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
  - Nejednotná nomenklatura genů
- Prohledávání databází podle podobnosti sekvencí
  - Výpočet lokálního/párového přiložení (alignment)
    = uspořádání do 2 pod sebou ležících řádků tak, aby identické zbytky ležely pod sebou
  - Výpočet mnohonásobného přiložení (multiple alignment) pro 3 a více sekvencí

Význam sekvenčního přiložení

Použití	Princip
Predikce, extrapolace	Klasifikace necharakterizovaných sekvencí do rodin / skupin
Hledání v databázích	Identifikace podobných sekvencí, charakterizace genů
Identifikace vzorů	Stanovení konzervovaných vzorů, profilů a identifikace funkčních oblastí a domén
Fylogenetická analýza	Rekonstrukce evoluce z ortologních sekvencí
Predikce struktury	Kvalitní přiložení umožňují predikci sekundární struktury využívané při konstrukci 3D modelů
Sestavení celogenomových sekvencí (assembly)	Využívá techniky přiložení pro vytváření kontigů ze sekvenačních dat
Analýza oligonukleotidů pro PCR	Design primerů a sond, posouzení sekundárních struktur

Terminologie použitá pro srovnávání sekvencí

- **Identita sekvencí** (Sequence identity), podíl identických aminokyselinových nebo nukleotidových zbytků ve stejné pozici
- **Podobnost sekvencí** (Sequence similarity), podíl identických plus substituovaných zbytků s podobnými chemickými vlastnostmi.
- **Homologie sekvencí** (Sequence homology), termín použitelný pouze u evolučně příbuzných sekvencí, např. stanovení ANI (average nucleotide identity) z celogenomových sekvencí nebo data z DNA-DNA hybridizací

Princip hledání podobnosti

- Sekvence jsou tvořeny symboly abecedy
- Komplexita sekvence je určena počtem různých znaků, které se mohou vyskytovat v sekvenci (DNA = 4, proteiny = 21)
- 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
  - Metodu dot-plot matic
  - Dynamické programování

Nástroje pro vyhledávání lokálních podobností sekvencí

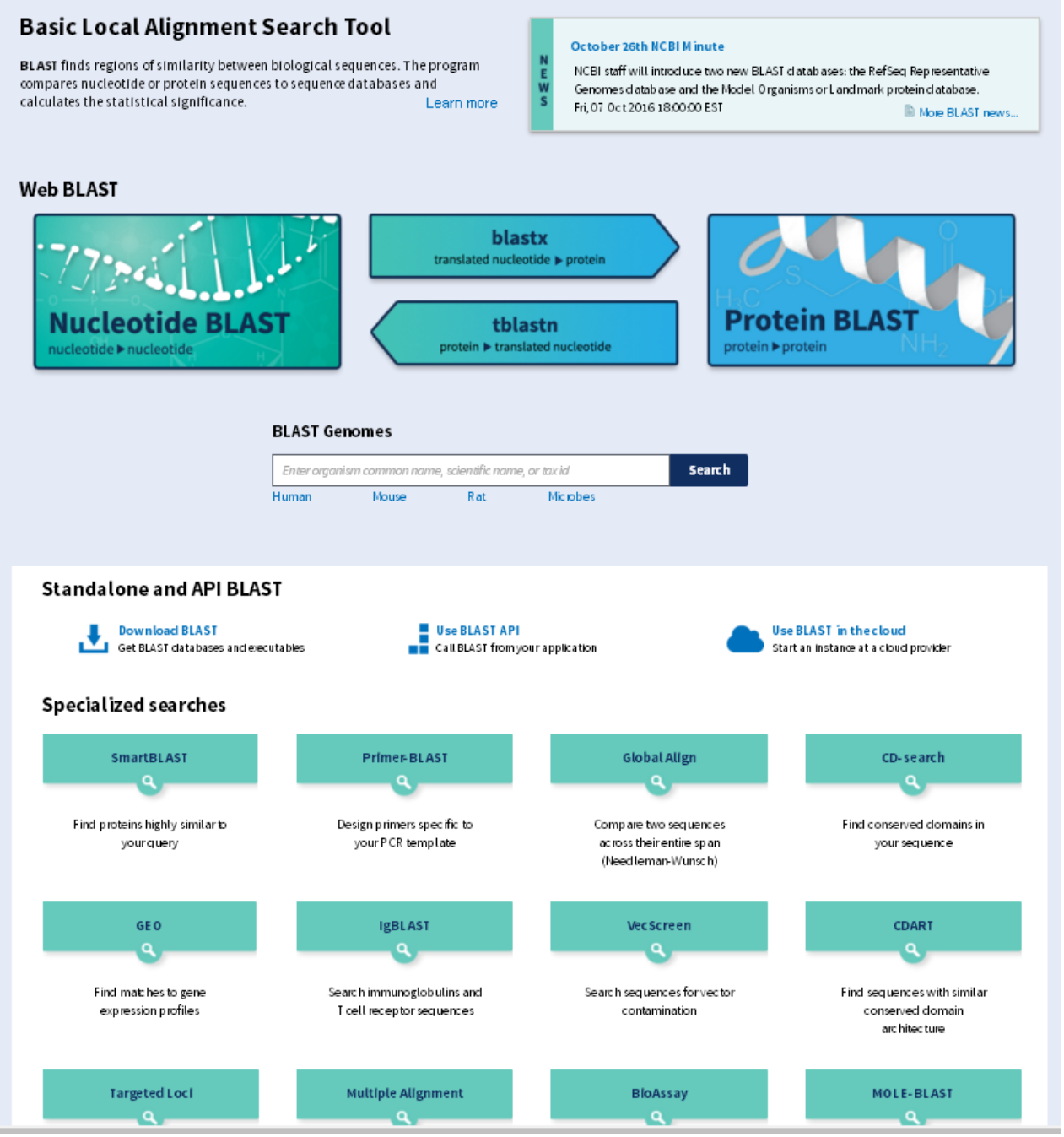
Sady programů zahrnujících algoritmy pro vyhledávání podobnosti v dostupných databázích sekvencí bez ohledu na to zdali dotazovaná sekvence je **DNA** nebo **protein**.

- BLAST
- Altschul et al., 1990
- dostupný na serveru NCBI

- FASTA
- Lipman a Pearson 1985
- dostupný na serveru EBI

Co je to BLAST?

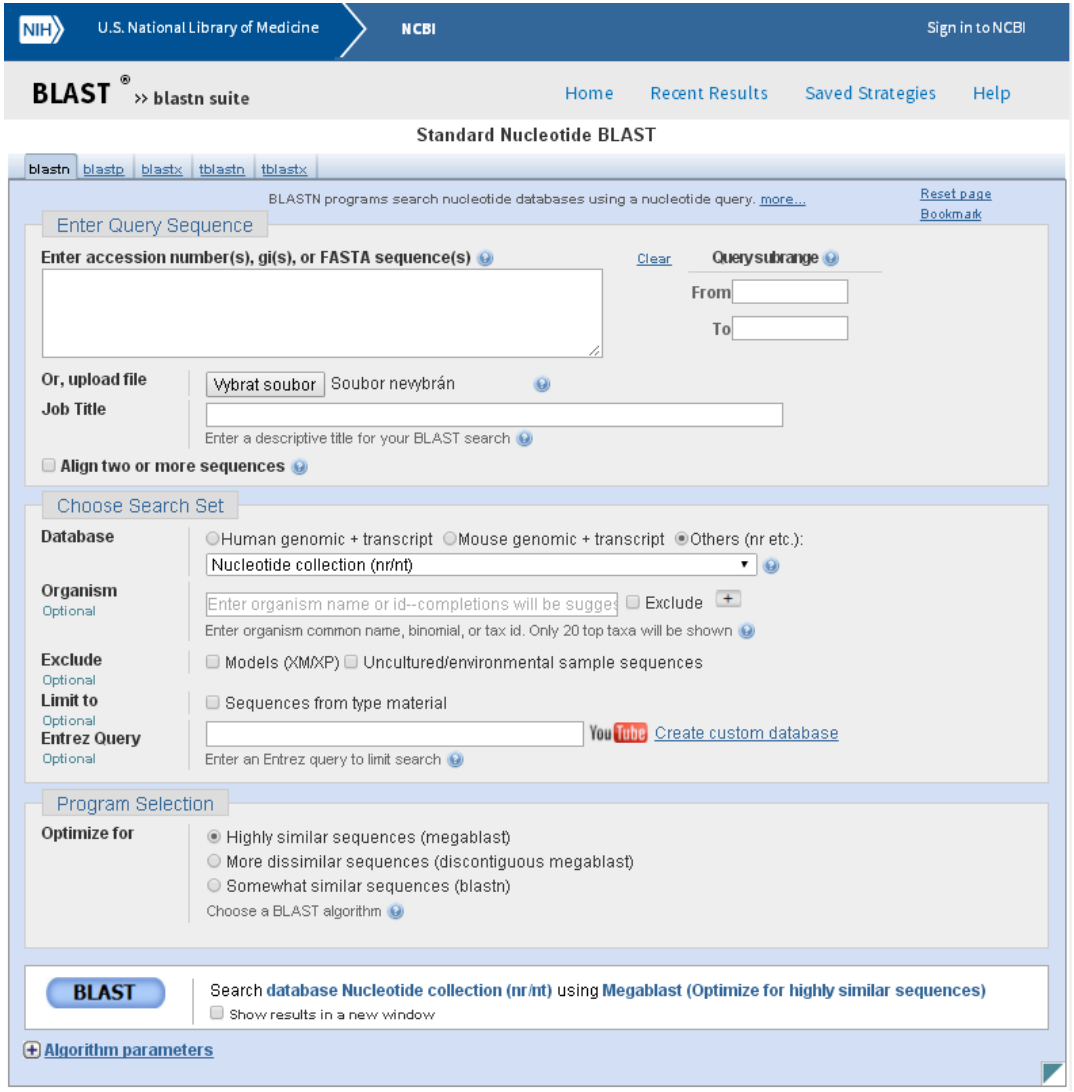
- **B**asic **L**ocal **A**lignment **S**earch **T**ool
  - Hledání lokálních podobností
  - Heuristický přístup založený na **Smith-Watermanově** algoritmu
  - Vyhledá nejoptimálnější **přiložení sekvencí**
  - Poskytuje data o statistické významnosti
  - Zobrazuje vzájemně párové přiložení sekvencí
  - Lokalizuje oblasti sekvencí s vysokou podobností a umožňuje zobrazení jejich primární struktury a funkce



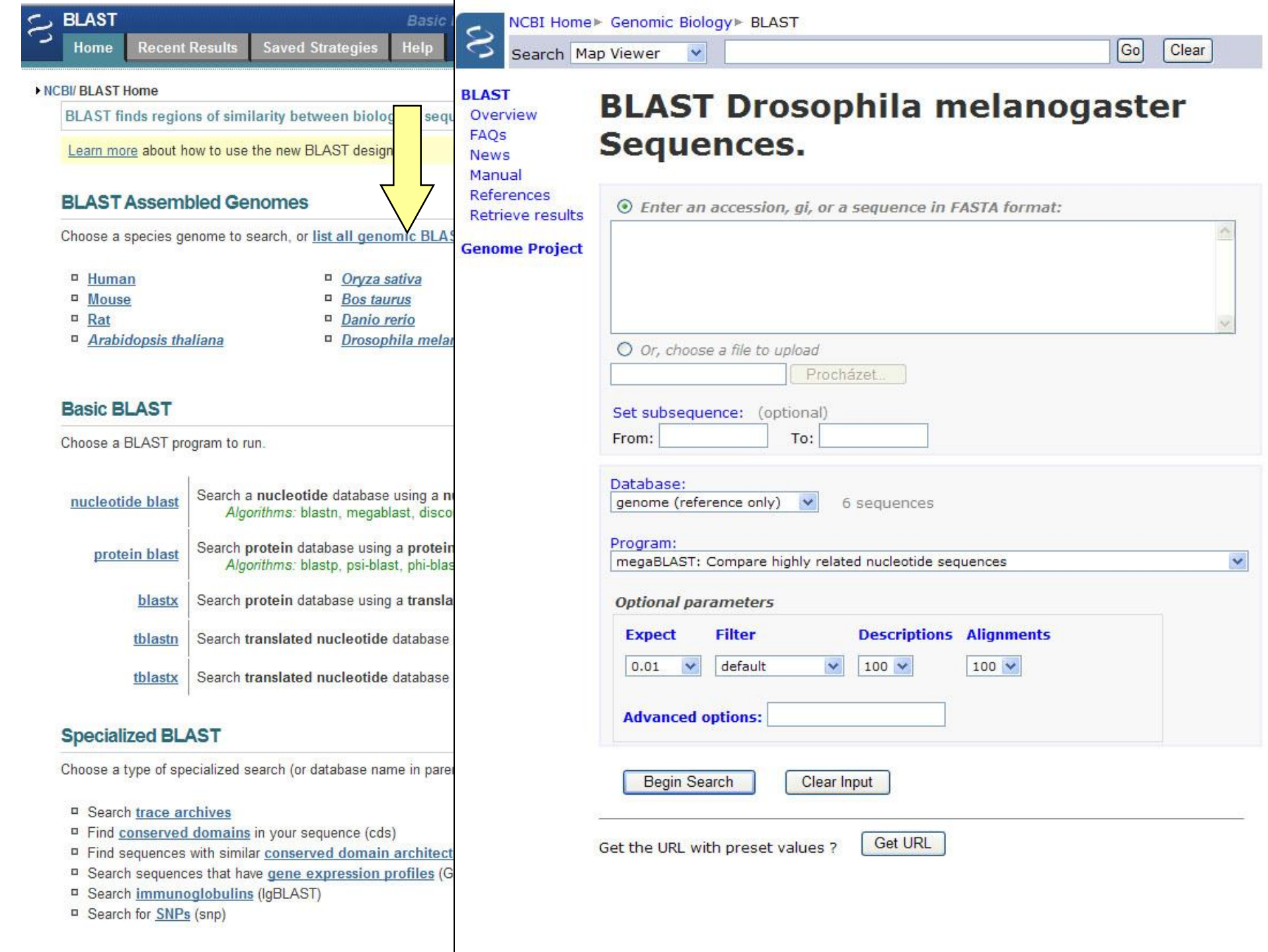
Výchozí stránka BLAST

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


BLAST
Basic
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST Home
BLAST finds regions of similarity between biolog
sequ
Learn more about how to use the new BLAST design
BLAST Assembled Genomes
Choose a species genome to search, or list all genomic BLAS
Human
Mouse
Rat
Arabidopsis thaliana
Oryza sativa
Bos taurus
Danio rerio
Drosophila melar
Basic BLAST
Choose a BLAST program to run.
nucleotide blast
Search a nucleotide database using a nu
Algorithms: blastn, megablast, disco
protein blast
Search protein database using a protein
Algorithms: blastp, psi-blast, phi-blas
blastx
Search protein database using a transla
tblastn
Search translated nucleotide database
tblastx
Search translated nucleotide database
Specialized BLAST
Choose a type of specialized search (or database name in parer
Search trace archives
Find conserved domains in your sequence (cds)
Find sequences with similar conserved domain architect
Search sequences that have gene expression profiles (G
Search immunoglobulins (IgBLAST)
Search for SNPs (snp)
NCBI Home
Genomic Biology
BLAST
Search
Map Viewer
Go
Clear
BLAST
Overview
FAQs
News
Manual
References
Retrieve results
Genome Project
BLAST Drosophila melanogaster Sequences.
Enter an accession, gi, or a sequence in FASTA format:
Or, choose a file to upload
Procházet...
Set subsequence: (optional)
From:
To:
Database:
genome (reference only)
6 sequences
Program:
megaBLAST: Compare highly related nucleotide sequences
Optional parameters
Expect
Filter
Descriptions
Alignments
0.01
default
100
100
Advanced options:
Begin Search
Clear Input
Get the URL with preset values ?
Get URL

Basic BLAST – výběr programů

Využití jednotlivých programů BLAST

Program	Dotaz	Databáze	Úroveň srovnání	Použití
blastn	DNA	DNA	DNA	Hledání identických sekvencí DNA
blasp	Protein	Protein	Protein	Hledání homologních proteinů
blastx	DNA	Protein	Protein	Hledání genů a homologních proteinů na DNA
tblastn	Protein	DNA	Protein	Hledání genů u necharakterizovaných DNA
tblastx	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**	MEGABLAST Standard BLAST (blastn)
		Vyhledání podobných sekvencí jako dotazovaná	Standard BLAST (blastn)
		Vyhledání podobných proteinů k překladu dotazované sekvence v přeložených databázích DNA	Translated BLAST (tblastx)
	Protein	**Vyhledání podobných proteinů k překladu dotazované sekvence v databázích proteinů**	Translated BLAST (blastx)
7 - 20 bp	**DNA**	**Vyhledání vazebných míst primerů nebo krátkých motivů**	Search for short, nearly exact matches

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**	Standard Protein BLAST (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	PSI-BLAST
		Vyhledání proteinů podobných dotazovanému v okolí určitého vzoru	PHI-BLAST
	Konzervativní domény	**Vyhledání konzervativních domén v dotazované sekvenci**	CD-search (RPS-BLAST)
	Konzervativní domény	**Vyhledání konzervativních domén v dotazované sekvenci a identifikace ostatních proteinů s podobnou architekturou domén**	Conserved Domain Architecture Retrieval Tool (CDART)
	DNA	**Vyhledání podobných proteinů v přeložených databázích DNA**	Translated BLAST (tblastn)
5-15 zbytků	Protein	**Hledání peptidových motivů**	Search for short, nearly exact matches

Jak používat BLAST?

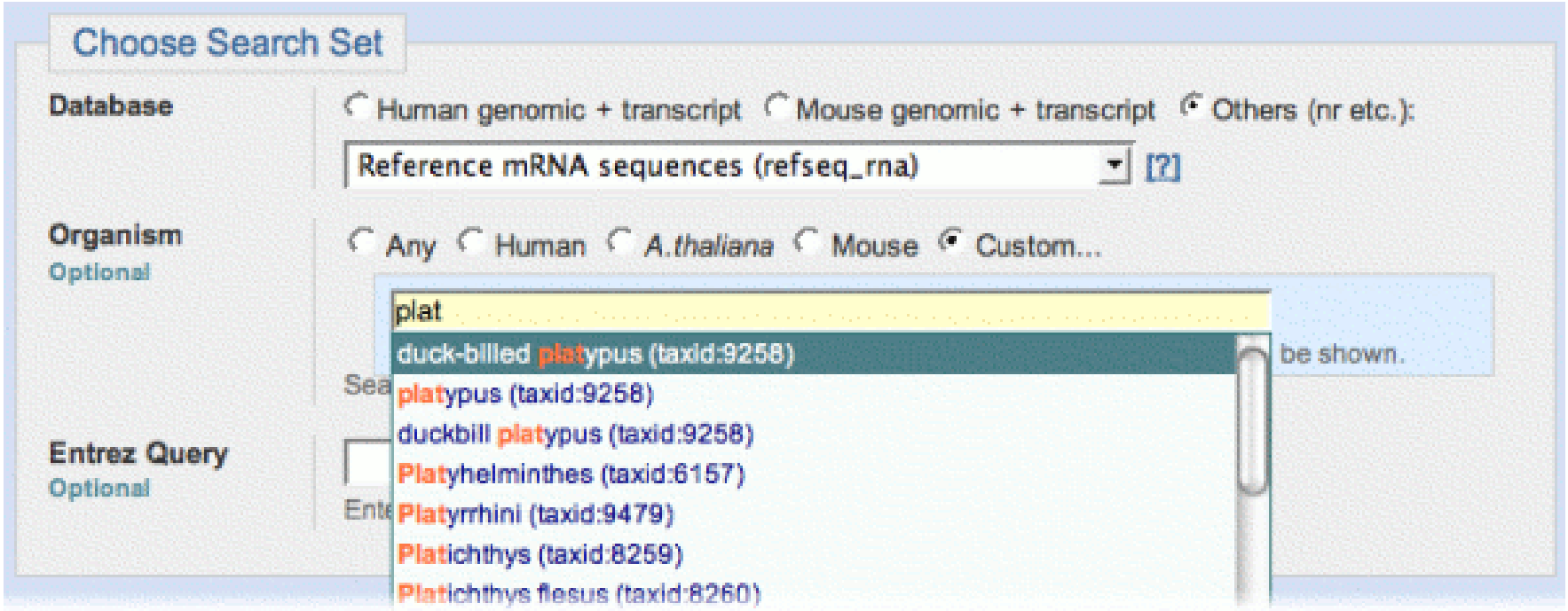
- http://blast.ncbi.nlm.nih.gov/Blast.cgi

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

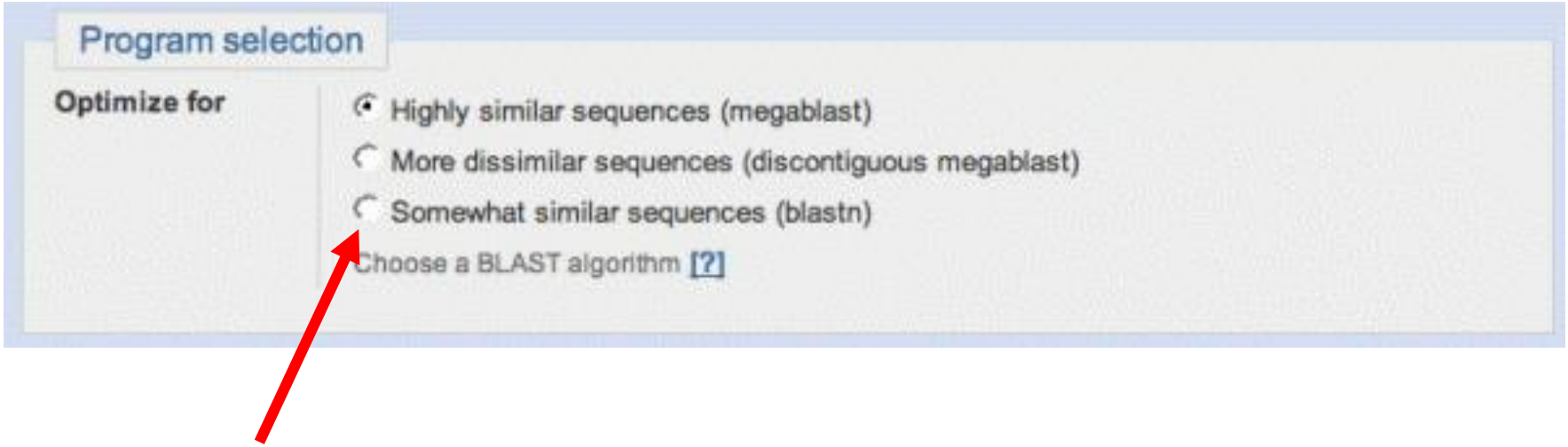

```
>priklad sekvence
GAATTCTTCAAAAAAGTATTCGTTGGATACACGGACAGTGAAGATCATTCAGAGGATTCTGCAAGTTCGTTACCCAGCTAACCCCCA
AAATGTTGAAGTAGCAGTTAATTCAAAATCTGCAACAGTTTCAGCAGAATAGGGGCTTTCAAAATAAATCAAAGGAGAATAATTTAT
GACTAAAACTTTAAAGGTTTATAAAGGAGACGACGTCGTAGCTTCTGAACAAGGTGAAGGCAAAGTGTCAGTAACTTTATCTAATTT
AGAAGCGGATACAACTTATCCAAAAGGTACTTACCAAGTGGCATGGGAAGAAAATGGTAAAGAATCTAGTAAAGTTGATGTACCTCA
ATTCAAAACCAATCCAATTCTAGTCTCAGGCGTATCATTTACACCCGAAACTAAATCAATCACGGTAAATGCTGATGACAATGTTGA
ACCAAACATTGCACCAAGTACAGCAACGAATAAAACGTTGAAATATACAAGTGAACATCCAGAGTTTGTTACTGTTGATGAGAGAAC
AGGAGCAATTCACGGTGTAGCTGAGGGAACTTCAGTTATCACTGCTACGTCTACTGACGGAAGTGACAAGTCTGGACAAATTACAGT
AACAGTAACAAATGGATAATTATTTGAGACGCAGAATATCTGCGTCT
```

Výběr databáze

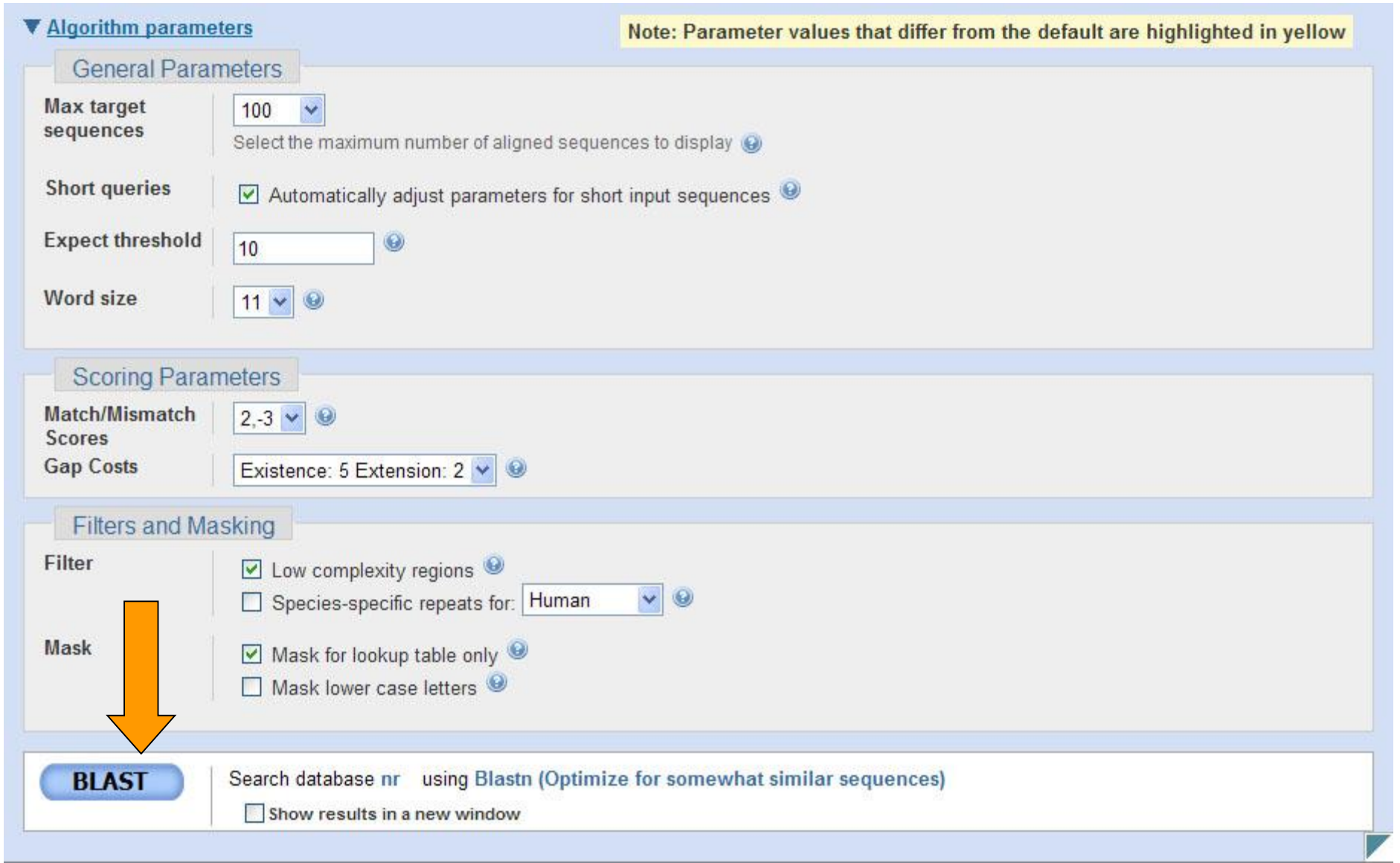


- Others (nr etc.) = celá databáze (neredundantní nukleotidová nr/nt)

Výběr podprogramu



Úprava parametrů algoritmu



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

TACTGGACATG

ACTGGACATGG

CTGGACATGGA

TGGACATGGAC

GGACATGGACC

GACATGGACCC

ACATGGACCCT

. . . . . .

minimální velikost = 7

blastn default = 11

megablast default = 28

tabulka se všemi slovy dotazu

Slova pro proteinové sekvence

Dotaz: **GTQITVEDLFYNIATRRKALKN**

GTQ Velikost = 3

Velikost slova může být 2, 3 (default = 6)

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

GTQITVEDLFYNI

SEI **YYN** sousedící slova

2 nalezené shody

přiložení sekvencí, které BLAST **může** nalézt

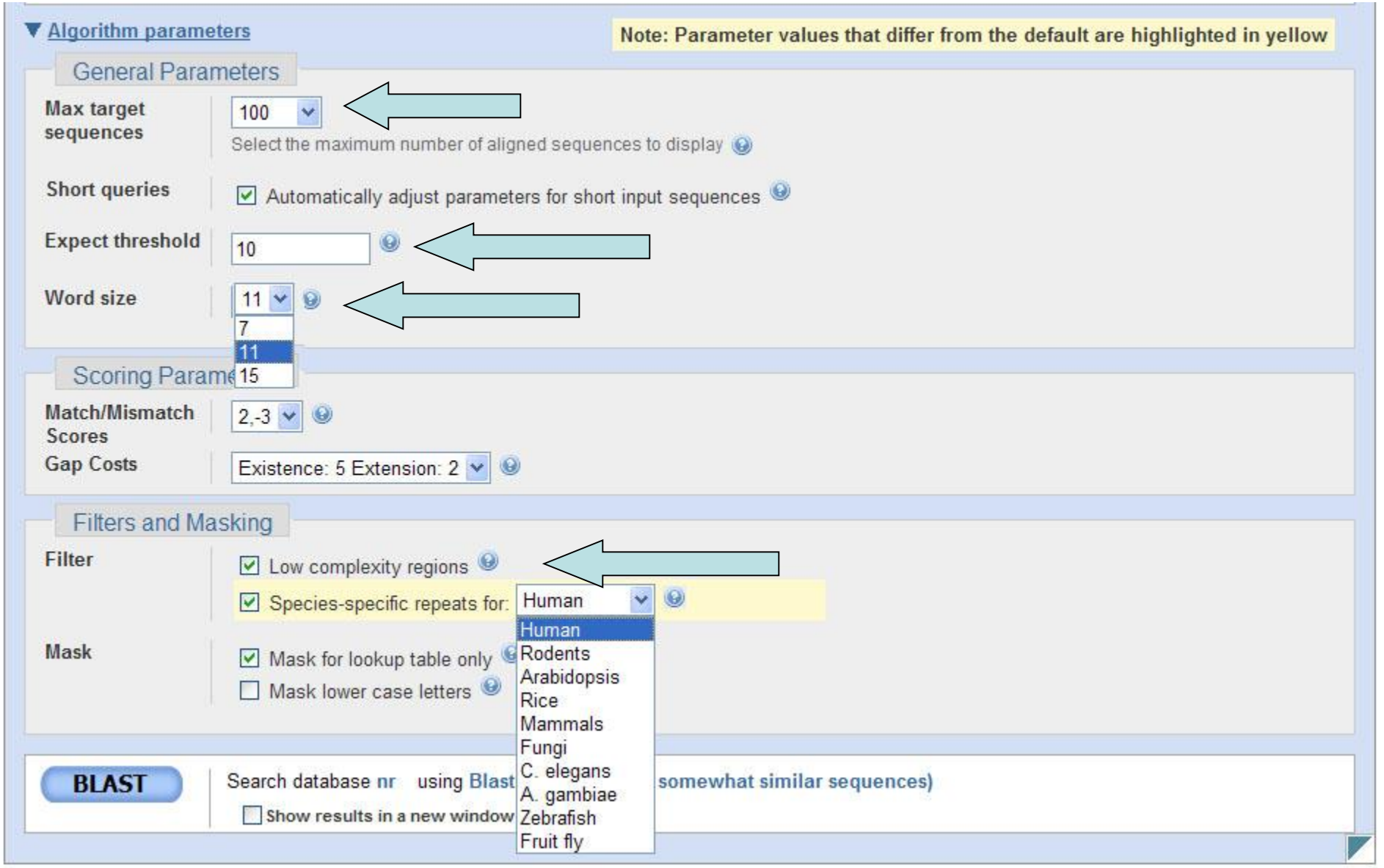
```
1 AATGGTAAAGACTACTGGATCATTAAGAACTCCTGGGGAG
  ||||| |||||||||||||||| || ||||||||||||||
1 AATGGAAAAGACTACTGGATCATCAAAAACTCCTGGGGAG
```

sekvence obsahují definovanou shodu slova

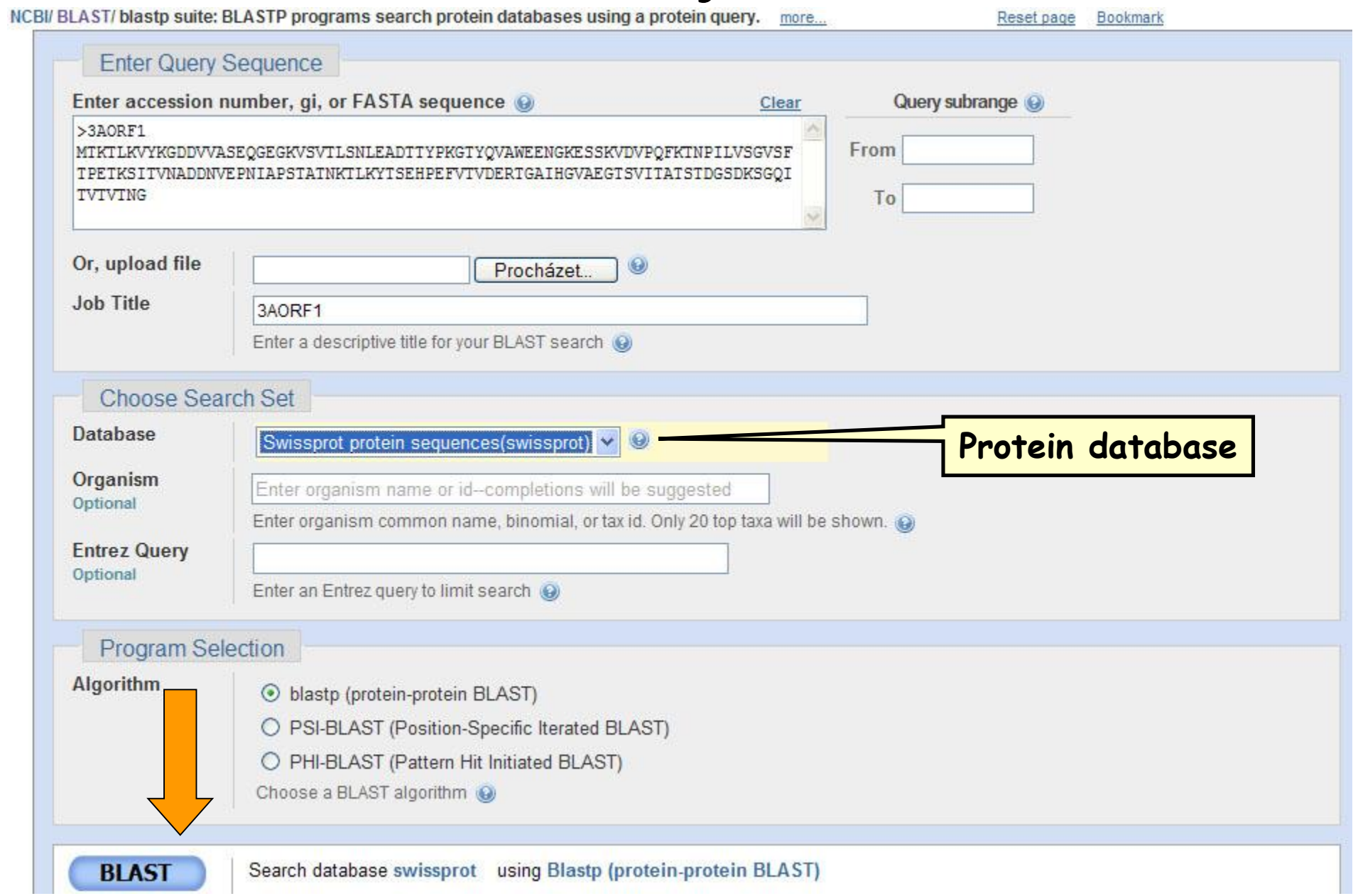
přiložení sekvencí, které BLAST **nemůže** nalézt

```
  1 GAATATATGAAGACCAAGATTGCAGTCCTGCTGGCCTGAACCACGCTATTCTTGCTGTTG
    ||  |  ||  ||  ||  |   ||  ||  ||    ||   |   |||  ||||||  |  |  ||  |  |||  |
  1 GAGTGTACGATGAGCCCGAGTGTAGCAGTGAAGATCTGGACCACGGTGTACTCGTTGTCG

 61 GTTACGGAACCGAGAATGGTAAAGACTACTGGATCATTAAGAACTCCTGGGGAGCCAGTT
    |  ||  ||      ||   |||  ||   |  ||||||  ||  |  ||||||   |||||   |      |
 61 GCTATGGTGTTAAGGGTGGGAAGAAGTACTGGCTCGTCAAGAACAGCTGGGCTGAATCCT

121 GGGGTGAACAAGGTTATTTCAGGCTTGCTCGTGGTAAAAAC
     ||||  ||  |||||  ||   ||     |   |  ||||   ||  |||
121 GGGGAGACCAAGGCTACATCCTTATGTCCCGTGACAACAAC
```

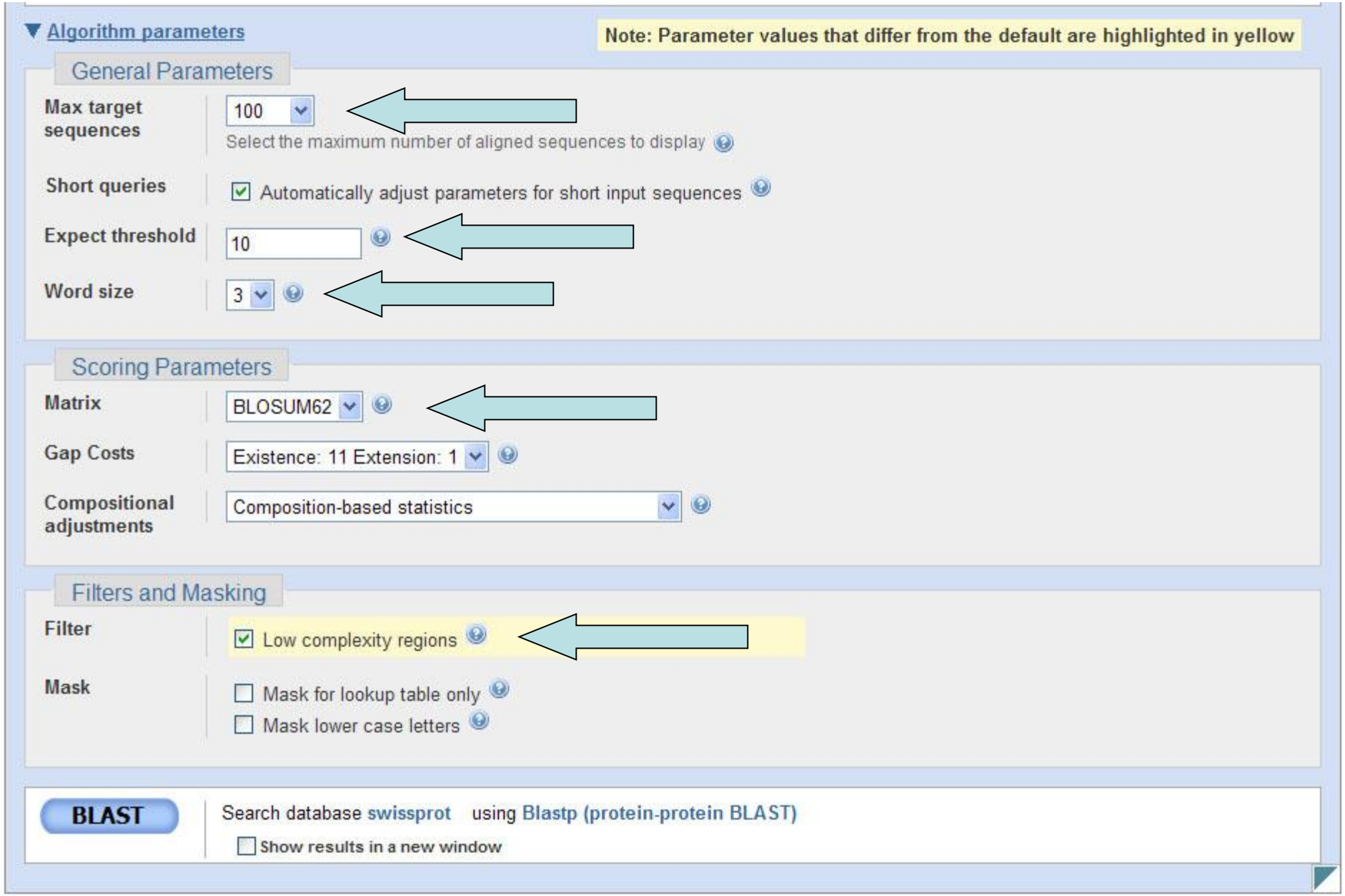
BLASTn - Možnosti nastavení



Proteinový BLAST



BLASTp - Možnosti nastavení



Hodnocení výsledků přiložení

- K posouzení významnosti shody nalezených úseků se používá numerická hodnota označovaná jako **skóre sekvenčního přiložení** (***S***)
  - Hrubé skóre (Raw score)
    - Suma skóre pro identity plus substituce minus penalizace mezer
  - Normalizované skóre (Normalised score)
    - Nezávislé na systému, umožňuje srovnání různých přiložení

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

Matice identity a možnosti výpočtu skóre

	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
||  ||||||||||||  |||||     skóre = 19-9 = 10
CACGTAGCAAGCTTG-GTGTCA
```

Penalizace mezer

- Kvalitní (optimální) přiložení podobných sekvencí
  - **maximalizuje** počet srovnatelných protějšků
  - **minimalizuje** počet mezer
- Povolení vložení mnoha 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 extension

gap opening

Výpočet penalizace za vložení mezery:

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 pravděpodobnot zachování v příslušné pozici sekvence proteinu než jiný.
  - Každá matrice je určená pro určitý typ vyhledávání – JE TŘEBA VĚDĚT CO HLEDÁME!

Substituční Matice

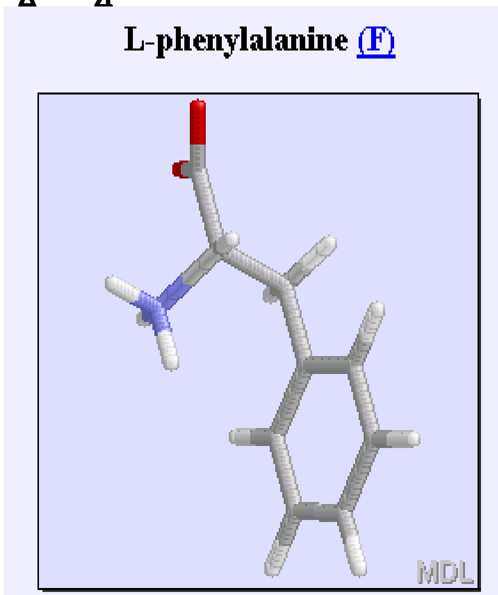
- Proč používat substituční matice?

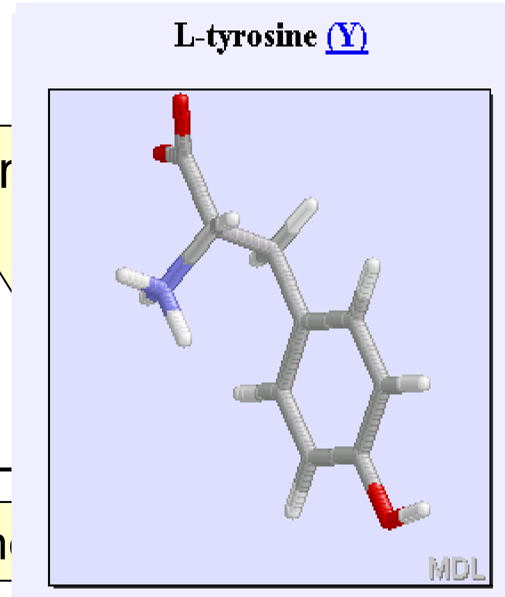
1. Stanovit pravděpodobnou homologii dvou proteinových sekvencí.
2. Substituce, které jsou více pravděpodobné získají vyšší skóre
3. Substituce, které jsou méně pravděpodobné obdrží nižší skóre.

Matice BLOSUM

- **BL**ocks **S**ubstitution **M**atrix
- Změny probíhající během dlouhodobé evoluce nejsou často vhodné pro výpočty a sledování malých recentních změn
- Matice BLOSUM jsou sestaveny na základě **analýzy mnohonásobných přiložení evolučně příbuzných proteinů v databázi BLOCKS**
- BLOSUM-x používá analýzu pouze těch proteinů, které mají **alespoň x %** identitu
  - BLOSUM45, BLOSUM50, BLOSUM62, BLOSUM80

Příklad matice BLOSUM62





Běžné ar… …ší významnost

Vzácn… …šší významnost

Negativní pro málo pravděpodobné substituce

Pozitivní pro více pravděpodobné substituce

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	X
A	4																				
E							5														
G							-2	6													
H							0	-2													
I							-3	-4	-												
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5								
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6							
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7						
S	1			0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4					
T													-1	-2	-1	1	5				
W													-1	1	-4	-3	-2	11			
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7		
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2		-1	-2	-2	0	-3	-1	4	
X	0	-1													-2	0	0	-2	-1	-1	-1

Matice PAM

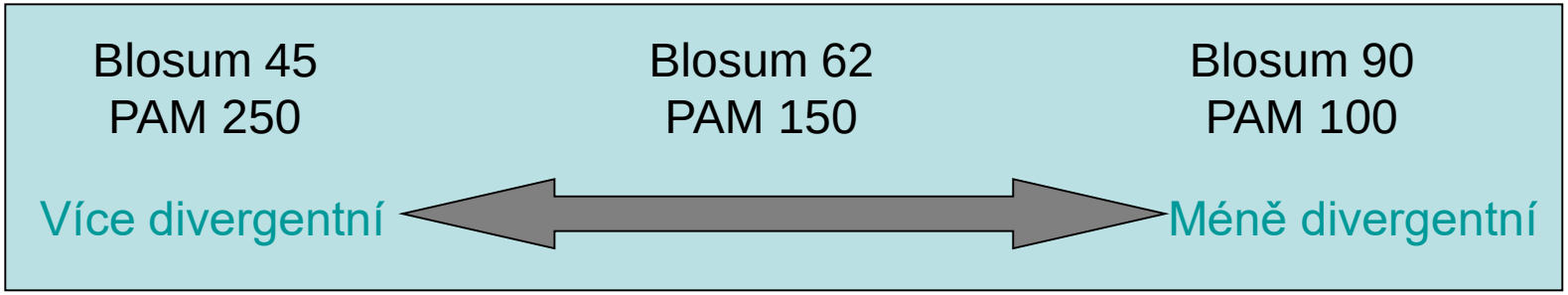
- PAM
  - **P**ercent **A**ccepted **M**utation
  - založeny na konceptu akceptovatelných bodových mutací za 10^8 let v globálních mnohonásobných přiloženích
  - Stanoveny na základě výpočtů u blízce příbuzných proteinů s identitou > 85%; 1572 proteinových sekvencí z 71 rodin
  - PAM1 reprezentuje 1% změn (1 mutace na 100 aminokyselinových zbytků)
  - PAM250 = $(PAM1)^{250}$

Nevýhody substitučních matic

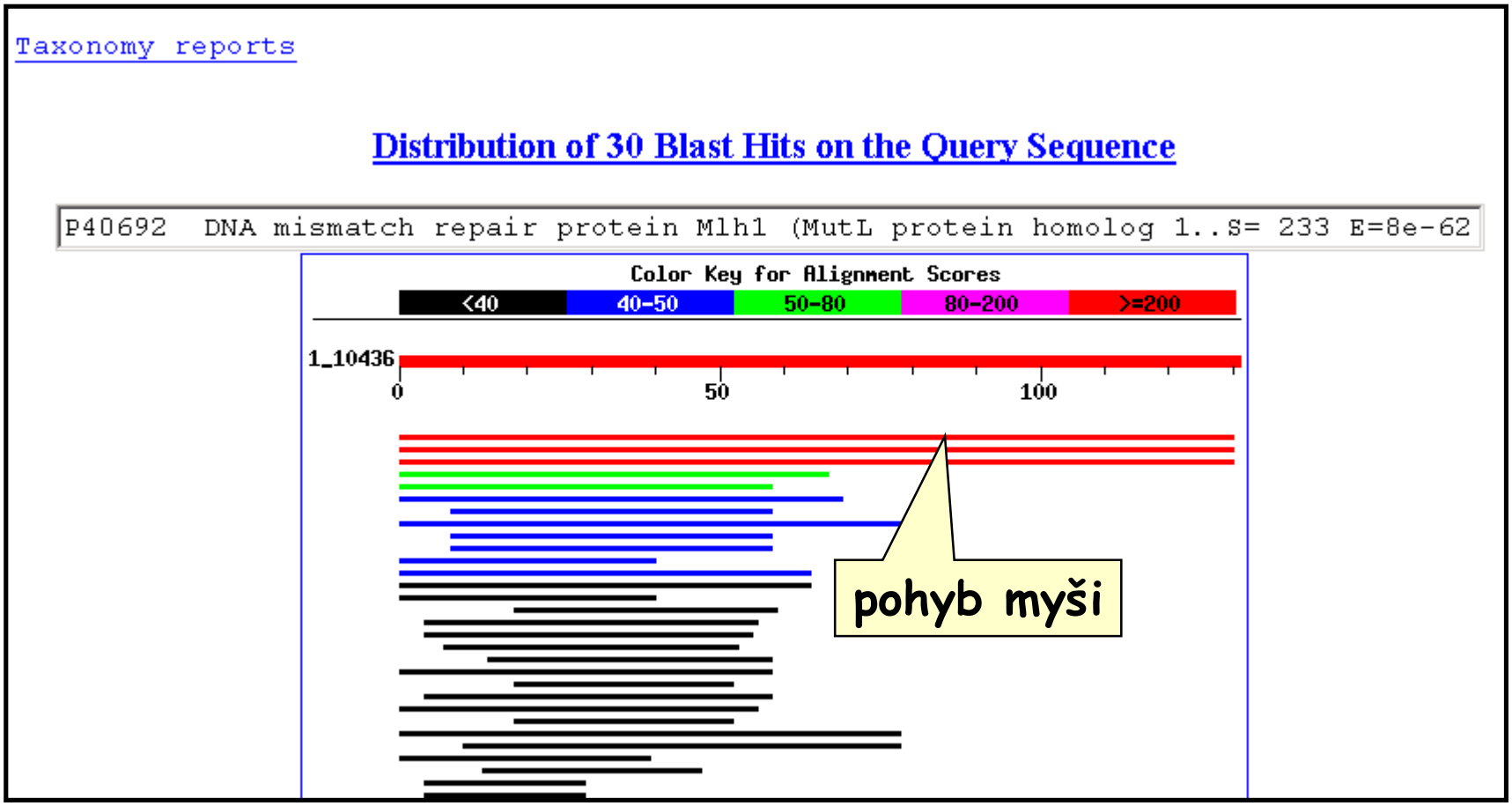
- Neberou v úvahu vzdálené interakce mezi aminokyselinovými zbytky
- Předpokládají, že identické zbytky v sekvenci jsou rovnocenné. Ve skutečnosti zbytky v aktivních místech enzymů podléhají jiným evolučním tlakům než stejné zbytky mimo tato místa
- Předpokládají, že evoluční rychlost je konstatní.

PAM versus BLOSUM

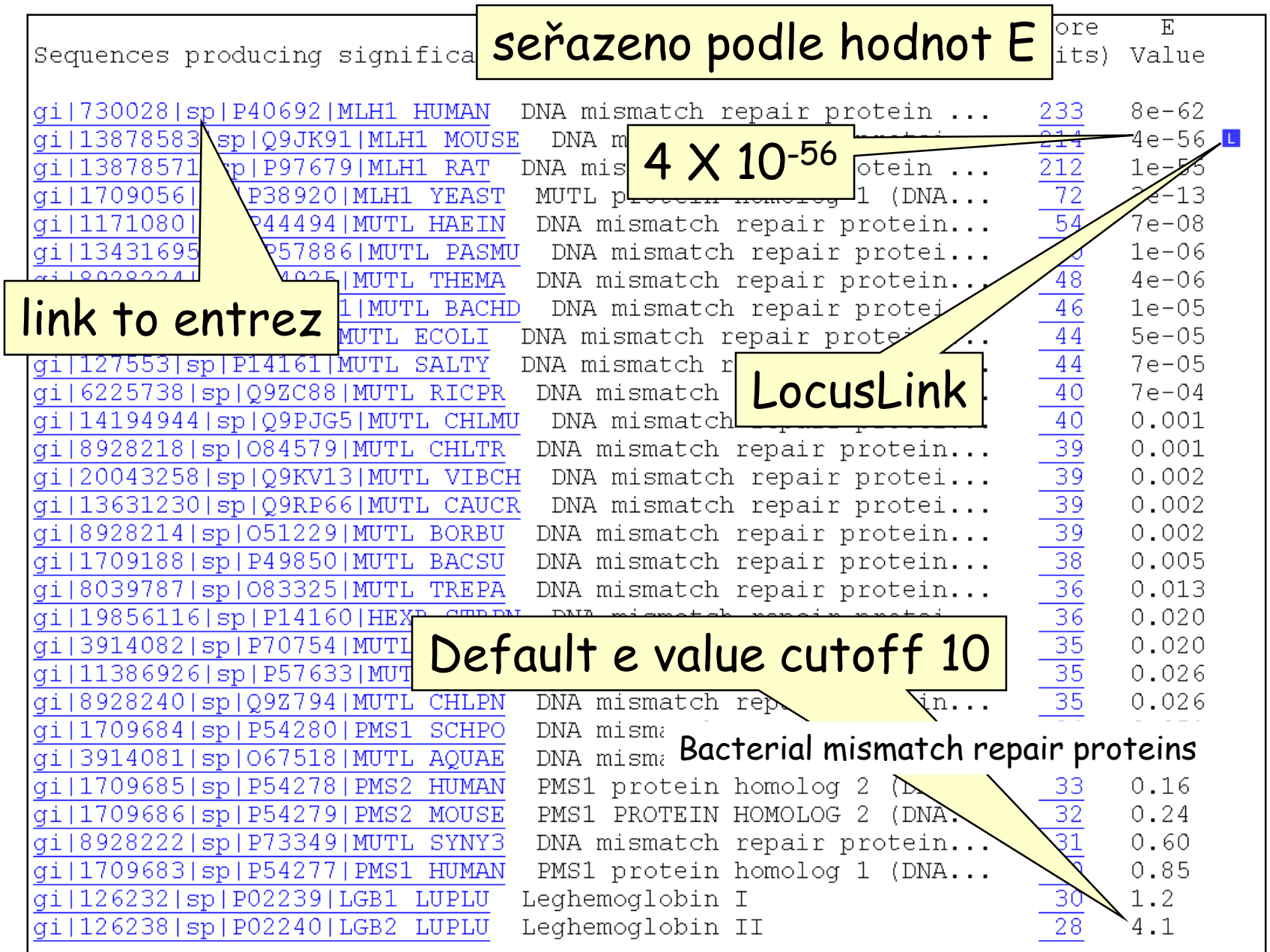
- PAM Matice (Percent Accepted Mutation)
  - Odvozené z pozorování; malé množství srovnávaných dat
  - vhodné pro evoluční modely
  - Všechny výpočty vycházejí z PAM1
  - PAM250 je nejpoužívanější
- BLOSUM (BLOck SUbstitution Matrices)
  - Odvozené z pozorování; velké množství vysoce konzervovaných sekvencí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



BLAST – grafický výstup



BLAST: popis výstupu



Sequences producing significa		ore its)	E Value
gi|730028|sp|P40692|MLH1_HUMAN	DNA mismatch repair protein ...	233	8e-62
gi|13878583|sp|Q9JK91|MLH1_MOUSE	DNA m	214	4e-56
gi|13878571|sp|P97679|MLH1_RAT	DNA mis otein ...	212	1e-55
gi|1709056| P38920|MLH1_YEAST	MUTL p 1 (DNA...	72	2e-13
gi|1171080| P44494|MUTL_HAEIN	DNA mismatch repair protein...	54	7e-08
gi|13431695 P57886|MUTL_PASMU	DNA mismatch repair protei...		1e-06
|MUTL_THEMA	DNA mismatch repair protein...	48	4e-06
1|MUTL_BACHD	DNA mismatch repair protei	46	1e-05
MUTL_ECOLI	DNA mismatch repair protei...	44	5e-05
gi|127553|sp|P14161|MUTL_SALTY	DNA mismatch r	44	7e-05
gi|6225738|sp|Q9ZC88|MUTL_RICPR	DNA mismatch	40	7e-04
gi|14194944|sp|Q9PJG5|MUTL_CHLMU	DNA mismatch	40	0.001
gi|8928218|sp|O84579|MUTL_CHLTR	DNA mismatch repair protein...	39	0.001
gi|20043258|sp|Q9KV13|MUTL_VIBCH	DNA mismatch repair protei...	39	0.002
gi|13631230|sp|Q9RP66|MUTL_CAUCR	DNA mismatch repair protei...	39	0.002
gi|8928214|sp|O51229|MUTL_BORBU	DNA mismatch repair protein...	39	0.002
gi|1709188|sp|P49850|MUTL_BACSU	DNA mismatch repair protein...	38	0.005
gi|8039787|sp|O83325|MUTL_TREPA	DNA mismatch repair protein...	36	0.013
gi|19856116|sp|P14160|HEXB_STRPN		36	0.020
gi|3914082|sp|P70754|MUTL		35	0.020
gi|11386926|sp|P57633|MUT		35	0.026
gi|8928240|sp|Q9Z794|MUTL_CHLPN	DNA mismatch rep in...	35	0.026
gi|1709684|sp|P54280|PMS1_SCHPO	DNA misma		
gi|3914081|sp|O67518|MUTL_AQUAE	DNA misma		
gi|1709685|sp|P54278|PMS2_HUMAN	PMS1 protein homolog 2 (D	33	0.16
gi|1709686|sp|P54279|PMS2_MOUSE	PMS1 PROTEIN HOMOLOG 2 (DNA.	32	0.24
gi|8928222|sp|P73349|MUTL_SYNY3	DNA mismatch repair protein...	31	0.60
gi|1709683|sp|P54277|PMS1_HUMAN	PMS1 protein homolog 1 (DNA...		0.85
gi|126232|sp|P02239|LGB1_LUPLU	Leghemoglobin I	30	1.2
gi|126238|sp|P02240|LGB2_LUPLU	Leghemoglobin II	28	4.1

Statistika lokálního přiložení

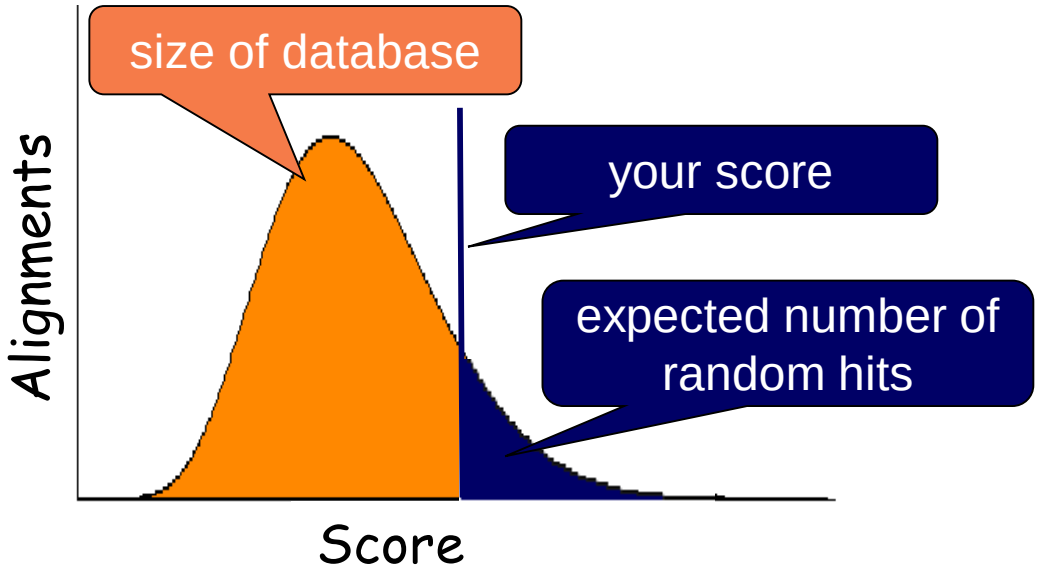
- Statistiky jsou odvozené ze skóre přiložení
- Popisují jeho celkovou kvalitu na základě porovnání pravděpodobnosti výskytu nalezených segmentů o určité sekvenční podobnosti s pravděpodobností, že se taková podobnost vyskytne mezi dvěma náhodnými sekvencemi
  - p-value (hodnota p)
    - Pravděpodobnost, že alignment s tímto skóre se vyskytne náhodně v databázi použité velikosti
  - e-value (hodnota e) („Expectation value“)
    - vyjadřuje počet různých sekvenčních přiložení se skórem shodným nebo vyšším než je dosažená hodnota, jejíž výskyt je očekáván při náhodném vyhledávání v databázi.
    - Čím blíže je hodnota *e* k 0, tím lepší je přiložení

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.



BLASTp – hledání konzervativních domén proteinů

NCBI formatting BLAST

Nucleotide Protein 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 1038888657-012751-16086

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**

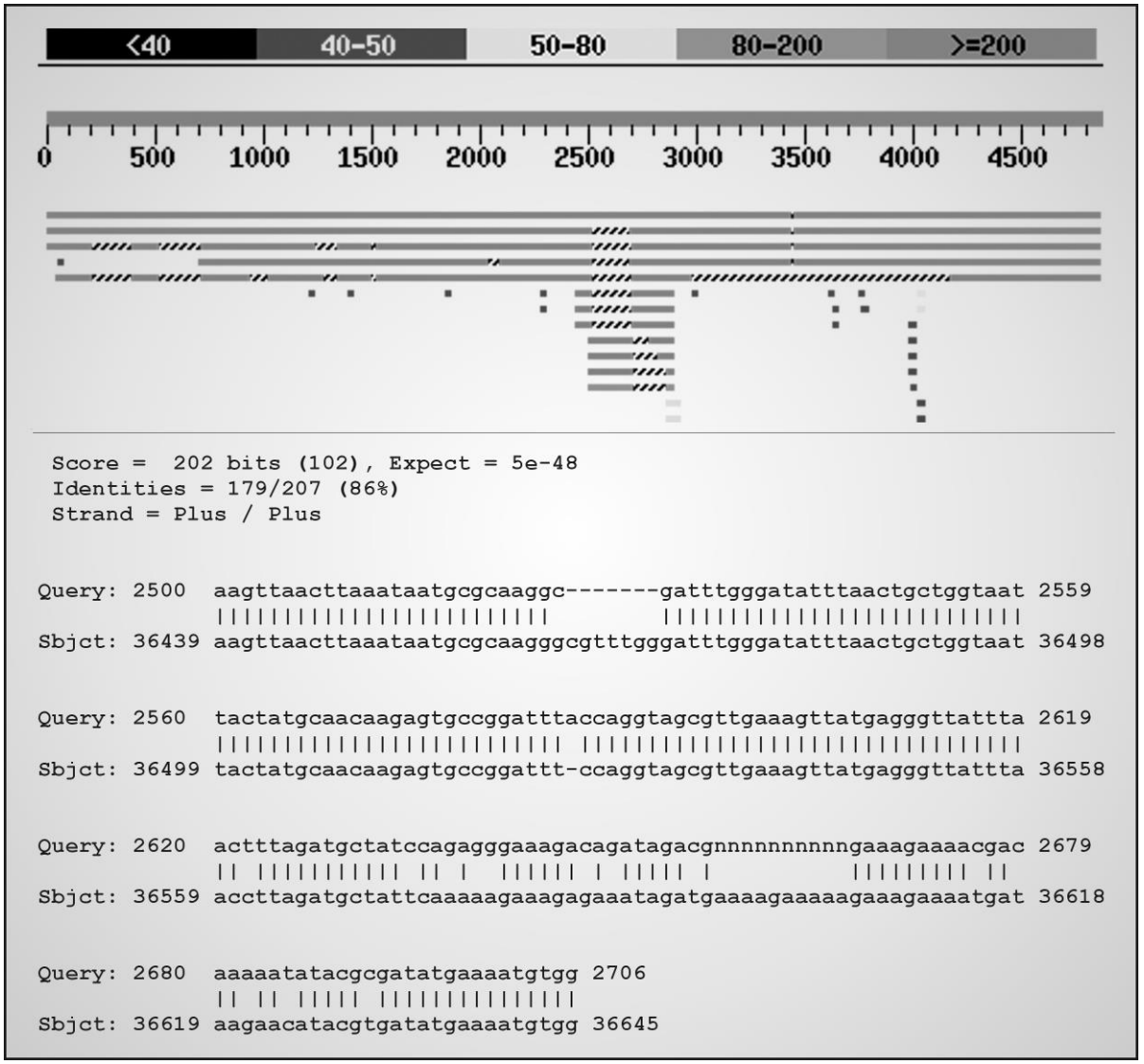
```
>gi|127552|sp|P23367|MUTL_ECOLI  DNA mismatch repair protein mutL
          Length = 615

 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 LGADQQPAFVLYLEIDPHQVDVNVHPAKHEVRFHQSRLVHDFIYQGVLSVLQQQLETPL 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  GSNSSRMYFTQTLLPGLAGPSGEMVKXXXXXXXXXXXXXXDKVYAHQMVRTDSREQKLDA 120
           GSNSSRMYFTQTLLPGLAGPSGEMVK              DKVYAHQMVRTDSREQKLDA
Sbjct: 336 GSNSSRMYFTQTLLPGLAGPSGEMVKSTTSLTSSSTSGSSDKVYAHQMVRTDSREQKLDA 395

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

sekvence s nízkou komplexitou

BLAST – příklad výstupu u DNA




```
Score =  202 bits (102), Expect = 5e-48
Identities = 179/207 (86%)
Strand = Plus / Plus


Query: 2500  aagttaacttaaataatgcgcaaggc-------gatttgggatatttaactgctggtaat 2559
              |||||||||||||||||||||||||       |||||||||||||||||||||||||||
Sbjct: 36439 aagttaacttaaataatgcgcaagggcgtttgggatttgggatatttaactgctggtaat 36498


Query: 2560  tactatgcaacaagagtgccggatttaccaggtagcgttgaaagttatgagggttattta 2619
              |||||||||||||||||||||||||| |||||||||||||||||||||||||||||||
Sbjct: 36499 tactatgcaacaagagtgccggattt-ccaggtagcgttgaaagttatgagggttattta 36558


Query: 2620  actttagatgctatccagagggaaagacagatagacgnnnnnnnnnngaaagaaaacgac 2679
              || |||||||||| || |  |||||| | ||||| |          ||||||||| ||
Sbjct: 36559 accttagatgctattcaaaaagaaagagaaatagatgaaaagaaaaagaaagaaaatgat 36618


Query: 2680  aaaaatatacgcgatatgaaaatgtgg 2706
              || || ||||| ||||||||||||||
Sbjct: 36619 aagaacatacgtgatatgaaaatgtgg 36645
```

Aplikace pro lokální přiložení na EBI

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.

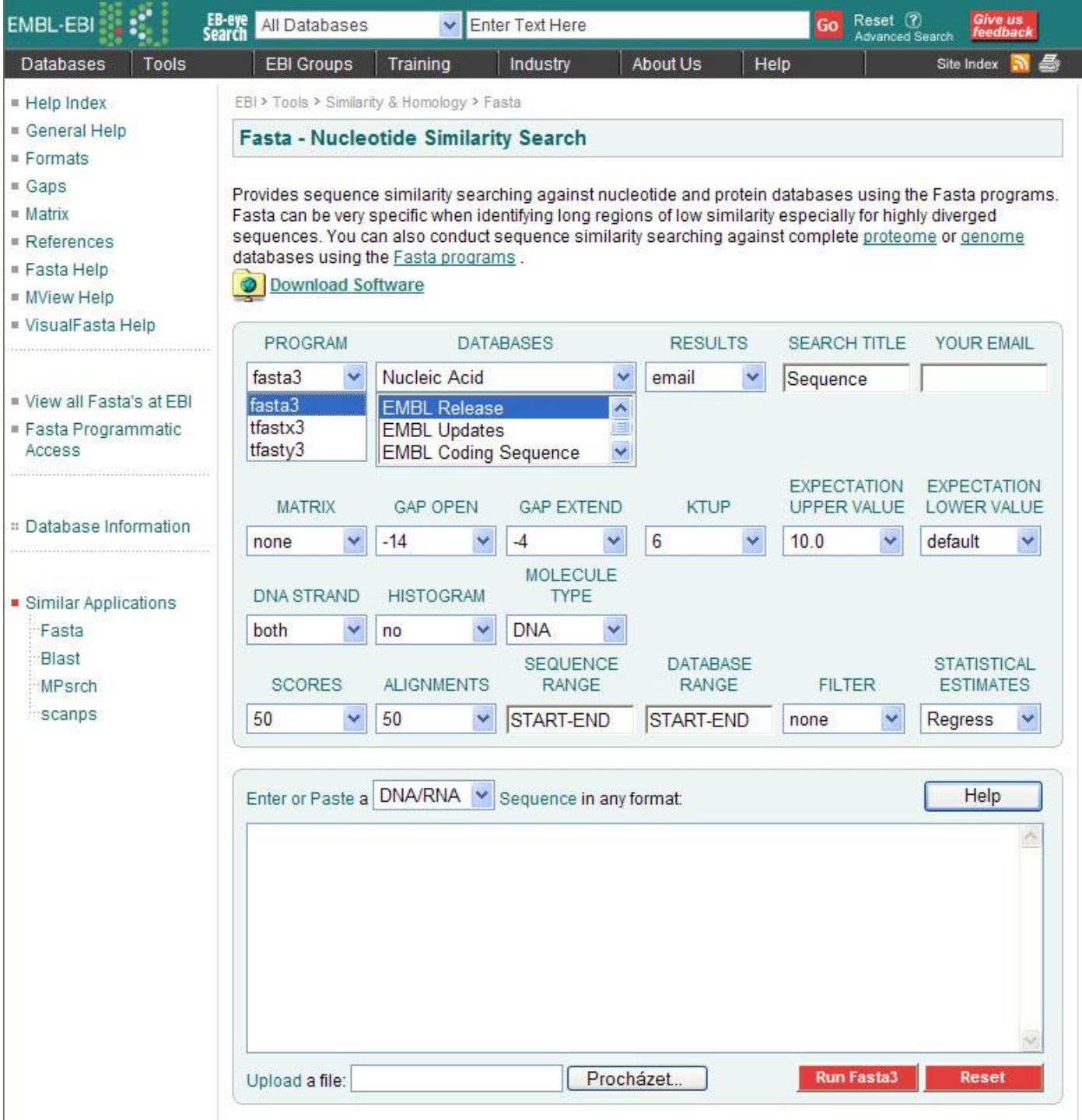
Protein

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.

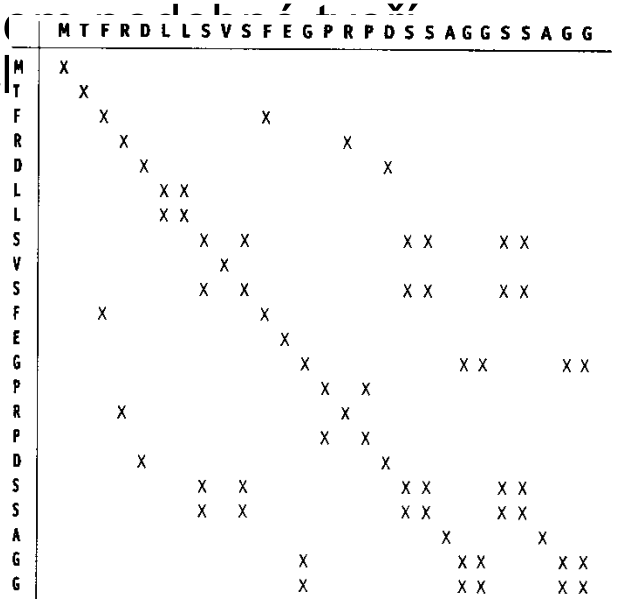
Try it out at ENA Sequence Search.

Fasta3 (EBI)

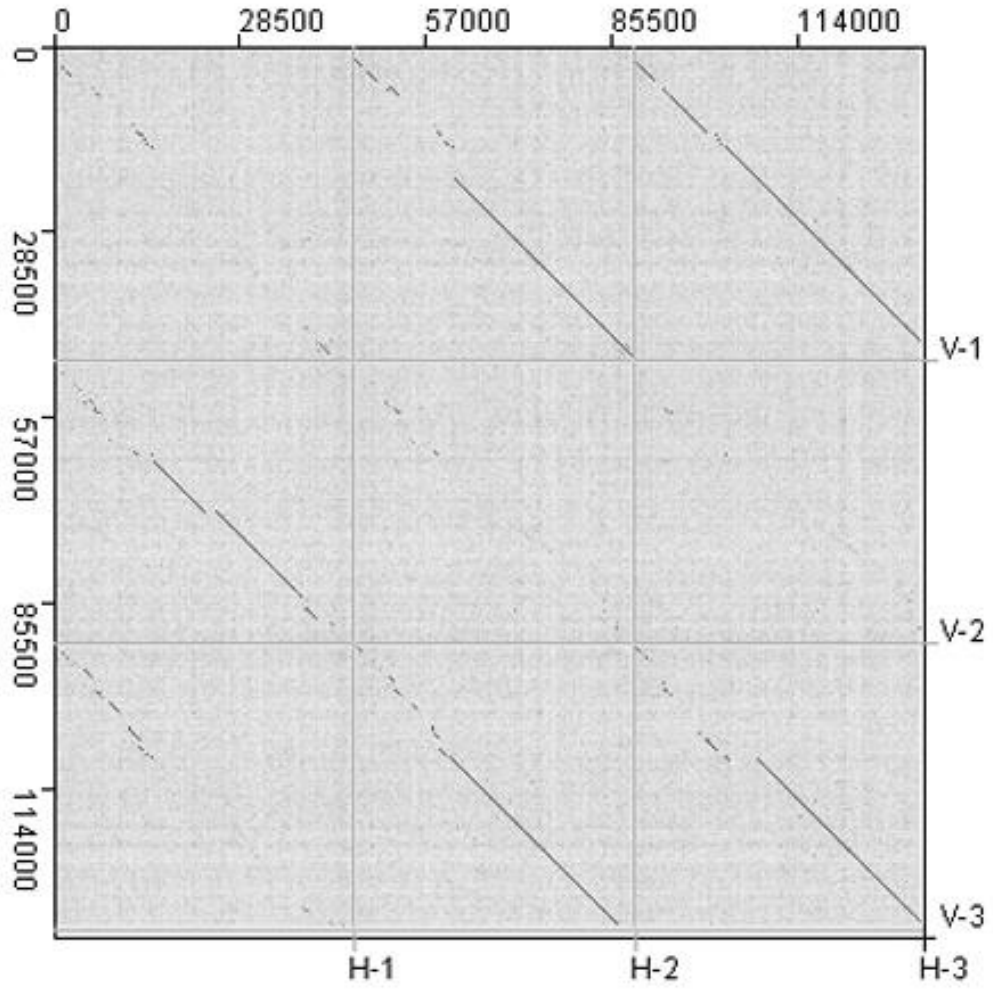

EMBL-EBI
EB-eye Search
All Databases
Enter Text Here
Go
Reset
Advanced Search
Give us feedback
Databases
Tools
EBI Groups
Training
Industry
About Us
Help
Site Index
Help Index
General Help
Formats
Gaps
Matrix
References
Fasta Help
MView Help
VisualFasta Help
View all Fasta's at EBI
Fasta Programmatic Access
Database Information
Similar Applications
Fasta
Blast
MPsrch
scanps
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
Nucleic Acid
email
Sequence
fasta3
tfastx3
tfasty3
EMBL Release
EMBL Updates
EMBL Coding 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
DNA/RNA
Sequence in any format:
Help
Upload a file:
Procházet...
Run Fasta3
Reset

Metoda tečkové (Dot-Plot) matice

- Bodový diagram vzájemné podobnosti sekvencí - nejjednodušší pomůcka pro grafické znázornění oblastí podobnosti.
- Horizontální a vertikální rozměry odpovídají porovnávaným sekvencím.
- 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 navzáj[illegible] řádek vysokých skóre diagonální linii přes teč[illegible]
- 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



Typy sekvenčních přiložení

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

  LOKÁLNÍ – nejlepší sekvenční přiložení segmentů bez ohledu na zbytek sekvence

  Smithův-Watermanův algoritmus

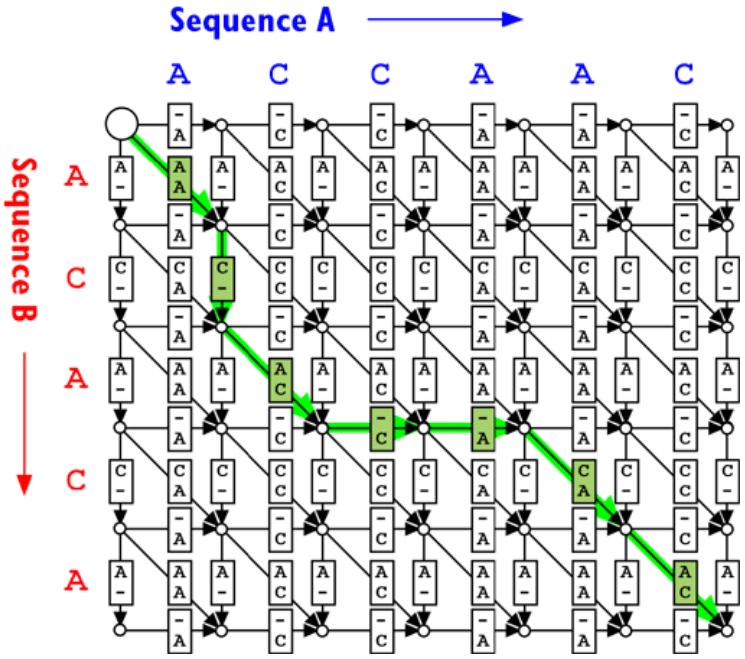
- Pro optimální globální alignment požadujeme dosažení nejlepšího skóre v celém řádku/sloupci

  GLOBÁLNÍ – nejlepší sekvenční přiložení celých sekvencí

  Needlemanův-Wunschův algoritmus

Sekvenční přiložení může posuzovat podobnost celých dlouhých sekvencí

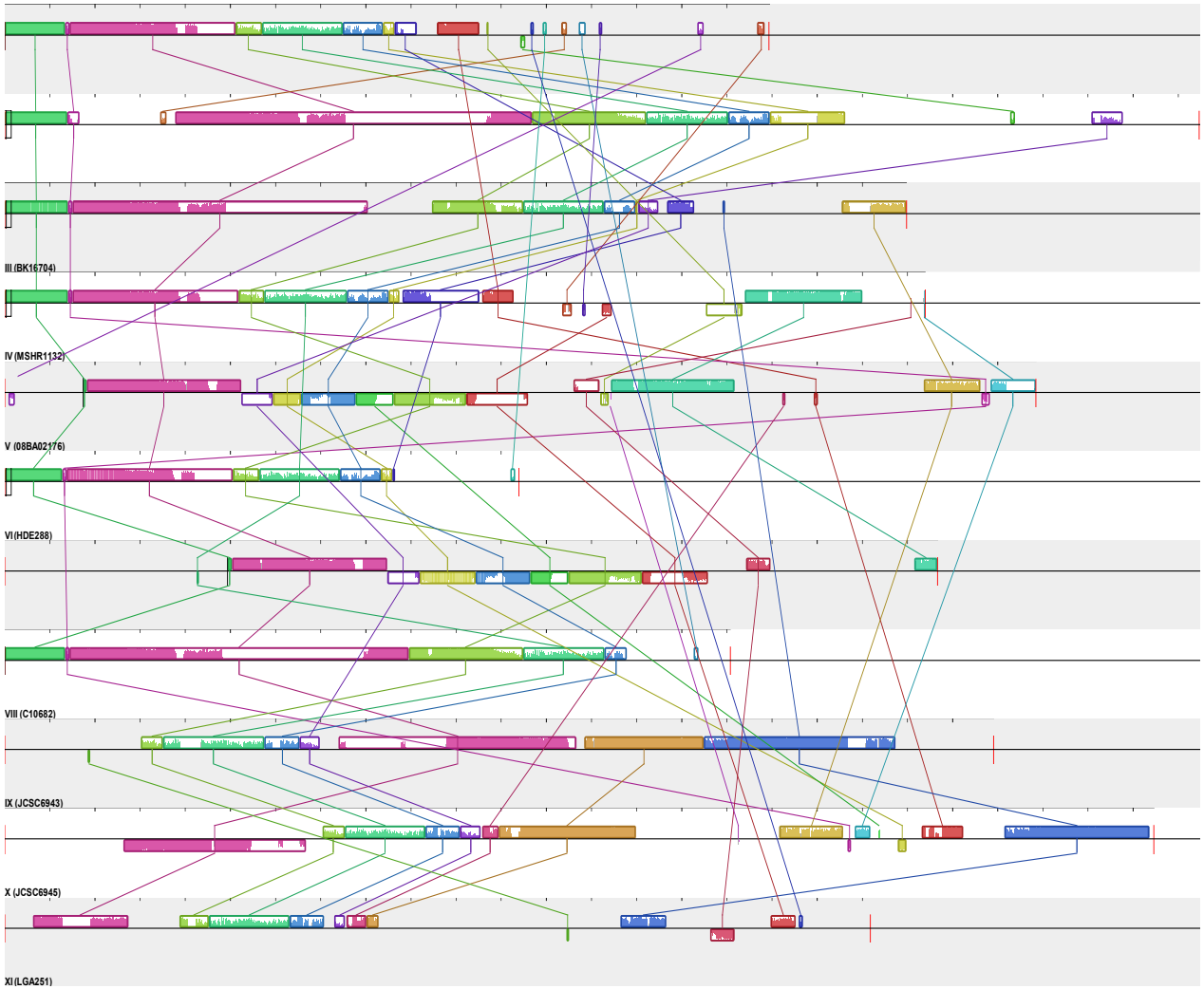
Nalezení nejefektivnější transformace jedné sekvence do druhé představuje využití dynamického programování pro konstrukci přiložení



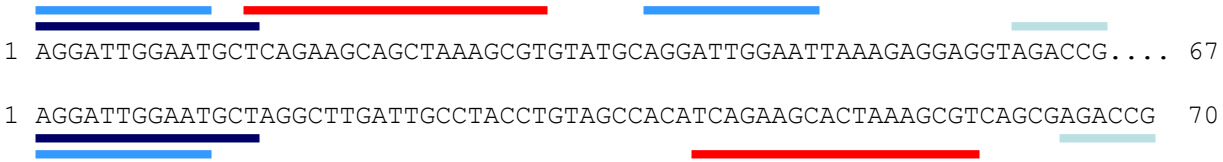
Seq B	ACA--CA
	|
Seq A	A-CCAAC

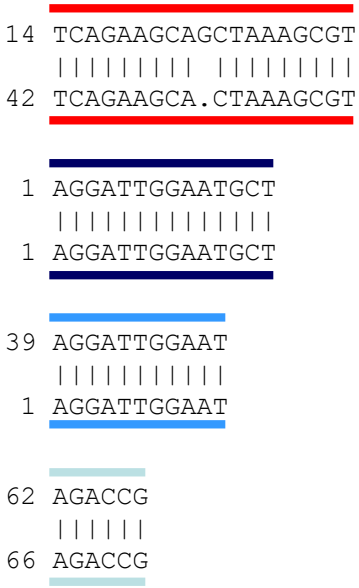
- Bodové změny, delece
- Inverze
- Translokace
- Duplikace
- Kombinace uvedených změn

Příklad globálního přiložení - MAUVE

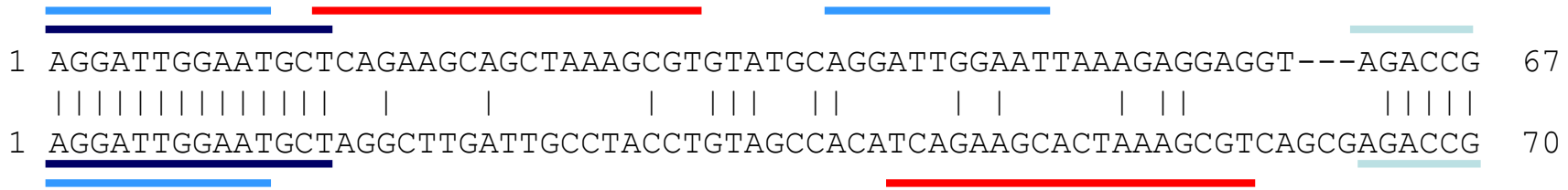


Lokální sekvenční přiložení





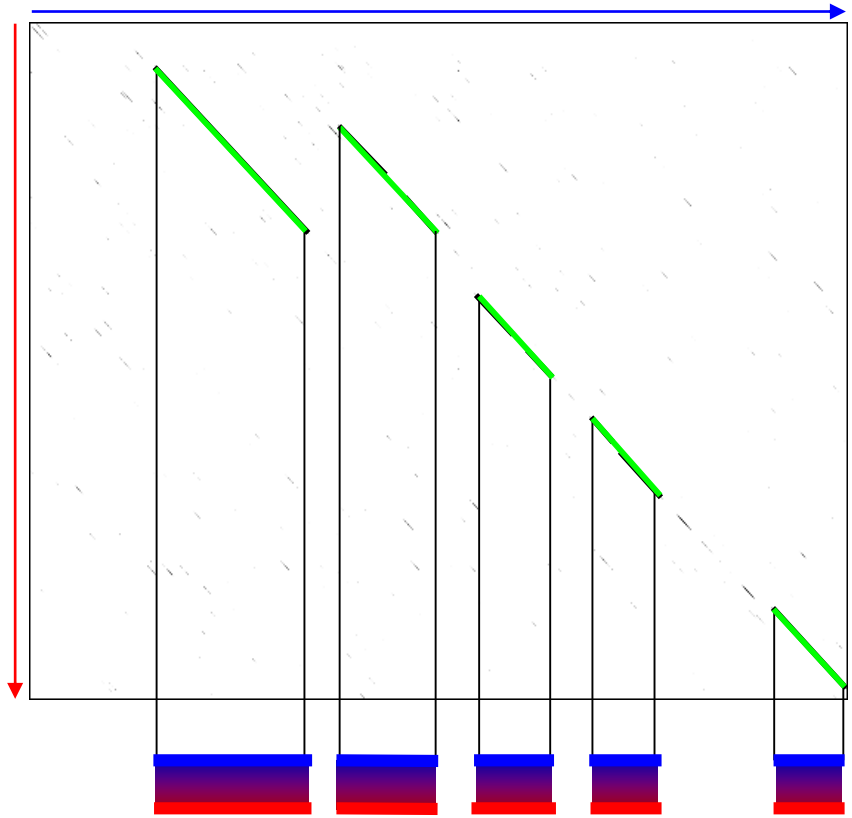
Globální sekvenční přiložení



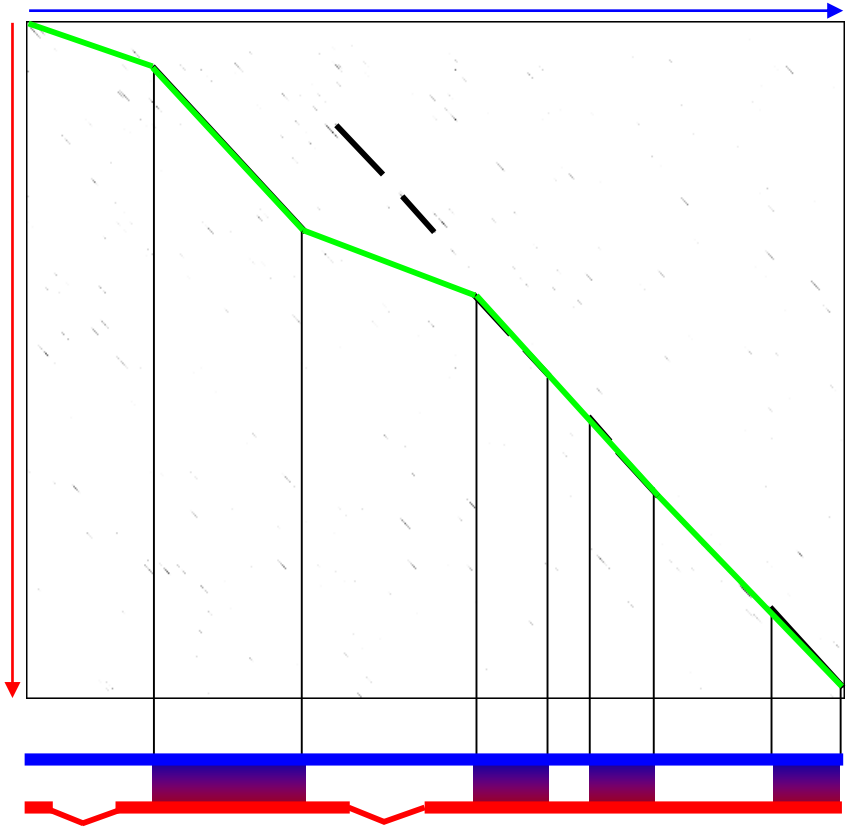
Dvě sekvence sdílejí oblasti s lokální podobností (end-to-end alignment)

Lokální vs. globální přiložení

Lokální:
- Citlivé k modulární struktuře
- Vhodné k hledání v databázích



Globální:
- Výsledek ovlivněn vkládáním mezer
- Neuvažuje modulární strukturu sekvencí



Poznámky na závěr

- Substituční matice a penalizace mezer vnáší do algoritmů pro přiložení biologický význam
- Existuje mnoho způsobů, jak přiložit dvě sekvence
- Přiložení ještě neznamená, že dvě sekvence sdílejí společnou biologickou historii. Významnost musí být posouzena statistickým skóre.
- Nevěřte, že vaše přiložení je jediné správné, zejména u sekvencí, které mají méně než 20% podobnost. K prozkoumání těchto sekvencí jsou potřebné další metody.