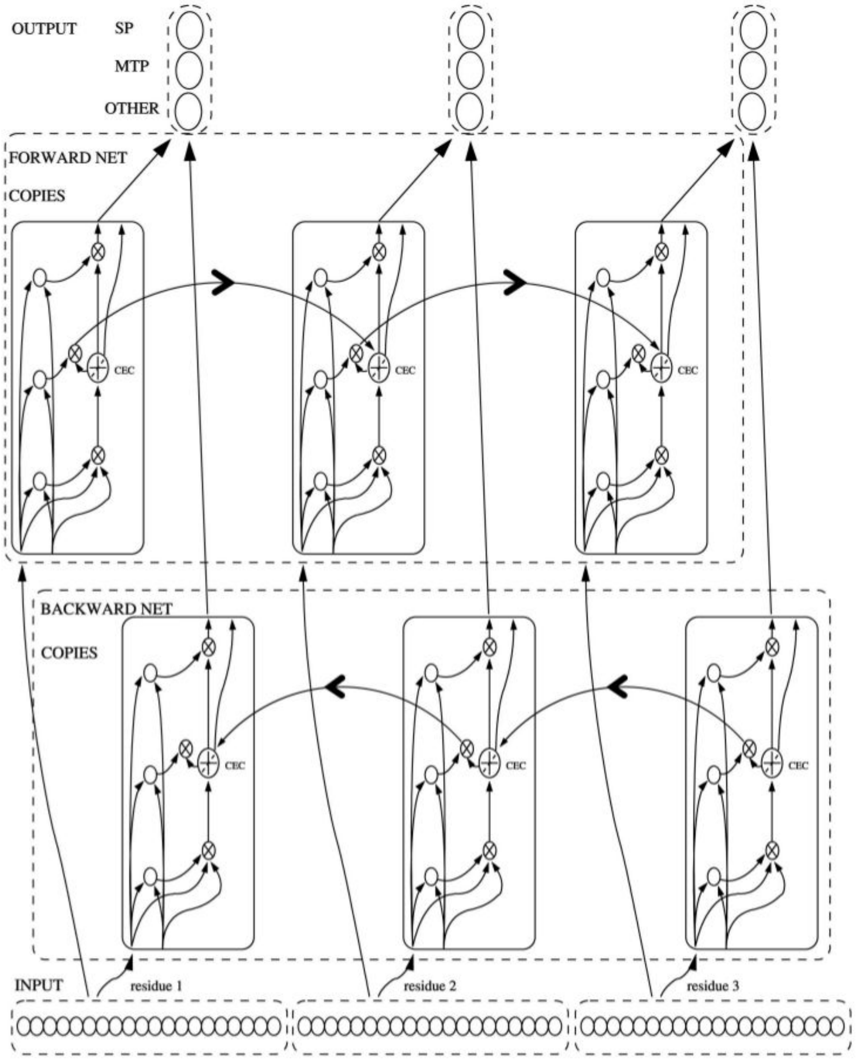

Promoter region
Stop codon
DNA
RNA polymerase
Transcription (takes place in nucleus)
Primary transcript
(RNA)
exon
intron
Splice junction consensus sites
Intron splicing (takes place at spliceosomes)
mRNA
Translation (takes place at ribosomes)
ATG
tRNA
met
Unfolded protein
Folding and post-translational modification
β-strand
α-helix
Hydrogen bonds
β-sheet
Disulphide bond
Folded protein
Coil
Glycosylation site

siRNA

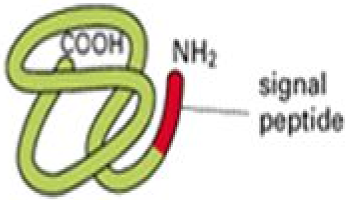
miRNA

©M. Reczko, reczko@ics.forth.gr

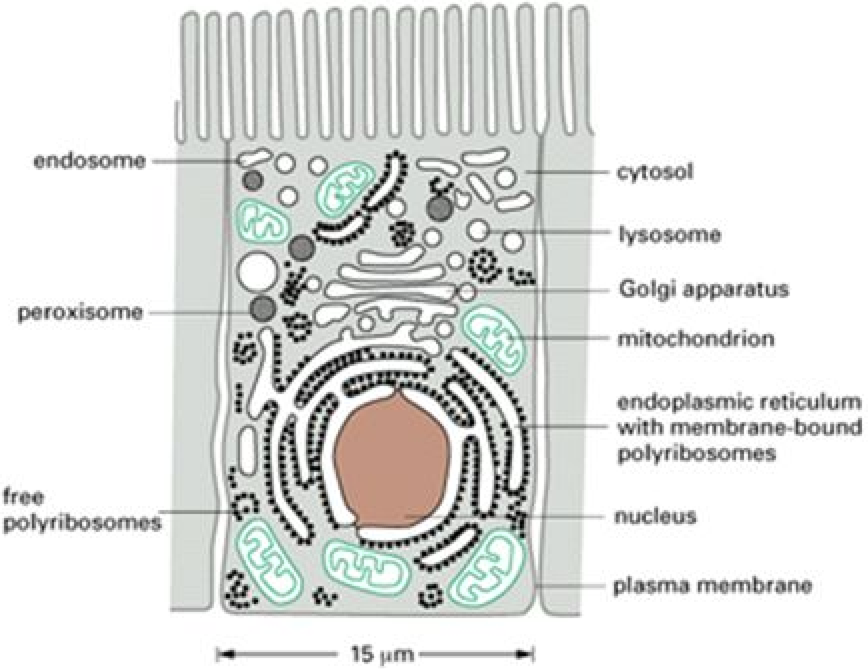
Introduction novel sequence learning algorithm (BLSTM)



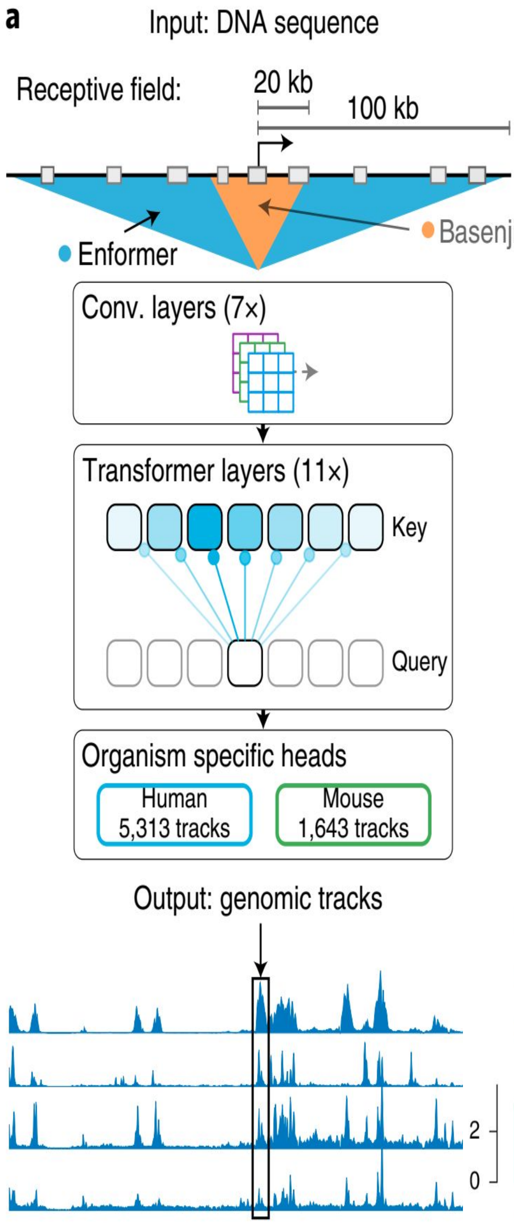
- Use start of proteinsequence to predict its compartment

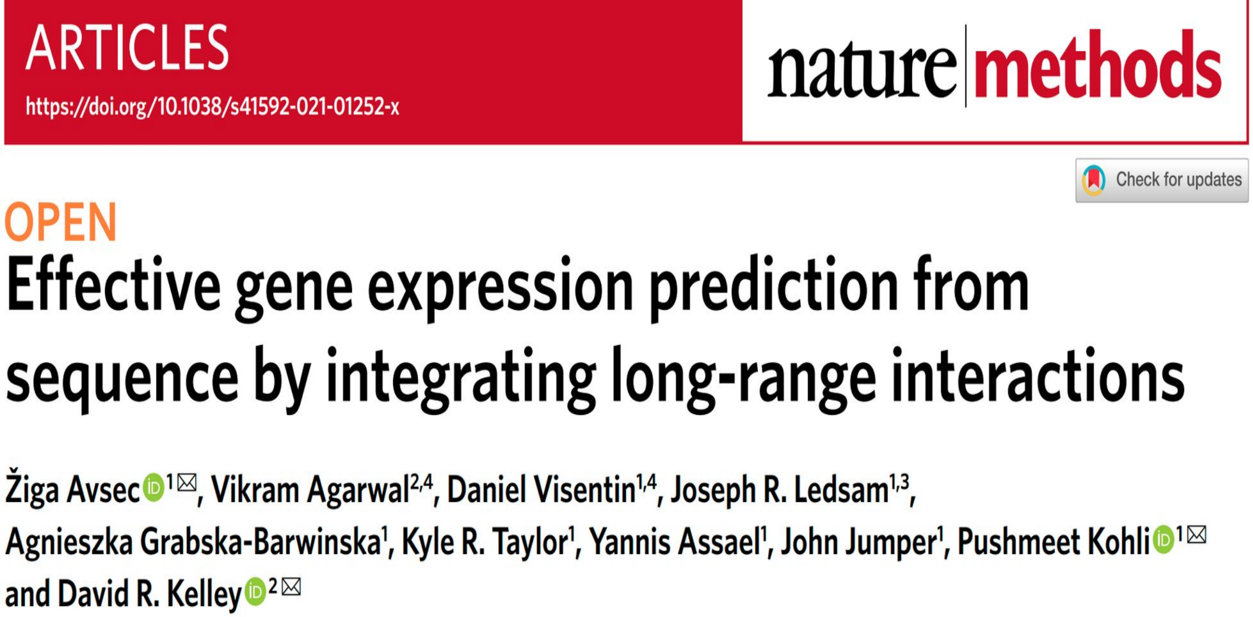


- BLSTMs precursors of transformer networks



IEEE Transactions on Computational Biology and Bioinformatics 2006, 4(3), pp.441-6.

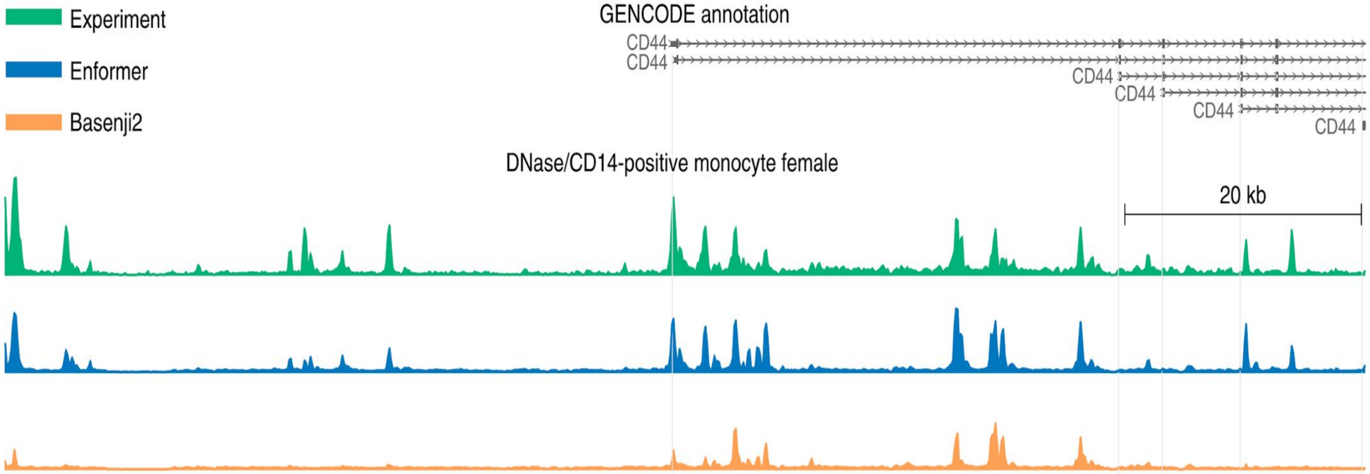



ARTICLES
https://doi.org/10.1038/s41592-021-01252-x
nature methods

OPEN

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

Žiga Avsec[1]✉, Vikram Agarwal[2,4], Daniel Visentin[1,4], Joseph R. Ledsam[1,3], Agnieszka Grabska-Barwinska[1], Kyle R. Taylor[1], Yannis Assael[1], John Jumper[1], Pushmeet Kohli[1]✉ and David R. Kelley[2]✉



DeepMind

About

Research

Impact

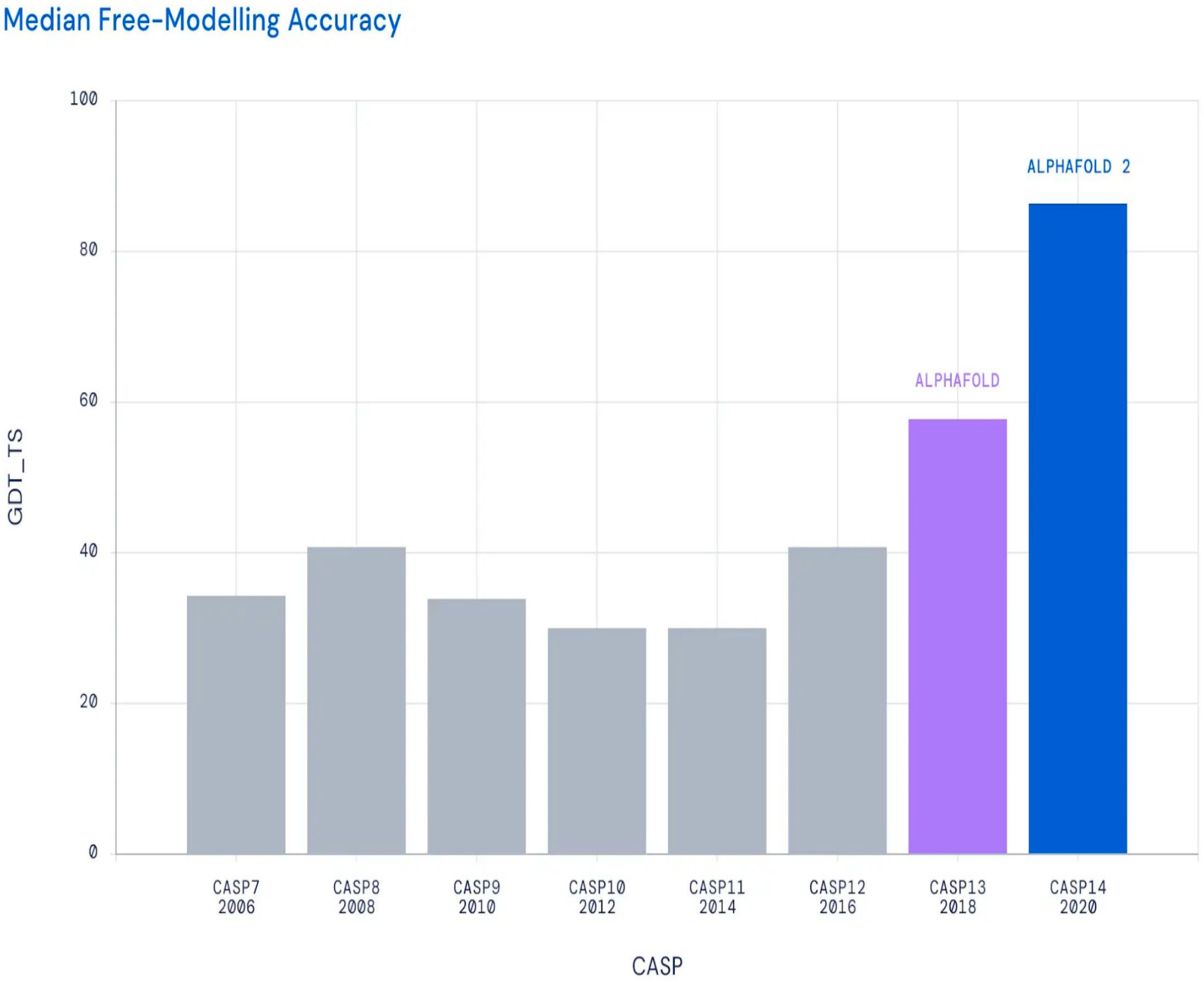
Blog

Safety & Ethics

Careers

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

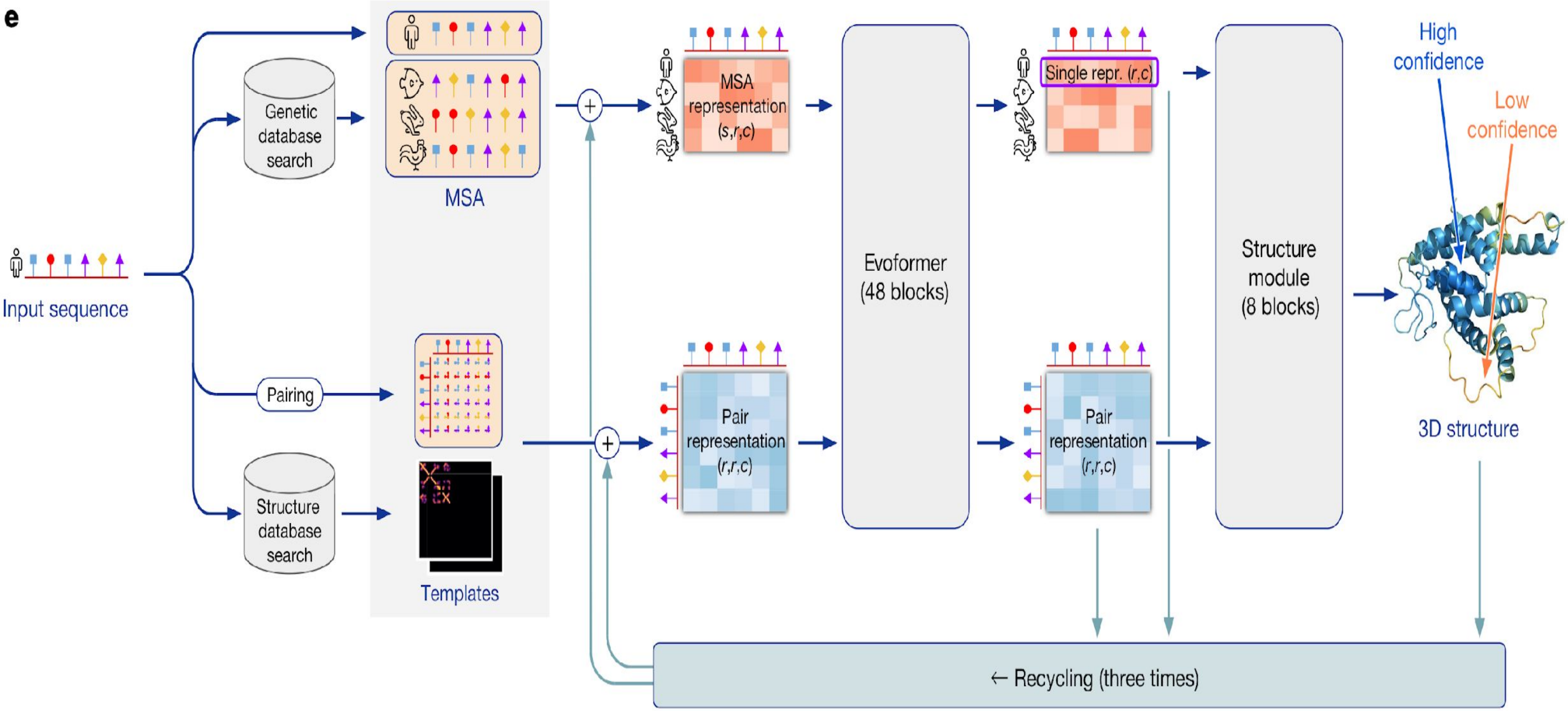
https://deepmind.com



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.

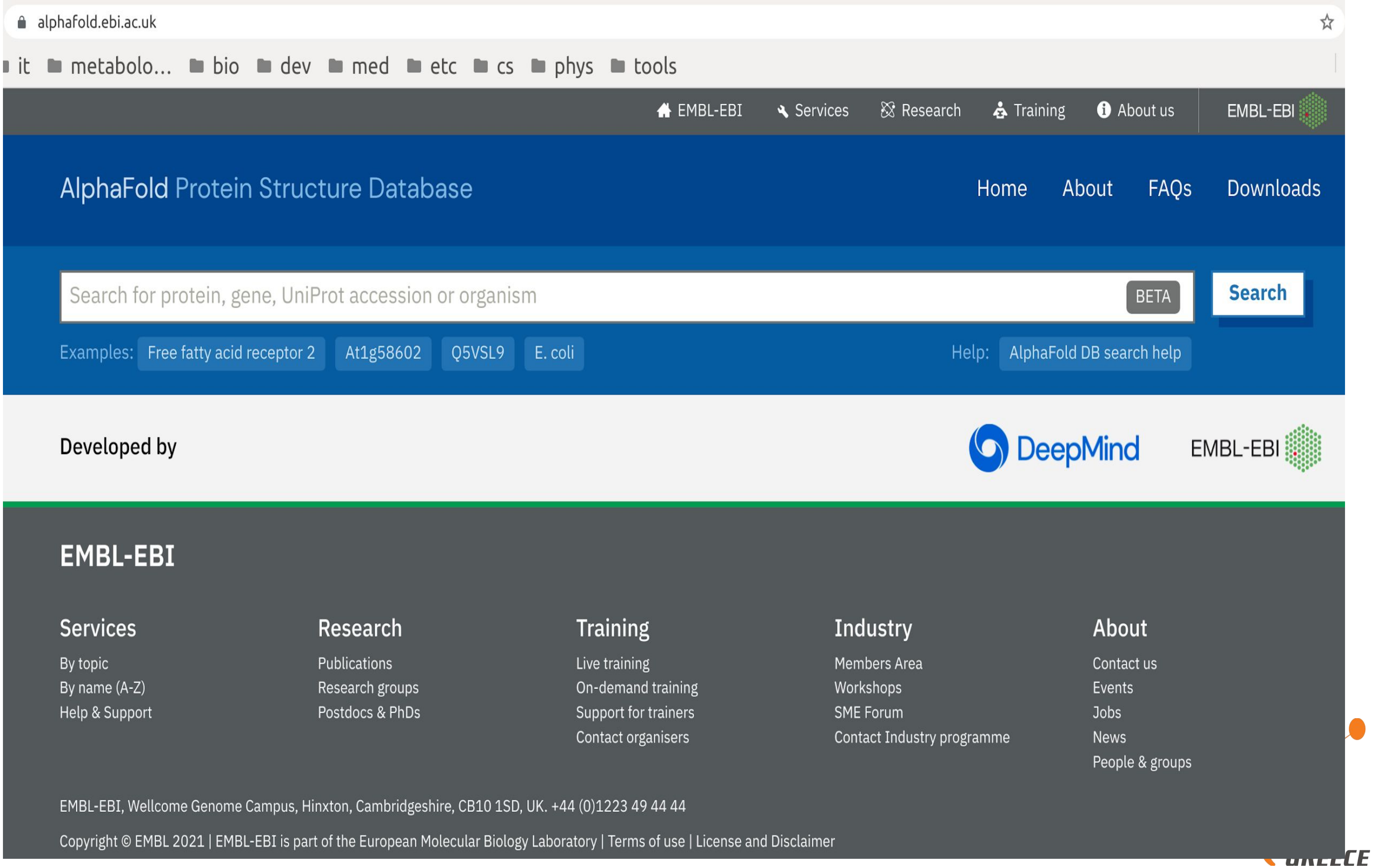
https://deepmind.com/blog/article/putting-the-power-of-alphafold-into-the-worlds-hands

AlphaFold2 architecture


e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

https://www.nature.com/articles/s41586-021-03819-2.pdf

AlphaFold2 database of predicted structures



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

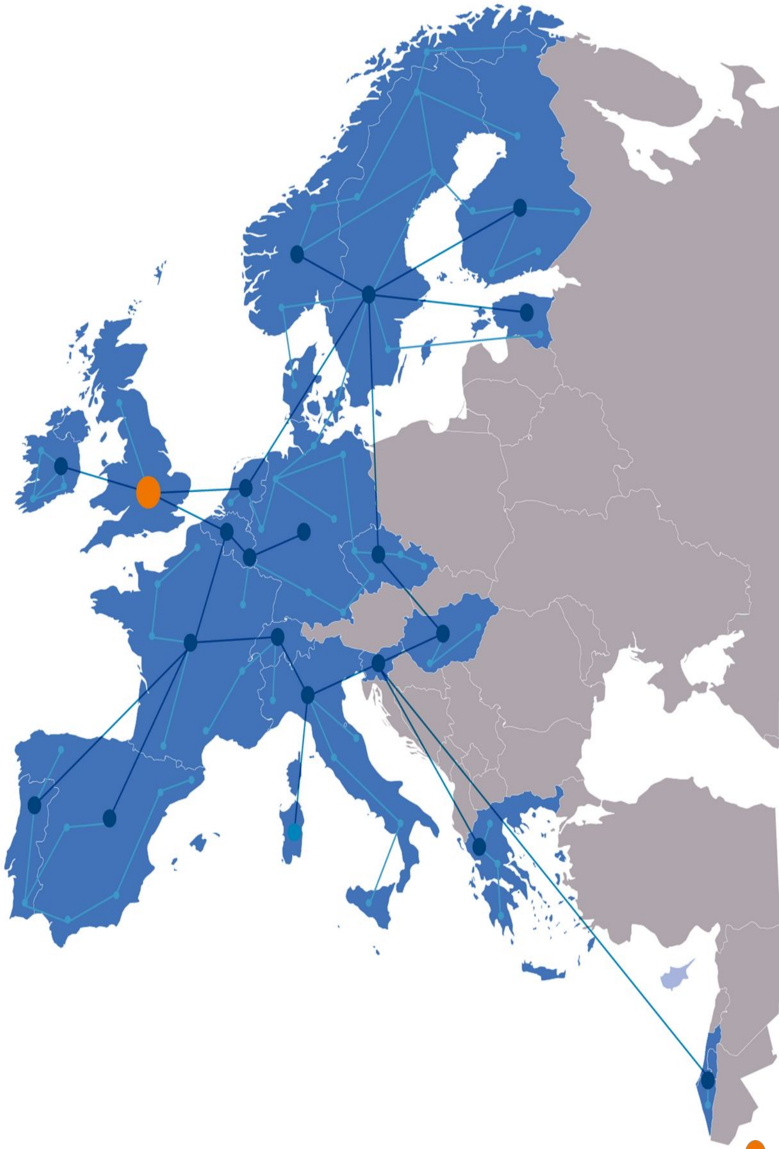
ELIXIR is an intergovernmental organisation that brings together life science resources such as databases, software tools, training materials, standards and compute resources, from across Europe.

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:

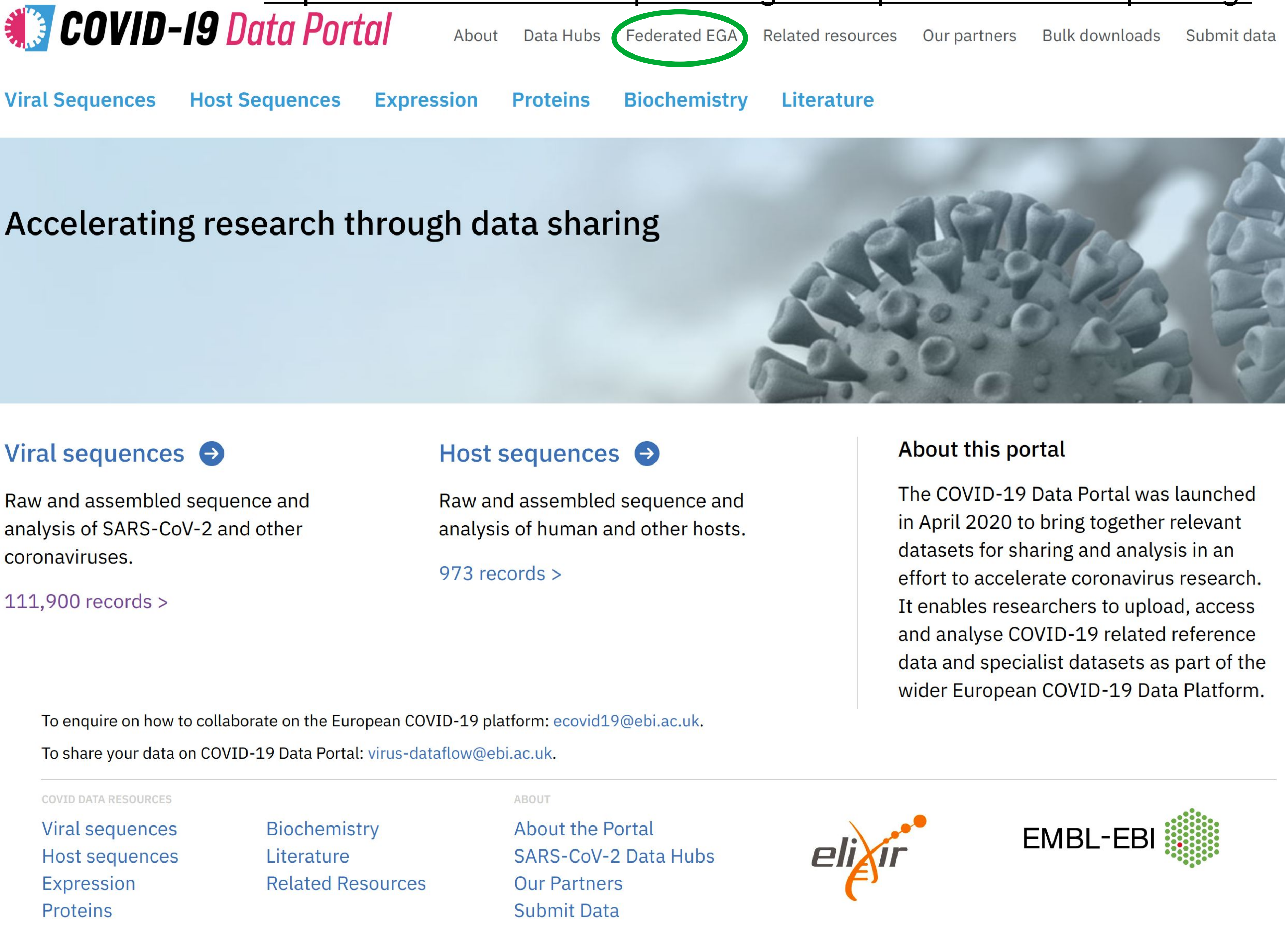
Find and share data
Exchange expertise
Agree on best practices in scientific research

Check: https://elixir-europe.org

https://elixir-greece.org



https://www.covid19dataportal.org & https://covid19dataportal.gr



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.

DATA

OPTIMISATION

MODEL

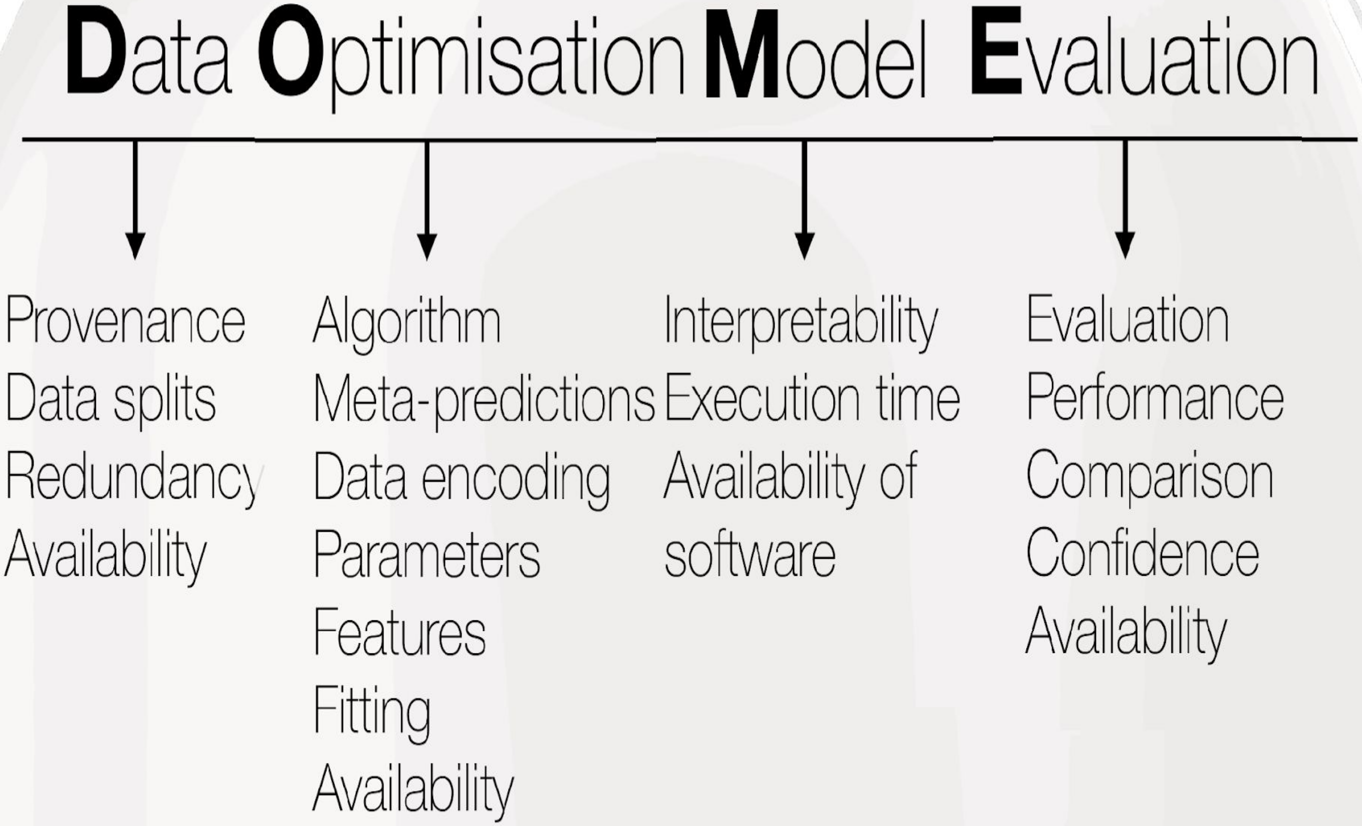
EVALUATION

MACHINE LEARNING
FOCUS GROUP

elixir

Website:
https://dome-ml.org/

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



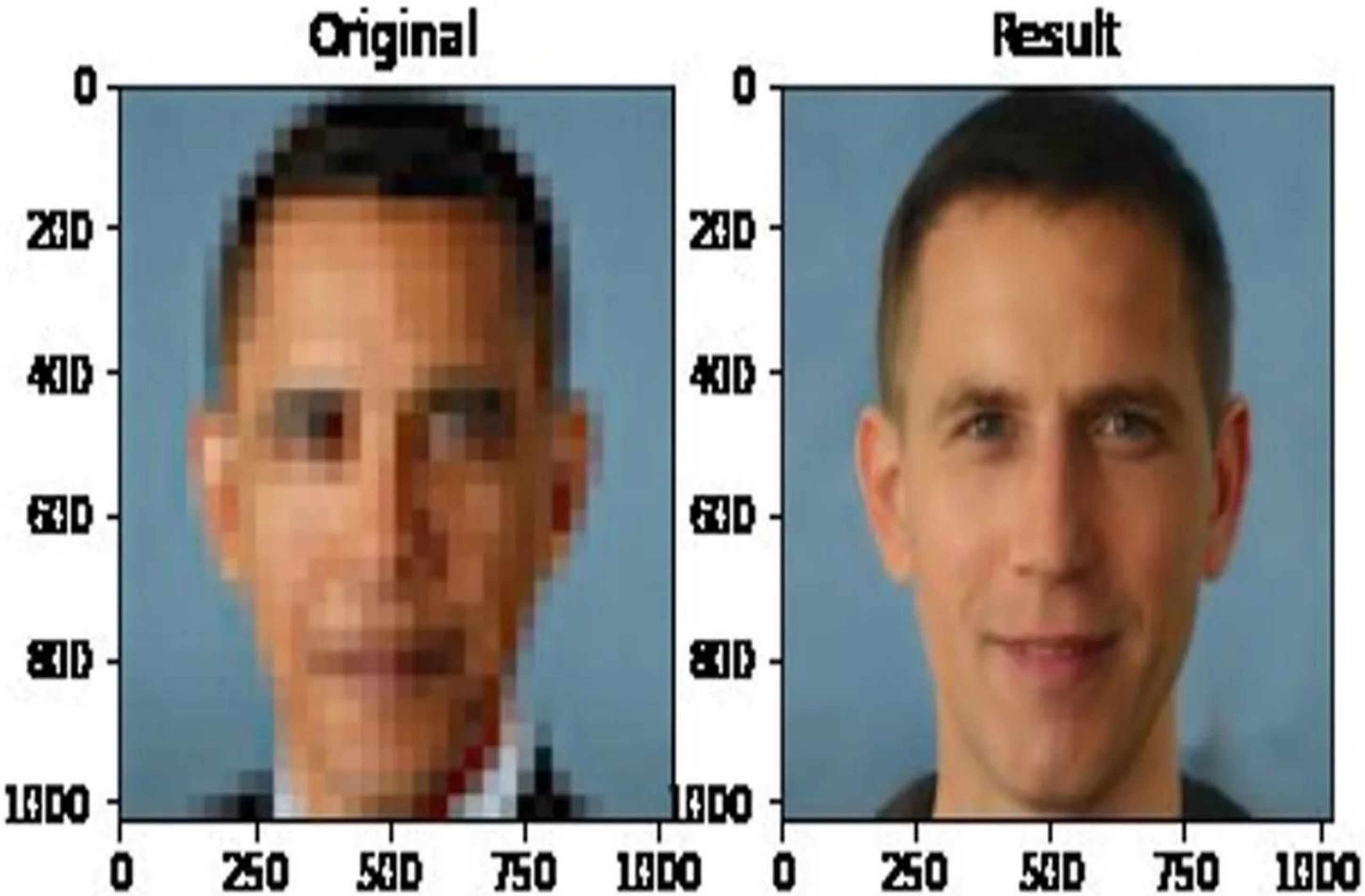
Provenance
Data splits
Redundancy
Availability

Algorithm
Meta-predictions
Data encoding
Parameters
Features
Fitting
Availability

Interpretability
Execution time
Availability of software

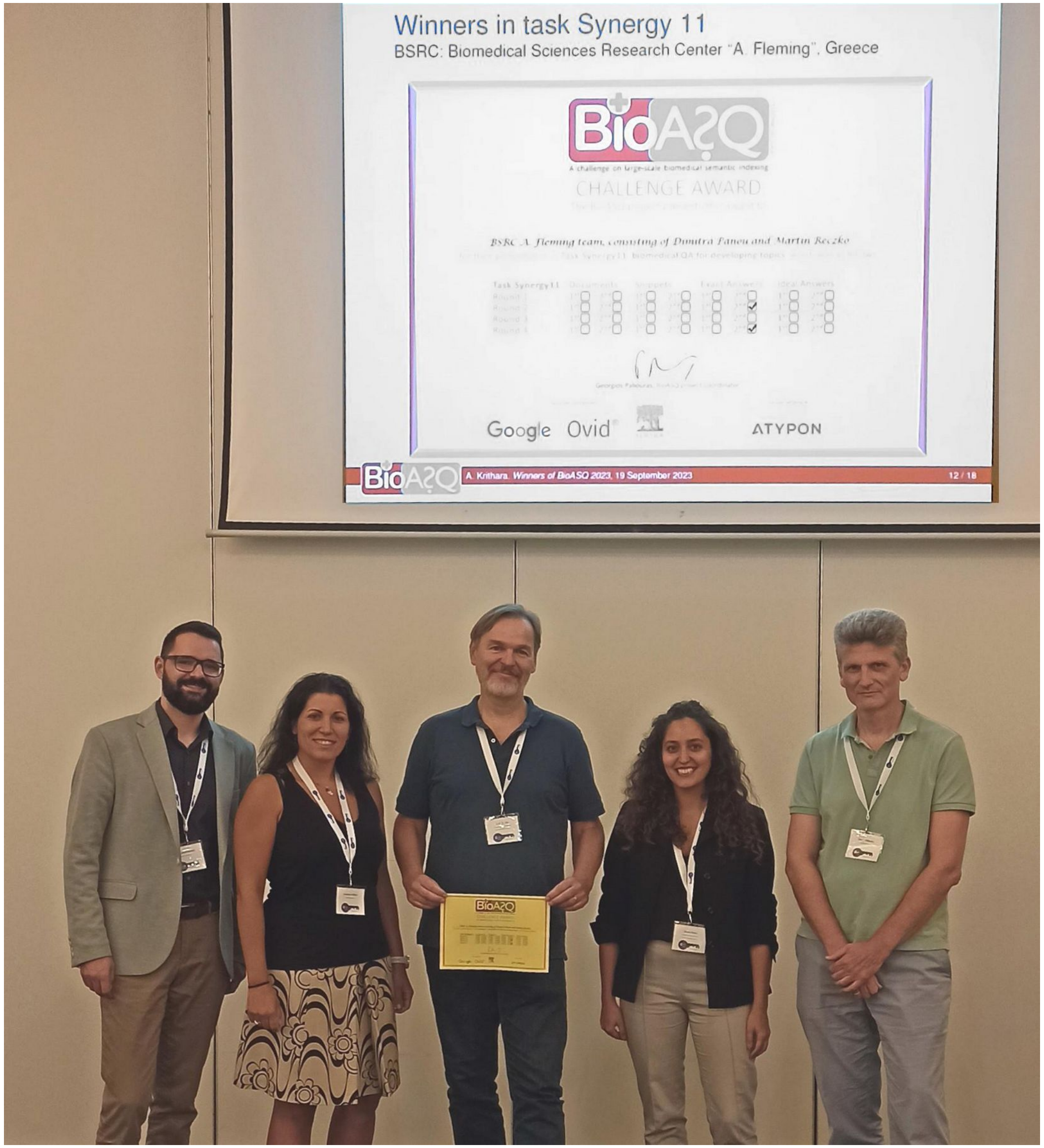
Evaluation
Performance
Comparison
Confidence
Availability

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   2nd places in 2023, with Dimitra Panou



Winners in task Synergy 11

BSRC: Biomedical Sciences Research Center "A. Fleming", Greece

BioASQ

A challenge on large-scale biomedical semantic indexing

CHALLENGE AWARD

BSRC A. Fleming team, consisting of Dimitra Panou and Martin Reczko

Task Synergy11 biomedical QA for developing topics

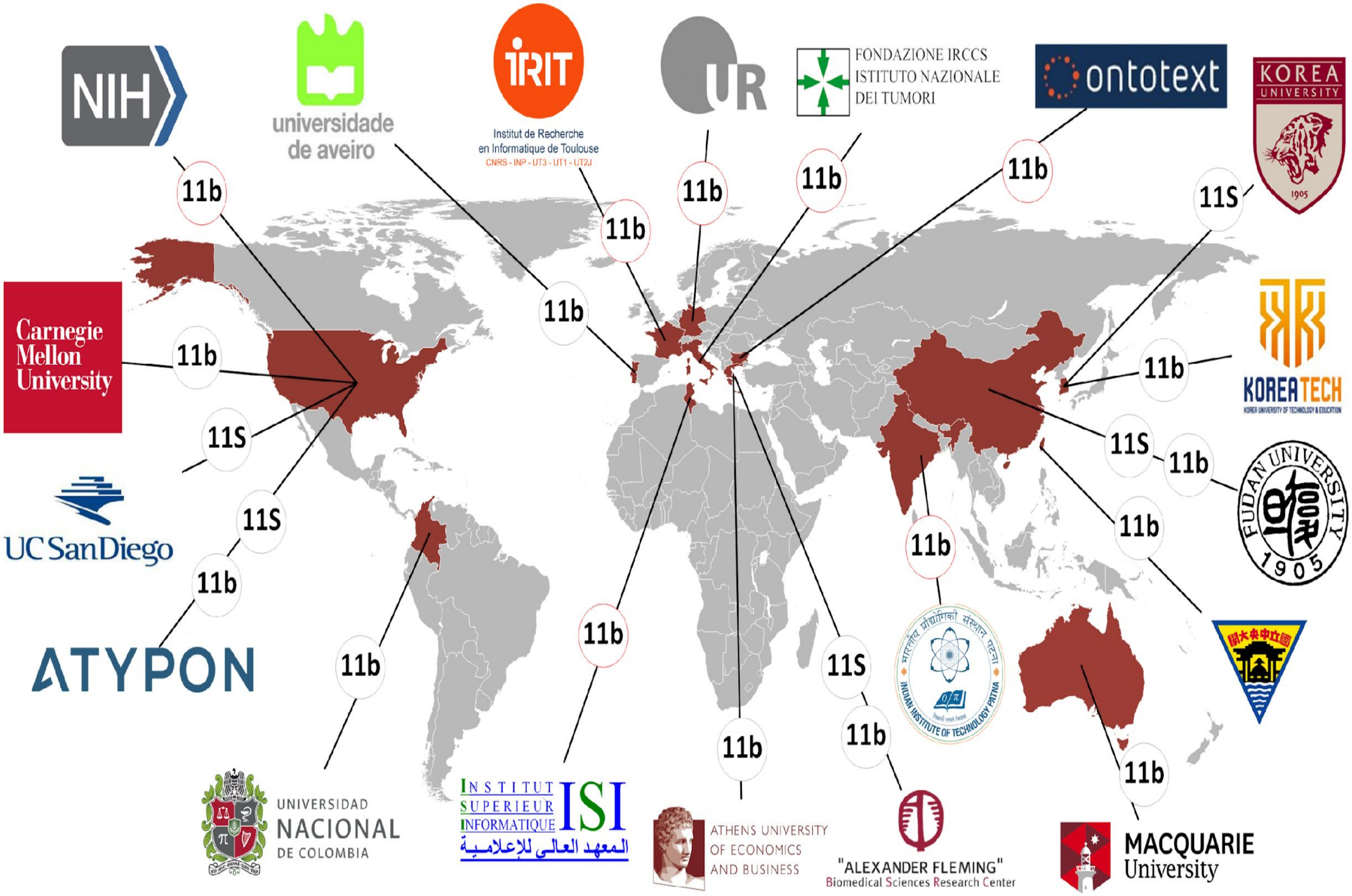
Task Synergy11	Documents		Snippets		Exact Answers		Ideal Answers	
Round 1	1st □	2nd □	1st □	2nd □	1st □	2nd □	1st □	2nd □
Round 2	1st □	2nd □	1st □	2nd □	1st □	2nd ☑	1st □	2nd □
Round 3	1st □	2nd □	1st □	2nd □	1st □	2nd □	1st □	2nd □
Round 4	1st □	2nd □	1st □	2nd □	1st □	2nd ☑	1st □	2nd □

Georgios Paliouras, BioASQ project coordinator

Google Ovid ATYPON

A. Krithara, *Winners of BioASQ 2023*, 19 September 2023 12 / 18

BioASQ: Int. competition for biomedical QA


NIH
11b
universidade de aveiro
11b
IRIT
Institut de Recherche en Informatique de Toulouse
CNRS - INP - UT3 - UT1 - UT2J
11b
UR
11b
FONDAZIONE IRCCS ISTITUTO NAZIONALE DEI TUMORI
11b
ontotext
11b
KOREA UNIVERSITY
1905
11S
Carnegie Mellon University
11b
11S
UC San Diego
11S
11b
ATYPON
11b
KOREATECH
KOREA UNIVERSITY OF TECHNOLOGY & EDUCATION
11b
11S
11b
FUDAN UNIVERSITY 1905
11b
11b
INDIAN INSTITUTE OF TECHNOLOGY PATNA
國立中央大學
11b
11b
11S
11b
11b
UNIVERSIDAD NACIONAL DE COLOMBIA
INSTITUT SUPERIEUR INFORMATIQUE ISI
المعهد العالي للإعلامية
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
"ALEXANDER FLEMING" Biomedical Sciences Research Center
MACQUARIE University

2024: Introduced a ‘Farm’ of LLMs

AI generated using Copilot

SCAN ME

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

"ALEXANDER FLEMING"
Biomedical Sciences Research Center

2024: 2   1st and 3   2nd places

- Our awards:

🏆 Batch 4 Snippet Identification

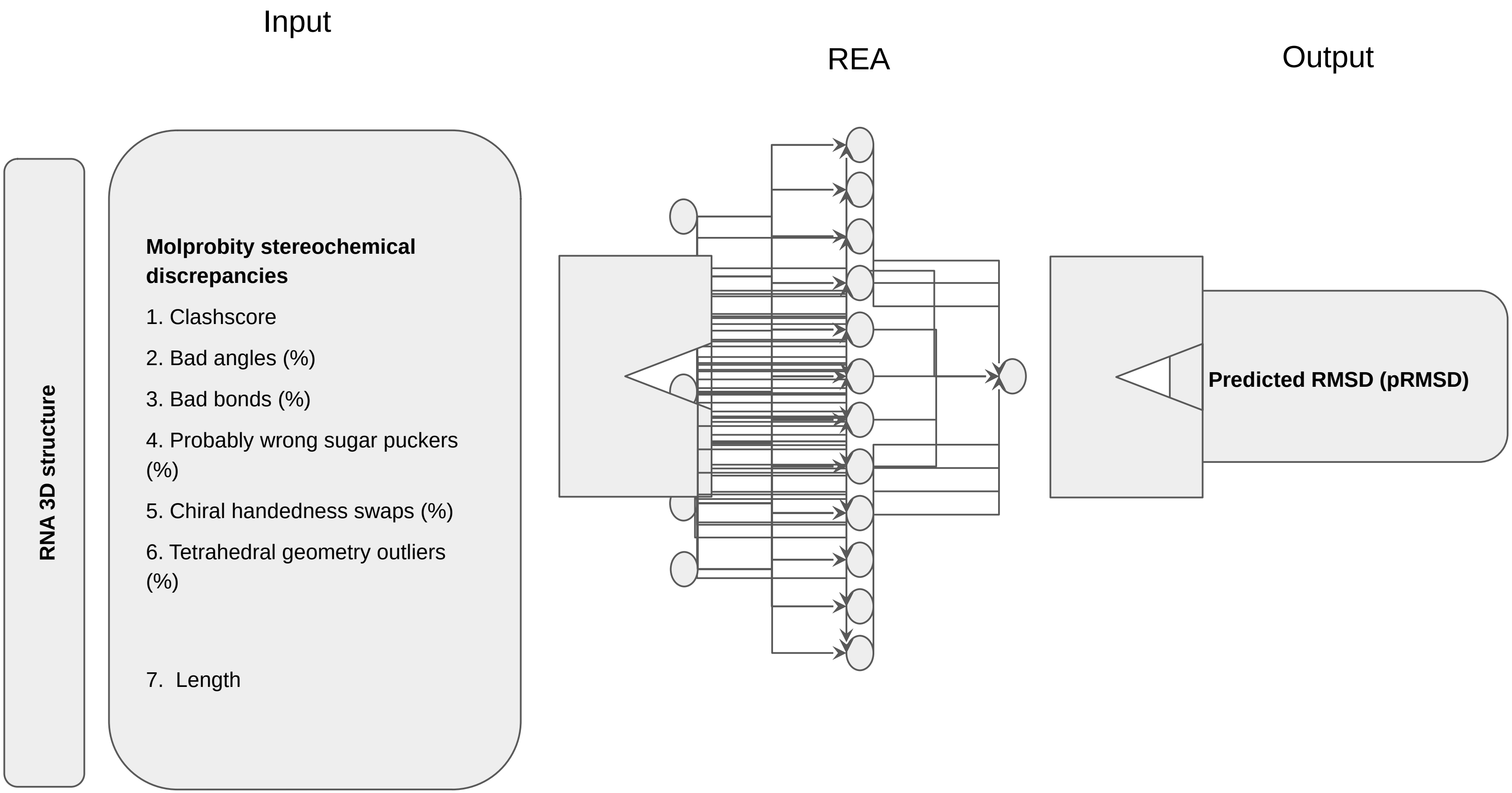
🏆 Batch 1 Exact Answers

🏆 Batch 2 Exact Answers

🏆 Batch 2 Ideal Answers

🏆 Batch 1 Documents retrieval

With Rea Kalampaliki: RMSD Estimation Algorithm (REA)



Input

REA

Output

RNA 3D structure

Molprobity stereochemical discrepancies

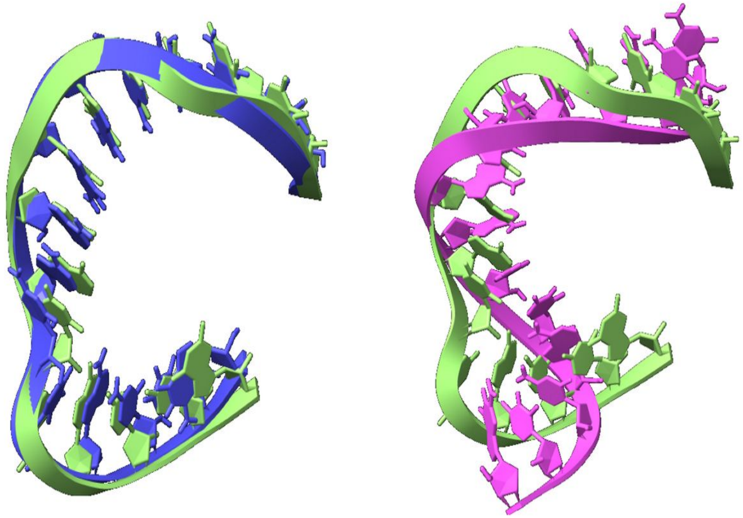
1. Clashscore
2. Bad angles (%)
3. Bad bonds (%)
4. Probably wrong sugar puckers (%)
5. Chiral handedness swaps (%)
6. Tetrahedral geometry outliers (%)

7. Length

Predicted RMSD (pRMSD)

https://biorxiv.org/cgi/content/short/2024.02.28.582508v1

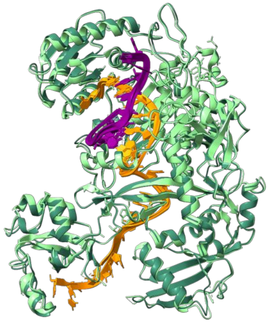
REA: Improvement in the accuracy of the predicted 3D structure of an RNA Nanosquare chain



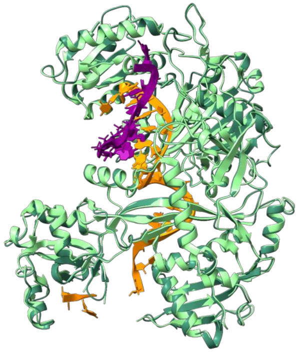
- DeepFoldRNA #4 3D model (left, purple, RMSD = 1.37Å), predicted by SumReaSVR
- trRosettaRNA 3D model (pink, RMSD = 7.38Å)
- Reference structure (green, PDB: 3P59, chain A)
- Improvement ~6Å in the accuracy of the predicted 3D model

Ago2 - miR - target AlphaFold3 models

4Z4F



4W5O

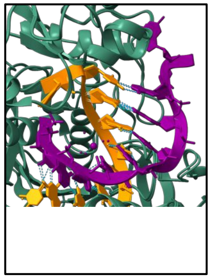


AF3 model	ipTM*	RMSD	RMSD_rnas
4Z4F	0.98	3.12	4.61
4W5O	0.97	4.20	2.00

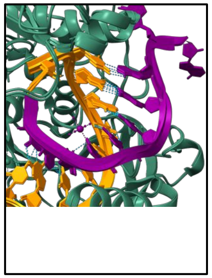
ipTM < 0.6: failed prediction

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

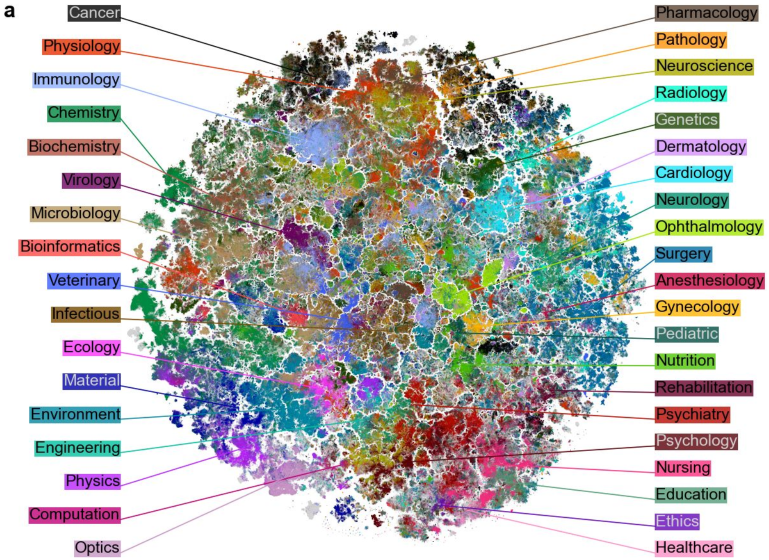
4Z4F target



4W5O target



Transformers help clustering all scientifc papers



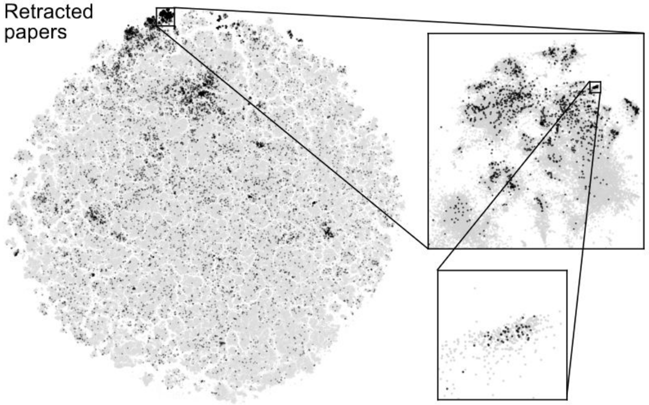


Figure 6: Retracted papers group together. All retracted papers with intact abstracts (11,756) are highlighted in black, plotted on top of the non-retracted papers. First inset corresponds to one of the regions with higher density of retracted papers (3.8%), covering research on cancer-related drugs, 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 EEST / Updated Sept. 12, 2023, 5:31 PM EEST / 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."

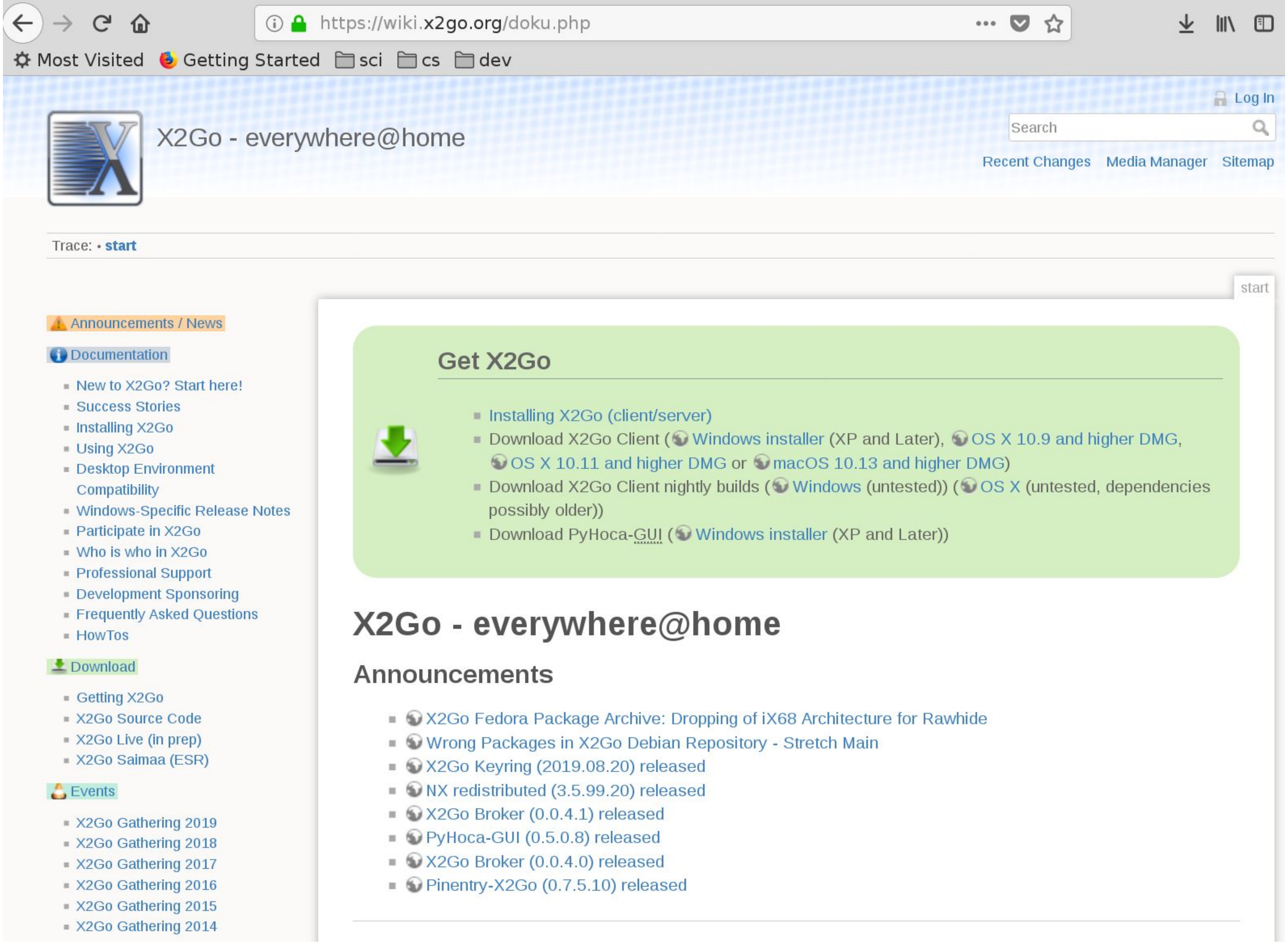
Then Alex began chewing things, so Courtney took him to the dentist. What followed was a three-year search for the cause of Alex's increasing pain and eventually other symptoms.

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

Install x2go to access graphical user interface



X2Go - everywhere@home

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

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

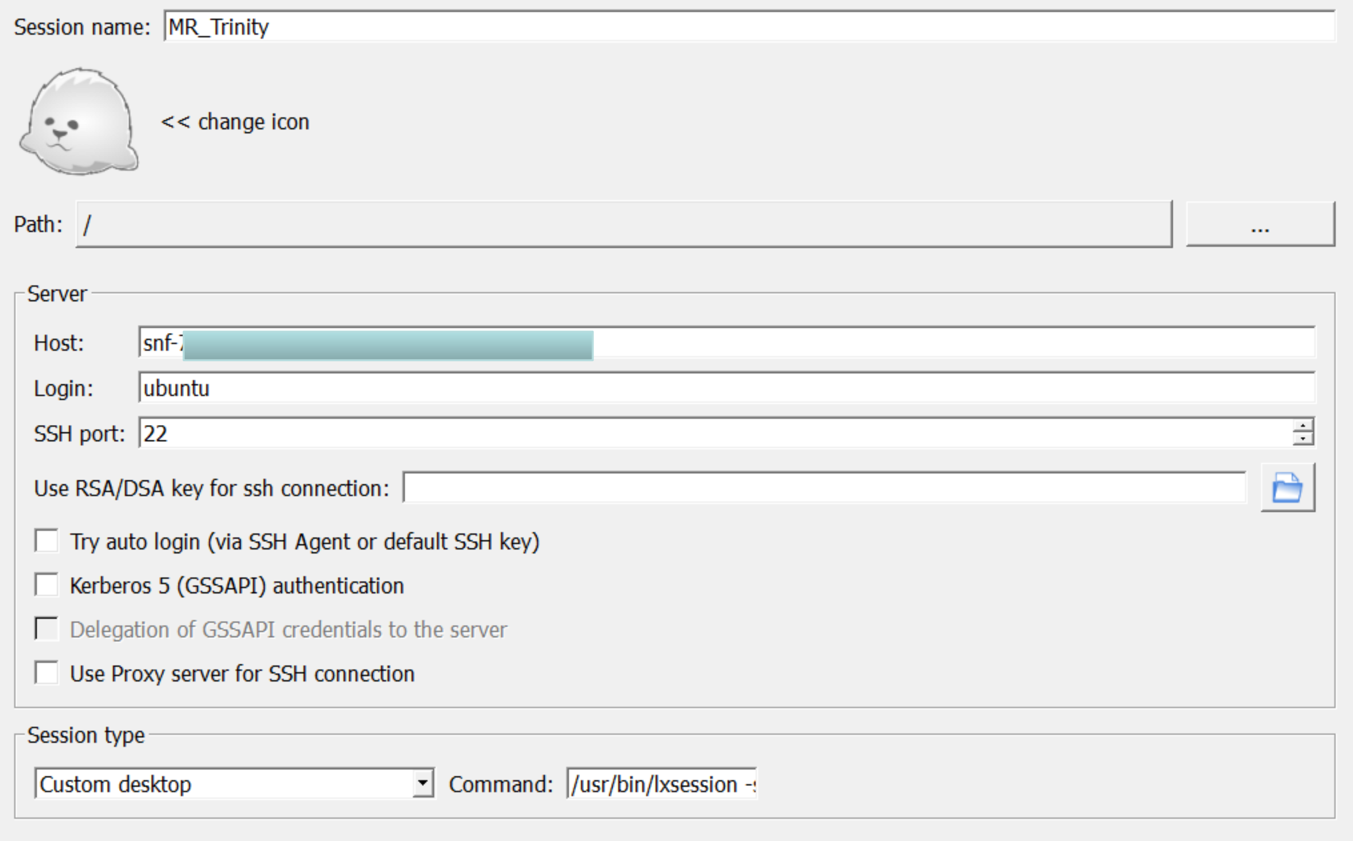
- X2Go Fedora Package Archive: Dropping of iX68 Architecture for Rawhide
- Wrong Packages in X2Go Debian Repository - Stretch Main
- X2Go Keyring (2019.08.20) released
- NX redistributed (3.5.99.20) released
- X2Go Broker (0.0.4.1) released
- PyHoca-GUI (0.5.0.8) released
- X2Go Broker (0.0.4.0) released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

Access to virtual machine

- Install x2go from: https://wiki.x2go.org/doku.php/download:start
- X2go:

L

- Session ty
  Lubuntu -e L
- (Virtualbox ya
)



Session name: MR_Trinity

<< change icon

Path: /

Server

Host: snf-[redacted]

Login: ubuntu

SSH port: 22

Use RSA/DSA key for ssh connection:

□ Try auto login (via SSH Agent or default SSH key)

□ Kerberos 5 (GSSAPI) authentication

□ Delegation of GSSAPI credentials to the server

□ Use Proxy server for SSH connection

Session type

Custom desktop Command: /usr/bin/lxsession -

Introduction to Bioinformatics 2024-2025

Exercise 1 (M. Reczko):

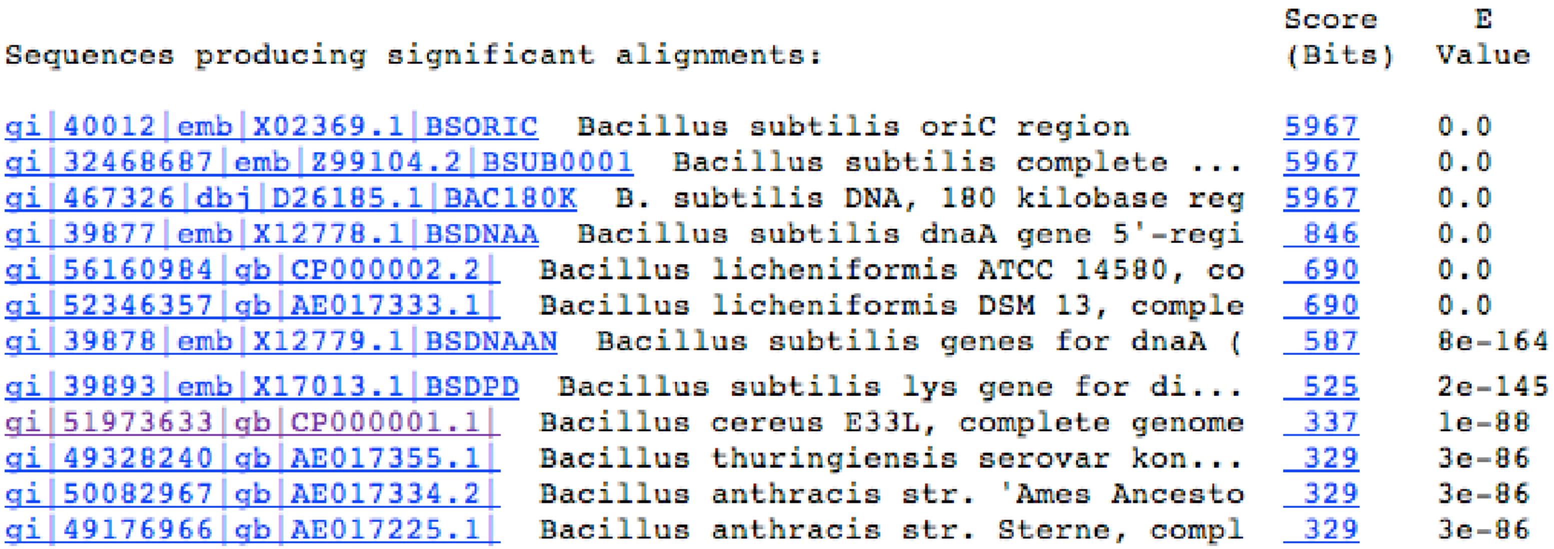
(Adapted from:
https://web.archive.org/web/20150425010121/http://www.ableweb.org/volumes/vol-28/v28reprint.php?ch=8
)

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:



```
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  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
TAAAAAATCACTAATNAATNTAAGAGCGGACTTAACNAGCGTAAAACTGTCTTACTAATTAAT
TGTCAGTTAGCTCGTTCAGGTAATGGTTCCTANCGGNCAATGCAGGAAGAGTTCTACCTGG
AACTGANAGACCGCTGGCGGTGACAACACACTACGTCAAAATAAGA
>outbreak15
TAGTCTTGTGCTNACGGAACTTATTTATGAGGTACCCACCGANTCTGAAAACCGCTAATANA
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://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