# Bioinformatika

http://bio.img.cas.cz

**Jan Pačes**
**Ústav molekulární genetiky**
*hpaces@img.cas.cz*


CZECH FOBIA

# Bioinformatics

## *statistics, alignments, phylogeny, clustering*

**Radek Zíka**
**Bioinformatics department**
**Institute of Molecular Genetics**


**e-mail: zikar@img.cas.cz**

# obsah

Biologické sekvence a jejich záznam
Zdroje na internetu
Alignment
Pairwise alignment
Substitution matrix
FastA
Blast
Psi-BLAST, Phi-BLAST, HMMER

# co je bioinformatika?

## Bioinformatics

From Wikipedia, the free encyclopedia

**Bioinformatics** and **computational biology** involve the use of techniques including applied mathematics, informatics, statistics, computer science, artificial intelligence, chemistry and biochemistry to solve biological problems usually on the molecular level. Research in computational biology often overlaps with systems biology. Major research efforts in the field include sequence alignment, gene finding, genome assembly, protein structure alignment, protein structure prediction, prediction of gene expression and protein-protein interactions, and the modeling of evolution.

- bomber
- bombers
- bonfire
- bonfires
- bumper
- bumpers
- bumpier

No definition found for "

You may wish to try an a

# bioinformatika

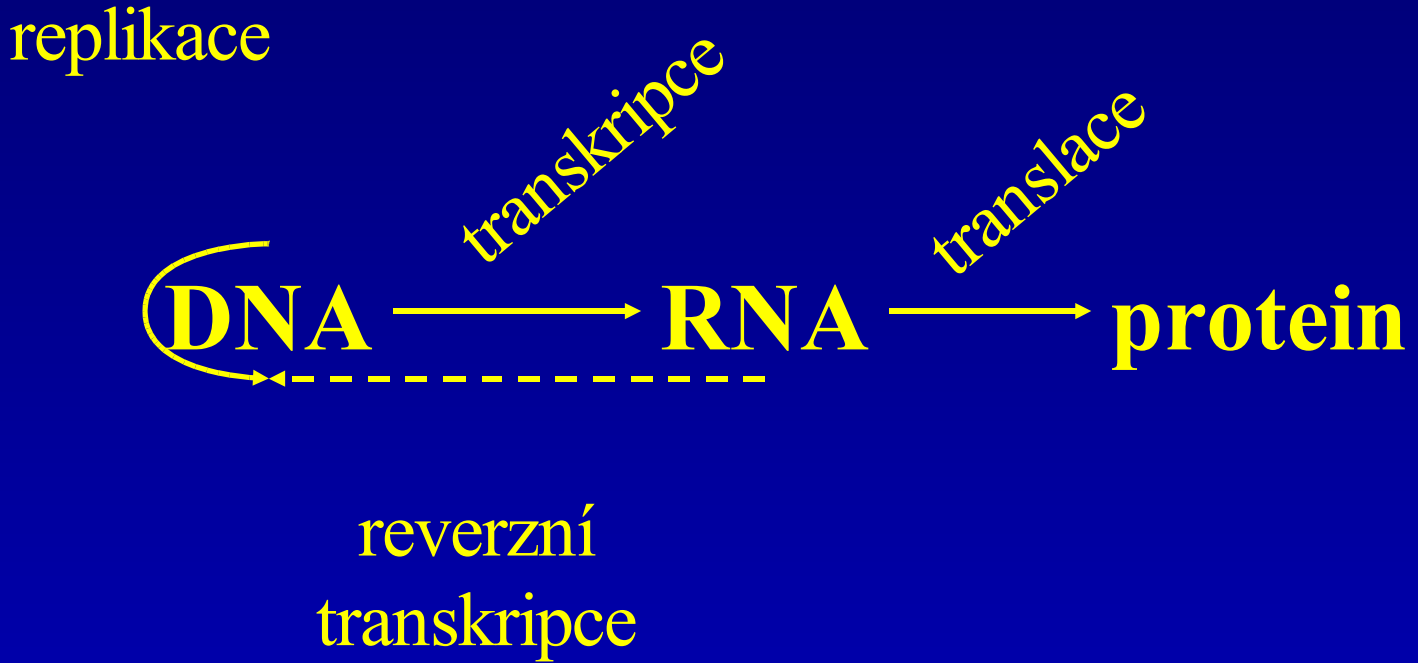
**Informatika nad biologickými molekulami (daty).**

**Bioinformatika extrahuje molekulární informační systém pro molekulární biologii.**

***Bioinformatika je konceptualizovaná molekulární biologie (ve smyslu fyzikálně chemickém) na níž je aplikována informatika (odvozená od matematické informatiky a statistiky).***

**Aplikace:** **teorie**
**biotechnologie**
**farmacie**
**medicína**
**genetické inženýrství**

# centrální dogma molekulární genetiky




**informace** → **funkce**

# velikosti genomů

*Mycoplasma genitalium* 0.58 Mbp

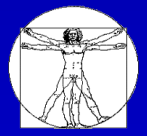
*Escherichia coli* 4.6 Mbp

*Caenorhabtitis elegans* 6 chr. 97 Mbp

*Saccharomyces cerevisiae* 16 chr. 11.2 Mbp

*Arabidopsis thaliana* 5 chr. 115.4 Mbp

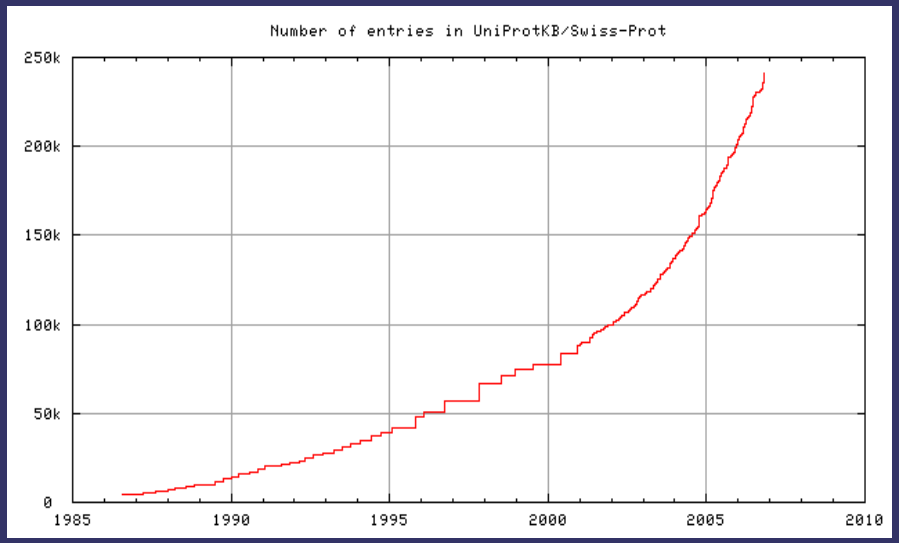
*Drosophila melanogaster* 5 chr. ~137.0 Mbp



*Homo sapiens* 24 chr. ~ 3.3 Gbp

# Internet, databases, www

- NAR web issue 2004: **137 webs**
  - Alignments
  - NA, protein sequence analysis
  - Structure analysis & description
  - Visualization




NAR DB collection 2004: **548 db**

1. Nucleotide seq
2. RNA seq
3. Protein sequence
4. Structure
5. Genomics (non-human)
6. Met. Enz. pathways, sig. pathways
7. Human and other vertebrate genomes

NAR DB collection 2005: **719 db**

1. Nucleotide seq
2. RNA seq
3. Protein sequence
4. Structure
5. Genomics (non-human)
6. Met. Enz. pathways, sig. pathways
7. Human and other vertebrate genomes
8. Imuno-polymorphism
9. Plants

# biologické sekvence a jejich záznam

**Způsoby záznamu sekvencí.**
**Jak uchovávat sekvence?**
**Odkud sekvence získat?**

# IUB kódy

## nukleotidy

| kód | nukleotidy | komplement |
|---|---|---|
| A | A | T |
| C | C | G |
| G | G | C |
| T | T | A |
| (U | U) | A |
| M | AC | K |
| R | AG | Y |
| W | AT | S |
| S | CG | W |
| Y | CT | R |
| K | GT | M |
| V | ACG | B |
| H | ACT | D |
| D | AGT | H |
| B | CGT | V |
| N | ACGT | N |
| - | mezera | - |

## aminokyseliny

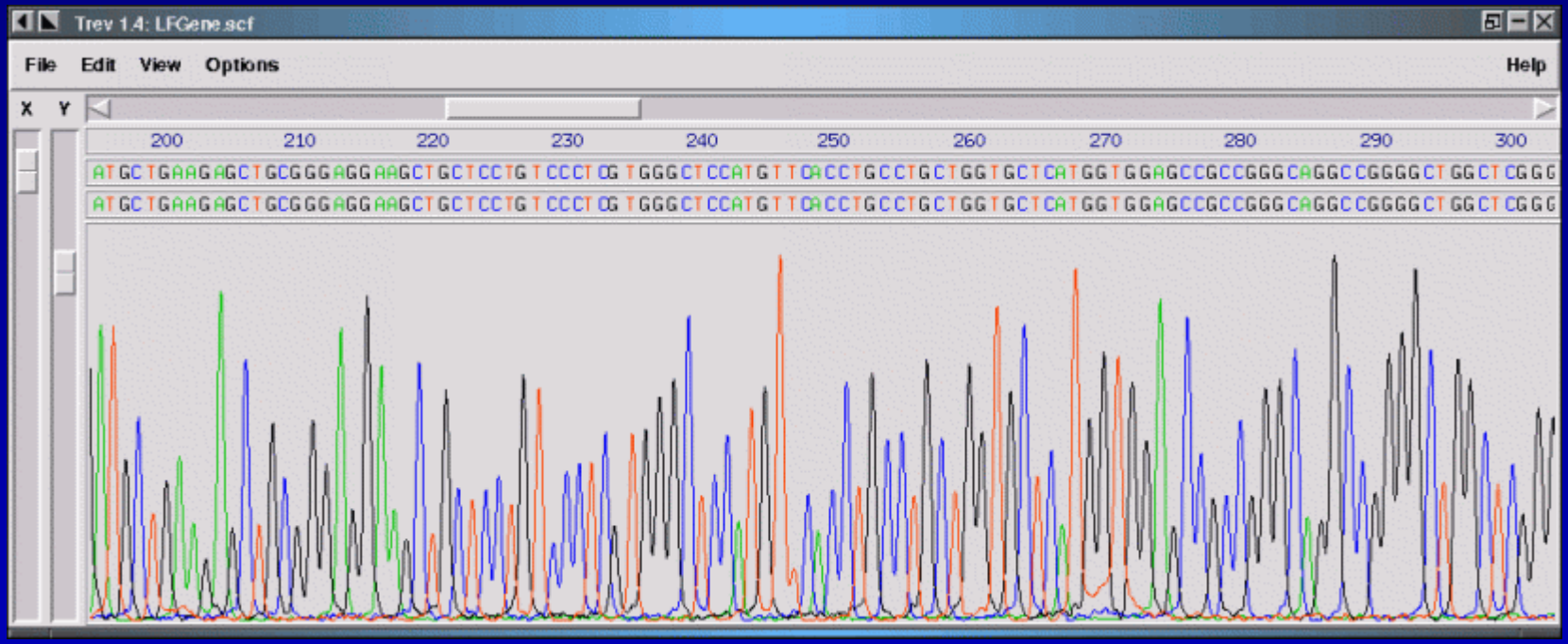
| kód | třípísmenný kód | aminokyselina |
|---|---|---|
| A | Ala | alanin |
| C | Cys | cystein |
| D | Asp | asparagová kyselina |
| G | Glu | glutamová kyselina |
| H | His | histidin |
| I | Ile | isoleucin |
| K | Lys | lysin |
| L | Leu | leucin |
| M | Met | methionin |
| N | Asn | asparagin |
| P | Pro | prolin |
| Q | Gln | glutamin |
| R | Arg | arginin |
| S | Ser | serin |
| T | Thr | threonin |
| V | Val | valin |
| W | Trp | tryptofan |
| Y | Tyr | tyrosin |
| B | Asx | asparagová kys. nebo asparagin |
| Z | Glx | glutamová kys. nebo glutamin |
| X | Xxx | jakákoliv aminokyselina |
| * | --- | stop |

# formáty sekvencí

| | | |
|---|---|---|
| binární | s chromatogramy | SCF<br>ALF<br>ABI |
| | pro programy | interní formáty databází |
| textové | minimální | text<br>fasta |
| | anotované | EMBL<br>GenBank<br>ASN<br>XML |

# formáty sekvencí - SCF

## SCF (standart chromatogram file)



# formáty sekvencí - FastA

## fasta

```
>gi|6102607|gb|AF145233.1|AF145233 Mus musculus transcription factor PAX4
TGGCAGGACTGAAGCAGCTGGAGGCTGTTACAAGACCAGACCACCAGCAAACCCTGGAGCCTGCACAGGA
CCCTGAGACCTCTTCCTGGAATTCCCACCTTTTTTCCTCCATCCAGAACCAGTCCCAAAGAGAAACTTCC
AGAAGGAGCTCTCCGTTTTCAGTTTGCCAGTTGGCTTCCTGTCCTTCTGTGAGGAGTACCAGTGTGAAGC
ATGCAGCAGGACGGACTCAGCAGTGTGAATCAGCTAGGGGGACTCTTTGTGAATGGCCGGCCCCTTCCTC
TGGACACCAGGCAGCAGATTGTGCAGCTAGCAATAAGAGGGATGCGACCCTGTGACATTTCACGGAGCCT
TAAGGTATCTAATGGCTGTGTGAGCAAGATCCTAGGACGCTACTACCGCACAGGTGTCTTGGAACCCAAG
TGTATTGGGGGAAGCAAACCACGTCTGGCCACACCTGCTGTGGTGGCTCGAATTGCCCAGCTAAAGGATG
AGTACCCTGCTCTTTTTGCCTGGGAGATCCAACACCAGCTTTGCACTGAAGGGCTTTGTACCCAGGACAA
GGCTCCCAGTGTGTCCTCTATCAATCGAGTACTTCGGGCACTTCAGGAAGACCAGAGCTTGCACTGGACT
CAACTCAGATCACCAGCTGTGTTGGCTCCAGTTCTTCCCAGTCCCCACAGTAACTGTGGGGCTCCCCGAG
GCCCCCACCCAGGAACCAGCCACAGGAATCGGACTATCTTCTCCCCGGGACAAGCCGAGGCACTGGAGAA
AGAGTTTCAGCGTGGGCAGTATCCAGATTCAGTGGCCCGTGGGAAGCTGGCTGCTGCCACCTCTCTGCCT
GAAGACACGGTGAGGGTTTGGTTTTCTAACAGAAGAGCCAAATGGCGCAGGCAAGAGAAGCTGAAATGGG
AAGCACAGCTGCCAGGTGCTTCCCAGGACCTGACAGTACCAAAAAATTCTCCAGGGATCATCTCTGCACA
GCAGTCCCCCGGCAGTGTACCCTCAGCTGCCTTGCCTGTGCTGGAACCATTGAGTCCTTCCTTCTGTCAG
CTATGCTGTGGGACAGCACCAGGCAGATGTTCCAGTGACACCTCATCCCAGGCCTATCTCCAACCCTACT
GGGACTGCCAATCCCTCCTTCCTGTGGCTTCCTCCTCATATGTGGAATTTGCCTGGCCCTGCCTCACCAC
CCATCCTGTGCATCATCTGATTGGAGGCCCAGGACAAGTGCCATCAACCCATTGCTCAAACTGGCCATAA
GAGGCCTCTATTTGACAGTAATAAAAACCTTTTCTTAGATGTTAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

# formáty sekvencí - EMBL

## EMBL (formát databáze EMBL)

```
ID   AF031150   standard; RNA; ROD; 1379 BP.
XX
AC   AF031150;
XX
SV   AF031150.1
XX
DT   27-FEB-1998 (Rel. 54, Created)
DT   27-FEB-1998 (Rel. 54, Last updated, Version 1)
XX
DE   Mus musculus paired-box transcription factor (Pax4) mRNA, complete cds.
XX
KW   .
XX
OS   Mus musculus (house mouse)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
XX
RN   [1]
RP   1-1379
RA   Inoue H., Nomiyama J., Nakai K., Matsutani A., Tanizawa Y., Oka Y.;
RT   Isolation of full-length cDNA of mouse PAX4 gene and identification of its
RT   human homologue;
RL   Biochem. Biophys. Res. Commun. 243:628-633(1998).
XX
FH   Key             Location/Qualifiers
…
```

# formáty sekvencí - EMBL

## EMBL (formát databáze EMBL)

```
…
FH   Key             Location/Qualifiers
FH
FT   source          1..1379
FT                   /db_xref=taxon:10090
FT                   /organism=Mus musculus
FT                   /cell_line=MIN6
FT   CDS             297..1346
FT                   /codon_start=1
FT                   /gene=Pax4
FT                   /product=paired-box transcription factor
FT                   /protein_id=AAC40046.1
FT                   /translation=MQQDGLSSVNQLGGLFVNGRPLPLDTRQQIVQLAIRGMRPCDISR
FT                   SLKVSNGCVSKILGRYYRTGVLEPKCIGGSKPRLATPAVVARIAQLKDEYPALFAWEIQ
…
FT                   PSTHCSNWP
XX
SQ   Sequence 1379 BP; 327 A; 402 C; 347 G; 303 T; 0 other;
     aaaaaaaaaa aaaaagcggc cgctgaattc tagcagaagg ctgccctctg ctcctgagtg        60
…
     gctgtgggac agcaccaggc agatgttcca gtgacacctc atcccaggcc tatctccaac      1200
     cctactggga ctgccaatcc ctccttcctg tggcttcctc ctcatatgtg gaatttgcct      1260
     ggccctgcct caccacccat cctgtgcatc atctgattgg aggcccagga caagtgccat      1320
     caacccattg ctcaaactgg ccataagagg cctctatttg acagtaataa aaacctttt       1379
//
```

# formáty sekvencí - GenBank

## Genbank

```
LOCUS       AF145233     1360 bp    mRNA            ROD       23-OCT-1999
DEFINITION  Mus musculus transcription factor PAX4 (Pax4) mRNA, complete cds.
ACCESSION   AF145233
VERSION     AF145233.1  GI:6102607
KEYWORDS    .
SOURCE      house mouse.
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 1360)
  AUTHORS   Kalousova,A., Benes,V., Paces,J., Paces,V. and Kozmik,Z.
  TITLE     DNA binding and transactivating properties of the paired and
            homeobox protein Pax4
  JOURNAL   Biochem. Biophys. Res. Commun. 259 (3), 510-518 (1999)
  MEDLINE   99294619
   PUBMED   10364449
FEATURES             Location/Qualifiers
     source          1..1360
                     /organism="Mus musculus"
                     /db_xref="taxon:10090"
     gene            1..1360
                     /gene="Pax4"
     CDS             211..1260
                     /gene="Pax4"
                     /note="DNA binding protein; paired box protein; homeobox
                     protein"
```

# formáty sekvencí - GenBank

## Genbank

```
CDS             211..1260
                     /gene="Pax4"
                     /note="DNA binding protein; paired box protein"
                     /codon_start=1
                     /product="transcription factor PAX4"
                     /protein_id="AAF03533.1"
                     /db_xref="GI:6102608"
                     /translation="MQQDGLSSVNQLGGLFVNGRPLPLDTRQQIVQLAIRGMRPCDIS
                     RSLKVSNGCVSKILGRYYRTGVLEPKCIGGSKPRLATPAVVARIAQLKDEYPALFAWE
                     IQHQLCTEGLCTQDKAPSVSSINRVLRALQEDQSLHWTQLRSPAVLAPVLPSPHSNCG
                     APRGPHPGTSHRNRTIFSPGQAEALEKEFQRGQYPDSVARGKLAAATSLPEDTVRVWF
                     SNRRAKWRRQEKLKWEAQLPGASQDLTVPKNSPGIISAQQSPGSVPSAALPVLEPLSP
                     SFCQLCCGTAPGRCSSDTSSQAYLQPYWDCQSLLPVASSSYVEFAWPCLTTHPVHHLI
                     GGPGQVPSTHCSNWP"
BASE COUNT      359 a    381 c    328 g    292 t
ORIGIN
        1 tggcaggact gaagcagctg gaggctgtta caagaccaga ccaccagcaa accctggagc
       61 ctgcacagga ccctgagacc tcttcctgga attcccacct tttttcctcc atccagaacc
      121 agtcccaaag agaaacttcc agaaggagct ctccgttttc agtttgccag ttggcttcct
      181 gtccttctgt gaggagtacc agtgtgaagc atgcagcagg acggactcag cagtgtgaat
…
     1201 catcatctga ttggaggccc aggacaagtg ccatcaaccc attgctcaaa ctggccataa
     1261 gaggcctcta tttgacagta ataaaaacct tttcttagat gttaaaaaaa aaaaaaaaaa
     1321 aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa
//
```

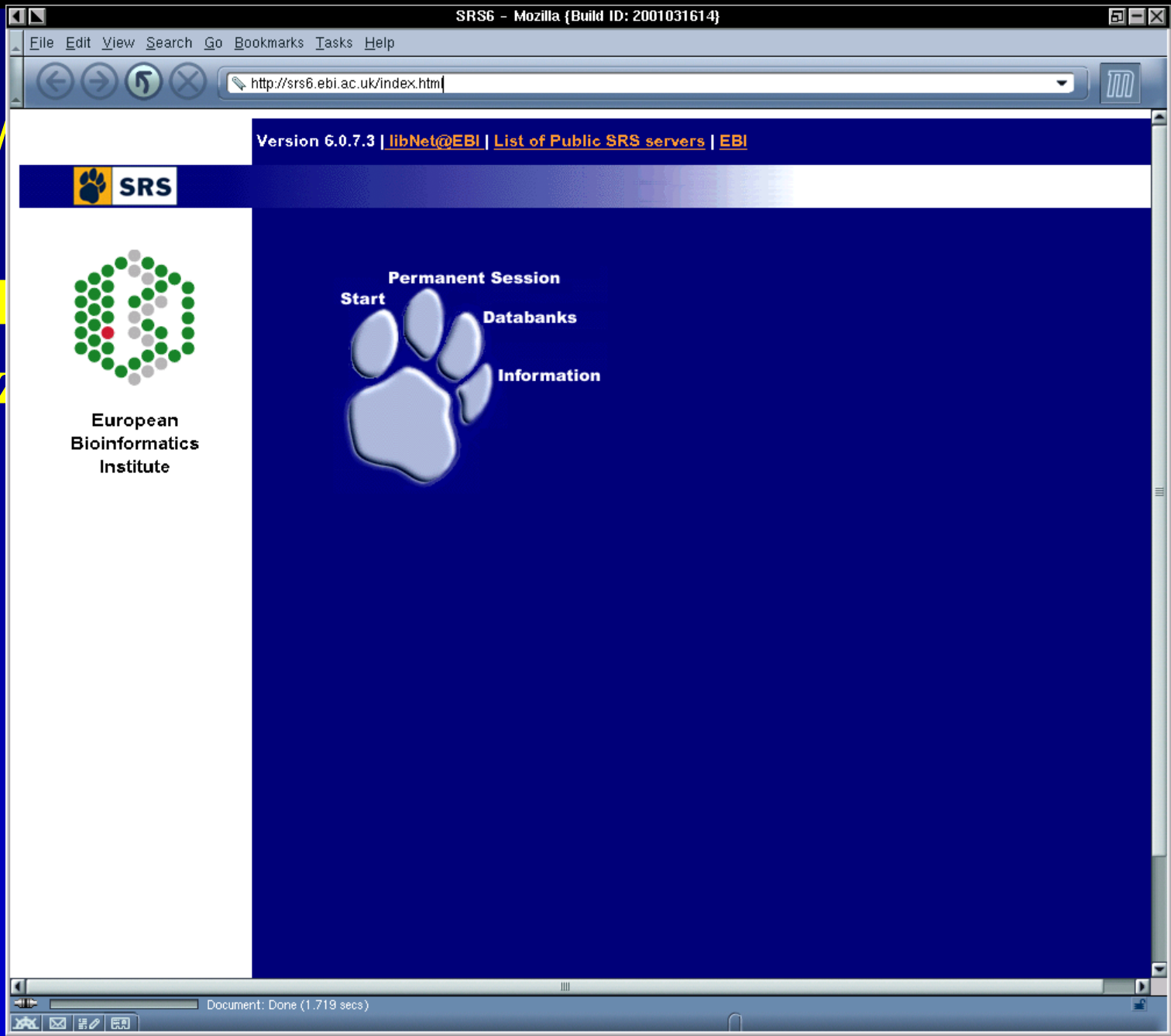
# zdroje na internetu

http://

Swissl

Entrez

SRS



# alignment

**Proč sekvence porovnáváme?**

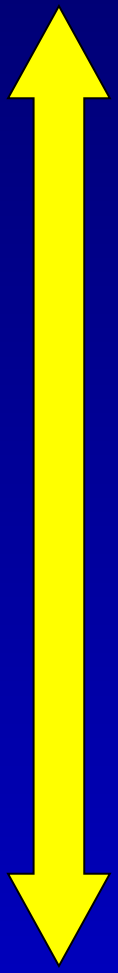
**Lokální vs. globální alignment.**

**Jaká je pravděpodobnost (statistická významnost) alignmentu.**

**Termíny:**

- **similarity (podobnost)**
- **homolog, paralog, ortholog**

# typy alignmentů



**1:1**

Pattern search
Dot plot

**1:n**

SSEARCH
BLITZ

**FASTA**
**BLAST**

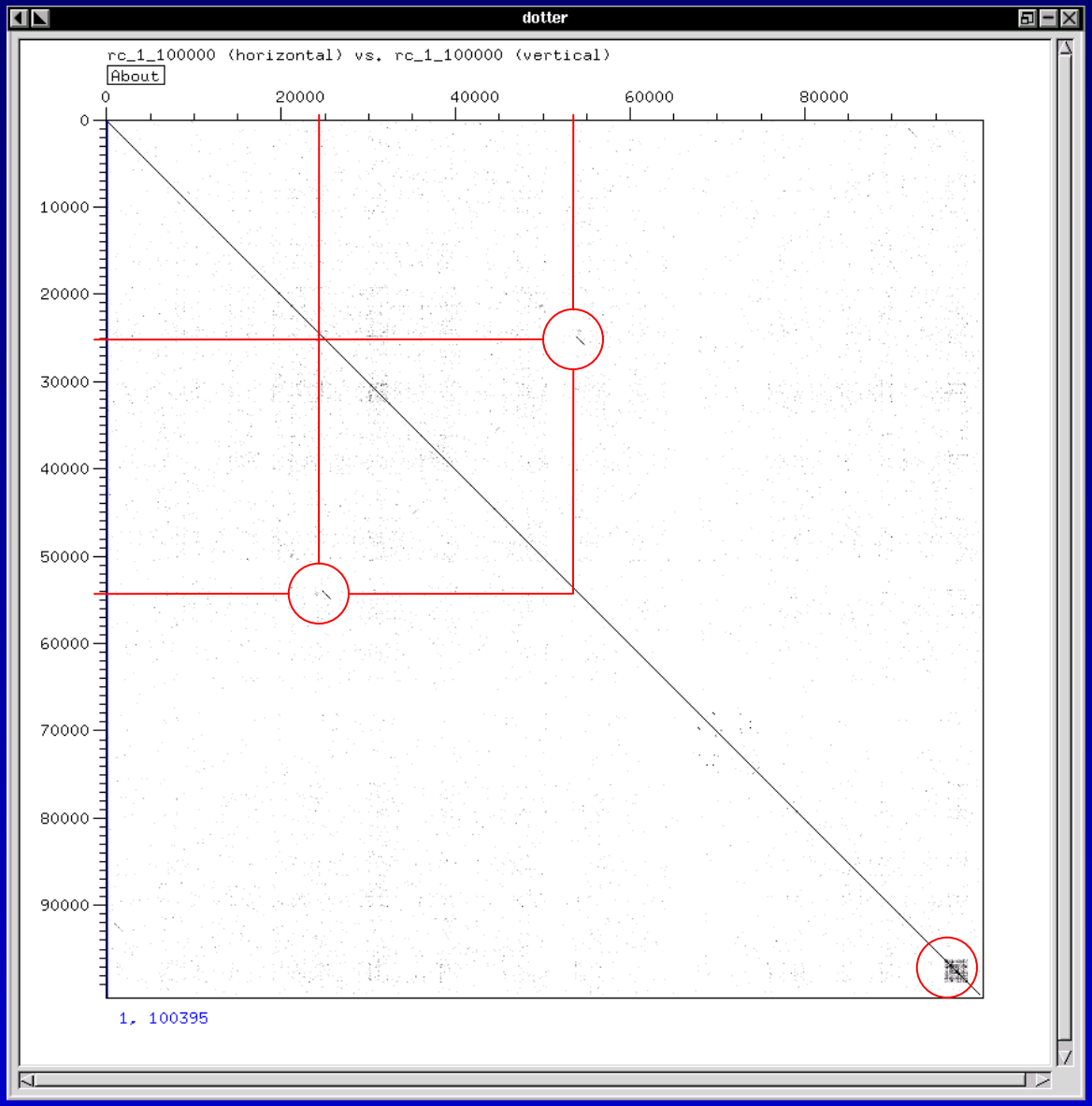
**n:n**

PSI-BLAST
HMMER

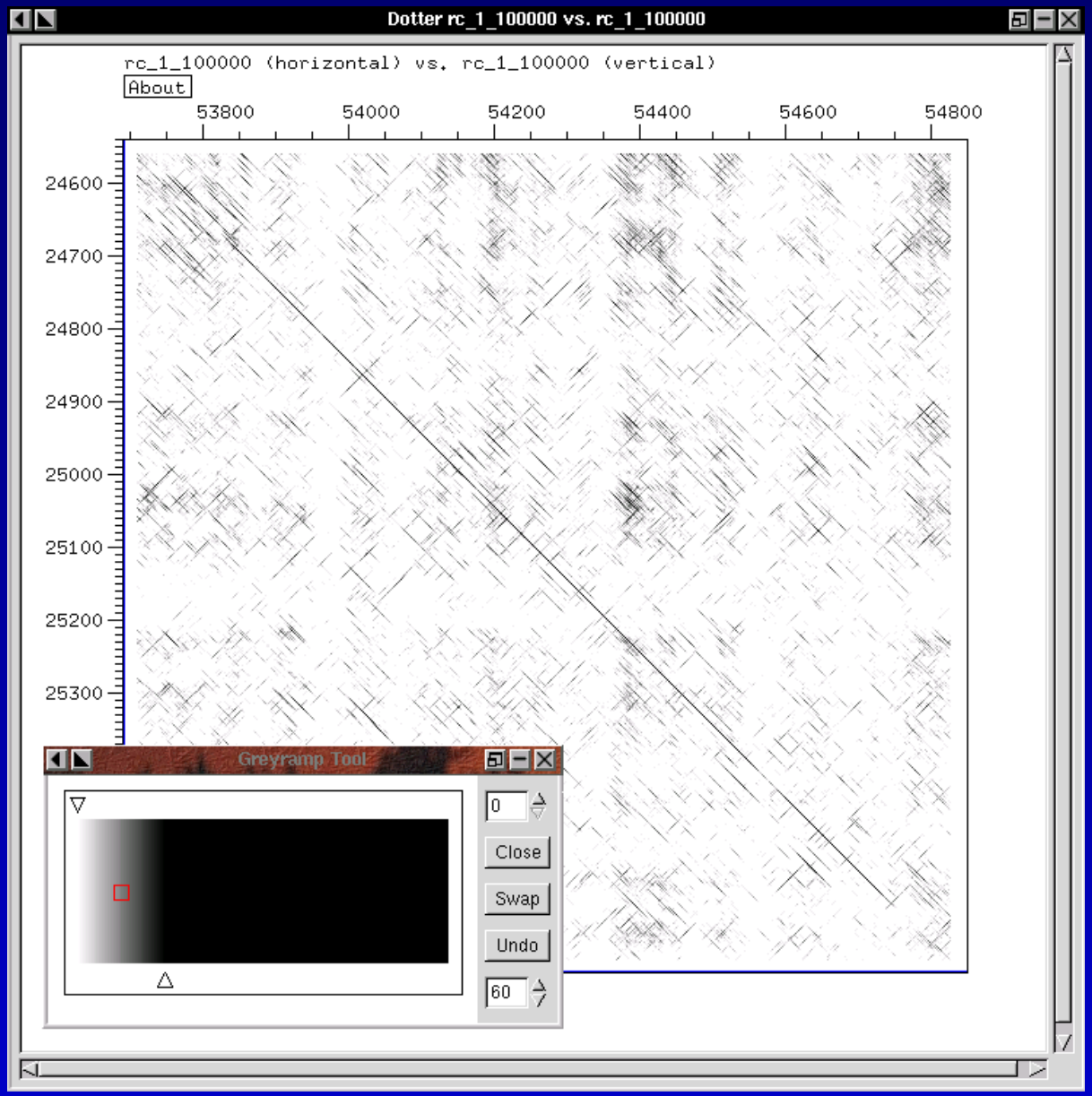
**n**

ClustalW
MultAlign
T-Coffee
Muscle

# dot plot

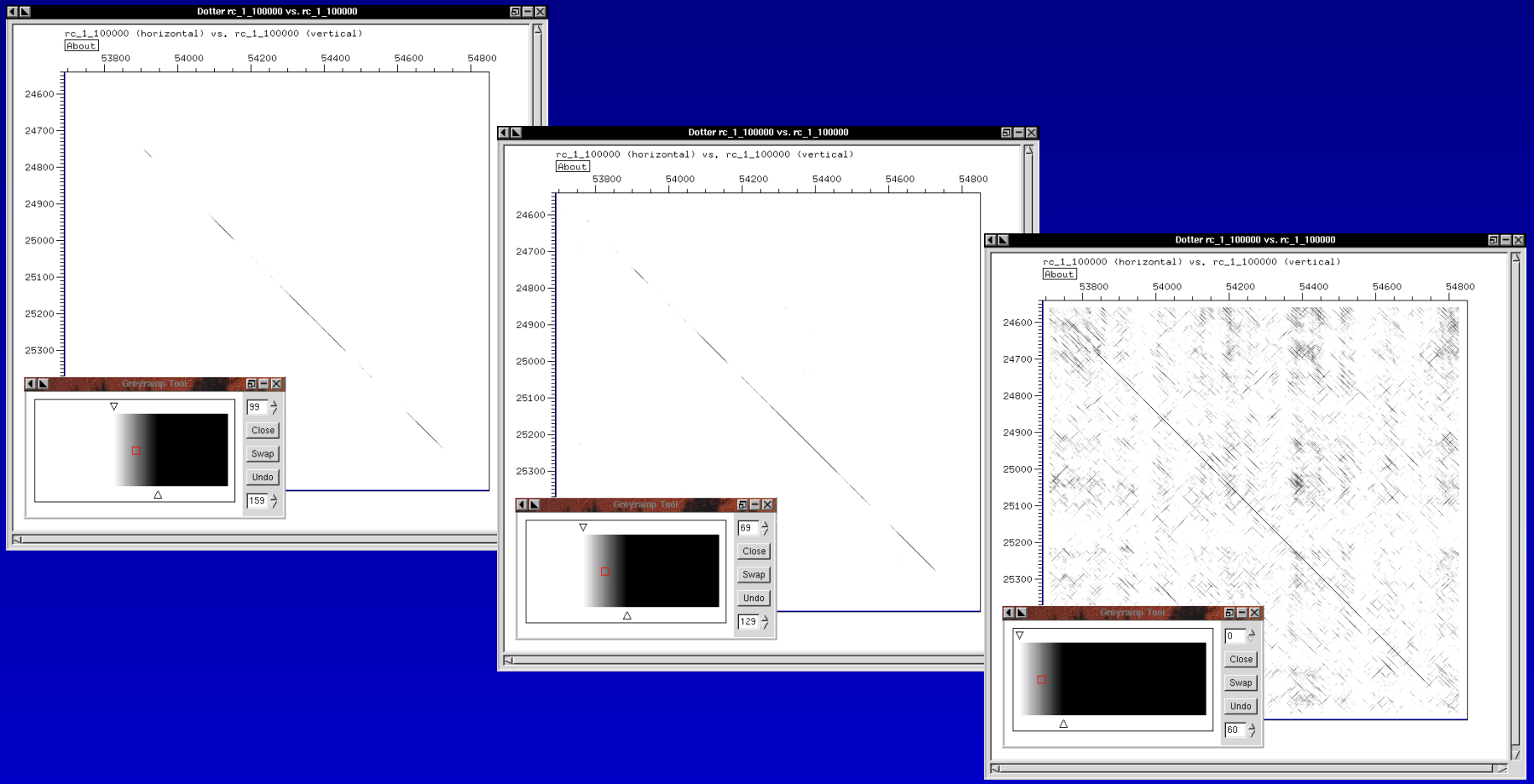

dotter
rc_1_100000 (horizontal) vs. rc_1_100000 (vertical)
About
0
20000
40000
60000
80000
0
10000
20000
30000
40000
50000
60000
70000
80000
90000
1, 100395

# dot plot

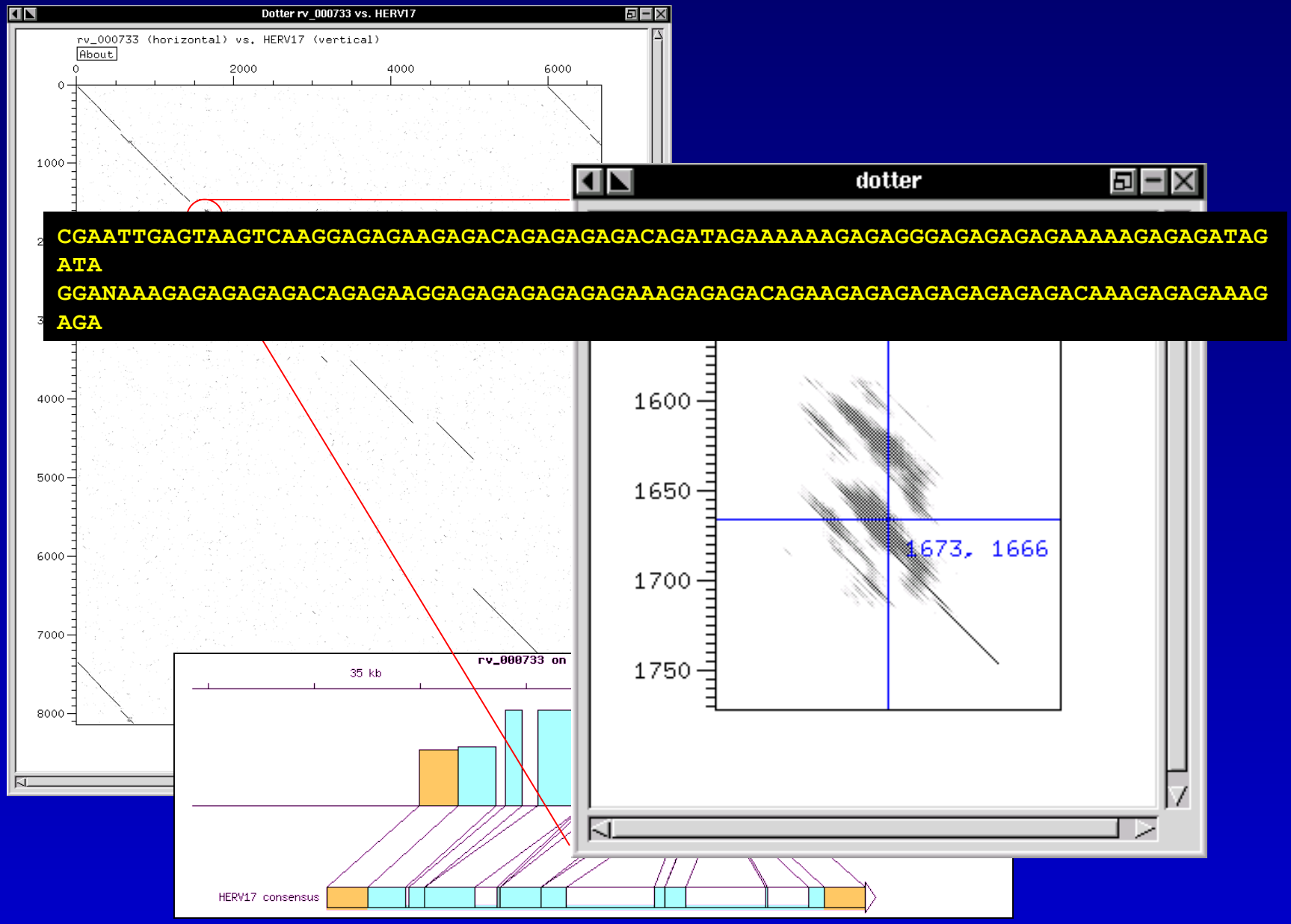

Dotter rc_1_100000 vs. rc_1_100000
rc_1_100000 (horizontal) vs. rc_1_100000 (vertical)
About
53800
54000
54200
54400
54600
54800
24600
24700
24800
24900
25000
25100
25200
25300
Greyramp Tool
0
Close
Swap
Undo
60

# dot plot

```
54268 CGGTCAGCCGGACCCGGACCACCACGGCAACGGGGCGCGATGTGGTGCGTCAGACCGCGCAGGCGATGGACGACATTGCCCGCAGCTCGGAACAGATCTCGCGCATCACCAGCGTCATCGACGAC
      ***** *** ******    ***** **  * ******** ******** *  **   **********  ****** * **  ******** ***** ** ********** **********
25111 CGGTCGGCCAGACCCGCGAGACCACCGCCGCCGGGCGCGAGGTGGTGCGCCGCACGCTGCAGGCGATGACCGACATCGCGCAAAGCTCGGAGCAGATTTCCCGCATCACCAGTGTCATCGACGAC
```



# dot plot


Dotter rv_000733 vs. HERV17
rv_000733 (horizontal) vs. HERV17 (vertical)
About
0
2000
4000
6000
1000
4000
5000
6000
7000
8000
dotter
CGAATTGAGTAAGTCAAGGAGAGAAGAGACAGAGAGAGACAGATAGAAAAAAGAGAGGGAGAGAGAGAAAAAGAGAGATAG
ATA
GGANAAAGAGAGAGAGACAGAGAAGGAGAGAGAGAGAGAAAGAGAGACAGAAGAGAGAGAGAGAGACAAAGAGAGAAAG
AGA
1600
1650
1700
1750
1673, 1666
rv_000733 on
35 kb
HERV17 consensus

# porovnávání textových řetězců

**data: TCATG a CATTG**

```
T C A T G          T C A - T G
      : :            : :   : :
C A T T G          . C A T T G
```

```
T C A T G .        T C A T - G
  : : :              : : :   :
. C A T T G        . C A T T G
```

# pairwise alignment

## formalizace problému

Vezměme dvě sekvence **a,b** (nukleotidové či aminokyselinové) délky **m,n:**

$$a = a_1, a_2, a_3 \; .. \; a_m$$
$$b = b_1, b_2, b_3 \; .. \; b_n$$

Chceme je porovnat mezi sebou a vytvořit alignment A, který sestává z řady párů

$$A = (a_i \; b_j) \; .. \; (a_k \; b_1) \quad \text{kde} \quad 1 \le i < .. < k \le m,$$
$$1 \le j < .. < l \le n$$

Pro výpočet skóre alignmentu A přiřadíme každému páru hodnotu $s(a_i, b_j)$ (pozitivní nebo negativní) v závislosti na tom, zda se jedná o totožný, příbuzný nebo nepříbuzný pár. Skóre subalignmentu $S_{i,j}$ získáme jako maximální skóre předcházejících subalignmentů plus skóre páru $s(a_i, b_j)$:

$$S_{i,j} = \max ( S_{i-1,j}, S_{i,j-1}, S_{i-1,j-1} ) + s_{i,j}$$

Celkové skóre alignmentu je tak

$$S = \Sigma \, s(a_i, b_j) \qquad \text{pro} \quad i = 1..m, \; j = 1..n$$

Hledání nejlepšího alignmentu je hledáním alignmentu s maximálním skóre ze všech možných alignmentů.

# pairwise alignment

## scoring matrix

| | G | G | A | C | T | C | T | T | G | G | A | A | A | G | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G | 1 | 1 | | | | | | | 1 | 1 | | | | 1 | 1 |
| G | 1 | 1 | | | | | | | 1 | 1 | | | | 1 | 1 |
| A | | | 1 | | | | | | | | 1 | 1 | 1 | | |
| C | | | | 1 | | 1 | | | | | | | | | |
| T | | | | | 1 | | 1 | 1 | | | | | | | |
| G | 1 | 1 | | | | | | | 1 | 1 | | | | 1 | 1 |
| G | 1 | 1 | | | | | | | 1 | 1 | | | | 1 | 1 |
| A | | | 1 | | | | | | | | 1 | 1 | 1 | | |
| A | | | 1 | | | | | | | | 1 | 1 | 1 | | |
| A | | | 1 | | | | | | | | 1 | 1 | 1 | | |
| G | 1 | 1 | | | | | | | 1 | 1 | | | | 1 | 1 |

## parametry: match 1; mismatch 0

# pairwise alignment

## sum matrix

| | G | G | A | C | T | C | T | T | G | G | A | A | A | G | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 4 | 4 | 4 | 4 | 5 | 6 |
| G | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 5 | 5 | 5 | 5 | 6 | 7 |
| A | 2 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 5 | 6 | 7 | 8 | 8 | 8 |
| C | 2 | 3 | 4 | 5 | 5 | 6 | 6 | 6 | 6 | 6 | 6 | 7 | 8 | 8 | 8 |
| T | 2 | 3 | 4 | 5 | 6 | 6 | 7 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |
| G | 3 | 4 | 4 | 5 | 6 | 6 | 7 | 8 | 9 | 10 | 10 | 10 | 10 | 11 | 12 |
| G | 4 | 5 | 5 | 5 | 6 | 6 | 7 | 8 | 10 | 11 | 11 | 11 | 11 | 12 | 13 |
| A | 4 | 5 | 6 | 6 | 6 | 6 | 7 | 8 | 10 | 11 | 12 | 13 | 14 | 14 | 14 |
| A | 4 | 5 | 7 | 7 | 7 | 7 | 7 | 8 | 10 | 11 | 13 | 14 | 15 | 15 | 15 |
| A | 4 | 5 | 8 | 8 | 8 | 8 | 8 | 8 | 10 | 11 | 14 | 15 | 16 | 16 | 16 |
| G | 5 | 6 | 8 | 8 | 8 | 8 | 8 | 8 | 11 | 12 | 14 | 15 | 16 | 17 | 18 |

# pairwise alignment

## zpětné hledání

| | G | G | A | C | T | C | T | T | G | G | A | A | A | G | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 4 | 4 | 4 | 4 | 5 | 6 |
| G | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 5 | 5 | 5 | 5 | 6 | 7 |
| A | 2 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 5 | 6 | 7 | 8 | 8 | 8 |
| C | 2 | 3 | 4 | 5 | 5 | 6 | 6 | 6 | 6 | 6 | 6 | 7 | 8 | 8 | 8 |
| T | 2 | 3 | 4 | 5 | 6 | 6 | 7 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |
| G | 3 | 4 | 4 | 5 | 6 | 6 | 7 | 8 | 9 | 10 | 10 | 10 | 10 | 11 | 12 |
| G | 4 | 5 | 5 | 5 | 6 | 6 | 7 | 8 | 10 | 11 | 11 | 11 | 11 | 12 | 13 |
| A | 4 | 5 | 6 | 6 | 6 | 6 | 7 | 8 | 10 | 11 | 12 | 13 | 14 | 14 | 14 |
| A | 4 | 5 | 7 | 7 | 7 | 7 | 7 | 8 | 10 | 11 | 13 | 14 | 15 | 15 | 15 |
| A | 4 | 5 | 8 | 8 | 8 | 8 | 8 | 8 | 10 | 11 | 14 | 15 | 16 | 16 | 16 |
| G | 5 | 6 | 8 | 8 | 8 | 8 | 8 | 8 | 11 | 12 | 14 | 15 | 16 | 17 | 18 |

```
GGACTCTTGGAAAGG
::::   :  ::::::
GGAC--T-GGAAAG-
```

# gaps - formalizace problému

Ohodnotíme mezery ("gaps") v alignmentu funkcí

$w_x = y + zx$ pro $x \geq 0;\ y,z \leq 0$

kde **x** je délka mezery ("gap"). Parametr **y** bývá nazýván "open gap penalty" nebo "gap existence penalty", parametr **z** "gap extension penalty" nebo "per residue gap penalty". Skóre subalignmentu $S_{i,j}$ získáme z:

$$S_{i,j} = \max \begin{cases} S_{k,j} - w_{i-k}, & \text{pro } k=1..i-1 \\ S_{i-1,j-1} & \\ S_{i,l} - w_{j-l}, & \text{pro } l=1..j-1 \end{cases} + s_{i,j}$$

# pairwise alignment

## scoring matrix

| | G | G | A | C | T | C | T | T | G | G | A | A | A | G | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G | 2 | 2 | -1 | -1 | -1 | -1 | -1 | -1 | 2 | 2 | -1 | -1 | -1 | 2 | 2 |
| G | 2 | 2 | -1 | -1 | -1 | -1 | -1 | -1 | 2 | 2 | -1 | -1 | -1 | 2 | 2 |
| A | -1 | -1 | 2 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | 2 | 2 | 2 | -1 | -1 |
| C | -1 | -1 | -1 | 2 | -1 | 2 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 |
| T | -1 | -1 | -1 | -1 | 2 | -1 | 2 | 2 | -1 | -1 | -1 | -1 | -1 | -1 | -1 |
| G | 2 | 2 | -1 | -1 | -1 | -1 | -1 | -1 | 2 | 2 | -1 | -1 | -1 | 2 | 2 |
| G | 2 | 2 | -1 | -1 | -1 | -1 | -1 | -1 | 2 | 2 | -1 | -1 | -1 | 2 | 2 |
| A | -1 | -1 | 2 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | 2 | 2 | 2 | -1 | -1 |
| A | -1 | -1 | 2 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | 2 | 2 | 2 | -1 | -1 |
| A | -1 | -1 | 2 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | 2 | 2 | 2 | -1 | -1 |
| G | 2 | 2 | -1 | -1 | -1 | -1 | -1 | -1 | 2 | 2 | -1 | -1 | -1 | 2 | 2 |

**parametry: match 2; mismatch -1**

# pairwise alignment

## sum matrix

| | G | G | A | C | T | C | T | T | G | G | A | A | A | G | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G | 2 | 4 | 3 | 2 | 1 | 0 | -1 | -2 | 0 | 2 | 1 | 0 | -1 | 1 | 3 |
| G | 4 | 4 | 3 | 2 | 1 | 0 | -1 | -2 | 0 | 2 | 1 | 0 | -1 | 1 | 3 |
| A | 3 | 3 | 6 | 3 | 1 | 0 | -1 | -2 | -3 | -1 | 4 | 4 | 4 | 1 | 0 |
| C | 2 | 2 | 3 | 8 | 5 | 5 | 2 | -1 | -3 | -4 | 1 | 3 | 3 | 3 | 0 |
| T | 1 | 1 | 1 | 5 | 10 | 7 | 8 | 7 | 4 | 1 | -2 | 0 | 2 | 2 | 2 |
| G | 3 | 3 | 0 | 2 | 7 | 9 | 6 | 6 | 9 | 9 | 6 | 3 | 0 | 4 | 4 |
| G | 5 | 5 | 2 | -1 | 4 | 6 | 8 | 5 | 9 | 11 | 8 | 5 | 2 | 4 | 6 |
| A | 4 | 4 | 7 | 4 | 1 | 3 | 5 | 7 | 6 | 8 | 13 | 13 | 13 | 10 | 7 |
| A | 3 | 3 | 7 | 6 | 3 | 0 | 2 | 4 | 6 | 5 | 13 | 15 | 15 | 12 | 9 |
| A | 2 | 2 | 7 | 6 | 5 | 2 | -1 | 1 | 3 | 5 | 13 | 15 | 17 | 14 | 11 |
| G | 4 | 4 | 4 | 6 | 5 | 4 | 1 | -2 | 3 | 5 | 10 | 12 | 14 | 19 | 19 |

**parametry: open gap -2**

**extended gap gap -2**

# pairwise alignment

## zpětné hledání

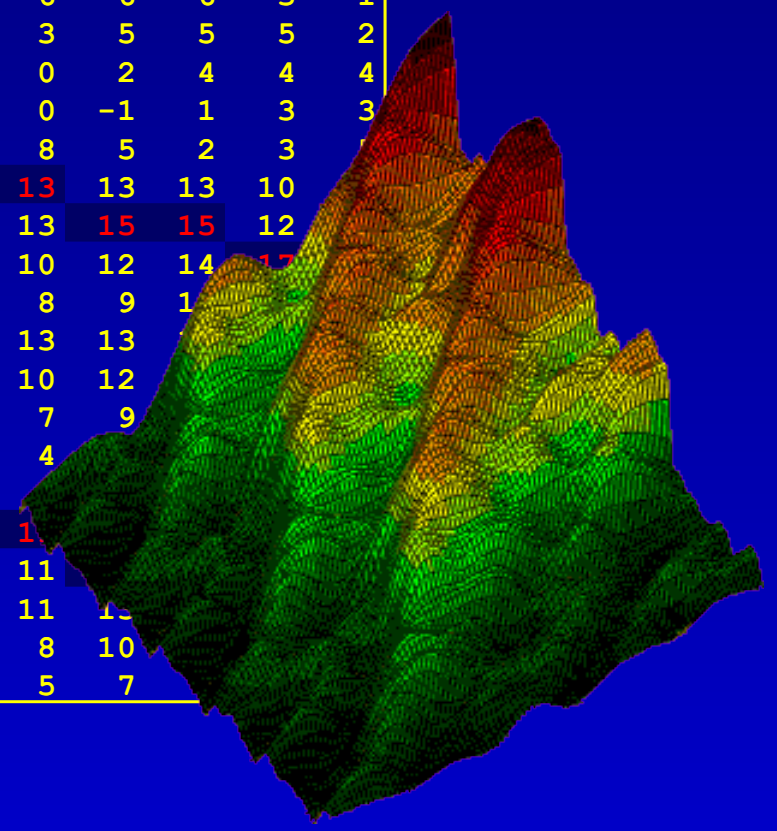
| | G | G | A | C | T | C | T | T | G | G | A | A | A | G | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G | 2 | 4 | 3 | 2 | 1 | 0 | -1 | -2 | 0 | 2 | 1 | 0 | -1 | 1 | 3 |
| G | 4 | 4 | 3 | 2 | 1 | 0 | -1 | -2 | 0 | 2 | 1 | 0 | -1 | 1 | 3 |
| A | 3 | 3 | 6 | 3 | 1 | 0 | -1 | -2 | -3 | -1 | 4 | 4 | 4 | 1 | 0 |
| C | 2 | 2 | 3 | 8 | 5 | 5 | 2 | -1 | -3 | -4 | 1 | 3 | 3 | 3 | 0 |
| T | 1 | 1 | 1 | 5 | 10 | 7 | 7 | 7 | 4 | 1 | -2 | 0 | 2 | 2 | 2 |
| G | 3 | 3 | 0 | 2 | 7 | 9 | 6 | 6 | 9 | 9 | 6 | 3 | 0 | 4 | 4 |
| G | 5 | 5 | 2 | -1 | 4 | 6 | 8 | 5 | 9 | 11 | 8 | 5 | 2 | 4 | 6 |
| A | 4 | 4 | 7 | 4 | 1 | 3 | 5 | 7 | 6 | 8 | 13 | 13 | 13 | 10 | 7 |
| A | 3 | 3 | 7 | 6 | 3 | 0 | 2 | 4 | 6 | 5 | 13 | 15 | 15 | 12 | 9 |
| A | 2 | 2 | 7 | 6 | 5 | 2 | -1 | 1 | 3 | 5 | 13 | 15 | 17 | 14 | 11 |
| G | 4 | 4 | 4 | 6 | 5 | 4 | 1 | -2 | 3 | 5 | 10 | 12 | 14 | 19 | 19 |

```
GGACTCTTGGAAAGG
:::::     :::::::
GGACT---GGAAAG-
```

# pairwise alignment - možné problémy

## parametry: match: 2; mismatch: -1; gap: -2

| | | G | G | A | C | T | C | T | T | G | G | A | A | A | G | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 2 | 2 | -1 | -1 | -1 | -1 | -1 | -1 | 2 | 2 | -1 | -1 | -1 | 2 | 2 |
| G | | 2 | 4 | 1 | -2 | -2 | -2 | -2 | -2 | 2 | 4 | 1 | -2 | -2 | 2 | 4 |
| G | | 2 | 4 | 3 | 0 | -3 | -3 | -3 | -3 | 2 | 4 | 3 | 0 | -3 | 2 | 4 |
| A | | -1 | 1 | 6 | 3 | 0 | -3 | -4 | -4 | -1 | 1 | 6 | 6 | 6 | 3 | 1 |
| C | | -1 | -2 | 3 | 8 | 5 | 5 | 2 | -1 | -4 | -2 | 3 | 5 | 5 | 5 | 2 |
| T | | -1 | -2 | 0 | 5 | 10 | 7 | 7 | 7 | 4 | 1 | 0 | 2 | 4 | 4 | 4 |
| T | | -1 | -2 | -3 | 2 | 10 | 9 | 9 | 9 | 6 | 3 | 0 | -1 | 1 | 3 | 3 |
| G | | 2 | 2 | -1 | -1 | 7 | 9 | 8 | 8 | 11 | 11 | 8 | 5 | 2 | 3 | |
| A | | -1 | 1 | 4 | 1 | 4 | 6 | 8 | 7 | 8 | 10 | 13 | 13 | 13 | 10 | |
| A | | -1 | -2 | 4 | 3 | 1 | 3 | 5 | 7 | 6 | 7 | 13 | 15 | 15 | 12 | |
| G | | 2 | 2 | 1 | 3 | 2 | 0 | 2 | 4 | 9 | 9 | 10 | 12 | 14 | | |
| G | | 2 | 4 | 1 | 0 | 2 | 1 | -1 | 1 | 9 | 11 | 8 | 9 | | | |
| A | | -1 | 1 | 6 | 3 | 0 | 1 | 0 | -2 | 6 | 8 | 13 | 13 | | | |
| T | | -1 | -2 | 3 | 5 | 5 | 2 | 3 | 3 | 3 | 5 | 10 | 12 | | | |
| T | | -1 | -2 | 0 | 2 | 7 | 4 | 4 | 5 | 2 | 2 | 7 | 9 | | | |
| G | | 2 | 2 | -1 | -1 | 4 | 6 | 3 | 3 | 7 | 7 | 4 | | | | |
| G | | 2 | 4 | 1 | -2 | 1 | 3 | 5 | 2 | 7 | 9 | | | | | |
| A | | -1 | 1 | 6 | 3 | 0 | 0 | 2 | 4 | 4 | 6 | | | | | |
| A | | -1 | -2 | 6 | 5 | 2 | -1 | -1 | 1 | 3 | 3 | 11 | | | | |
| A | | -1 | -2 | 6 | 5 | 4 | 1 | -2 | -2 | 0 | 2 | 11 | | | | |
| G | | 2 | 2 | 3 | 5 | 4 | 3 | 0 | -3 | 0 | 2 | 8 | 10 | | | |
| G | | 2 | 4 | 1 | 2 | 4 | 3 | 2 | -1 | 0 | 2 | 5 | 7 | | | |



# pairwise alignment - možné problémy

```
jana.- ---TGCCG-TTG--GAATCGACCCCGATCGCCGTCTCGACCACGTAGCTCATGCGGTTGCAGAGCTTGAGGTCGGCGTGGGCGGCGGAGGTGAGGGCGAG
           ::: :::  :  :::::: :: : ::     ::    :  ::::  :: ::: :    :: ::::  : :::  :: :  ::::     :: ::
Contig CCACCCCGATTGCCGTCTCGACCACG-TAGCTCATGCGGTTGCACAGCT--TGAGGTCG----GCGTGAG--CCGCGGAGGTGAGGGAGAGAGTGGTGAC
        44680     44690     44700     44710     44720     44730          44740     44750     44760

jana.- GGTGGTGACAATGGCG-GCGAGGGC--GCGCGATCTCCACCCCGTCATGCCCGG---CCTTGTGCCGGGCATCCACGTCTTGCTGAGAAAACCGCCGGAA
         ::::: : : :::: : ::   :  :: ::   : : :  : :  :::: ::   ::  ::   :   :::::::::::::::: ::::::::::::
Contig AATGGTGGCGAGGGCGCGTGATCTCCAGCCCGTCATGCCCGGCCTTGTGCCGGGCATCCACGTCTTGCTGATCCACGTCTTGCTGAGGAAACCGCCGGAA
       44770     44780     44790     44800     44810     44820     44830     44840     44850     44860

jana.- AGACGTGGATGGCCGGGACGAGCCCGGCCATGACGGATGTTGTCGCCGCAGCCTCGCATCACTTGTGGATCAGCGTGCCGGTGCCCTGGTTGGTGAACAA
       :::::::::::::::::::::::::::::::::::::::: ::::::: ::: ::::::::::::::::::::::::::::::::::::::::::::
Contig AGACGTGGATGGCCGGGACGAGCCCGGCCATGACGGATGTTATCGCCGCCGCCCCGCATCACTTGTGGATCAGCGTGCCGGTGCCCTGGTTGGTGAACAG
       44870     44880     44890     44900     44910     44920     44930     44940     44950     44960
```

```
jana.s CTCATGCGGTTGCAGAGCTTGAGGTCGGCGTGGGCGGCGGAGGTGAGGGCGAGGGTGGTGACAATGGCGGCGAGGGCGCGCGATCTCCACCCCGTCATGC
       :::::::::::::: :::::::::::::::: :: :::::::::::: ::: :::::::::::: :::::::::::: :::::::: ::::::::::
Contig CTCATGCGGTTGCACAGCTTGAGGTCGGCGTGAGCCGCGGAGGTGAGGGAGAGAGTGGTGACAATGGTGGCGAGGGCGCGTGATCTCCAGCCCGTCATGC
         44710     44720     44730     44740     44750     44760     44770     44780     44790     44800

jana.s CCGGCCTTGTGCCGGGCAT---------------CCACGTCTTGCTGAGAAAACCGCCGGAAAGACGTGGATGGCCGGGACGAGCCCGGCCATGACGGAT
       ::::::::::::::::::               :::::::::::::: :::::::::::::::::::::::::::::::::::::::::::::::::::
Contig CCGGCCTTGTGCCGGGCATCCACGTCTTGCTGATCCACGTCTTGCTGAGGAAACCGCCGGAAAGACGTGGATGGCCGGGACGAGCCCGGCCATGACGGAT
         44810     44820     44830     44840     44850     44860     44870     44880     44890     44900

jana.s GTTGTCGCCGCAGCCTCGCATCACTTGTGGATCAGCGTGCCGGTGCCCTGGTTGGTGAACAATTCGAGCAGCACTGCGTGCGGGACCTTGCCGTCGAGGA
       ::: ::::::: ::: :::::::::::::::::::::::::::::::::::::::: :::::::::::::::::::::::::::::::::::::::
Contig GTTATCGCCGCCGCCCCGCATCACTTGTGGATCAGCGTGCCGGTGCCCTGGTTGGTGAACAGTTCGAGCAGCACTGCGTGCGGGACCTTGCCGTCGAGGA
         44910     44920     44930     44940     44950     44960     44970     44980     44990     45000
```

# genetický kód

| | T | | C | | A | | G | | |
|---|---|---|---|---|---|---|---|---|---|
| T | TTT | Phe | TCT | Ser | TAT | Tyr | TGT | Cys | T |
| | TTC | Phe | TCC | Ser | TAC | Tyr | TGC | Cys | C |
| | TTA | Leu | TCA | Ser | TAA | Stop | TGA | Stop | A |
| | TTG | Leu | TCG | Ser | TAG | Stop | TGG | Trp | G |
| C | CTT | Leu | CCT | Pro | CAT | His | CGT | Arg | T |
| | CTC | Leu | CCC | Pro | CAC | His | CGC | Arg | C |
| | CTA | Leu | CCA | Pro | CAA | Gln | CGA | Arg | A |
| | CTG | Leu | CCG | Pro | CAG | Gln | CGG | Arg | G |
| A | ATT | Ile | ACT | Thr | AAT | Asn | AGT | Ser | T |
| | ATC | Ile | ACC | Thr | AAC | Asn | AGC | Ser | C |
| | ATA | Ile | ACA | Thr | AAA | Lys | AGA | Arg | A |
| | ATG | Met | ACG | Thr | AAG | Lys | AGG | Arg | G |
| G | GTT | Val | GCT | Ala | GAT | Asp | GGT | Gly | T |
| | GTC | Val | GCC | Ala | GAC | Asp | GGC | Gly | C |
| | GTA | Val | GCA | Ala | GAA | Glu | GGA | Gly | A |
| | GTG | Val | GCG | Ala | GAG | Glu | GGG | Gly | G |

# substitution matrix

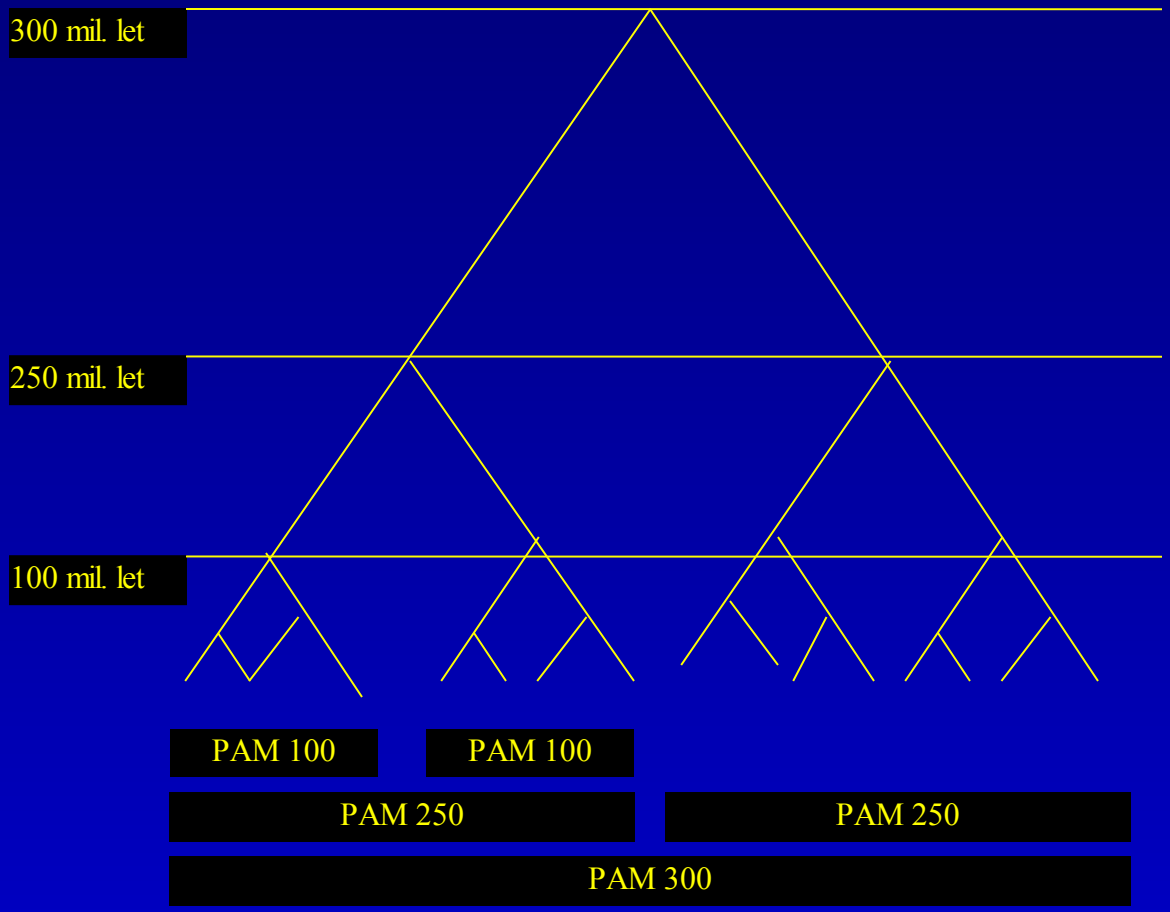
**Princip: proteiny se vyvíjejí pomocí nezávislých mutací a jsou fixovány postupně**

## PAM (Percent Accepted Mutation)

**1 PAM = jedna mutace na cestě mezi dvěma sekvencemi na 100 nukleotidů**

## BLOSSUM

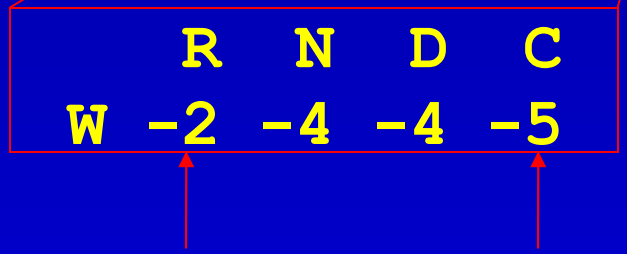
# substitution matrix



# substitution matrix

## BLOSUM 45

```
#  Entropy =   0.3795, Expected =  -0.2789
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  5 -2 -1 -2 -1 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -2 -2  0 -1 -1  0 -5
R -2  7  0 -1 -3  1  0 -2  0 -3 -2  3 -1 -2 -2 -1 -1 -2 -1 -2 -1  0 -1 -5
N -1  0  6  2 -2  0  0  0  1 -2 -3  0 -2 -2 -2  1  0 -4 -2 -3  4  0 -1 -5
D -2 -1  2  7 -3  0  2 -1  0 -4 -3  0 -3 -4 -1  0 -1 -4 -2 -3  5  1 -1 -5
C -1 -3 -2 -3 12 -3 -3 -3 -3 -3 -2 -3 -2 -2 -4 -1 -1 -5 -3 -1 -2 -3 -2 -5
Q -1  1  0  0 -3  6  2 -2  1 -2 -2  1  0 -4 -1  0 -1 -2 -1 -3  0  4 -1 -5
E -1  0  0  2 -3  2  6 -2  0 -3 -2  1 -2 -3  0  0 -1 -3 -2 -3  1  4 -1 -5
G  0 -2  0 -1 -3 -2 -2  7 -2 -4 -3 -2 -2 -3 -2  0 -2 -2 -3 -3 -1 -2 -1 -5
H -2  0  1  0 -3  1  0 -2 10 -3 -2 -1  0 -2 -2 -1 -2 -3  2 -3  0  0 -1 -5
I -1 -3 -2 -4 -3 -2 -3 -4 -3  5  2 -3  2  0 -2 -2 -1 -2  0  3 -3 -3 -1 -5
L -1 -2 -3 -3 -2 -2 -2 -3 -2  2  5 -3  2  1 -3 -3 -1 -2  0  1 -3 -2 -1 -5
K -1  3  0  0 -3  1  1 -2 -1 -3 -3  5 -1 -3 -1 -1 -1 -2 -1 -2  0  1 -1 -5
M -1 -1 -2 -3 -2  0 -2 -2  0  2  2 -1  6  0 -2 -2 -1 -2  0  1 -2 -1 -1 -5
F -2 -2 -2 -4 -2 -4 -3 -3 -2  0  1 -3  0  8 -3 -2 -1  1  3  0 -3 -3 -1 -5
P -1 -2 -2 -1 -4 -1  0 -2 -2 -2 -3 -1 -2 -3  9 -1 -1 -3 -3 -3 -2 -1 -1 -5
S  1 -1  1  0 -1  0  0  0 -1 -2 -3 -1 -2 -2 -1  4  2 -4 -2 -1  0  0  0 -5
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -1 -1  2  5 -3 -1  0  0 -1  0 -5
W -2 -2 -4 -4 -5 -2 -3 -2 -3 -2 -2 -2 -2  1 -3 -4 -3 15  3 -3 -4 -2 -2 -5
Y -2 -1 -2 -2 -3 -1 -2 -3  2  0  0 -1  0  3 -3 -2 -1  3  8 -1 -2 -2 -1 -5
V  0 -2 -3 -3 -1 -3 -3 -3 -3  3  1 -2  1  0 -3 -1  0 -3 -1  5 -3 -3 -1 -5
B -1 -1  4  5 -2  0  1 -1  0 -3 -3  0 -2 -3 -2  0  0 -4 -2 -3  4  2 -1 -5
Z -1  0  0  1 -3  4  4 -2  0 -3 -2  1 -1 -3 -1  0 -1 -2 -2 -3  2  4 -1 -5
X  0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1  0  0 -2 -1 -1 -1 -1 -1 -5
* -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5  1
```



# substitution matrix

## BLOSUM 62

```
#  Entropy =   0.6979, Expected =  -0.5209
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1  0 -4
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1  0 -1 -4
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  3  0 -1 -4
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4  1 -1 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0  3 -1 -4
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3 -1 -2 -1 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3  0  0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1 -4 -3 -1 -4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2  0  1 -1 -4
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2  0  0  0 -4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0 -1 -1  0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1 -3 -2 -1 -4
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4 -3 -2 -1 -4
B -2 -1  3  4 -3  0  1 -1  0 -3 -4  0 -3 -3 -2  0 -1 -4 -3 -3  4  1 -1 -4
Z -1  0  0  1 -3  3  4 -2  0 -3 -3  1 -1 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
X  0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -2  0  0 -2 -1 -1 -1 -1 -1 -4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4  1
```

# substitution matrix

| W | R | N | D | C | W |
|---|---|---|---|---|---|
| **PAM 50** | -1 | -7 | -12 | -13 | 13 |
| **PAM 100** | 1 | -5 | -9 | -9 | 12 |
| **PAM 250** | 2 | -4 | -7 | -8 | 17 |
| **BLOSUM 100** | -7 | -8 | -10 | -7 | 17 |
| **BLOSUM 62** | -3 | -4 | -4 | -2 | 11 |
| **BLOSUM 30** | 0 | -7 | -4 | -2 | 20 |

# global vs. local alignment

**Globální:**

- Porovnáváme kompletní geny (proteiny) - zajímá nás, do jaké míry si jsou příbuzné.
- Přítomnost nehomologních párů je neutrální, aby nebylo ovlivněno celkové skóre.
- aka Needleman-Wunsch.

**Lokální:**

- Hledáme podobné oblasti uvnitř delších sekvencí (domény) - zajímá nás, jestli obsahují konzervované úseky.
- Negativní skóre pro nehomologní páry (se vzdáleností od domény skóre klesá).
- Nejvyšší skóre nehledáme pouze v posledním sloupci/řádku, ale v celé sum matrix. Postupujeme na obě strany k nule.
- aka Smitch-Waterman.

# optimalizace pro hledání v databázích

## Efektivita hledání je řádu $N^2*L$

(N je délka prohledávající sekvence, L velikost prohledávané databáze.)

GenBank (April 2006): ~130 000 000 000 nt

Swiss-prot (Rel. 51.0; 31-Oct-06): 88 541 632 aa

## Zlepšení:

**Výchozí úvaha:** oblasti, které si jsou podobné, budou pravděpodobně obsahovat krátké identické úseky.

**Hledáme:**

- Oblasti, kde následuje několik identických "slov" (words) ve stejném pořadí za sebou.
- Použijeme předpočítanou tabulku výskytu běžných "slov" v databázi - **hashing**. Výpočet tabulky je řádu L (velikost databáze), ale použití pouze řádu N (délka prohledávané sekvence).
- Nalezený úsek s okolím použijeme pro přesný alignment.

# FastA - princip

## FastA - Fast Algorhitm

1. Najdeme diagonály krátkých identických sekvencí.
2. Získáme alignment a spočteme jeho skóre bez mezer (initn).
3. Jednotlivé části spojíme a získáme neoptimalizovaný alignment, do skóre započítáme i gaps (init1).
4. Prodloužíme alignment na obě strany a použitím "pairwise" algoritmu získáme optimalizovaný alignment (opt).
5. Spočteme z-skóre (bit-skóre) a expectancy

# FastA - použití

**zdrojový kód:** **ftp://ftp.virginia.edu/pub/fasta**

(Zdrojový kód pro akademické použití volný, kompilace pod UNIXy bez problémů, lze kompilovat i pod windows.

**www:** **http://www.ebi.ac.uk/fasta**

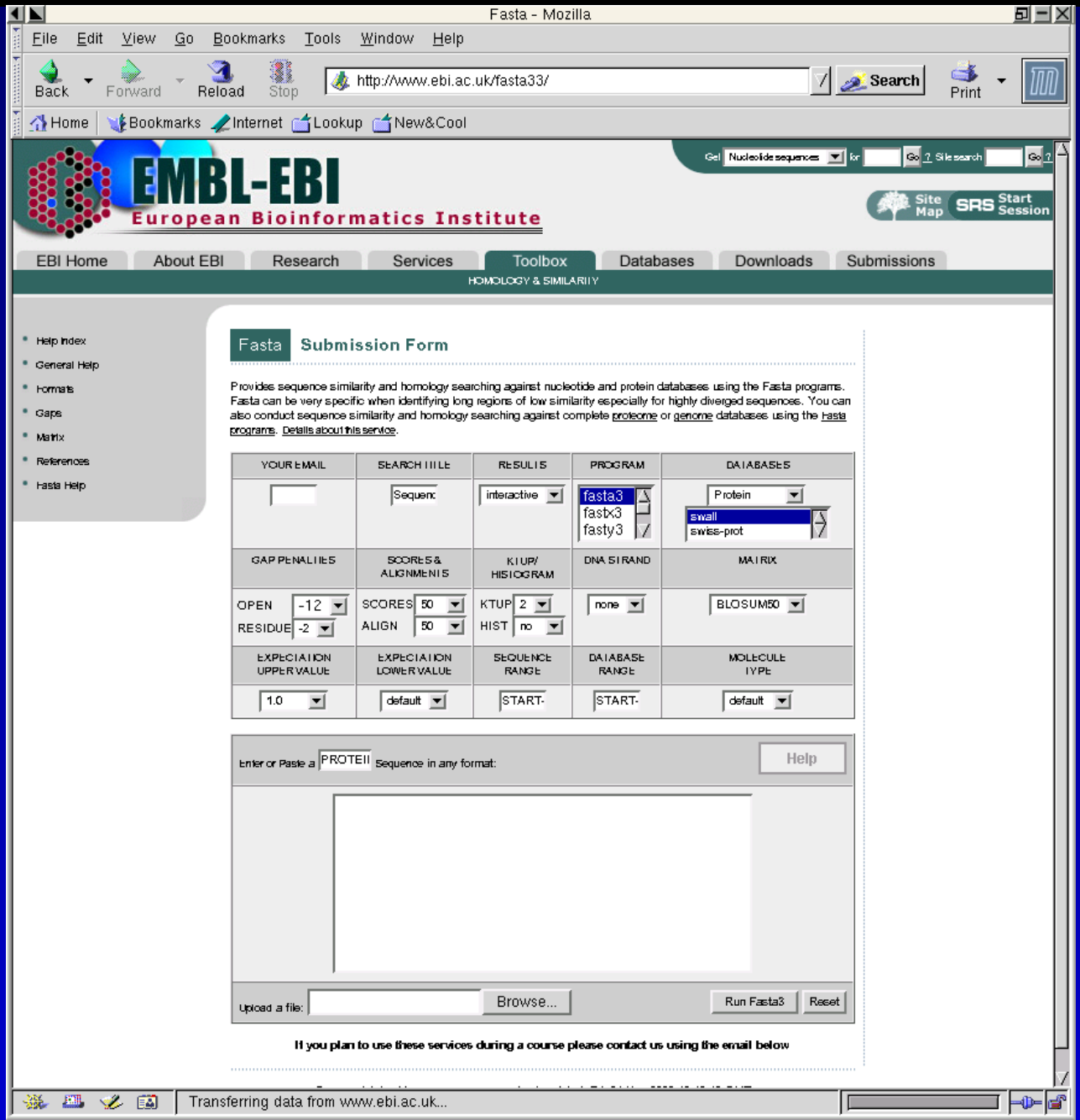
**vstupní parametry:**

k-tuple (velikost slova)
similarity matrix
gap open penalty
extended gap penalty.

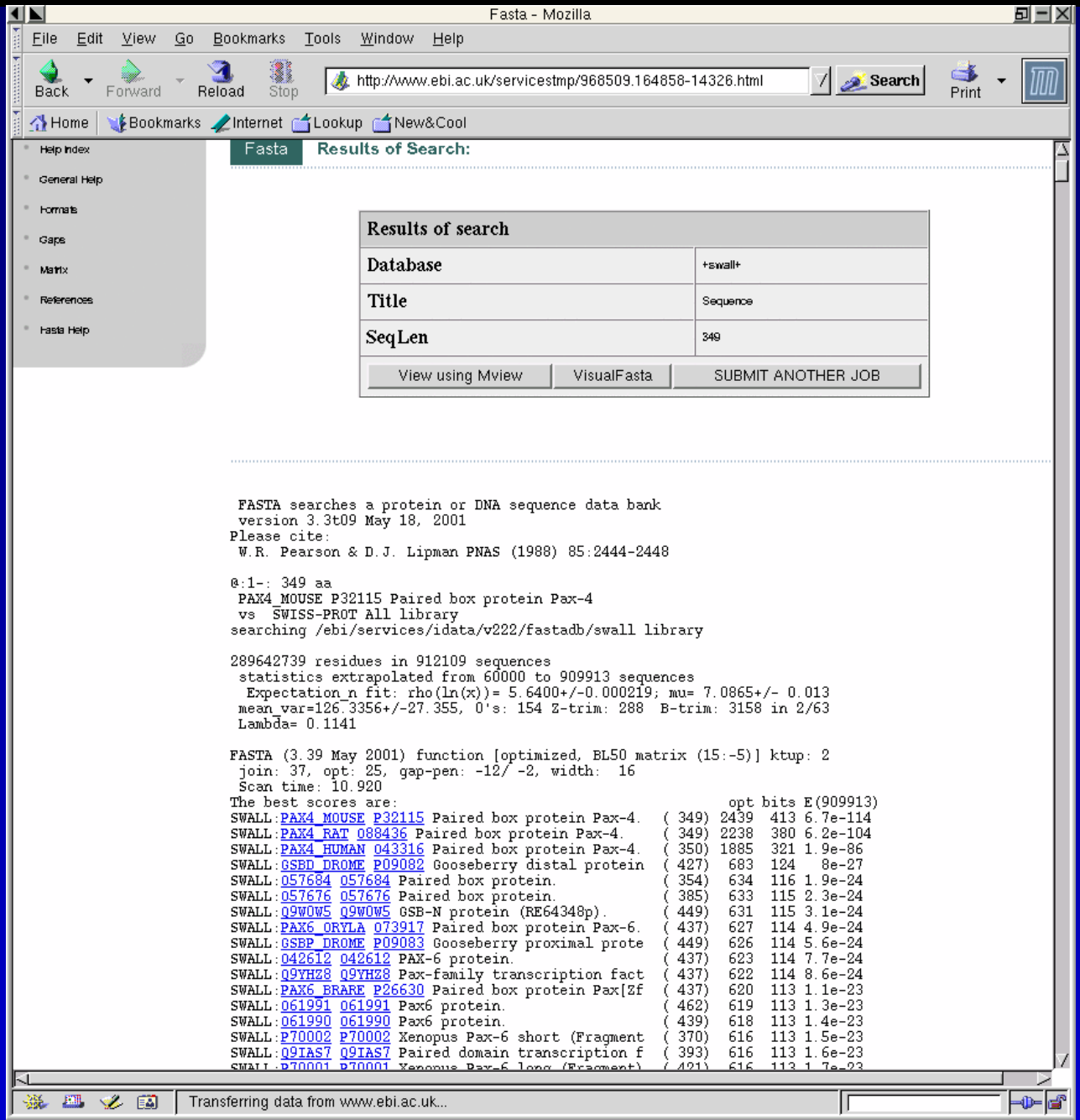
**programy:**

| | |
|---|---|
| fasta3 | DNA x DNAdb nebo AA x AAdb |
| tfasta3 | AA x DNAdb přeloženou do AA v šesti možných framech |
| fastx/y3 | DNA přeloženou x AAdb |
| tfastx/y3 | AA x DNAdb přeloženou |
| (t)fastf3 | seřazené peptidy (Edman) x DNAdb nebo AAdb |
| (t)fasts3 | peptidy (hmotová spektroskopie) x DNAdb nebo AAdb |
| ssearch | DNA x DNA nebo AA x AA, Smith-Waterman bez optimalizace |

# FastA - www


Fasta - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://www.ebi.ac.uk/fasta33/
Search
Print
Home Bookmarks Internet Lookup New&Cool
EMBL-EBI
European Bioinformatics Institute
Site Map
SRS Start Session
EBI Home
About EBI
Research
Services
Toolbox
Databases
Downloads
Submissions
HOMOLOGY & SIMILARITY
Help Index
General Help
Formats
Gaps
Matrix
References
Fasta Help
Fasta
Submission Form
Provides sequence similarity and homology searching against nucleotide and protein databases using the Fasta programs. Fasta can be very specific when identifying long regions of low similarity especially for highly diverged sequences. You can also conduct sequence similarity and homology searching against complete proteome or genome databases using the fasta programs. Details about this service.
YOUR EMAIL
SEARCH TITLE
RESULTS
PROGRAM
DATABASES
interactive
fasta3
fastx3
fasty3
Protein
swall
swiss-prot
GAP PENALTIES
SCORES & ALIGNMENTS
KTUP/ HISTOGRAM
DNA STRAND
MATRIX
OPEN -12
RESIDUE -2
SCORES 50
ALIGN 50
KTUP 2
HIST no
none
BLOSUM50
EXPECTATION UPPER VALUE
EXPECTATION LOWER VALUE
SEQUENCE RANGE
DATABASE RANGE
MOLECULE TYPE
1.0
default
START-
START-
default
Enter or Paste a PROTEIN Sequence in any format:
Help
Upload a file:
Browse...
Run Fasta3
Reset
If you plan to use these services during a course please contact us using the email below
Transferring data from www.ebi.ac.uk...

# FastA - výsledky



# FastA - výsledky

```
>>SWALL:GSBD_DROME P09082 Gooseberry distal protein (BSH  (427 aa)
 initn: 706 init1: 500 opt: 683  Z-score: 621.0  bits: 123.7 E(): 8e-27
Smith-Waterman score: 683;  40.000% identity (41.424% ungapped) in 320 aa overlap (5-318:19-
333)

                          10        20        30        40
PAX4_M               MQQDGLSSVNQLGGLFVNGRPLPLDTRQQIVQLAIRGMRPCDISRS
                       : . ::::::.:::::::   :.::::.:  :.::: :::.
SWALL: MAVSALNMTPYFGGYPFQGQGRVNQLGGVFINGRPLPNHIRRQIVEMAAAGVRPCVISRQ
              10        20        30        40        50        60

        50        60        70        80        90       100
PAX4_M LKVSNGCVSKILGRYYRTGVLEPKCIGGSKPRLATPAVVARIAQLKDEYPALFAWEIQHQ
       :.:::::::::.:. .:: ..:  ::::::::.::: . .:: .::.  :.:.:::. .
SWALL: LRVSHGCVSKILNRFQETGSIRPGVIGGSKPRVATPDIESRIEELKQSQPGIFSWEIRAK
              70        80        90       100       110       120

       110       120       130       140       150       160
PAX4_M LCTEGLCTQDKAPSVSSINRVLRALQEDQSLHWTQLRSPAVLAPVLPSPHSNCGA-PRGP
       :   :.: ..:::::::::.::. . . . :  .    .  . :    .:.  : :
SWALL: LIEAGVCDKQNAPSVSSISRLLRGSSGSGTSHSIDGILGGGAGSVGSEDESEDDAEPSVQ
             130       140       150       160       170       180

        170       180       190       200       210       220
PAX4_M HPGTSHRNRTIFSPGQAEALEKEFQRGQYPDSVARGKLAAATSLPEDTVRVWFSNRRAKW
          .:.:: ::  : .:::. : : ::::  .: .:: .:.: :  :.:::::::.
SWALL: LKRKQRRSRTTFSNDQIDALERIFARTQYPDVYTREELAQSTGLTEARVQVWFSNRRARL
             190       200       210       220       230       240
```

# FastA - poznámky

Zvýšením k-tuple se zvýší rychlost, ale sníží senzitivita.

Může minout pozitivní signál:

- sekvence **GGtTCtACgAAg** a **GGcTCcACaAAa** kódují stejný peptid **Gly-Ser-Thr-Lys**, ale při k-tuple > 2 nebude podobnost nalezena
- peptidy **Asp-Lys-Val** a **Glu-Arg-Ile** jsou si biochemicky podobné, aminokyseliny jsou různé
- podobnost mezi peptidy **Gly-Asp-Gly-Lys-Gly** a **Gly-Glu-Gly-Arg-Gly** pro k-tuple 2 a více nebude nalezena

# FastA - reference

**W. J. Wilbur and D. J. Lipman.** Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. U.S.A. 80:*726-730 (1983)

**D. J. Lipman and W. R. Pearson.** Rapid and sensitive protein similarity searches. *Science 227:*1435-1441 (1985)

**W. R. Pearson and D. J. Lipman.** Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A. 85:*2444-2448 (1988)

# BLAST - princip

## BLAST - Basic Local Alignment Tool

1. **Definujeme HSP (high segment scoring pair) jako úsek stejné délky dvou sekvencí se skóre, které nelze zlepšit prodloužením.**
2. **Předkompilujeme všechna slova o délce w se skóre lepším než T k dané sekvenci.**
3. **Hledáme v databázi zásahy ("hits") těchto slov.**
4. **Prodloužíme zásahy až do HSP.** (Pro NCBI-BLAST2 uvažujeme alespoň dva nepřekrývající se zásahy ve vzdálenosti A na diagonále.)
5. **Spočteme bit-skóre a expectancy**
6. **(Pro DNA použijeme čtyř bitovou kompresi.)**

# NCBI-BLAST - použití

**zdrojový kód: ftp://ncbi.nlm.nih.gov/tools**

**program: ftp://ncbi.nlm.nih.gov/blast/executables**

(UNIXy i windows, akademické použití zdarma)

**www: http://www.ncbi.nlm.nih.gov/blast/blast.cgi**

**vstupní parametry:**

similarity matrix
gap existence cost
per residue gap cost
lambda ratio

**programy:**

blastn DNA x DNAdb
blastp AA x AAdb
blastx AA x DNAdb přeloženou do AA v šesti možných framech

# WU-BLAST - použití

**zdrojový kód:** **pouze verze 1.x**

**program:**

**http://sapiens.wustl.edu/blast/blast/executables**

(Pouze UNIXy, pro akademické užití zdarma.)

**www:** **http://www.ebi.ac.uk/blast2**

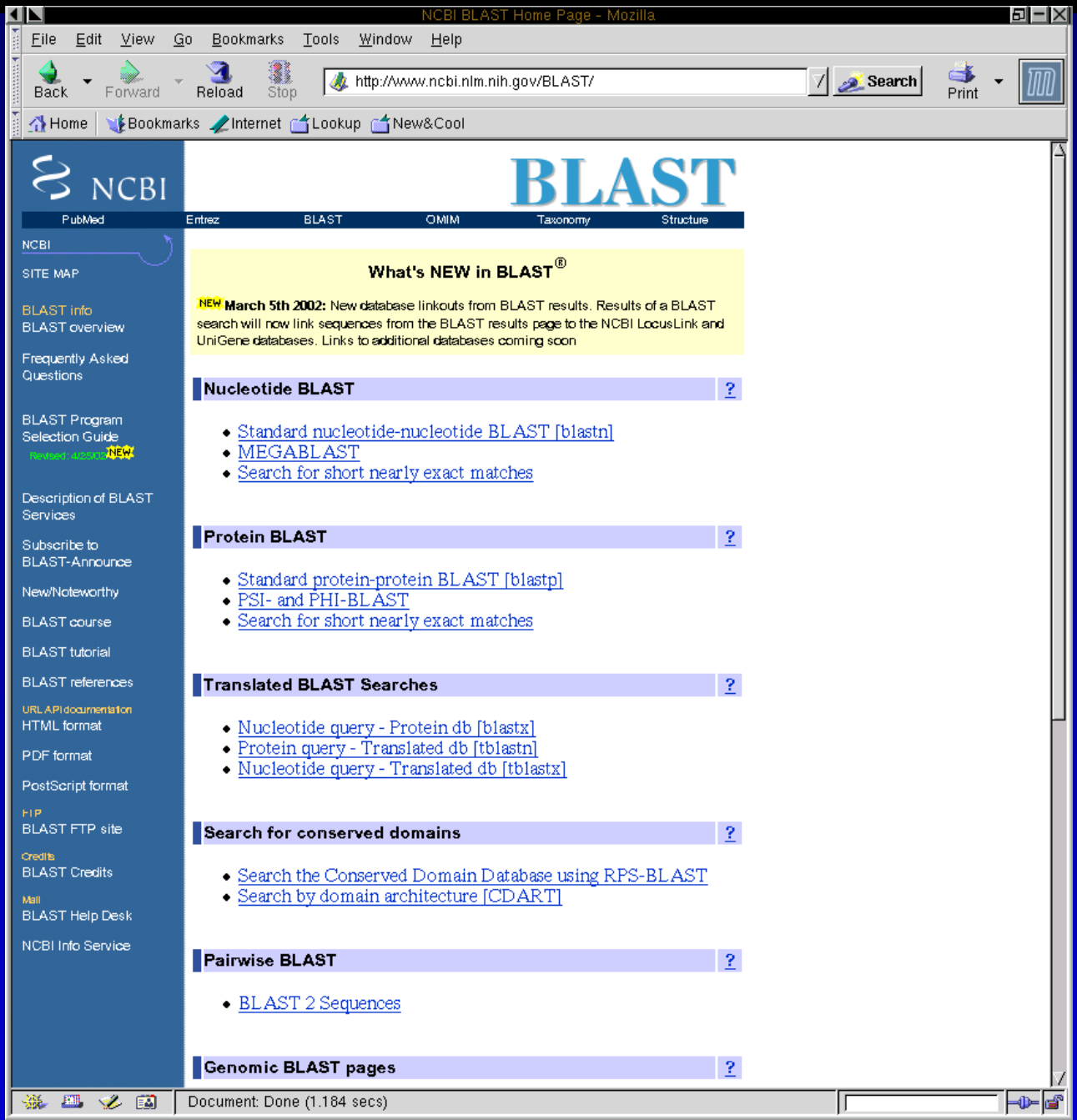
**vstupní parametry:**

similarity matrix
gap existence cost
per residue gap cost

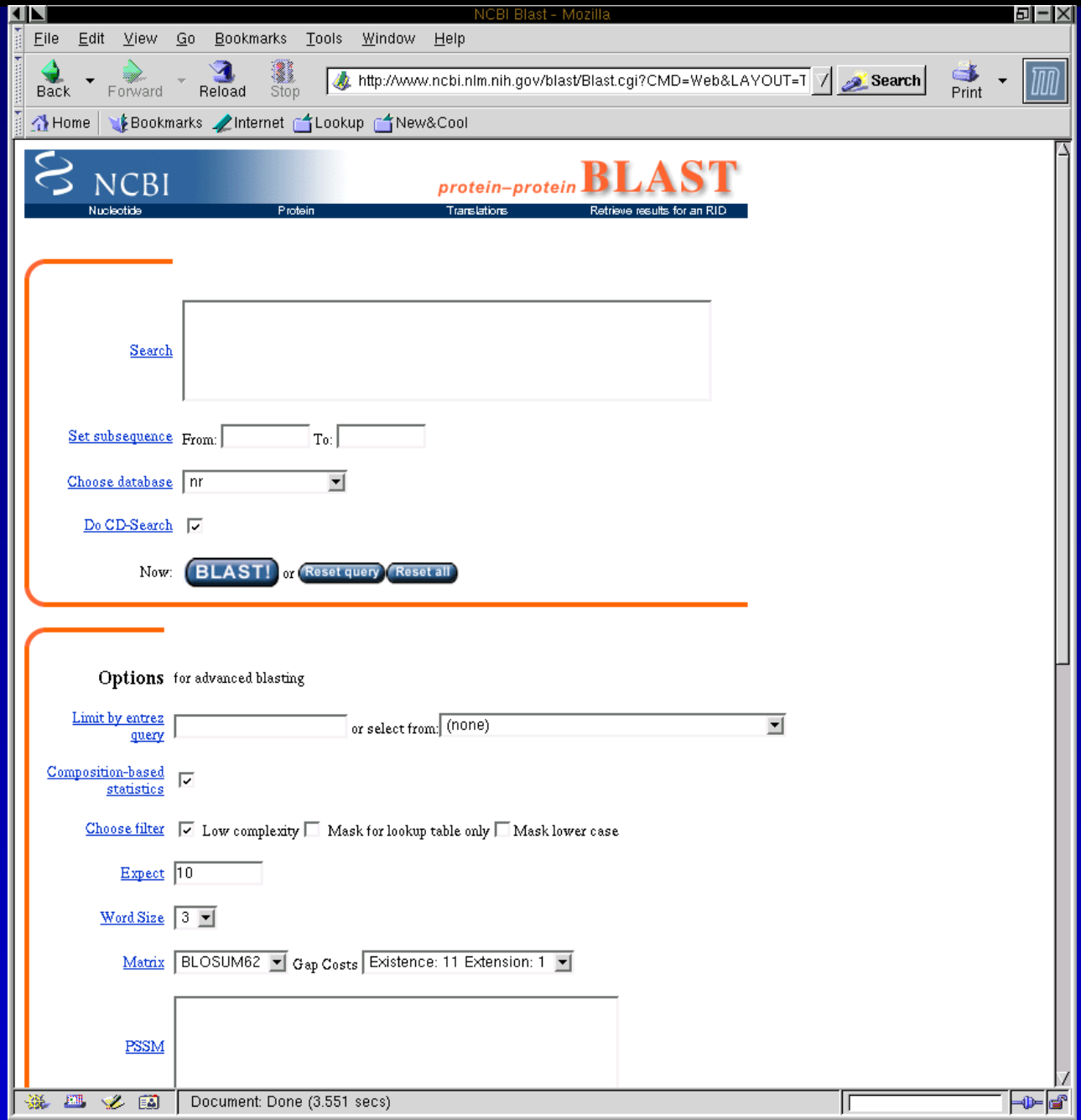
**programy:**

| | |
|---|---|
| blastn | DNA x DNAdb |
| blastp | AA x AAdb |
| blastx | AA x DNAdb přeloženou do AA v šesti možných framech |
| tblastn | DNA x AAdb |
| tblastx | DNA x DNAdb přeloženou |

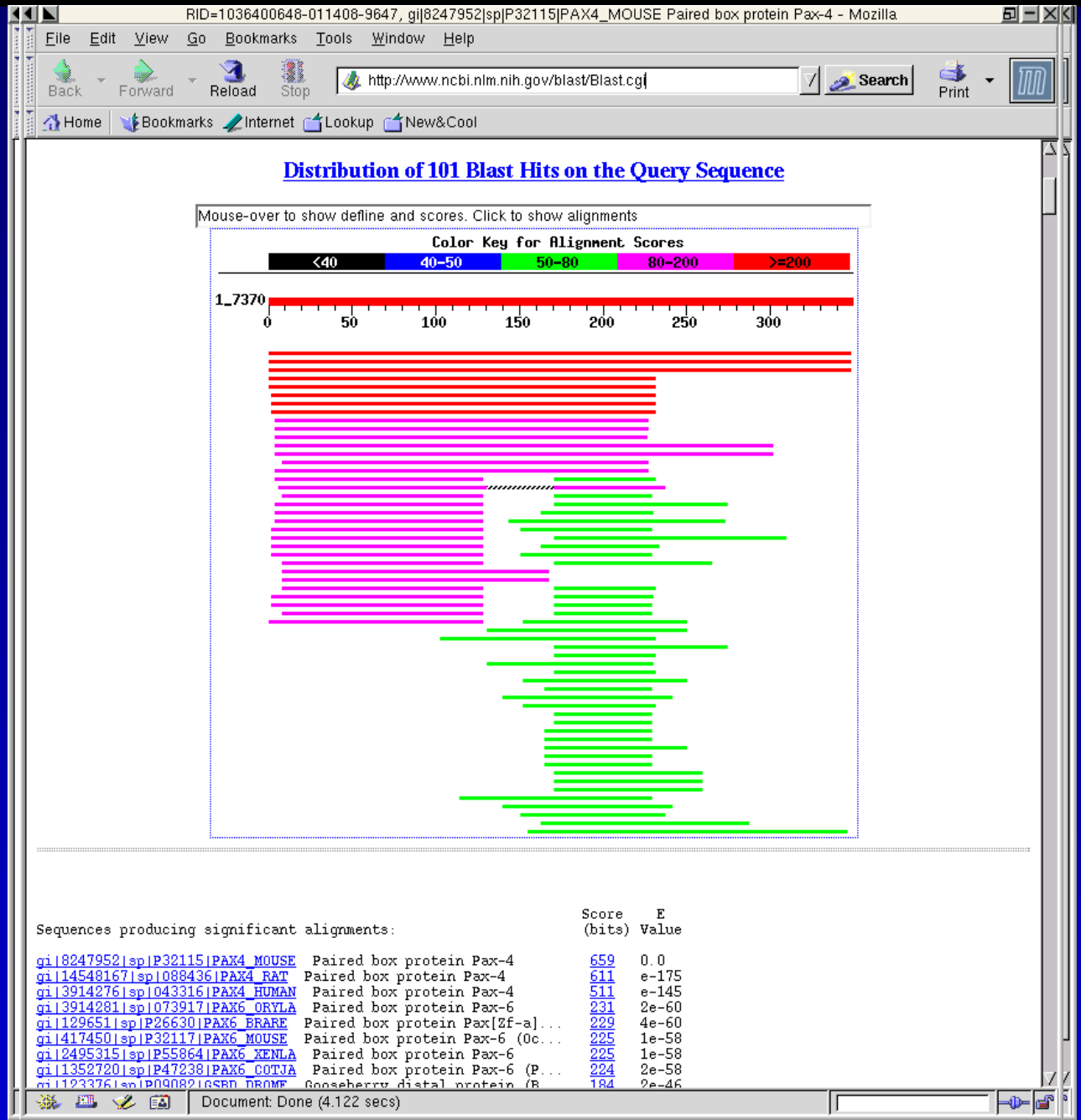
# BLAST - www


NCBI BLAST Home Page - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://www.ncbi.nlm.nih.gov/BLAST/
Search
Print
Home Bookmarks Internet Lookup New&Cool
NCBI
BLAST
PubMed Entrez BLAST OMIM Taxonomy Structure
NCBI
SITE MAP
BLAST info
BLAST overview
Frequently Asked Questions
BLAST Program Selection Guide
Description of BLAST Services
Subscribe to BLAST-Announce
New/Noteworthy
BLAST course
BLAST tutorial
BLAST references
URL API documentation
HTML format
PDF format
PostScript format
FTP
BLAST FTP site
Credits
BLAST Credits
Mail
BLAST Help Desk
NCBI Info Service
What's NEW in BLAST®
NEW March 5th 2002: New database linkouts from BLAST results. Results of a BLAST search will now link sequences from the BLAST results page to the NCBI LocusLink and UniGene databases. Links to additional databases coming soon
Nucleotide BLAST
Standard nucleotide-nucleotide BLAST [blastn]
MEGABLAST
Search for short nearly exact matches
Protein BLAST
Standard protein-protein BLAST [blastp]
PSI- and PHI-BLAST
Search for short nearly exact matches
Translated BLAST Searches
Nucleotide query - Protein db [blastx]
Protein query - Translated db [tblastn]
Nucleotide query - Translated db [tblastx]
Search for conserved domains
Search the Conserved Domain Database using RPS-BLAST
Search by domain architecture [CDART]
Pairwise BLAST
BLAST 2 Sequences
Genomic BLAST pages
Document: Done (1.184 secs)

# BLAST - www


NCBI Blast - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&LAYOUT=T
Search
Print
Home Bookmarks Internet Lookup New&Cool
NCBI
protein–protein BLAST
Nucleotide Protein Translations Retrieve results for an RID
Search
Set subsequence From: To:
Choose database nr
Do CD-Search
Now: BLAST! or Reset query Reset all
Options for advanced blasting
Limit by entrez query or select from: (none)
Composition-based statistics
Choose filter Low complexity Mask for lookup table only Mask lower case
Expect 10
Word Size 3
Matrix BLOSUM62 Gap Costs Existence: 11 Extension: 1
PSSM
Document: Done (3.551 secs)

# BLAST - výsledky



# BLAST - výsledky

```
>gi|3914281|sp|O73917|PAX6_ORYLA   Paired box protein Pax-6
          Length = 437

 Score =  231 bits (589), Expect = 2e-60
 Identities = 142/274 (51%), Positives = 169/274 (60%), Gaps = 43/274 (15%)

Query: 1   MQQDGLSSVNQLGGLFVNGRPLPLDTRQQIVQLAIRGMRPCDISRSLKVSNGCVSKILGR 60
           M Q+  S VNQLGG+FVNGRPLP  TRQ+IV+LA  G RPCDISR L+VSNGCVSKILGR
Sbjct: 19  MMQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGR 78

Query: 61  YYRTGVLEPKCIGGSKPRLATPAVVARIAQLKDEYPALFAWEIQHQLCTEGLCTQDKAPS 120
           YY TG + P+ IGGSKPR+ATP VVA+IAQ K E P++FAWEI+ +L +EG+CT D  PS
Sbjct: 79  YYETGSIRPRAIGGSKPRVATPEVVAKIAQYKRECPSIFAWEIRDRLLSEGICTNDNIPS 138

Query: 121 VSSINRVLRAL-QEDQSL----HWTQLRS---------------PAVLAPVLPSPHSNCG 160
           VSSINRVLR L  E Q +     + +LR                 P    P  P+    C
Sbjct: 139 VSSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGTWGTRPGWYPGTSVPGQPN-QDGCQ 197

Query: 161 APRGPHPGTS--------------------HRNRTIFSPGQAEALEKEFQRGQYPDSV 198
              G    T+                    RNRT F+  Q EALEKEF+R  YPD
Sbjct: 198 QQDGAGENTNSISSNGEDSEETQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVF 257

Query: 199 ARGKLAAATSLPEDTVRVWFSNRRAKWRRQEKLK 232
           AR +LAA   LPE  ++VWFSNRRAKWRR+EKL+
Sbjct: 258 ARERLAAKIDLPEARIQVWFSNRRAKWRREEKLR 291
```

# BLAST - reference

**S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman.** Basic Local Alignment Search Tool. *J. Mol. Biol. 215*:403-410 (1990)

**Karlin, Samuel and Stephen F. Altschul**. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. USA 90*:5873-7 (1993)

**Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ.** Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res. 25(17)*:3389-402. (1997)

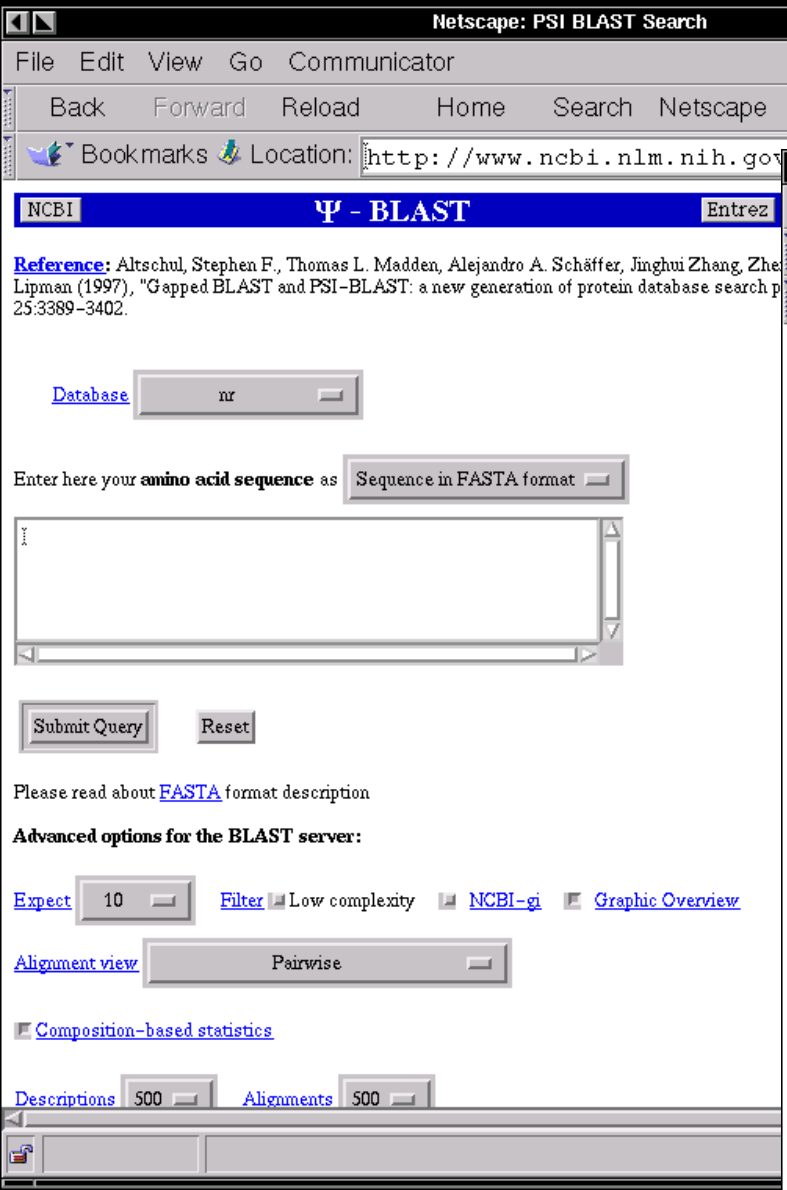
# PSI-BLAST - princip

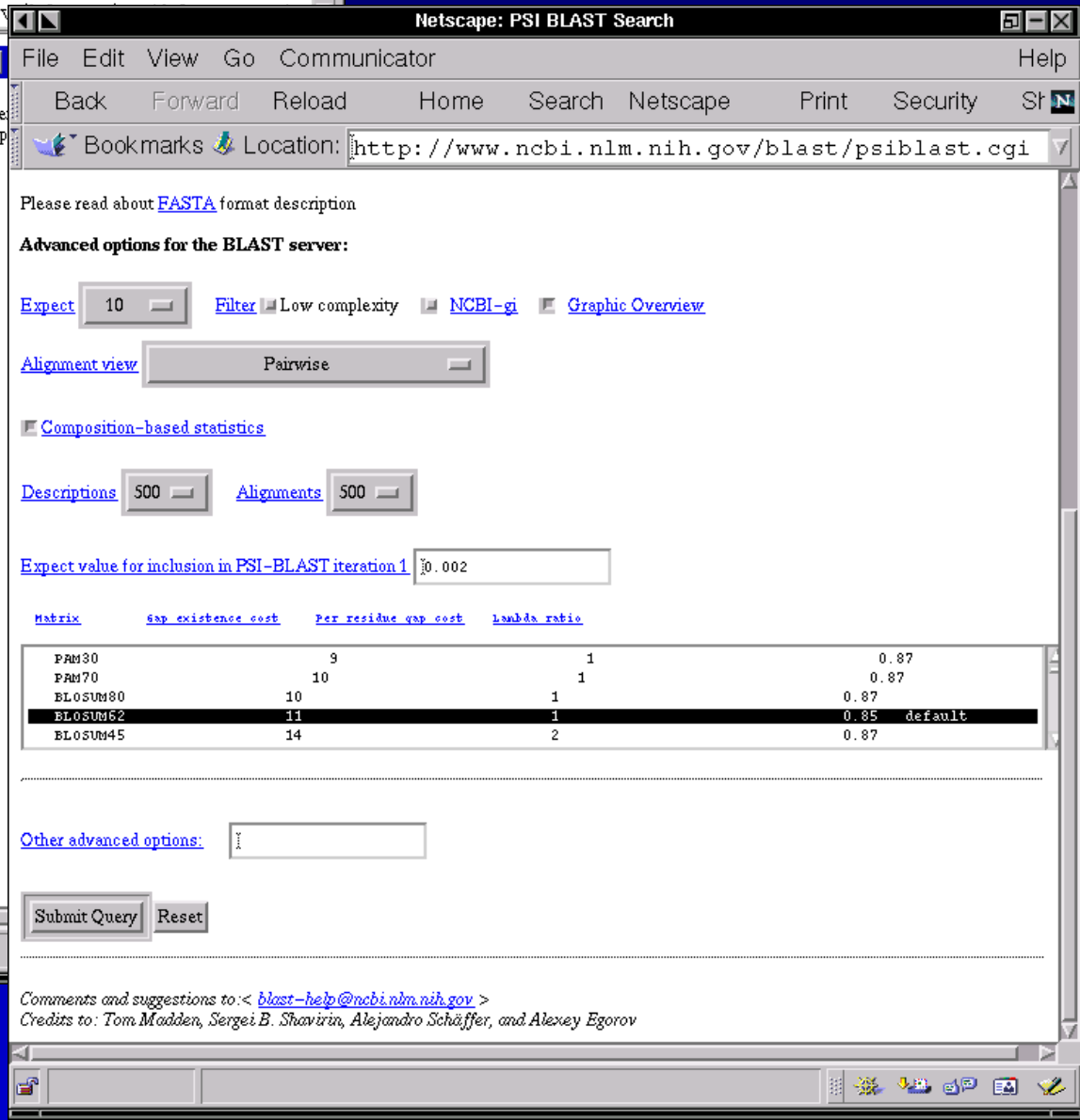
## PSI-BLAST - Position Specific Iterated BLAST

3. Pomocí BLAST získáme sadu sekvencí se skóre lepším než T.
4. Sestrojíme multiple alignment.
5. Identity matrix o velikosti 20x20 nahradíme matrix o velikosti Lx20 (kde L je délka použité sekvence), kterou spočteme z multiple alignmentu
6. Získáme novou sadu sekvencí.
7. Iterujeme přes kroky 2-4.

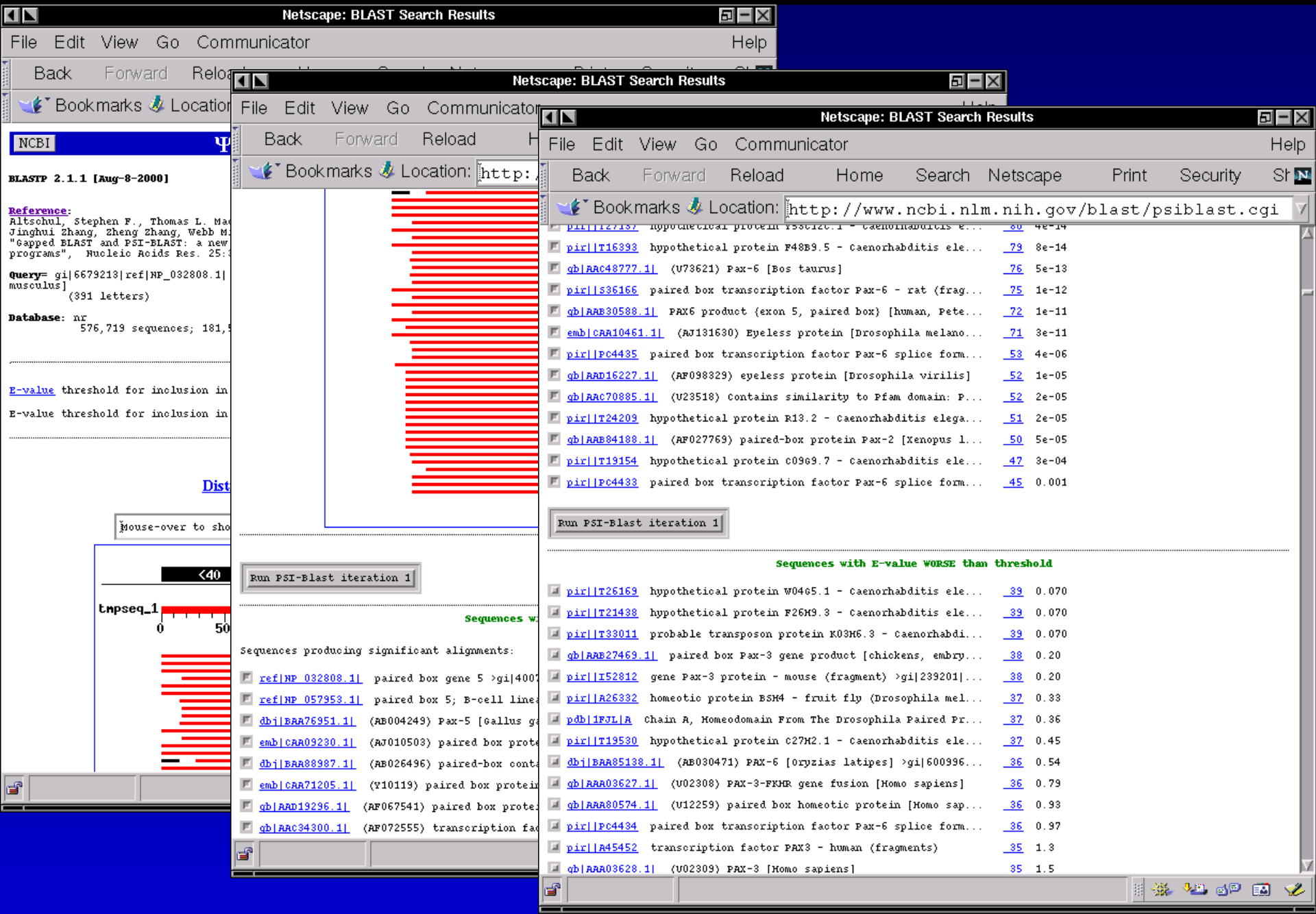
**www: http://www.ncbi.nlm.nih.gov/blast/psiblast.cgi**

# PSI-BLAST - www

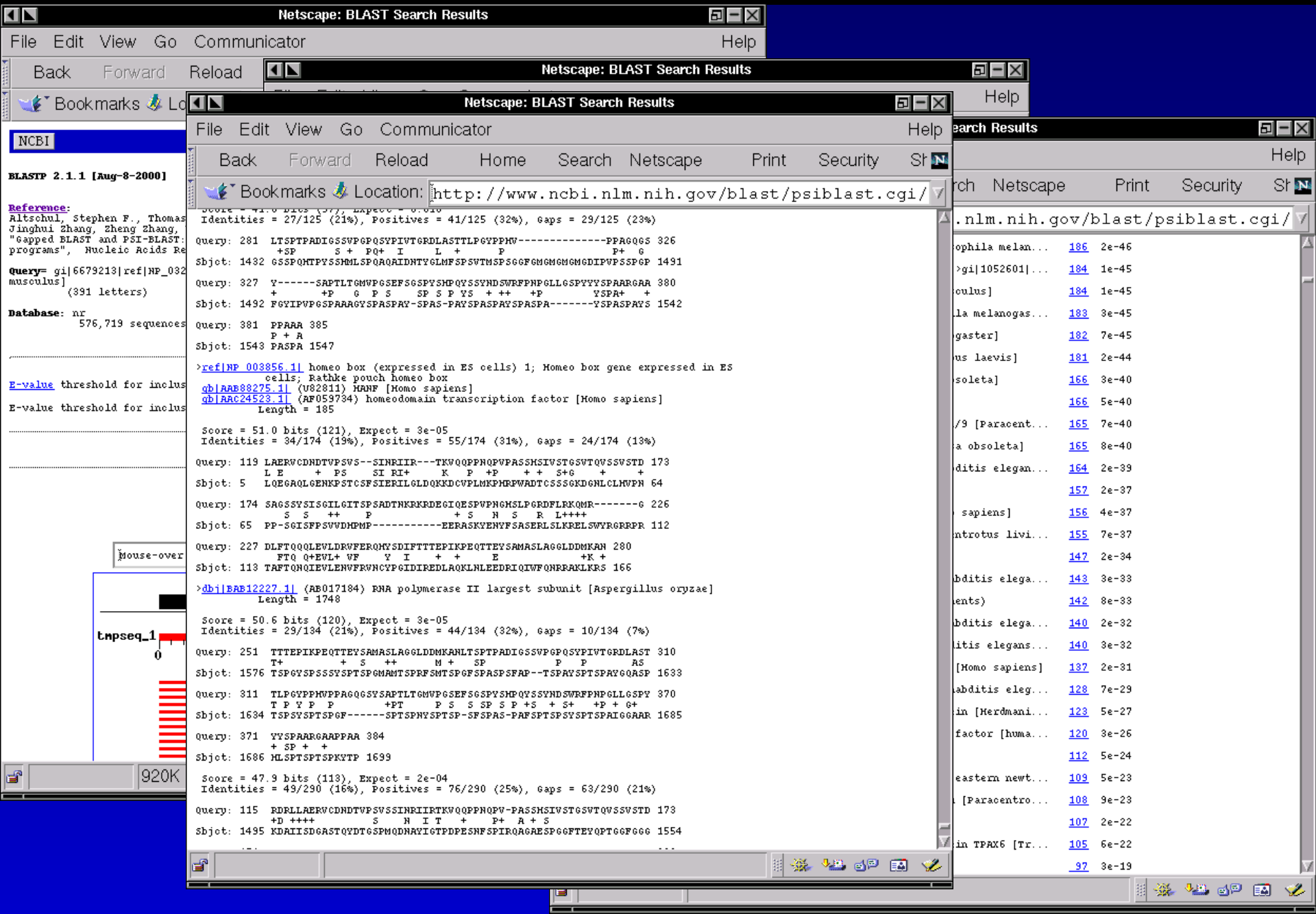


# PSI-BLAST - výsledky



# PSI-BLAST - výsledky



# PHI-BLAST - princip

## PHI-BLAST - Pattern Hit Initiated BLAST

3. Pomocí BLASTu získáme sadu sekvencí se skóre lepším než S. Pro výpočet skóre použijeme "pattern", krátkou sekvenci ve formátu PROSITE.
4. Skóre Sestrojíme multiple alignment a pro iterace použijeme PSI-BLAST
5. patterny: (ve stylu PROSITE)
   [LFYT] výběr z několika aminokyselin
   x(2,5) 2 až 5 libovolných aminokyselin
   - nic (oddělovač)
   například [LIVMF]-G-E-x(5,11)-A(3)-x-[STACV]

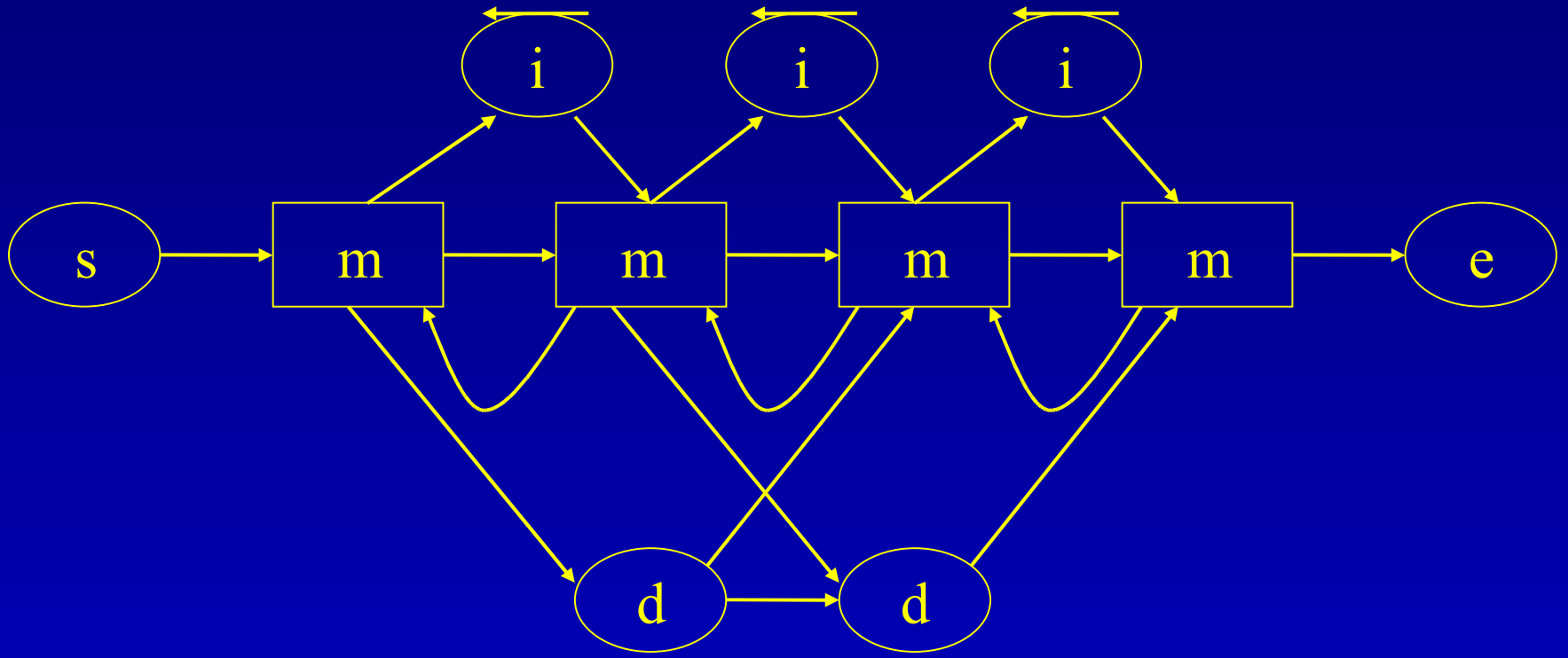
# PHI-BLAST - princip

**www: http://www.ncbi.nlm.nih.gov/blast/**
**/psiblast.cgi?Jform=1**

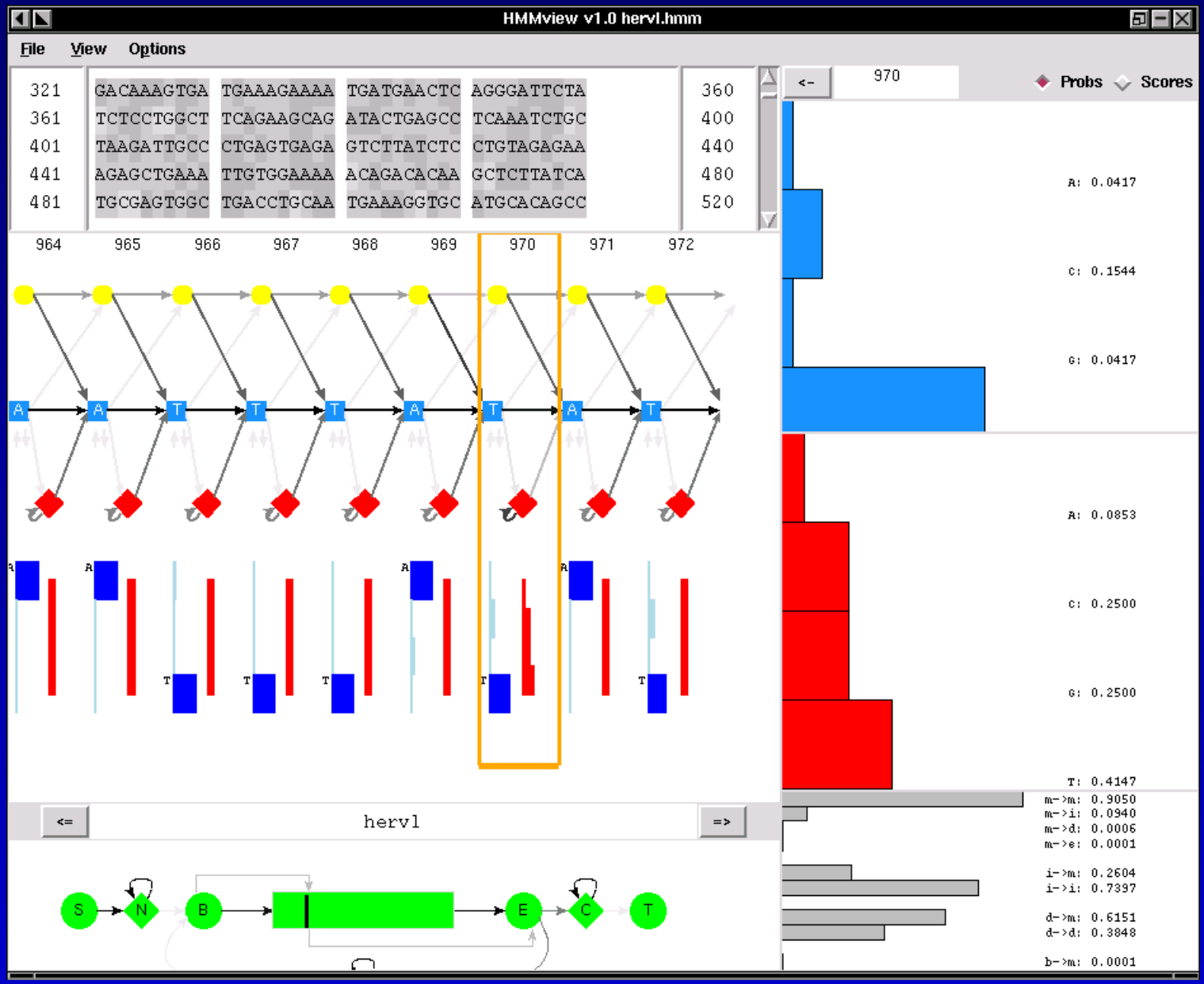
**reference:**

**Zhang, Zheng, Alejandro A. Schäffer, Webb Miller, Thomas L. Madden, David J. Lipman, Eugene V. Koonin, and Stephen F. Altschul**, Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res. 26:*3986-3990. (1998)

# HMMER - princip



# HMMER - vizualizace



# HMMER - použití

**zdrojový kód:** **http://hmmer.wustl.edu**

(Zdrojový kód pro akademické použití volný, kompilace pod UNIXy bez problémů)

**www:** **http://pfam.wustl.edu**

**programy:**

| | |
|---|---|
| hmmsearch | prohledává modelem (hmmerem) databázi sekvencí |
| hmmerpfam | prohledává sekvencí databázi modelů |