

# Introduction to applied bioinformatics

---

PETRA MATOUŠKOVÁ

2023/2024

6/10

# „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

## Local Alignment

Target Sequence

5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'

||||| ||||| ||||| ||||| |||||

Query Sequence

5' TACTCACGGATGAGGTACTTTAGAGGC 3'

## Global Alignment

Target Sequence

5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'

||||| ||||| ||||| |||||

5' ACTACTAGATT---ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'

Query Sequence

# Nucleotide sequence comparison

-Analogous to protein comparison

e.g. multalin

→ Comparing/looking for short sequences (primers)

Default matrix: BLOSUM62

input:

>sequence

```
GGGCACACTCCAGCAGACGCCCGAATTCAAATCCTGGAAGGATGGAAGAAACGCCTGGAGAATATTTGGG
ATGAGACACCACGTATTTTGGCTCCAAGCAGCCTCTTTGACCTAAACTTCCAGGCAGGATTCTTAATGAA
AAAAGAGGTACAGGATGAGGAGAAAAACAAGAAATTTGGCCTTTCTGTGGGCCATCACTTGGGCAAGTCC
ATCCCAACTGACAACCAGATCAAAGCTAGAAAATGAGATTTCCTTAGCCTGGATTTCTTCTAACATGTTA
TCAAATCTGGGTATCTTTCCAGGCTTCCCTGACTTGCTTTAGTTTTTAAGATTTGTGTTTTTCTTTTTCC
ACAAGGAATAAATGAGAGGGAATCGACTGTATTCGTGCATTTTTGGATCATTTTTAACTGATTCTTATGA
TTACTATCATGGCATATAACCAAATCCGACTGGGCTCAAGAGGCCACTTAGGGAAAGATGTAGAAAGAT
>2.exon
```

```
CAGGATGAGGAGAAAAACA
```



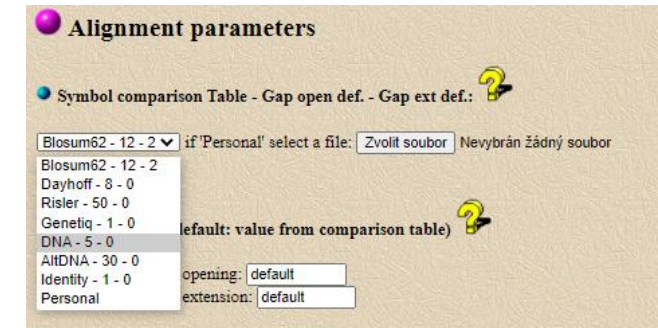
	1	10	20	30	40	50	60	70	80	90	100	110	120	130
sequence	GGGCACACTCCAGCAGACGCCCGAATTCAAATCCTGGAAGGATGGAAGAAACGCCTGGAGAATATTTGGG													
Primer														
Consensus	.....													
131	140	150	160	170	180	190	200	210	220	230	240	250	260	
sequence	TCTTAAATGAAAAAGAGGTCAGGATGAGGAGAAAAACAAGAAATTTGGCCTTTCTGTGGGCCATCACTTGGCAGTCCATCCCACTGACACAGATCAAGCTAGAAATGAGATTCTTAGCCTG													
Primer														
Consensus	.....													
261	270	280	290	300	310	320	330	340	350	360	370	380	390	
sequence	GATTCCTTCTAARATGTTATCAATCTGGGTATCTTCCAGGCTTCCCTGACTTGTGTTAGTTTTAGATTTGTGTTTTCTTTTCCACAGGATTAATGAGAGGATCGACTGATTCTGCTGAT													
Primer														
Consensus	.....													
391	400	410	420	430	440	450	460	470	480	490				
sequence	TTTGGATCATTTTTAARCTGATTCCTTATGATTACTATCATGGCATATAACCAAAATCCGACTGGGCTCAGAGGCCACTTAGGGAAAGATGTAGAAAGAT													
Primer														
Consensus	.....													

# 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 GTCTTTGCCCTCGCGACGCCGCCACCTCCGGF
69BF16      NGGCATTCTCCGCACTGTGTGGGGC
Consensus .....gCcCagcCgaCccCaCcaccTccGGz
    
```

```

131     140     150     160
|-----|-----|-----|
NM_015696.4 TCACATCCGGGGCAAACTGGTGTGCTGGAG
69BF16      AAAATTCTAGTATTTTGATTAATTTGARTCTT
Consensus aaAAcaccaGgagcaaaaTgaTgTcaaTcgaz
    
```

```

261     270     280     290
|-----|-----|-----|
NM_015696.4 GGGCCCCACCACCTTTAACGTGCTCGCCTTCC
69BF16      AGTATATCAAG-CAATAATCTCCCACCCAGG
Consensus aGgacacCAac.CaaTAAccTcCcacCCAacc
    
```

```

391     400     410     420
|-----|-----|-----|
NM_015696.4 AAGATTGCAGTCAACGGTACTGGTGCCATCC
69BF16      AAGCATTCAGTGAACATTTTTTG---CATATF
Consensus RaGaaTgCRAaTcRaCagTacTgG...Caaacc
    
```

```

521     530     540     550
|-----|-----|-----|
NM_015696.4 ACCCAACTGTGTCAGTGGAGGAGGTCAAGACC
69BF16      CTGTATTATTTCTTCATTACAAGAAGAAATG
Consensus .....TTCCTTCATTACAAGAAGAAATG
    
```

```

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

```

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

```

521     530     540     550     560     570     580     590     600     610     620     630     640
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
NM_015696.4 CAAGTGTGTCAGTGGAGGAGGTCAGACCCAGATCACAGCGCTCGTGAGGAGGTCATCTACTGAAAGCGAAGACTTATAACACCACCGCTCCTCCTCCACCACTCATCCCGCCACCTGTG
69BF16      NGGCATTCTCCCGCAC---TGTG
Consensus .....ncaCaTcaTCCCGCaC...TGTG
    
```

```

651     660     670     680     690     700     710     720     730     740     750     760     770
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
NM_015696.4 GCTG-ACCAA-TGCARACTCAATGGTGTTCARAGGGAGAGACCCACTGACTCTCCTTCTTACTCTTATGCCATTGGTCCCATCATTCTTGTGGGGGAAAAATCTAGTATTTTGATTATTG
69BF16      GCTGGACCAATGCAACTCAA-TGGTGTTCAR-GGGAGAGACCCACTGACTCTCCTTCTTACTCTTATGCCATTGGTCCCATCATTCTTGTGGGGGAAAAATCTAGTATTTTGATTATTG
Consensus GCTG.ACCAA.TGCARACTCAA.TGGTGTTCAR.GGGAGAGACCCACTGACTCTCCTTCTTACTCTTATGCCATTGGTCCCATCATTCTTGTGGGGGAAAAATCTAGTATTTTGATTATTG
    
```

```

781     790     800     810     820     830     840     850     860     870     880     890     900
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
NM_015696.4 TTACAGCAACAATAGGAATCCTGGCCAAATGAGAGCTTTGACCAGTGAATCACCAGCCGATACGAGCGCTTGCACACAAAATGTGTGGCAATAGAGTATATCAGCAATATCTCCACC
69BF16      TTACAGCAACAATAGGAATCCTGGCCAAATGAGAGCTTTGACCAGTGAATCACCAGCCGATACGAGCGCTTGCACACAAAATGTGTGGCAATAGAGTATATCAGCAATATCTCCACC
Consensus .....
    
```

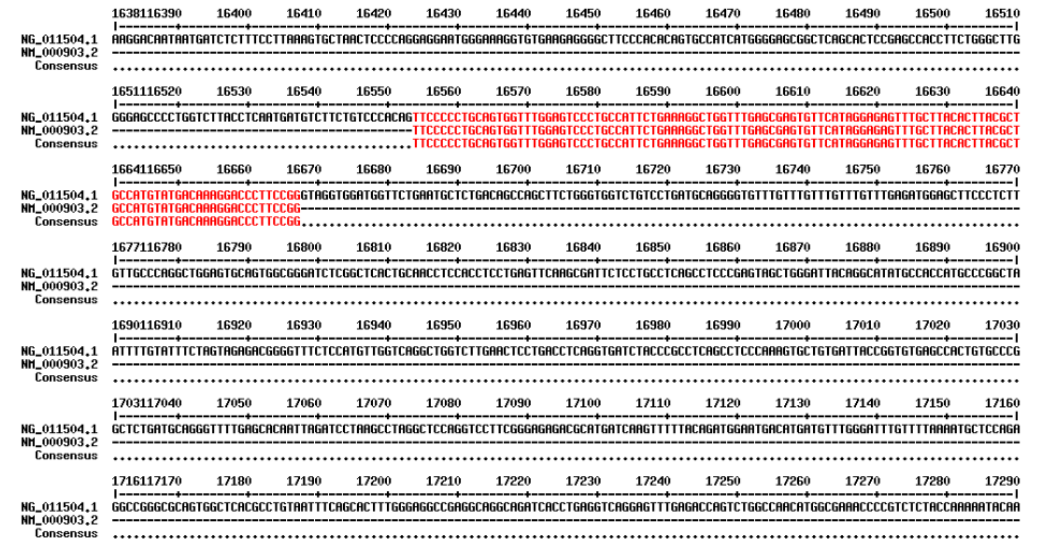
# Nucleotide sequence comparison

-Analogous to protein comparison

e.g. multalin

→ Comparing genomic DNA and cDNA (mRNA)

Default: BLOSUM62 → DNA 5-0



# Practical part

---

Try to compare sequences EX\_0 in Multalin, using both BLOSUM62(default) and DNA-5-0

- - see the difference, what do you see? which one is correct?

HW: Compare your sequence CDS and mRNA



# „reading“ DNA = translation

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

## 1. DNA sequence:

5' - **ATCGA** AGTATT TAAAGCGCCACCTATT TAA - 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

		Second nucleotide					
		U	C	A	G		
U	UUU	Phe	UCU	Tyr	UGU	Cys	U
	UUC		UCC	Ser	UGC		C
	UUA	Leu	UCA	STOP	UGA	STOP	A
	UUG		UCG	STOP	UGG	Trp	G
C	CUU		CCU	His	CGU		U
	CUC	Leu	CCC	Pro	CGC	Arg	C
	CUA		CCA	Gln	CGA		A
	CUG		CCG		CGG		G
A	AUU	Ile	ACU	Asn	AGU	Ser	U
	AUC		ACC	Thr	AGC		C
	AUA		ACA	Lys	AGA	Arg	A
	AUG	Met	ACG		AGG		G
G	GUU		GCU	Asp	GGU		U
	GUC	Val	GCC	Ala	GGC	Gly	C
	GUA		GCA		GGA		A
	GUG		GCG	Glu	GGG		G



# „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' -TACCTTCATAAATTTTCGCGGTGGATAAATT-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

		Second nucleotide					
		U	C	A	G		
U	UUU	Phe	UCU	UAU	Tyr	UGU	Cys
	UUC		UCC	UAC		UGC	
	UUA	Leu	UCA	UAA	STOP	UGA	STOP
	UUG		UCG	UAG	STOP	UGG	Trp
C	CUU		CCU	CAU	His	CGU	
	CUC	Leu	CCC	CAC		CGC	Arg
	CUA		CCA	CAA	Gln	CGA	
	CUG		CCG	CAG		CGG	
A	AUU	Ile	ACU	AAU	Asn	AGU	Ser
	AUC		ACC	AAC		AGC	
	AUA		ACA	AAA	Lys	AGA	Arg
	AUG	Met	ACG	AAG		AGG	
G	GUU		GCU	GAU	Asp	GGU	
	GUC <td>Val</td> <td>GCC</td> <td>GAC</td> <td></td> <td>GGC</td> <td>Gly</td>	Val	GCC	GAC		GGC	Gly
	GUA		GCA	GAA	Glu	GGA	
	GUG		GCG	GAG		GGG	

# „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 **TTT** 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

		Second nucleotide					
		U	C	A	G		
U	UUU	Phe	UCU	Tyr	UGU	Cys	U
	UUC		Ser	UAC	UGC		C
	UUA	Leu	UCA	<b>STOP</b>	<b>STOP</b>	<b>STOP</b>	A
	UUG		UCG	<b>STOP</b>	UGG	Trp	G
C	CUU		CCU	His	CGU	Arg	U
	CUC	Leu	CCC	Pro	CGC		C
	CUA		CCA	Gln	CGA		A
	CUG		CCG		CGG		G
A	AUU	Ile	ACU	Asn	AGU	Ser	U
	AUC		ACC	Thr	AGC		C
	AUA		ACA	Lys	AGA	Arg	A
	AUG	Met	ACG		AGG		G
G	GUU		GCU	Asp	GGU		U
	GUC	Val	GCC	Ala	GGC	Gly	C
	GUA		GCA		GGA		A
	GUG		GCG	Glu	GGG		G

# „reading“ DNA = translation

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

**SMS** 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 reading frame 1 reading frame 2 reading frame 3

Submit Clear direct reverse

• Translate  
• Use the

\*This page rec  
\*You can mirr

Fri Jun 17 16:17:06 20  
Valid XHTML 1.0; Valid CSS

**Translate results**

```
>rf 1 Untitled
LSQQIQLASVIRPCSPFFL*PLLCCYHGSDTQFH HQ*SDRFHQDVSSLVQRNCLLHQQL
SSRFGQ*SCLFRYRLYDMVFR*SQIRSHVHNLIMS*LNQCLDPQVFRLDTYDHSSESSGRH
VLPFVHVHASNHQVLMLIDDVCVSTQQVGS L
```

**Translate results**

```
>rf 2 Untitled
CRSRYNLLR*YVHVHFFFHSHCCVVTTEATHNFIINKATVFIKT*VRLCDNVTVFFISS*
VVDLVSDRVCFDIDETIWCFDKAKFVHTCIT***VD*TNVWTLRCFDWTHTTIVRVVDVT
YFHLCTFTRQTTRS*C**TTFVCQLSKWVRL
```

# „reading“ DNA = translation

Insert both CDS and mRNA in fasta format

- 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
  - Restriction Digest
  - Restriction Summary
  - Reverse Translate
  - Translate
- Sequence Figures
  - Color Align Conservation
  - Color Align Properties
  - Group DNA
  - Group Protein
  - Primer Map
  - Restriction Map
  - Translation Map
- Random Sequences
  - Mutate DNA
  - Mutate Protein
  - Random Coding DNA
  - Random DNA Sequence
  - Random DNA Regions
  - Random Protein Sequence
  - Random Protein Regions
  - Sample DNA
  - Sample Protein
  - Shuffle DNA
  - Shuffle Protein
- Miscellaneous
  - Home
  - IUPAC codes

## Sequence Manipulation Suite:

### Translate

Translate accepts a DNA sequence and converts it into a protein in the reading frame y

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

```
>NM_000903.3:122-946 Homo sapiens NAD(P)H quinone
dehydrogenase 1 (NQO1), transcript variant 1, mRNA
ATGGTCGGCAGAAAGAGCACTGATCGTACTGGCTCACTCAGAGAGGACGTCCTTCAACTATG
CCATGAAGG
AGGCTGCTGCAGCGGCTTTGAAGAAGAAAGGATGGGAGGTGGTGAAGTCGGACCTCTATGC
CATGAACCT
```

- Translate in  on the  strand.
- Use the  genetic code.

\*This page requires JavaScript. See [browser compatibility](#).  
\*You can [mirror this page](#) or use it off-line.

Sun 1 Oct 12:00:08 2023  
Valid XHTML 1.0; Valid CSS

[new window](#) | [h](#)

Sequence Manipulation Suite – Pracovní – Microsoft Edge

about:blank

### Translate results

```
>rf 1 NM_000903.3:122-946 Homo sapiens NAD(P)H quinone dehydrogenase 1 (NQO1), transcript variant 1, mRNA
MVGRRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVEVDLYAMNPNPIISRKDIITGKL
KDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVIFQFPPLQWFGVPAILKGFPERV
FIGEFAYTYAAMYDKGPFPRSKAVLSITGGSGSMYSLOGIHGDMNVILWPIQSGILHFC
GFQVLEPQLTYSIGHTPADARIQILEGWKKRLENIWDETPLYFAPSSLFDLNLFQAGFLMK
KEVQDEEKNNKFFGLSVGHLLGKSIPTDNQIKARK*
```

```
>rf 1 NM_000903.3 Homo sapiens NAD(P)H quinone dehydrogenase 1 (NQO1), transcript variant 1, mRNA
TRDSHKVAAGAAQLTESLVPARVAPATTSPANQRPLHQSHGRQKSTDRGTGSLREDVLQL
CHEGGCCSGFEEERMGGGGVGPLCHELQSHHFQKGHHR*TEGPCELSVSCRVCSSL*RRP
SEPRYCG*TKEAGSRPCDIPVPPAVVWSPCHSERLV*ASVHRRVCLHLRCHV*QRTLPE
*EGSAFHWHHWQWLHVLRSARDPRGHECHSLANSEWHS AFLWLPLSLRTSTDI*HWAHSSRR
PNSNPGRMEETPGEYLG*DTTVFCCKQPL*PKLPGRIILNEKRG TG*GEKQEIWPPFCGSSL
GQVHPN*QPDQS*KMRFSLDFLLTCYQIWSVFAASLTCFSF*DLCFSFSTRNK*EGIDC
IRAFLDHF*LILMITIMAYNQNP TGLKRLRERCRKMLEKCSL KASTQFNSSF*G*SRV
QFG*VSFNPMFY*SPLCSLWQKGI AQRKRRLNLPALRDLTCLVVS HMLVYDISCFQLQ
SSY*YA*HKYHSWAFVVIQYTDTLKGRANKSLLCCSHLLFF*LKKIFF*SSLALLPRL
ECSGVI SAHCNLC L PGSSNS PASASLVAGMTGACHHA*LI FVFLVETAFHHV GQAGL KLL
TSGDPPTSAQSAGITGVIIHHTWPLQSSTLRFAEINQ*IHTVHLQYEFKKNSTFNT*K*
SSTKNTLFLIYTNFQKVIFFIIAKLMTYYGMGSSPMTLGYNCKPRVLS TLVNSFGI IVNF
YFWKSSHSTVGII*FKENMIKHCPLVVH*KKR*EMKRLPEKWTETASYLPRK*RDWTELE
NLLYQMLTGTGGFCRSRYPQ*LTAGCFSLKIFCCLHLHLCNFV*ISQRSELNK*NSFLQTH
```

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

NCBI Resources How To jostovap My NCBI Sign Out

ORFfinder PubMed  Search

### Open Reading Frame Finder

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein



#### Choose Search Parameters

Minimal ORF length (nt): 75

Genetic code: 1. Standard

ORF start code: 2. Vertebrate Mitochondrial

- "ATG" only
- "ATG" and
- Any sense

- Ignore nested
- 3. Yeast Mitochondrial
  - 4. Mold, Protozoan and Coelenterate Mitochondrial, and the Mycoplasma/Spiroplasma
  - 5. Invertebrate Mitochondrial
  - 6. Ciliate, Dasycladacean and Hexamita Nuclear
  - 9. Echinoderm and Flatworm Mitochondrial
  - 10. Euplotid Nuclear
  - 11. Bacterial, Archaeal and Plant Plastid
  - 12. Alternative Yeast Nuclear
  - 13. Ascidian Mitochondrial
  - 14. Alternative Flatworm Mitochondrial
  - 16. Chlorophycean Mitochondrial
  - 21. Trematode Mitochondrial
  - 22. Scenedesmus obliquus Mitochondrial
  - 23. Thraustochytrium Mitochondrial
  - 24. Pterobranchia Mitochondrial
  - 25. Candidate Division SR1 and Gracilibacteria

Start Search /

Submit

Clear



# „reading“ DNA = translation

ORFfinder → looking for the longest ORF

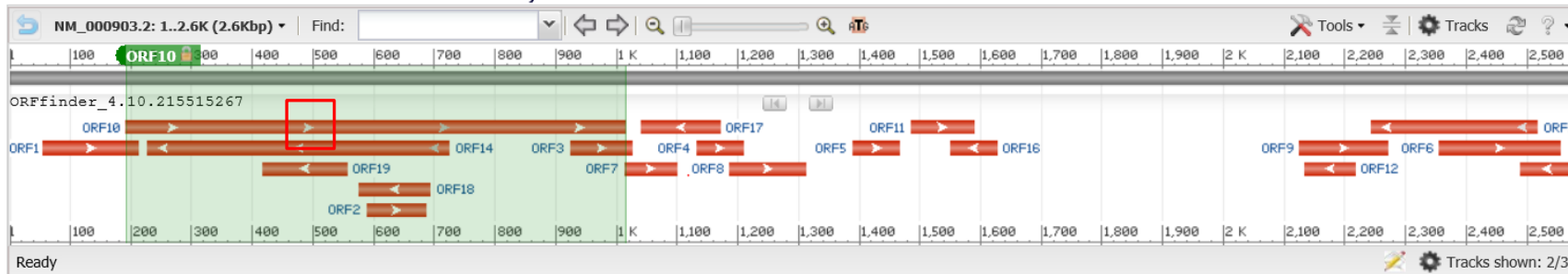
NCBI Resources How To jostovap My NCBI Sign Out

ORFfinder PubMed Search

## Open Reading Frame Viewer

NM\_000903.2 Homo sapiens NAD(P)H quinone dehydrogenase 1 (NQO1), transcript variant 1, mRNA

ORFs found: 19 Genetic code: 1 Start codon: 'ATG' only



Add six-frame translation track

ORF10 (274 aa) Display ORF as... Mark

```
>lcl|ORF10
MVGRRALIVLAHSERTSFNYAMKEAAAAALKKGWVVES
DLYAMNFPNII SRKIDITGKLDKDPANFYPAESVLAYKEGH
LSPDIVAEQKKLEAADLIVFQFPILQWFGVPAILKGWFERV
FIGEFAYTYAAMYDKGPPFRSKAVLSITGGSGSMYSLQG
IHGDMNVILWPIQSGILHFCGFQVLEPQLTYSIGHTPAD
RIQILEGWKKRLENIWDETPPLYFAPSSLFDLNFQAGFLMK
KEVQDEEKNNKFGLSVGHHLGKSIPTDNQIKARK
```

Mark subset... Marked: 0 Download marked set as Protein FASTA

Label	Strand	Frame	Start	Stop	Length (nt   aa)
ORF10	+	3	192	1016	825   274
ORF14	-	2	725	228	498   165
ORF15	-	3	2518	2243	276   91
ORF6	+	1	2356	2556	201   66
ORF1	+	1	55	213	159   52
ORF9	+	2	2126	2272	147   48
ORF19	-	3	556	416	141   46
ORF17	-	3	1174	1016	100   40

BLAST ORF10

# „reading“ DNA = translation

**BLAST**® » blastp suite

## Standard Protein BLAST

blastn | **blastp** | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. [more...](#)

### Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) Query subrange [Query subrange](#)

```
>lcl|ORF10_NM_000903.2:192:1016 unnamed protein product
MVGRRALIVLÄHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFPNPIISRKIDITGKLKDPANFQ
YPA
ESVLAYKEGHLSPDIVAEQKLEADLVIFQFPLQWFGVPAILKGWFERVFIGEFAYTYAAMYDKGP
FRS
```

From  To

Or, upload file

Job Title   
Enter a descriptive title for

Align two or more sequences [Align two or more sequences](#)

### Choose Search Set

Database

Organism   
Optional Enter organism common r

Exclude  Models (XM/XP)  t  
Optional

Entrez Query   
Optional Enter an Entrez query to li

### Program Selection

Algorithm  blastp (protein-prote

Putative conserved domains have been detected, click on the image below for detailed results.

Query seq.

Specific hits

Superfamilies

### Distribution of the top 100 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments

**Color key for alignment scores**

Score Range	Color
<40	Black
40-50	Blue
50-80	Green
80-200	Magenta
>=200	Red

# „reading“ DNA = translation

ORFfinder → looking for the longest ORF

NCBI Resources How To jostovap My NCBI Sign Out

ORFfinder PubMed Search

### Open Reading Frame Viewer

Help

Sequence

ORFs found: 18 Genetic code: 1 Start codon: 'ATG' only

ORFfinder\_4.5.18628361

ORF15 (43 aa) Display ORF as... Mark

```
>1c1|ORF15
MIQKCTNTVDLSFIPCGKRKTKQLKTKASQGSLELYPDLITC
```

ORF4

Marked set (0)

SmartBLAST

BLAST

SmartBLAST best hit titles...

BLAST

### BLAST

blastp suite » RID-8M12WFMM01R

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 NEW Click here to use the new BLAST results page

Job title: Protein Sequence

RID 8M12WFMM01R (Expires on 04-07 00:16 am)

Query ID Icl|Query\_57926

Description Icl|ORF15:1101:970 unnamed protein product

Molecule type amino acid

Query Length 43

Database Name swissprot

Description Non-redundant UniProtKB/SwissProt sequences

Program BLASTP 2.10.0+ Citation

No significant similarity found. For reasons why, click here

Other reports: Search Summary

Mark subset... Marked: 0 Download marked set as Protein FASTA

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?**
- in sequence (**ex2**) from Moodle and identify the correct ORF and the **source organism ? - impossible?**



# Unknown sequence identification – BLASTn !

- if the sequence does not have easily recognizable ORF ?

→ looking for similarity: BLASTn (or BLASTx)

The screenshot shows the NCBI BLAST website. At the top, there is a navigation bar with the NIH logo, "U.S. National Library of Medicine", "NCBI National Center for Biotechnology Information", and user links for "jostovap", "My NCBI", and "Sign Out". Below this is a secondary navigation bar with "BLAST" and links for "Home", "Recent Results", "Saved Strategies", and "Help".

The main content area features a "Basic Local Alignment Search Tool" section with a description: "BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance." A "Learn more" link is provided. To the right is a "NEWS" box titled "Magic-BLAST 1.2.0 released" with the text: "A new version of the BLAST RNA-seq mapping tool is now available. Mon, 27 Feb 2017 14:00:00 EST" and a link to "More BLAST news...".

Below this is the "Web BLAST" section, which is highlighted with a red box. It contains three main options:

- Nucleotide BLAST**: nucleotide ▶ nucleotide (highlighted with a red box)
- blastx**: translated nucleotide ▶ protein
- tblastn**: protein ▶ translated nucleotide

To the right of these options is a "Protein BLAST" section with the text: "protein ▶ protein".

At the bottom, there is a "BLAST Genomes" section with a search input field containing the placeholder text "Enter organism common name, scientific name, or tax id" and a "Search" button. Below the input field are links for "Human", "Mouse", "Rat", and "Microbes".



# Unknown sequence identification

- if the sequence does not have easily recognizable ORF ?
- 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 [Icl|Query\\_28989](#) Database Name [nr](#)  
Description [None](#) Description [Nucleotide collection \(nt\)](#)  
Molecule type [nucleic acid](#) Program [BLASTN 2.6.0+](#) [Citation](#)

[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     TTAGCAATGATAAAAGAAAATAACCTTCTGAAAATTTGTATAGATCAGAAATAAAGTATTT 120
           |||
Sbjct 1943   TTAGCAATGATAAAAGAAAATAACCTTCTGAAAATTTGTATAGATCAGAAATAAAGTATTT 1884

Query 121    TTTGTGGAAGACTATTTTTAAGTATTGAAGGTAATTTCTTTCTTGAATTCATATTGC 180
           |||
Sbjct 1883   TTTGTGGAAGACTATTTTTAAGTATTGAAGGTAATTTCTTTCTTGAATTCATATTGC 1824

Query 181    AGATGTACGGTGTGGATTTATTGGTTTATCTCTGCAAACCTTAAAGTAGAAGATTGCAAG 240
```

**Related Information**  
[Gene](#) - associated gene details

# Try

---

## Identify:

- sequence (**ex2**) from Moodle

# „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**

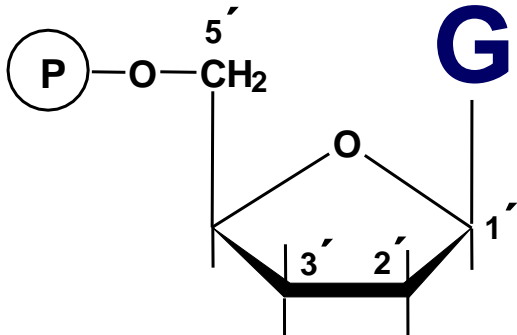
Gene expression

microRNA

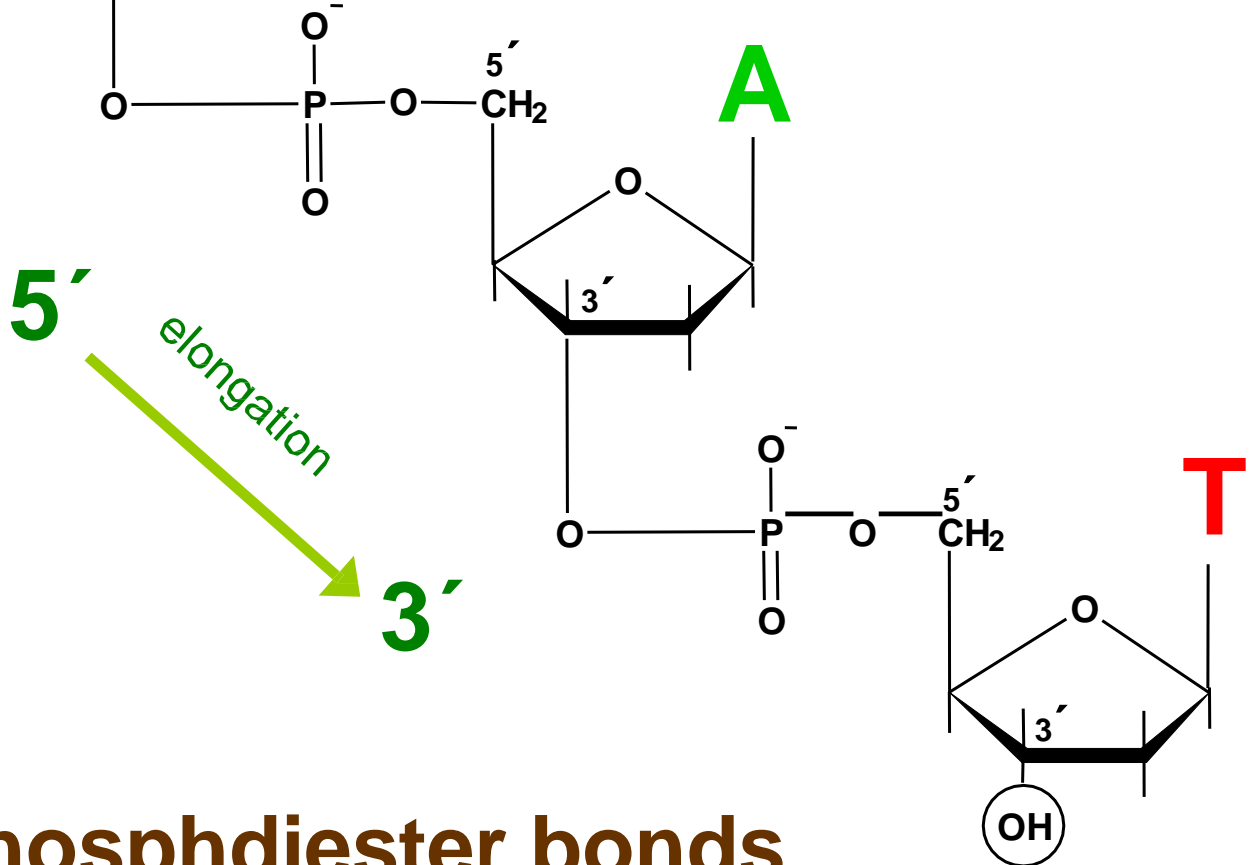
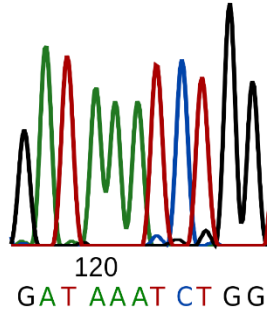
Genomes....

....

# DNA primary structure

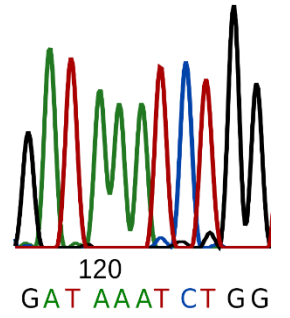


# DNA sequencing



# Phosphodiester bonds

# DNA sequencing

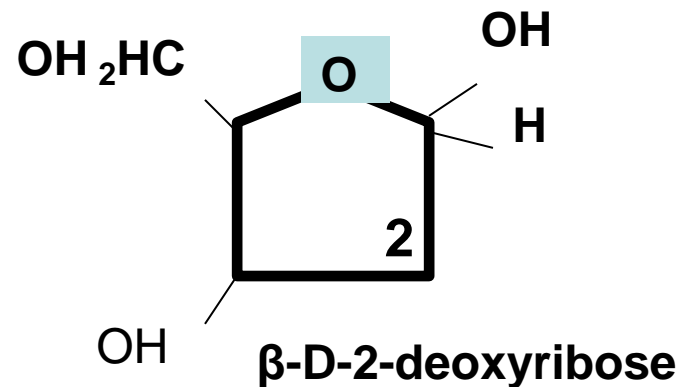


„Clasic“ Sanger sequencing (1977)

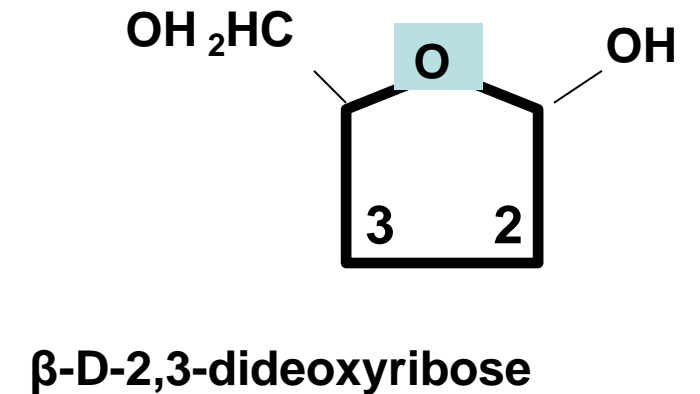
→ Length of sequence: ~1000nt

→ Output: „text“ 4 letters (ACTG)

## Deoxynucleotides x **dideoxynucleotides**

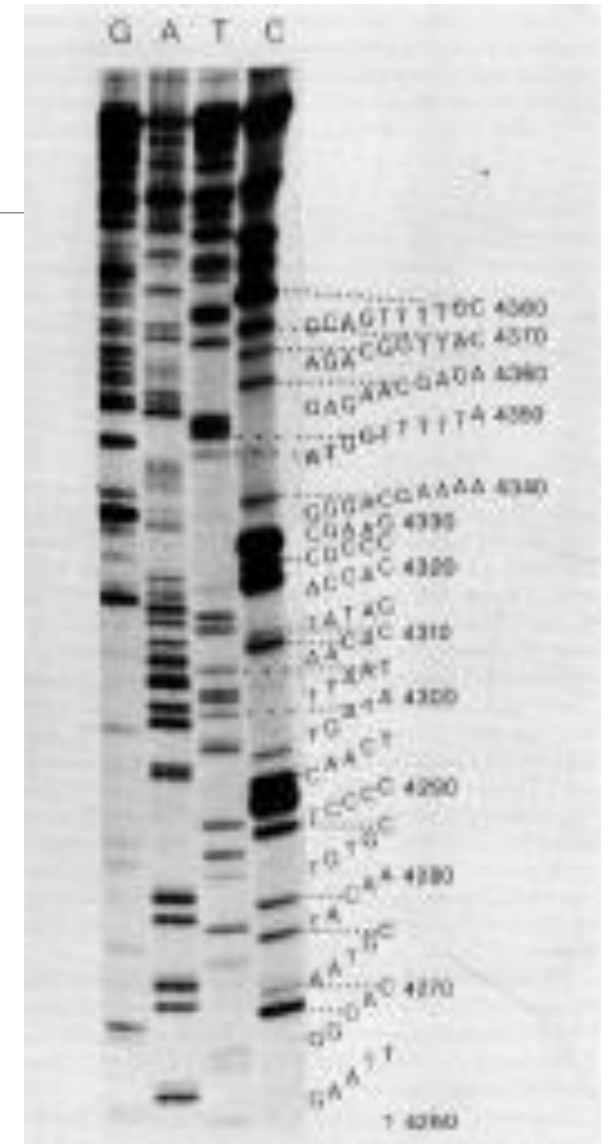
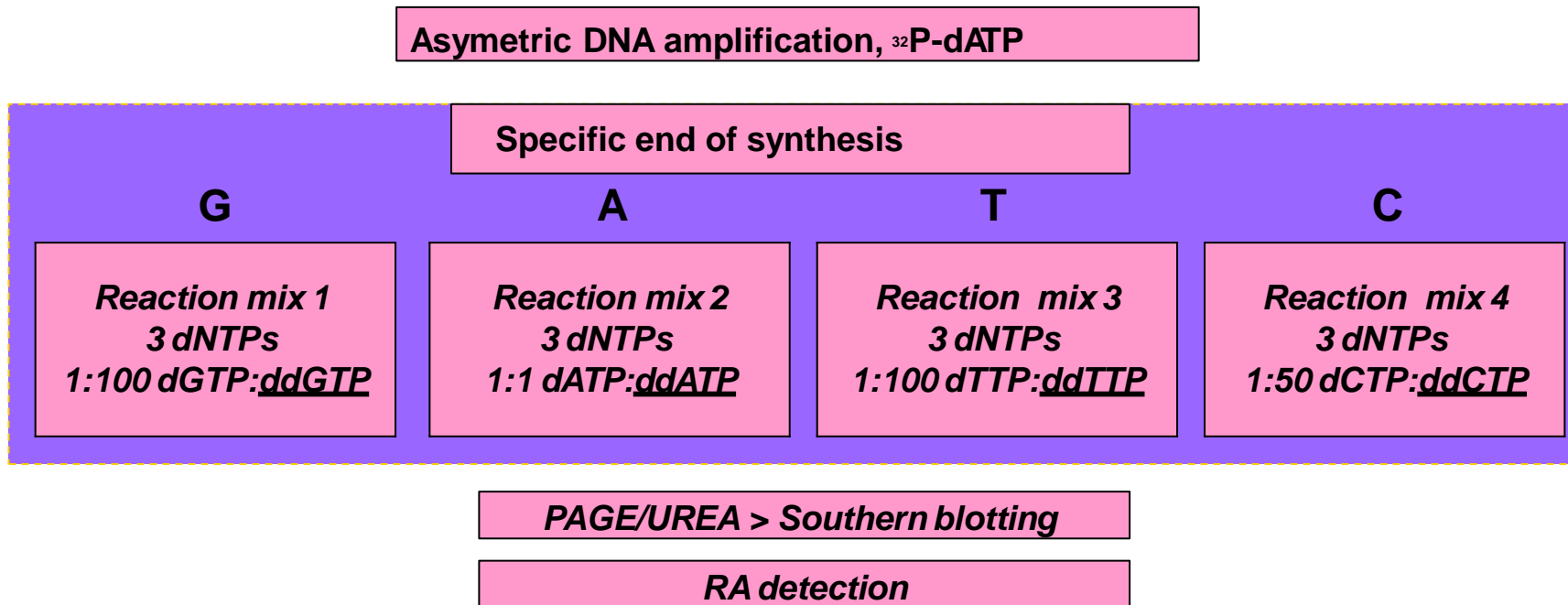


→  
~1:100



# DNA sequencing

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

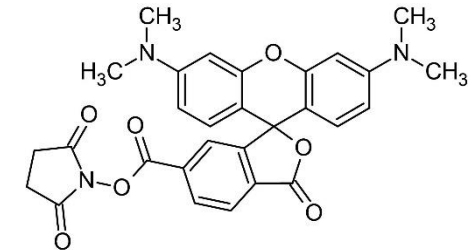
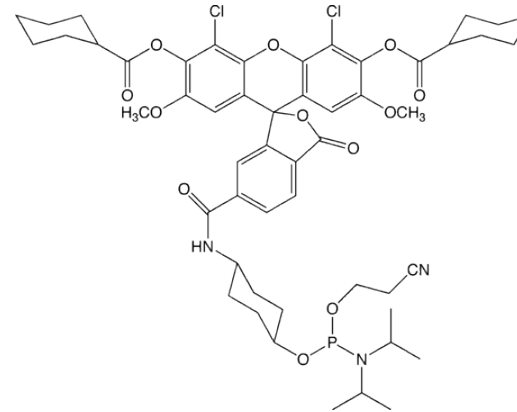
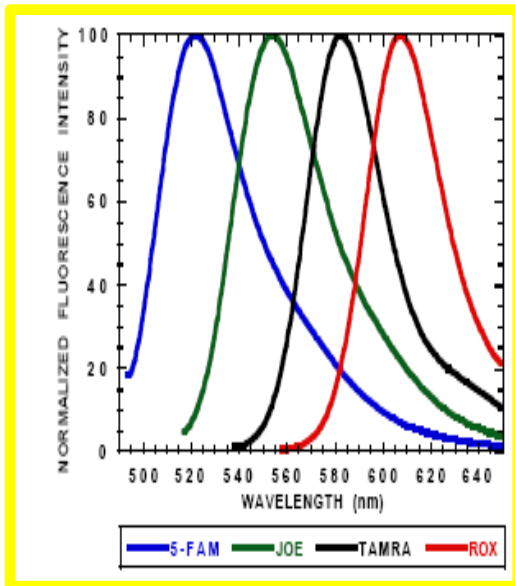




# DNA sequencing

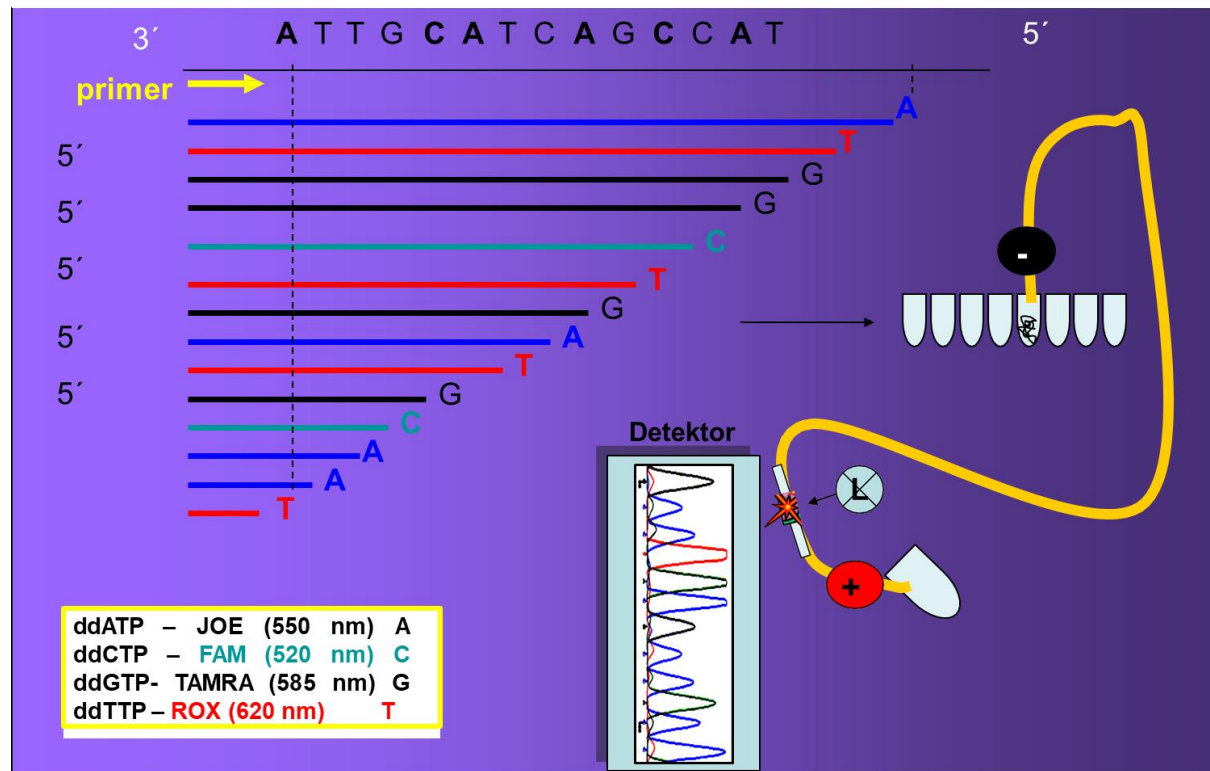
1980s new fluorophores enabled automatization

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

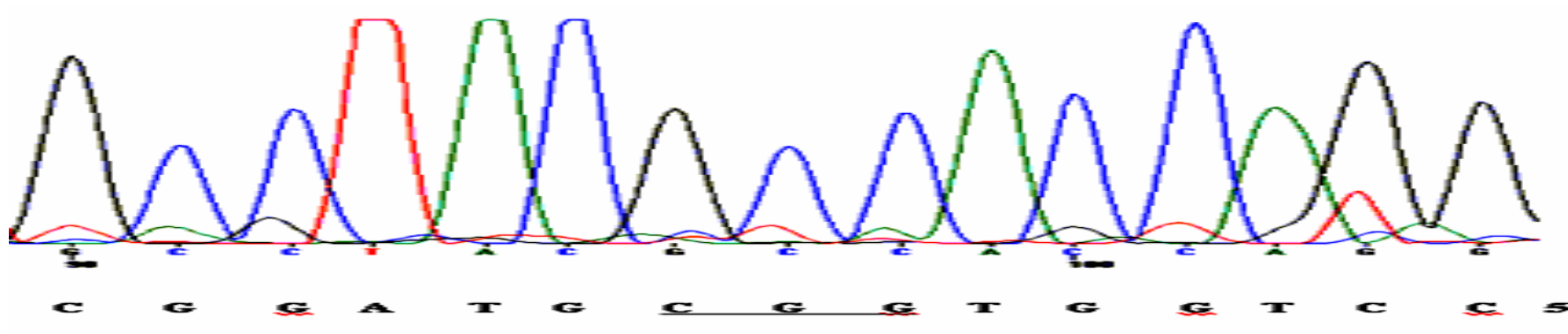
---

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

„plain text“: CGGATGCGGTGGTCG

„fasta“: >identifier  
CGGATGCGGTGGTCG

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



# Sequence data analysis

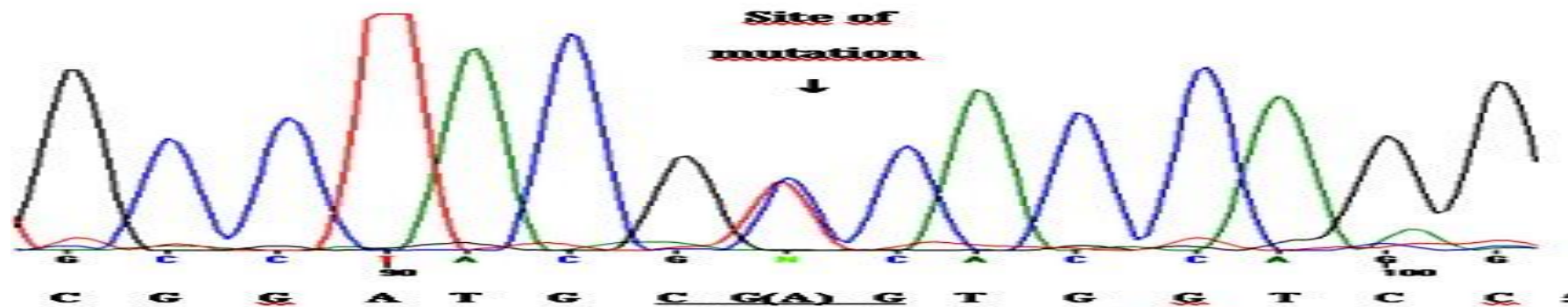
---

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

„plain text“: CGGATGCNGTGGTCG

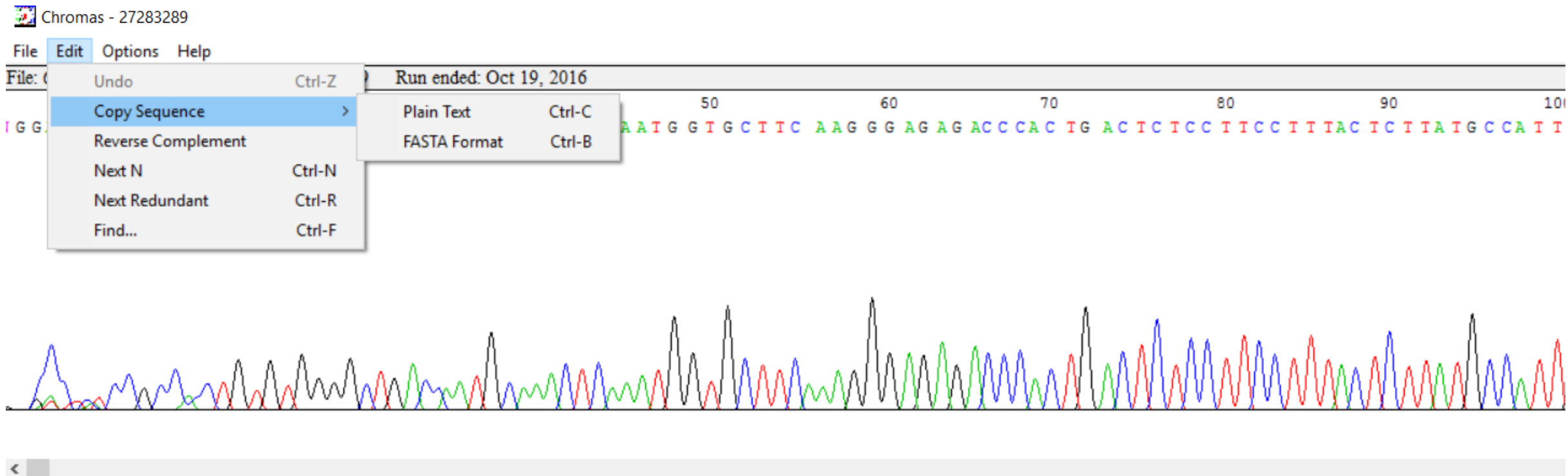
„fasta“: >identifikace  
CGGATGCNGTGGTCG

„sekvenační formát“(.scf, .abi, .ab1)



# Sequence data analysis

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



Try

---

## Run „chromas“ from Moodle

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

What does it code?

From which organism does it probably come from?

# Sequence data analysis (Ex3) - Blastn

BLAST<sup>®</sup> » blastn suite » RID-ES1YNVDF016

[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 (1167 letters)

RID [ES1YNVDF016](#) (Expires on 04-12 06:31 am)

Query ID [Id|Query\\_42683](#)

Description None

Molecule type nucleic acid

Query Length 1167

Database Name nr

Description Nucleotide collection (nt)

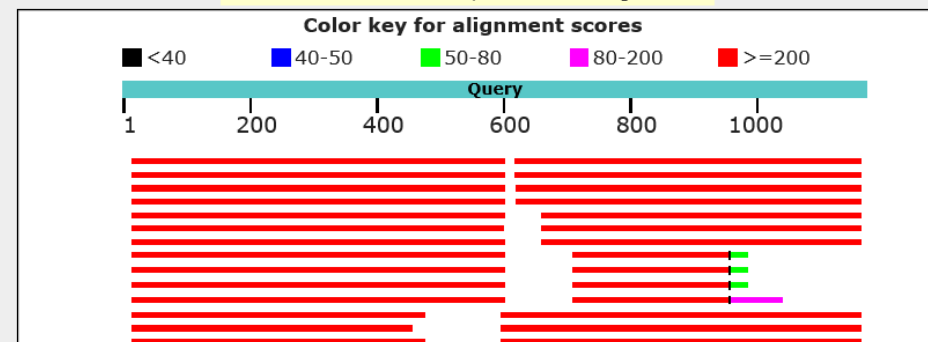
Program BLASTN 2.6.0+ [Citation](#)

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#) [MSA viewer](#)

### Graphic Summary

Distribution of the top 114 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments



# Sequence data analysis (Ex3) - Vecscreen

NCBI Resources How To jostovap My NCBI Sign Out

VecScreen All Databases  Search

BLAST® » vector contamination » RID-ER5PR0JA016 Home Recent Results Saved Strategies Help

## BLAST Results

Formatting options Download YouTube How to read this page Blast report description

Vecscreen

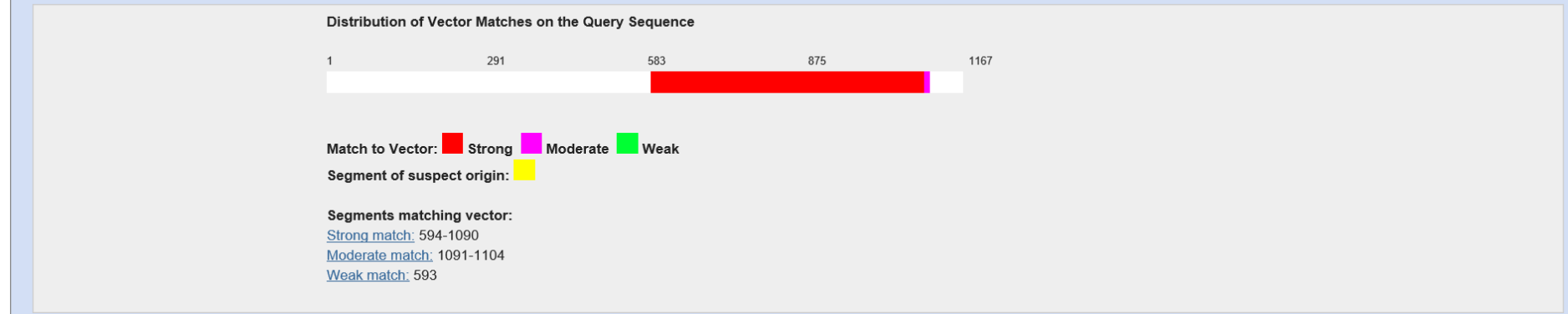
Job title: 69BF16 sequence exported from chromatogram

Interpretation of VecScreen Results

RID	ER5PR0JA016 (Expires on 04-11 22:29 pm)	Database Name	screen/UniVec
Query ID	lcl Query_123453	Description	UniVec (build 9.0)
Description	69BF16 sequence exported from chromatogram file	Program	BLASTN 2.6.0+ Citation
Molecule type	nucleic acid		
Query Length	1167		

Other reports: Search Summary Taxonomy reports Distance tree of results MSA viewer

### Graphic Summary





# Sequence data analysis: purifying sequence of vector

## SMS „Range Extractor DNA“

**SMS** Sequence Manipulation Suite:  
Range Extractor DNA

Range Extractor DNA accepts a DNA sequence along with a set of positions or ranges. The bases corresponding to the positions or ranges are returned as a sequence, a set of FASTA records, as uppercase text, or as lowercase text. Use Range Extractor DNA to obtain subsequences using position information.

Paste a raw sequence or one or more FASTA sequences into the text area below. Input limit is 500000 characters.

```
AATAAACAAAGTTAACAAACAACAATTCATTTCATTTTATGTTTTAGGTTTCAGGGGGAGATG
TGGGAAGGTTTTTTTAAAGCAAGTAAAACCTCTACAAATGTGGTAAAAATCGAATTTTAAACA
AAATATTAACGCTTACAATTTCTGATGCGGTATTTCTCCTTACGCATCTGTGCGGTAT
TTCACCCCGCATACGCGGGATCTGCCAACACCATGGCCCTGAAATAACCTTCTGAAAAG
AGGAACTTGGGTTAGGTACCTTCTTGAGGCTGAAAAAAACCATCTTGGGAAATGTGTGCT
ACATTTAGGGTTGTAGAATTCTCAAAG
```

Enter the base positions or ranges to be extracted. Use ".." to represent a range, and use a comma to separate entries. The words 'start', 'end', 'center', 'beginning', 'end', 'middle', and 'length' of the sequence. Arithmetic expressions can be included in the ranges. For example, to obtain the 30 bases on either side of the center base along with the center base, the ranges '(center - 1)..(center + 1)' and '(center - 30)..(center + 30)' can be used.

Please check the [browser compatibility](#) page before using this program.

- Obtain bases from the  strand.
- Sequence segments should be returned as

\*This page requires JavaScript. See [browser compatibility](#).  
\*You can [mirror this page](#) or [use it off line](#).

# Homework 6

Work with „your“ nucleotide sequence.

---

- 1) Compare your mRNA and CDS sequence (repetition from Lesson 5)
- 2) Translate „your“ nucleotide sequence (mRNA), in which ORF is the CDS?
- 3) Download unknown sequence „Homework 6.ab1“and open it in chromas.

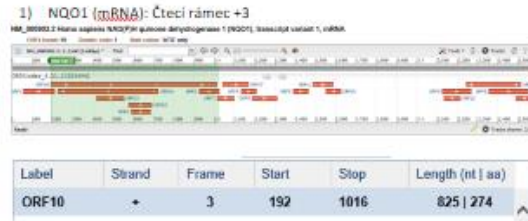
\*Check for vector contamination and identify „pure“ sequence

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

- 4) View „PCR Primer Design“ [https://www.youtube.com/watch?v=c-f1H07D\\_70](https://www.youtube.com/watch?v=c-f1H07D_70)

# Homework 6-example

D06



2)



3)

