Introduction to applied bioinformatics

PETRA MATOUŠKOVÁ

2024/2025

„Nucleotide bioinformatics II“

Retrieving nucleotide sequences from databases (Genbank/NCBI)

Feature analysis: statistics, reverse complement, restriction analysis

Translation, identifying open reading frame

PCR primer design, rt-PCR

Secondary structure prediction

Sequence comparison, unknown sequence identification

Single Nucleotide Polymorphisms

DNA sequencing

Gene expression

microRNA

Genomes….

….

Nucleotide sequence comparison

-Analogous to protein comparison

Nucleotide sequence comparison

-Analogous to protein comparison

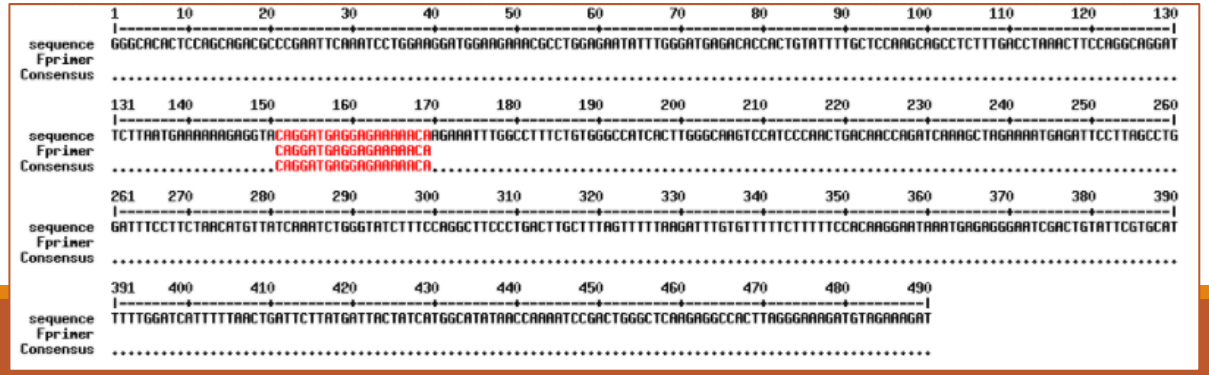
e.g. multalin

→ Comparing/looking for short sequences (primers)

Default matrix:BLOSUM62

input:

```
>sequence
GGGCACACTCCAGCAGACGCCCGAATTCAAATCCTGGAAGGATGGAAGAAACGCCTGGAGAATATTTGGG
ATGAGACACCACTGTATTTTGCTCCAAGCAGCCTCTTTGACCTAAACTTCCAGGCAGGATTCTTAATGAA
AAAAGAGGTACAGGATGAGGAGAAAAACAAGAAATTTGGCCTTTCTGTGGGCCATCACTTGGGCAAGTCC
ATCCCAACTGACAACCAGATCAAAGCTAGAAAATGAGATTCCTTAGCCTGGATTTCCTTCTAACATGTTA
TCAAATCTGGGTATCTTTCCAGGCTTCCCTGACTTGCTTTAGTTTTTAAGATTTGTGTTTTTCTTTTTCC
ACAAGGAATAAATGAGAGGGAATCGACTGTATTCGTGCATTTTTGGATCATTTTTAACTGATTCTTATGA
TTACTATCATGGCATATAACCAAAATCCGACTGGGCTCAAGAGGCCACTTAGGGAAAGATGTAGAAAGAT
>2.exon
CAGGATGAGGAGAAAAACA
```



Sequence comparison

open Multalin

from Examples:

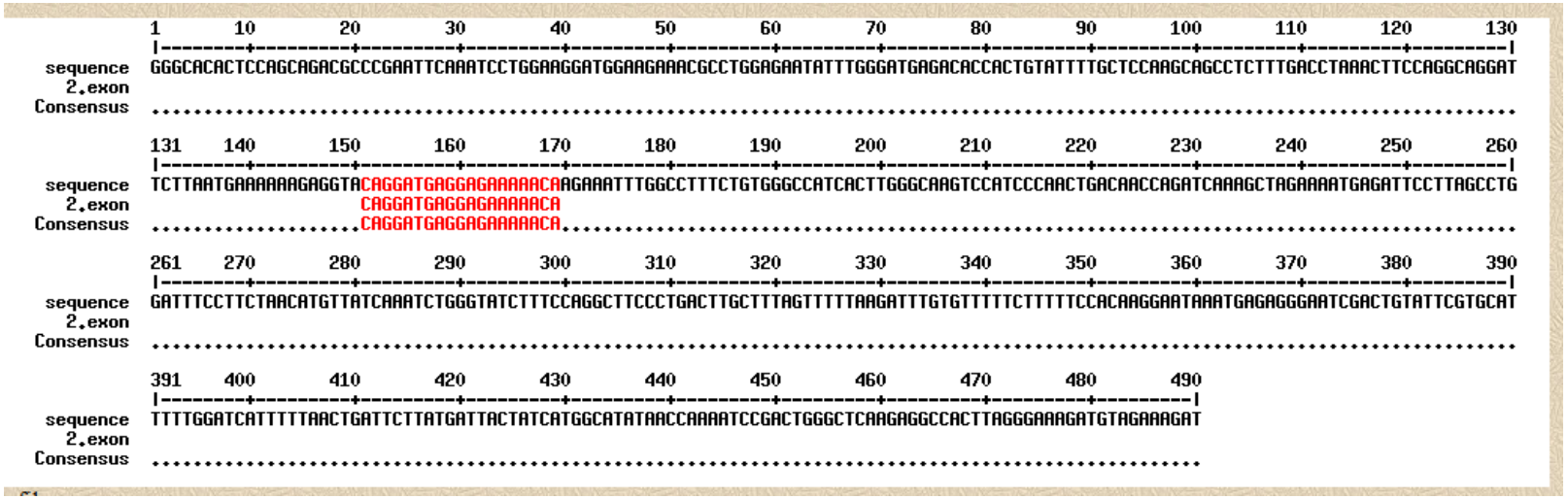
- A: compare long and short sequence (default Blosum matrix)

Sequence comparison

open Multalin

from Examples:

- A: compare long and short sequence (default Blosum matrix)



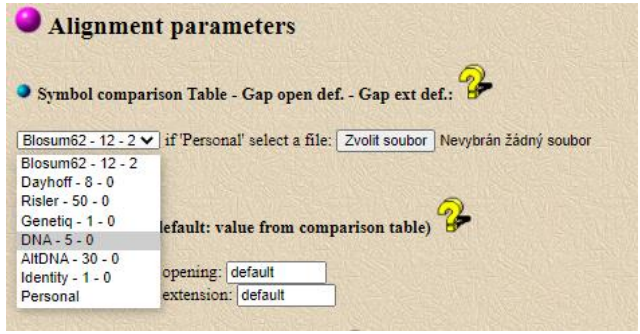
- Using both matrices would look like the same

Nucleotide sequence comparison

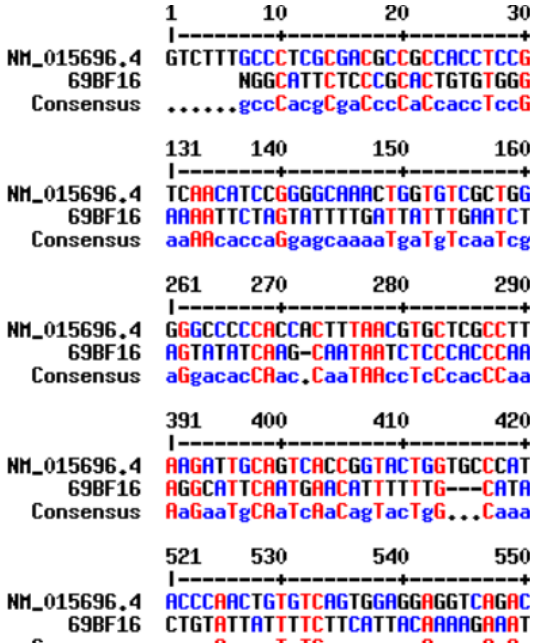
-Analogous to protein comparison

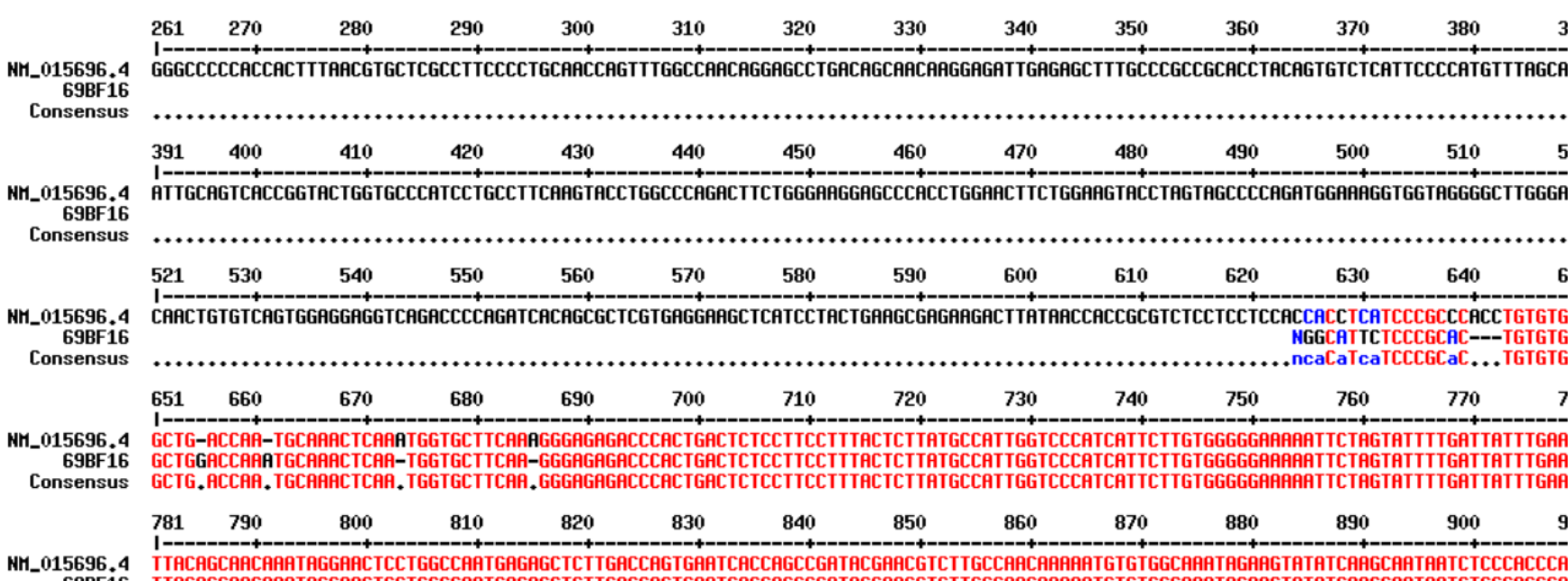
e.g. multalin

→ Comparing sequences identified and unknown



Default matrix:BLOSUM62 – often does not work for DNA sequence →**DNA-5-0**





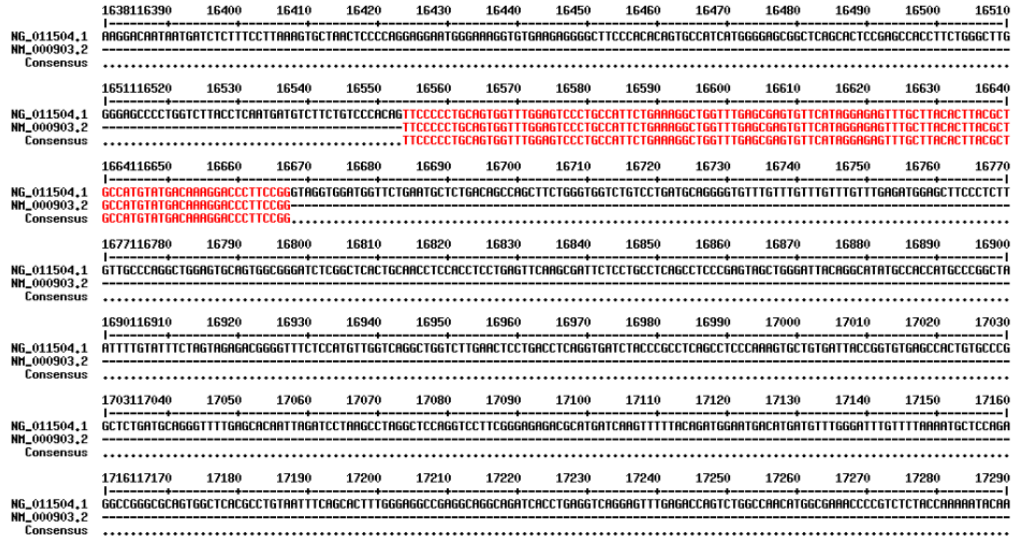
Nucleotide sequence comparison

-Analogous to protein comparison

e.g. multalin

→ Comparing genomic DNA and cDNA (mRNA)

Default:BLOSUM62 **→DNA 5-0**



Sequence comparison

open Multalin

from Examples:

- A: compare long and short sequence (default Blosum matrix)
- B: compare two long sequences (default Blosum matrix)
- change parametres: matrix **DNA-5-0**

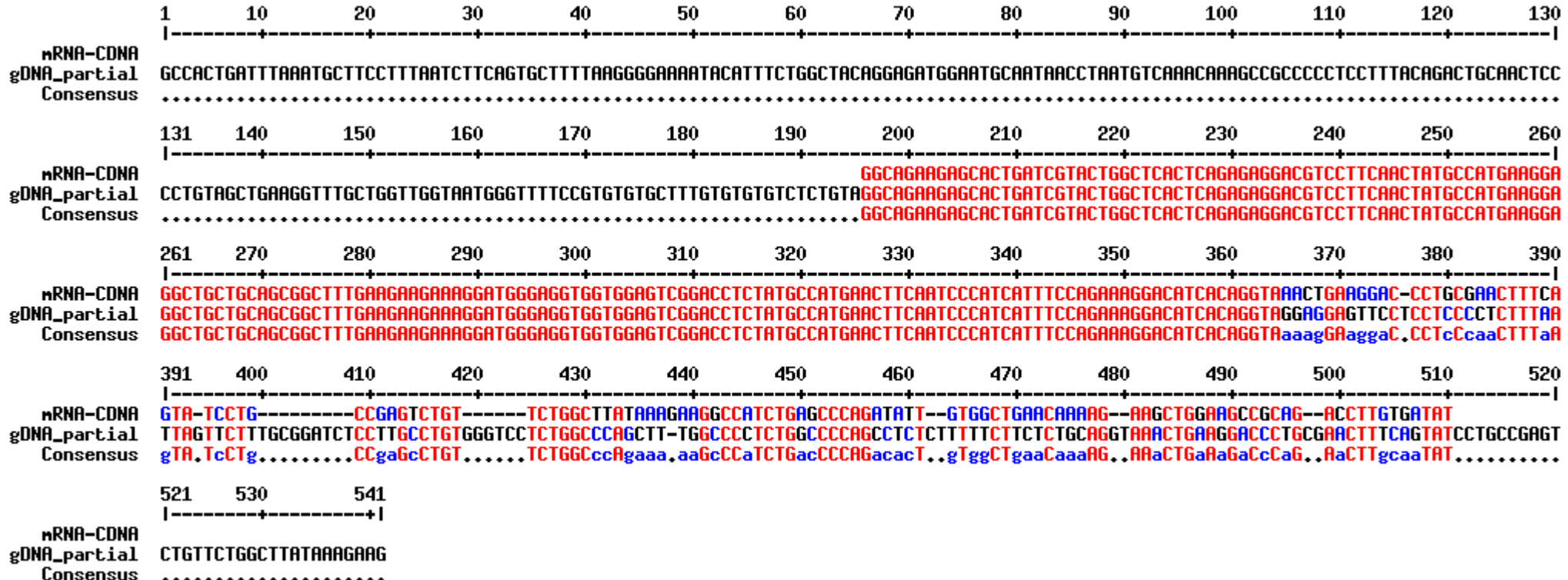
  see the difference?

Sequence comparison

open Multalin

from Examples:

- B: compare two long sequences (default Blosum matrix)

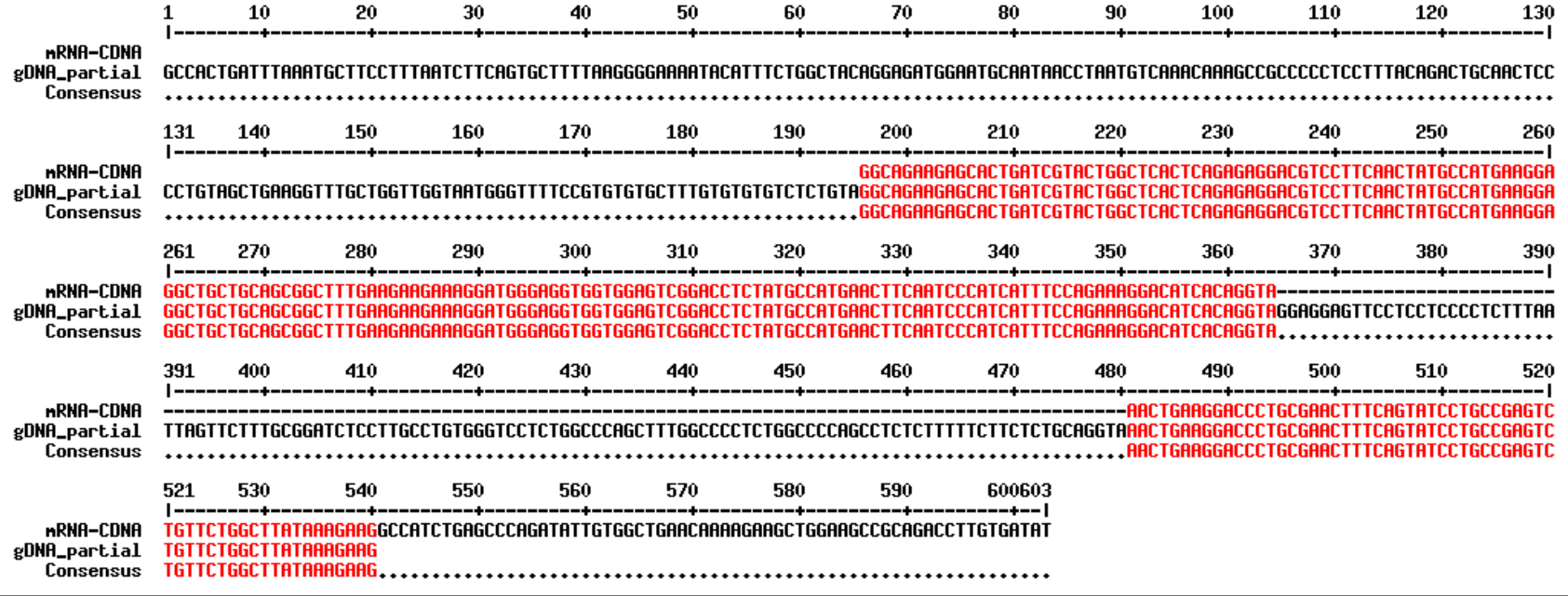


Sequence comparison

open Multalin

from Examples:

- B: compare two long sequences (default Blosum matrix)
- change parametres: matrix **DNA-5-0**



Sequence comparison

open Multalin

from Examples:

- A: compare long and short sequence (default Blosum matrix)
- B: compare two long sequences (default Blosum matrix)
- change parametres: matrix **DNA-5-0**

  see the difference?

- **HW: compare your CDS and mRNA**

„reading“ DNA = translation

Genetic code: triplets (codons)→3 possibilities of reading= ORF (open reading frame)

1. DNA sequence:

5´-ATGGAAGTATTTAAAGCGCCACCTATTTAA-3´

2. Divided into triplets:

ATG GAA GTA TTT AAA GCG CCA CCT ATT TAA

A **TGG** AAG **TAT** TTA **AAG** CGC **CAC** CTA **TT**T AA

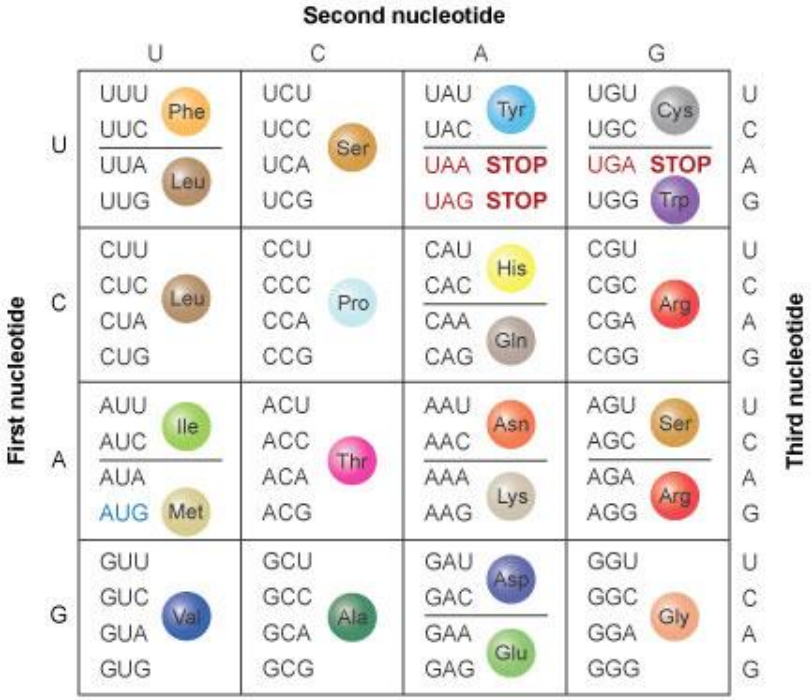
AT **GGA** AGT ATT TAA AGC GCC ACC TAT TTA A

3. Each triplet translated into aminoacid (decoded):

M E V F K A P P I STOP(*)

W K Y L K R H L F

G R I * S A T Y L



„reading“ DNA = translation

DNA sequence: 5' - 3', protein sequence from **N-** to **C-** terminus.

X we don´t know which strand is coding:

5´-ATGGAAGTATTTAAAGCGCCACCTATTTAA-3´

3´-TACCTTCATAAATTTCGCGGTGGATAAATT-5´

5´-TTAAATAGGTGGCGCTTTAAATACTTCCAT-3´

TTA AAT AGG TGG CGC TTT AAA TAC TTC CAT

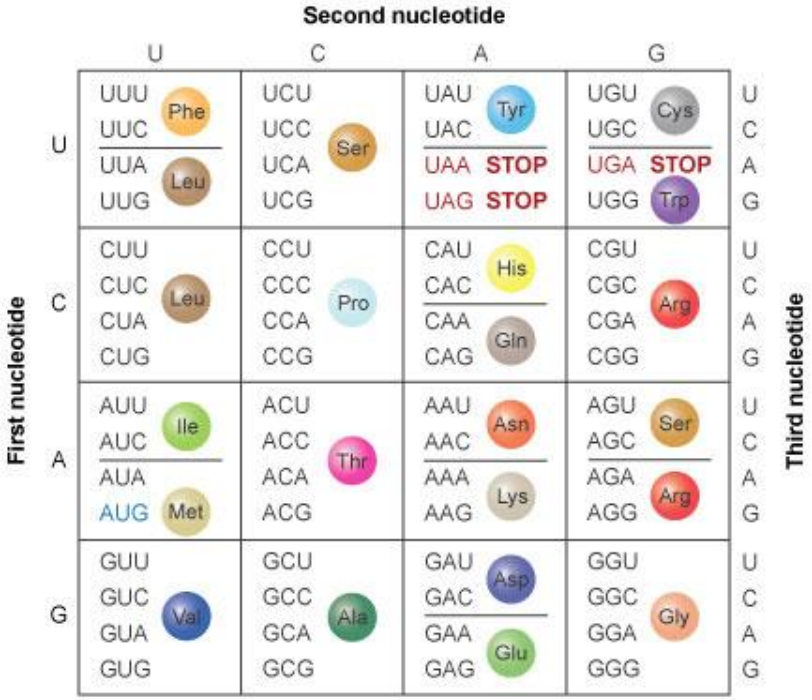
T TAA ATA GGT GGC GCT TTA AAT ACT TCC AT

TT AAA TAG GTG GCG CTT TAA ATA CTT CCA T

L N R W R F K Y F H

* I G G A L N T S

K * V A L * I L P



„reading“ DNA = translation

→ there are 6(!) potential open reading frames = **ORFs**

5´- ATGGAAGTATTTAAAGCGCCACCTATTTAA-3´

ATG GAA GTA TTT AAA GCG CCA CCT ATT TAA

A **TGG** AAG **TAT** TTA **AAG** CGC **CAC** CTA **TT**T AA

AT **GGA** AGT ATT TAA AGC GCC ACC TAT TTA A

M E V F K A P P I STOP(*)

W K Y L K R H L F

G R I * S A T Y L

5´-TTAAATAGGTGGCGCTTTAAATACTTCCAT-3´

TTA AAT AGG TGG CGC TTT AAA TAC TTC CAT

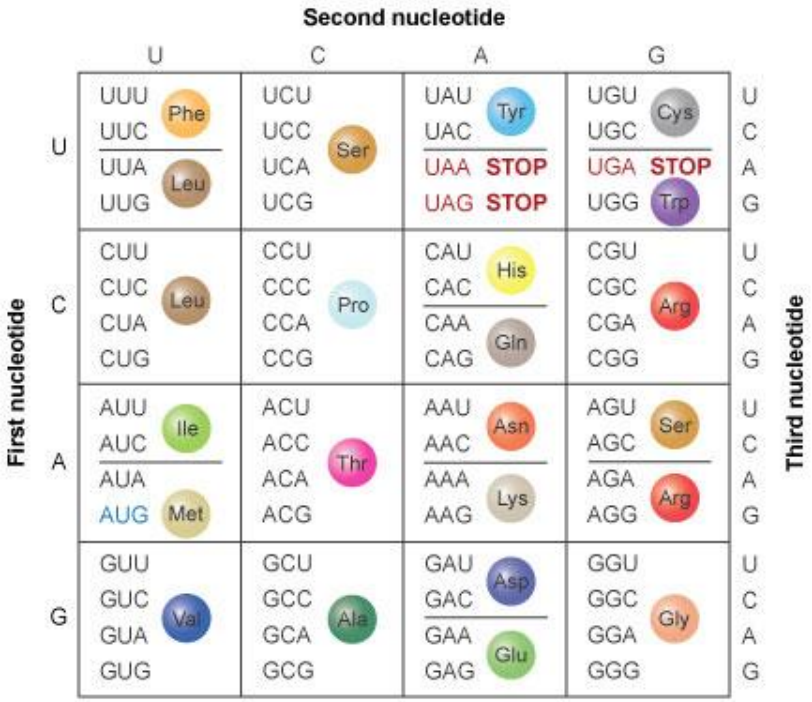
T TAA ATA GGT GGC GCT TTA AAT ACT TCC AT

TT AAA TAG GTG GCG CTT TAA ATA CTT CCA T

L N R W R F K Y F H

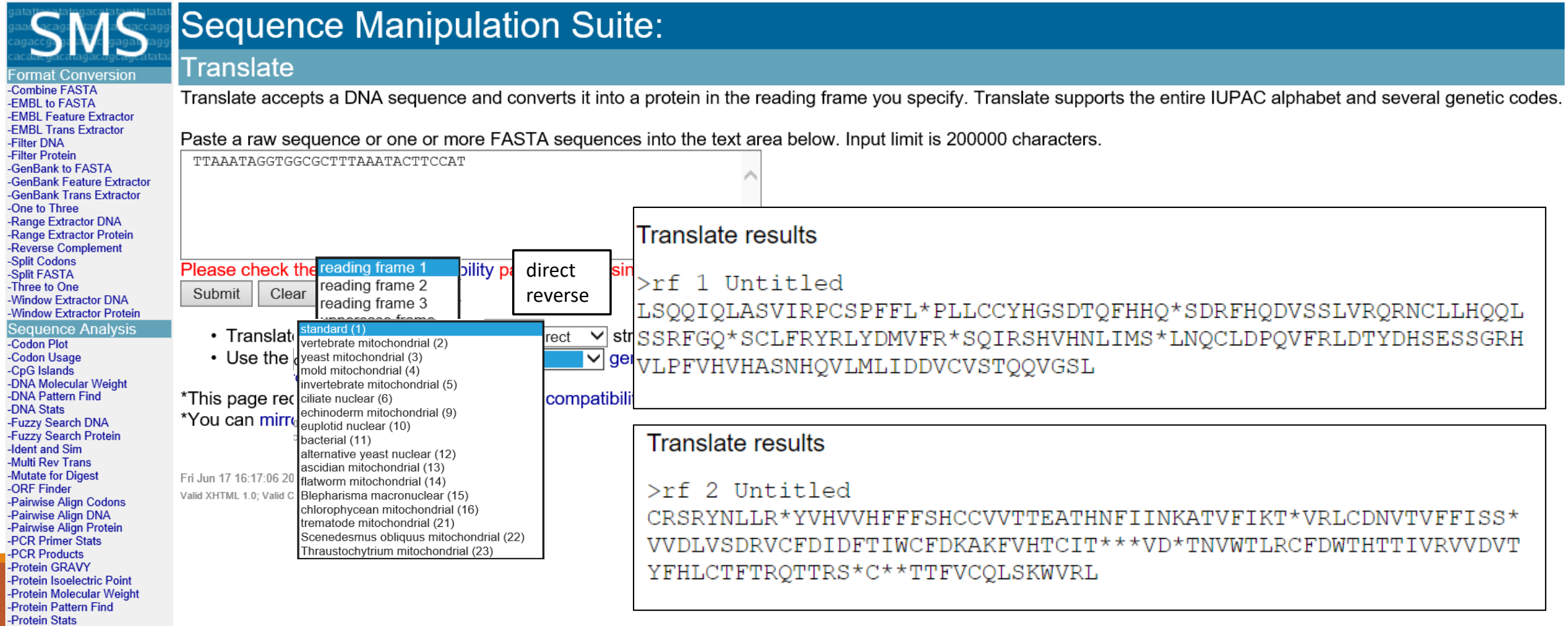
* I G G A L N T S

K * V A L * I L P



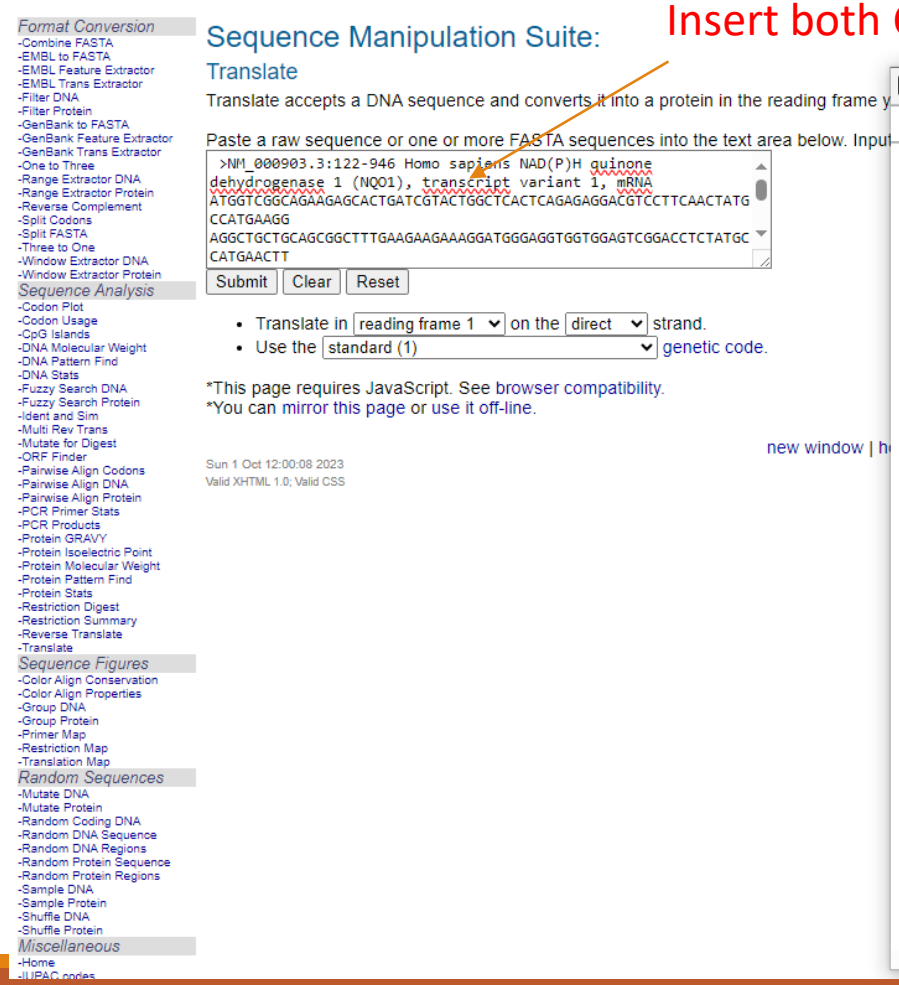
„reading“ DNA = translation

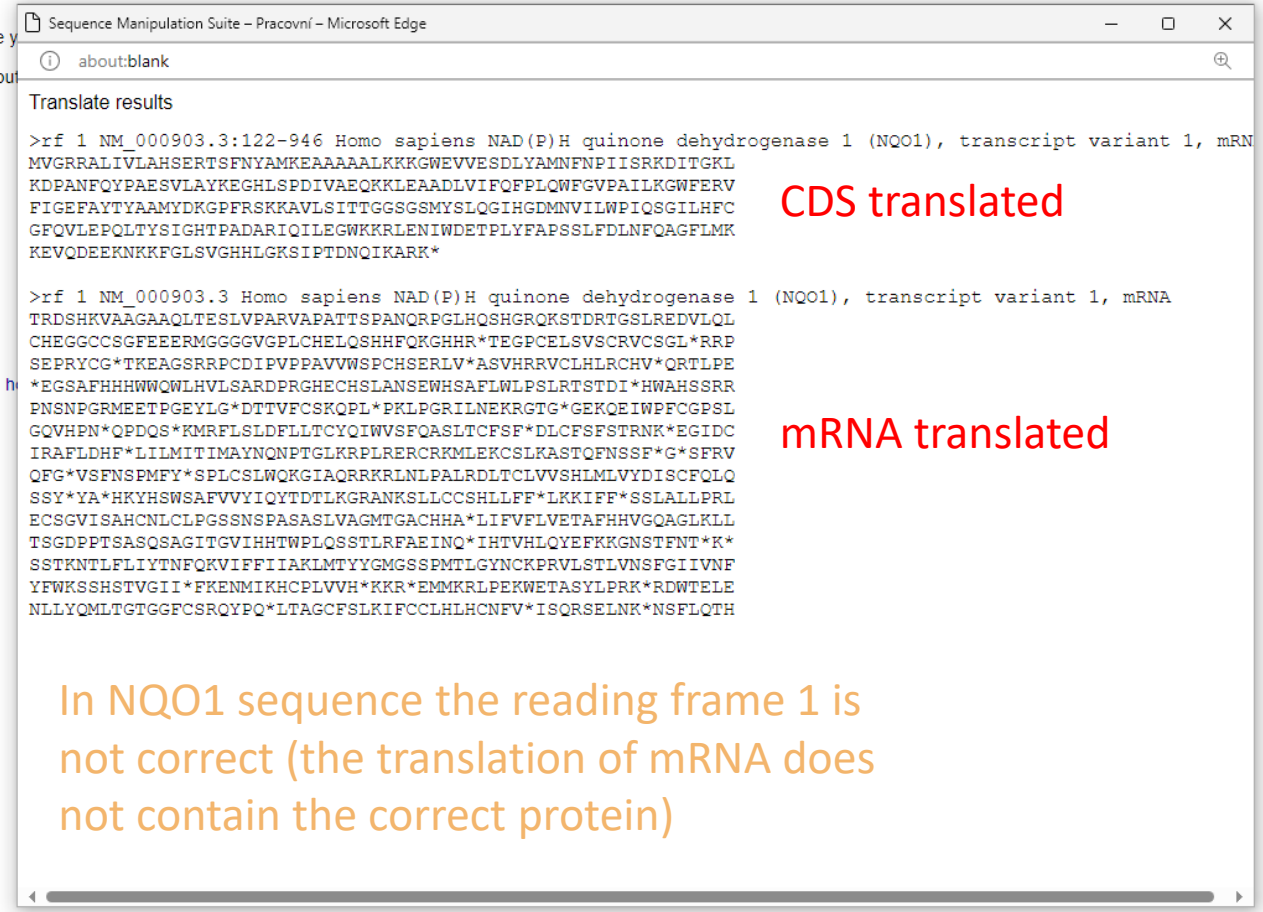
SMS/Translate → suitable for **full length CDS only**=starting with „ATG“ (or when we know which reading frame to use)



„reading“ DNA = translation

Insert both CDS and mRNA in fasta format





CDS translated

mRNA translated

In NQO1 sequence the reading frame 1 is not correct (the translation of mRNA does not contain the correct protein)

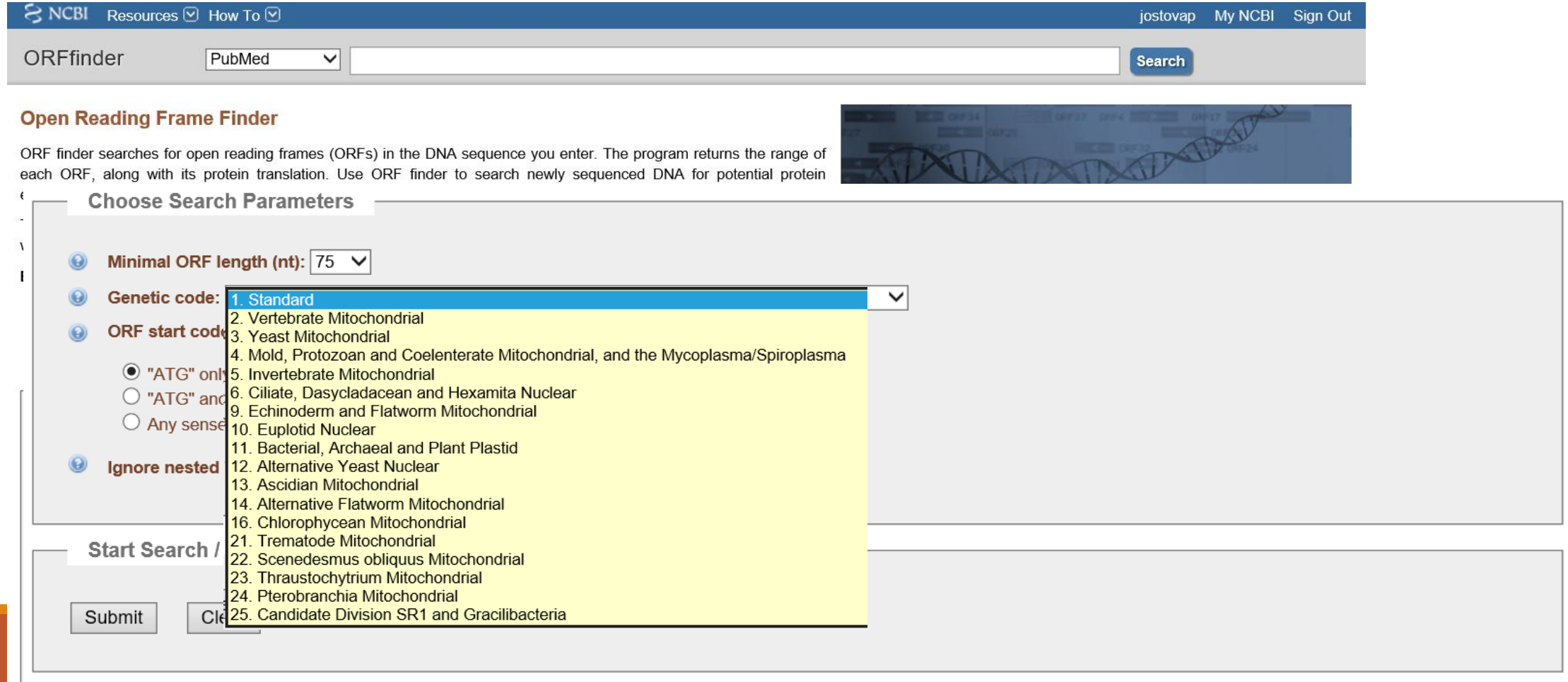
Try

-translate you full length **mRNA** and **CDS** in *SMS translate*

-spot the difference

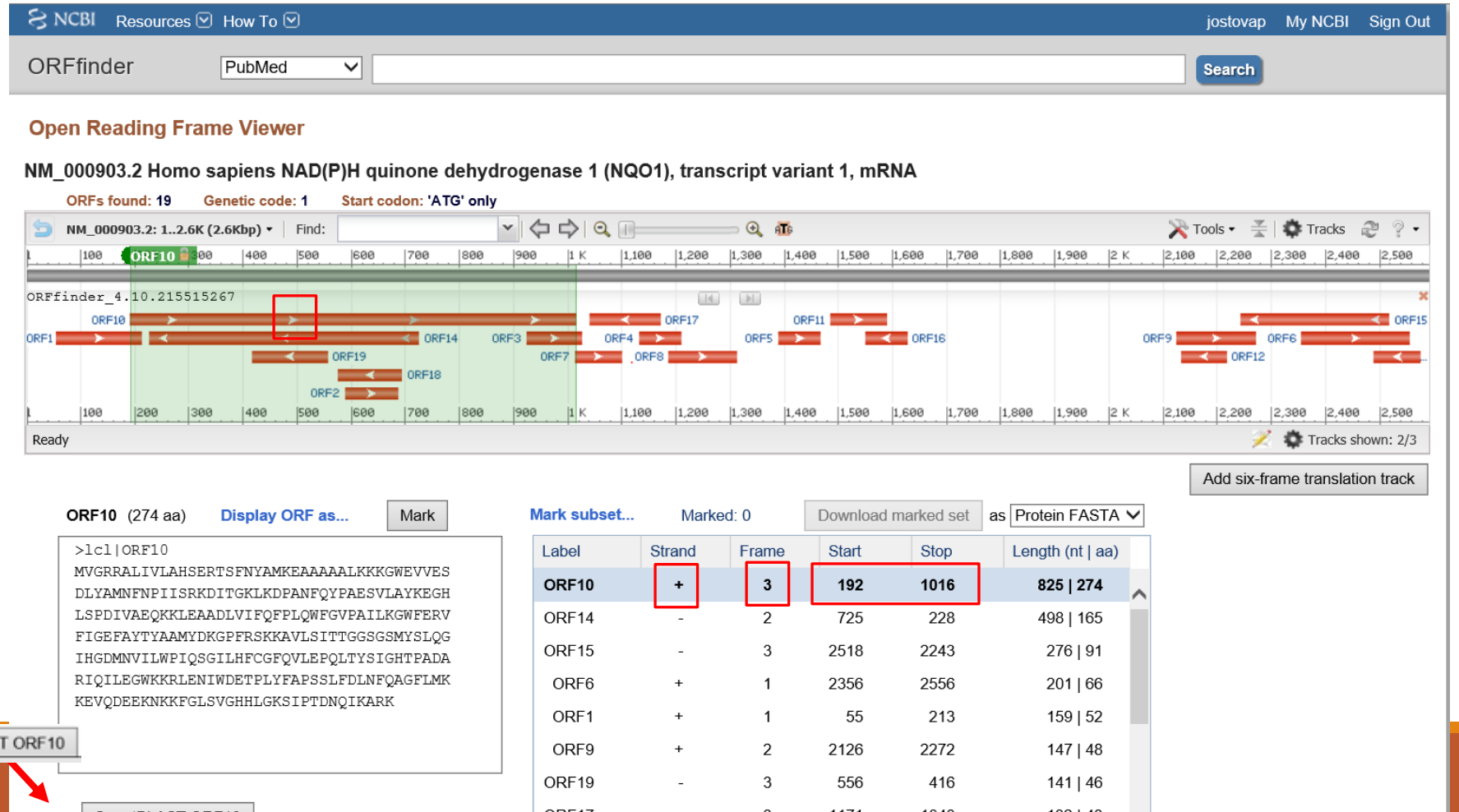
„reading“ DNA = translation

ORFfinder → looking for the longest ORF **Open Reading Frame**

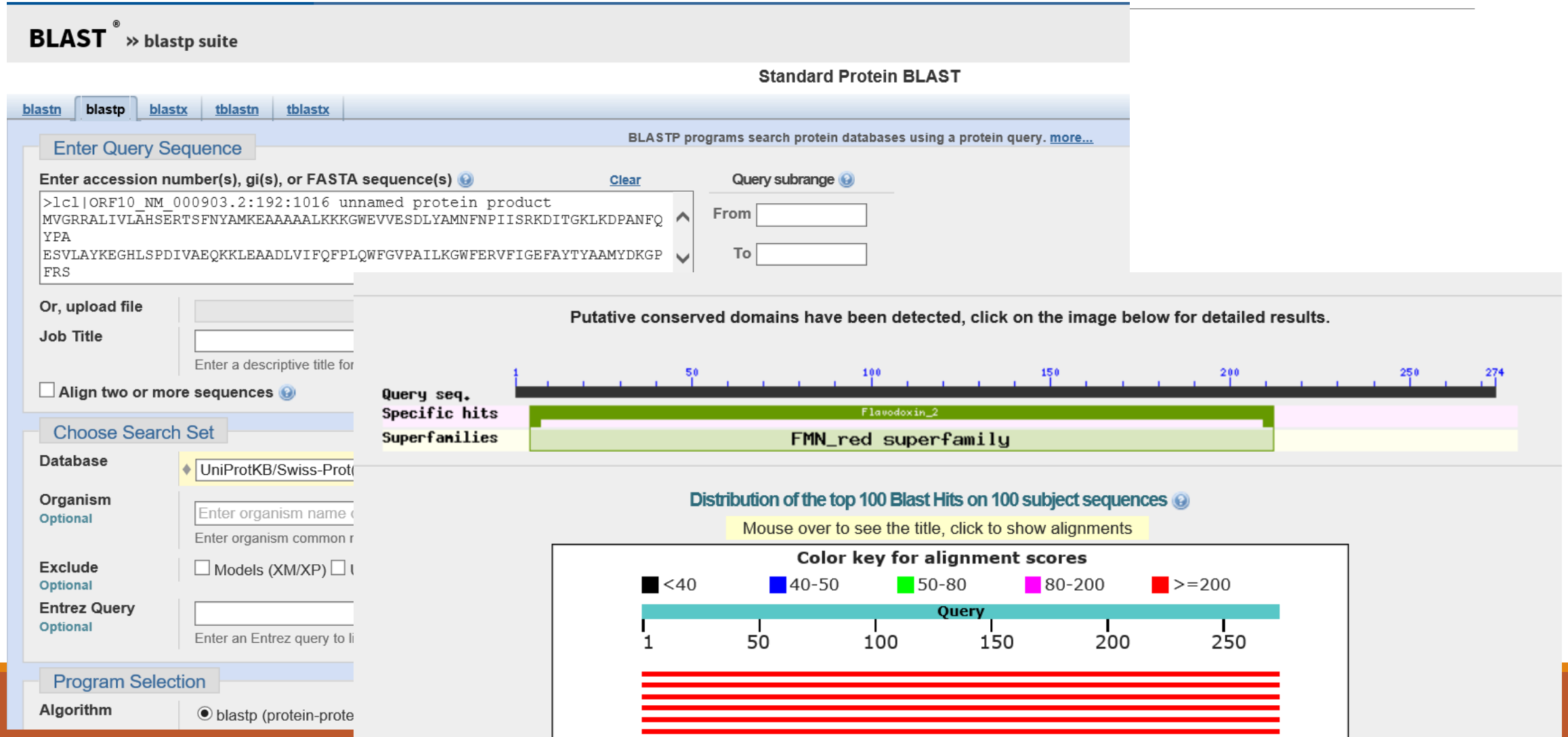


„reading“ DNA = translation

ORFfinder → looking for the longest ORF

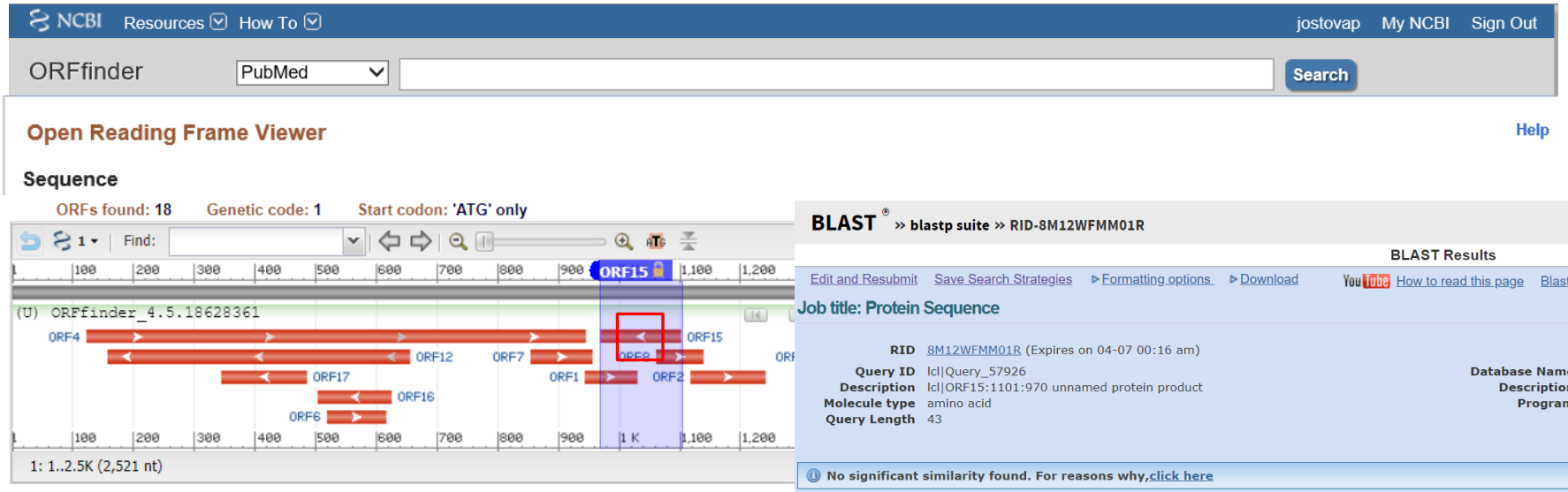


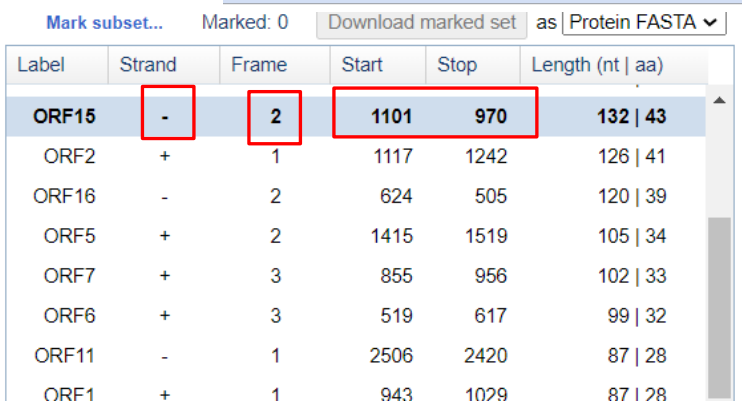
„reading“ DNA = translation



„reading“ DNA = translation

ORFfinder → looking for the longest ORF





Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF15	**-**	**2**	**1101**	**970**	**132 | 43**
ORF2	+	1	1117	1242	126 | 41
ORF16	-	2	624	505	120 | 39
ORF5	+	2	1415	1519	105 | 34
ORF7	+	3	855	956	102 | 33
ORF6	+	3	519	617	99 | 32
ORF11	-	1	2506	2420	87 | 28
ORF1	+	1	943	1029	87 | 28

Try

Look for ORFs:

- in „your“ nucleotide sequence (complete mRNA)

- in sequence (**ex1**) from Moodle and identify the correct ORF and the **source organism?**

„Nucleotide bioinformatics II“

Retrieving nucleotide sequences from databases (Genbank/NCBI)

Feature analysis: statistics, reverse complement, restriction analysis

Translation, identifying open reading frame

PCR primer design, rt-PCR

Secondary structure prediction

Sequence comparison, unknown sequence identification

Single Nucleotide Polymorphisms

DNA sequencing

Gene expression

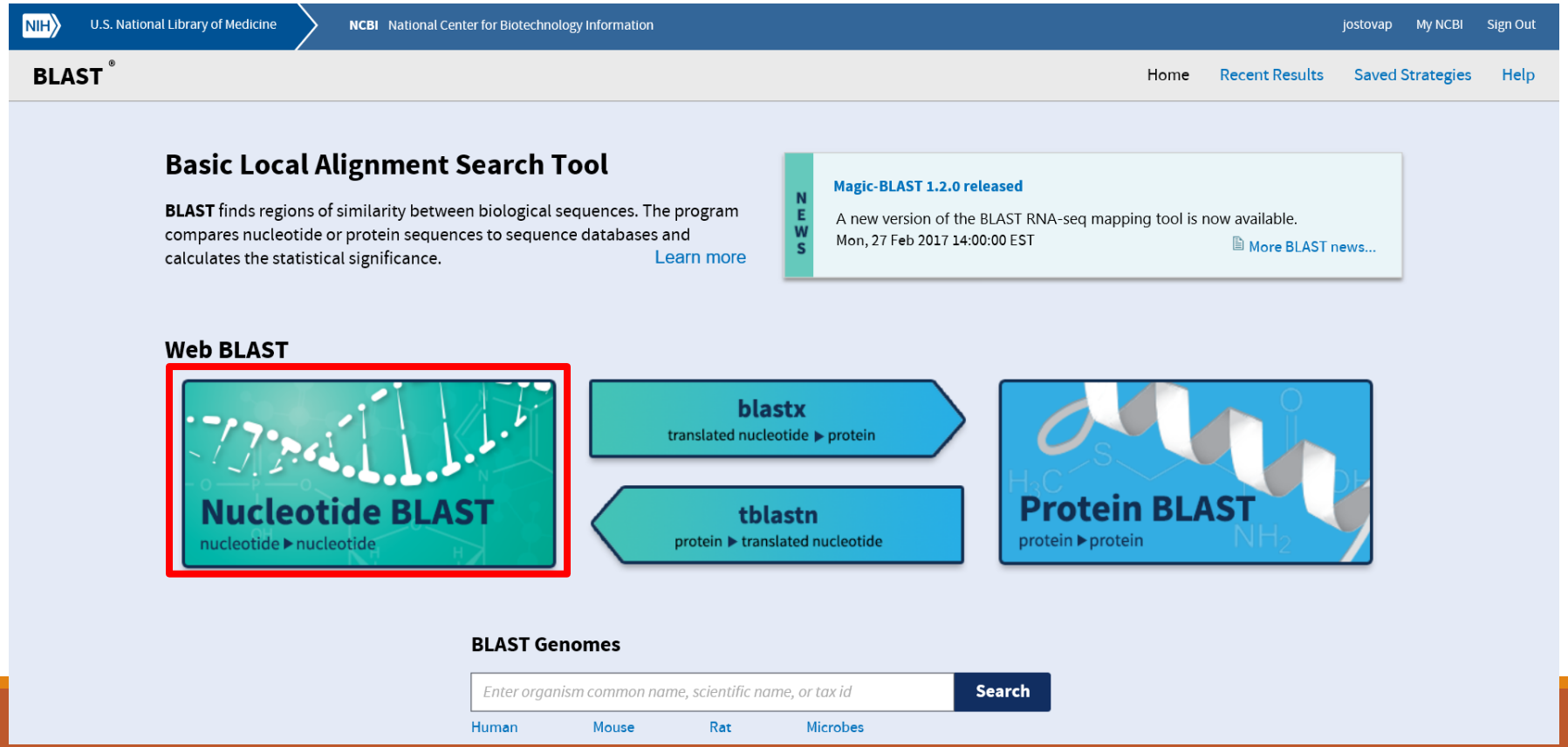
microRNA

Genomes….

….

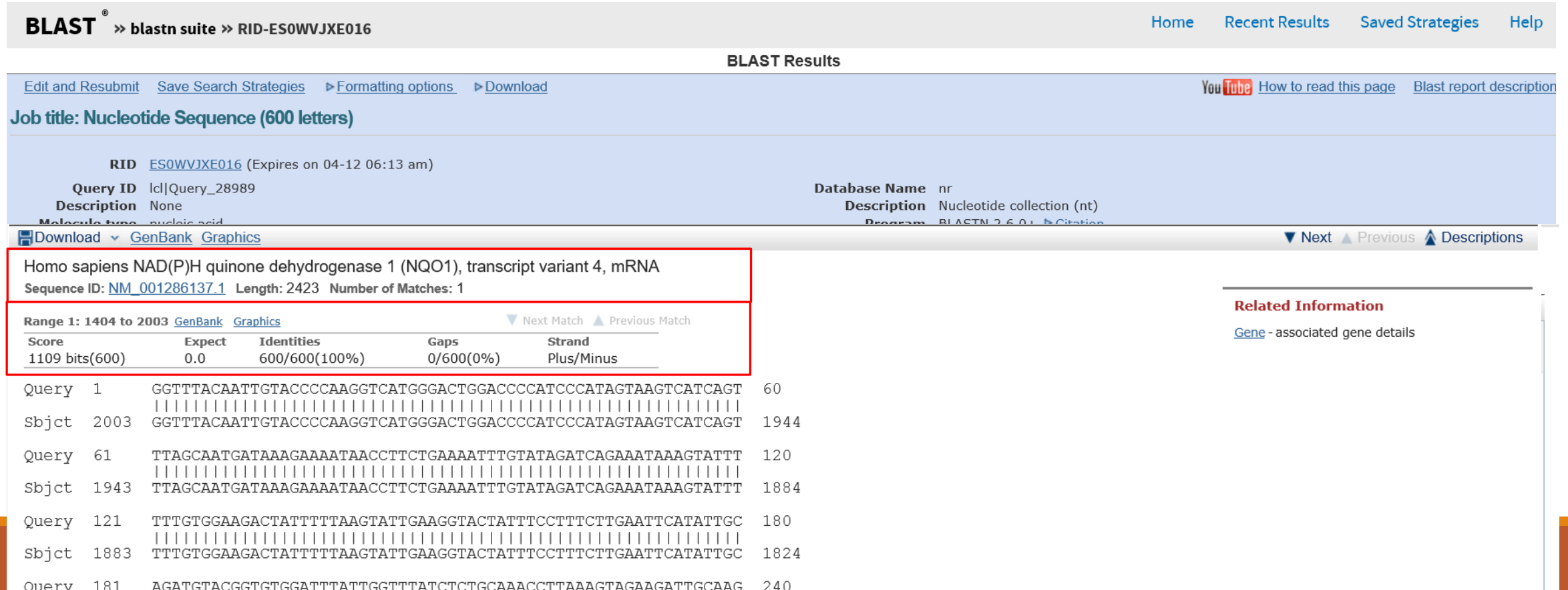
Unknown sequence identification – **BLASTn !**

→ looking for similarity: **BLASTn**



Unknown sequence identification

→ looking for similarity: BLASTn (or BLASTx)



Try

to identify sequence from Ex2 in Moodle
what is it and what is the **source organism?**

„Nucleotide bioinformatics II“

Retrieving nucleotide sequences from databases (Genbank/NCBI)

Feature analysis: statistics, reverse complement, restriction analysis

Translation, identifying open reading frame

PCR primer design, rt-PCR

Secondary structure prediction

Sequence comparison, unknown sequence identification

Single Nucleotide Polymorphisms

DNA sequencing

Gene expression

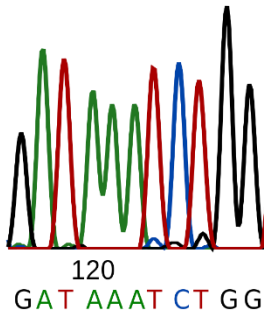
microRNA

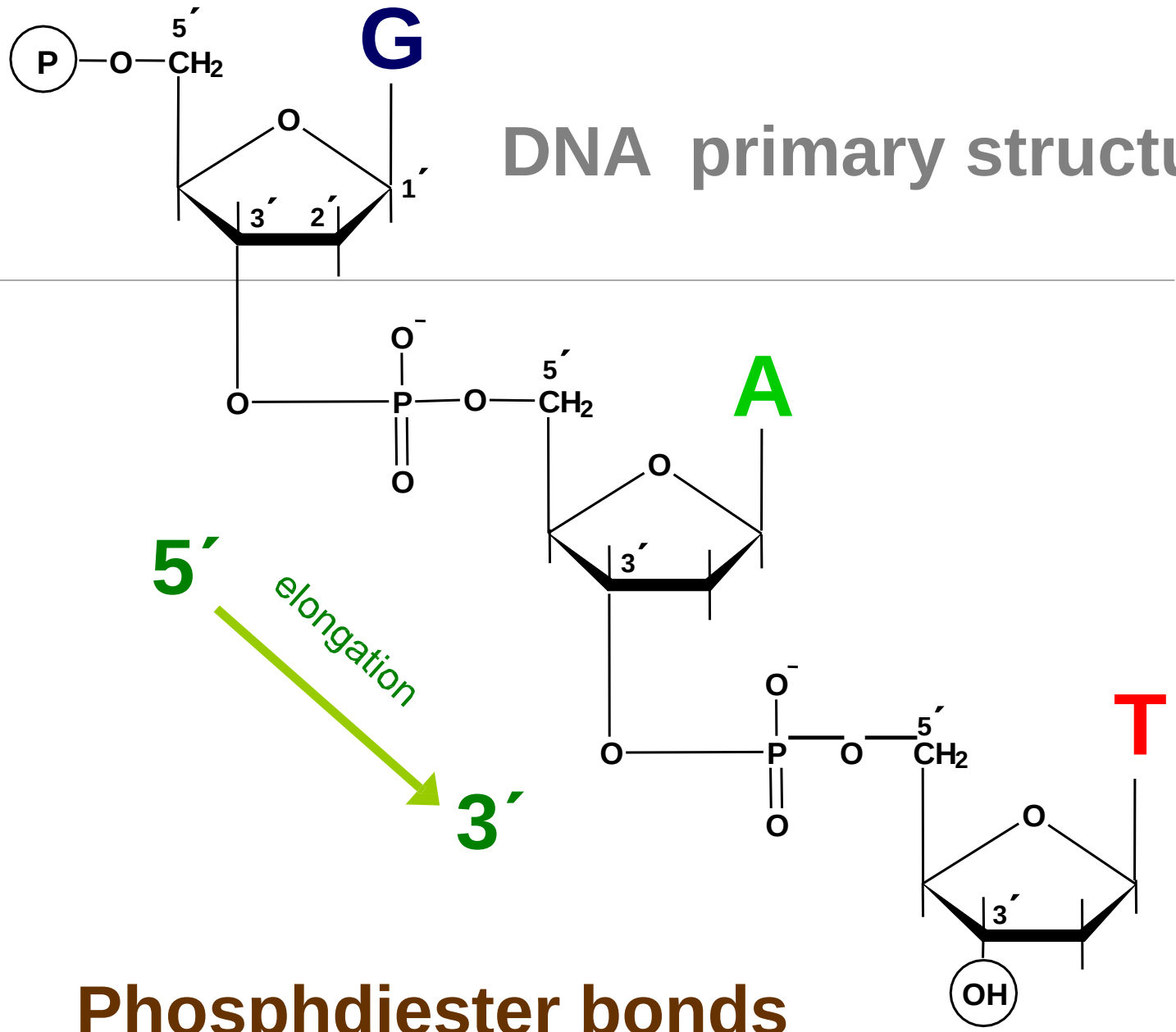
Genomes….

….

DNA primary structure

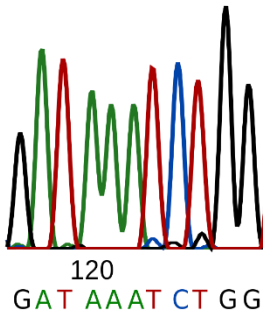
DNA sequencing





Phosphdiester bonds

DNA sequencing

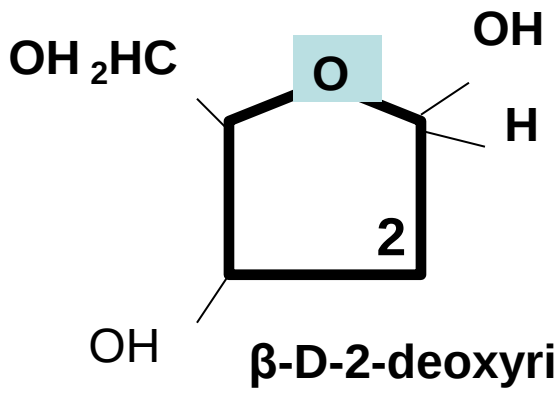


„Clasic“ Sanger sequencing (1977)

→ Lenght of sequence: ~1000nt

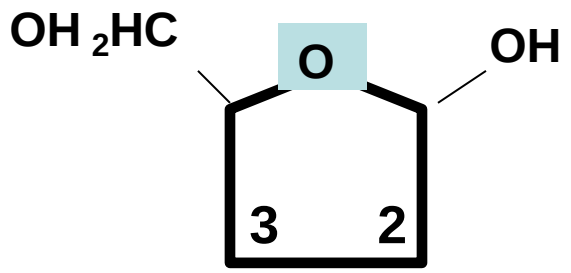
→ Output: „text“ 4 letters (ACTG)

Deoxynucleotides x **dideoxynucleotides**



β-D-2-deoxyribose

~1:100



β-D-2,3-dideoxyribose

DNA sequencing

Enzymatic „Sanger“ sequencing (1977) – sequencing by synthesis

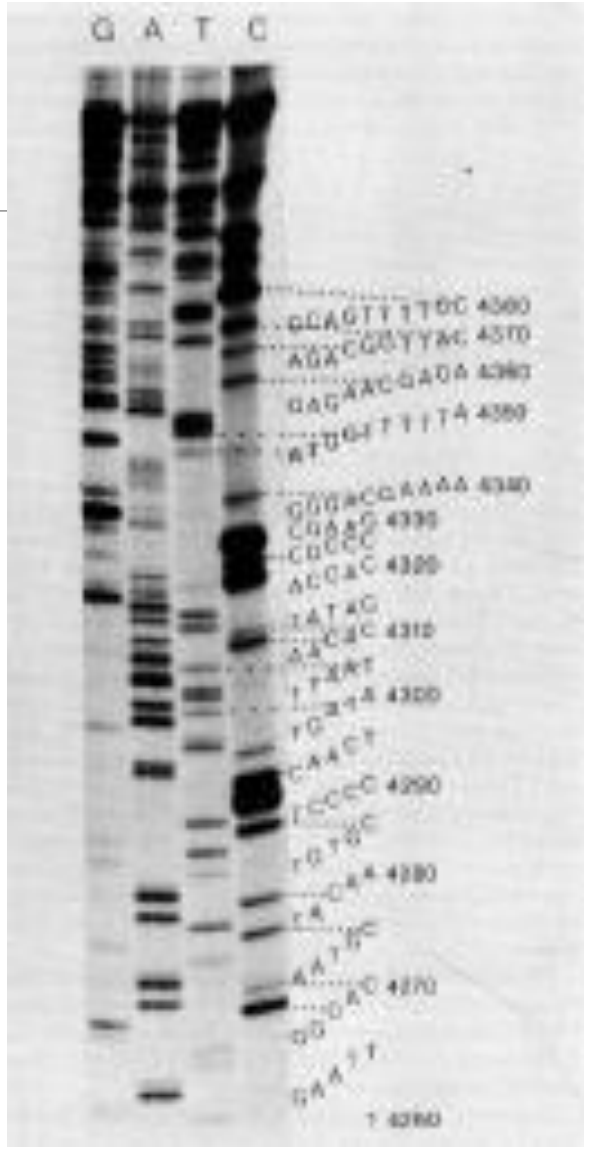
Asymetric DNA amplification, ^{32}P-dATP

Specific end of synthesis

G	A	T	C
Reaction mix 1 3 dNTPs 1:100 dGTP:ddGTP	***Reaction mix 2 3 dNTPs 1:1 dATP:ddATP***	***Reaction mix 3 3 dNTPs 1:100 dTTP:ddTTP***	***Reaction mix 4 3 dNTPs 1:50 dCTP:ddCTP***

PAGE/UREA > Southern blotting

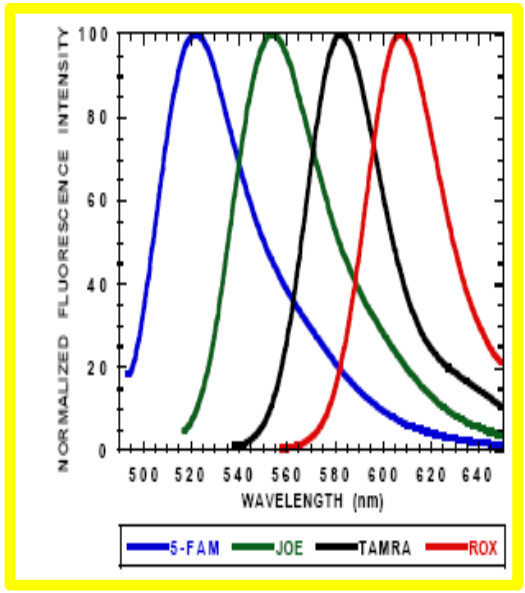
RA detection

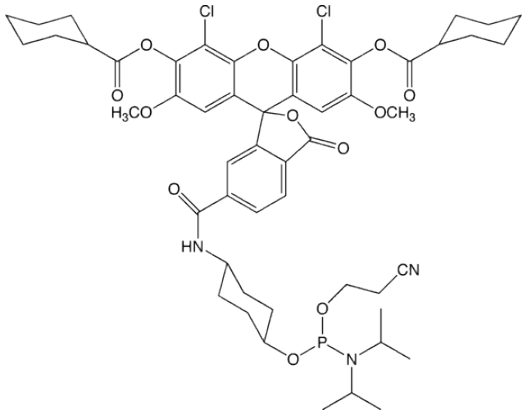


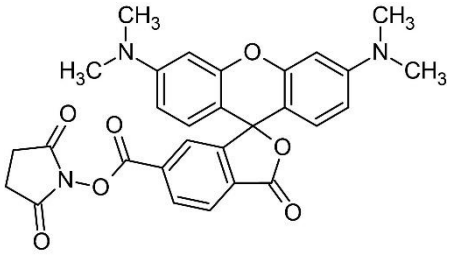
DNA sequencing

1980s new fluorophores enabled automatization

ddATP – JOE (550 nm) **A**
ddCTP – FAM (520 nm) **C**
ddGTP- TAMRA (585 nm) **G**
ddTTP – ROX (620 nm) **T**

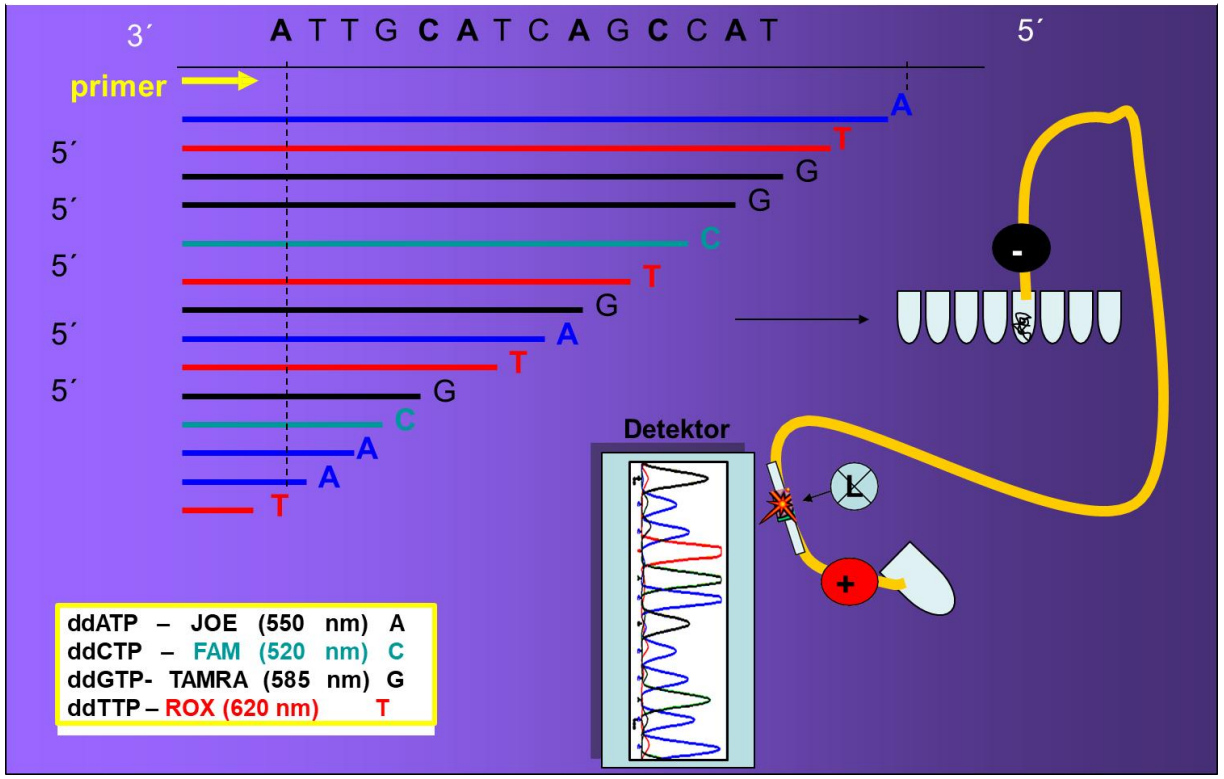






DNA sequencing

Princip: Sanger sequencing and capillary electrophoresis



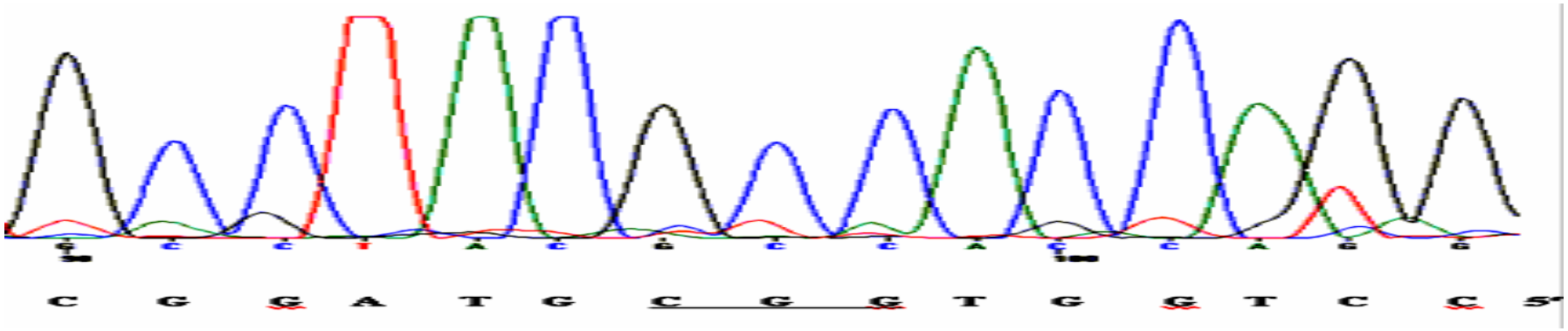
Sequence data analysis

Chromas → enable direct sequence data analysis (.scf, .abi, .ab1)

„plain text“: CGGATGCGGTGGTCG

„fasta“: >identificator
CGGATGCGGTGGTCG

„sequencing formate“(.scf, .abi, .ab1)



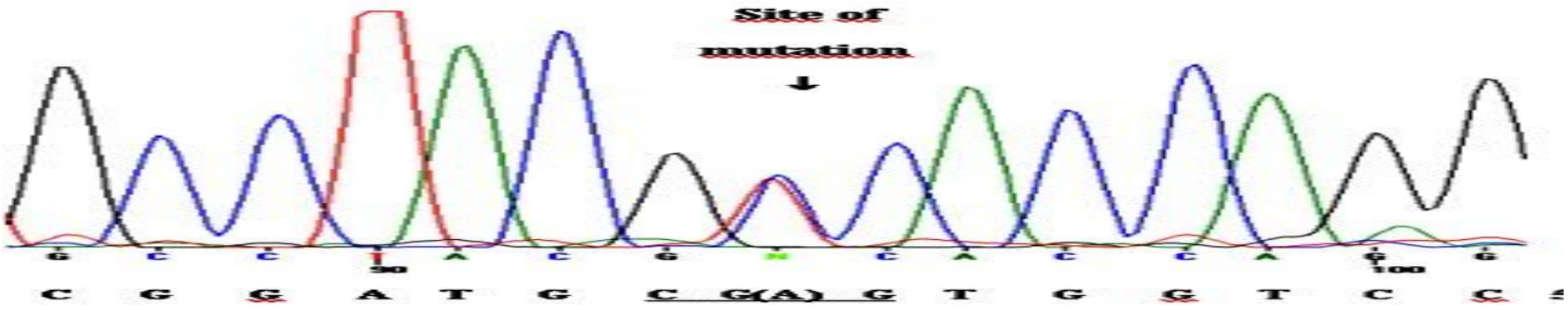
Sequence data analysis

Chromas → enable direct sequence data analysis (.scf, .abi, .ab1)

„plain text“: CGGATGCNGTGGTCG

„fasta“: >identifikace
CGGATGC**N**GTGGTCG

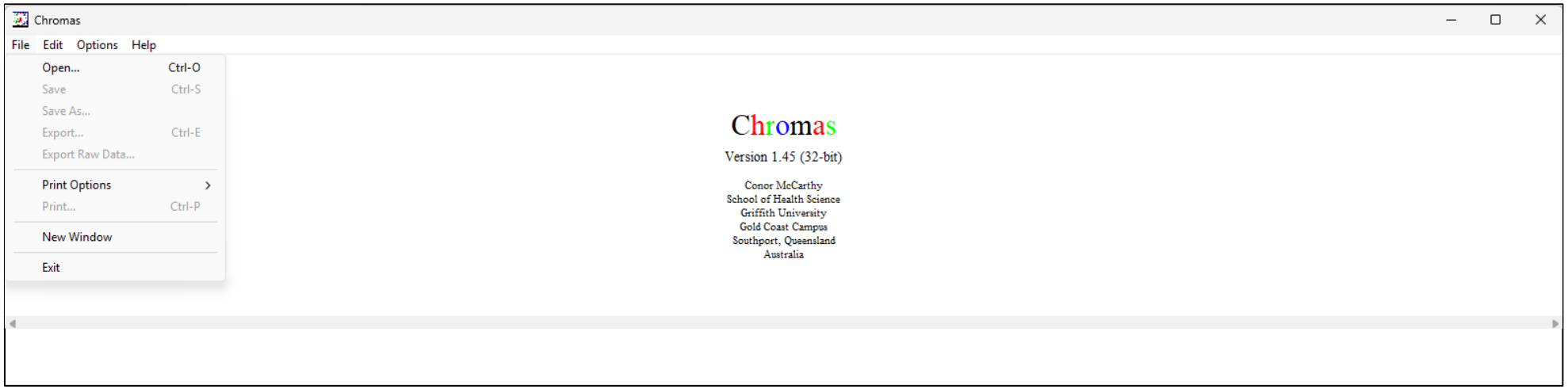
„sekvenační formát“(.scf, .abi, .ab1)



Sequence data analysis

Chromas → enable direct sequence data analysis (.scf, .abi, .ab1)

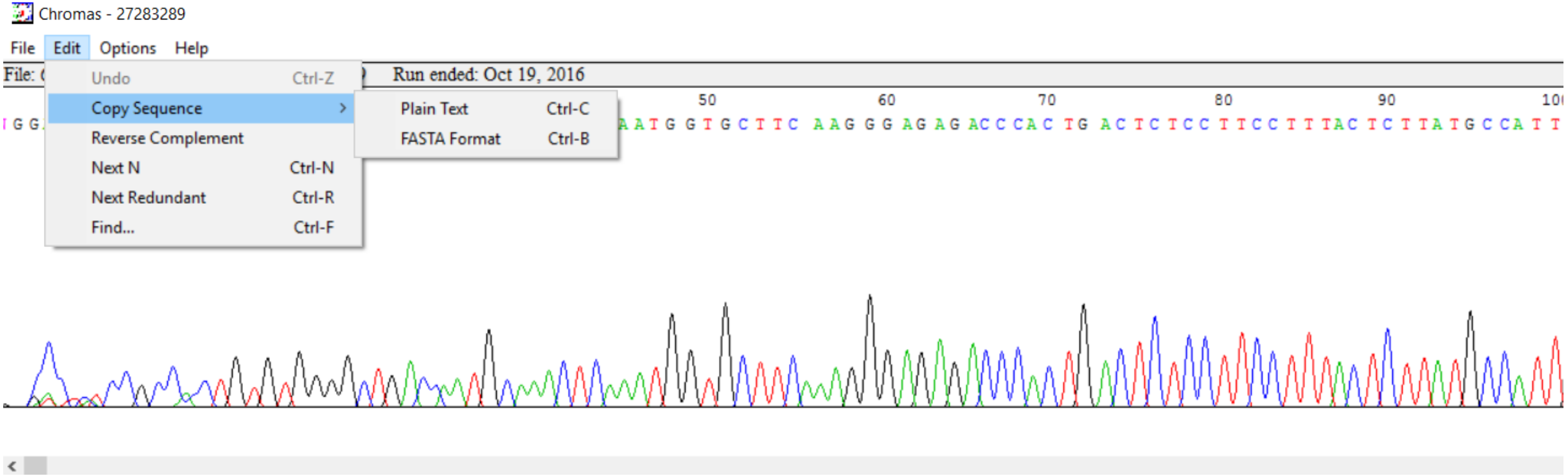
- the sequence has to be first saved and then open in chromas



Sequence data analysis

Chromas → enable direct sequence data analysis (.scf, .abi, .ab1)

Edit-copy sequence-FASTA or plain (is copied into memory and can be directly pasted into a program)



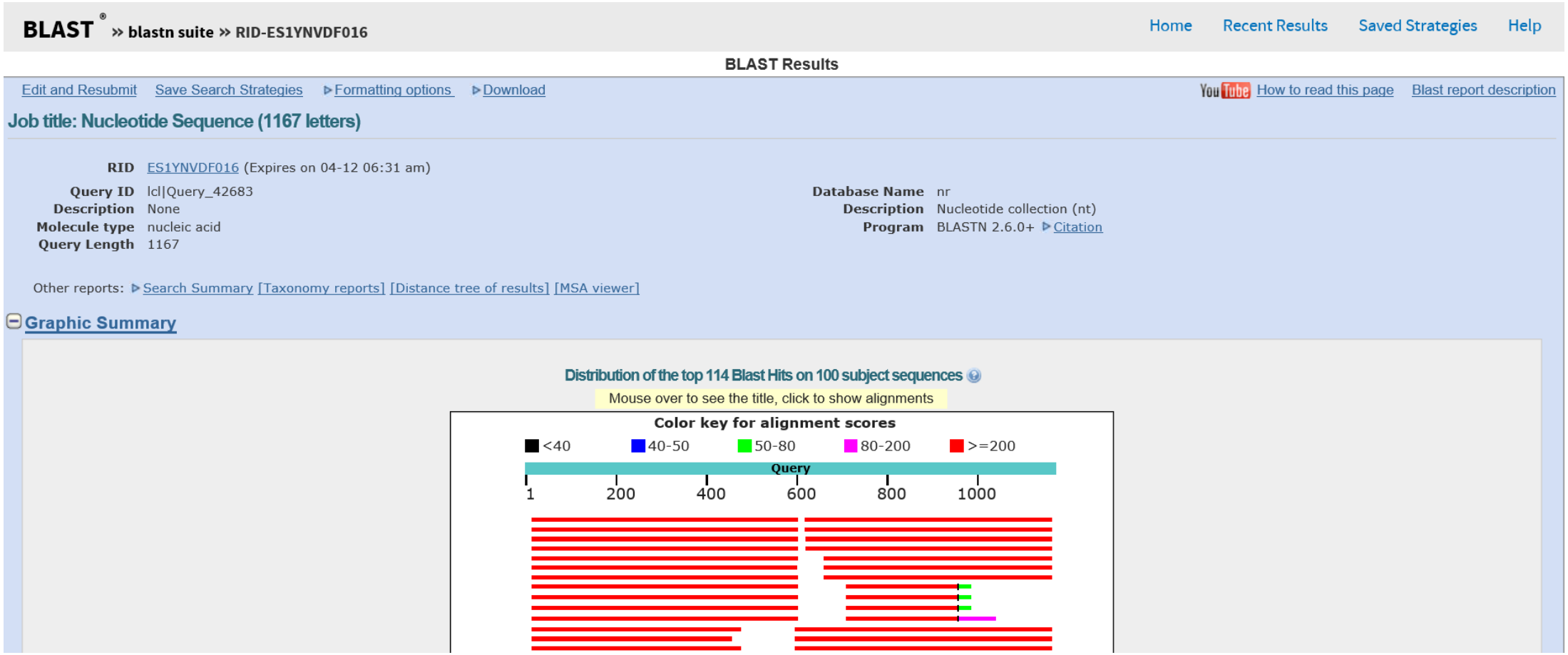
Try

Run „chromas“ from Moodle

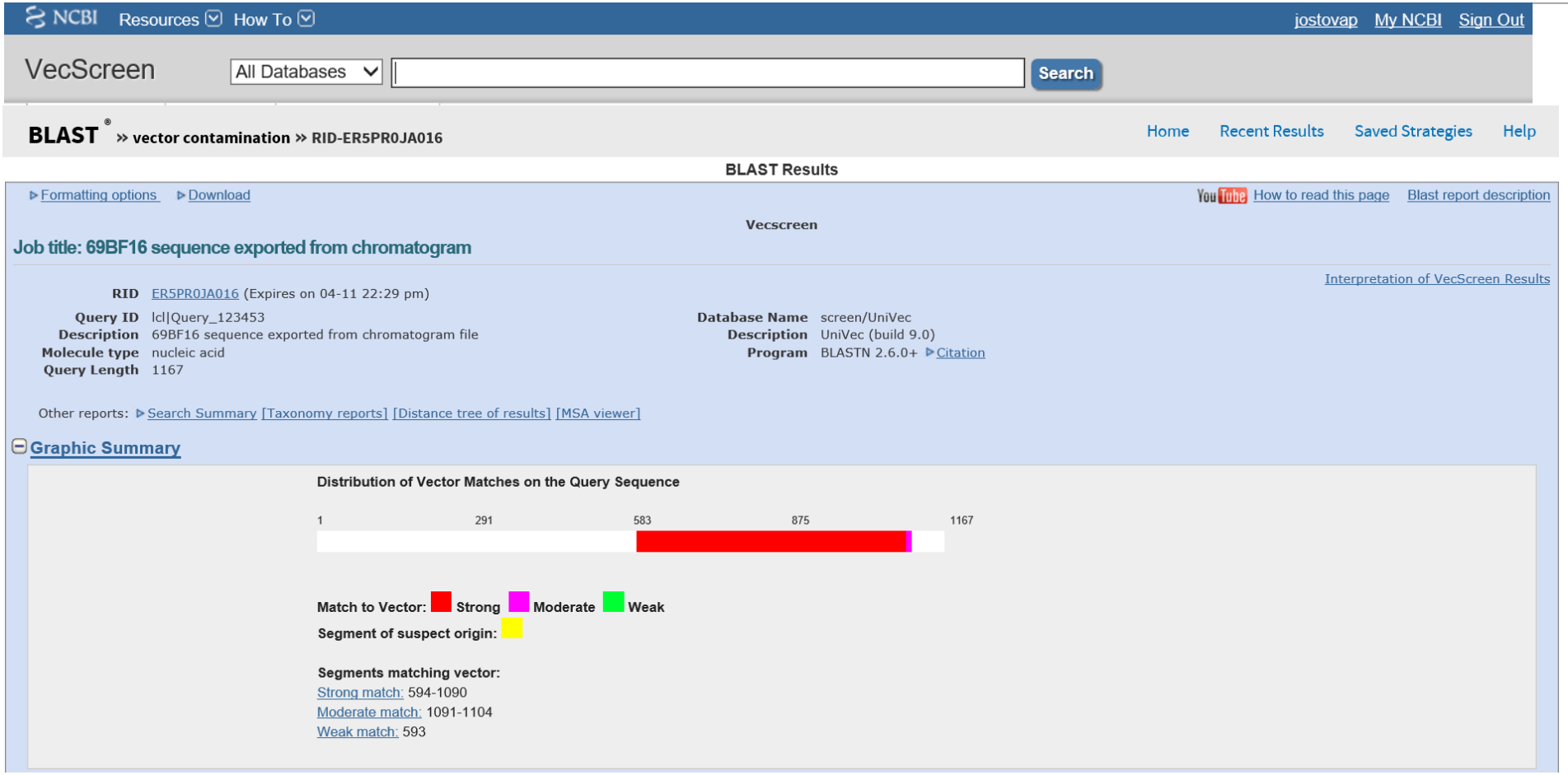
Download and store Ex3 sequence data from Moodle.
Open the sequence in chromas.
Export the sequence from chromas and identify it.

What does it code?
From which organism does it probably come from?

Sequence data analysis (Ex3) - Blastn

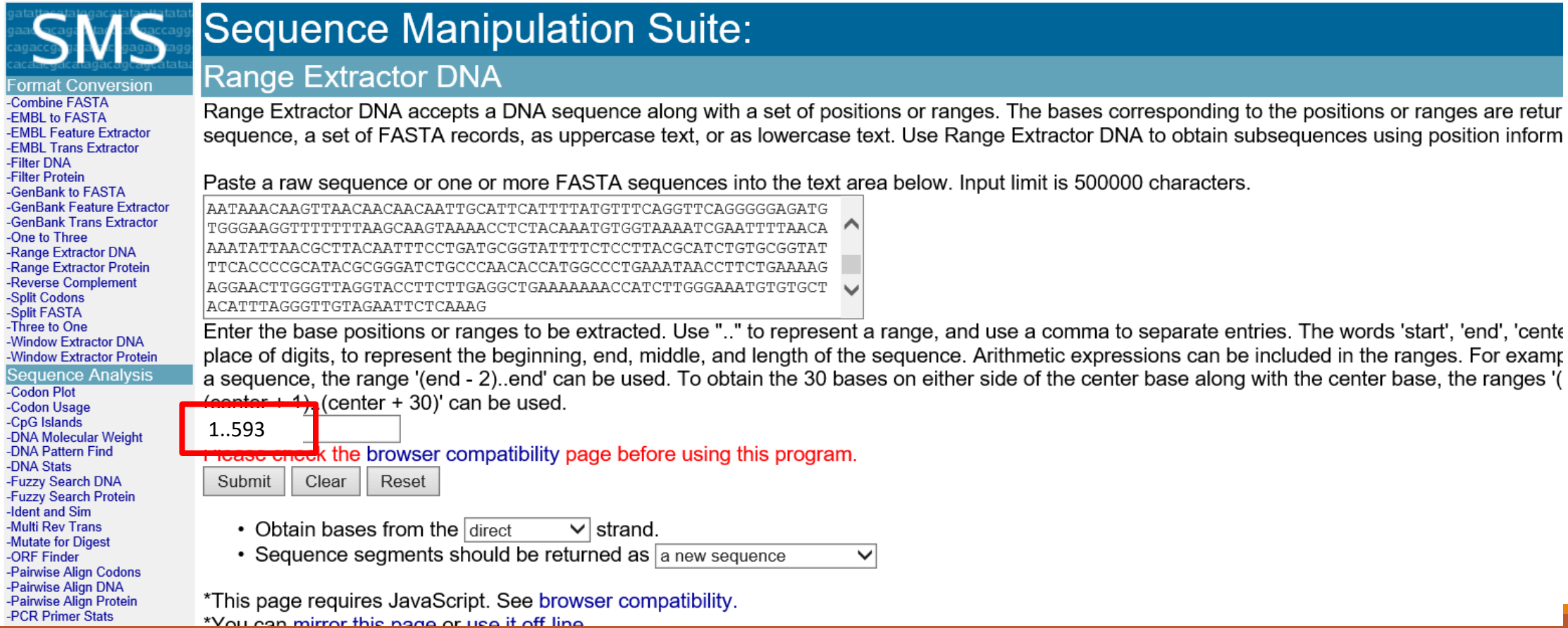


Sequence data analysis (Ex3) - Vecsreen



Sequence data analysis: purifing sequence of vector

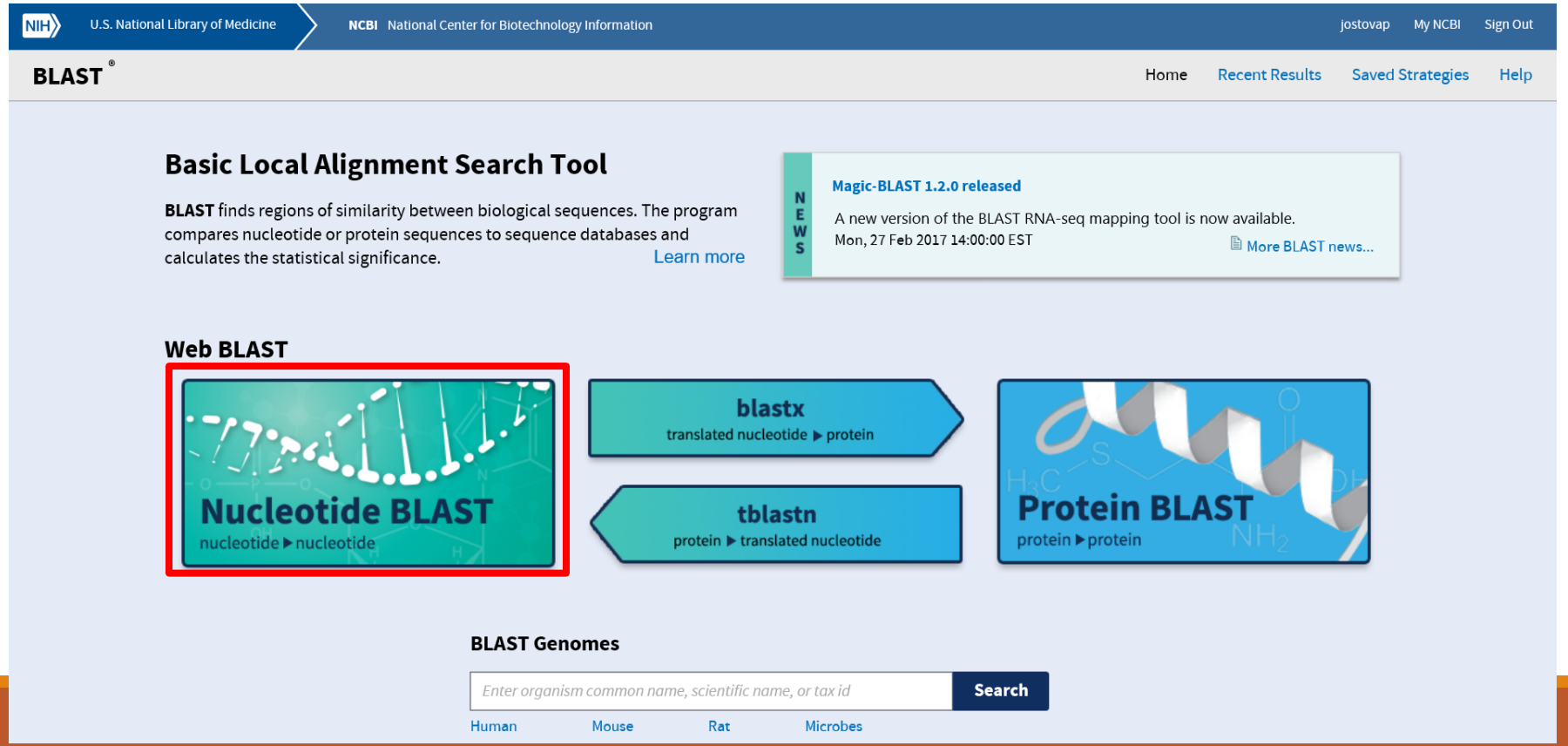
SMS „Range Extractor DNA“



Unknown sequence identification – BLASTn !

- if the sequence does not have easily recognizable ORF ?

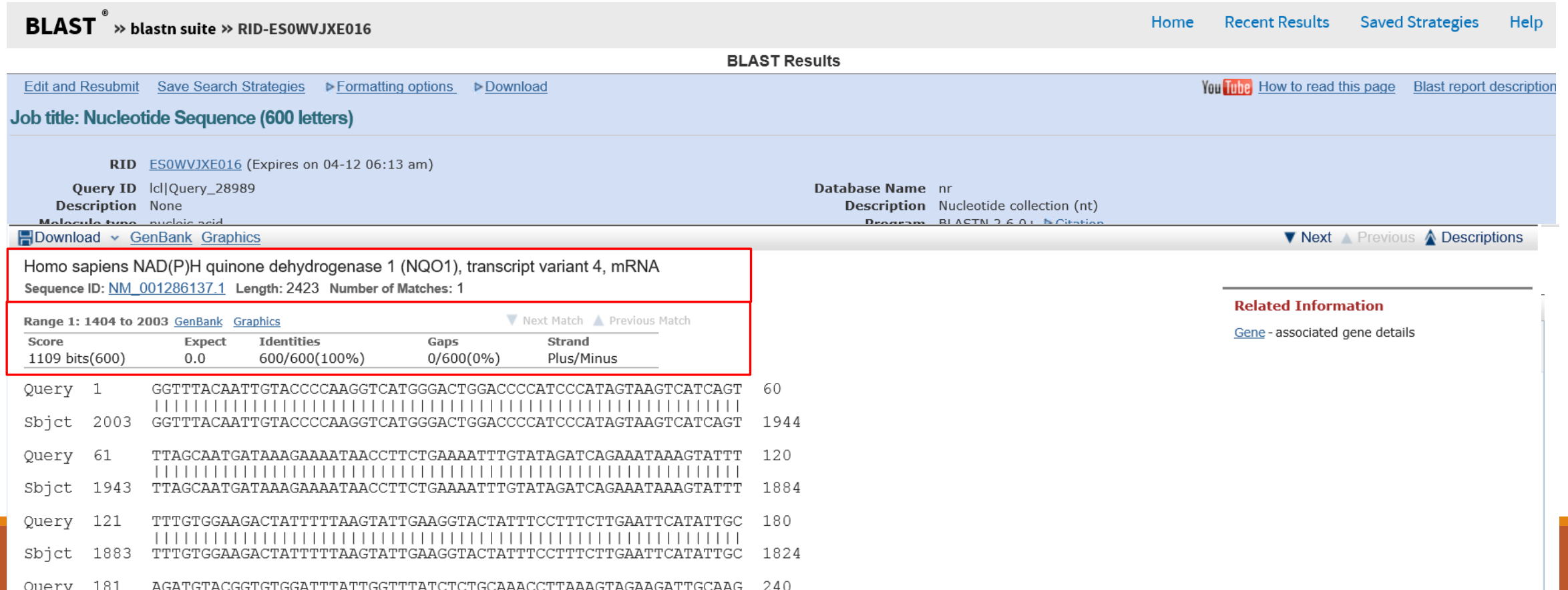
→ looking for similarity: BLASTn (or BLASTx)



Unknown sequence identification

- if the sequence does not have easily recognizable ORF ?

→ looking for similarity: BLASTn (or BLASTx)



Homework 6

Work with „your“ nucleotide sequence.

---

1) Compare your mRNA and CDS sequence

2) Translate „your“ nucleotide sequence (mRNA), in which ORF is the CDS?

3) Download unknown sequence „**Homework 6.ab1**“and open it in chromas.

*Check for vector contamination and identify „pure“ sequence

*Identify the sequence (pure) and the organism with BLASTn.

4) View „PCR Primer Design“ https://www.youtube.com/watch?v=c-f1H07D_70

Homework 6-example

DÚ6

1) NQO1 (mRNA): Čtecí rámec +3

Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF10	+	3	192	1016	825 | 274

2)

3)