

Introduction to applied bioinformatics

PETRA MATOUŠKOVÁ
2024/2025

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' ACTACTAGATTACTTACGGATCAGGTACTTAGAGGCTTGCAACCA 3'
||||| ||||| ||| ||||||| ||||| |||||

Query Sequence

5' TACTCACGGATGAGGTACTTAGAGGC 3'

Global Alignment

Target Sequence

5' ACTACTAGATTACTTACGGATCAGGTACTTAGAGGCTTGCAACCA 3'
||||| ||||| ||| ||||||| ||||| |||||
5' ACTACTAGATT----ACGGATC--GTACTTAGAGGCTAGCAACCA 3'

Query Sequence

Nucleotide sequence comparison

-Analogous to protein comparison

e.g. multalin

→ Comparing/looking for short sequences (primers)

Default matrix:BLOSUM62

input:

>sequence

GGGCACACTCCAGCAGACGCCGAATTCAAATCCTGGAAGGATGGAAGAAACGCCCTGGAGAATATTTGGG
ATGAGACACCCTGTATTTGCTCCAAGCAGCCTTTGACCTAAACTCCAGGCAGGATTCTTAATGAA
AAAAGAGGTACAGGATGAGGAGAAAAACAAGAAATTGGCCTTCTGTGGGCCATCACTGGCAAGTC
ATCCC ACTGACAACCAAGATCAAAGCTAGAAAATGAGATTCTTAGCCTGGATTTCTTAACATGTTA
TCAAATCTGGGTATCTTCCAGGCTTCCCTGACTTGCTTAGTTAACAGATTGTGTTTCTTTCC
ACAAGGAATAATGAGAGGGAATCGACTGTATT CGTCATTTGGATCATTTAAC TGATTCTTATGA
TTACTATCATGGCATATAACCAAAATCCGACTGGCTCAAGAGGCCACTAGGGAAAGATGTAGAAAGAT
>2.exon
CAGGATGAGGAGAAAAACAA

Alignment parameters

- #### • Symbol comparison Table - Gap open def. - Gap ext def.:

Blosum62 - 12 - 2 ▾ if 'Personal' select a file: Nevybrán žádný soubor

Sequence comparison

open Multalin

from Examples:

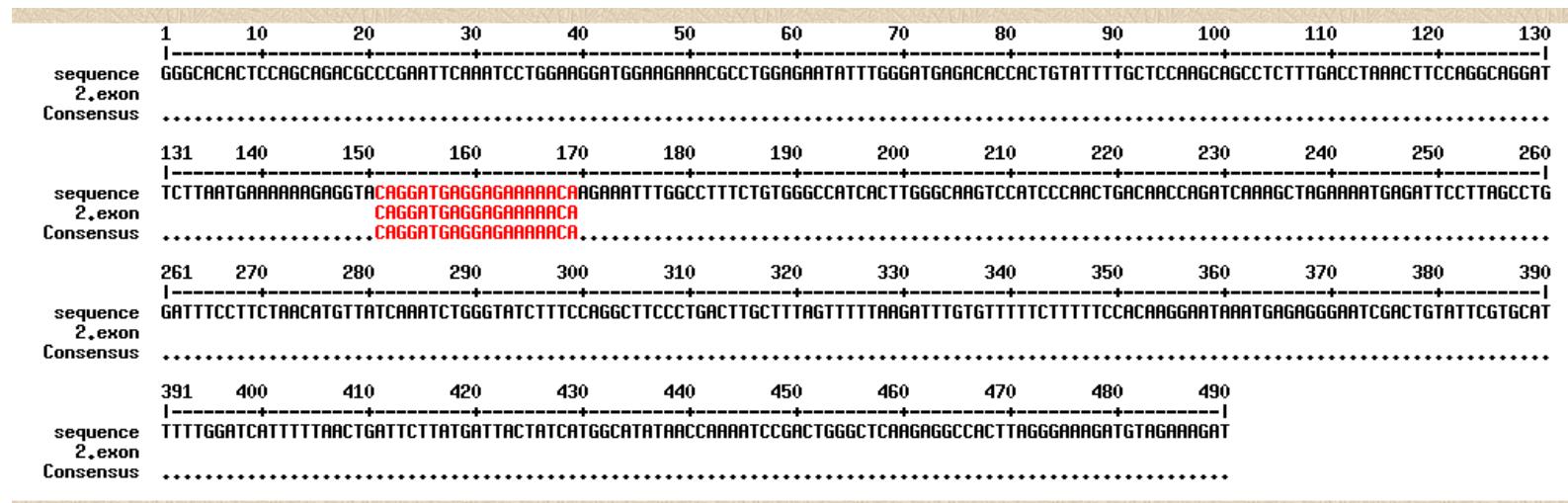
- A: compare long and short sequence (default Blosum matrix)

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

The screenshot shows the 'Alignment parameters' dialog box. At the top, there is a dropdown menu labeled 'bol comparison Table - Gap open def. - Gap ext def.'. Below it, there is a button labeled 'Zvolit soubor' (Select file) and another button labeled 'Nevybrán žádný soubor' (No file selected). In the center, there is a label 'default: value from comparison table'. At the bottom, there are two input fields: 'opening:' containing 'default' and 'extension:' also containing 'default'. A yellow question mark icon is positioned next to the 'gap open def.' and 'gap ext def.' labels.

→DNA-5-0

```

  1       10      20      30
+---+ +---+ +---+
NM_015696.4 GTCTTTGCCATCGCAGCGGCCACCTCCG
69BF16          NGGCATTCTCCCGCAGCTGTGTGGG
Consensus .....gccAacgCgaCccCaCaccTccG

```

```

  131   140   150   160
|-----+-----+-----+
NM_015696.4  TCACATCCGGGGCARTGTTCGTGG
  69BF16      AATTTCTAGTATTTTGATTTTTGARTCT
Consensus    aaAcaccaGgagccaaaTgaTgcatcg

```

	261	270	280	290
NM_015696.4	GGGCCCCACCACTTAACTGTCGCGCTT			
69BF16	AGTATATCARG-CATAATCTCCACCCRA			
Consensus	aGacacC A c.CaaTA <u>CC</u> Tc <u>C</u> ac <u>CC</u> aa			

```

  391   400   410   420
+---+---+---+---+
NM_015696.4  AGGATTGCAGTCACCGGTATGGTGCCCAT
69BF16        AGGCATTCATGACCATTTTTG--CATA
Consensus     AaGaaTcGaTcCaCagTacTG...Caaa

```

NM_015696.4 **69BF16** **521** **530** **540** **550**
ACCCCACTGTGTCAGTGAGGGAGGTCAAGAC
CTGTATTATTTCTCATTCACAAAAGAAAT

NM_015696.4
69BF16 ATTCGAGTCACCGGTACTGGTCCCCATCCTGCCTTCAGTACCTGGCCCGAGCTCTGGGAGGGAGGCCACCTGGAACTTCTGGAGTGACCTAGTAGCCCCAGATGGAAAGGTGGTAGGGGGCTTGGGA

521 530 540 550 560 570 580 590 600 610 620 630 640 6

NM_015696.4
69BF16

CAACTGTGTCAGTGGAGGAGGTCAAGACCCCCAGATCAGCGCTCGTGAGGAAGCTCATCTACTGAAAGCGAGAAGACTTTATAACCACCGCGTCTCCCTCCAC
CACCTCTATCCGCCACCTGTGTC
NGGCATTCTCCCGCAC---TGTGTG

651 660 670 680 690 700 710 720 730 740 750 760 770 780
NM_015696.4 GCTG-ACCAAA-TGCAAACTCAA-TGGTGTTCAAA-GGGAGAGACCCACTGACTCTCTTCTTACTCTTGCCATTGGTCCCACATTCTGTGGGGGAAAAATTCTAGTATTTGATTATTGAA
69BF16 GCTGGACCAAA-TGCAAACTCAA-TGGTGTTCAAA-GGGAGAGACCCACTGACTCTCTTCTTACTCTTGCCATTGGTCCCACATTCTGTGGGGGAAAAATTCTAGTATTTGATTATTGAA

Nucleotide sequence comparison

-Analogous to protein comparison

e.g. multalin

→ Comparing genomic DNA and cDNA (mRNA)

Default:BLOSUM62 →DNA 5-0

	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026						
NG_011504.1	AGACRACRAC	CTG	GANGCT	TAAA	TTGGT	TRACGC	CCTG	GRRGGT	TRAGRG	RGRGC	CTG	TCCTC	ACTGAT	TCTCTGTG	TCCCT	GRGGCC	CTCTTA	-TCAGA
NM_000903.2	---	CCATCT	TC	ARGGCT	GTTT	CGC	CGT	TCTRC	GRGG	GT	TTC	---	RC	TCCTC	TCCTG	TGAC	GRGGAC	CCCT
Consensus	-----	Caacaa	CTG	ARGGCT	Taaa	Taa	Tca	Gc	Gc	Gc	Ga	Ga	RC	Taa	Cc	TGaa	Gaa	Taa

The figure displays a sequence alignment between three entries: NG_011504.1, NM_000903.2, and Consensus. The x-axis represents the genomic position from 20671 to 20770. A dashed line indicates regions of high conservation, while arrows above the sequence show the direction of divergence for each entry.

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

„reading“ DNA = translation

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

X we don't know which strand is coding:

5'-ATGGAAGTATTAAAGCGCCACCTATTAA-3'

3' -TACCTTCATAAAATT CGCGGTGGATAAAATT-5'

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

„reading“ DNA = translation

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

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

„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.

```
TTAAATAGGTGGCGCTTAAATACTTCAT
```

Please check the compatibility panel for your sequence.

Submit Clear

reading frame 1 reading frame 2 reading frame 3 reading frame 4 reading frame 5 reading frame 6 reading frame 7 reading frame 8 reading frame 9 reading frame 10 reading frame 11 reading frame 12 reading frame 13 reading frame 14 reading frame 15 reading frame 16 reading frame 17 reading frame 18 reading frame 19 reading frame 20 reading frame 21 reading frame 22 reading frame 23

direct reverse

standard (1) vertebrate mitochondrial (2) yeast mitochondrial (3) mold mitochondrial (4) invertebrate mitochondrial (5) ciliate nuclear (6) echinoderm mitochondrial (9) eukaryotic nuclear (7) 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)

compatibility panel

Translate results

```
>rf 1 Untitled
LSQQIQLASVIRPCSPFFL*PLLCCYHGSDTQFHQ*SDRFHQDVSSLVRQRNCLLHQQL
SSRFGQ*SCLFRYRLYDMVFR*SQIRSHVHNLLIMS*LNQCLDPQVFRLDTYDHSESSGRH
VLPFVHVHASNHQVLMLIDDVCVSTQQVGSL
```

Translate results

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

Fri Jun 17 16:17:06 2017
Valid XHTML 1.0; Valid CSS

„reading“ DNA = translation

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
-Identical Seq
-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

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 you specify.

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  
ATGGTCGGAGAGAGCACTGATCGTACTGCCCTCACTCAGAGAGGAGCTTCAACTATG  
CCATGAAGG  
AGGCTGCTGCAGCGCTTGAAGAAAGAAAGGATGGAGGTGGAGTCGGACCTCTATGC  
CATGAACCT
```

Submit Clear Reset

- Translate in reading frame 1 on the direct strand.
- Use the standard (1) 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

Insert both CDS and mRNA in fasta format

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
MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPPIISRKDITGKL
KDPANQYPAESVLAYKEGHLSPLDEAQKKLEAADLVIQFPLQWFGVPAILKGWFERV
FIGEFAYTYAAMYDKGPFRSKKAVLSITTGGSGSMYSLQGIHDMNVILWPIQSGLHLFC
GFQVLEPQLTYSIGHTPADARIQILEGWKKRLENIWEDETLFYPAPSSLFDLNFEQAGFLMK
KEVQDEEKNKKFGLSVGHHLGKSIPTDNQIKARK*

>rf 1 NM_000903.3 Homo sapiens NAD(P)H quinone dehydrogenase 1 (NQO1), transcript variant 1, mRNA
TRDSHKVAAGAAQLTESLVPARVAPATTSPANQRGLHQSHGRQKSTDRTGSLREDVLQL
CHEGGCCSGFEEERMGGGGVGPLCHELQSHHFQKGHH*TEGPCELSVSCRVCSSG*RRP
SEPRYCG*TKEAGSRRPCDIPVPPAVVWSPCHSERLV*ASVHRRVCLHLRCHV*QRTLPE
*EGSAFHWWQWLHVSARDFRGHECHSLANSEWHSALFLWLPSLRISTSDI*HWAHSSRR
PNSNPNRMEETPGYEYLG*DTTVFCSKQPL*PKLPGRLNEKRGTG*GEKQEIWFCGCPSL
GQVHPN*QPQDS*KMRLSLDFLLTCYQIWIWSFQASLTFCFF*DLCFSFSTRNK*EGIDC
IRAFLDHF*LILMITIMAYQNQNPGLKRPLRERCRKMLEKCSLKASTQFNSSF*G*SFRV
QFG*VSFNSPMFY*SPLCSLWQKGIAQRKRLNLPALRDLTCLVVSMLVYDISCFQLQ
SSY*YA*HKYHSWSAFVVYIQTDTLKGANKSLLCCSHILLFF*LKKIFF*SSLALLPRL
ECSGVISAHCNLCCLPGSSNNSPASASLVAGMTGACHHA*LIFVFLVETAFHHVGQAGLKLL
TSGDDPTTSQSQSAGITGVIIHTWPLQSSTLRFAEIINQ*IHTVHLQYEFKKGNSTFNT*K*
SSTKNTLFLIYTNFQKVIIIAKLMTYYGMGSSPMTLGYNCKPRLSTLVNSFGIIIVNF
YFWKSSHSTVGII*FKENMIKHCPLVVH*KKR*EMMKRLPEKWTASYLPRK*RDWTELE
NLLYQMLTGTGGFCSRQYPQ*LTAGCFSLKIFCCLLHHCNFV*ISQRSELNK*NSQLQTH

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 your full length mRNA and CDS in *SMS translate*
- spot the difference

„reading“ DNA = translation

ORFfinder → looking for the longest ORF

Open Reading Frame

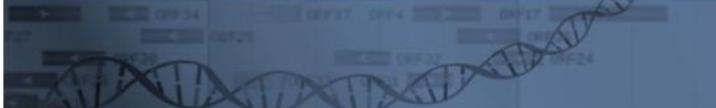
NCBI Resources How To

ORFfinder PubMed Search

jostovap My NCBI Sign Out

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:

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

Ignore nested

Start Search /

Submit Clear

Detailed description: This screenshot shows the 'Open Reading Frame Finder' interface on the NCBI website. It includes a navigation bar with links to NCBI, Resources, How To, and user-specific options like jostovap, My NCBI, and Sign Out. Below the navigation is a search bar with dropdown menus for 'ORFfinder' and 'PubMed', and a 'Search' button. The main content area is titled 'Open Reading Frame Finder' and explains the function of the tool: searching for open reading frames (ORFs) in a DNA sequence. It also mentions returning the range of each ORF and its protein translation. A background image of a DNA double helix is visible. On the left, there's a sidebar titled 'Choose Search Parameters' containing several configuration options. One option, 'Genetic code', has a dropdown menu open, showing a list of 25 different genetic codes numbered 1 through 25. The first item, 'Standard', is highlighted with a blue background. Other options in the sidebar include 'Minimal ORF length (nt)' set to 75, 'ORF start code' with three radio button options ('"ATG" only', '"ATG" and others', 'Any sense'), and 'Ignore nested'. At the bottom, there are 'Start Search /' buttons and 'Submit' and 'Clear' buttons.

„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

NM_000903.2: 1..2.6K (2.6Kbp) Find: Tools Tracks ?

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

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

Add six-frame translation track

ORF10 (274 aa)

>lcl|ORF10
MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGEVVE
DLYAMNFNPPIISRKDITGKLKDPMFQYPAESVLAYKEGH
LSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWF
FIGEFAYTYAAMYDKGPFRSKKA
VLSITGGSGSMYSLQG
IHGDMNVILWPIQSGILHF
CGFQVLEPQLTYSIGHTP
ADA
RIQILEGWKKRLEN
IWI
DETPLYFAPSSLFDLN
FQAGFLMK
KEVQDEEKNNK
FGLSVGH
HLGKSIPTDN
QIKARK

BLAST ORF10

Label Strand Frame Start Stop Length (nt | aa)

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	-	2	1171	1040	100 40

„reading“ DNA = translation

BLAST® » blastp suite

Standard Protein BLAST

[blastn](#) **blastp** [blastx](#) [tblastn](#) [tblastx](#)

Enter Query Sequence

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

>1cl|ORF10_NM_000903.2:192:1016 unnamed protein product
MVGRRALIVLAHSDTSFNYAMKEAAAALKKGWEVVESDLYAMNFPNIISRKDITGKLKDPAFQ
YPA
ESVLAYKEGHLSPLDIVAEQKKLEAADLVIFQFFLQWFGVPAILKGWFERVFIGEFAYTYAAMYDKGP
FRS

From _____ To _____

Or, upload file _____

Job Title _____
Enter a descriptive title for _____

Align two or more sequences [?](#)

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

Query seq. Specific hits Superfamilies

Flavodoxin_2

FMN_red superfamily

1 50 100 150 200 250 274

Choose Search Set

Database: UniProtKB/Swiss-Prote

Organism: Enter organism name
Optional: Enter organism common name
Enter organism common name

Exclude: Models (XM/XP)

Entrez Query: Enter an Entrez query to li
Optional: Enter an Entrez query to li

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

■ <40	■ 40-50	■ 50-80	■ 80-200	■ >=200
-------	---------	---------	----------	---------

Query

1 50 100 150 200 250

Program Selection

Algorithm: blastp (protein-protein)

„reading“ DNA = translation

ORFfinder → looking for the longest ORF

The screenshot shows the NCBI ORFfinder interface and a BLAST search results page.

NCBI ORFfinder: The top navigation bar includes links for NCBI, Resources, How To, jostovap, My NCBI, and Sign Out. The main search bar has "ORFfinder" selected and "PubMed" dropdown. A search button is present. Below the search bar, the title "Open Reading Frame Viewer" is displayed. The "Sequence" section shows a sequence length of 1: 1.2.5K (2,521 nt). It displays multiple overlapping reading frames (ORF4 to ORF15) with arrows indicating direction. A red box highlights ORF15. The "Mark" button is highlighted with a red box.

BLAST Results: The title is "BLAST® > blastp suite > RID-8M12WFMM01R". The "Job title: Protein Sequence" section shows the Query ID as Icl|Query_57926, Description as Icl|ORF15:1101:970 unnamed protein product, Molecule type as amino acid, and Query Length as 43. The Database Name is swissprot, Description is Non-redundant UniProtKB/SwissProt sequences, and Program is BLASTP 2.10.0+. The results table shows the following data:

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

At the bottom left, there is a "SmartBLAST" link and a "BLAST" button, both highlighted with red boxes.

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**

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information jostovap My NCBI Sign Out

BLAST® Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

[Learn more](#)

NEWS

Magic-BLAST 1.2.0 released

A new version of the BLAST RNA-seq mapping tool is now available.
Mon, 27 Feb 2017 14:00:00 EST

[More BLAST news...](#)

Web BLAST

Nucleotide BLAST nucleotide ▶ nucleotide

blastx translated nucleotide ▶ protein

tblastn protein ▶ translated nucleotide

Protein BLAST protein ▶ protein

BLAST Genomes

Enter organism common name, scientific name, or tax id

Search

Human Mouse Rat Microbes

Unknown sequence identification

→ looking for similarity: BLASTn (or BLASTx)

BLAST® » blastn suite » RID-ES0WVJXE016

Home Recent Results Saved Strategies Help

[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
Molecule type nucleic acid

Database Name nr
Description Nucleotide collection (nt)
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 GGTTTACAATTGTACCCCAAGGTATGGACTGGACCCCATCCATAGTAAGTCATCAGT 60
Sbjct 2003 GGTTTACAATTGTACCCCAAGGTATGGACTGGACCCCATCCATAGTAAGTCATCAGT 1944

Query 61 TTGCAATGATAAAGAAAATAACCTTCTGAAAATTGTATAGATCAGAAATAAAGTATTT 120
Sbjct 1943 TTGCAATGATAAAGAAAATAACCTTCTGAAAATTGTATAGATCAGAAATAAAGTATTT 1884

Query 121 TTTGTGGAAGACTATTTAAGTATTGAAGGTACTATTCCTTCTGAATTCATATTGC 180
Sbjct 1883 TTTGTGGAAGACTATTTAAGTATTGAAGGTACTATTCCTTCTGAATTCATATTGC 1824

Query 181 AGATGTACGGTGTGGATTATTGGTTATCTCTGCAAACCTTAAAGTAGAAGATGCAAG 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

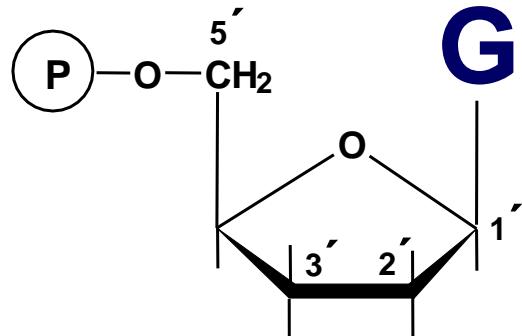
DNA sequencing

Gene expression

microRNA

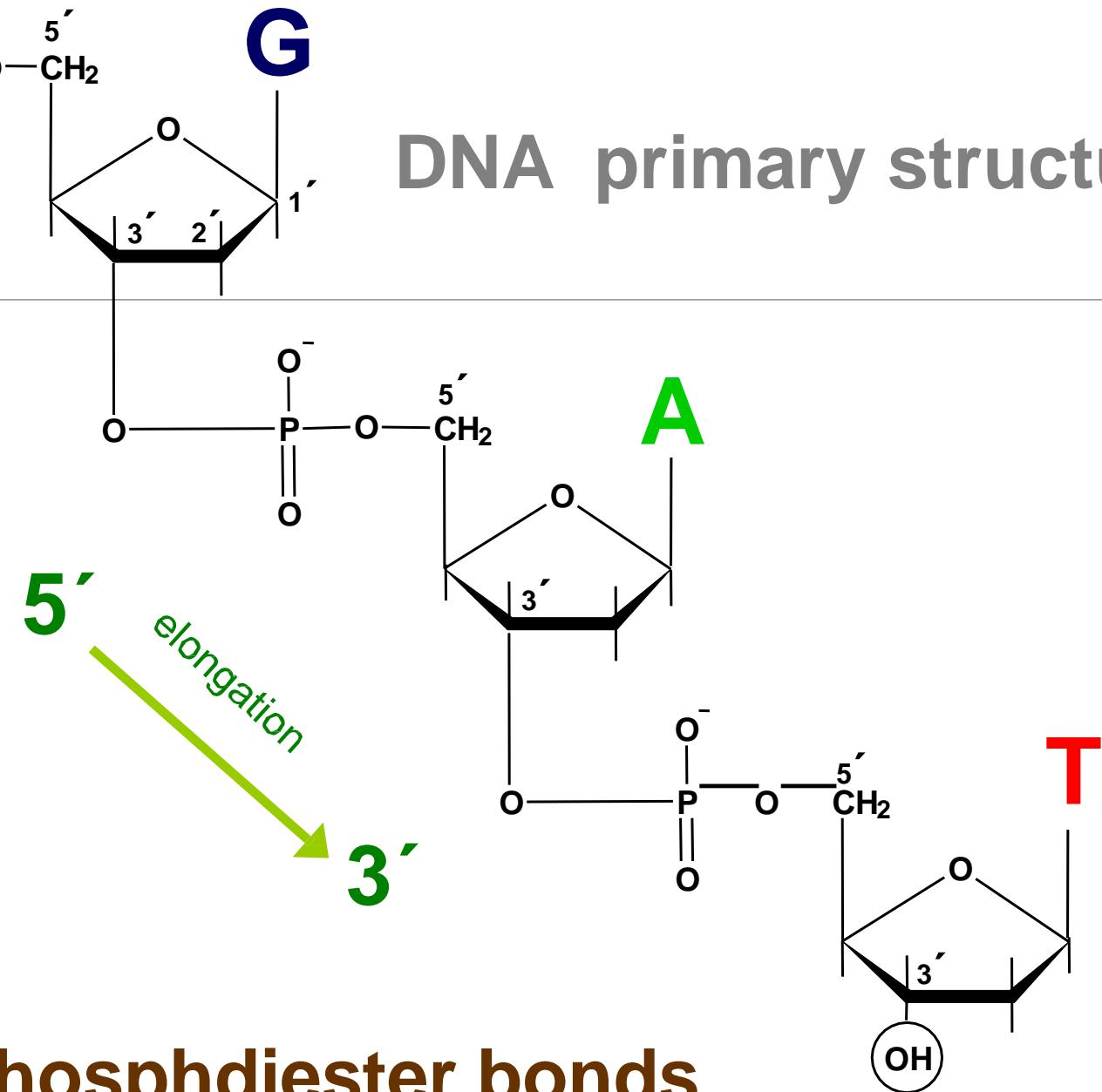
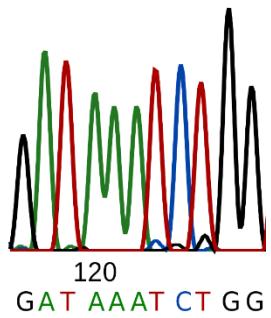
Genomes....

....



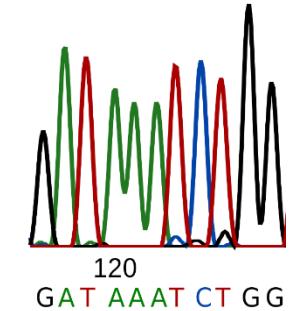
DNA primary structure

DNA sequencing



Phosphodiester bonds

DNA sequencing

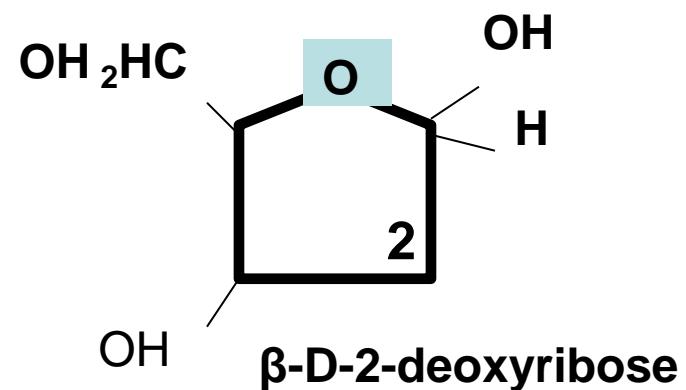


„Clasic“ Sanger sequencing (1977)

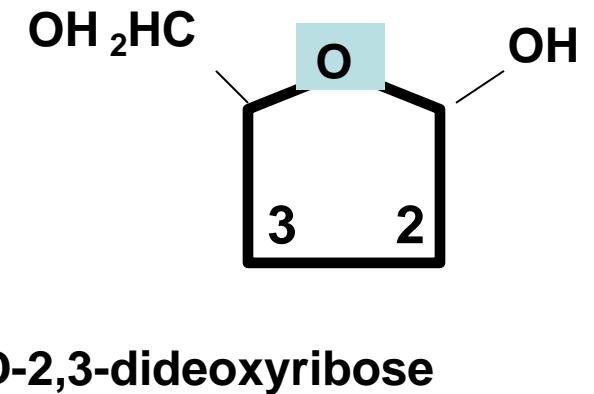
→ Length of sequence: ~1000nt

→ Output: „text“ 4 letters (ACTG)

Deoxynucleotides x dideoxynucleotides

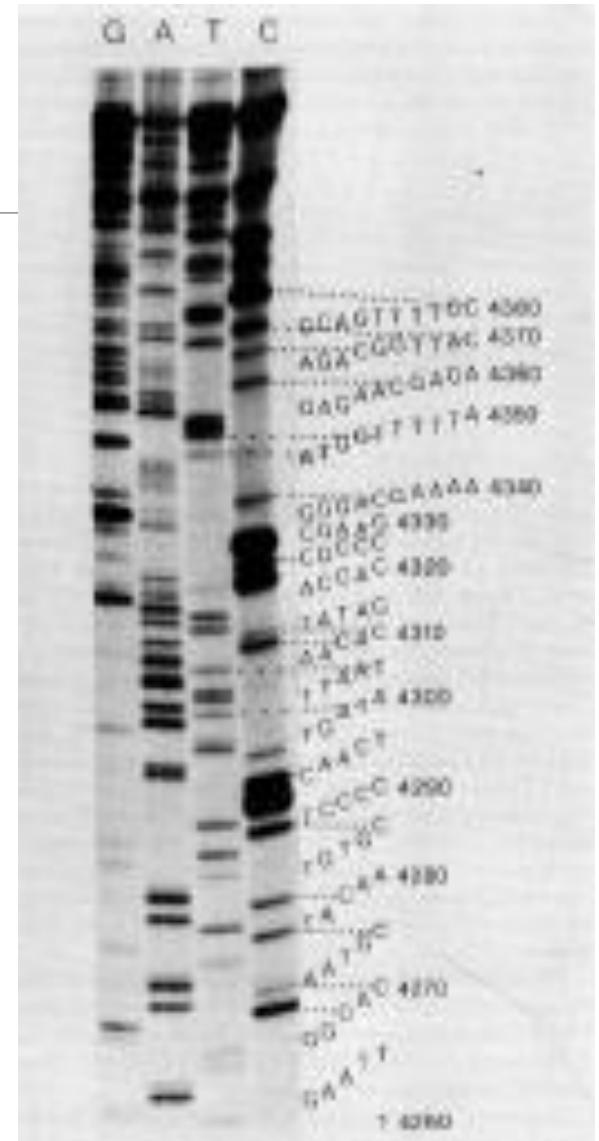
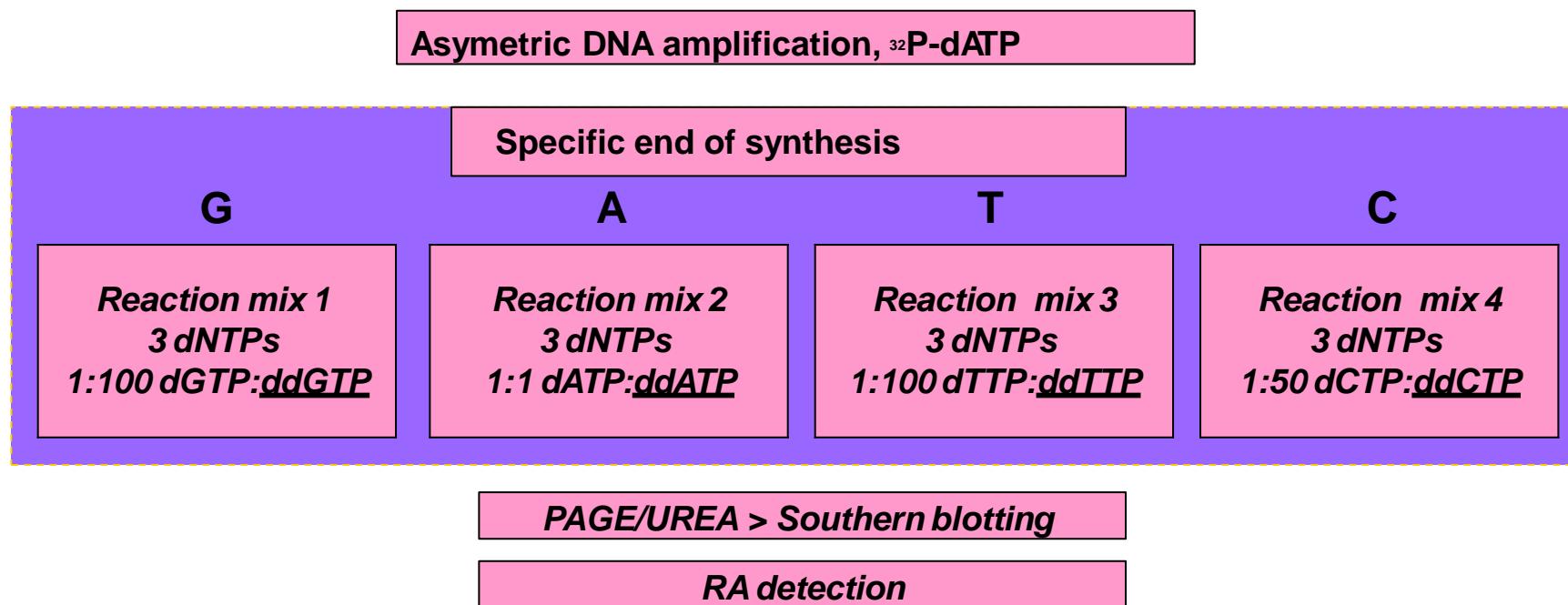


~1:100



DNA sequencing

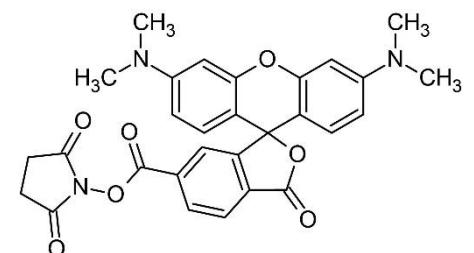
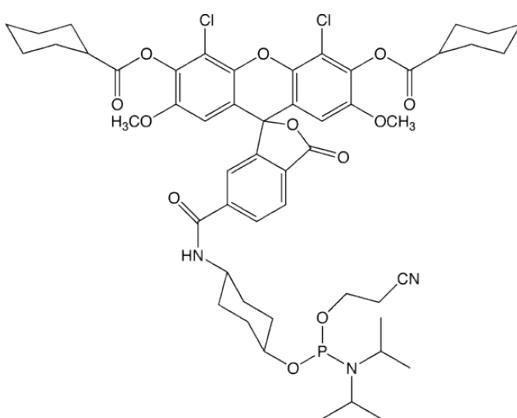
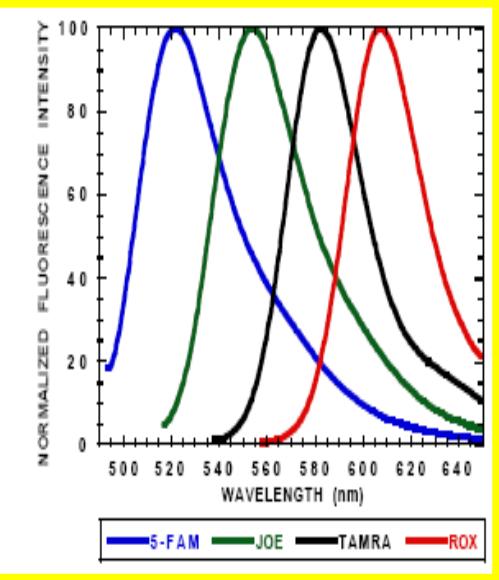
Enzymatic „Sanger“ sequencing (1977) – sequencing by synthesis



DNA sequencing

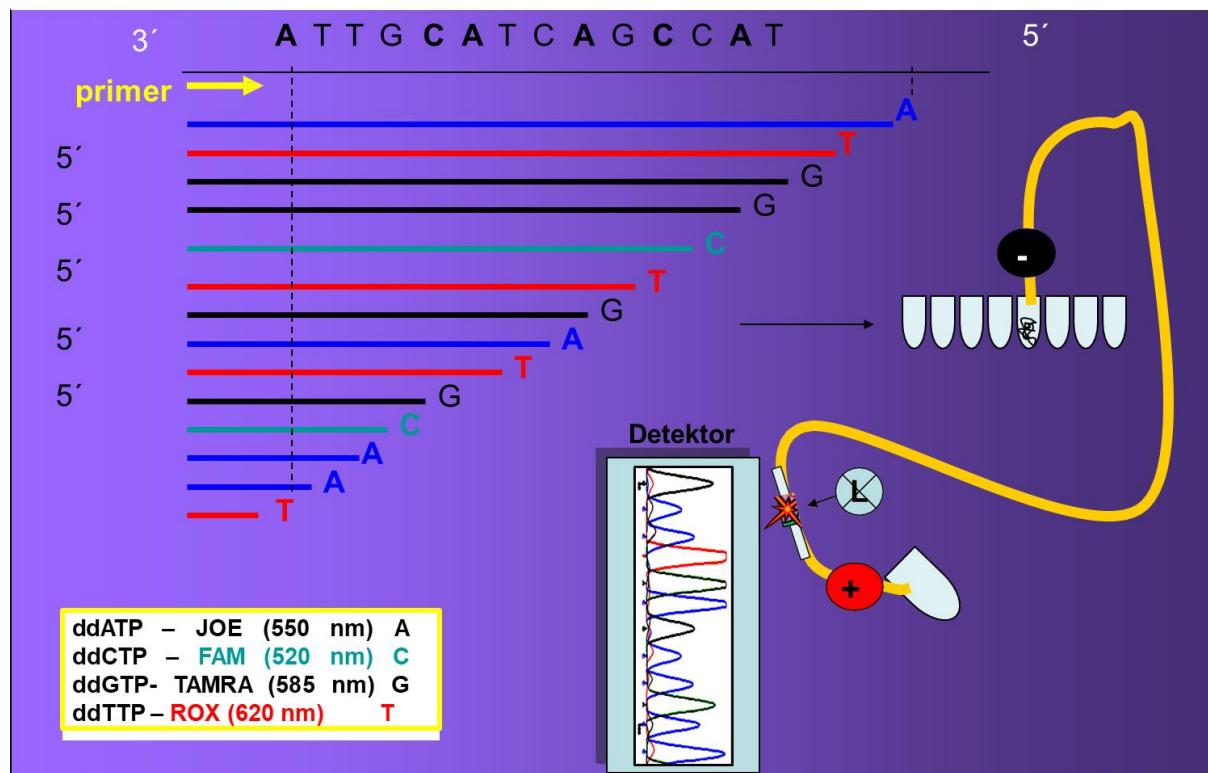
1980s new fluorophores enabled automatization

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



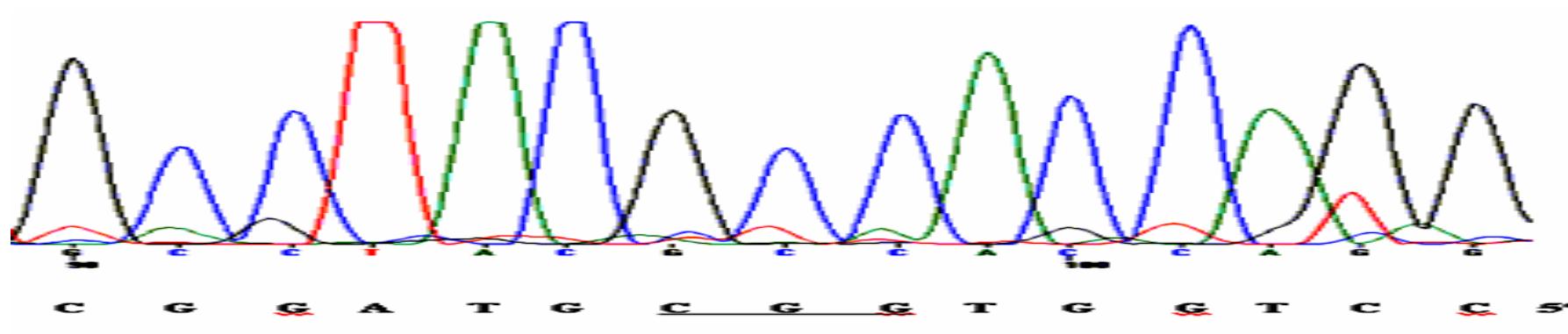
Sequence data analysis

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

„plain text“: CGGATGCGGTGGTCG

„fasta“: >identifier
CGGATGCGGTGGTCG

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



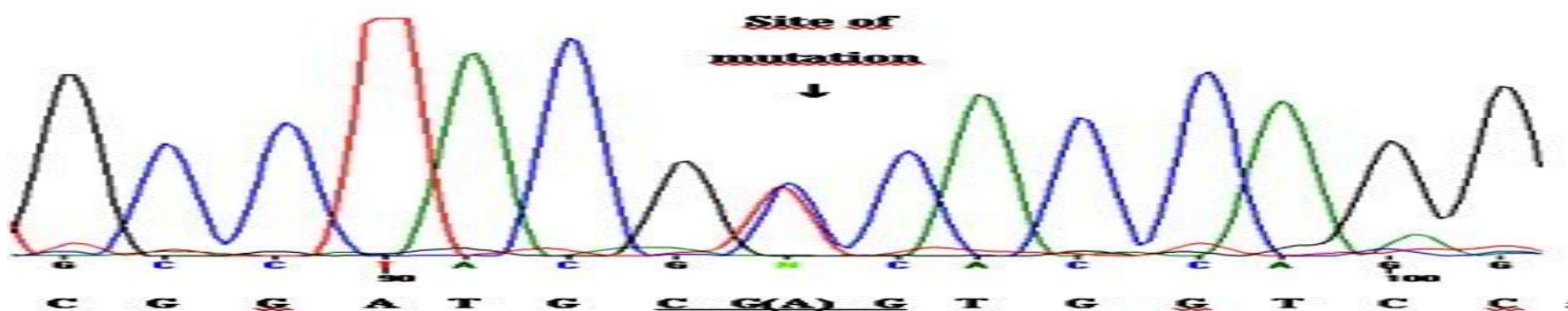
Sequence data analysis

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

„plain text“: CGGATGCNGTGGTCG

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

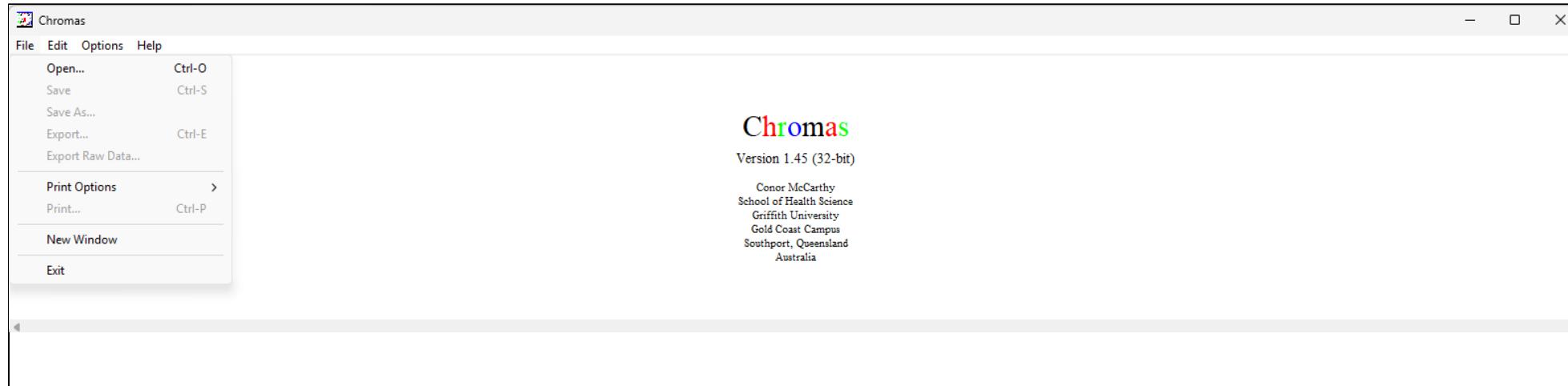
„sekvenační formát“(.scf, .abi, .ab1)



Sequence data analysis

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

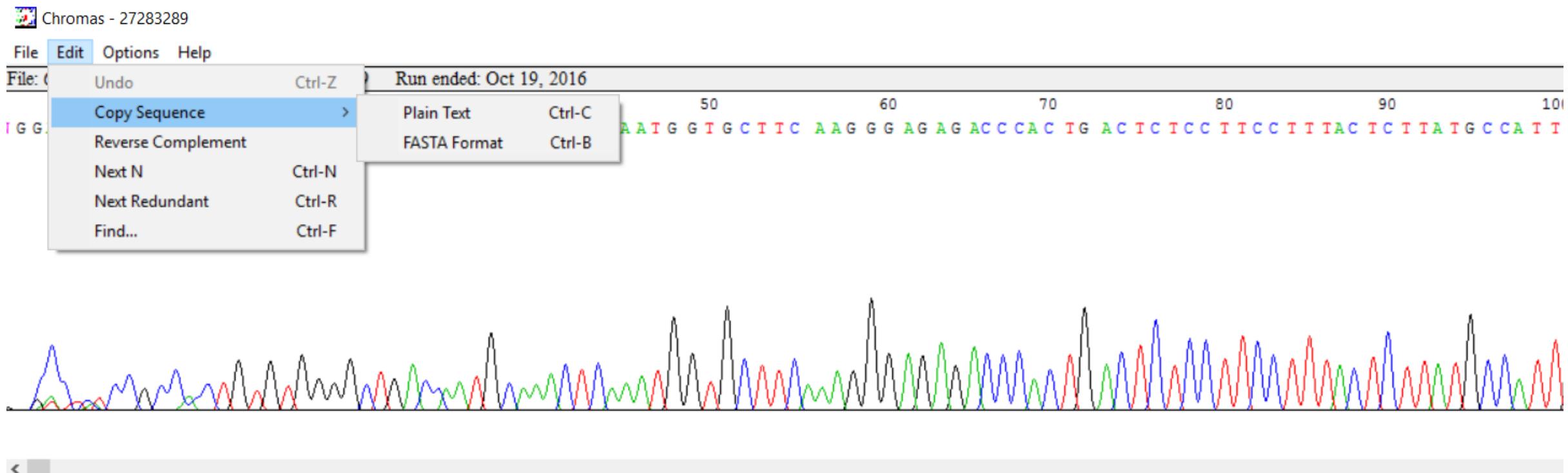
- the sequence has to be first saved and then open in chromas



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

BLAST® » 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 lcl|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

Color key for alignment scores
■ <40 ■ 40-50 ■ 50-80 ■ 80-200 ■ >=200

Query 1 200 400 600 800 1000

Sequence data analysis (Ex3) - Vecsreen

NCBI Resources ▾ How To ▾ jostovap My NCBI Sign Out

VecScreen All Databases ▾ Search

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

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

Job title: 69BF16 sequence exported from chromatogram

RID ER5PR0JA016 (Expires on 04-11 22:29 pm)

Query ID Icl|Query_123453

Description 69BF16 sequence exported from chromatogram file

Molecule type nucleic acid

Query Length 1167

Database Name screen/UniVec

Description UniVec (build 9.0)

Program BLASTN 2.6.0+ ▶ Citation Interpretation of VecScreen Results

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

Graphic Summary

Distribution of Vector Matches on the Query Sequence

Match to Vector: Strong Moderate Weak

Segment of suspect origin:

Segments matching vector:
Strong match: 594-1090
Moderate match: 1091-1104
Weak match: 593

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.

```
AATAAACAAAGTTAACACAAACAATTGCATTTCATTATGTTTCAGGTTCAAGGGGGAGATG  
TGGGAAGGTTTTTAAGCAAGTAAAACCTCTACAAATGTGGTAAATCGAATTAAACA  
AAATATTAAACGCTTACAATTTCCTGATGCCGTATTTCTCCTTACGCATCTGTGCCGTAT  
TTCACCCCGCATACGCCGGATCTGCCAACACCATGCCCTGAAAAAACCTCTGAAAAG  
AGGAACATTGGGTTAGGTACCTCTTGAGGCTGAAAAAAACATCTGGGAAATGTGTGCT  
ACATTAGGGTTAGAATTCTCAAAG
```

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' and 'length' can be used as place of digits, to represent the beginning, end, middle, and length of the sequence. Arithmetic expressions can be included in the ranges. For example, to extract the 30 bases on either side of the center base from a sequence, the range '(end - 2)..end' can be used. To obtain the 30 bases on either side of the center base along with the center base, the ranges '(center - 1)..(center + 30)' can be used.

1..593

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.

Unknown sequence identification – BLASTn !

- if the sequence does not have easily recognizable ORF ?

→ looking for similarity: BLASTn (or BLASTx)

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information jostovap My NCBI Sign Out

BLAST® Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

Magic-BLAST 1.2.0 released
A new version of the BLAST RNA-seq mapping tool is now available.
Mon, 27 Feb 2017 14:00:00 EST [More BLAST news...](#)

Web BLAST

Nucleotide BLAST nucleotide ▶ nucleotide

blastx translated nucleotide ▶ protein

tblastn protein ▶ translated nucleotide

Protein BLAST protein ▶ protein

BLAST Genomes

Enter organism common name, scientific name, or tax id Search

Human Mouse Rat 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 lcl|Query_28989
Description None
Molecule type nucleic acid

Database Name nr
Description Nucleotide collection (nt)
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 GGTTTACAATTGTACCCCAAGGTATGGACTGGACCCCATCCATAGTAAGTCATCAGT 60
Sbjct 2003 GGTTTACAATTGTACCCCAAGGTATGGACTGGACCCCATCCATAGTAAGTCATCAGT 1944

Query 61 TTGCAATGATAAAGAAAATAACCTTCTGAAAATTGTATAGATCAGAAATAAAGTATTT 120
Sbjct 1943 TTGCAATGATAAAGAAAATAACCTTCTGAAAATTGTATAGATCAGAAATAAAGTATTT 1884

Query 121 TTTGTGGAAGACTATTTAAGTATTGAAGGTACTATTCCTTCTGAATTCATATTGC 180
Sbjct 1883 TTTGTGGAAGACTATTTAAGTATTGAAGGTACTATTCCTTCTGAATTCATATTGC 1824

Query 181 AGATGTACGGTGTGGATTATTGGTTATCTCTGCAAACCTTAAAGTAGAAGATGCAAG 240

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

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



2)



3)

