

Bioinformatika 1

Heurističko poravnavanje. BLAST

Mirjana Domazet-Lošo
FER, 2021./2022.



Creative Commons Imenovanje-Nekomercijalno-Dijeli pod istim uvjetima 3.0

Zašto pretraživati baze podataka sljedova?

Usporedbe gene, genoma, proteina, proteoma

- novi genom
 - sličnost s prethodno sekvenciranim genomima
 - identifikacija gena
- kada je pronađen gen u genomu
 - postoji li taj gen u nekom drugom organizmu?
 - je li poznata funkcija tog gena u drugom organizmu?
 - za novootkriveni protein odrediti funkciju na temelju sličnosti s proteinima kojima je poznata funkcija

Pretraživanje baza podataka sljedova

Osnovna ideja

- baza podataka se pretražuje kako bi se pronašli slični sljedovi
- određivanje sličnosti između sljedova
→ poravnavanjem
(najčešći postupak u bioinformatici)

Usporedbe velikih količina podataka (1)

- *exhaustive search*

→ ispitivanje svih mogućih rješenja kako bi se pronašlo optimalno rješenje

- dinamičko programiranje

- poravnanje 2 niza duljine n i m :

Smith-Watermanov, Needleman–Wunschov algoritam

- vremenska složenost: **$O(nm)$**

→ presporo?

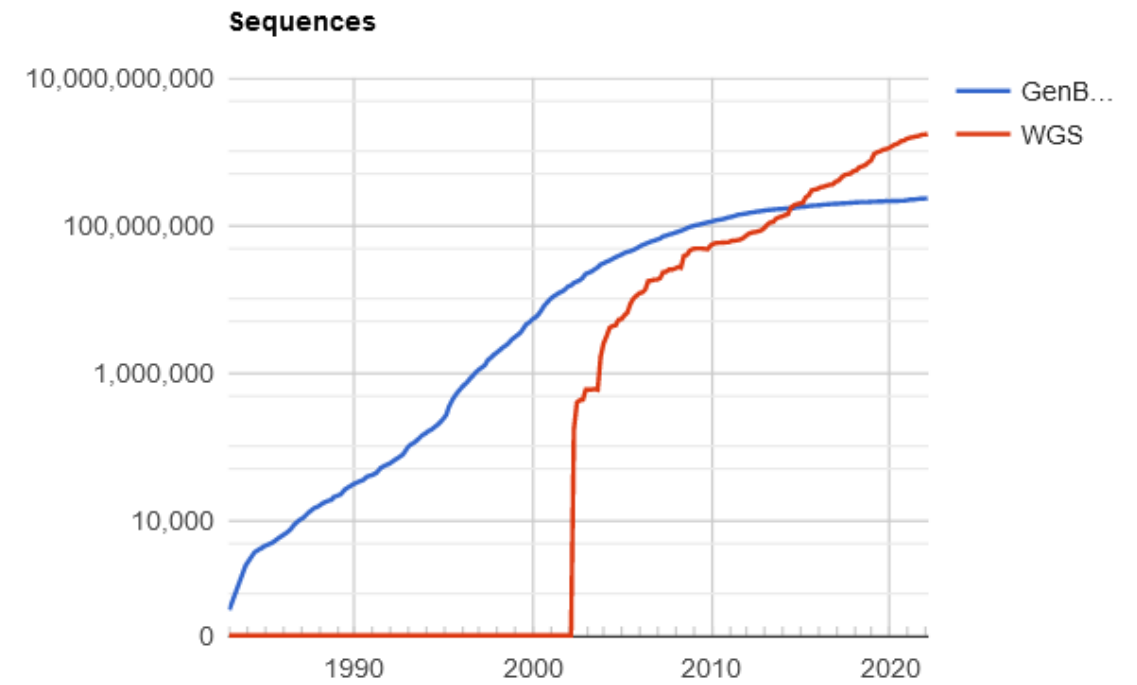
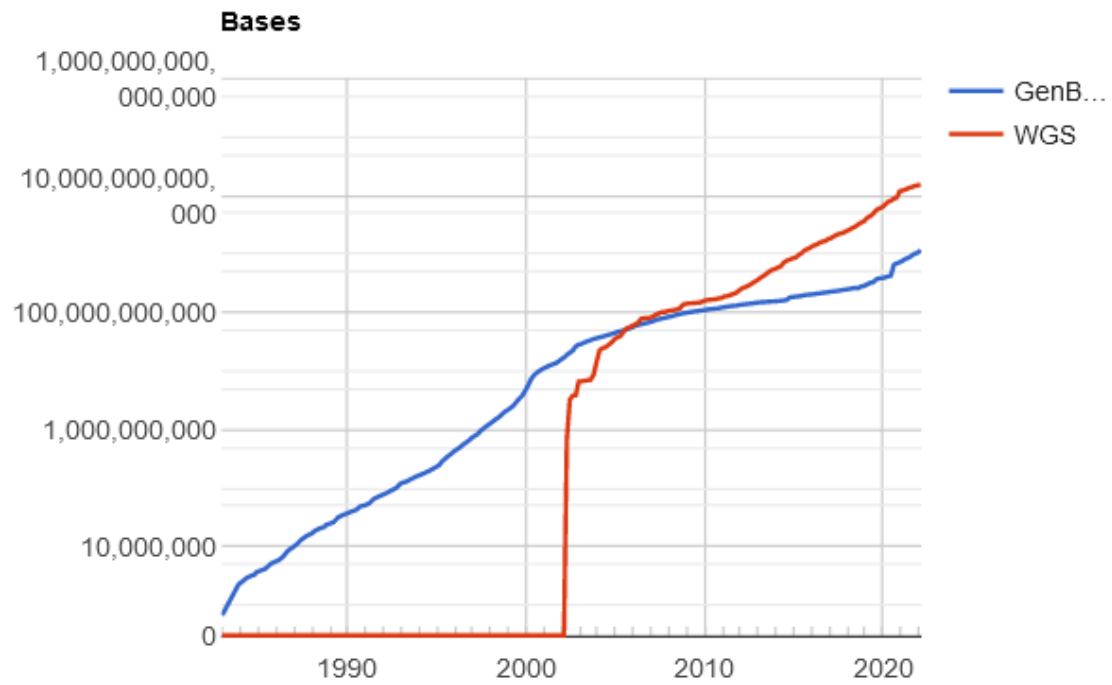
Usporedbe velikih količina podataka (2)

- Primjer:

Usporedba (i) gena (ii) genoma

s bazom podataka nukleotida, npr. *GenBank*

- duljina gena (čovjek) $\approx 10^3\text{bp}$
- duljina ljudskog genoma $\approx 3 \cdot 10^9\text{bp}$
- veličina GenBank (NCBI-GenBank, 2./2022.): **$1.1 \cdot 10^{12}\text{bp}$**
- veličina WGS (NCBI-GenBank, 2./2022.): **$15 \cdot 10^{12}\text{bp}$**
- <http://www.ncbi.nlm.nih.gov/genbank/statistics>
- <http://www.ncbi.nlm.nih.gov/genbank/wgs>



NCBI (National Center for Biotechnology Information)
<http://www.ncbi.nlm.nih.gov/genbank/statistics>

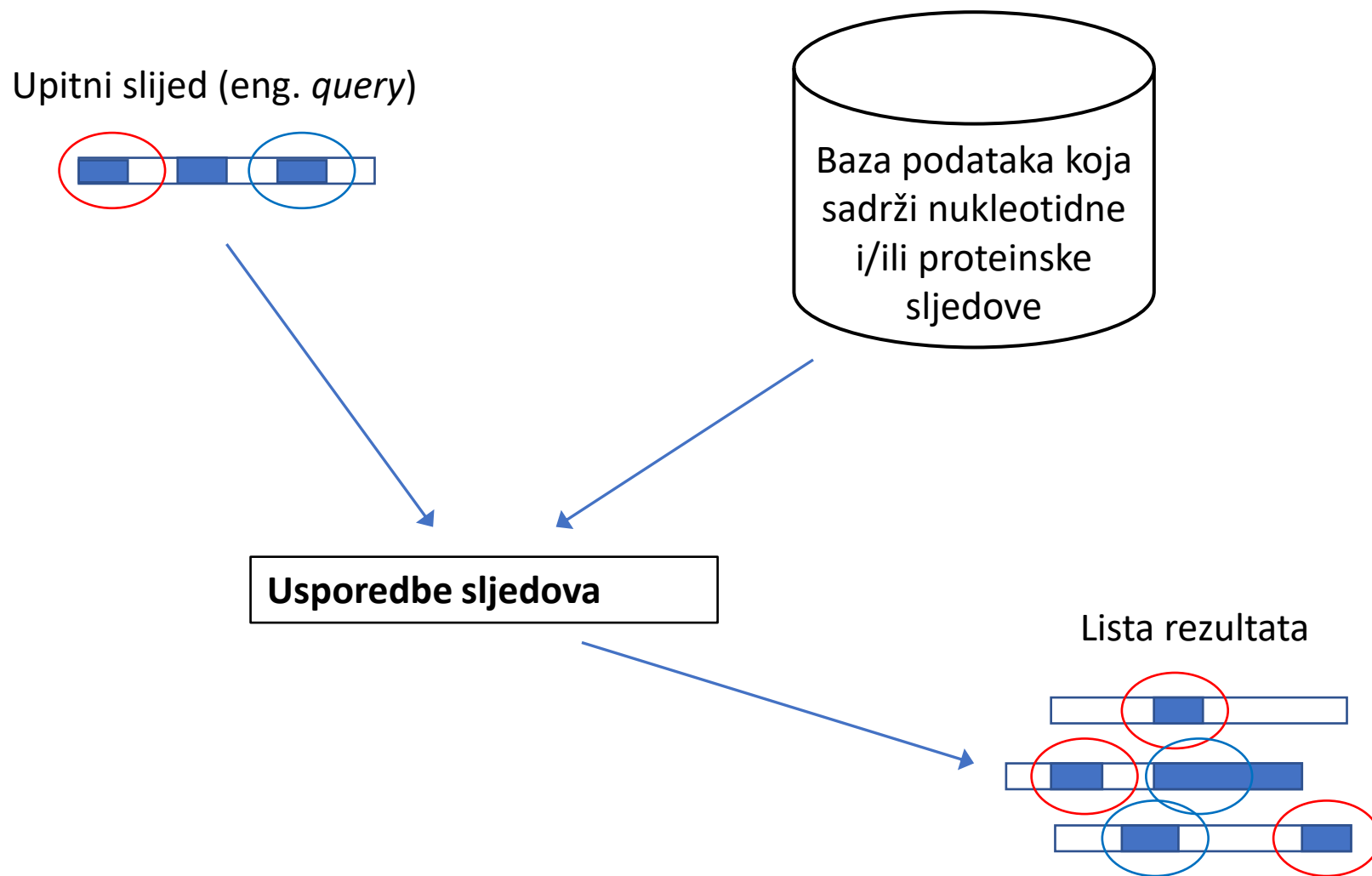
Heuristički pristup

- pronalazak optimalnog rješenja nije zajamčen, koristi se heuristika
- odvagivanje između osjetljivosti i brzine (eng. *sensitivity vs speed*)
- metode temeljene na pronalasku identičnih ili skoro identičnih riječi (eng. *words*) ili *k*-torki (eng. *k-tuples*) između parova sljedova:
 - **BLAST** (Altschul *et al.*, 1990. *Basic local alignment search tool*)
 - **FASTA** (Lipman & Pearson, 1985. *Rapid and sensitive protein similarity searches*)

BLAST - citiranost

- Citiranost (*Web of Knowledge*, ožujak 2021.)
 - BLAST (*Basic Local Alignment Search Tool*; Altschul *et al.*, 1990):
61 934 citata
 - Gapped BLAST & PSI-BLAST (Altschul *et al.*, 1997):
54 299 citata
 - usporedba:
48 911 citata (10./2019.)
44 061 citata (10./2017.)
40 016 citata (10./2015.)

Pretraživanje baze podataka



Program BLAST

- BLAST (*Basic Local Alignment Search Tool*)
 - glavni alat NCBI-a za usporedbe proteinskih i nukleotidnih sljedova NCBI-a (*National Center for Biotechnology Information*)
- Lokalno poravnanje
 - upitni slijed (ili njegovi dijelovi) se poravnavaju sa sljedovima u bazi podataka (Altschul *et al.*, 1990; 1997)

BLAST – ideja

- Ideja
 - uspoređivanje upitnog slijeda (eng. *query*) s bazom podataka sljedova (eng. *target*)
 - rezultat:
lista najboljih podudaranja (eng. *matches*)

BLAST – postupak pretraživanja BP


1. odabir BLAST programa
 - blastp, blastn, blastx, tblastx, tblastn
2. odabir upitnog slijeda (eng. *query*)
 - *accession number*, *gi* identifikator, fasta format
3. odabir baze podataka za pretraživanje
 - najčešće: *nonredundant database* (*nr*)
4. odabir parametara za pretraživanje
 - pregled prikazanih rezultata

BLAST - programi

Program	Upit	Baza podataka
blastp	Protein	Protein
blastn	DNA	DNA
blastx	DNA* (6 mogućih čitanja)	Protein
tblastn	Protein	DNA* (6 mogućih čitanja)
tblastx**	DNA* (6 mogućih čitanja)	DNA* (6 mogućih čitanja)

- **P** → protein
- **N** → nukleotid
- **X** → DNA dinamički translirana u 6 proteinskih sljedova
- **T** → DNA bazu podataka se pretražuje prema 6 okvira čitanja
- ****tblastx** – 36 mogućih protein-protein kombinacija

6 okvira čitanja

- Primjer: 
AAACCCGGG
TTTGGGCCC

Gornji lanac - 3 okvira čitanja:

AAA CCC GGG → Lys Pro Gly

AAC CCG → Asn Pro

ACC CGG → Thr Arg

Donji lanac - 3 okvira čitanja:

CCC GGG TTT → Pro Gly Phe

CCG GGT → Pro Gly

CGG GTT → Arg Val

Npr. za standardni genetički kod vidjeti: http://en.wikipedia.org/wiki/Genetic_code

BLAST: Basic Local Alignment Search Tool

blast.ncbi.nlm.nih.gov/Blast.cgi

Most Visited Getting Started

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI [Sign In] [Register]

NCBI/ BLAST Home

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

New DELTA-BLAST, a more sensitive protein-protein search [Go](#)

BLAST Assembled RefSeq Genomes

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

Human	Oryza sativa	Gallus gallus
Mouse	Bos taurus	Pan troglodytes
Rat	Danio rerio	Microbes
Arabidopsis thaliana	Drosophila melanogaster	Apis mellifera

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
protein blast	Search protein database using a protein query <i>Algorithms: blastp, psi-blast, phi-blast, delta-blast</i>
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)

News

[Update to SRA-BLAST](#)

SRA-BLAST has undergone a dramatic update, both in terms of user interface and search performance.
Thu, 20 Jun 2013 11:00:00 EST

[More BLAST news...](#)

Tip of the Day

[Use Genomic BLAST to see the genomic context](#)

If you are interested in the evolution of a particular gene or gene family it is often interesting to examine the intro-exon structure even across species.

[More tips...](#)

BLAST – primjer (1)

Nucleotide BLAST: Search nucleotide da... +

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&BLAST_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome

Most Visited Getting Started

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastn suite

Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

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

Enter Query Sequence

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

From

To

TGACAGTCAACTAGCACACAGACAGGGCCCGGAACATCCGGAGTTTACAAAGAC
TGCTGACACAGAGGGACTTTCCGCGGGGACTTTCCACTGGGGCGTTCTAGGAGGTGTGGT
CTGGCGGACTGGGAGTGGTCAACCCCTCAAATGCTGCATATAAGCAGCTGCTTTTCGCGTG
TACTGGGTCTCTCTAGTCAGACCATCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGAAC
CCACTGCTTAAGCCTCAATAAGCTTGCCCTGAGGGGCTAGAGCGGCGCCACCGCGGTGG
AGCTCCAGCTTTTGTCCCTTTAGTGAGGTTAATTGCGCGCTGGCGATC

Or, upload file [Pretraži...](#) Datoteka nije odabrana.

Job Title

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database ☐ Human genomic + transcript ☐ Mouse genomic + transcript ☒ Others (nr etc.):
Nucleotide collection (nr/nt)

Organism [Optional](#)
 Enter organism name or id--completions will be suggested ☐ Exclude +
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude [Optional](#)
☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

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

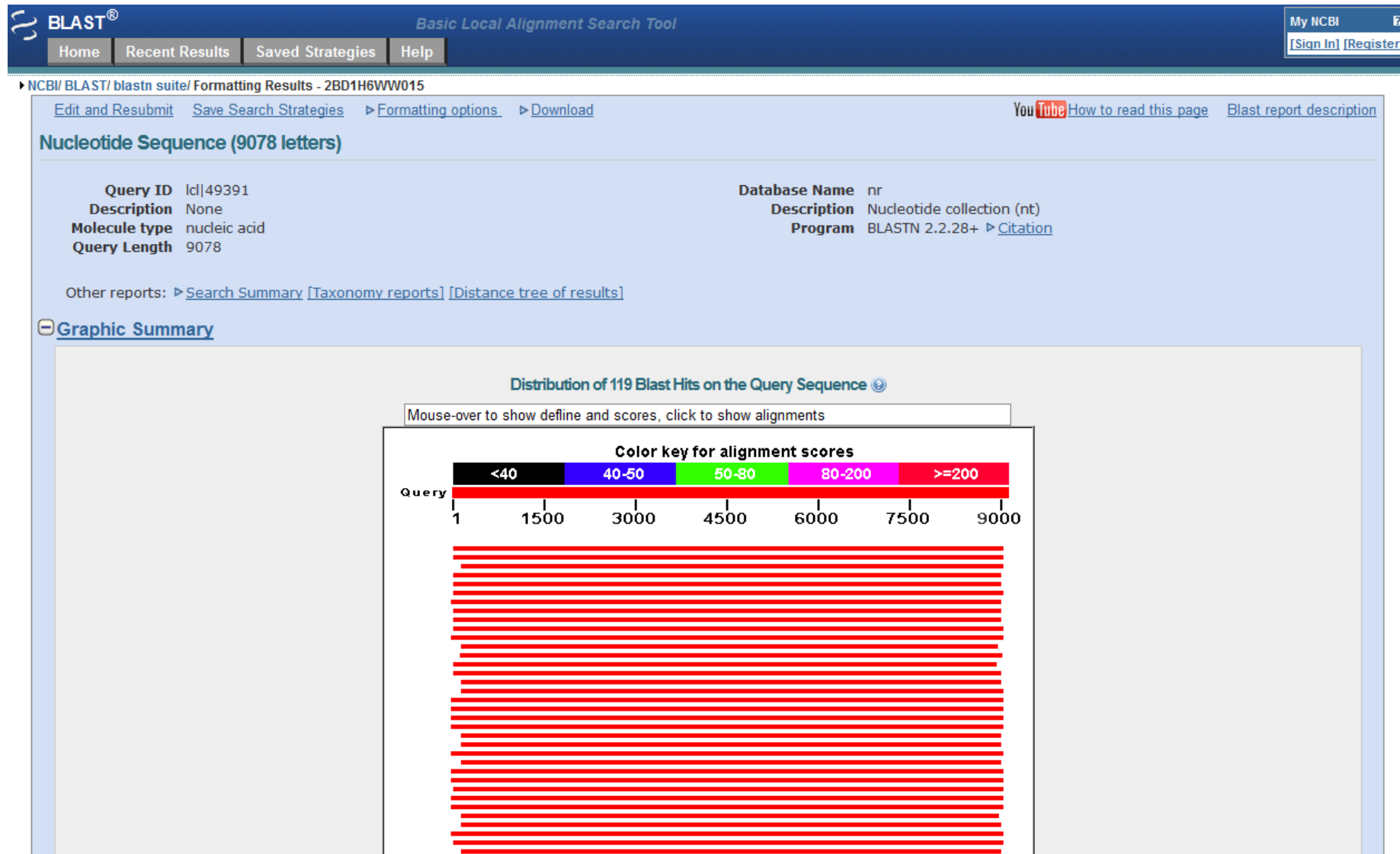
Program Selection

Optimize for ☒ Highly similar sequences (megablast)
☐ More dissimilar sequences (discontiguous megablast)
☐ Somewhat similar sequences (blastn)
Choose a BLAST algorithm

BLAST Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)
☐ Show results in a new window

[+ Algorithm parameters](#)

BLAST – primjer (2)



BLAST – primjer (3)

NCBI Blast:Nucleotide Sequence (9078 le... +

blast.ncbi.nlm.nih.gov/Blast.cgi

Most Visited Getting Started

Descriptions

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

[Alignments](#) [Download](#) [GenBank](#) [Graphics](#) [Distance tree of results](#)

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	HIV-1 strain 97CN001 from China, complete genome	16188	16188	98%	0.0	99%	AF286226.1
<input type="checkbox"/>	HIV-1 strain 98CN009 from China, complete genome	15653	15653	98%	0.0	98%	AF286230.1
<input type="checkbox"/>	HIV-1 strain CNGL179 from China, complete genome	15531	15531	97%	0.0	98%	AF503396.1
<input type="checkbox"/>	HIV-1 isolate Sichuan_2006_SC025 from China qag protein (qag) gene, complete cds; pol protein (pol) gene, partial cds; vif protein (vif), vpr prot	14846	14846	98%	0.0	97%	JX392381.1
<input type="checkbox"/>	HIV-1 clone XJN0084 from China, complete genome	14713	14713	98%	0.0	96%	EF368371.1
<input type="checkbox"/>	HIV-1 isolate Xinjiang_2006_713 from China qag protein (qag) gene, complete cds; pol protein (pol) gene, partial cds; vif protein (vif), vpr protein	14694	14694	98%	0.0	96%	JX392384.1
<input type="checkbox"/>	HIV-1 strain pXJDC6441-2 from China, complete genome	14694	14694	98%	0.0	96%	EF420986.1
<input type="checkbox"/>	HIV-1 isolate Sichuan_2006_SC008 from China qag protein (qag) gene, complete cds; pol protein (pol) gene, partial cds; vif protein (vif), vpr prot	14670	14670	98%	0.0	96%	JX392379.1
<input type="checkbox"/>	HIV-1 clone XJDC6441 from China, complete genome	14628	14628	98%	0.0	96%	EF368370.1
<input type="checkbox"/>	HIV-1 isolate CNGZD from China, partial genome	14421	14421	98%	0.0	96%	JQ423923.1
<input type="checkbox"/>	HIV-1 strain NLXJDC6441X2 from China, complete genome	14368	14917	99%	0.0	96%	EF420987.1
<input type="checkbox"/>	HIV-1 isolate TW_D60 from Taiwan qag protein (qag) gene, partial cds; nonfunctional pol protein (pol) gene, partial sequence; and vif protein (vif)	14200	14200	96%	0.0	96%	DQ230842.1
<input type="checkbox"/>	HIV-1 isolate 1114 from China, partial genome	14176	14176	97%	0.0	96%	HQ215552.1
<input type="checkbox"/>	HIV-1 isolate Xinjiang_2006_709 from China qag protein (qag) gene, complete cds; pol protein (pol) gene, partial cds; vif protein (vif), vpr protein	14111	14111	97%	0.0	95%	JX392383.1
<input type="checkbox"/>	HIV-1 clone XJDC6431-2 from China, complete genome	13965	13965	98%	0.0	95%	EF368372.1
<input type="checkbox"/>	Human immunodeficiency virus 1 genomic RNA, complete genome, isolate:HH069	13586	13586	97%	0.0	95%	AP005206.1
<input type="checkbox"/>	Human immunodeficiency virus 1 genomic RNA, complete genome, isolate:HH086	13568	13568	97%	0.0	94%	AP005207.1
<input type="checkbox"/>	HIV-1 isolate pBRGX from China, complete genome	13542	14435	99%	0.0	94%	JF719818.1
<input type="checkbox"/>	Human immunodeficiency virus 1 proviral DNA, nearly complete genome, clone: p00CH-HH090_08_BC30	13505	14404	99%	0.0	94%	AB773885.1
<input type="checkbox"/>	HIV-1 isolate 2007CNGX-HK from China, complete genome	13501	14400	99%	0.0	94%	JF719819.1

BLAST – primjer (4)

NCBI Blast:Nucleotide Sequence (9078 le... +

blast.ncbi.nlm.nih.gov/Blast.cgi#alnHdr_13569237

Most Visited Getting Started

Download ▾ GenBank Graphics

HIV-1 strain 97CN001 from China, complete genome
Sequence ID: [gb|AF286226.1|AF286226](#) Length: 8978 Number of Matches: 1

Range 1: 2 to 8972 [GenBank](#) [Graphics](#) ▾ Next Match ▲ Previous Match

	Score	Expect	Identities	Gaps	Strand
	16188 bits(8766)	0.0	8904/8972(99%)	3/8972(0%)	Plus/Plus
Query 32	TGAAAGCGAAAGTAAGACCAGAGGAGATCTCTCGACGCAGGACTCGGCTTGCTGAAGTGC	91			
Sbjct 2	TGAAAGCGAAAGTAAGACCAGAGGAGATCTCTCGACGCAGGACTCGGCTTGCTGAAGTGC	61			
Query 92	ACTCGGCAAGAGGCGAGAGCGGCGACTGGTGAGTACGCCAATTATATTGACTAGCGGAG	151			
Sbjct 62	ACTCGGCAAGAGGCGAGAGCGGCGACTGGTGAGTACGCCAATTATATTGACTAGCGGAG	121			
Query 152	GCTAGAAGGAGAGAGATGGGTGCGAGAGCGTCAATATTAAGAGGGGGAAAATTAGATAAA	211			
Sbjct 122	GCTAGAAGGAGAGAGATGGGTGCGAGAGCGTCAATATTAAGAGGGGGAAAATTAGATAAA	181			
Query 212	TGGGAAAAAATTAGGTTAAGGCCAGGGGGAAAGAAACACTATATGCTAAAACACCTAGTA	271			
Sbjct 182	TGGGAAAAAATTAGGTTAAGGCCAGGGGGAAAGAAACACTATATGCTAAAACACCTAGTA	241			
Query 272	TGGGCAAGCAGGGAGCTGGAAAGATTTGCACTTAACCCTGGCCTTTTAGAGACATCAGAA	331			
Sbjct 242	TGGGCAAGCAGGGAGCTGGAAAGATTTGCACTTAACCCTGGCCTTTTAGAGACATCAGAA	301			
Query 332	GGCTGTAAACAAATAATGAAACAGCTACAATCAGCTCTTCAGACAGGAACAGAGGAACTT	391			
Sbjct 302	GGCTGTAAACAAATAATGAAACAGCTACAACCAGCTCTTCAGACAGGAACAGAGGAACTT	361			

BLAST – parametri pretraživanja (1)

Algorithm parameters

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 | 28 [?](#)

Max matches in a query range | 0 [?](#)

Scoring Parameters

Match/Mismatch Scores | 1,-2 [?](#)

Gap Costs | Linear [?](#)

Filters and Masking

Filter | ☒ Low complexity regions [?](#)
☐ Species-specific repeats for:

Mask | ☒ Mask for lookup table only [?](#)
☐ Mask lower case letters [?](#)

BLAST – parametri pretraživanja (2)

- očekivani prag (eng. *expect threshold*): *E-value*
 - broj pogodaka (zapisa) čiji je rezultat $\geq S$ (eng. *score*) za koji se očekuju da će biti pronađen u bazi podataka slučajnim putem
 - npr. rezultat $S = 18$ uz $E = 0.025$ znači da se, pretraživanjem baze podataka uz zadane parametre, očekuje 0.025 pojavljivanja rezultata $S \geq 18$
 - obično se kao statistički pouzdano uzima $E \leq 0.05$, ali ...

BLAST – parametri pretraživanja (3)

- duljina riječi (eng. *word size*)
 - kod pretraživanja, upitni slijed se dijeli u kraće sljedove/*riječi* (eng. *words*) zadane duljine (eng. *word size*), prema kojima se onda određuje sličnost sa sljedovima u bazi podataka
 - inicijalne vrijednosti:
 - za amino kiseline: 3 (2 za vrlo kratke peptide)
 - za nukleotidne sljedove: 11 (7 - 15)
- kratki upiti (eng. *short queries*)
 - ako se odabere ova opcija, onda su duljina riječi i prag očekivanja automatski postavljeni

BLAST - parametri pretraživanja (4)

- supstitucijska matrica (eng. *substitution matrix*) – za proteinske sljedove
 - inicijalno: BLOSUM 62
 - može se izabrati:
PAM30, PAM70, PAM250, BLOSUM45, BLOSUM80, itd.
- kazne za uvođenje procijepa (praznina) (eng. *gap costs*)
 - tipično: pojava (otvaranje) procijepa kažnjava se više od proširenja procijepa
 - linearni model (Ln)
 - afini model ($G + L(n-1)$)

BLAST - parametri pretraživanja (5)

- statistika temeljena na nukleotidnom/aminokiselinskom sastavu sljedova (eng. *composition-based statistics*)
 - neki proteini imaju atipičan sastav nukleotida ili aminokiselina, koji se onda kompenzira uključivanjem ove opcije

Primjer:

- *Plasmodium falciparum* ima 80.6% AT nukleotida
- neki proteini sadrže visok postotak hidrofobnih dijelova
- popravljaja *E* statistiku

BLAST - parametri pretraživanja (6)

- filtri i maske (eng. *filters and masking*)
 - filtriranje se primijenjuje na upit, a ne na cijelu bazu podataka
 - filtriraju se dijelovi niske složenosti (eng. *low complexity*), npr. dinukleotidna ponavljanja (npr. $(AT)_n$)
 - za nukleotide: program DUST
 - za aminokiseline: program SEG
 - filtriranje ponavljanja – kako bi se izbjegla lažna podudaranja
 - maskiranje malih slova
 - samo se velika slova koriste za pretraživanje baze podataka

BLAST – parametri ispisa

Alignments Download GenBank Graphics Distance tree of results							
	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	HIV-1 strain 97CN001 from China, complete genome	16188	16188	98%	0.0	99%	AF286226.1
<input type="checkbox"/>	HIV-1 strain 98CN009 from China, complete genome	15653	15653	98%	0.0	98%	AF286230.1
<input type="checkbox"/>	HIV-1 strain CNG1179 from China, complete genome	15531	15531	97%	0.0	98%	AF503396.1
<input type="checkbox"/>	HIV-1 isolate Sichuan_2006_SC025 from China gag protein (gag) gene, complete cds; pol protein (pol) gene, partial cds; vif protein (vif), vpr prot	14846	14846	98%	0.0	97%	JX392381.1

- identifikator slijeda
 - npr. slijed iz RefSeq (npr. **ref**|NP_006735.1), GenBank (npr. **gb**|AAF69622.1)
- kratki opis slijeda
- maksimalan rezultat poravnanja (*Max bit-score*)
- ukupan rezultat (*Total bit-score*) \geq maksimalan rezultat poravnanja
- *E*-vrijednost

BLAST – evaluacija dobivenih rezultata

- Kako evaluirati rezultate?
 - Kako odrediti značaj rezultata?
 - Što napraviti kada je puno rezultata?
 - Što napraviti kada je malo rezultata?

BLAST – statistika (1)

- **rezultat poravnanja** (eng. *score*) S :

$$S = (\sum M_{ij}) - cP_{postojanje} - dP_{pojedinačno}$$

M_{ij} – rezultat prema matrici sličnosti

$P_{postojanje}$ – kazna za postojanje/otvaranje procijepa (BLAST: 11)

$P_{pojedinačno}$ – kazna za pojedinačnu prazninu (proširenje procijepa) (BLAST: 1)

d – ukupna duljina procijepa

c – broj procijepa

- **E-vrijednost** (eng. *E-value*) - određivanje statističkog značaja poravnanja
 - ima li poravnanje biološko značenje (homologija) ili je samo rezultat slučajnosti?

BLAST – statistika (2)

- statistički temelji: Altschul *et al.* (1990, 1994, 1997)
- neka su zadana sljedovi duljine m (upitni slijed) i n (slijed ili baza podataka sljedova s kojim se upitni slijed uspoređuje) te konstanta K

- neka je rezultat poravnanja ta dva slijeda jednak S

- vrijedi:

$$P(S < x) = \exp(-e^{-\lambda(x-u)})$$

$$u = \ln Kmn / \lambda$$

λ , K – Karlin-Altschulovi statistički parametri

(konstante ovisne o odabranoj supstitucijskoj matrici)

BLAST – statistika (3)

- vjerojatnost da rezultat poravnanja S bude $\geq x$ za neka 2 slučajna slijeda je:

$$P(S \geq x) = 1 - \exp(-Kmn e^{-\lambda x})$$

- K i λ su Karlin-Altschulovi statistički parametri

BLAST – statistika (4)

Pitanje:

Kolika je vjerojatnost da će se pretraživanjem baze podataka duljine n za upitni slijed duljine m

pronaći slijed koji će samo zbog slučajnosti,
a ne stvarne homologije,

imati rezultat poravnanja s upitnim slijedom $\geq S$?

→ Procijeniti broj lažnih pozitivnih rezultata (eng. *false positives*)

BLAST – E i p vrijednosti (1)

Povezanost između E -vrijednosti i vjerojatnosti p

- rezultat poravnanja: S
- bitovni rezultat poravnanja: $S' = (\lambda S - \ln K) / \ln 2$
- očekivanje: $E = mn \cdot 2^{-S'}$
- vjerojatnost p da je dobiveni rezultat S slučajan:

$$p = 1 - e^{-E}$$

E	p
10	0.99995460
5	0.99326205
1	0.63212056
0.1	0.09516258
0.05	0.04877058
0.001	0.00099950

BLAST – E i p vrijednosti (2)

- Koliki je E značajan?
 - $E \approx 0.05$
 - vrlo srodni sljedovi mogu imati $E \leq 10^{-20}$ i manje
 - npr. $E \leq 10^{-100}$ za homologe ili identične gene
 - analize genoma mikroorganizama pokazuju da su značajne vrijednosti oko 10^{-4} (10^{-5} i manje)

BLAST – ideja

List, Scan, Extend

BLAST – algoritam (1)

Predkorak:

Isključiti područja niske složenosti (eng. *low-complexity regions*)

1. Iz upitnog slijeda odrediti listu riječi L , gdje su riječi unaprijed zadane duljine
 - obično 3 aminokiseline ili 11 nukleotida
2. Za svaku riječ W iz L pronaći slične riječi, tj. izgraditi listu riječi L' čije poravnanje s W daje rezultat $\geq T$
 - rezultat poravnanja računa se korištenjem supstitucijskih matrica ili bodovanjem podudaranja/nepodudaranja nukleotida

BLAST – algoritam (2)

3. Pretražiti bazu podataka kako bi se pronašle sljedovi koji sadrže riječi (eng. *hits*) iz L'
4. Proširiti poravnanje oko pogodaka ulijevo i udesno, sve dok rezultat poravnanja ne počne padati ispod zadanog praga
→ područja zvana **HSP** (eng. *high-scoring segment pair*)
 - oduzima 90% vremena!
 - prag (rezultat poravnanja) npr. 22 za proteine i 20 za nukleotide
5. Odabrati *HSP*-ove s najvišim rezultatom te odrediti njihov statistički značaj (E i p vrijednosti)

BLAST – algoritam (3)

Gapped BLAST

(Gapped Extension)

6. Povezati 2 ili više inicijalnih pogodaka koji se nalaze na istoj dijagonali i međusobno su udaljeni $< A$
7. Statistički značajna podudaranja se ponovno poravnavaju korištenjem SW algoritma

Supstitucijske matrice (1)

- supstitucijska matrica S : matrica 20×20 koja sadrži brojeve koji predstavljaju stope mutacija aminokiselina u nekom vremenu
 - temelji se na vjerojatnosti
 - ocrtava svojstva aminokiselina
 - Primjer:
veća je vjerojatnost da će hidrofilna aminokiselina mutirati u drugu hidrofilnu aminokiselinu, a manja da će mutirati u hidrofobnu

Supstitucijske matrice (2)

- PAM matrice (Dayhoff et al. 1978)
 - PAM/APM (*Accepted Point Mutation*)
 - temeljene na promatranju globalnog poravnanja blisko srodnih vrsta (što uključuje i očuvana područja i područja veće mutacije)
 - filogenetski pristup: promatranje zajedničkog pretka poravnatih aminokiselina
- PAM 1 - promijenjeno je 1% aminokiselina
 - ostale PAM matrice se računaju iz PAM1 (sve do PAM250)
- $M_{ij} = 10 \cdot \log (q_{ij} / p_i)$
 - q_{ij} - promatrana frekvencija supstitucije, tj. promjene $j \rightarrow i$
 - p_i – očekivana frekvencija pojave i -te aminokiseline

Supstitucijske matrice (3)

- BLOSUM matrice (Henikoff & Henikoff, 1992)
 - BLOSUM (*BLOCK Substitution Matrix*)
 - poravnanje vrlo očuvanih područja u poravnanjima; nema eksplicitnog evolucijskog modela
 - bolje za usporedbe udaljenijih proteina; obično se koristi BLOSUM62

BLOSUM62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Hung et al. BMC Genomics 2010 11(Suppl 3):S14

BLAST – izgradnja liste sličnih riječi

Koraci 1 & 2:

Konstruirati listu riječi duljine n čiji zbroj daje vrijednost $\geq T$.

blastp: $n = 3$ (2 - 3) \rightarrow za 20 aminokiselina: $20^3 = 8000$ mogućih riječi

blastn: $n = 11$ (7 - 11)

Primjer:

upit = GEIIGCT

- rastavlja se na riječi **GEI** EII IIG IGC GCT
- za svaku od riječi r generirati listu riječi čiji je rezultat poravnanja s $r \geq T = 12$ (prema BLOSUM62)

Riječ	Zbroj (riječ GEI)
GEI	$6 + 5 + 4 = 15$
GEE	$6 + 5 - 3 = 8$
EEE	$-2 + 5 + (-3) = 0$
GEL	$6 + 5 + 2 = 13$

BLAST – pretraživanje baze podataka (1)

Korak 3:

- konstruira se automat s konačnim brojem stanja (eng. *finite state automaton*; FSA) na temelju liste riječi L' (čiji je zbroj $\geq T$)
→ automat omogućuje prepoznavanje riječi iz L'
- Ulaz: sljedovi iz baze podataka
 - u slijedu se promatra znak po znak (svaki ulazni znak inicira prijelaz u neko od stanja automata)
- Izlaz: dojava o pronađenim riječima ili odbacivanje ulaznog slijeda

BLAST – pretraživanje baze podataka (2)

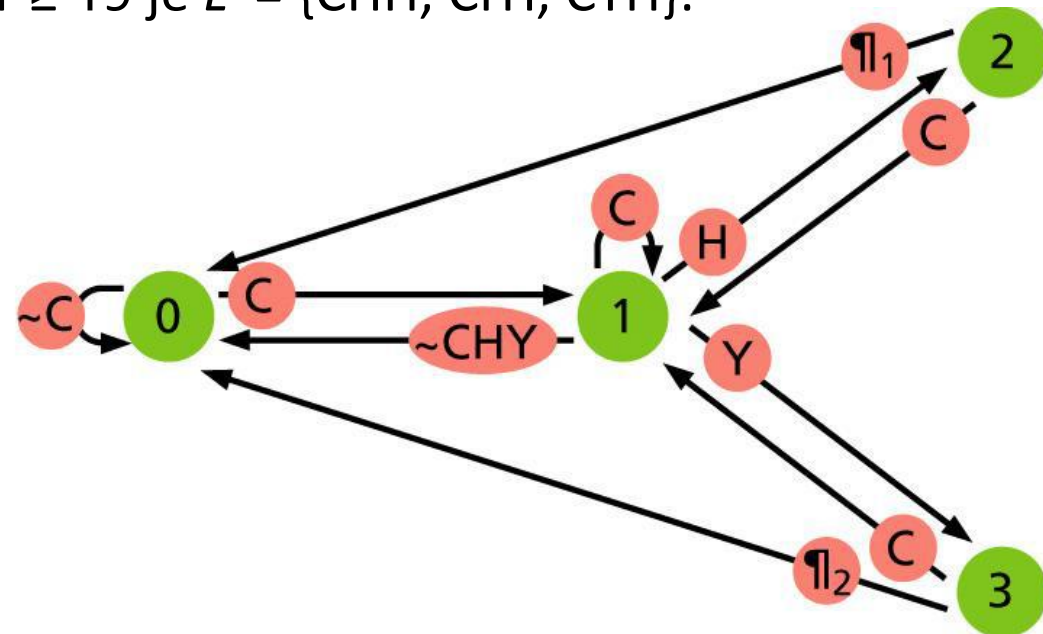
Primjer:

Zadana je riječ CHH i $T = 19$.

Lista riječi (prema BLOSUM62) za koje je $T \geq 19$ je $L' = \{CHH, CHY, CYH\}$.

Automat izgrađen na temelju L' :

	INPUT	OUTPUT
Π_1 :	\sim CHY	none
	H	CHH
	Y	CHY
Π_2 :	\sim CH	none
	H	CYH



Zvelebil & Marketa, 2008
(Understanding Bioinformatics, Ch.5, Fig. 5.23)

BLAST – *High-Scoring Segment Pairs*

Korak 4:

- pronalaženje pogodaka (eng. *hit*) u bazi podataka i njihovo proširenje
→ generiranje HSP (*High-scoring Segment Pair*)

- Primjer:

Upit = E E T **P Q I** A V E

Slijed iz BP = L I T **P Q E** L V C

E	E	T	P	Q	I	A	V	E
L	I	T	P	Q	E	L	V	C

- pogodak se proširuje sve dok trenutno najbolji rezultat ne padne za $\geq X = 3$
- koristi se supstitucijska matrica BLOSUM62

PQI || **PQE** $\rightarrow 7 + 5 - 3 = 9$ (pronađen je "pogodak" za PQI)

TPQIA || TPQEL = $5 + 7 + 5 - 3 - 1 = 13$

ET**PQ**IAV || IT**PQ**ELV = $-3 + 5 + 7 + 5 - 3 - 1 + 4 = 14$

EET**PQ**IAVE || LIT**PQ**ELVC = $-3 - 3 + 5 + 7 + 5 - 3 - 1 - 4 = 7$

BLAST – odabir supstitucijskih matrica

- za blisko srodne proteine:
PAM1, BLOSUM80
- za *srednje* srodne proteine:
PAM120, BLOSUM62
- za udaljene proteine:
PAM250, BLOSUM45

Napomena:

Kod vrlo udaljenih homologa, kojima je inicijalni rezultat poravnanja malen, ali imaju vrlo sličnu 3D strukturu, koristiti PSI-BLAST.

PSI-BLAST (1)

Position-Specific Iterated BLAST (Altschul *et al.*, 1997; Schäffer *et al.* 2001)

1. Osnovno blastp pretraživanje
2. PSI-BLAST gradi MSA prema inicijalnom rezultatu blastp korištenjem statistike temeljene na aminokiselinskom sastavu sljedova (eng. *composition-based statistics*) → izgraditi profil **PSSM** (*Position-Specific Score Matrix*)
3. PSSM (L x 20) postaje upitni slijed i onda se koristi za daljnje pretraživanje baze podataka (umjesto supstitucijske matrice)
4. PSI-BLAST procjenjuje statistički značajne rezultate (*E*-vrijednost)
→ novi profil se koristi kao novi upit

Koraci 3-4 se ponavljaju više puta (obično 5 puta); novi se profil koristi kao upit.

L = duljina upita, 20 – broj aminokiselina

MSA (*Multiple Sequence Alignment*) = poravnanje više sljedova odjednom

<u>730496</u>	66	FTVDENGQMSATAKGRVRLFNNUWVDCADMIGSFTDTE	125
<u>200679</u>	63	FSVDEKGHMSATAKGRVRLLSNUEVCADMVGTFTDTE	122
<u>206589</u>	34	FSVDEKGHMSATAKGRVRLLSNUEVCADMVGTFTDTE	93
<u>2136812</u>	2	MSATAKGRVRLLSNNUWVDCADMVGTFTDTE	53
<u>132408</u>	65	FKIEDNGKTTATAKGRVRILDKLELCANMVGTFIETN	124
<u>267584</u>	44	FSVDESGKVTATAHGRVILNNWHECANMFGTFTEDTP	103
<u>267585</u>	44	FSVDGSGKVTATAQGRVILNNWHECANMFGTFTEDTP	103
<u>8777608</u>	63	FTIHEDGANTATAKGRVILNNWHECADHMFETTPDPA	122
<u>6687453</u>	60	FKVEEDGTHTATAIGRVILNNWHECANMFGTFTEDTE	119
<u>10697027</u>	81	FKVQEDGTHTATATGRVILNNWHECANMFGTFTEDTE	140
<u>13645517</u>	1	MVGTFTDTE	32
<u>13925316</u>	38	FSVDGSGKMTATAQGRVILNNWHECANMFGTFTEDTP	97
<u>131649</u>	65	YTVEEDGTHTASSKGRVKLFGFVVICADMAAQYTDPT	126

MSA (Multiple Sequence Alignment)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1 M	-1	-2	-2	-3	-2	-1	-2	-3	-2	1	2	-2	6	0	-3	-2	-1	-2	-1	1
2 K	-1	1	0	1	-4	2	4	-2	0	-3	-3	3	-2	-4	-1	0	-1	-3	-2	-3
3 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
4 V	0	-3	-3	-4	-1	-3	-3	-4	-4	3	1	-3	1	-1	-3	-2	0	-3	-1	4
5 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
6 A	5	-2	-2	-2	-1	-1	0	-2	-1	-1	-1	-1	-1	-3	-1	1	0	-3	-2	0
7 L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
8 L	-1	-3	-3	-4	-1	-3	-3	-4	-3	2	2	-3	1	3	-3	-2	-1	-2	0	3
9 L	-1	-3	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	2
10 L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
11 A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
12 A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
13 W	-2	-3	-4	-4	-2	-2	-3	-4	-3	1	4	-3	2	1	-3	-3	-2	7	0	0
14 A	3	-2	-1	-2	-1	-1	-2	4	-2	-2	-2	-1	-2	-3	-1	1	-1	-3	-3	-1
15 A	2	-1	0	-1	-2	2	0	2	-1	-3	-3	0	-2	-3	-1	3	0	-3	-2	-2
16 A	4	-2	-1	-2	-1	-1	-1	3	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	-1
...																				
37 S	2	-1	0	-1	-1	0	0	0	-1	-2	-3	0	-2	-3	-1	4	1	-3	-2	-2
38 G	0	-3	-1	-2	-3	-2	-2	6	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
39 T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-3	-2	0
40 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
41 Y	-2	-2	-2	-3	-3	-2	-2	-3	2	-2	-1	-2	-1	3	-3	-2	-2	2	7	-1
42 A	4	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
...																				

PSSM (Position-Specific Score Matrix)

PSI-BLAST (2)

PSSM (*Position-Specific Score Matrix*)

Algoritam:

Odredi PSSM na temelju upitnog slijeda

Ponavljaj dok ima novih homologa ili je broj pretraživanja $< n$

PSSM' := PSSM uz informacije dobivene
pronalaskom homologa

Pretraži bazu korištenjem PSSM' matrice

Kraj

Ispiši homologe

PSI-BLAST (3)

1. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs (Altschul *et al.* 1997.)
2. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements (Schäffer *et al.* 2001)

→ Najveće poboljšanje (2) u odnosu na (1):
izračun λ ovisno o sastavu upitnog slijeda i baze podataka, a ne temeljem prethodno izračunatih vrijednosti za λ

MegaBLAST

- uspoređivanje jako dugih upitnih sljedova
- usporedbe vrlo sličnih sljedova (unutar vrste ili između blisko srodnih vrsta)

Primjer:

usporedba cijelog ljudskog kromosoma s drugim dugačkim kromosomima (npr. kromosomom miša)

Popis literature

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J., 1990. Basic local alignment search tool, J. Mol. Biol. 215:403-410.
2. Altschul S.F. *et al.* 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389-3402.
3. Lipman & Pearson, 1988. Improved Tools for Biological Sequence Comparison. Proc Natl Acad Sci USA. 1988 Apr;85(8):2444-8.
4. Kerfeld CA, Scott KM (2011) Using BLAST to Teach “E-value-tionary” Concepts. PLoS Biol 9(2): e1001014. doi: 10.1371/journal.pbio.1001014
5. Pevsner J., 2009. Bioinformatics and Functional Genomics, 2nd Ed., Ch. 4 & 5
6. Zvelebil & Baum. 2008. Understanding Bioinformatics, Ch.5
7. <http://petang.cgu.edu.tw/Bioinfomatics/MANUALS/NCBIblast/psi1.html>