

# Posuzování podobnosti sekvencí

Nástroje pro párové přiložení (pairwise alignment) a vyhledávání lokálních podobností sekvencí

# Hledání v databázích

- Textové vyhledávání příbuzných sekvencí v databázích
  - Neefektivní - chybí anotace řady sekvencí
- Prohledávání databází podle podobnosti sekvencí
  - Výpočet lokálního přiložení (alignment)  
= uspořádání do 2 pod sebou ležících řádků tak, aby identické zbytky ležely pod sebou
  - Identifikace podobnosti a evoluční vzdálenosti

# Nástroje pro vyhledávání lokálních podobností sekvencí

Sady programů zahrnujících algoritmy pro vyhledávání podobnosti v dostupných databázích sekvencí bez ohledu na to zdali dotazovaná sekvence je **DNA** nebo **protein**.

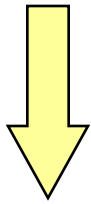
- BLAST
- Altschul et al., [1990](#)
- dostupný na serveru NCBI
- FASTA
- Lipman a Pearson [1985](#)
- dostupný na serveru EBI

# Princip hledání podobnosti

- Sekvence jsou tvořeny symboly abecedy
- Komplexita sekvence je určena počtem různých znaků, které se mohou vyskytovat v sekvenci (DNA = 4, proteiny = 20)
- Algoritmy využívají heuristickou analýzu pro identifikaci krátkých homologických subsekvencí bez mezer s následným rozšiřováním vyhledávání v okolí subsekvencí s cílem získat lokálně uspořádané sekvence, do nichž mohou být vloženy mezery tak, aby přiložení bylo optimální

# Co je to BLAST?

- **Basic Local Alignment Search Tool**
  - Hledání lokálních podobností
  - Heuristický přístup založený na Smith-Watermanově algoritmu
  - Vyhledá neoptimálnější **přiložení sekvencí**
  - Poskytuje data o statistické významnosti
  - Zobrazuje vzájemně párové přiložení sekvencí
  - Lokalizuje oblasti sekvencí s vysokou podobností a umožňuje zobrazení jejich primární struktury a funkce



# Výchozí stránka BLAST

**BLAST** Basic Local Alignment Search Tool

My NCBI [Sign In] [Register]

Home Recent Results Saved Strategies Help

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

[Learn more](#) about how to use the new BLAST design

### BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

- [Human](#)
- [Mouse](#)
- [Rat](#)
- [Arabidopsis thaliana](#)
- [Oryza sativa](#)
- [Bos taurus](#)
- [Danio rerio](#)
- [Drosophila melanogaster](#)
- [Gallus gallus](#)
- [Pan troglodytes](#)
- [Microbes](#)
- [Apis mellifera](#)

### Basic BLAST

Choose a BLAST program to run.

- [nucleotide blast](#) Search a **nucleotide** database using a **nucleotide** query  
*Algorithms:* blastn, megablast, discontinuous megablast
- [protein blast](#) Search **protein** database using a **protein** query  
*Algorithms:* blastp, psi-blast, phi-blast
- [blastx](#) Search **protein** database using a **translated nucleotide** query
- [tblastn](#) Search **translated nucleotide** database using a **protein** query
- [tblastx](#) Search **translated nucleotide** database using a **translated nucleotide** query

### Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Search [trace archives](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)
- Search sequences that have [gene expression profiles](#) (GEO)
- Search [immunoglobulins](#) (IgBLAST)
- Search for [SNPs](#) (snp)

### News

[New Human and Mouse pre-indexed databases](#)  
Human and mouse genomic + transcript megablast searches now use a faster, indexed algorithm that typically reduces run time by two thirds, as compared with standard megablast.  
2007-09-04 10:55:00  
[More BLAST news...](#)

### Tip of the Day

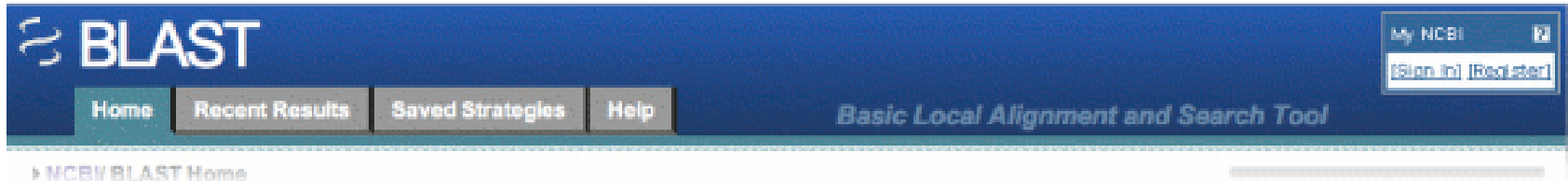
**Using Genomic BLAST**

Genomic BLAST pages are helpful because they allow the genomic context of a BLAST search to be displayed in the Map Viewer. For example, discontinuous (cross-species) MegaBLAST against the human RefSeq transcript for albumin (NM\_000477) can be used to identify the homolog in the rat genome.

[More tips...](#)

<http://www.ncbi.nlm.nih.gov/BLAST>

# Uživatelské rozhraní BLAST



- [Home Tab](#): Odkaz na úvodní stránku
- [Recent Results Tab](#): Odkaz na výsledky, které jste získali za posledních 36 hodin
- [Saved Strategies Tab](#): Vyplněné vstupní formuláře pro hledání, které jste uložili do *MyNCBI*
- [Help Tab](#): Katalog s dokumentací a nápovědou

NCBI/BLAST Home

BLAST finds regions of similarity between biological sequences

Learn more about how to use the new BLAST design

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST

- Human, Mouse, Rat, Arabidopsis thaliana, Oryza sativa, Bos taurus, Danio rerio, Drosophila melanogaster

Basic BLAST

Choose a BLAST program to run.

- nucleotide blast, protein blast, blastx, tblastn, tblastx

Specialized BLAST

Choose a type of specialized search (or database name in parentheses)

- Search trace archives, Find conserved domains, Find sequences with similar conserved domain architecture, Search sequences that have gene expression profiles, Search immunoglobulins, Search for SNPs

- BLAST Overview, FAQs, News, Manual, References, Retrieve results

Genome Project

BLAST Drosophila melanogaster Sequences.

Enter an accession, gi, or a sequence in FASTA format:

Text input field for sequence or accession

Or, choose a file to upload

File upload field with Procházet.. button

Set subsequence: (optional)

From: To: range selection fields

Database:

genome (reference only) 6 sequences

Program:

megaBLAST: Compare highly related nucleotide sequences

Optional parameters

Expect: 0.01, Filter: default, Descriptions: 100, Alignments: 100

Advanced options: text input field

Begin Search, Clear Input buttons

Get the URL with preset values? Get URL button





# Basic BLAST – výběr programů

## Využití jednotlivých programů BLAST

Program	Dotaz	Databáze	Úroveň srovnání	Použití
<a href="#"><u>blastn</u></a>	DNA	DNA	DNA	Hledání identických sekvencí DNA
<a href="#"><u>blasp</u></a>	Protein	Protein	Protein	Hledání homologních proteinů
<a href="#"><u>blastx</u></a>	DNA	Protein	Protein	Hledání genů a homologních proteinů na DNA
<a href="#"><u>tblastn</u></a>	Protein	DNA	Protein	Hledání genů u necharakterizovaných DNA
<a href="#"><u>tblastx</u></a>	DNA	DNA	Protein	Studium struktury genů

# Příklady využití algoritmů BLAST

Volba programu, jestliže Vaše sekvence je NUKLEOTIDOVÁ			
Délka	Data-báze	Účel vyhledávání	BLAST Program
20 bp nebo delší	DNA	Identifikace dotazované sekvence	<a href="#">MEGABLAST</a> <a href="#">Standard BLAST</a> (blastn)
		Vyhledání podobných sekvencí jako dotazovaná	<a href="#">Standard BLAST</a> (blastn)
		Vyhledání podobných proteinů k překladu dotazované sekvence v přeložených databázích DNA	<a href="#">Translated BLAST</a> (tblastx)
	Protein	Vyhledání podobných proteinů k překladu dotazované sekvence v databázích proteinů	<a href="#">Translated BLAST</a> (blastx)
7 - 20 bp	DNA	Vyhledání vazebných míst primerů nebo krátkých motivů	<a href="#">Search for short, nearly exact matches</a>

# Příklady využití algoritmů BLAST

Volba programu, jestliže Vaše sekvence je PROTEIN			
Délka	Data-báze	Účel vyhledávání	BLAST program
15 aminokyselino-vých zbytků nebo delší	Protein	Identifikace dotazované sekvence nebo vyhledání sekvencí podobných proteinů	<a href="#">Standard Protein BLAST</a> (blastp)
		Vyhledání členů proteinové rodiny, tvorba vlastní pozičně-specifické matice a konstrukce profilu → profil je potom srovnán a lokálně přiřazen k sekvencím v proteinové databázi	<a href="#">PSI-BLAST</a>
		Vyhledání proteinů podobných dotazovanému v okolí určitého vzoru	<a href="#">PHI-BLAST</a>
	Konzervativní domény	Vyhledání konzervativních domén v dotazované sekvenci	<a href="#">CD-search</a> (RPS-BLAST)
	Konzervativní domény	Vyhledání konzervativních domén v dotazované sekvenci a identifikace ostatních proteinů s podobnou architekturou domén	<a href="#">Conserved Domain Architecture Retrieval Tool</a> (CDART)
	DNA	Vyhledání podobných proteinů v přeložených databázích DNA	<a href="#">Translated BLAST</a> (tblastn)
5-15 zbytků	Protein	Hledání peptidových motivů	<a href="#">Search for short, nearly exact matches</a>

# Jak používat BLAST?

- <http://blast.ncbi.nlm.nih.gov/Blast.cgi>
1. Vybrat příslušný BLAST-program (blastn, blastp, blastx, tblastn, tblastx, specializované varianty algoritmů)
  2. Vložit sekvenci (DNA nebo protein nebo Accession number)
  3. Vybrat databázi, která má být prohledána
  4. Upřesnit nastavení parametrů algoritmu
  5. Odeslat požadavek na vyhledání

# Vložení sekvence

**Enter Query Sequence**

Enter accession number, gi, or FASTA sequence [\[?\]](#) [Clear](#)

Query subrange [\[?\]](#)

From

To

Or, upload file  [Browse...](#) [\[?\]](#)

Job Title

Enter a descriptive title for your BLAST search [\[?\]](#)

>příklad sekvence

```
GAATTCTTCAAAAAAGTATTCGTTGGATACACGGACAGTGAAGATCATTGAGAGGATTCTGCAAGTTCGTTACCCAGCTAACCCCA
AAATGTTGAAGTAGCAGTTAATTCAAATCTGCAACAGTTTCAGCAGAATAGGGGCTTTCAAAATAAATCAAAGGAGAATAATTTAT
GACTAAAACTTTAAAGGTTTATAAAGGAGACGACGTCGTAGCTTCTGAACAAGGTGAAGGCAAAGTGTGAGTAACTTTATCTAATTT
AGAAGCGGATACAACCTTATCCAAAAGGTAATACCAAGTGGCATGGGAAGAAAATGGTAAAGAATCTAGTAAAGTTGATGTACCTCA
ATTCAAACCAATCCAATTCTAGTCTCAGGCGTATCATTTACACCCGAAACTAAATCAATCACGGTAAATGCTGATGACAATGTTGA
ACCAACATTGCACCAAGTACAGCAACGAATAAAACGTTGAAATATACAAGTGAACATCCAGAGTTTGTACTGTTGATGAGAGAAC
AGGAGCAATTCACGGTGTAGCTGAGGGAAGTTGAGTTATCACTGCTACGTCTACTGACGGAAGTGACAAGTCTGGACAAATTACAGT
AACAGTAACAAATGGATAATTTTGGAGACGCAGAATATCTGCGTCT
```

# Výběr databáze

**Choose Search Set**

**Database**  
 Human genomic + transcript  Mouse genomic + transcript  Others (nr etc.):  
Reference mRNA sequences (refseq\_rna) [?]

**Organism**  
Optional  
 Any  Human  *A.thaliana*  Mouse  Custom...

**Entrez Query**  
Optional

Search: plat  
Seznam výsledků:  
duck-billed platypus (taxid:9258) be shown.  
platypus (taxid:9258)  
duckbill platypus (taxid:9258)  
Platyhelminthes (taxid:6157)  
Platyrrhini (taxid:9479)  
Platichthys (taxid:8259)  
Platichthys flesus (taxid:8260)

- Others (nr etc.) = celá databáze (neredundantní nukleotidová nr/nt)


# Výběr podprogramu

Program selection

Optimize for

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm [\[?\]](#)



# Úprava parametrů algoritmu

## Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow

### General Parameters

Max target sequences

100

Select the maximum number of aligned sequences to display

Short queries

Automatically adjust parameters for short input sequences

Expect threshold

10

Word size

11

### Scoring Parameters

Match/Mismatch Scores

2,-3

Gap Costs

Existence: 5 Extension: 2

### Filters and Masking

Filter

Low complexity regions

Species-specific repeats for: Human

Mask

Mask for lookup table only

Mask lower case letters

**BLAST**

Search database nr using Blastn (Optimize for somewhat similar sequences)

Show results in a new window



# Jak BLAST pracuje?

- Proces zahrnuje 3 kroky
  1. Příprava dotazu
    - rozseká zkoumanou sekvenci na krátké úseky a sestaví z nich vhodnou tabulku
  2. Vyhledává shody v databázi
  3. Rozšiřuje vyhledávání v oblasti nalezených shod, tak aby byla splněna zadaná kritéria

# Slova pro nukleotidové sekvence

Dotaz: **GTACTGGACATGGACCCTACAGGAA**

~~GTACTGGACAT~~

Velikost slova = 11

minimální velikost = 7

**TACTGGACATG**

blastn default = 11

tabulka se všemi slovy dotazu  
slovy dotazu

**ACTGGACATGG**

megablast default = 28

**CTGGACATGGA**

**TGGACATGGAC**

**GGACATGGACC**

**GACATGGACCC**

**ACATGGACCCT**

• • • • • • • • • •

# Slova pro proteinové sekvence

Dotaz: **GTQITVEDLIFYNIATRRKALKN**

**GTQ**  
Velikost = 3

Velikost slova může být 2 nebo 3 (default = 3)

**TQI**

tabulka se všemi slovy dotazu

**QIT**

Sousedící slova

**ITV** → LTV, MTV, ISV, LSV, etc.

**TVE**

**VED**

**EDL**

**DLF**

• • •

# Minimální požadavek pro shodu

ATCGCCATGCTTAATTGGGCTT

CATGCTTAATT

přesná shoda slova

1 nalezená shoda

- Nucleotidový BLAST vyžaduje **jednu přesnou shodu**
- Proteinový BLAST vyžaduje **dvě sousedící shody v úseku 40 aa**

GTQITVEDLIFYNI

SEI

YYN

sousedící slova

2 nalezené shody

# přiložení sekvencí, které BLAST může nalézt

```
1 AATGGTAAAGACTACTGGATCATTAAAGAACTCCTGGGGAG
  ||||| ||||||||||||||||| || |||||||||||||
1 AATGGAAAAGACTACTGGATCATCAAAAACCTCCTGGGGAG
```

sekvence obsahují definovanou shodu slova

# přiložení sekvencí, které BLAST nemůže nalézt

1 GAATATATGAAGACCAAGATTGCAGTCCTGCTGGCCTGAACCACGCTATTCTTGCTGTTG

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

1 GAGTGTACGATGAGCCCGAGTGTAGCAGTGAAGATCTGGACCACGGTGTACTCGTTGTCTG

61 GTTACGGAACCGAGAATGGTAAAGACTACTGGATCATTAAAGAACTCCTGGGGAGCCAGTT

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

61 GCTATGGTGTTAAGGGTGGGAAGAAGTACTGGCTCGTCAAGAACAGCTGGGCTGAATCCT

121 GGGGTGAACAAGGTTATTTTCAGGCTTGCTCGTGGTAAAAC

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

121 GGGGAGACCAAGGCTACATCCTTATGTCCCGTGACAACAAC

# BLASTn - Možnosti nastavení

## Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow

### General Parameters

Max target sequences

100

Select the maximum number of aligned sequences to display

Short queries

Automatically adjust parameters for short input sequences

Expect threshold

10

Word size

11

7

11

15

### Scoring Parameters

Match/Mismatch Scores

2,-3

Gap Costs

Existence: 5 Extension: 2

### Filters and Masking

Filter

Low complexity regions

Species-specific repeats for: Human

Mask

Mask for lookup table only

Mask lower case letters

BLAST

Search database nr using Blast

Show results in a new window

(somewhat similar sequences)

Human  
Human  
Rodents  
Arabidopsis  
Rice  
Mammals  
Fungi  
C. elegans  
A. gambiae  
Zebrafish  
Fruit fly

# Proteinový BLAST

NCBI/ BLAST/ blastp suite: BLASTP programs search protein databases using a protein query. [more...](#)

[Reset page](#) [Bookmark](#)

## Enter Query Sequence

Enter accession number, gi, or FASTA sequence [?](#)

[Clear](#)

```
>3AORF1
MTKTLKVYKGGDDVASEQEGEGKVSVILSNLEADTTYPKGTQVAVEENGKSSKVDVDPQFKTNPILVSGVSF
TPETKSITVNADDNVEPNIA PSTATNKLKYTSEHPEFVTVDERTGAIHGVAEGTSVITATSTDGSDKSGQI
TVIVTNG
```

Query subrange [?](#)

From

To

Or, upload file

Procházet... [?](#)

Job Title

3AORF1

Enter a descriptive title for your BLAST search [?](#)

## Choose Search Set

Database

Swissprot protein sequences(swissprot) [?](#)

**Protein database**

Organism

Optional

Enter organism name or id--completions will be suggested

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

Entrez Query

Optional

Enter an Entrez query to limit search [?](#)

## Program Selection

Algorithm

- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm [?](#)

**BLAST**

Search database **swissprot** using **Blastp (protein-protein BLAST)**




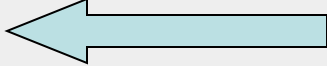

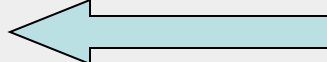



# BLASTp - Možnosti nastavení

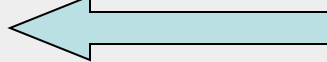



## Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow

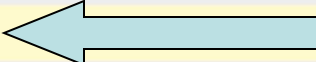



### General Parameters

- Max target sequences   Select the maximum number of aligned sequences to display 
- Short queries  Automatically adjust parameters for short input sequences 
- Expect threshold   
- Word size   

### Scoring Parameters

- Matrix   
- Gap Costs Existence: 11 Extension: 1 
- Compositional adjustments  

### Filters and Masking

- Filter  Low complexity regions  
- Mask  Mask for lookup table only   
 Mask lower case letters 

**BLAST**

Search database **swissprot** using **Blastp (protein-protein BLAST)**

Show results in a new window

# Typy matic pro výpočet skóre

- Matice identity
  - Především pro nukleotidové sekvence
  - Neschopné transformovat na jiné zbytky
  - Pro přiložení velmi podobných sekvencí
- Matice podobnosti
  - Používané u proteinových sekvencí
  - Vyjadřují biochemické/biologické vlastnosti aminokyselin
  - Vyšší účinnost při srovnávání sekvencí

# Matrice identity

	A	G	C	T
A	+1	-3	-3	-3
G	-3	+1	-3	-3
C	-3	-3	+1	-3
T	-3	-3	-3	+1

CAGGTAGCAAGCTTGCATGTCA

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

CACGTAGCAAGCTTG-GTGTCA

skóre = 19-9 = 10

# Penalizace mezer

Kvalitní (optimální) přiložení podobných sekvencí

- **maximalizuje** počet srovnatelných protějšků
- **minimalizuje** počet mezer

Povolení vložení mnohá náhodných mezer vede k dosažení přiložení nehomologních sekvencí s vysokými skóre

Penalizace mezer má za výsledek přiložení s relativně málo mezerami a nehomologní sekvence nejsou přiloženy

# Příklad penalizace mezer

Celkové skóre:

4

```
T A T G T G C G T A T A
| | | |
A T G T T A T A C
```

Celkové skóre:  $8 + (-3.2) = 4.8$

```
T A T G T G C G T A T A
| | | |           | | | |
A T G T - - - T A T A C
```

Gap parameters:

$d = 3$  (gap opening)

$e = 0.1$  (gap extension)

$g = 3$  (gap length)

$$\gamma(g) = -d - (g - 1) e = -3 - (3 - 1) 0.1 = -3.2$$

match	= 1
mismatch	= 0

# Substituční Matice

- Co je substituční matice?
  - Kompletní sada skóre pro všechny kombinace párů zbytků se nazývá substituční matice
  - Stanovuje frekvenci při které každý možný zbytek v sekvencích může být změněn za kterýkoli jiný zbytek během času (evoluce)
  - Např., hydrofobní zbytek má vyšší pravděpodobnost zachování v příslušné pozici sekvence než jiný.
  - Každá matice je určena pro určitý typ vyhledávání –  
**JE TŘEBA VĚDĚT CO HLEDÁME!**

# Substituční Matice

- Proč používat substituční matice?
  1. Stanovit pravděpodobnou homologii dvou proteinových sekvencí.
  2. Substituce, které jsou více pravděpodobné získají vyšší skóre
  3. Substituce, které jsou méně pravděpodobné obdrží nižší skóre.

# Matrice BLOSUM

- **Blocks Substitution Matrix**
- Změny probíhající během dlouhodobé evoluce nejsou často vhodné pro výpočty a sledování malých recentních změn
- Matrice BLOSUM jsou sestaveny na základě analýzy mnohonásobných příložení evolučně příbuzných proteinů v databázi BLOCKS
- BLOSUM-x používá analýzu pouze těch proteinů, které mají alespoň x % identitu



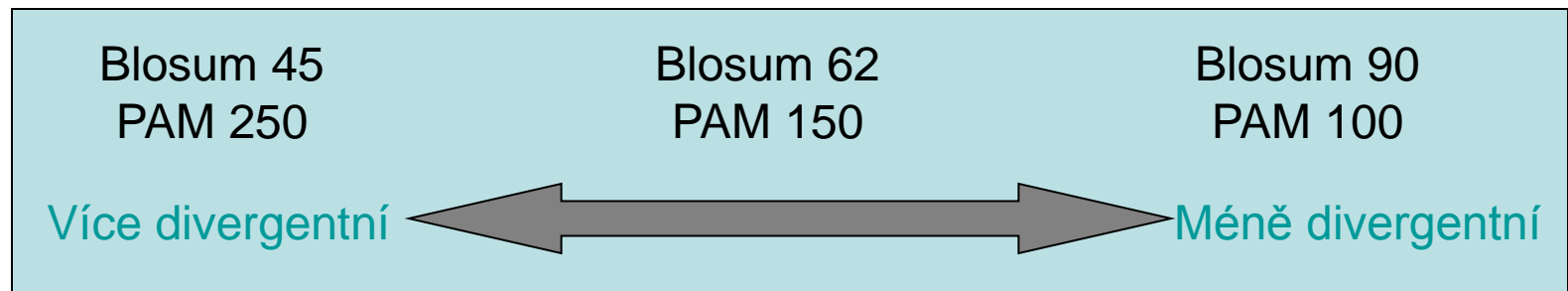


# Matrice PAM

- PAM
  - Percent Accepted Mutation
  - založeny na konceptu akceptovatelných bodových mutací za  $10^8$  let v globálních mnohonásobných příloženích
  - Stanoveny na základě výpočtů u blízce příbuzných proteinů s identitou  $> 85\%$
  - PAM1 reprezentuje 1% změn (1 mutace na 100 aminokyselinových zbytků)
  - $PAM250 = (PAM1)^{250}$

# PAM versus BLOSUM

- PAM Matice (Percent Accepted Mutation)
  - Odvozené z pozorování; malé množství srovnávaných dat
  - vhodné pro evoluční modely
  - Všechny výpočty vycházejí z PAM1
  - PAM250 je nejpoužívanější
- BLOSUM (BLOck SUBstitution Matrices)
  - Odvozené z pozorování; velké množství vysoce konzervovaných sekvencí (BLOCKS)
  - Každá matice odvozená samostatně podle definované procentuální identity
  - BLOSUM62 – výchozí matice pro BLAST



# Obecné závěry

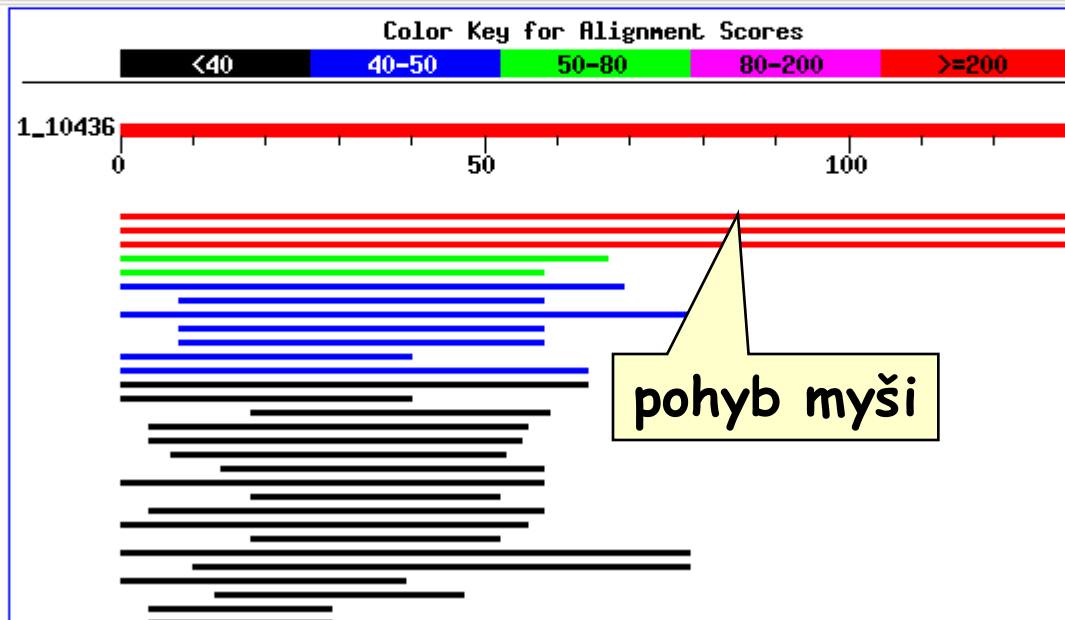
- Klíčovým elementem vyhodnocujícím výsledky srovnání aminokyselinových sekvencí je substituční matice
- Různé matice jsou přizpůsobené pro detekci podobností u sekvencí, které se vyznačují různým stupněm divergence
- BLOSUM je vhodnější pro lokální srovnání
  - BLOSUM-62 je optimální pro detekci nízkých podobností proteinů
  - BLOSUM-45 je vhodnější pro detekci nízkých podobností u dlouhých sekvencí

# BLAST – grafický výstup

[Taxonomy reports](#)

## Distribution of 30 Blast Hits on the Query Sequence

P40692 DNA mismatch repair protein Mlh1 (MutL protein homolog 1..S= 233 E=8e-62



# Statistika lokálního přiložení

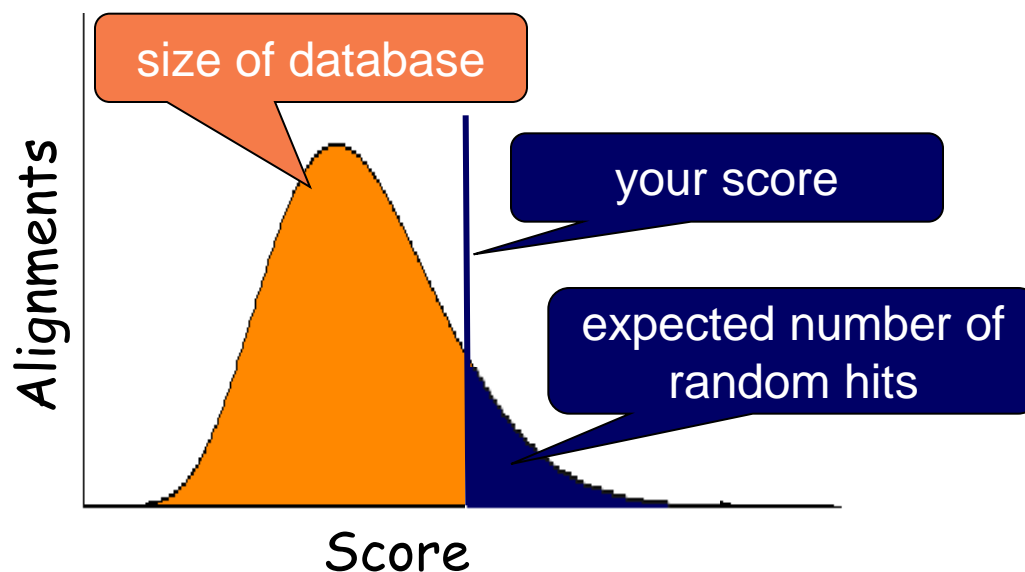
- K posouzení významnosti shody nalezených úseků se používá numerická hodnota označovaná jako **skóre sekvenčního přiložení (S)**
- Popisuje jeho celkovou kvalitu na základě porovnání pravděpodobnosti výskytu nalezených segmentů o určité sekvenční podobnosti s pravděpodobností, že se taková podobnost vyskytne mezi dvěma náhodnými sekvencemi
- Vyšší číslo odpovídá vyšší podobnosti
- Ekvivalentem skóre  $S$  je **hodnota E** („Expectation value“), která vyjadřuje počet různých sekvenčních přiložení se skórem shodným nebo vyšším než je hodnota  $S$ , jejíž výskyt je očekáván při náhodném vyhledávání v databázi.

# Výpočet hodnoty E (Expectation value)

E = počet nálezů v databázi, které mohou být ojeveny náhodně

$$E = mn 2^{-S}$$

Potom platí, že čím je hodnota E nižší, tím je skóre významnější.



# BLAST: popis výstupu

Sequences producing significant alignments	Score	E Value
<a href="#">gi 730028 sp P40692 MLH1 HUMAN</a> DNA mismatch repair protein ...	<a href="#">233</a>	8e-62
<a href="#">gi 13878583 sp Q9JK91 MLH1 MOUSE</a> DNA mismatch repair protein ...	<a href="#">214</a>	4e-56
<a href="#">gi 13878571 sp P97679 MLH1 RAT</a> DNA mismatch repair protein ...	<a href="#">212</a>	1e-55
<a href="#">gi 1709056 sp P38920 MLH1 YEAST</a> MUTL protein homolog 1 (DNA...	<a href="#">72</a>	7e-13
<a href="#">gi 1171080 sp P44494 MUTL HAEIN</a> DNA mismatch repair protein...	<a href="#">54</a>	7e-08
<a href="#">gi 13431695 sp P57886 MUTL PASMU</a> DNA mismatch repair protei...	<a href="#">48</a>	1e-06
<a href="#">gi 18928241 sp P40925 MUTL THEMA</a> DNA mismatch repair protein...	<a href="#">48</a>	4e-06
<a href="#">gi 18928241 sp P40925 MUTL BACHD</a> DNA mismatch repair protei...	<a href="#">46</a>	1e-05
<a href="#">gi 18928241 sp P40925 MUTL ECOLI</a> DNA mismatch repair protei...	<a href="#">44</a>	5e-05
<a href="#">gi 127553 sp P14161 MUTL SALTY</a> DNA mismatch repair protei...	<a href="#">44</a>	7e-05
<a href="#">gi 6225738 sp Q9ZC88 MUTL RICPR</a> DNA mismatch repair protei...	<a href="#">40</a>	7e-04
<a href="#">gi 14194944 sp Q9PJG5 MUTL CHLMU</a> DNA mismatch repair protei...	<a href="#">40</a>	0.001
<a href="#">gi 8928218 sp O84579 MUTL CHLTR</a> DNA mismatch repair protein...	<a href="#">39</a>	0.001
<a href="#">gi 20043258 sp Q9KV13 MUTL VIBCH</a> DNA mismatch repair protei...	<a href="#">39</a>	0.002
<a href="#">gi 13631230 sp Q9RP66 MUTL CAUCR</a> DNA mismatch repair protei...	<a href="#">39</a>	0.002
<a href="#">gi 8928214 sp O51229 MUTL BORBU</a> DNA mismatch repair protein...	<a href="#">39</a>	0.002
<a href="#">gi 1709188 sp P49850 MUTL BACSU</a> DNA mismatch repair protein...	<a href="#">38</a>	0.005
<a href="#">gi 8039787 sp O83325 MUTL TREPA</a> DNA mismatch repair protein...	<a href="#">36</a>	0.013
<a href="#">gi 19856116 sp P14160 HEXP</a> DNA mismatch repair protei...	<a href="#">36</a>	0.020
<a href="#">gi 3914082 sp P70754 MUTL</a> DNA mismatch repair protei...	<a href="#">35</a>	0.020
<a href="#">gi 11386926 sp P57633 MUTL</a> DNA mismatch repair protei...	<a href="#">35</a>	0.026
<a href="#">gi 8928240 sp Q9Z794 MUTL CHLPN</a> DNA mismatch repair protei...	<a href="#">35</a>	0.026
<a href="#">gi 1709684 sp P54280 PMS1 SCHPO</a> DNA mismatch repair protei...		
<a href="#">gi 3914081 sp O67518 MUTL AQUAE</a> DNA mismatch repair protei...		
<a href="#">gi 1709685 sp P54278 PMS2 HUMAN</a> PMS1 protein homolog 2 (DNA...	<a href="#">33</a>	0.16
<a href="#">gi 1709686 sp P54279 PMS2 MOUSE</a> PMS1 PROTEIN HOMOLOG 2 (DNA...	<a href="#">32</a>	0.24
<a href="#">gi 8928222 sp P73349 MUTL SYNY3</a> DNA mismatch repair protein...	<a href="#">31</a>	0.60
<a href="#">gi 1709683 sp P54277 PMS1 HUMAN</a> PMS1 protein homolog 1 (DNA...	<a href="#">30</a>	0.85
<a href="#">gi 126232 sp P02239 LGB1 LUPLU</a> Leghemoglobin I	<a href="#">30</a>	1.2
<a href="#">gi 126238 sp P02240 LGB2 LUPLU</a> Leghemoglobin II	<a href="#">28</a>	4.1

seřazeno podle hodnot E

$4 \times 10^{-56}$

link to entrez

LocusLink

Default e value cutoff 10

Bacterial mismatch repair proteins



# BLASTp – hledání konzervativních domén proteinů



Nucleotide

Protein

*formatting* **BLAST**

Translations

Retrieve results for an RID

Your request has been successfully submitted and put into the Blast Queue.

Query = Mutated in Colon Cancer (131 letters)

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



The request ID is

[Format!](#) or [Reset all](#)

The results are estimated to be ready in 36 seconds but may be done sooner.

# BLAST – výstup u srovnání proteinových sekvencí

```
>gi|127552|sp|P23367|MUTL_ECOLI DNA mismatch repair protein mutL  
Length = 615
```

```
Score = 44.3 bits (103), Expect = 5e-05
```

```
Identities = 25/59 (42%), Positives = 33/59 (55%), Gaps = 8/59 (13%)
```

```
Query: 9 LPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHF-----LHE---ESILERVQQHIESKL 59  
L + P L LEI P VDVNVHP KHEV F +H+ + +L +QQ +E+ L  
Sbjct: 280 LGADQQPAFVLYLEIDPHQVDVNVHPAKHEVRFHQSRVLVHDFIYQGVLSVLQQQLETP 338
```

# BLAST – výstup filtrování sekvencí

```
>gi|730028|sp|P40692|MLH1_HUMAN DNA mismatch repair protein Mlh1 1)
      Length = 756
```

```
Score = 233 bits (593), Expect = 8e-62
Identities = 117/131 (89%), Positives = 117/131 (89%)
```

```
Query: 1 IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL 60
        IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL
Sbjct: 276 IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL 335
```

```
Query: 61 GSNSSRMYFTQTLLPGLAGPSGEMVXXXXXXXXXXXXXXXXXDKVYAHQMVRTDSREQK LDA 120
        GSNSSRMYFTQTLLPGLAGPSGEMVK DKVYAHQMVRTDSREQK LDA
Sbjct: 336 GSNSSRMYFTQTLLPGLAGPSGEMVKSTTSLTSSSTSGSSDKVYAHQMVRTDSREQK LDA 395
```

```
Query: 121 FLQPLSKPLSS 131
        FLQPLSKPLSS
Sbjct: 396 FLQPLSKPLSS 406
```

*sekvence s nízkou komplexitou*



# Fasta3 (EBI)

EMBL-EBI EB-eye Search All Databases Enter Text Here Go Reset Advanced Search Give us feedback

Databases Tools EBI Groups Training Industry About Us Help Site Index

- Help Index
- General Help
- Formats
- Gaps
- Matrix
- References
- Fasta Help
- MView Help
- VisualFasta Help

View all Fasta's at EBI

Fasta Programmatic Access

Database Information

Similar Applications

- Fasta
- Blast
- MPsrch
- scansp

EBI > Tools > Similarity & Homology > Fasta

### Fasta - Nucleotide Similarity Search

Provides sequence similarity 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 searching against complete [proteome](#) or [genome](#) databases using the [Fasta programs](#).

[Download Software](#)

PROGRAM	DATABASES	RESULTS	SEARCH TITLE	YOUR EMAIL
fasta3 fasta3 tfastx3 tfasty3	Nucleic Acid EMBL Release EMBL Updates EMBL Coding Sequence	email	Sequence	

MATRIX	GAP OPEN	GAP EXTEND	KTUP	EXPECTATION UPPER VALUE	EXPECTATION LOWER VALUE
none	-14	-4	6	10.0	default

DNA STRAND	HISTOGRAM	MOLECULE TYPE
both	no	DNA

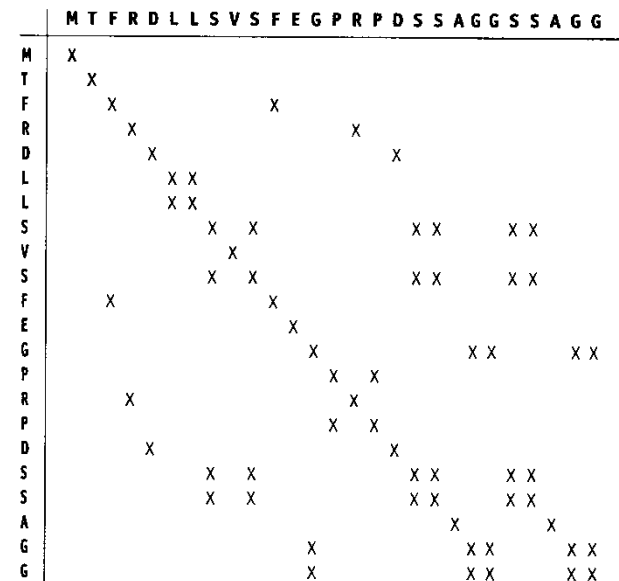
SCORES	ALIGNMENTS	SEQUENCE RANGE	DATABASE RANGE	FILTER	STATISTICAL ESTIMATES
50	50	START-END	START-END	none	Regress

Enter or Paste a  Sequence in any format:

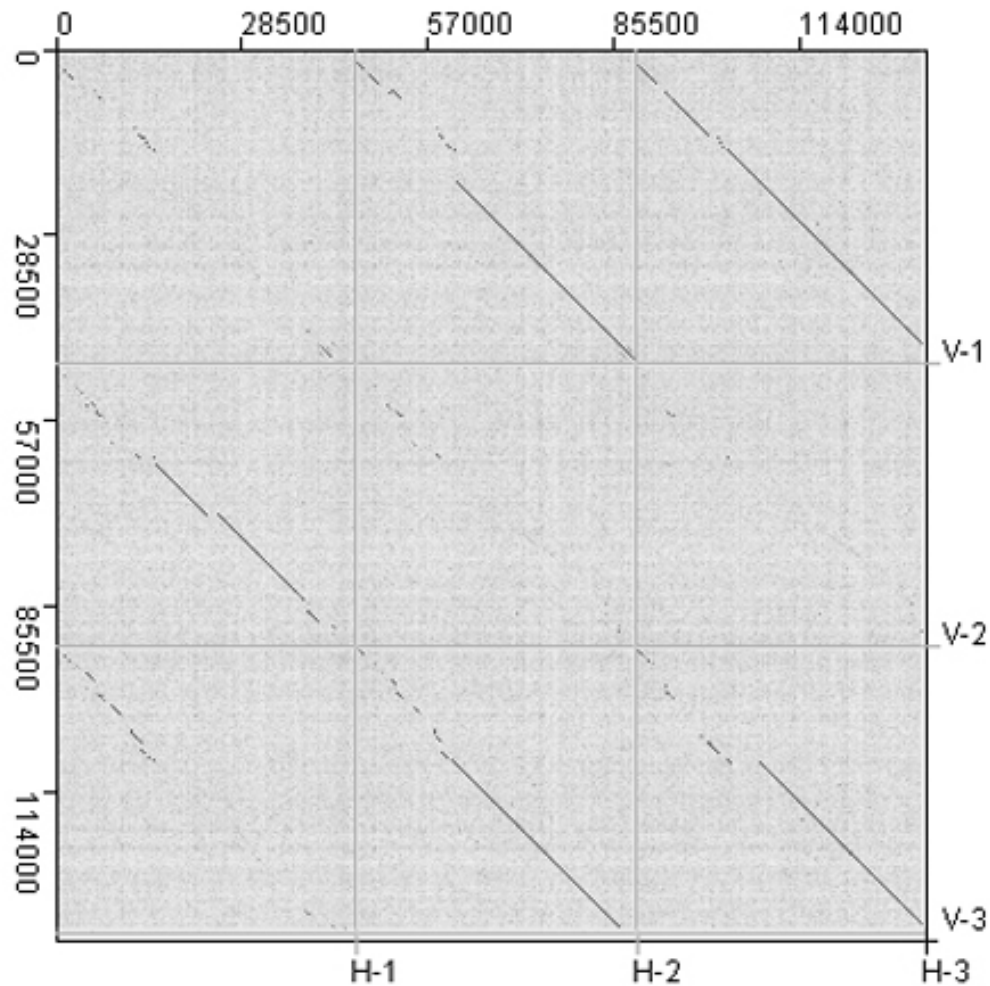
Upload a file:

# Metoda tečkové (Dot-Plot) matice

- Bodový diagram vzájemné podobnosti sekvencí - nejjednodušší pomůcka pro posouzení podobnosti
- Každý zbytek z jedné sekvence je srovnáván s každým zbytkem ve druhé sekvenci
- První sekvence tvoří osu x a druhá sekvence osu y; shoda je vyjádřena **tečkou**
- V oblastech, kde jsou si obě sekvence navzájem podobné tvoří řádek vysokých skóre diagonální linii přes tečkovou matici
- Podobné sekvence pak tvoří přerušované diagonální linie.
- Po odfiltrování diagonál kratších než 3 tečky je výsledkem grafické vykreslení podobností sekvencí ve formě čtvercové nebo trojúhelníkové matice zobrazené v šedé škále

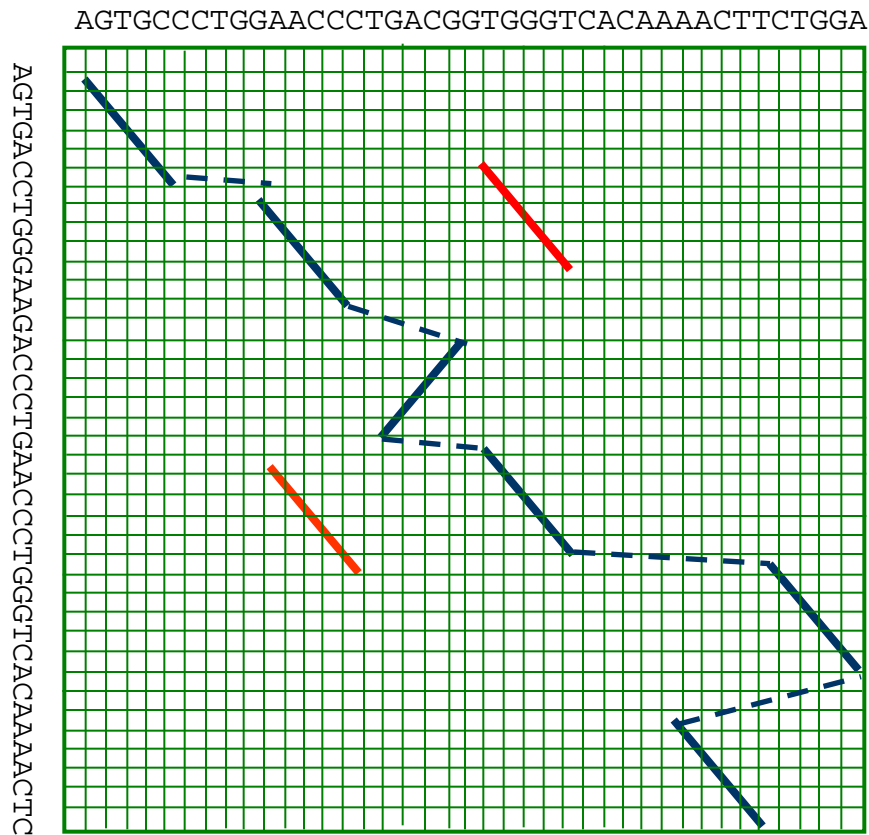


# Příklad: Dot-plot pro 3 virové genomy s různým stupněm podobnosti



# Globální sekvenční přiložení posuzuje podobnost celých dlouhých sekvencí

Nalezení nejefektivnější transformace jedné sekvence do druhé vyžaduje využití dynamického programování



- Bodové změny, delece
- Inverze
- Translokace
- Duplikace
- Kombinace uvedených změn



# Lokální versus globalizované sekvenční přiložení

- Pro optimální lokální alignment požadujeme dosažení nejlepšího skóre kdekoli v matici

**LOKÁLNÍ – nejlepší sekvenční přiložení segmentů bez ohledu na zbytek sekvence**

Smithův-Watermanův algoritmus

- Pro optimální globální alignment požadujeme dosažení nejlepšího skóre v celém řádku/sloupci

**GLOBÁLNÍ – nejlepší sekvenční přiložení celých sekvencí**

Needlemanův-Wunschův algoritmus

# Lokální sekvenční příložení

1 AGGATTGGAATGCTCAGAAGCAGCTAAAGCGTGTATGCAGGATTGGAATTAAGAGGAGGTAGACCG.... 67

1 AGGATTGGAATGCTAGGCTTGATTGCCTACCTGTAGCCACATCAGAAGCACTAAAGCGTCAGCGAGACCG 70

14 TCAGAAGCAGCTAAAGCGT  
| | | | | | | | | | | | | | | |  
42 TCAGAAGCA . CTAAAGCGT

1 AGGATTGGAATGCT  
| | | | | | | | | | | | | | | |  
1 AGGATTGGAATGCT

39 AGGATTGGAAT  
| | | | | | | | | | | | | | | |  
1 AGGATTGGAAT

62 AGACCG  
| | | | | |  
66 AGACCG

# Globální sekvenční příložen

```
1 AGGATTGGAATGCTCAGAAGCAGCTAAAGCGTGTATGCAGGATTGGAATTAAAGAGGAGGT---AGACCG 67
  ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
1 AGGATTGGAATGCTAGGCTTGATTGCCTACCTGTAGCCACATCAGAAGCACTAAAGCGTCAGCGAGACCG 70
```

Dvě sekvence sdílejí oblasti s lokální podobností (end-to-end alignment)

# Adresy služeb pro lokální přiložení

<http://www.ebi.ac.uk/Tools/sss/>

## FASTA

### FASTA

FASTA is another commonly used sequence similarity search tool which uses heuristics for fast **local** alignment searching.

[Protein](#) [Nucleotide](#) [Genomes](#) [Whole Genome Shotgun](#)

### SSEARCH

SSEARCH is an optimal (as opposed to heuristics-based) **local** alignment search tool using the Smith-Waterman algorithm. Optimal searches guarantee you find the best alignment score for your given parameters.

[Protein](#) [Nucleotide](#) [Genomes](#) [Whole Genome Shotgun](#)

### PSI-Search

PSI-Search combines the sensitivity of the Smith-Waterman search algorithm (SSEARCH) with the PSI-BLAST profile construction strategy to find distantly related protein sequences.

[Protein](#)

### GGSEARCH

GGSEARCH performs optimal **global-global** alignment searches using the Needleman-Wunsch algorithm.

[Protein](#) [Nucleotide](#)

### GLSEARCH

GLSEARCH performs an optimal sequence search using alignments that are **global** in the query but **local** in the database sequence. This can be useful when you want to match all of a short query sequence to part of a larger database sequence.

[Protein](#) [Nucleotide](#)

## BLAST

### NCBI BLAST

NCBI BLAST is the most commonly used sequence similarity search tool. It uses heuristics to perform fast **local** alignment searches.

[Protein](#) [Nucleotide](#) [Vectors](#)

### WU-BLAST

WU-BLAST is similar to NCBI BLAST but combines multiple parameter options into a simpler 'sensitivity' setting.

[Protein](#) [Nucleotide](#)

### PSI-BLAST

PSI-BLAST allows users to construct and perform a BLAST search with a custom, position-specific, scoring matrix which can help find distant evolutionary relationships. PHI-BLAST functionality is also available to restrict results using patterns.

[Protein](#)

## ENA Sequence Search

EMBL-EBI has a new nucleotide search tool which is far faster than BLAST for large datasets, with only a marginal loss in search sensitivity.

Try it out at [ENA Sequence Search](#).