Introduction to applied bioinformatics

PETRA MATOUŠKOVÁ

2025/2026

„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

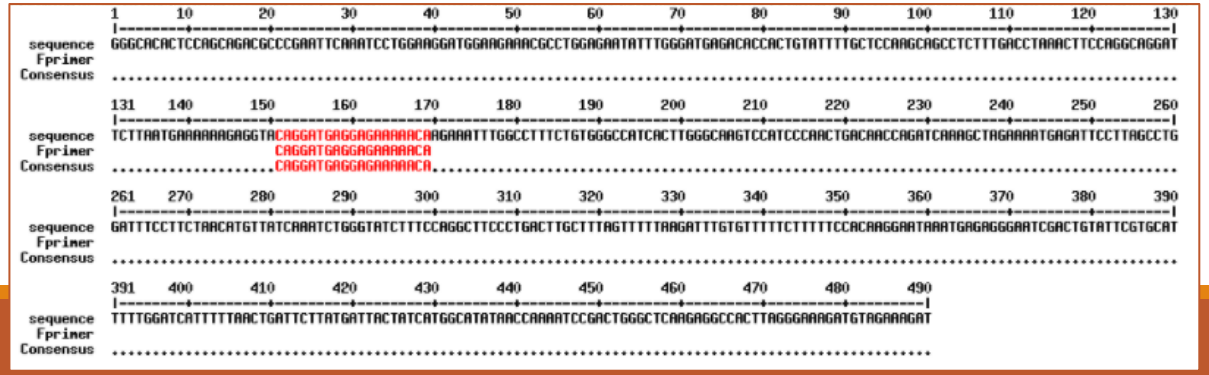
e.g. multalin

→ 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:

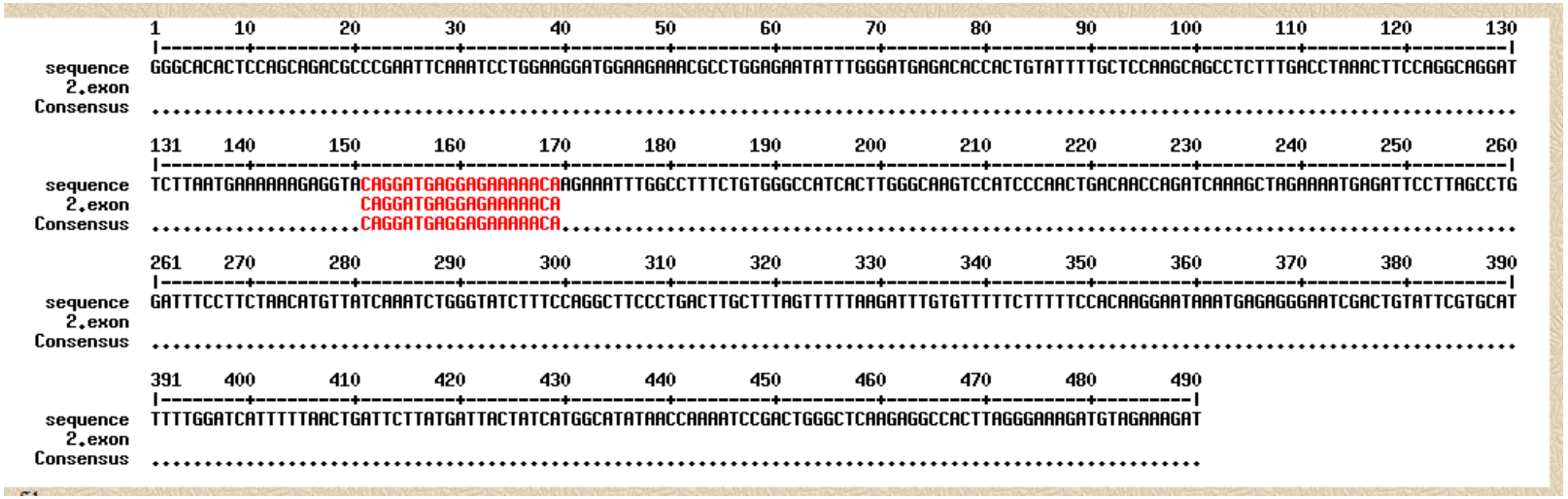
- A: compare long and short sequence (default Blosum matrix)

Sequence comparison

open Multalin

from Examples:

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



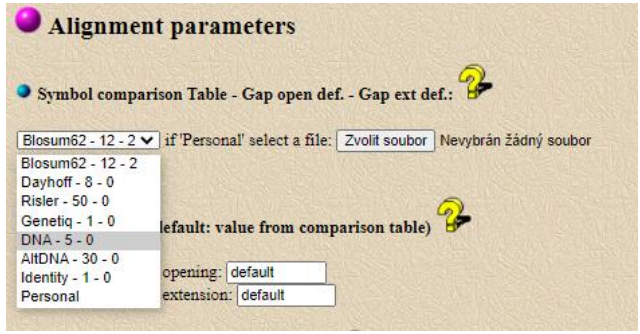
- Using both matrices would look like the same

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**

```
                1       10        20        30
                |-------+---------+---------+
NM_015696.4     GTCTTTGCCCTCGCGACGCCGCCACCTCCG
69BF16                NGGCATTCTCCCGCACTGTGTGGG
Consensus       ......gccCacgCgaCccCaCcaccTccG

              131     140       150       160
                |-------+---------+---------+
NM_015696.4     TCAACATCCGGGGCAAACTGGTGTCGCTGG
69BF16          AAAATTCTAGTATTTTGATTATTTGAATCT
Consensus       aaAAcaccaGgagcaaaaTgaTgTcaaTcg

              261     270       280       290
                |-------+---------+---------+
NM_015696.4     GGGCCCCCACCACTTTAACGTGCTCGCCTT
69BF16          AGTATATCAAG-CAATAATCTCCCACCCAA
Consensus       aGgacacCAac.CaaTAAccTcCcacCCaa

              391     400       410       420
                |-------+---------+---------+
NM_015696.4     AAGATTGCAGTCACCGGTACTGGTGCCCAT
69BF16          AGGCATTCAATGAACATTTTTTG---CATA
Consensus       AaGaaTgCAaTcAaCagTacTgG...Caaa

              521     530       540       550
                |-------+---------+---------+
NM_015696.4     ACCCAACTGTGTCAGTGGAGGAGGTCAGAC
69BF16          CTGTATTATTTTCTTCATTACAAAGAAAT
```

```
              261     270       280       290       300       310       320       330       340       350       360       370       380       3
                |-------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
NM_015696.4     GGGCCCCCACCACTTTAACGTGCTCGCCTTCCCCTGCAACCAGTTTGGCCAACAGGAGCCTGACAGCAACAAGGAGATTGAGAGCTTTGCCCGCCGCACCTACAGTGTCTCATTCCCCATGTTTAGCA
69BF16
Consensus       ..........................................................................................................................................

              391     400       410       420       430       440       450       460       470       480       490       500       510       5
                |-------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
NM_015696.4     ATTGCAGTCACCGGTACTGGTGCCCATCCTGCCTTCAAGTACCTGGCCCAGACTTCTGGGAAGGAGCCCACCTGGAACTTCTGGAAGTACCTAGTAGCCCCAGATGGAAAGGTGGTAGGGGCTTGGGA
69BF16
Consensus       ..........................................................................................................................................

              521     530       540       550       560       570       580       590       600       610       620       630       640       6
                |-------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
NM_015696.4     CAACTGTGTCAGTGGAGGAGGTCAGACCCCAGATCACAGCGCTCGTGAGGAAGCTCATCCTACTGAAGCGAGAAGACTTATAACCACCGCGTCTCCTCCTCCACCACCTCATCCCGCCCACCTGTGTG
69BF16                                                                                                                  NGGCATTCTCCCGCAC---TGTGTG
Consensus       ..........................................................................................ncaCaTcaTCCCGCaC...TGTGTG

              651     660       670       680       690       700       710       720       730       740       750       760       770       7
                |-------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
NM_015696.4     GCTG-ACCAA-TGCAAACTCAAATGGTGCTTCAAGGGAGAGACCCACTGACTCTCCTTCCTTTACTCTTATGCCATTGGTCCCATCATTCTTGTGGGGGAAAAATTCTAGTATTTTGATTATTTGAA
69BF16          GCTGGACCAAATGCAAACTCAA-TGGTGCTTCAA-GGGAGAGACCCACTGACTCTCCTTCCTTTACTCTTATGCCATTGGTCCCATCATTCTTGTGGGGGAAAAATTCTAGTATTTTGATTATTTGAA
Consensus       GCTG.ACCAA.TGCAAACTCAA.TGGTGCTTCAA.GGGAGAGACCCACTGACTCTCCTTCCTTTACTCTTATGCCATTGGTCCCATCATTCTTGTGGGGGAAAAATTCTAGTATTTTGATTATTTGAA

              781     790       800       810       820       830       840       850       860       870       880       890       900       9
                |-------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
NM_015696.4     TTACAGCAACAAATAGGAACTCCTGGCCAATGAGAGCTCTTGACCAGTGAATCACCAGCCGATACGAACGTCTTGCCAACAAAAATGTGTGGCAAATAGAAGTATATCAAGCAATAATCTCCCACCCA
```

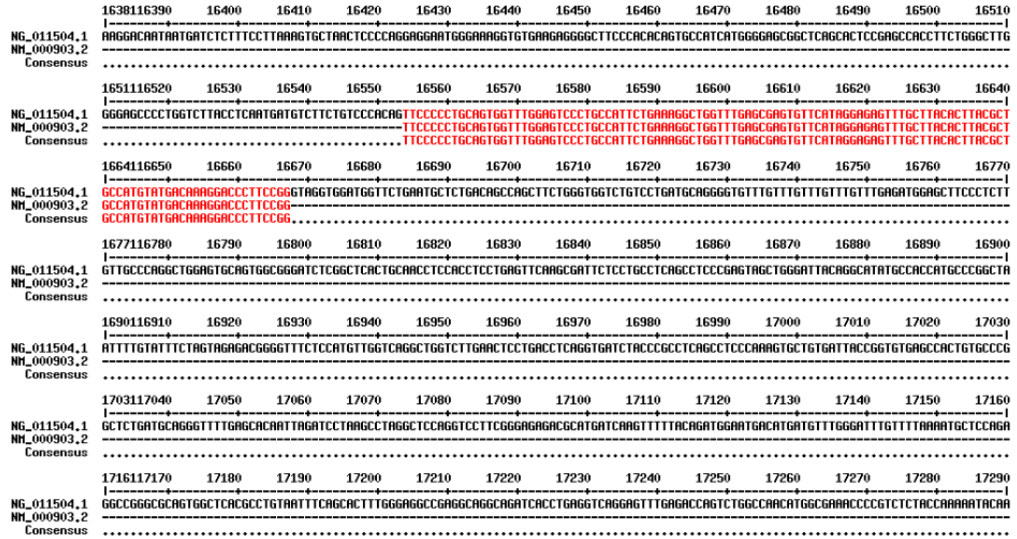
Nucleotide sequence comparison

-Analogous to protein comparison

e.g. multalin

→ Comparing genomic DNA and CDS (or 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**

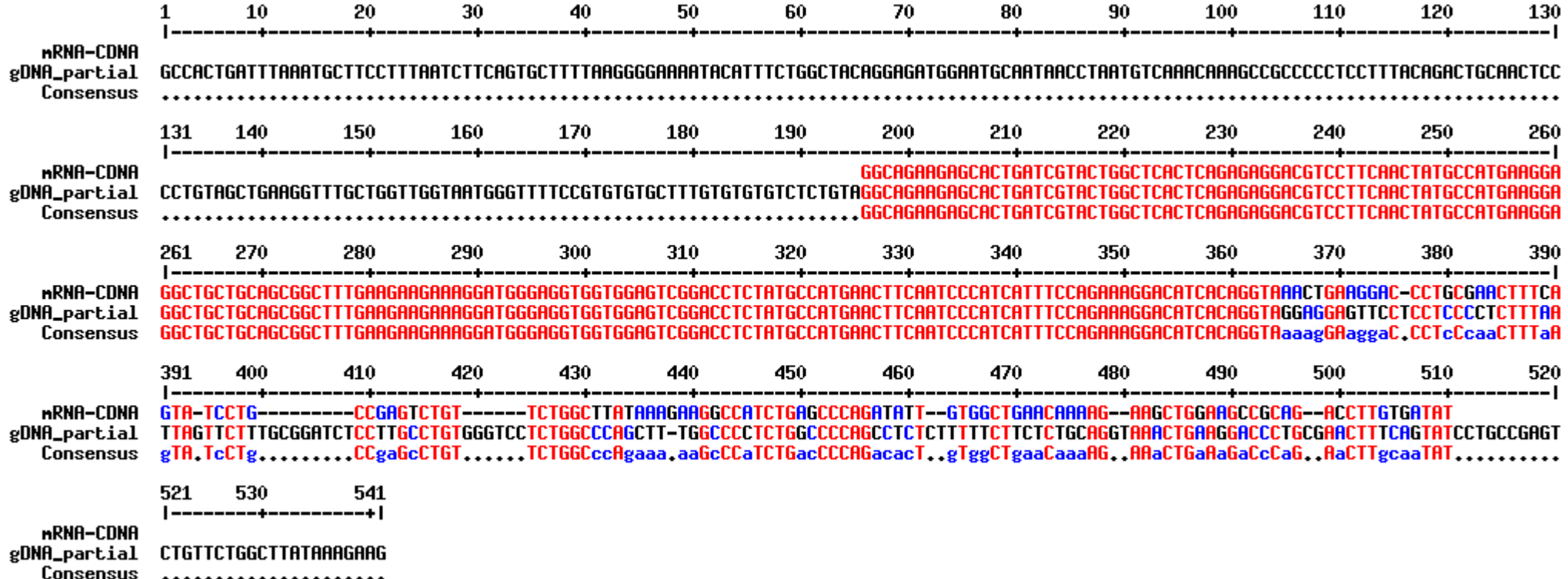
  see the difference?

Sequence comparison

open Multalin

from Examples:

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

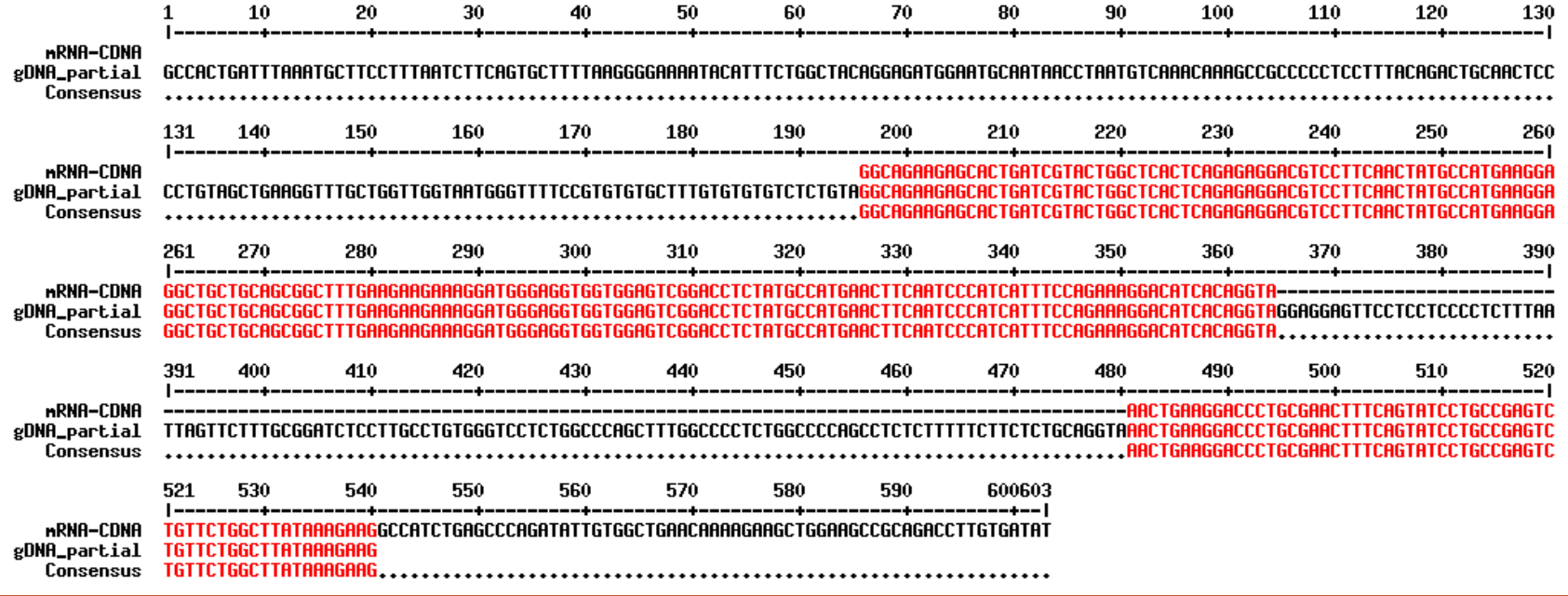


Sequence comparison

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 **TTT** AA

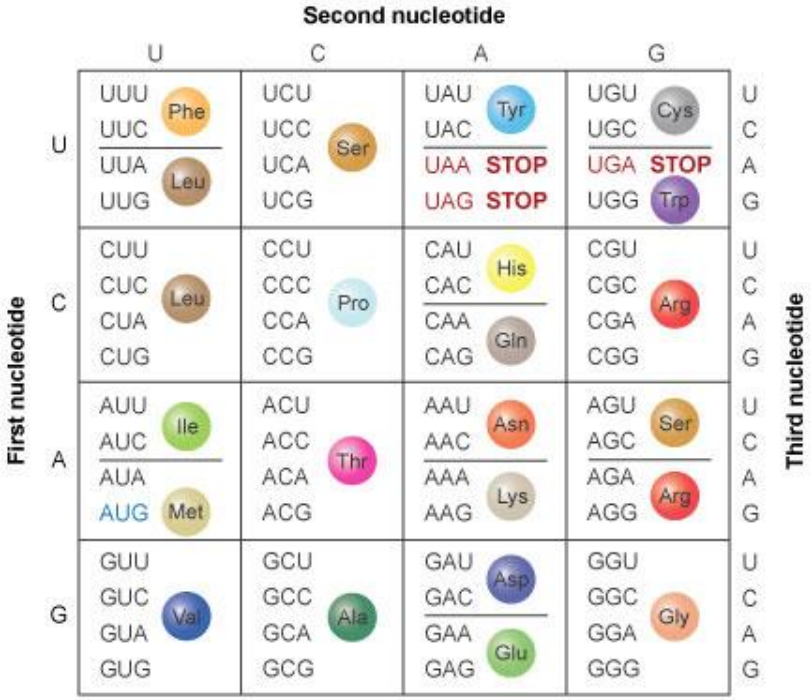
AT **GGA** AGT ATT TAA AGC GCC ACC TAT TTA A

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

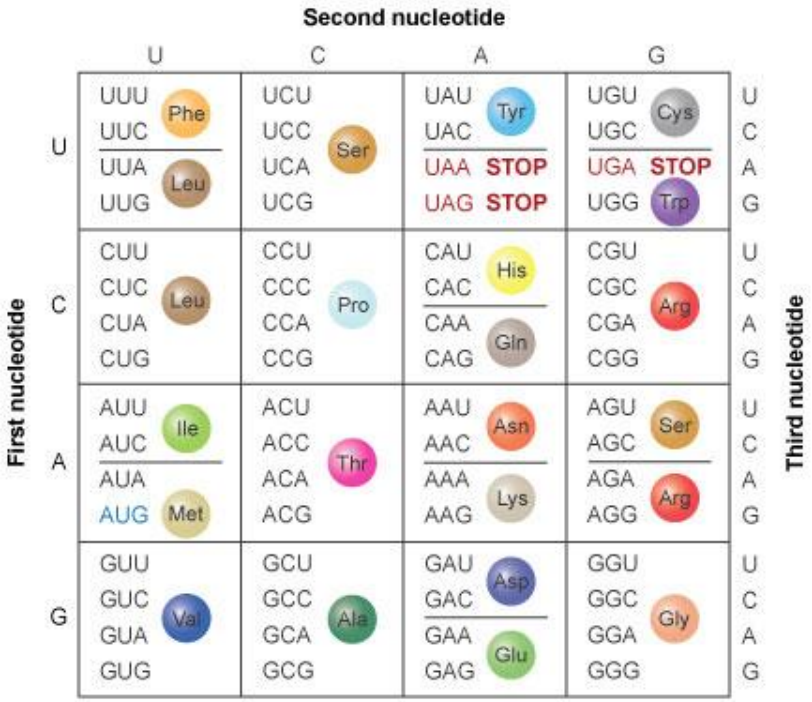
T TAA ATA GGT GGC GCT TTA AAT ACT TCC AT

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

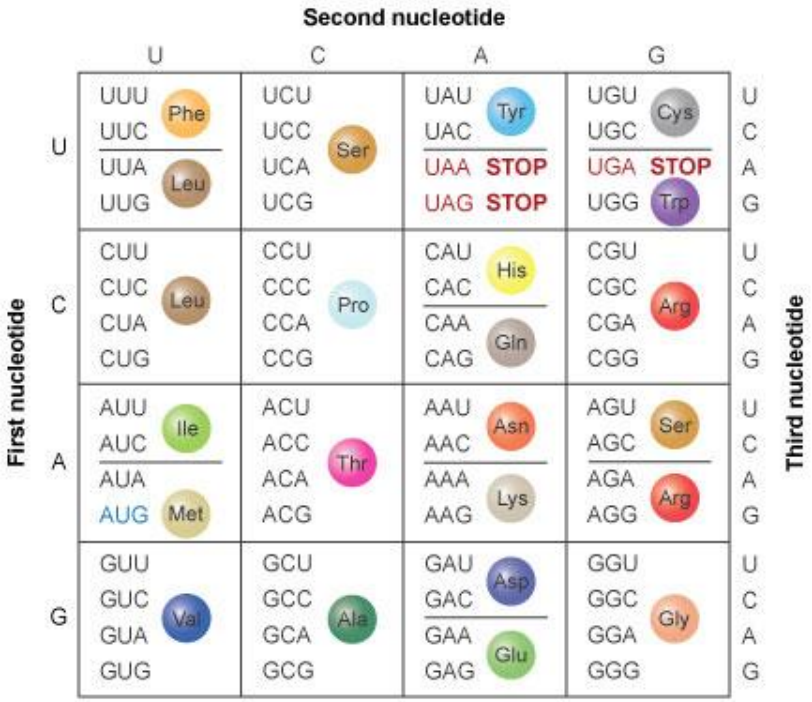
T TAA ATA GGT GGC GCT TTA AAT ACT TCC AT

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)

SMS

Format Conversion
-Combine FASTA
-EMBL to FASTA
-EMBL Feature Extractor
-EMBL Trans Extractor
-Filter DNA
-Filter Protein
-GenBank to FASTA
-GenBank Feature Extractor
-GenBank Trans Extractor
-One to Three
-Range Extractor DNA
-Range Extractor Protein
-Reverse Complement
-Split Codons
-Split FASTA
-Three to One
-Window Extractor DNA
-Window Extractor Protein

Sequence Analysis
-Codon Plot
-Codon Usage
-CpG Islands
-DNA Molecular Weight
-DNA Pattern Find
-DNA Stats
-Fuzzy Search DNA
-Fuzzy Search Protein
-Ident and Sim
-Multi Rev Trans
-Mutate for Digest
-ORF Finder
-Pairwise Align Codons
-Pairwise Align DNA
-Pairwise Align Protein
-PCR Primer Stats
-PCR Products
-Protein GRAVY
-Protein Isoelectric Point
-Protein Molecular Weight
-Protein Pattern Find
-Protein Stats

Sequence Manipulation Suite:

Translate

Translate accepts a DNA sequence and converts it into a protein in the reading frame you specify. Translate supports the entire IUPAC alphabet and several genetic codes.

Paste a raw sequence or one or more FASTA sequences into the text area below. Input limit is 200000 characters.

```
TTAAATAGGTGGCGCTTTAAATACTTCCAT
```

Please check the … bility pa … sin

Submit Clear

reading frame 1
reading frame 2
reading frame 3

direct
reverse

standard (1)
vertebrate mitochondrial (2)
yeast mitochondrial (3)
mold mitochondrial (4)
invertebrate mitochondrial (5)
ciliate nuclear (6)
echinoderm mitochondrial (9)
euplotid nuclear (10)
bacterial (11)
alternative yeast nuclear (12)
ascidian mitochondrial (13)
flatworm mitochondrial (14)
Blepharisma macronuclear (15)
chlorophycean mitochondrial (16)
trematode mitochondrial (21)
Scenedesmus obliquus mitochondrial (22)
Thraustochytrium mitochondrial (23)

• Translat… rect … str…
• Use the … ge…

*This page rec… compatibili…
*You can mirr…

Fri Jun 17 16:17:06 20…
Valid XHTML 1.0; Valid C…

Translate results

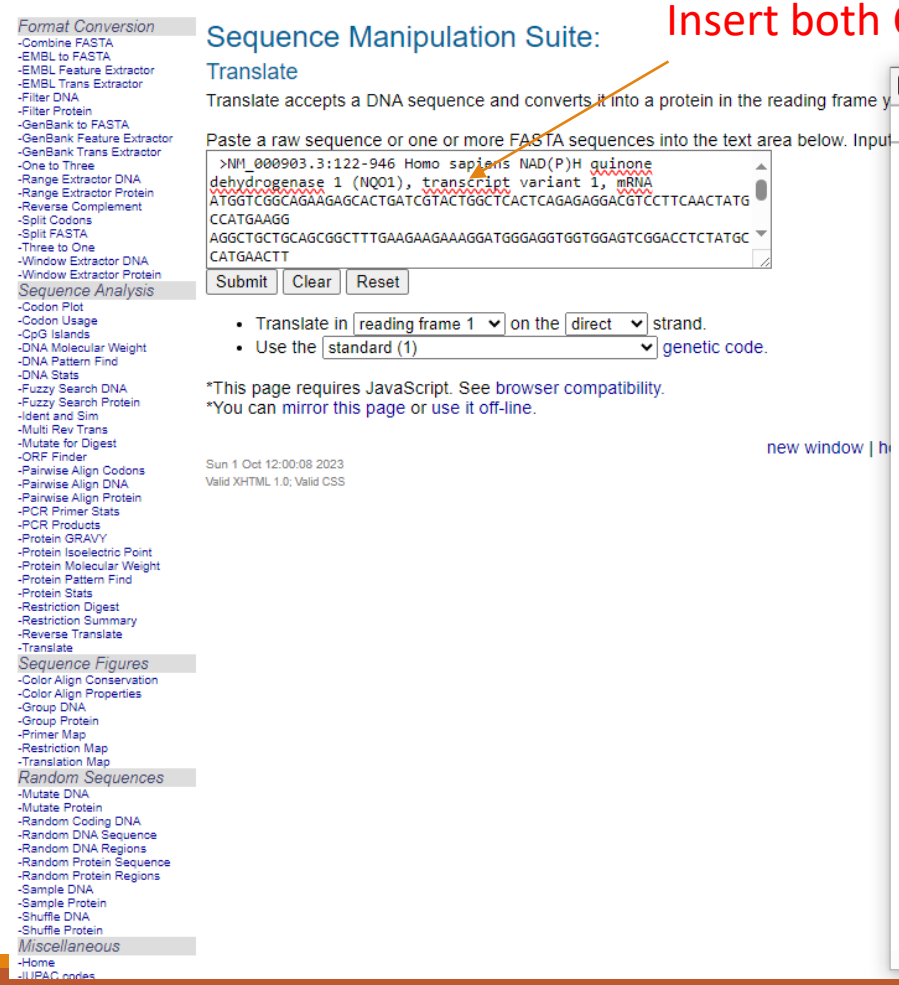
```
>rf 1 Untitled
LSQQIQLASVIRPCSPFFL*PLLCCYHGSDTQFHHQ*SDRFHQDVSSLVRQRNCLLHQQL
SSRFGQ*SCLFRYRLYDMVFR*SQIRSHVHNLIMS*LNQCLDPQVFRLDTYDHSESSGRH
VLPFVHVHASNHQVLMLIDDVCVSTQQVGSL
```

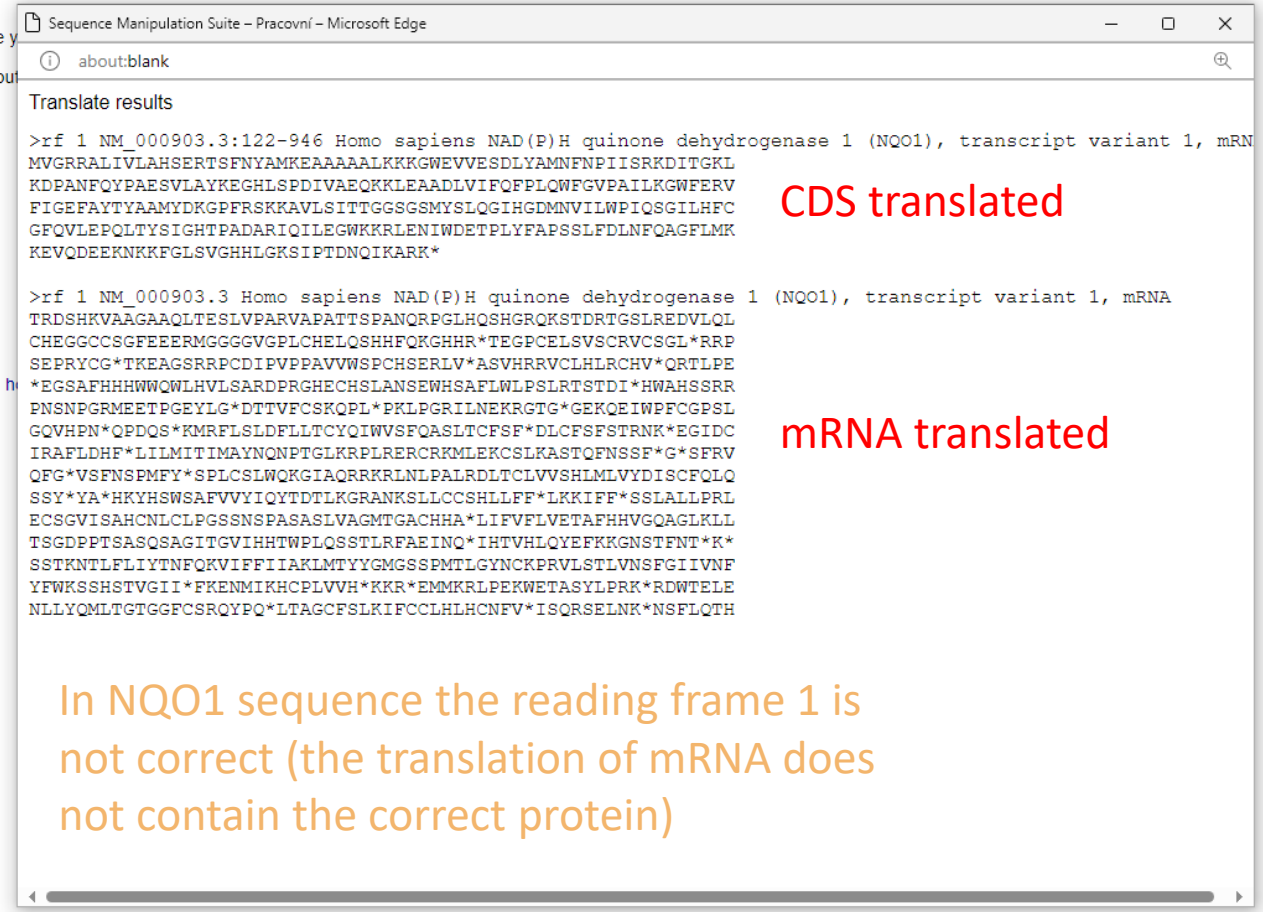
Translate results

```
>rf 2 Untitled
CRSRYNLLR*YVHVVHFFFSHCCVVTTEATHNFIINKATVFIKT*VRLCDNVTVFFISS*
VVDLVSDRVCFDIDFTIWCFDKAKFVHTCIT***VD*TNVWTLRCFDWTHTTIVRVVDVT
YFHLCTFTRQTTRS*C**TTFVCQLSKWVRL
```

„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)

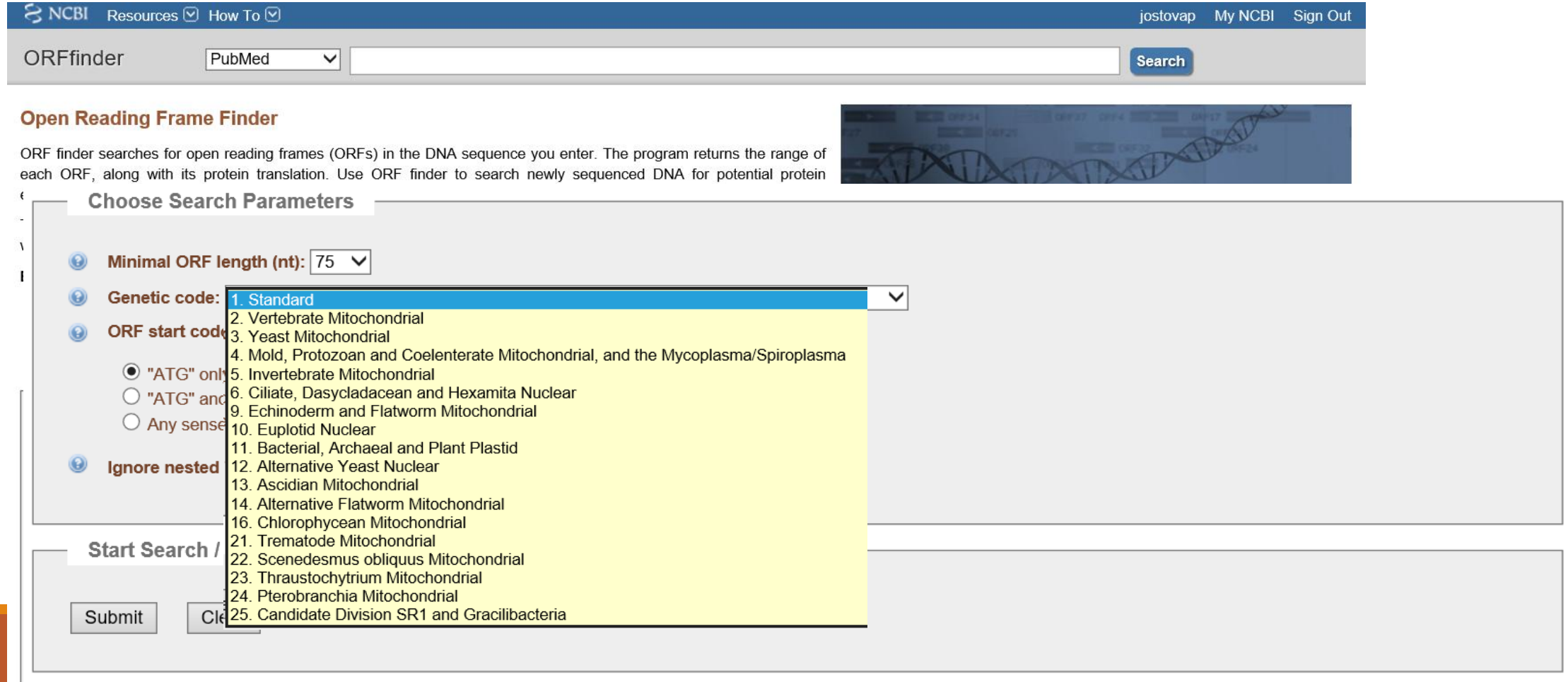
Try

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

-see the difference

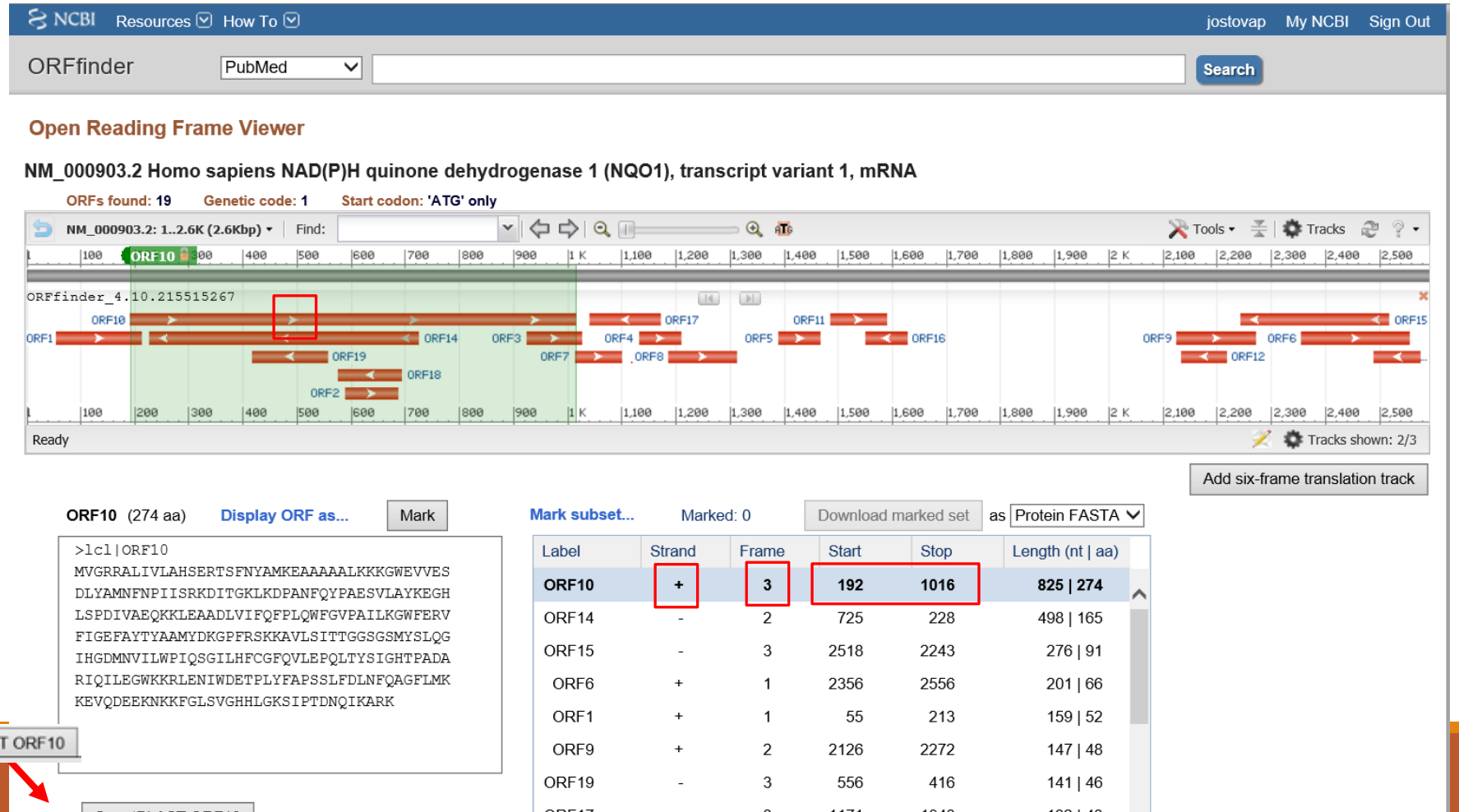
„reading“ DNA = translation

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

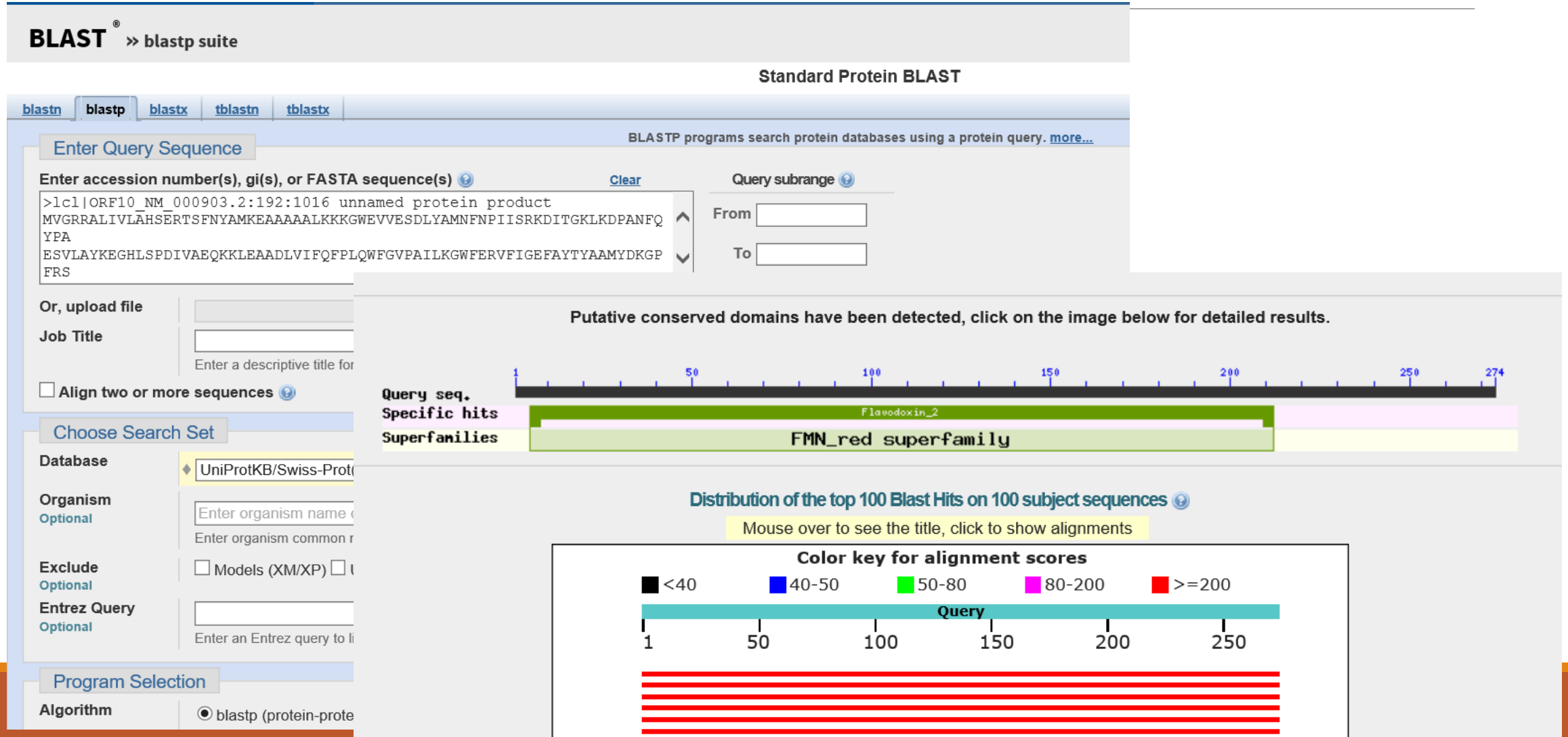


„reading“ DNA = translation

ORFfinder → looking for the longest ORF

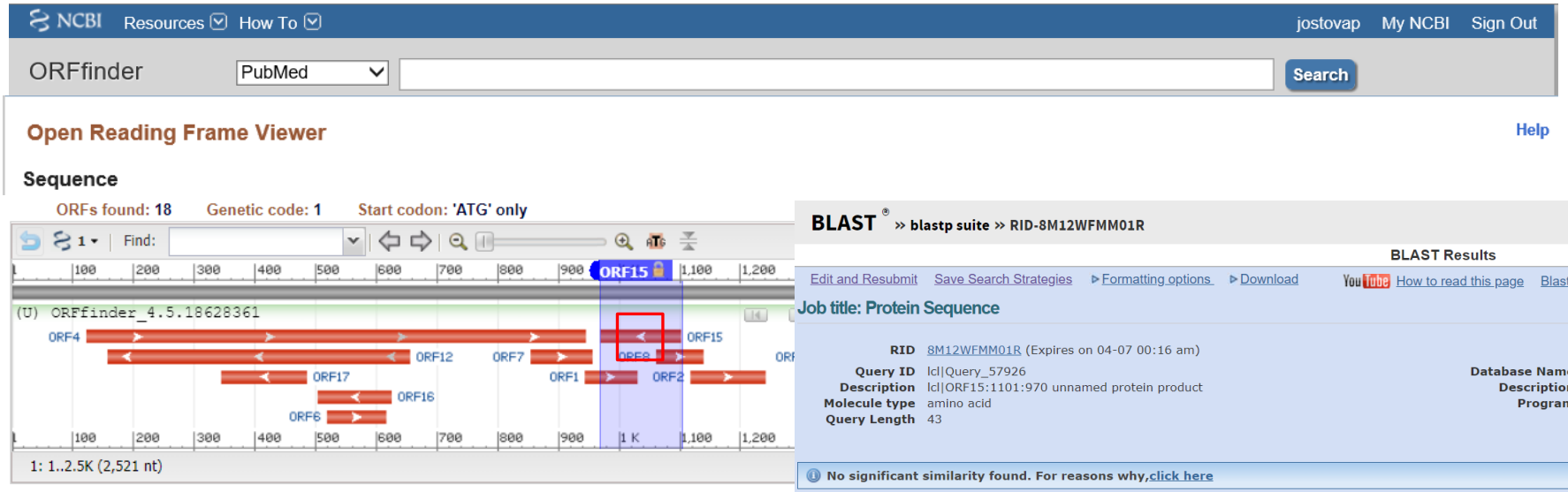


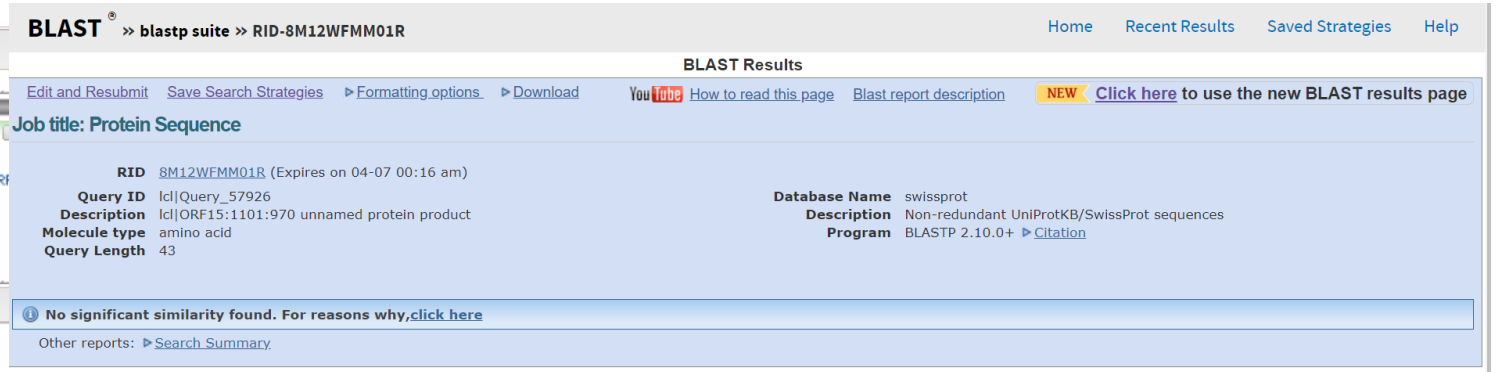
„reading“ DNA = translation

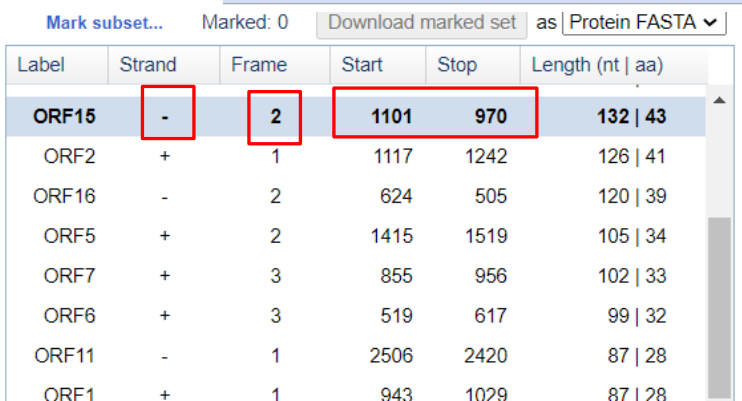


„reading“ DNA = translation

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**?**

„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

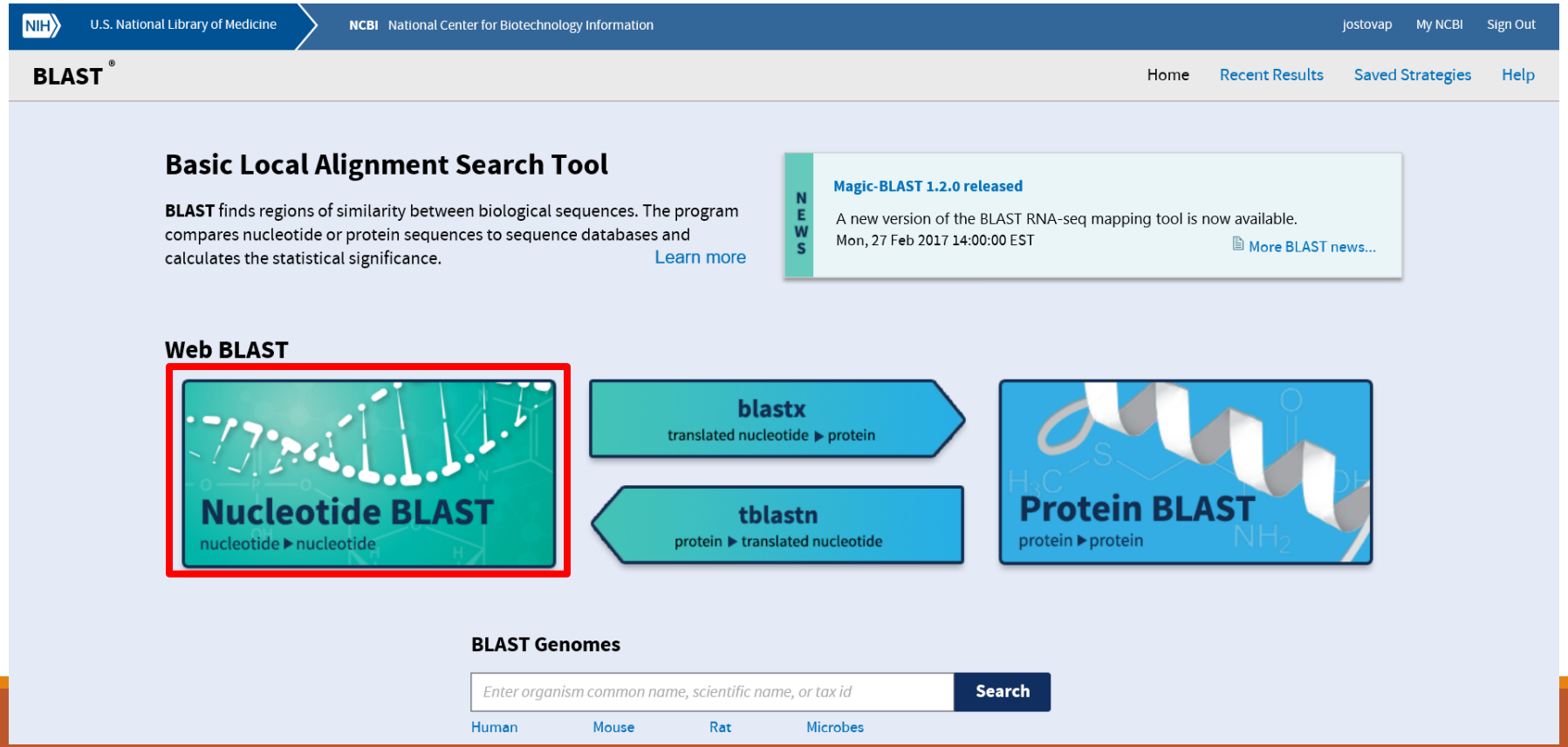
microRNA

Genomes….

….

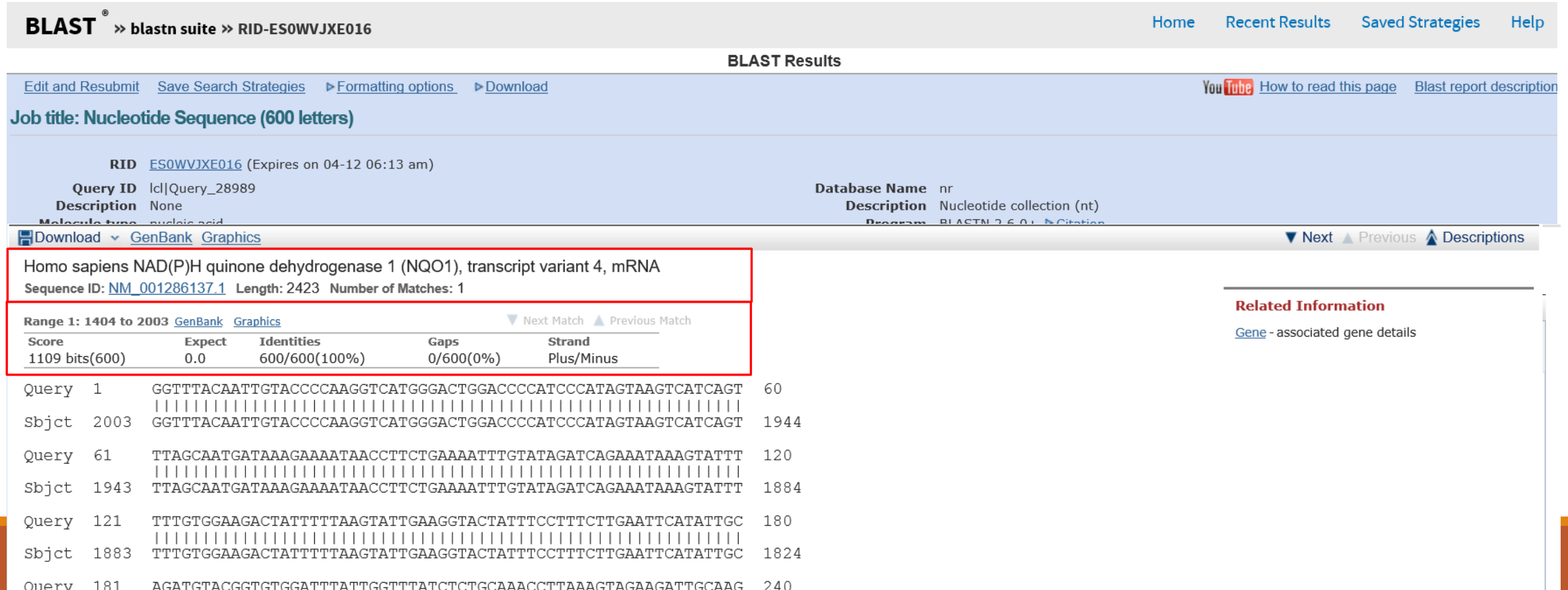
Unknown sequence identification – **BLASTn !**

→ looking for similarity: **BLASTn**



Unknown sequence identification

→ looking for similarity: BLASTn (or BLASTx)



BLAST® » blastn suite » RID-ES0WVJXE016

Home Recent Results Saved Strategies Help

BLAST Results

Edit and Resubmit | Save Search Strategies | ▸Formatting options | ▸Download | YouTube How to read this page | Blast report description

Job title: Nucleotide Sequence (600 letters)

RID ES0WVJXE016 (Expires on 04-12 06:13 am)

Query ID lcl|Query_28989
Description None

Database Name nr
Description Nucleotide collection (nt)

Download | GenBank | Graphics ▼ Next ▲ Previous ▲ Descriptions

Homo sapiens NAD(P)H quinone dehydrogenase 1 (NQO1), transcript variant 4, mRNA

Sequence ID: NM_001286137.1 Length: 2423 Number of Matches: 1

Range 1: 1404 to 2003 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
1109 bits(600)	0.0	600/600(100%)	0/600(0%)	Plus/Minus

```
Query  1     GGTTTACAATTGTACCCCAAGGTCATGGGACTGGACCCCATCCCATAGTAAGTCATCAGT  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2003  GGTTTACAATTGTACCCCAAGGTCATGGGACTGGACCCCATCCCATAGTAAGTCATCAGT  1944

Query  61    TTAGCAATGATAAAGAAAATAACCTTCTGAAAATTTGTATAGATCAGAAATAAAGTATTT  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1943  TTAGCAATGATAAAGAAAATAACCTTCTGAAAATTTGTATAGATCAGAAATAAAGTATTT  1884

Query  121   TTTGTGGAAGACTATTTTTAAGTATTGAAGGTACTATTTCCTTTCTTGAATTCATATTGC  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1883  TTTGTGGAAGACTATTTTTAAGTATTGAAGGTACTATTTCCTTTCTTGAATTCATATTGC  1824

Query  181   AGATGTACGGTGTGGATTTATTGGTTTATCTCTGCAAACCTTAAAGTAGAAGATTGCAAG  240
```

Related Information

Gene - associated gene details

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

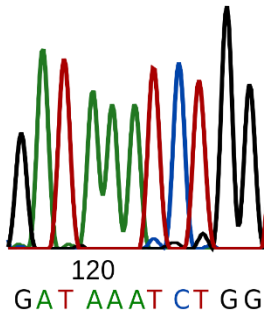
microRNA

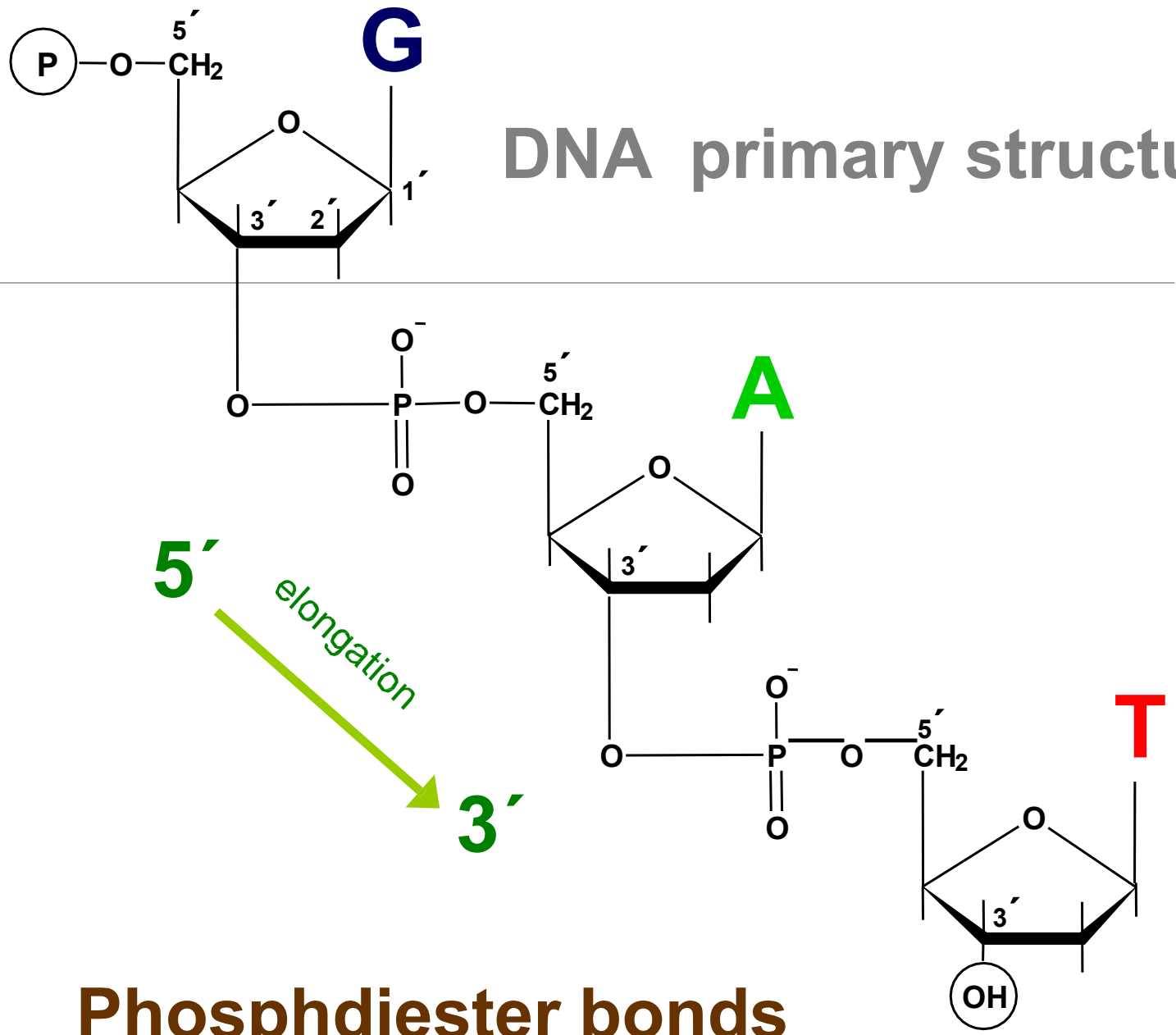
Genomes….

….

DNA primary structure

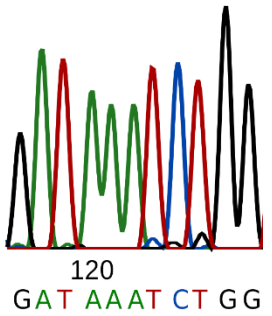
DNA sequencing





Phosphdiester bonds

DNA sequencing

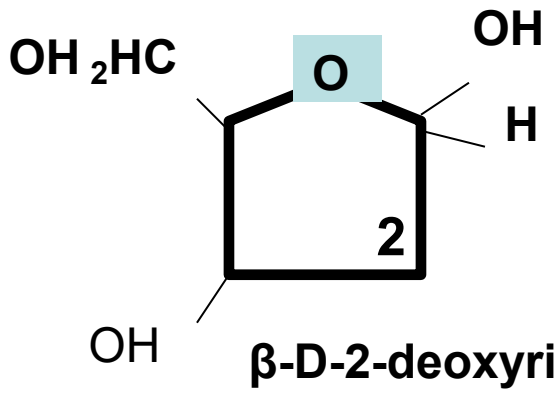


„Clasic“ Sanger sequencing (1977)

→ Lenght of sequence: ~1000nt

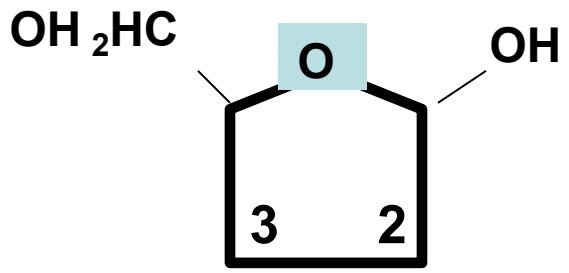
→ Output: „text“ 4 letters (ACTG)

Deoxynucleotides x **dideoxynucleotides**



β-D-2-deoxyribose

~1:100



β-D-2,3-dideoxyribose

DNA sequencing

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

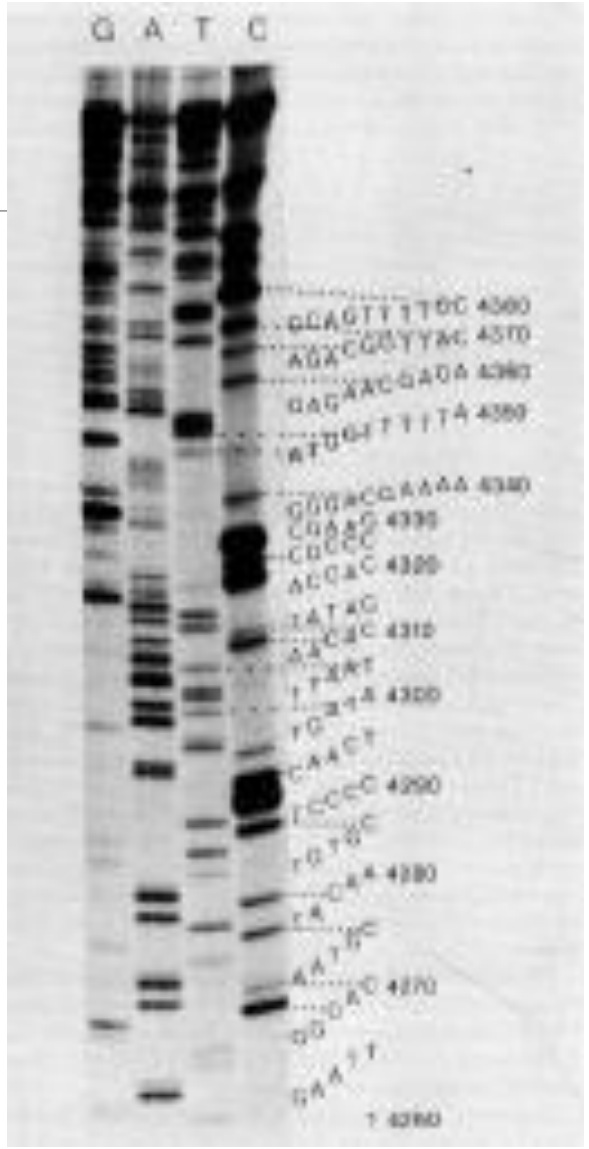
Asymetric DNA amplification, ^{32}P-dATP

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

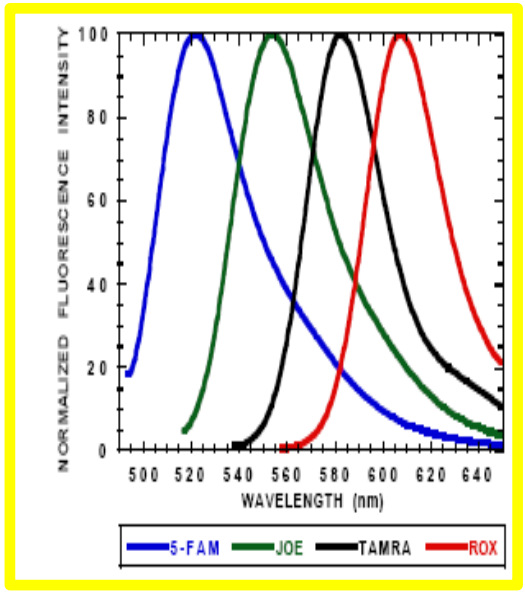
RA detection

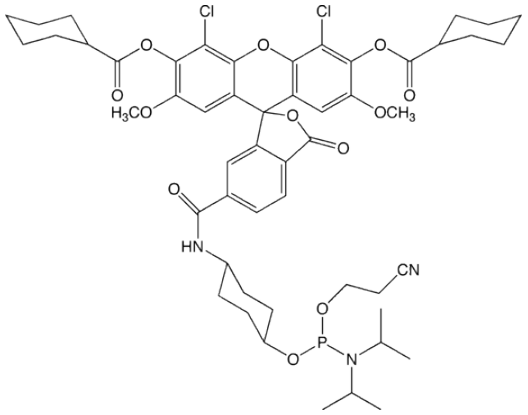


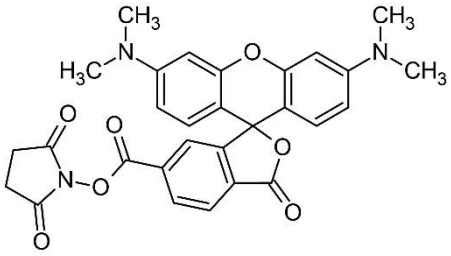
DNA sequencing

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**

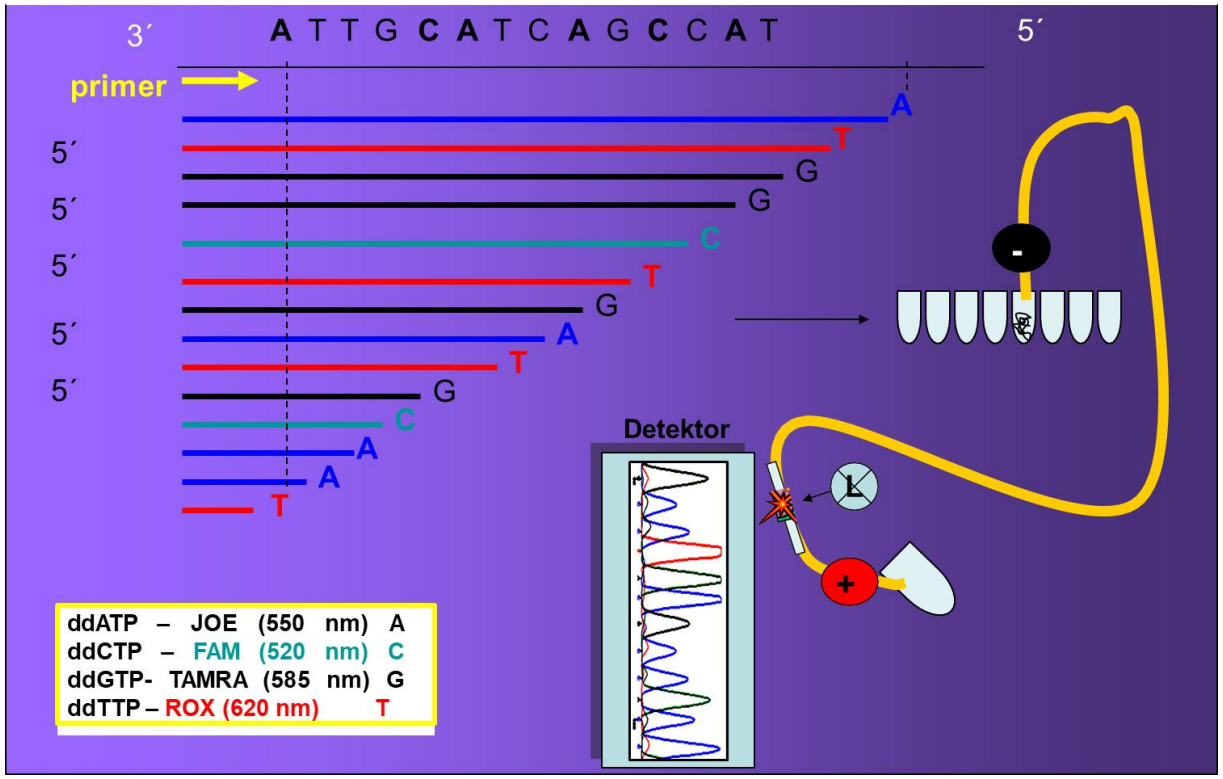






DNA sequencing

Princip: Sanger sequencing and capillary electrophoresis



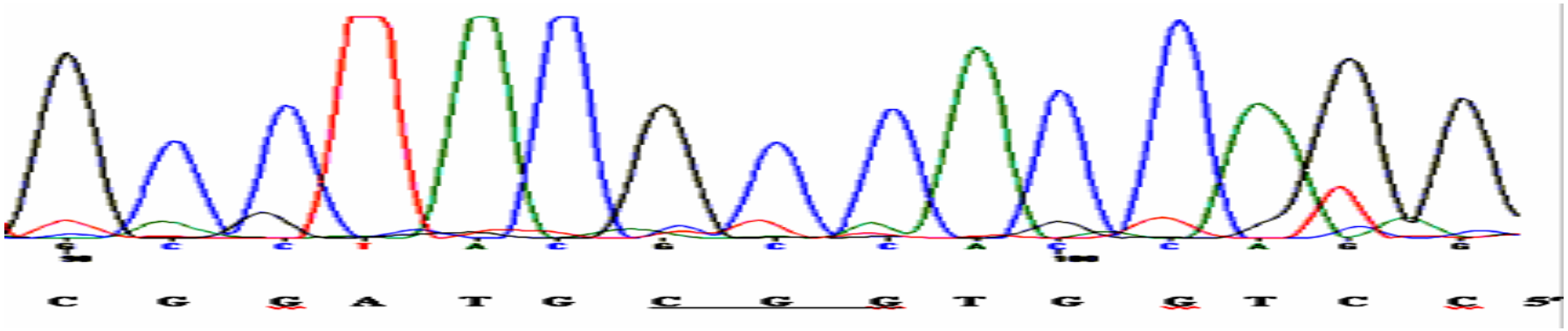
Sequence data analysis

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

„plain text“: CGGATGCGGTGGTCG

„fasta“: >identificator
CGGATGCGGTGGTCG

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



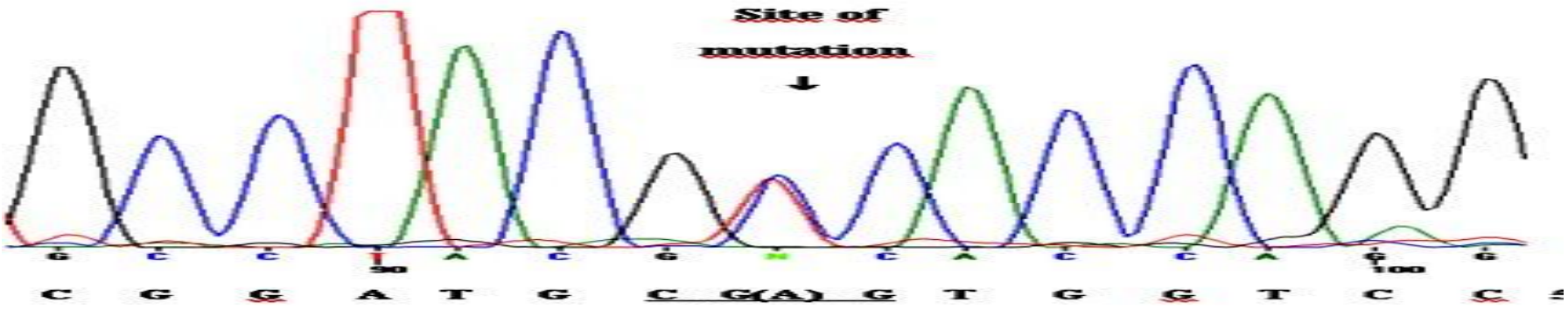
Sequence data analysis

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

„plain text“: CGGATGCNGTGGTCG

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

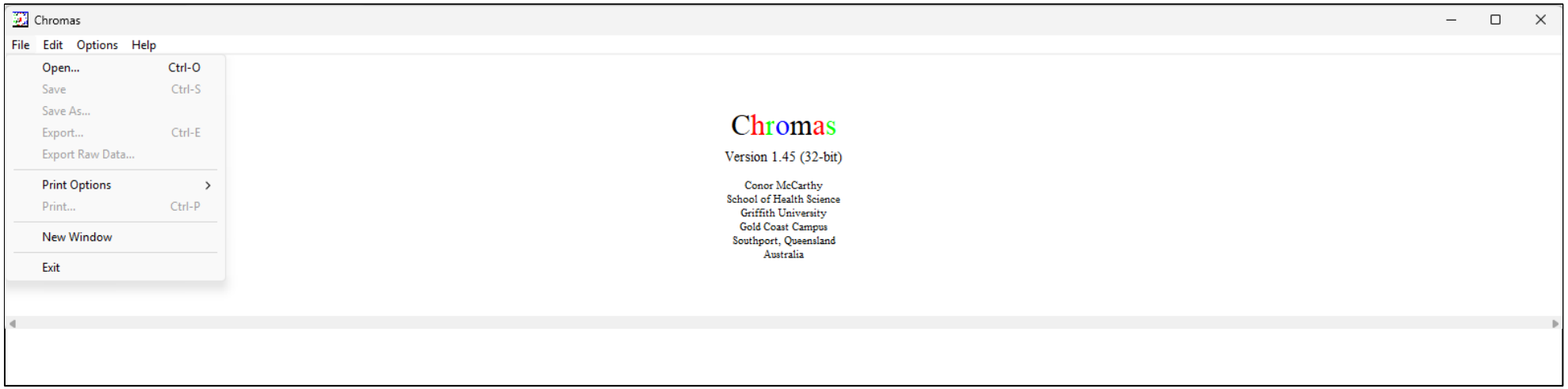
„sekvenační formát“(.scf, .abi, .ab1)



Sequence data analysis

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

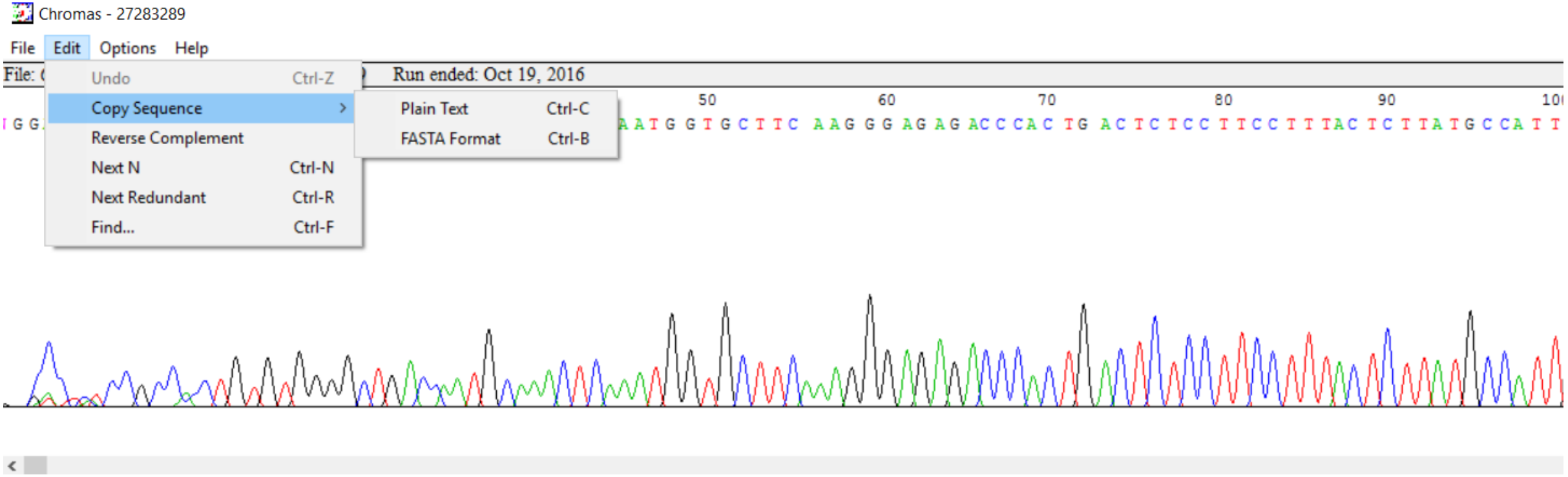
- the sequence has to be first saved and then open in chromas



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)



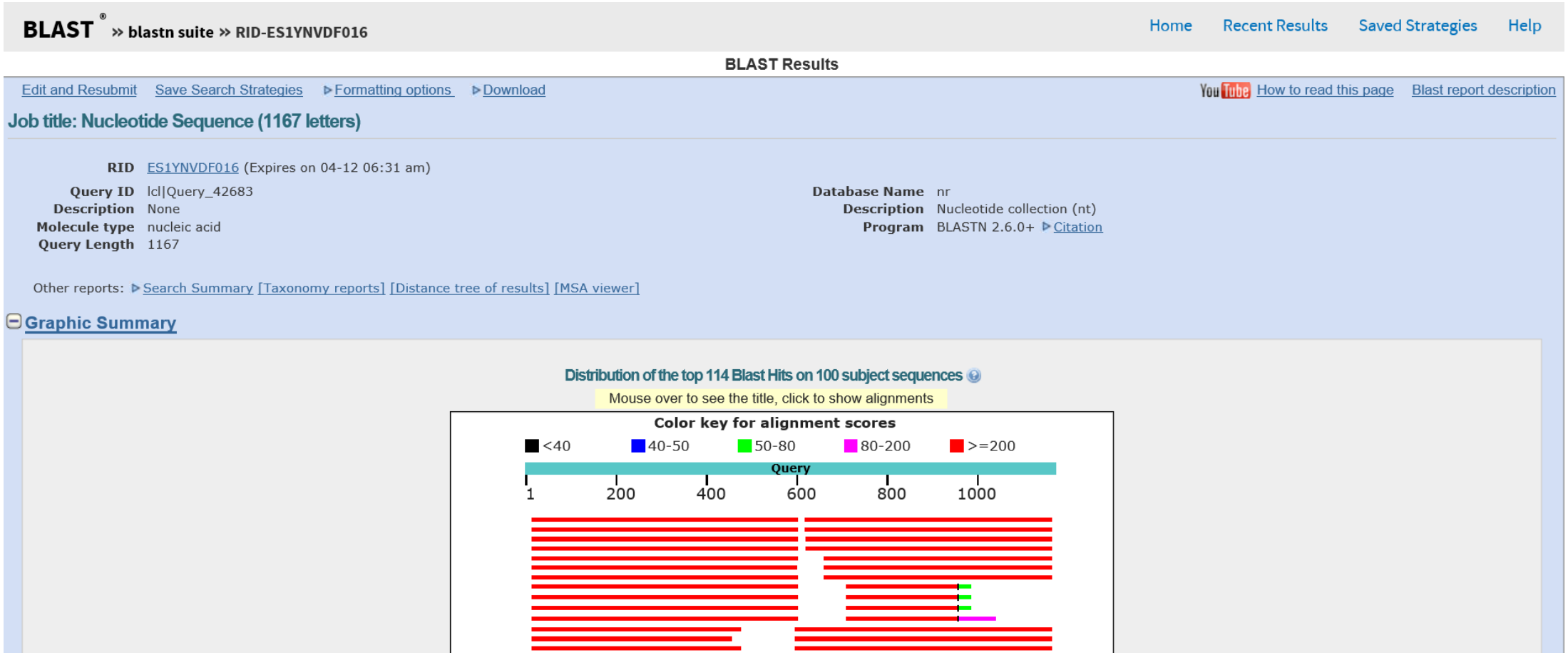
Try

Run „chromas“ from Moodle

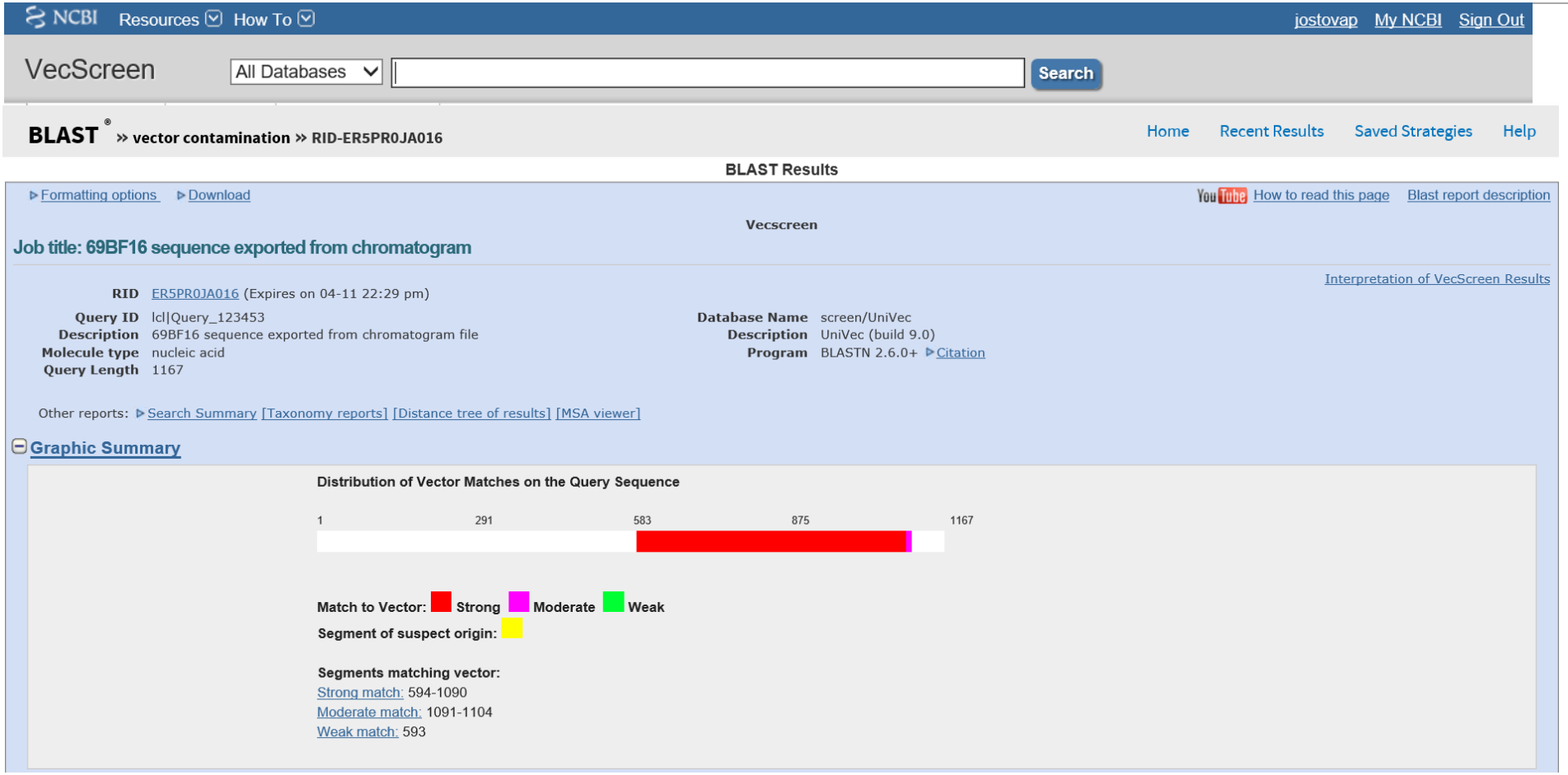
Download and store Ex3 sequence data from Moodle.
Open the sequence in chromas.
Export the sequence from chromas and identify it.

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

Sequence data analysis (Ex3) - Blastn

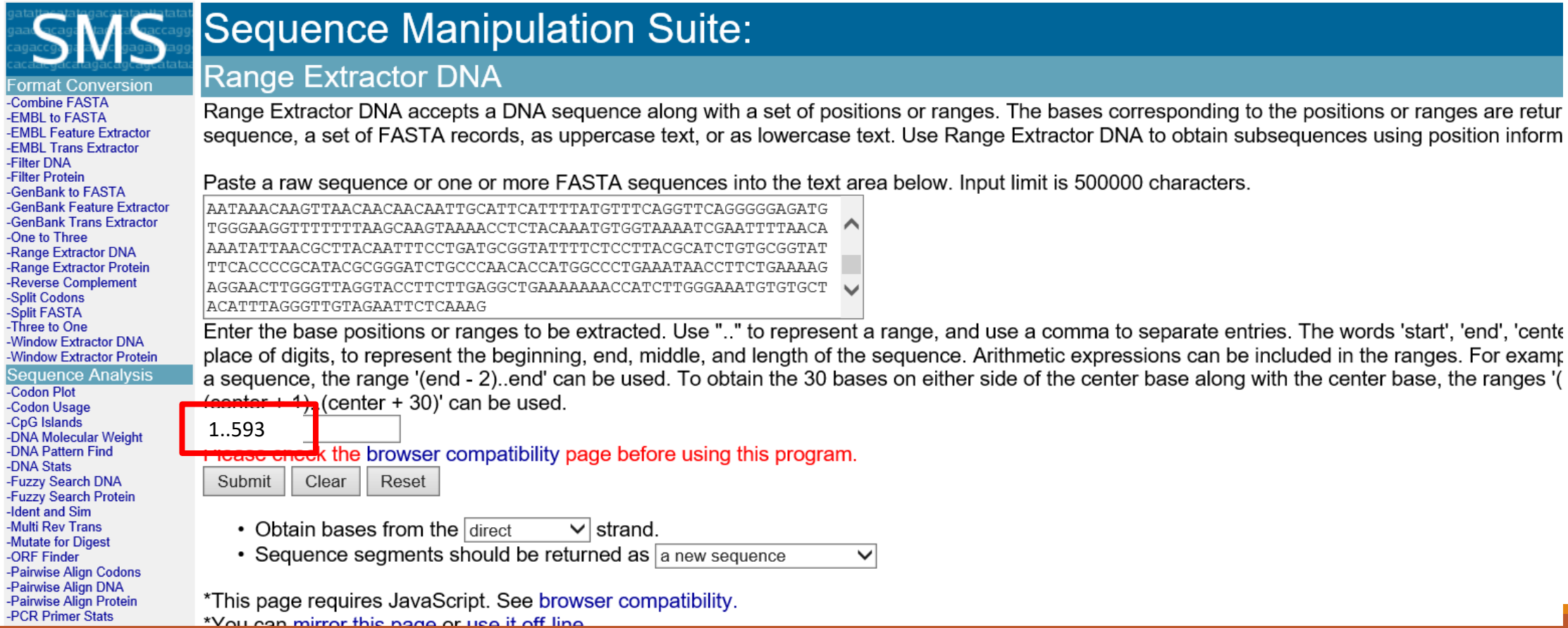


Sequence data analysis (Ex3) - Vecsreen



Sequence data analysis: purifing sequence of vector

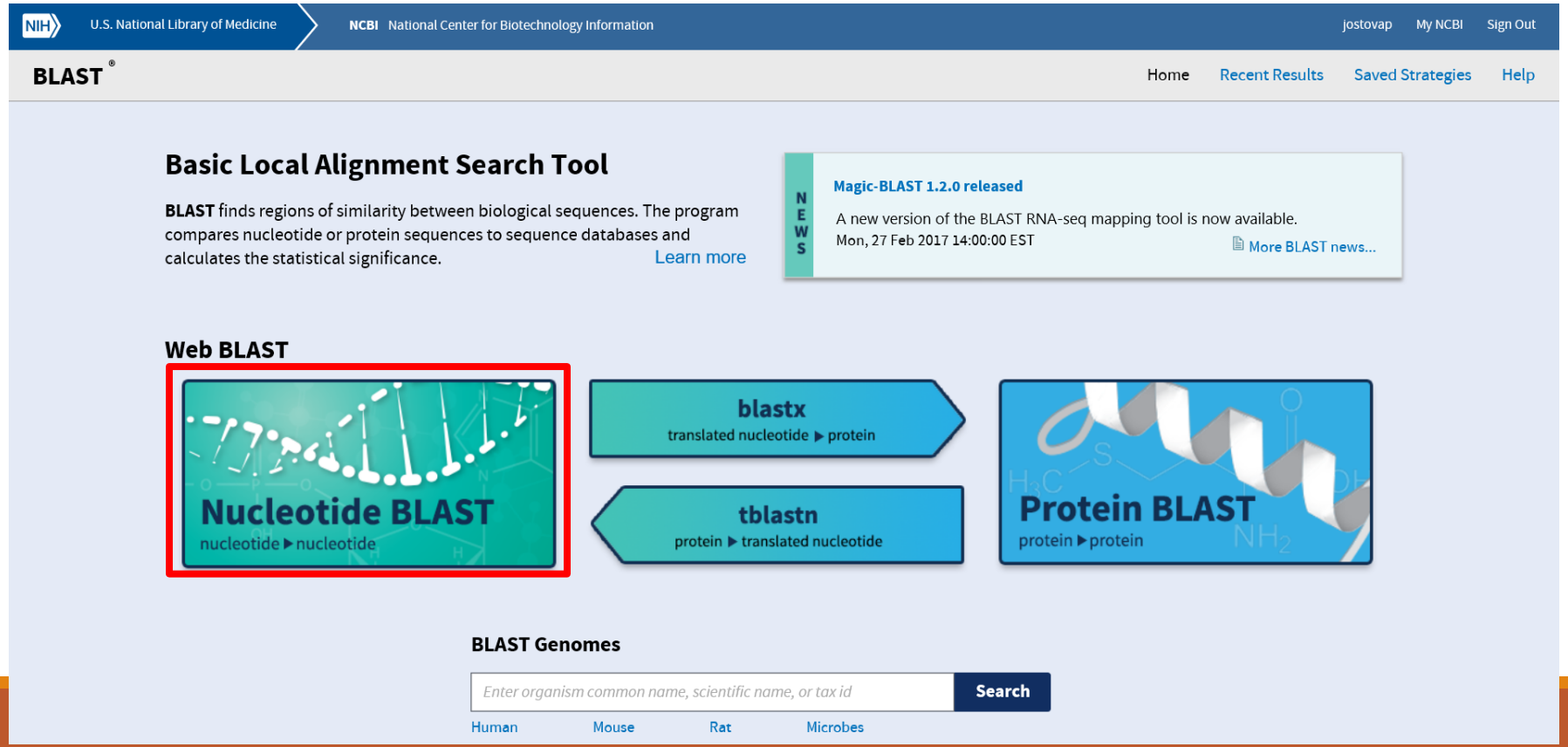
SMS „Range Extractor DNA“



Unknown sequence identification – BLASTn !

- if the sequence does not have easily recognizable ORF ?

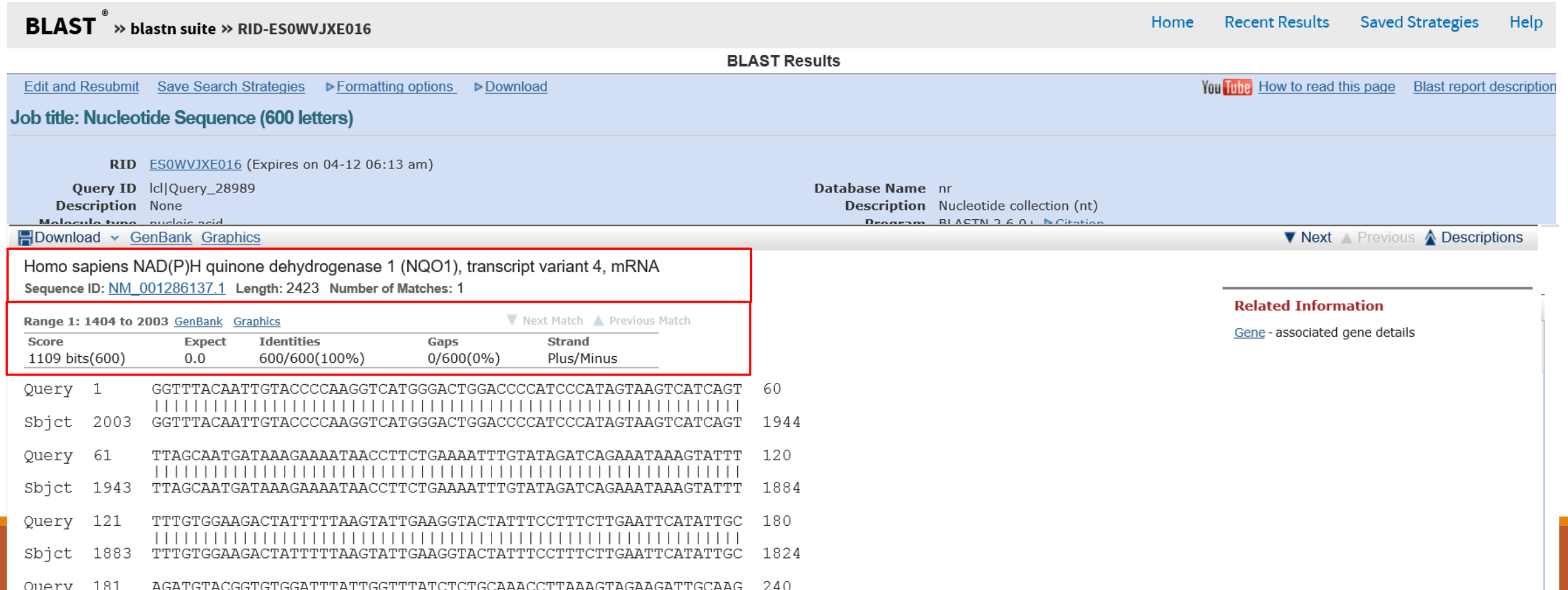
→ looking for similarity: BLASTn (or BLASTx)



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)