

Bioinformatika

<http://bio.img.cas.cz>

Jan Pačes
Ústav molekulární genetiky
hpacesimg.cas.cz



CZECH FOBIA

Bioinformatics

statistics, alignments, phylogeny, clustering

Radek Zíka

Bioinformatics department
Institute of Molecular Genetics



e-mail: [zikar!\[\]\(0f848bbd71cef6b345273b16f905912a_img.jpg\)img.cas.cz](mailto:zikar@img.cas.cz)

obsah

**Biologické sekvence a jejich záznam
Zdroje na internetu**

Alignment

Pairwise alignment

Substitution matrix

FastA

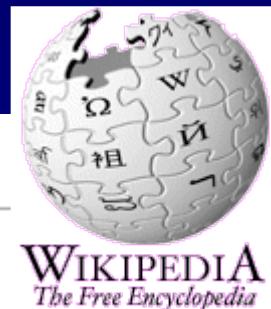
Blast

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**.

Addition

Enter word here

Exact Approximate

[Look up definition](#)

Hyperlinks

No definition found for "

You may wish to try an alternative search

- [bomber](#)
- [bombers](#)
- [bonfire](#)
- [bonfires](#)
- [bumper](#)
- [bumpers](#)
- [bumpier](#)

Load the new
Cambridge
Dictionary
and look up
anywhere on the
Web — it's free!

Dictionary

Thesaurus

Put a
Webster dictionary
on your Palm Pilot!

bioinformatika

Informatika nad biologickými molekulami (daty).

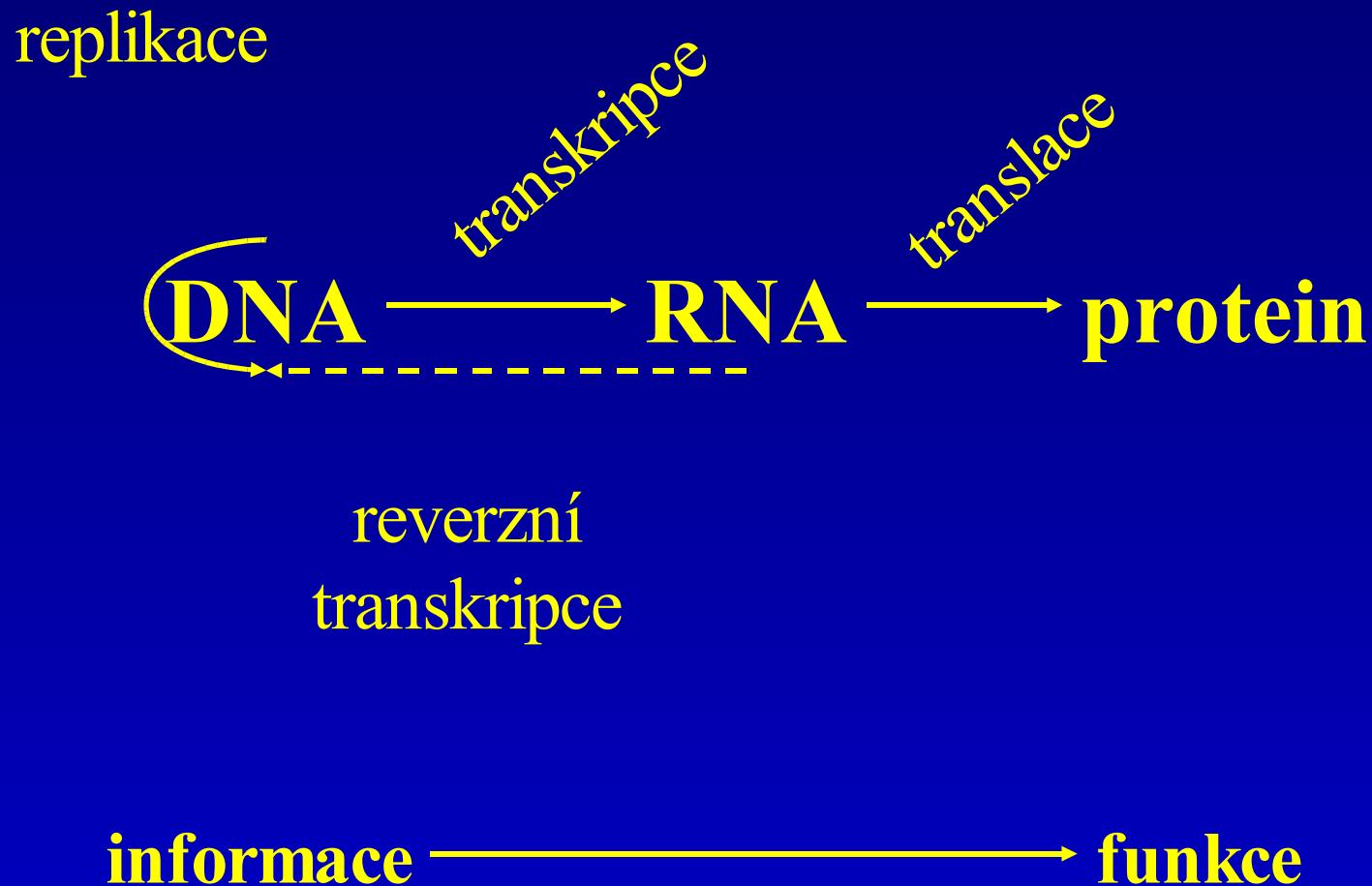
Bioinformatika extrahuje molekulární informační systém pro molekulární biologii.

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



velikosti genomů

	<i>Mycoplasma genitalium</i>	0.58 Mbp
	<i>Escherichia coli</i>	4.6 Mbp
	<i>Caenorhabditis elegans</i>	6 chr. 97 Mbp
	<i>Saccharomyces cerevisiae</i>	16 chr. 11.2 Mbp
	<i>Arabidopsis thaliana</i>	5 chr. 115.4 Mbp
	<i>Drosophila melanogaster</i>	5 chr. ~137.0 Mbp
	<i>Homo sapiens</i>	24 chr. ~ 3.3 Gbp

Internet, databases, www

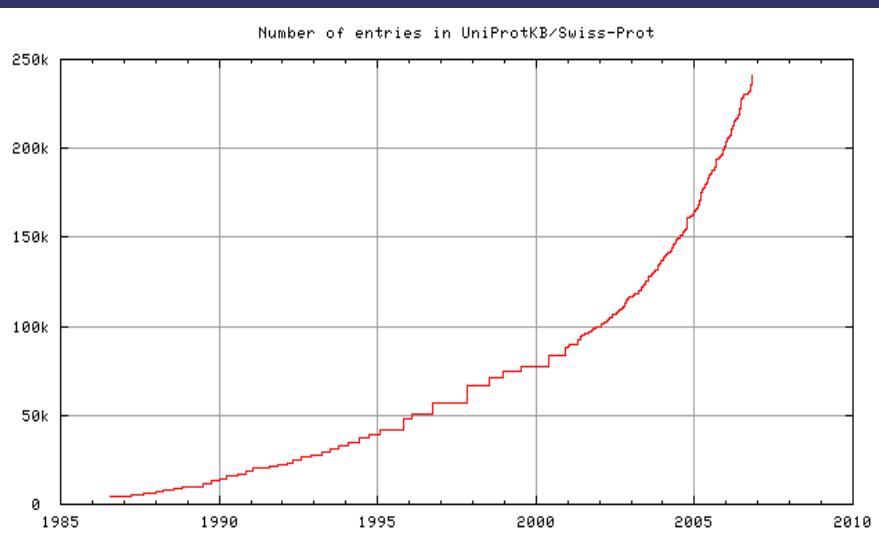
- NAR web issue 2004: **137 webs**
 - Alignments
 - NA, protein sequence analysis
 - Structure analysis & description
 - Visualization

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	jakákoliž aminokyselina
*	---	stop

formáty sekvencí

binární s chromatogramy

**SCF
ALF
ABI**

pro programy

interní formáty databází

textové minimální

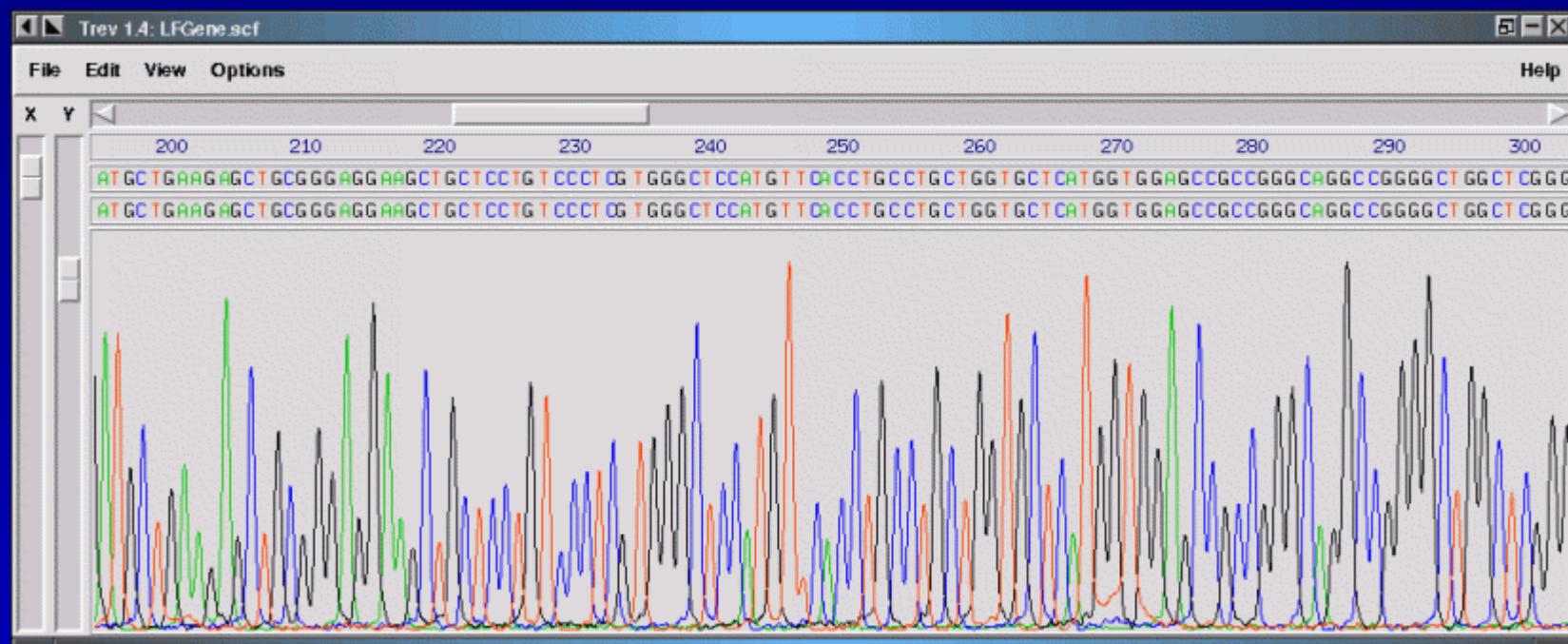
**text
fasta**

anotované

**EMBL
GenBank
ASN
XML**

formáty sekvencí - SCF

SCF (standart chromatogram file)



formáty sekvencí - FastA

fastA

```
>gi|6102607|gb|AF145233.1|AF145233 Mus musculus transcription factor PAX4
TGGCAGGACTGAAGCAGCTGGAGGCTGTTACAAGACCAAGACCACAGCAAACCTGGAGCCTGCACAGGA
CCCTGAGACCTCTCCTGGAATTCCCACCTTTCTCCATCCAGAACCGAGTCCAAAGAGAAACTTCC
AGAAGGAGCTCTCCGTTTCAGTTGCCAGTTGGCTTCCTGTGAGGAGTACCAAGTGTGAAGC
ATGCAGCAGGACGGACTCAGCAGTGTGAATCAGCTAGGGGACTCTTGTAATGCCGGCCCTTCCTC
TGGACACCAGGCAGCAGATTGTGCAGCTAGCAATAAGAGGGATGCGACCCCTGTGACATTACGGAGCCT
TAAGGTATCTAATGGCTGTGAGCAAGATCCTAGGACGCTACTACCGCACAGGTGTCTGGAACCCAAG
TGTATTGGGGAAAGCAAACACGTCTGGCCACACCTGCTGTGGCTCGAATTGCCAGCTAAAGGATG
AGTACCCCTGCTCTTTGCCTGGAGATCCAACACCAAGCTTGCACTGAAGGGCTTGTACCCAGGACAA
GGCTCCCAGTGTGCCTCTATCAATCGACTACTTCGGGCACTTCAGGAAGACCAGAGCTGCACTGGACT
CAACTCAGATCACCAAGCTGTGGCTCCAGTTCTCCAGTCCCCACAGTAACGTGGGGCTCCCGAG
GCCCCACCCAGGAACCAGCCACAGGAATCGGACTATCTCTCCCCGGACAAGCCGAGGCACTGGAGAA
AGAGTTTCAGCGTGGGCAGTATCCAGATTCACTGGCCGTGGGAAGCTGGCTGCTGCCACCTCTGCCT
GAAGACACGGTGAGGGTTGGTTCTAACAGAAGGCCAAATGGCGCAGGCAAGAGAAGCTGAAATGGG
AAGCACAGCTGCCAGGTGCTCCAGGACCTGACAGTACCAAAAAATTCTCCAGGGATCATCTCTGCACA
GCAGTCCCCGGCAGTGTACCCCTCAGCTGCCCTGCTGGAAACCATTGAGTCCTCCTCTGTCA
CTATGCTGTGGCACACCAGGAGATCTCCAGTGACACCTCATCCCAGCCATCTCCAACCTACT
GGGACTGCCAATCCCTCCTGTGGCTTCCCTCATATGTGGAATTGCTGGCCCTGCCTCACCA
CCATCCTGTGCATCATGTGATTGGAGGCCAGGACAAGTGCATCAACCCATTGCTCAAACGTGGCATAA
GAGGCCTCTATTGACAGTAATAAAAACCTTTCTTAGATGTTAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAA
```

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
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=MQQDGLSSVNQLGGLFVNNGRPLPLDTRQQIVQLAIRGMRPCDISR
FT  SLKVSNGCVSKILGRYYRTGVLEPKCIGGSKPRLATPAVVARIAQLKDEYPALFAWEIQ
...
FT          PSTHCSNWP
XX
SQ  Sequence 1379 BP; 327 A; 402 C; 347 G; 303 T; 0 other;
aaaaaaa     aaaaagcgac cgctgaattc tagcagaagg ctgcacctctg ctccctgagtg 60
...
gctgtggac agcaccaggc agatgttcca gtgacacactc atcccaggcc tatctccaaac 1200
cctactggga ctgccaatcc ctccttcctg tggcttcctc ctcatatgtg gaatttgccct 1260
ggccctgcct caccaccat cctgtgcattc atctgattgg aggcccaggaa caagtgcatt 1320
caacccattg ctcattttttt 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
IQHQLCTEGLCTQDKAPS VSSINRVLRALQEDQSLHWTQLRSPAVLAPVLPSPHSNCG
APRGPHPGTSHRNRTIFSPGQAEALEKEFQRGQYPDSVARGK LAAATSLPEDTVRVWF
SNRRAKWRRQEKLKWEAQLPGASQDLTVPKNSPGIISAQQSPGSVPSAALPVLEPLSP
SFCQLCCGTAPGRCSSDTSSQAYLQPYWDCQSLLPVASSSYVEFAWPCLTTHPVHHLI
GGPGQVPSTHCSNWP"

BASE COUNT

359 a 381 c 328 g 292 t

ORIGIN

1 tggcaggact gaagcagctg gaggctgtta caagaccaga ccaccagcaa accctggagc
61 ctgcacagga ccctgagacc tttccctgga attcccacct ttttcctcc atccagaacc
121 agtcccaaag agaaacttcc agaaggagct ctccgtttc agttgccag ttggcttcct
181 gtccttctgt gaggagtacc agtgtgaagc atgcagcagg acggactcag cagtgtgaat
...
1201 catcatctga ttggaggccc aggacaagtg ccatcaaccc attgctcaaa ctggccataa
1261 gaggcctcta tttgacagta ataaaaaacct tttcttagat gttaaaaaaaaaaaaaaaa
1321 aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa

//

zdroje na internetu

http://

Swiss
Entrez
SRS

The screenshot shows a Mozilla browser window displaying the SRS6 interface at <http://srs6.ebi.ac.uk/index.htm>. The page version is 6.0.7.3. On the left, there's a sidebar with the EBI logo and text: "European Bioinformatics Institute". The main area features a large blue paw print graphic with three menu options: "Start", "Permanent Session", and "Information". The "Start" option is highlighted. At the bottom, there's a toolbar with icons for file operations like Open, Save, Print, and a search bar.

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

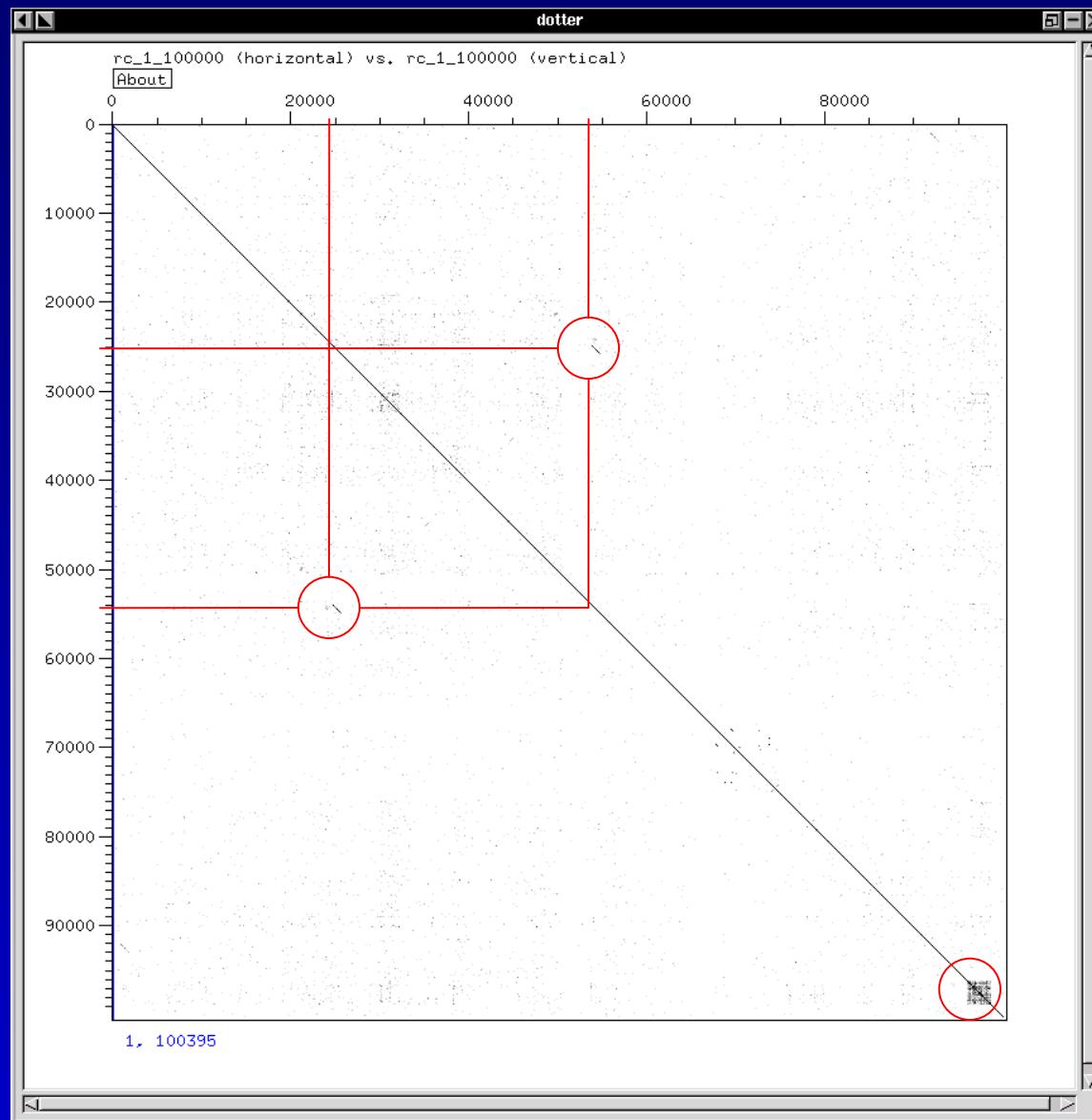
n:n

PSI-BLAST
HMMER

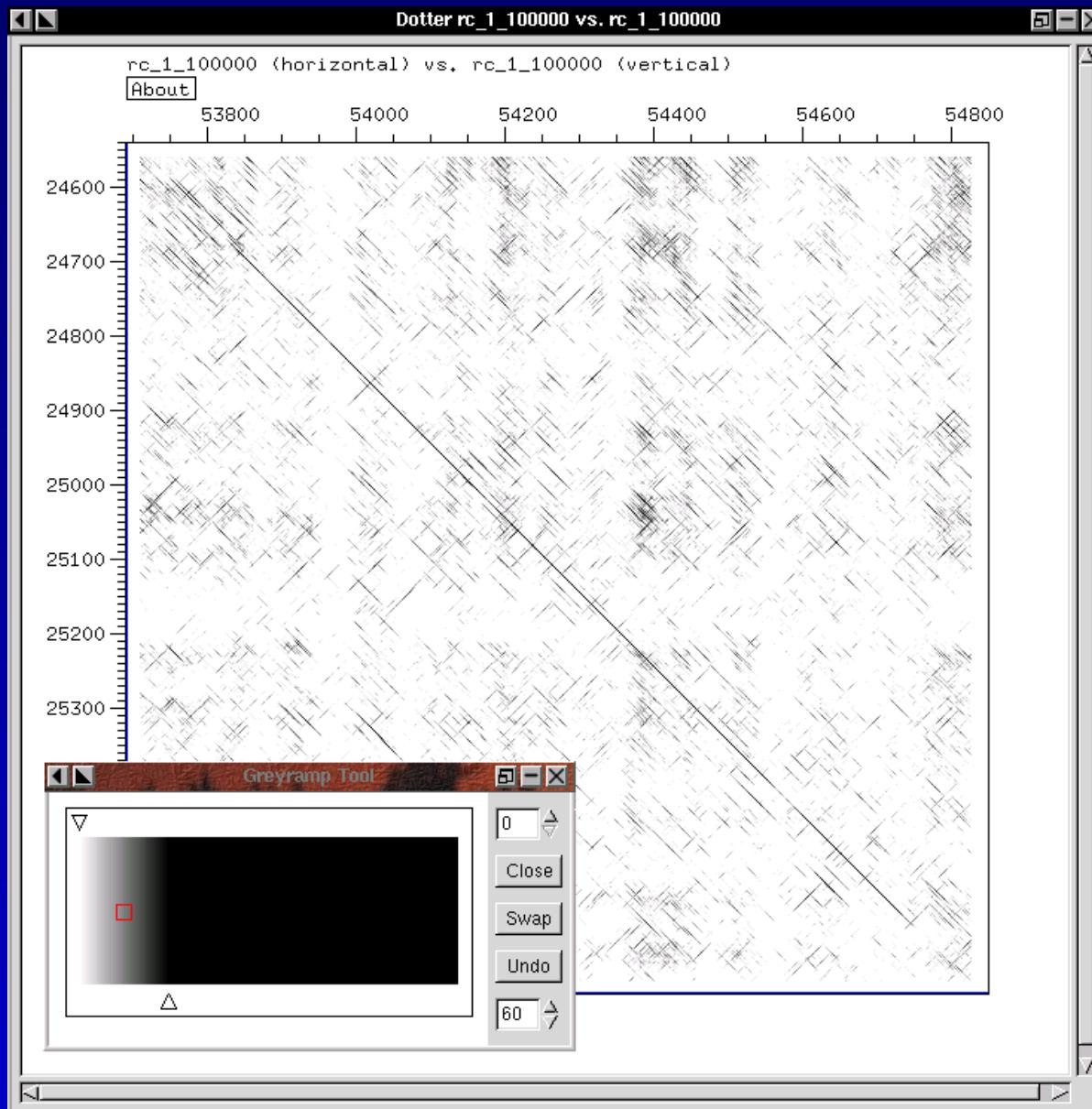
n

ClustalW
MultAlign
T-Coffee
Muscle

dot plot



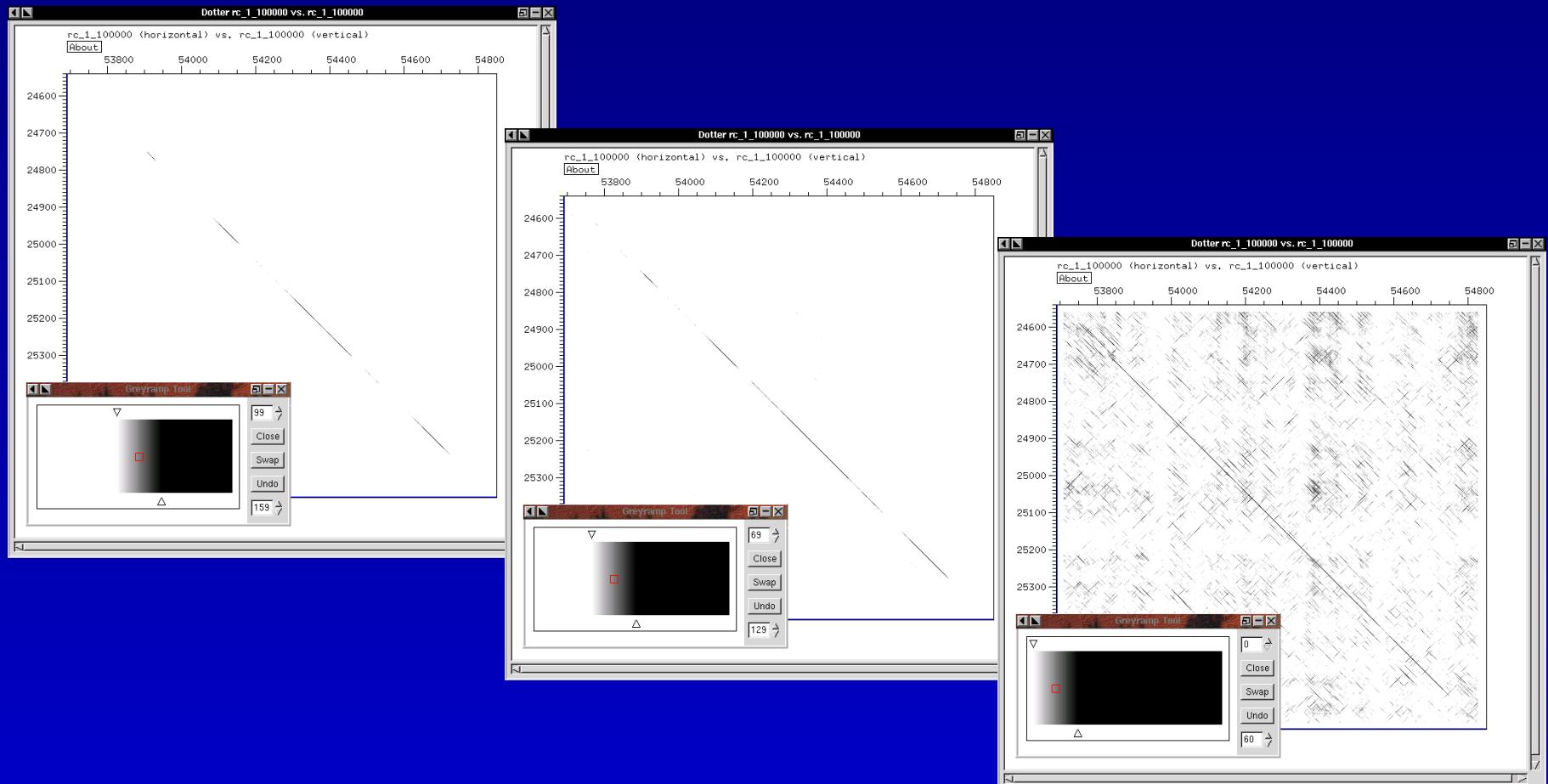
dot plot



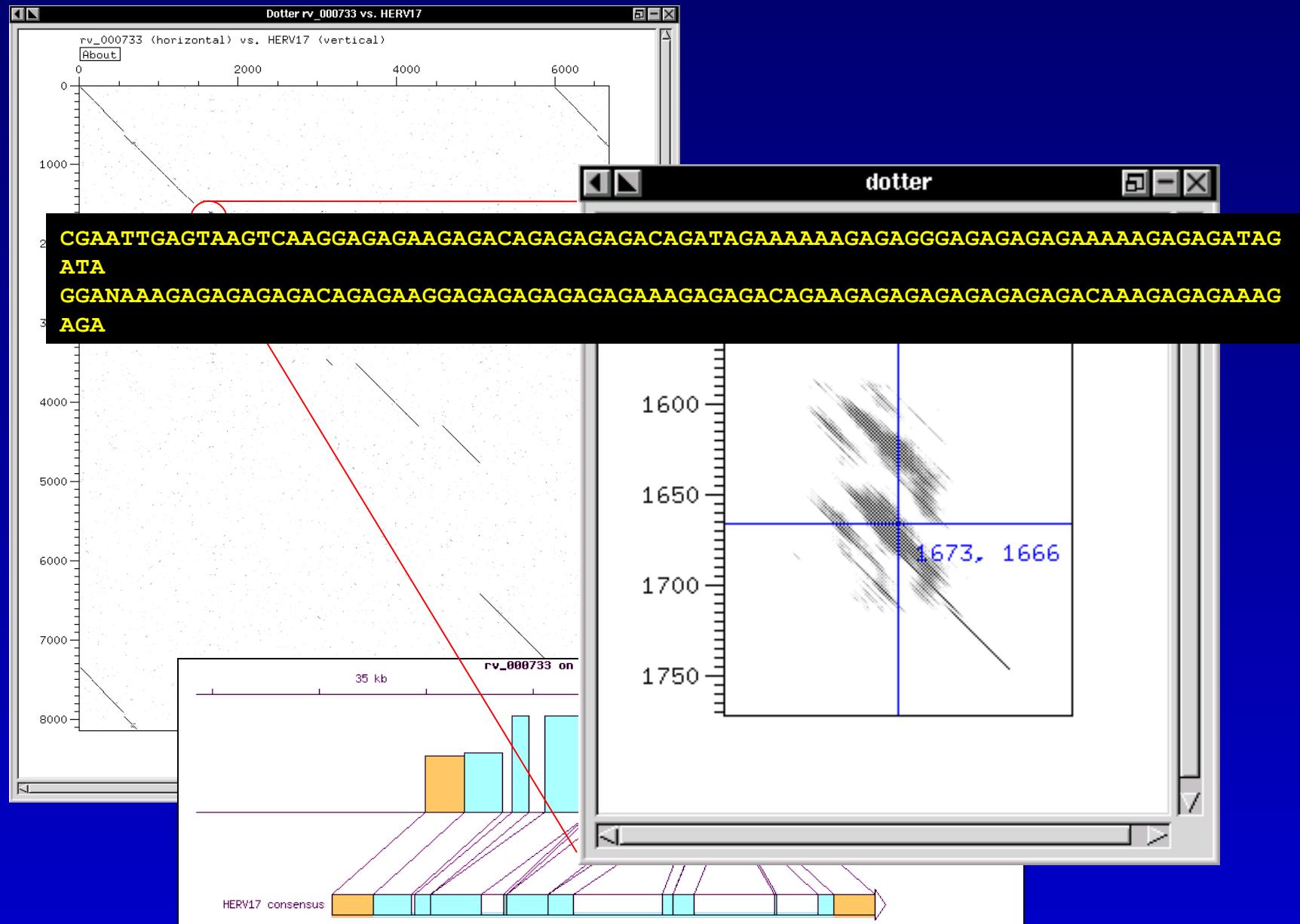
dot plot

54268 CGGTCAGCCGGACCCGGACCACCGAACGGCAACGGGCGCGATGTGGTGCCTCAGACCGCGCAGGGCATGGACGACATTGCCCGCAGCTCGGAACAGATCTCGGCATCACCAAGCGTCATCGACGAC

25111 CGGTGGCCAGACCCGCGAGACCACCGCCGCCGGCGCAGGTGGTGCCTGCAGCGATGACCGACATCGCGAAAGCTCGGAGCAGATTTCCCGCATCACCAAGTGTATCGACGAC



dot plot



porovnávání textových řetězců

data: TCATG a CATTG

T C A T G

: :

C A T T G

T C A - T G

: : : :

. C A T T G

T C A T G .

: : :

. C A T T G

T C A T - G

: : : :

. C A T T G

pairwise alignment

formalizace problému

Vezměme dvě sekvence \mathbf{a}, \mathbf{b} (nukleotidové či aminokyselinové) délky m, n :

$$\begin{aligned}\mathbf{a} &= a_1, a_2, a_3 \dots a_m \\ \mathbf{b} &= b_1, b_2, b_3 \dots b_n\end{aligned}$$

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

$$A = (a_i \ b_j) \dots (a_k \ b_l) \quad \text{kde } 1 \leq i < \dots < k \leq m, \\ 1 \leq j < \dots < l \leq n$$

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

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

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-

pairwise alignment

gaps - formalizace problému

Ohodnotíme mezery ("gaps") v alignmentu funkcí

$$w_x = y + zx \quad \text{pro } x \geq 0; \quad y, z \leq 0$$

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:

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

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	2	2	2	-1	-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	2	2	2	-1	-1	-1
A	-1	-1	2	-1	-1	-1	-1	-1	-1	2	2	2	-1	-1	-1
A	-1	-1	2	-1	-1	-1	-1	-1	-1	2	2	2	-1	-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

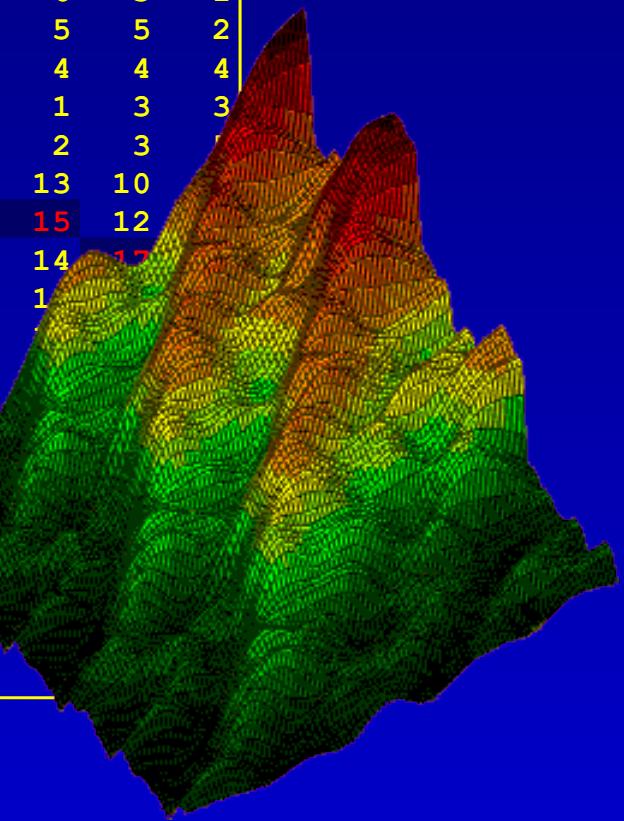
: : : : : : : : : :

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



pairwise alignment - možné problémy

jana.- ---TGCG-TTG--GAATCGACCCCGATGCCGTCTGACCACGTAGCTCATGCCGGTGCAGAGCTTGAGGTCGGCGTGGCGGAGGTGAGGGCGAG
::: :::: : :::::: :: : :::: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: :

Contig CCACCCCGATTGCCGTCTGCCACACG-TAGCTCATGCCGGTGCACAGCT--TGAGGTCG---GCGTGAG--CCGCGGAGGTGAGGGAGAGAGTGGTGAC
44680 44690 44700 44710 44720 44730 44740 44750 44760

jana.- GGTGGTGACAATGGCG-GCGAGGGC--GCGCGATCTCCACCCCGTCATGCCCGG---CCTTGTGCCGGCATCCACGTCTTGCTGAGAAAACCGCCGGAA
::: :: :

Contig AATGGTGGCGAGGGCGCGTGTATCCAGCCGTATGCCCGGCCCTTGCTGCCGGCATCCACGTCTTGCTGATCCACGTCTTGCTGAGGAAACCGCCGGAA
44770 44780 44790 44800 44810 44820 44830 44840 44850 44860

jana.- AGACGTGGATGGCCGGGACGAGCCCCGGCCATGACGGATTTGTCGCCGCAGCCTCGCATCACTTGTGGATCAGCGTGCCGGTGCCTGGTTGGTGAACAA
::: :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: :

Contig AGACGTGGATGGCCGGGACGAGCCCCGGCCATGACGGATTTATGCCGCCGCCGCATCACTTGTGGATCAGCGTGCCGGTGCCTGGTTGGTGAACAG
44870 44880 44890 44900 44910 44920 44930 44940 44950 44960

jana.s CTCATGCCGGTGCAGAGCTTGAGGTCGGCGTGGCGCGGAGGTGAGGGCGAGGGTGGTACAATGGCGGAGGGCGCGATCTCCACCCGTATGC
::: :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: :

Contig CTCATGCCGGTGCACAGCTTGAGGTCGGCGTGAGCCCGGAGGTGAGGGAGAGAGTGGTACAATGGTGGCGAGGGCGCGTGTATCTCCAGCCGTATGC
44710 44720 44730 44740 44750 44760 44770 44780 44790 44800

jana.s CCGGCCTTGTGCCGGCAT-----CCACGTCTTGCTGAGAAAACGCCGAAAGACGTGGATGCCGGACGAGCCGGCCATGACGGAT
::: :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: :

Contig CCGGCCTTGTGCCGGCATCCACGTCTTGCTGATCCACGTCTTGCTGAGGAAACGCCGAAAGACGTGGATGCCGGACGAGCCGGCCATGACGGAT
44810 44820 44830 44840 44850 44860 44870 44880 44890 44900

jana.s GTTGTGCCCGAGCCTCGCATCACTTGAGTCAAGCGTGCCGGTGCCTGGTTGGTAACAATTGAGCAGCACTGCGTGCGGACCTTGCCGTGAGGA
::: :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: : :: :

Contig GTTATGCCGCCGCCCGCATCACTTGAGTCAAGCGTGCCGGTGCCTGGTTGGTAACAGTCGAGCAGCACTGCGTGCGGACCTTGCCGTGAGGA
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

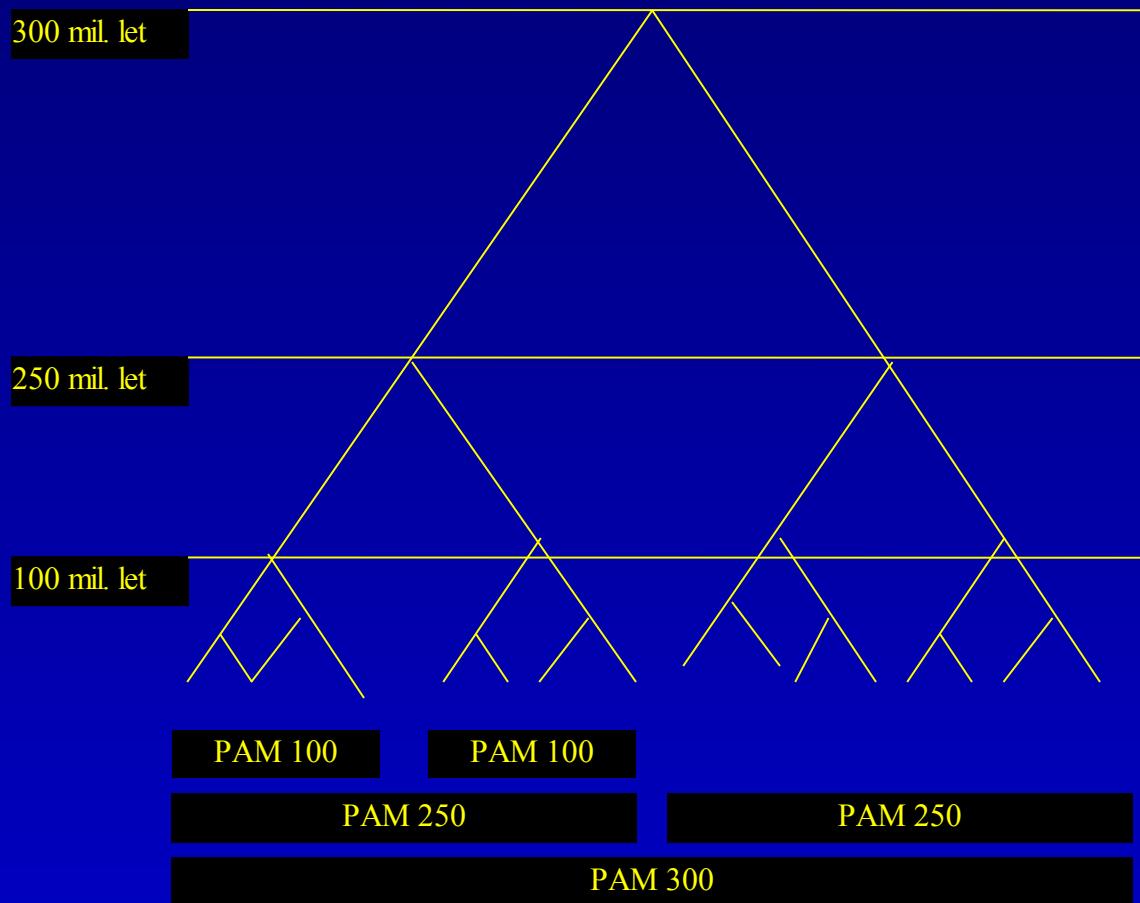
Princip: proteiny se vyvíjejí pomocí nezávislých mutací a jsou fixovány postupně

PAM (Percent Accepted Mutation)

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

BLOSSUM

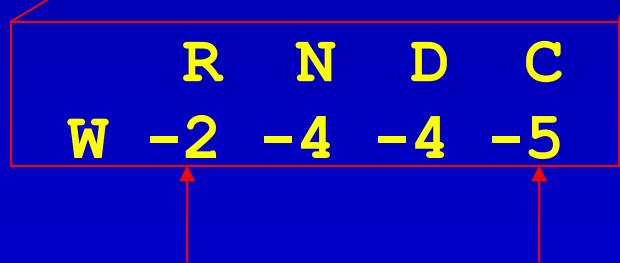
substitution matrix



substitution matrix

BLOSUM 45

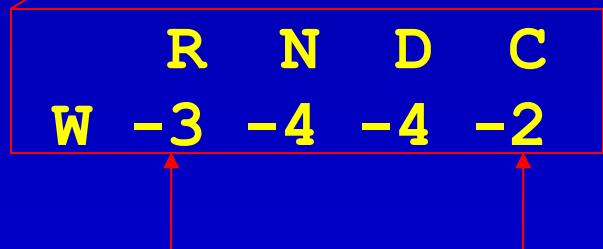
#	Entropy	=	0.3795,	Expected	=	-0.2789	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	5	-2	-1	-2	-1	-1	0	-2	-1	-1	-1	-1	-1	-2	-1	1	0	-2	-2	0	-1	-1	0	-5						
R	-2	7	0	-1	-3	1	0	-2	0	-3	-2	3	-1	-2	-2	-1	-1	-2	-1	-2	-1	0	-1	-5						
N	-1	0	6	2	-2	0	0	0	1	-2	-3	0	-2	-2	-2	1	0	-4	-2	-3	4	0	-1	-5						
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	-2	-3	-2	-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	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	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4	-4		
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-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	-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	-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	-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	-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	-2	0	0	-2	-1	-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	-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 páru 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 Smith-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, komplikace 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/fasta3/ Search Print

Home Bookmarks Internet Lookup New&Cool

Gel Nucleotide sequences for Go Site search Go ?

EMBL-EBI European Bioinformatics Institute

Site Map SRS Start Session

EBI Home About EBI Research Services Toolbox Databases Downloads Submissions HOMOLOGY & SIMILARITY

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 [lastz](#) programs. [Details about this service](#).

YOUR EMAIL	SEARCH TITLE	RESULTS	PROGRAM	DATABASES		
<input type="text"/>	Sequence	interactive	fasta3 fastx3 fasty3	Protein swall swiss-prot		
GAP PENALTIES	SCORES & ALIGNMENTS	KTUP / HISTOGRAM	DNA STRAND	MATRIX		
OPEN RESIDUE	-12 -2	SCORES ALIGN	50 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 - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop http://www.ebi.ac.uk/servicestmp/968509.164858-14326.html Search Print

Home Bookmarks Internet Lookup New&Cool

Fasta Results of Search:

Results of search

Database	+swall+
Title	Sequence
SeqLen	349

[View using Mview](#) [VisualFasta](#) [SUBMIT ANOTHER JOB](#)

FASTA searches a protein or DNA sequence data bank
version 3.3t09 May 18, 2001
Please cite:
W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

0:1:- 349 aa
PAX4_MOUSE P32115 Paired box protein Pax-4
vs SWISS-PROT All library
searching /ebi/services/idata/v222/fastadb/swall library

289642739 residues in 912109 sequences
statistics extrapolated from 60000 to 909913 sequences
Expectation n fit: rho(ln(x))= 5.6400+/-0.000219; mu= 7.0865+/- 0.013
mean_var=126_3356+/-27.355, 0's: 154 z-trim: 288 B-trim: 3158 in 2/63
Lambda= 0.1141

FASTA (3.39 May 2001) function [optimized, BL50 matrix (15:-5)] ktup: 2
join: 37, opt: 25, gap-pen: -12/-2, width: 16
Scan time: 10.920

The best scores are:

	opt	bits	E(909913)
SWALL:PAX4_MOUSE P32115 Paired box protein Pax-4.	(349)	2439	413 6.7e-114
SWALL:PAX4_RAT Q88436 Paired box protein Pax-4.	(349)	2238	380 6.2e-104
SWALL:PAX4_HUMAN O43316 Paired box protein Pax-4.	(350)	1885	321 1.9e-86
SWALL:GSBD_DROME P09082 Gooseberry distal protein	(427)	683	124 8e-27
SWALL:Q57684 Q57684 Paired box protein.	(354)	634	116 1.9e-24
SWALL:Q57676 Q57676 Paired box protein.	(385)	633	115 2.3e-24
SWALL:Q9WUWS Q9WUWS GSB-N protein (RE64348p).	(449)	631	115 3.1e-24
SWALL:PAX6_ORYLA Q73917 Paired box protein Pax-6.	(437)	627	114 4.9e-24
SWALL:Q5BP_DROME P09083 Gooseberry proximal prote	(449)	626	114 5.6e-24
SWALL:Q42612 Q42612 Pax-6 protein.	(437)	623	114 7.7e-24
SWALL:Q9YH28 Q9YH28 Pax-family transcription fact	(437)	622	114 8.6e-24
SWALL:PAX6_BRARE P26630 Paired box protein Pax[Zf	(437)	620	113 1.1e-23
SWALL:O61991 O61991 Pax6 protein.	(462)	619	113 1.3e-23
SWALL:O61990 O61990 Pax6 protein.	(439)	618	113 1.4e-23
SWALL:P70002 P70002 Xenopus Pax-6 short (Fragment	(370)	616	113 1.5e-23
SWALL:Q9IAST Q9IAST Paired domain transcription f	(393)	616	113 1.6e-23
SWALL:P70001 P70001 Xenopus Pax-6 long (Fragment)	(421)	616	113 1.7e-23

Transferring data from www.ebi.ac.uk...

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)

PAX4_M 10 20 30 40
MQQDGGLSSVNQLGGLFVNNGRPLPLDTRQQIVQLAIRGMRPCDISRS
: . :::::::::::::::::::: : ::::::: . ::::: :::
SWALL: MAVSALNMTPYFGGYPFQGQGRVNQLGGVFINGRPLPNHIRQIVEMAAAGVRPCVISRQ
10 20 30 40 50 60

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

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

PAX4_M 170 180 190 200 210 220
HPGTSHRNRTIFSPGQAEALEKEFQRGQYPDSVARGKLAAATSLPEDTVRVWFNSNRRAKW
.
SWALL: LKRKQRRSRRTFSNDQIDALERIFARTQYPDVYTREELAQSTGLTEARVQVWFNSNRARL
190 200 210 220 230 240

FastA - poznámky

Zvýšením k-tuple se zvýší rychlosť, ale sníží senzitivitu.

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 DNAb

blastp AA x AAdb

blastx AA x DNAb přeloženou do AA v šesti možných framech

tblastn DNA x AAdb

tblastx DNA x DNAb 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 PubMed Entrez BLAST OMIM Taxonomy Structure

BLAST

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

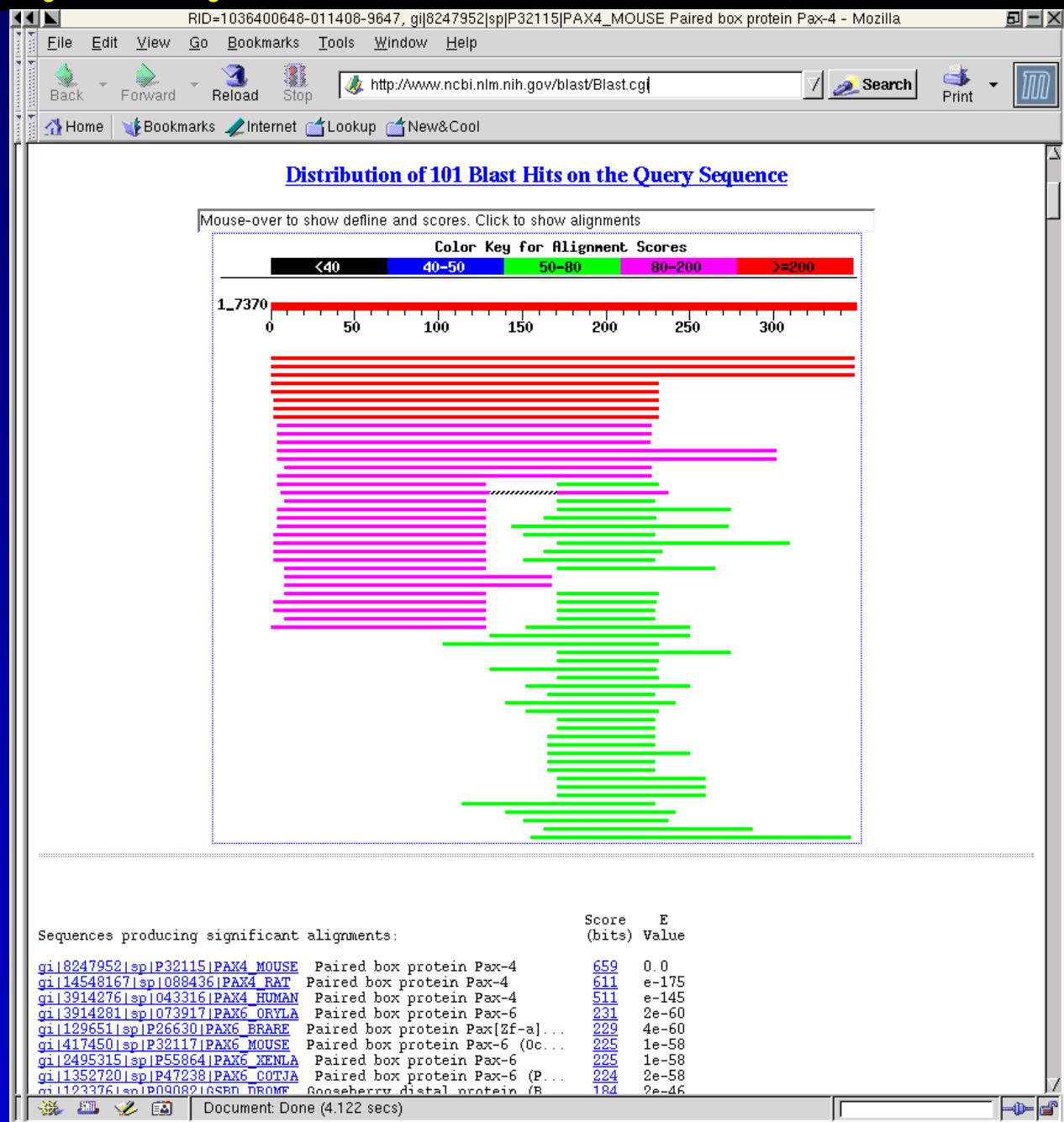
Word Size 3

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

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

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

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

Query: 199 ARGKLAATSLPEDTVRVWFSNRRAKWRRQEKLK 232
          AR +LAA    LPE  ++VWF SNRRAKWR+EK+L+
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

Netscape: PSI BLAST Search

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Sh N

Bookmarks Location: <http://www.ncbi.nlm.nih.gov>

NCBI **Ψ - BLAST** Entrez

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zhe Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". *Nature Biotechnology* 15:3389-3402.

Database: nr

Enter here your **amino acid sequence** as Sequence in FASTA format

Please read about [FASTA](#) format description

Advanced options for the BLAST server:

Expect: 10 Filter Low complexity NCBI-gi Graphic Overview

Alignment view: Pairwise

Composition-based statistics

Descriptions: 500 Alignments: 500

Expect value for inclusion in PSI-BLAST iteration 1: 0.002

Matrix: gap existence cost per residue gap cost Lambda ratio

Matrix	gap existence cost	per residue gap cost	Lambda ratio
PAM30	9	1	0.87
PAM70	10	1	0.87
BLOSUM80	10	1	0.87
BLOSUM62	11	1	0.85 default
BLOSUM45	14	2	0.87

Other advanced options:

Submit Query Reset

Comments and suggestions to: <blast-help@ncbi.nlm.nih.gov>
Credits to: Tom Madden, Sergei B. Shavrin, Alejandro Schäffer, and Alexey Egorov

PSI-BLAST - výsledky

Netscape: BLAST Search Results

File Edit View Go Communicator Help

Back Forward Reload Bookmarks Location NCBI Help

NCBI

BLASTP 2.1.1 [Aug-8-2000]

Reference:
Altschul, Stephen F., Thomas L. Ma Jinhui Zhang, Zheng Zhang, Webb M. "Gapped BLAST and PSI-BLAST: a new programs", Nucleic Acids Res. 25: 3389-3402.

query= gi|6679213|ref|NP_032808.1|musculus| (391 letters)

Database: nr
576,719 sequences; 181,300 alignments

E-value threshold for inclusion in

E-value threshold for inclusion in

Dist

Mouse-over to show sequence

<40

tmpseq_1

Sequences producing significant alignments:

- ref|NP_032808.1| paired box gene 5 >gi|400...
- ref|NP_057953.1| paired box 5; B-cell lineage...
- dbj|BAR76951.1| (AB004249) Pax-5 [Gallus ga...
- emb|CAR09230.1| (AJ010503) paired box protein...
- dbj|BAR88987.1| (AB026496) paired-box cont...
- emb|CAR71205.1| (Y10119) paired box protein...
- gb|RAD19296.1| (AF067541) paired box protein...
- gb|RAC34300.1| (AF072555) transcription fac...

Run PSI-Blast iteration 1

Netscape: BLAST Search Results

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Help

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

pir|IT112137 hypothetical protein Y35T12.1 - Caenorhabditis elegans | 80 4e-14

pir|IT16393 hypothetical protein F48B9.5 - Caenorhabditis elegans | 79 8e-14

qb|RAC48777.1| (U73621) Pax-6 [Bos taurus] | 76 5e-13

pir|IS36166 paired box transcription factor Pax-6 - rat (fragment) | 75 1e-12

qb|RAB30588.1| PAX6 product (exon 5, paired box) [human, Homo sapiens] | 72 1e-11

emb|CRA10461.1| (AJ131630) Eyeless protein [Drosophila melanogaster] | 71 3e-11

pir|IPC4435 paired box transcription factor Pax-6 splice form | 53 4e-06

qb|RAD16227.1| (AF098329) eyeless protein [Drosophila virilis] | 52 1e-05

qb|RAC70885.1| (U23518) Contains similarity to Pfam domain: P...

pir|IT24209 hypothetical protein R13.2 - Caenorhabditis elegans | 51 2e-05

qb|RAB84188.1| (AF027769) paired-box protein Pax-2 [Xenopus laevis] | 50 5e-05

pir|IT19154 hypothetical protein C09G9.7 - Caenorhabditis elegans | 47 3e-04

pir|IPC4433 paired box transcription factor Pax-6 splice form | 45 0.001

Run PSI-Blast iteration 1

Sequences with E-value WORSE than threshold

- pir|IT26169 hypothetical protein W0465.1 - Caenorhabditis elegans | 39 0.070
- pir|IT21438 hypothetical protein F26M9.3 - Caenorhabditis elegans | 39 0.070
- pir|IT33011 probable transposon protein K03H6.3 - Caenorhabditis elegans | 39 0.070
- qb|RAB27469.1| paired box Pax-3 gene product [chickens, embryonic tissue] | 38 0.20
- pir|IT52812 gene Pax-3 protein - mouse (fragment) >gi|239201|... | 38 0.20
- pir|IA26332 homeotic protein BSM4 - fruit fly (Drosophila melanogaster) | 37 0.33
- pdb|1FJL|A Chain A, Homeodomain From The Drosophila Paired Protein | 37 0.36
- pir|IT19530 hypothetical protein C27H2.1 - Caenorhabditis elegans | 37 0.45
- dbj|BAR885138.1| (AB030471) PAX-6 [Oryzias latipes] >gi|600996|... | 36 0.54
- qb|RAA03627.1| (U02308) PAX-3-FKHR gene fusion [Homo sapiens] | 36 0.79
- qb|RAA80574.1| (U12259) paired box homeotic protein [Homo sapiens] | 36 0.93
- pir|IPC4434 paired box transcription factor Pax-6 splice form | 36 0.97
- pir|IA45452 transcription factor PAX3 - human (fragments) | 35 1.3
- qb|RAA03628.1| (U02309) PAX-3 [Homo sapiens] | 35 1.5

PSI-BLAST - výsledky

Netscape: BLAST Search Results

File Edit View Go Communicator Help

Back Forward Reload

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

NCBI

BLASTP 2.1.1 [Aug-8-2000]

Reference: Altschul, Stephen F., Thomas Jinhui Zhang, Zheng Zhang, "Gapped BLAST and PSI-BLAST: programs", Nucleic Acids Research, 25, 3389-3402 (2000)

Query= gi|6679213|ref|NP_032 musculus] (391 letters)

Database: nr
576,719 sequences

E-value threshold for inclusion

E-value threshold for inclusion

Score = 71.0 bits (177), Expect = 0.010
Identities = 27/125 (21%), Positives = 41/125 (32%), Gaps = 29/125 (23%)
Query: 281 LTSPTPADIGSSVPGPQSYPIVTGRDLASTLPGYPHPH-----PPRAQGGS 326
+SP S + P+ I L + P P+ G
Sbjct: 1432 GSSPQPTPYSSMMLSPQQAIDNTYGLMFSPSVTMSPSGGFMGMGMGMGDIWPSSPGP 1491
Query: 327 Y-----SAPLTGWMWPGSEFSGSPYSHPQYSSYNSDWRFPNGLLLGSPPYYSPARARGA 380
+ P G S SP S P YS + ++ +P YSPR+ +
Sbjct: 1492 FGYIPVPGSPARAGYSPPSPRAY-SPAS-PAYSPPSPSPASPA-----YSPASPPAYS 1542
Query: 381 PPARY 385
P + A
Sbjct: 1543 PASPA 1547

>ref|NP_003856_1| homeo box (expressed in ES cells) 1; Homeo box gene expressed in ES cells; Rathke pouch homeo box
obj|BAB88275_1| (U82811) HANF [Homo sapiens]
obj|BAC24523_1| (AF059734) homeodomain transcription factor [Homo sapiens]
Length = 185

Score = 51.0 bits (121), Expect = 3e-05
Identities = 34/174 (19%), Positives = 55/174 (31%), Gaps = 24/174 (13%)
Query: 119 LAERVCNDNTVPSVS--SINRIIR---TKVQOPPNQPVPAASSHSIVSTGWSVTQVSSVSTD 173
L E + PS SI RI+ K P +P + + S+G + + +
Sbjct: 5 LQE6AQOLGENKPSTCSFSSTERILGLDQKKDGPPLMKPMPHRFWADTCSSSGKDGNLCLHVPN 64
Query: 174 SAGGSSYISGILSITSPSADTNKRKDRDEGIQESPVPHNGMSLPGRDPLRKQMR-----G 226
S S ++ P + + S + H S R L++++
Sbjct: 65 PP-SGISFFPSVVDHPMP-----EERASKVENYFSSAERLSLKRELWSVYRGRRPR 112
Query: 227 DLFTQQOLELDDNVFERQVSDIFTTPEPIKPEQTTEYSMASLAQGLDDMMVN 280
FTQ Q+EVL+ VF Y I + + E + + K +
Sbjct: 113 TAFTONICIEVLENWFRWNCPGIDIREDLAOKLNLEEDRIQIWQNERAKLRS 166

>obj|BAB12227_1| (AB017184) RNA polymerase II largest subunit [Aspergillus oryzae]
Length = 1748

Score = 50.6 bits (120), Expect = 3e-05
Identities = 29/134 (21%), Positives = 44/134 (32%), Gaps = 10/134 (7%)
Query: 251 TTPEPIKPEQTTTEYSMASLRRGGGLDDMKRHLTSPTPADIGSSVPGPQSYPIVTGRDLAST 310
T+ + S ++ M + SP P P AS
Sbjct: 1576 TSPGVSPSSSYSPSPGNTSPRFSMTSPSPSPSPFAP--TSPPSPSPSPAYGQASP 1633
Query: 311 TLPGYPPWPPRGQGYSYSAPLITGWMWPGSEFS6SPYSHPQYSSYNSDWRFPNGLLLGSPPY 370
T P Y P P +PT P S S SP S P +S + S+ +P + G+
Sbjct: 1634 TSPPSPSPGF----SPTSPHYSPTSP-SFSPAS-PAFSPSPSPSPSPATGGRAR 1685
Query: 371 YVSPPARGRAPPAA 384
+ SP + +
Sbjct: 1686 HLSPTSPTPKXTP 1699

Score = 47.9 bits (113), Expect = 2e-04
Identities = 49/290 (16%), Positives = 76/290 (25%), Gaps = 63/290 (21%)
Query: 115 RDRLLERVCNDNTVPSVSSINRIIRTKVQOPPNQPV-PASSHSIVSTGWSVTQVSSVSTD 173
+D ++++ S N I T + P+ R + S
Sbjct: 1495 KDRATISDGRTSYDTGSPMQDNAYIGTPDPDESHFSPIRQAGRESPGGFTEYOPTGGFGGG 1554
...

Netscape: BLAST Search Results

File Edit View Go Communicator Help

Back Forward Reload

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

Search Results

Help

Search Netscape Print Security Stop

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

ophila melan... 186 2e-46
>gi|1052601|... 184 1e-45
culus] 184 1e-45
la melanogas... 183 3e-45
gaster] 182 7e-45
us laevis] 181 2e-44
soleta] 166 3e-40
 166 5e-40
/9 [Paracent... 165 7e-40
a obsoleta] 165 8e-40
ditis elegan... 154 2e-39
 157 2e-37
sapiens] 156 4e-37
ntrotus livi... 155 7e-37
 147 2e-34
bditis elega... 143 3e-33
ents) 142 8e-33
bditis elega... 140 2e-32
itidis elegans... 140 3e-32
[Homo sapiens] 137 2e-31
abdritis eleg... 128 7e-29
in [Herdmani... 123 5e-27
factor [huma... 120 3e-26
 112 5e-24
eastern newt... 109 5e-23
. [Paracentro... 108 9e-23
 107 2e-22
in TPAX6 [Tr... 105 6e-22
 97 3e-19

920K

Mouse-over

tmpseq_1

0

920K

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]

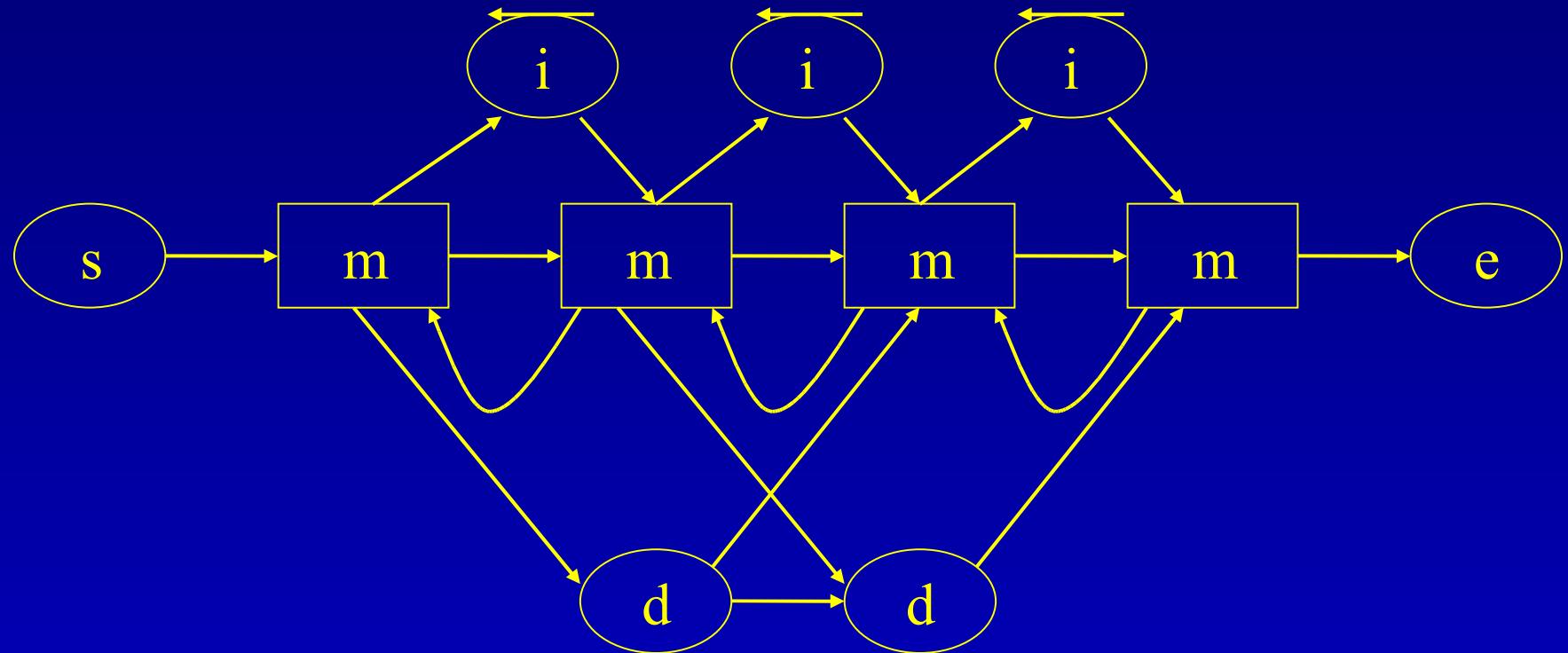
PHI-BLAST - princip

www: <http://www.ncbi.nlm.nih.gov/blast/>
[/psiblast.cgi?Jform=1](http://www.ncbi.nlm.nih.gov/blast/psiblast.cgi?Jform=1)

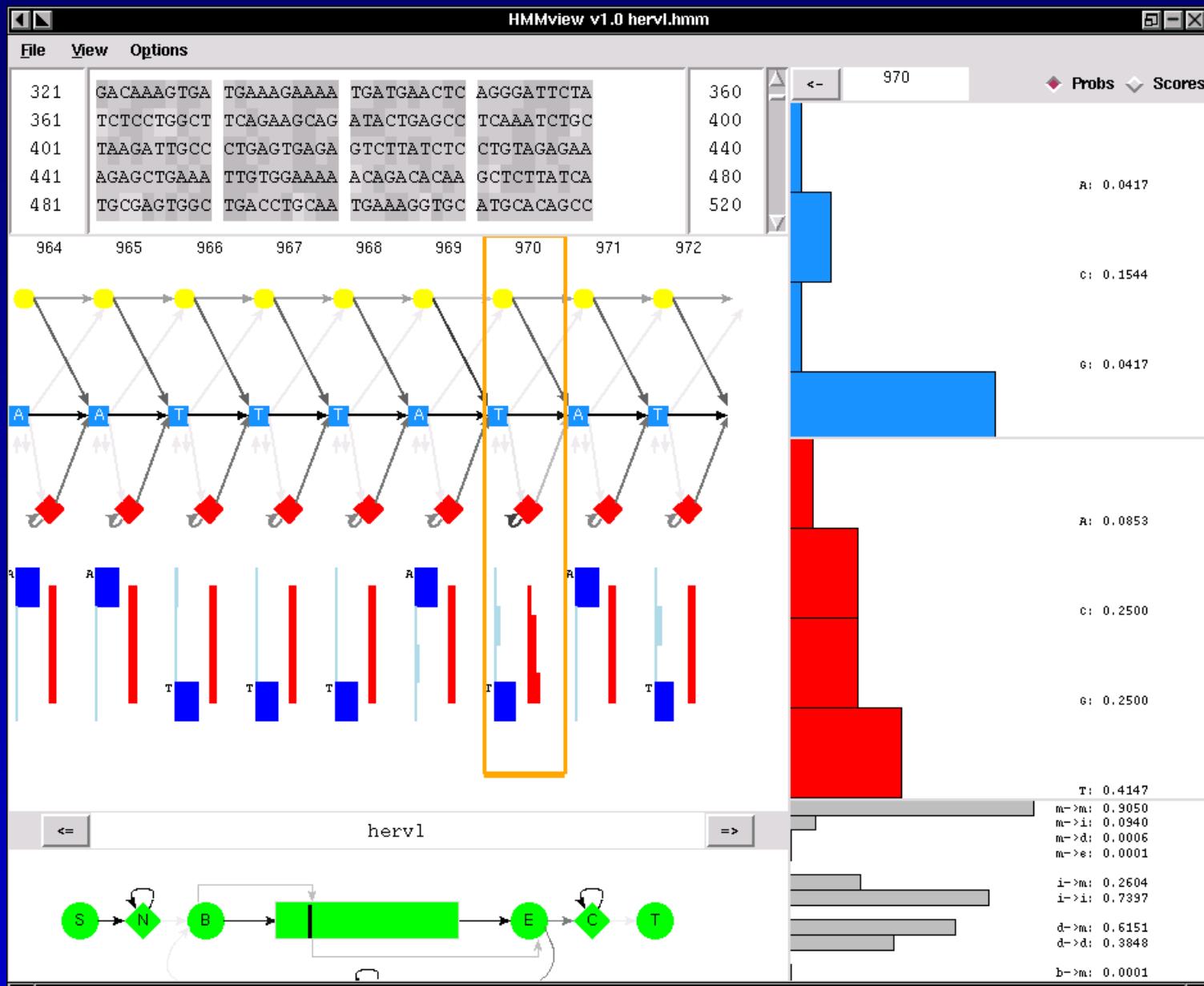
reference:

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, komplikace pod UNIXy bez problémů)

www:

<http://pfam.wustl.edu>

programy:

hmmsearch

hmmerpfam

prohledává modelem (hmmerem) databázi sekvencí

prohledává sekvencí databázi modelů