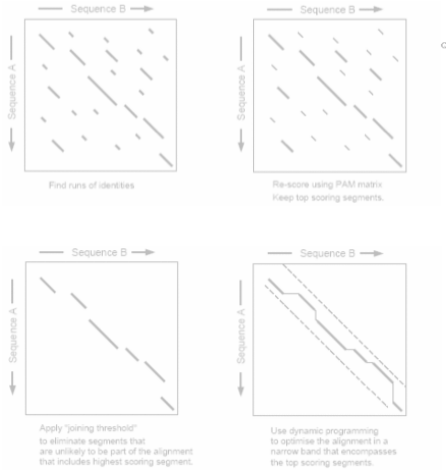
Praktične vaje pri predmetu
Biokemijska informatika

8. vaja

Iskanje podobnih zaporedij



Miha Pavšič
april 2014

Na vaji bomo spoznali dva programa za iskanje podobnih zaporedij:

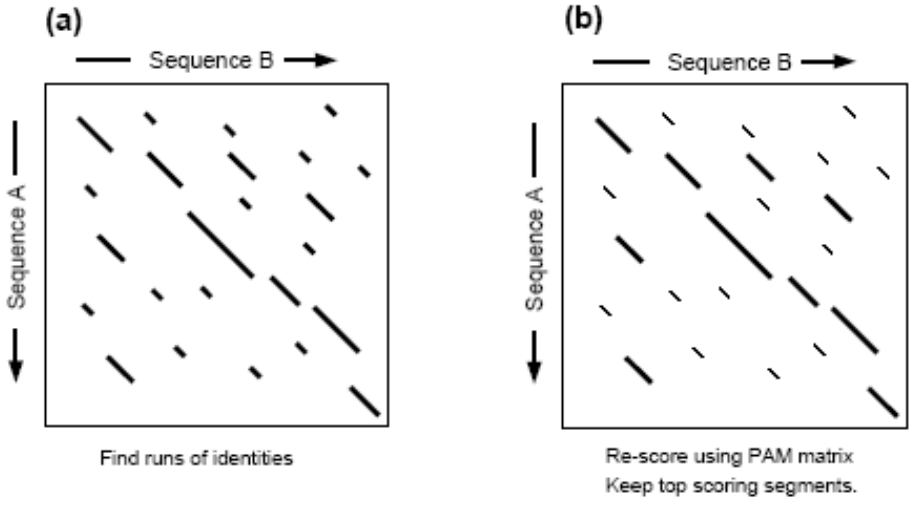
- **FASTA** – FAST-All (FAST-P & FAST-N)
  FASTA je prvi hiter algoritem za iskanje podobnih zaporedij oz. primerjavo danega zaporedja z zaporedji v bazi.
  Po tem programu se imenuje format FASTA, saj ga ta program uporablja za zaporedja v bazi.
- **BLAST** – Basic Local Alignment Search Technique (Tool)
  Izboljšan algoritem v smislu hitrosti, enostavnosti in statistične obravnave rezultatov iskanja.

Osnovna ideja (velja za FASTA in BLAST): dobro prileganje vsebuje odseke z veliko stopnjo podobnosti/identičnosti.

1. Identifikacija kratkih odsekov identičnosti.
2. Razširitev/podaljšanje odsekov iz prejšnje stopnje, tako da dobimo daljše odseke z določeno mero podobnosti.
3. Optimizacija najboljših zadetkov.

FASTA - algoritem (1/2)

- temelji na točkovnem diagramu (*dot plot*) – izračun najboljših diagonal na osnovi vseh okvirov prileganja
- najprej iskanje točnih ponovitev med našim zaporedjem (A) in zaporedji v bazi (B)
  - za nt zaporedja: tipično besede dolžine 6 nt (ktup = 6)
  - za ak zaporedja: tipično besede dolžine 2 ak (ktup = 2)
- te kratke ponovitve se nato točkujejo z uporabo matrike zamenjav

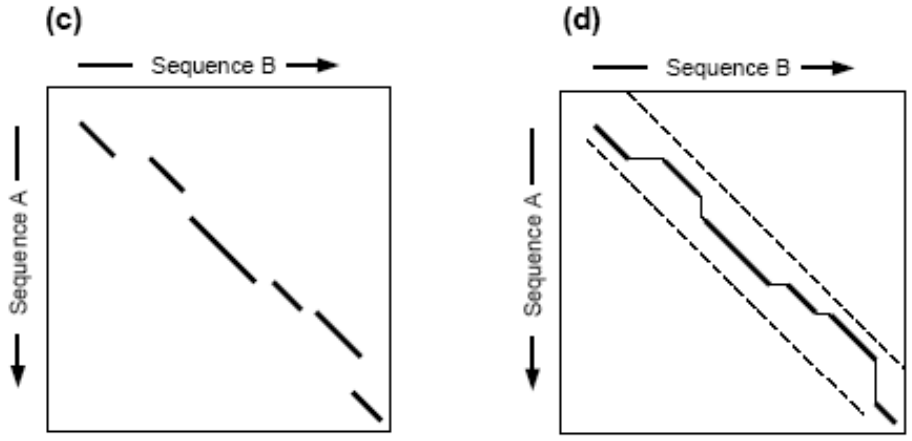


Find runs of identities

Re-score using PAM matrix
Keep top scoring segments.

FASTA - algoritem (2/2)

- z dinamični programiranjem se nato optimizira prileganje, ki zajema segmente z najvišjo vrednostjo



Apply "joining threshold"
to eliminate segments that
are unlikely to be part of the alignment
that includes highest scoring segment.

Use dynamic programming
to optimise the alignment in a
narrow band that encompasses
the top scoring segments.

FASTA

Več izvedb:

- **FASTA** – osnoven algoritem
- **FASTX** in **FASTY** – primerjava nt zaporedja z bazo ak zaporedij (nt zaporedje se prevede v ak v vseh 6 bralnih okvirih); pri FATSY so lahko premiki bralnega okvira znotraj kodonov (počasneje, a boljša prileganja)
- **TFASTX** in **TFASTY** – primerjava ak zaporedja z bazo nt zaporedij (vsako nt zaporedje v bazi se prevede v vseh 6 bralnih okvirih); pri TFATSY so lahko premiki bralnega okvira znotraj kodonov (počasneje, a boljša prileganja)

Na strani EBI so na voljo tudi:

- **SSEARCH** – iskanje lokalnega prileganja z algoritmom Smith-Waterman
- **GGSEARCH** – iskanje globalnega prileganja z algoritmom Needleman-Wunsch

BLAST - algoritem (1/2)

1. Generate words from sequence above threshold (e.g. T=11)

Query Sequence:

```
>gi|16329320 (residues 412 to 594)
SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGVVTTQTTG
RHQGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNKAFVLNLLQEGRS
VRRHFDHQPLPKDGDNPFSRLEHYSTQNGCLILAEALAYLECLVQSWSNI
GDHVLVYATVQAGQVLQPNGITAIRHRKSGGQY
```

Fragmentation into words:

SWVSQASFTPPGIM → SWV WVS VSQ SQA QAS ASF SFT ...

Selection of words scoring above threshold (for word SWV):

Substitution Matrix*

	R	G	I	K	F	S	T	W	V
R	5	0	-1	-1	-2	1	0	-3	0
G		6	-4	-2	-3	0	-2	-2	-3
I			4	-3	0	-2	-1	-3	3
K				5	-3	0	-1	-3	-2
F					6	-2	-2	1	-1
S						4	1	-3	-2
T							5	-2	0
W								11	-3
V									4

*A portion of the BLOSUM 62 matrix

SWV (4+11+4 = 19)
SWI (4+11+3 = 18)
TWV (1+11+4 = 16)
GWV (0+11+4 = 15)
KWV (0+11+4 = 15)
SWS (4+11-2 = 13)

Synonyms above threshold 11... (others not shown)

SFV (4+1+4 = 9)
SRV (4-3+4 = 5)

Synonyms below threshold 11... (others not shown)

Vir: 10.1371/journal.pbio.1001014.g001

BLAST - algoritem (2/2)

2. Search the database for words matching those generated

3. Extend matching hits in both directions

```
RHQGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNK
    .| ||||||||| |||: :|:  |   | ||
..TAMLVSWVSQASFNPPGLTIALAKE.RAEGLDHSGD
```

Word match from Step 1 — Extension until score drops

4. Generate alignment and calculate statistics

```
>ref|YP_002482587.1| flavin reductase domain protein FMN-binding [Cyanothece sp.
PCC 7425]
 gb|ACL44226.1| flavin reductase domain protein FMN-binding [Cyanothece sp. PCC
7425]
Length=585

 Score =  176 bits (446),  Expect = 1e-42, Method: Compositional matrix adjust.
 Identities = 95/196 (48%), Positives = 125/196 (63%), Gaps = 16/196 (8%)

Query  1    SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGVVTTQTTGRH--------  52
            +G++FA+ L+  K+QR  RQ+  E Q+DRT+QAVGRIIGS+ V+T +    H
Sbjct  393  AGSDFAQVLKKAKKQRSPRQSILEVQSDRTEQAVGRIIGSLCVLTAKQQQTHPHPEVEEP  452

Query  53   -----QGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNKAFVLNLLQEGRSVRRHFDH  107
                 +L SWVSQASF PPG+ +A+  E  A GL     AFVLN+L+EG ++RRHF
Sbjct  453  QLEVPTAMLVSWVSQASFNPPGLTIALAKE-RAEGLDHSGDAFVLNVLKEGMNLRRHFSK  511

Query  108  QPLPKDGDNPFSRLEHYSTQNGCLILAEALAYLECLVQSWSNIGDHVLVYATVQAGQVLQ  167
               P  G++ F+ L      +NGC +L + LAYLEC VQS    GDH L+YATV  G+VLQ
Sbjct  512  SFAP--GEDRFAGLNIQWAENGCPVLQDCLAYLECTVQSRMECGDHWLIYATVNNGKVLQ  569

Query  168  PNGITAIRHRKSGGQY  183
            P G TA++HRKSG QY
Sbjct  570  PTGTTAVQHRKSGNQY  585
```

Vir: 10.1371/journal.pbio.1001014.g001

BLAST – variante (http://blast.ncbi.nlm.nih.gov)

program	iskalno zaporedje (*query sequence*)	baza (*database*)
BLASTN	nt	nt
BLASTP	ak	ak
BLASTX	nt (→ak)	ak
TBLASTN	ak	nt (→ak)
TBLASTX	nt (→ak)	nt (→ak)

iskalno zaporedje se prevede v vseh 6 bralnih okvirih

vsa zaporedja v bazi so prevedena v veh 6 bralnih okvirih

iskalno zaporedje in vsa zaporedja v bazi so prevedena v vseh 6 bralnih okvirih

Specialized BLAST

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

- Make specific primers with Primer-BLAST
- 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 and T cell receptor sequences (IgBLAST)
- Screen sequence for vector contamination (vecscreen)
- Align two (or more) sequences using BLAST (bl2seq)
- Search protein or nucleotide targets in PubChem BioAssay
- Search SRA by experiment
- Constraint Based Protein Multiple Alignment Tool
- Needleman-Wunsch Global Sequence Alignment Tool
- Search RefSeqGene

BLAST – primeri uporabe

blastp
- želimo izvedeti funkcijo proteina, za katerega poznamo ak zaporedje
- iskanje podobnih regij v proteinih → funkcijska identifikacija, filogenetske analize

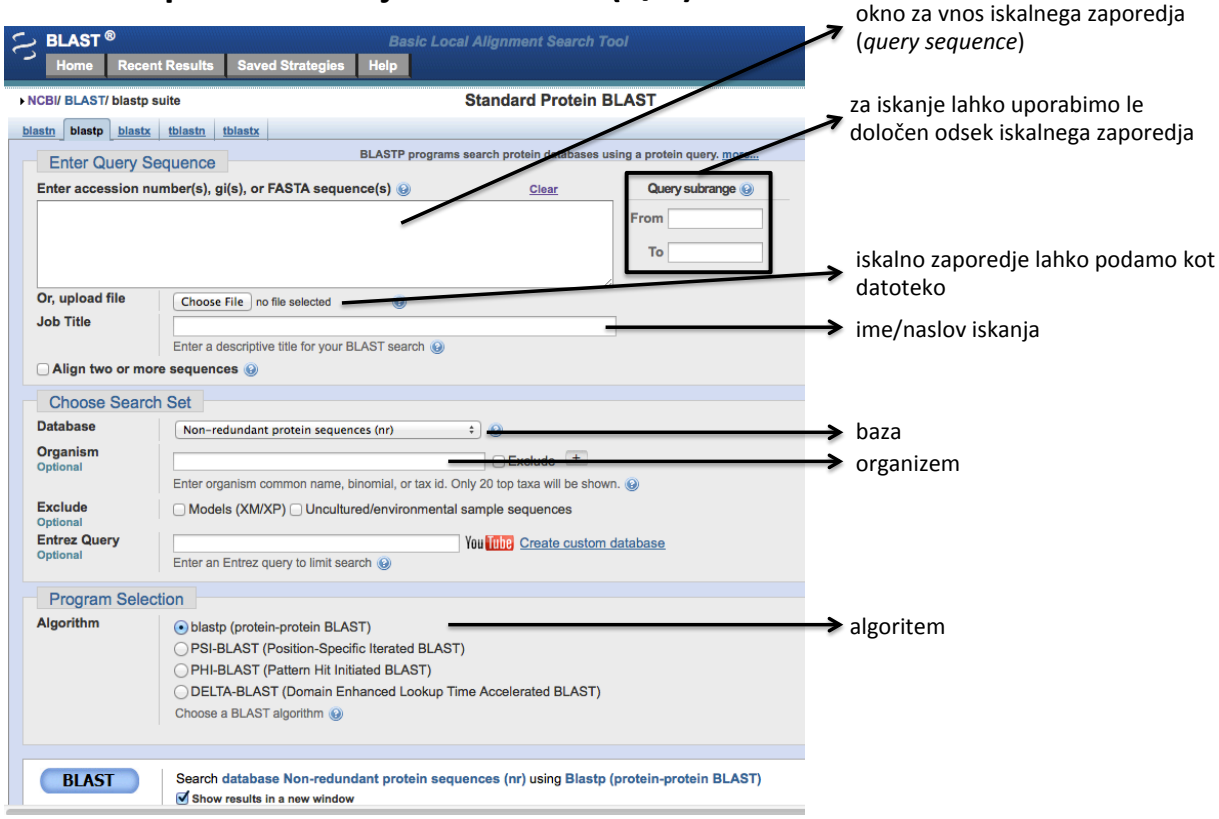
tblastn
- iskanje novig protein-kodirajočih genov
- mapiranje gena za protein na genomsko DNA

blastn
- za zelo podobna nt zaporedja
- mapiranje oligonukleotidov in PCR produktov v genomu
- iskanje ponavljajočih zaporedij
- medvrstna primerjava zaporedij

tblastx
- odkrivanje genov/proteinov in EST

blastx
- analiza iskalnega nt zaporedja
- iskanje genov, ki kodirajo za protein, v genomskih bazah
- iskanje, če cDNA nosi zapis za znan protein

BLAST – primer iskanja za BLASTP (1/5): začetek



okno za vnos iskalnega zaporedja (*query sequence*)

za iskanje lahko uporabimo le določen odsek iskalnega zaporedja

iskalno zaporedje lahko podamo kot datoteko

ime/naslov iskanja

baza

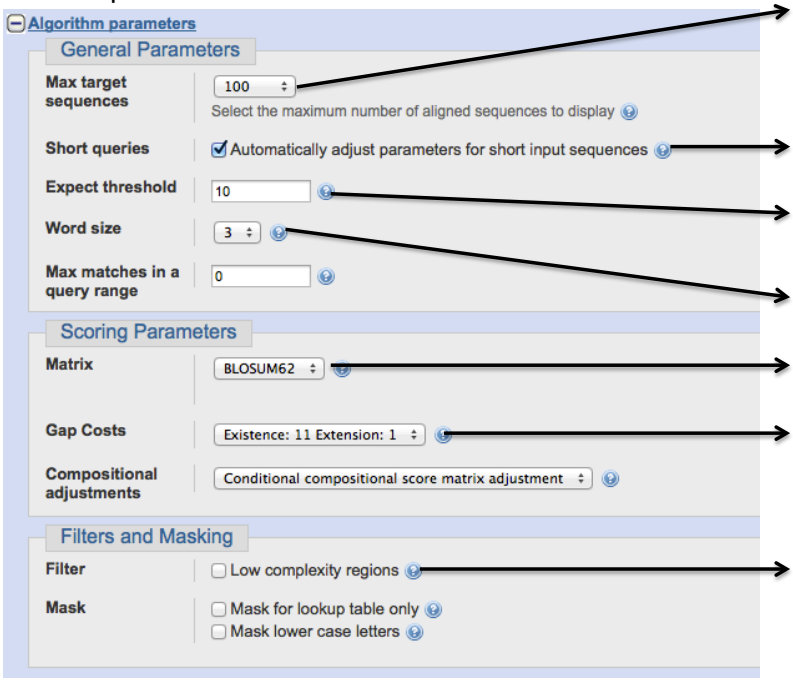
organizem

algoritem

Podrobneje: http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml

BLAST – primer iskanja za BLASTP (2/5): začetek

Dodatni parametri:



največje število zaporedij, prikazanih kot rezultat iskanja

avtomatska prilagotitev parametrov za kratka zaporedja (npr. pri iskanju N-končnega zaporedja proteinov)

največja vrednost E zadetkov, ki ji še prikaže med rezultati

velikost besede za iniciacijo prileganja (vpliva na občutljivost in hitrost)

matrika zamenjav za točkovanje

"kazni" za vrzeli

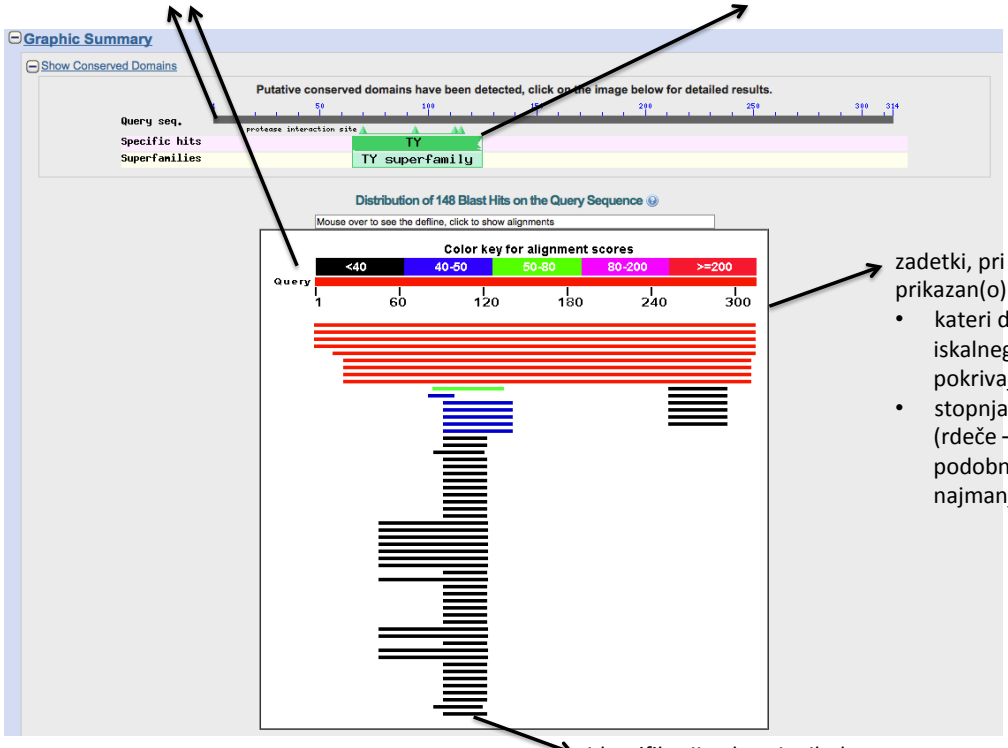
prilagoditve za npr. regije z nizko kompleksnostjo (primer je zaporedje GGGSSSSGGGG)

Podrobneje: http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml

BLAST – primer iskanja za BLASTP (3/5): rezultati

iskalno zaporedje prikazano kot "trak" (**QUERY**)

identificirane ohranjene domene



zadetki, pri katerih je prikazan(o):

- kateri del našega iskalnega zaporedja pokrivajo
- stopnja podobnosti (rdeče – najbolj podobno, črno – najmanj podobno)

identifikacija ohranjenih domen

BLAST – primer iskanja za BLASTP (4/5): rezultati

Tabelaričen prikaz zadetkov

MAX SCORE – pove nam, kako podobni sta si zaporedji

TOTAL SCORE – vsebuje točke/vrednosti s posameznih nepovezanih odsekov zaporedij

kolikšen del iskalnega zaporedja pokriva prileganje s tem zaporedjem

E-vrednost

% identičnosti (prilegan del)

koda v bazi za zadetek

opis posameznega zadetka

Sequences producing significant alignments:

Select: All None Selected:0

Alignments | Download | GenPept | Graphics | Distance tree of results | Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
epithelial cell adhesion molecule precursor [Homo sapiens] >sp|P16422.2|EPCAM_HUMAN RecName: Full=Epithelial cell adhesion	652	652	100%	0.0	100%	NP_002345.2
TACSTD1 [Homo sapiens]	652	652	100%	0.0	99%	CAG47055.1
carcinoma-associated antigen GA733-2 [Homo sapiens] >gb|AAB00775.1| carcinoma-associated antigen GA733-2 [Homo sapiens] >	650	650	100%	0.0	99%	AAA35861.1
unnamed protein product [Homo sapiens] >gb|AAA36151.1| adenocarcinoma-associated antigen precursor (KSA) [Homo sapiens] >g	649	649	100%	0.0	99%	CAA32870.1
tumor-associated calcium signal transducer 1, isoform CRA_b [Homo sapiens]	561	561	95%	0.0	97%	EAX00219.1
tumor-associated calcium signal transducer 2 precursor [Homo sapiens] >sp|P09758.3|TACD2_HUMAN RecName: Full=Tumor-assc	265	265	92%	2e-85	51%	NP_002344.2
TACSTD2 [Homo sapiens]	264	264	92%	6e-85	51%	CAG47056.1
gp50/Trop-2 [Homo sapiens] >gb|EAX06630.1| tumor-associated calcium signal transducer 2 [Homo sapiens] >gb|ACR78188.1| tum	263	263	92%	1e-84	50%	CAA54799.1
GA733-1 protein precursor [Homo sapiens]	262	262	92%	4e-84	50%	AAA52505.1
tumor-associated calcium signal transducer 2 [Homo sapiens]	78.2	78.2	16%	2e-17	67%	ACZ01960.1
C215 antigen, GA733-2 protein=non-mucin type integral membrane glycoprotein {N-terminal} [human, colon adenocarcinoma COLO	43.9	43.9	6%	5e-05	95%	AAB28754.1
testican [Homo sapiens]	40.8	40.8	15%	0.004	38%	AAC24945.1

MAX SCORE = TOTAL SCORE

(ko sta zaporedji podobni v eni regiji – v skladu s parametri (lahko pokriva celotno zaporedje))

BLAST – primer iskanja za BLASTP (5/5): rezultati

Prileganja

1. prileganje (1. zadetek)

identičnost, identičnost+podobnost, vrzeli

prileganje

2. prileganje (2. zadetek)

...itd.

Download GenPept Graphics

epithelial cell adhesion molecule precursor [Homo sapiens]

Sequence ID: ref|NP_002345.2| Length: 314 Number of Matches: 1

See 5 more title(s)

Range 1: 1 to 314 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
652 bits(1682)	0.0	Compositional matrix adjust.	314/314(100%)	314/314(100%)	0/314(0%)

```
Query  1    MAPPQVLAFGLLLAAATATFAAAQEECVCENYKLAVNCFVNNNRQCQCTSVGAQNTVICS  60
            MAPPQVLAFGLLLAAATATFAAAQEECVCENYKLAVNCFVNNNRQCQCTSVGAQNTVICS
Sbjct  1    MAPPQVLAFGLLLAAATATFAAAQEECVCENYKLAVNCFVNNNRQCQCTSVGAQNTVICS  60

Query  61   KLAAKCLVMKAEMNGSKLGRRAKPEGALQNNDGLYDPDCDESGLFKAKQCNGTSMCWCVN  120
            KLAAKCLVMKAEMNGSKLGRRAKPEGALQNNDGLYDPDCDESGLFKAKQCNGTSMCWCVN
Sbjct  61   KLAAKCLVMKAEMNGSKLGRRAKPEGALQNNDGLYDPDCDESGLFKAKQCNGTSMCWCVN  120

Query  121  TAGVRRTDKDTEITCSERVRTYWIIIELKHKAREKPYDSKSLRTALQKEITTRYQLDPKF  180
            TAGVRRTDKDTEITCSERVRTYWIIIELKHKAREKPYDSKSLRTALQKEITTRYQLDPKF
Sbjct  121  TAGVRRTDKDTEITCSERVRTYWIIIELKHKAREKPYDSKSLRTALQKEITTRYQLDPKF  180

Query  181  ITSILYENNVITIDLVQNSSQKTQNDVDIADVAYYFEKDVKGESLFHSKKMDLTVNGEQL  240
            ITSILYENNVITIDLVQNSSQKTQNDVDIADVAYYFEKDVKGESLFHSKKMDLTVNGEQL
Sbjct  181  ITSILYENNVITIDLVQNSSQKTQNDVDIADVAYYFEKDVKGESLFHSKKMDLTVNGEQL  240

Query  241  DLDPGQTLIYYVDEKAPEFSMQGLKAGVIAVIVVVVIAVVAGIVVLVISRKKRMAKYEKA  300
            DLDPGQTLIYYVDEKAPEFSMQGLKAGVIAVIVVVVIAVVAGIVVLVISRKKRMAKYEKA
Sbjct  241  DLDPGQTLIYYVDEKAPEFSMQGLKAGVIAVIVVVVIAVVAGIVVLVISRKKRMAKYEKA  300

Query  301  EIKEMGEMHRELNA  314
            EIKEMGEMHRELNA
Sbjct  301  EIKEMGEMHRELNA  314
```

Download GenPept Graphics

TACSTD1 [Homo sapiens]

Sequence ID: emb|CAG47055.1| Length: 314 Number of Matches: 1

Range 1: 1 to 314 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
652 bits(1681)	0.0	Compositional matrix adjust.	313/314(99%)	314/314(100%)	0/314(0%)

```
Query  1    MAPPQVLAFGLLLAAATATFAAAQEECVCENYKLAVNCFVNNNRQCQCTSVGAQNTVICS  60
            MAPPQVLAFGLLLAAATATFAAAQEECVCENYKLAVNCFVNNNRQCQCTSVGAQNTVICS
Sbjct  1    MAPPQVLAFGLLLAAATATFAAAQEECVCENYKLAVNCFVNNNRQCQCTSVGAQNTVICS  60

Query  61   KLAAKCLVMKAEMNGSKLGRRAKPEGALQNNDGLYDPDCDESGLFKAKQCNGTSMCWCVN  120
```

BLAST – vrednost E (Expect)

Vrednost E nam pove, koliko NAKLJUČNIH zadetkov lahko pričakujemo pri iskanju po bazi določene velikosti.

vrednost prileganja: $S = (\sum M_{ij}) - cO - dG$

M_{ij} – vrednost za zamenjavo med ostankoma i in j na osnovi substitucijske matrike
c – število vrzelo
O – vredost za odprtje vrzeli
d – dolžina vseh vrzeli
G – vrednost za vrzel

korigirana vrednost (bit score): $S' = \frac{\lambda S - \ln K}{\ln 2}$

λ in **K** – parametra, odvisna od uporabljene substitucijske matrike

to nam poda BLAST

vrednost E (Expect value): $E = \frac{n \times m}{2^{S'}}$

m – dolžina iskalnega zaporedja (query sequence)
n – celotna dolžina baze (št. vseh nt oz. ak ostankov v bazi)

8. vaja: Iskanje podobnih zaporedij

BLAST – vrednost E (Expect)

Primer: če je vrednost E za nek zadetek v bazi enaka 1 to pomeni, da lahko pričakujemo, da v bazi z enako velikostjo najdemo po NAKLJUČJU točno 1 zadetek z enako vrednostjo S'.

↓ E večja signifikatnost (pomen) zadetka

Identična zaporedja: E = 0

Kratka prileganja z visoko identičnostjo imajo relativno visoke vrednosti E, saj izračun vrednosti E upošteva dolžino iskalnega zaporedja, za kratka zaporedja pa je verjetnost, da najdemo zadetek naključno, večja.

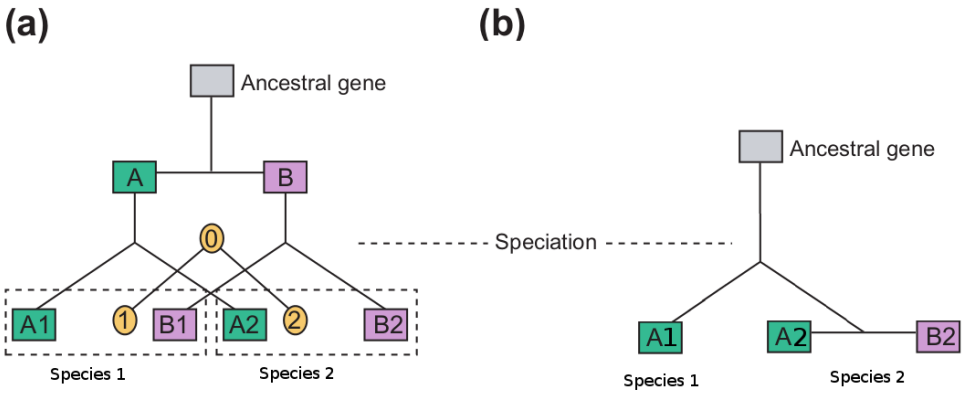
Na splošno rečemo, da sta zaporedji homologni, če je $E < 10^{-5}$.
Meja za homologijo je odvisna od primera do primera:

- nižja E za ločevanje med ortologi in paralogi
- višja E za evolucijsko bolj oddaljena zaporedja

Homologi, ortologi, paralogi

HOMOLOGNA gena– gena, ki imata skupnega prednika, nastala pa sta bodisi z:

- ločitvijo vrst (speciacijo) → **ORTOLOGI** (ponavadi se funkcija ohrani)
- duplikacijo znotraj genoma → **PARALOGI** (ponavadi se funkcija ne ohrani (popolnoma))

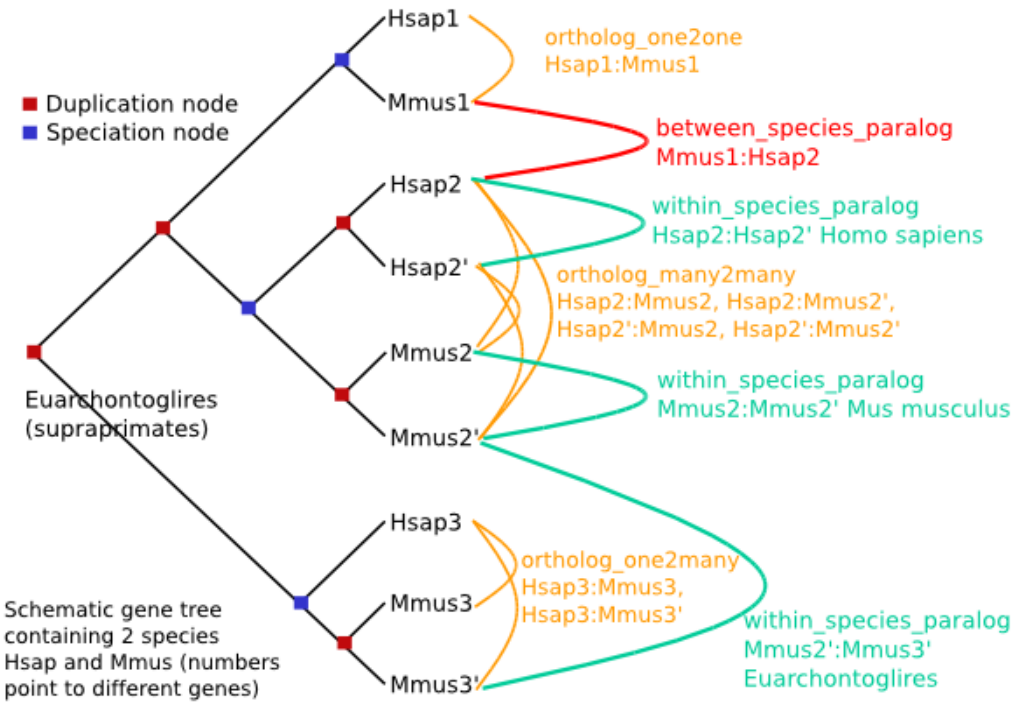


- A in B - paraloga
- A1 in B1 - paraloga
- A1 in B2 - paraloga
- A2 in B1 - paraloga
- A2 in B2 - paraloga
- A1 in A2 - ortologa
- B1 in B2 - ortologa

- A2 in B2 sta ortologa A1
- A2 in B2 - paraloga

8. vaja: Iskanje podobnih zaporedij

Homologi, ortologi, paralogi



Več: http://www.ensembl.org/info/genome/compara/homology_method.html