PRILEGANJE ZAPOREDIJ

Primerjave zaporedij

- čemu poravnave
- rezultat poravnave, ocenjevalne matrike, teža vrzeli
- načini poravnav (globalno, lokalno)

Točkovni grafi (*Dotplot*)

- *Praktična naloga 2*

Algoritmi za poravnave zaporedij

- Needleman-Wunsch
- Smith-Waterman

Praktično iskanje po podatkovnih zbirkah

- BLASTA
- FAST
- Interpretacija rezultatov

Primer:

------**CC**TTCAGAATACAGAA**T**AGGGACATAGAGA
ATCCCA**CC**CAGCCCCCTGGACC**T**GTAT---------

Poravnava “na oko”, ki vnese vrzeli, da zviša število ujemanj. Ampak takšni zaporedji se da poravnati tudi na drug način:

CCTT**CA**GAATA**CAG**AATAGG**G**ACATA**G**AGA
ATCC**CA**---CC**CAG**CCCCCT**G**GACCT**G**TAT

Poveča število baz, ki se ujemajo za 133% in zniža število vrzeli za 80%.

Torej:
Potrebujemo takšno poravnavo, ki maksimizira število identičnih elementov in hkrati minimizira število vrzeli.

Potrebujemo neka dobro določena pravila (algoritme) in racunalnik, ki po teh pravilih poišče najboljšo poravnavo upoštevajoč vse možnosti.
Algoritmi
NEEDLEMAN-WUNSCH GLOBALNA poravnava
SMITH-WATERMAN LOKALNA poravnava

NEEDLEMAN-WUNSCH-ev ALGORITEM

Needleman SB and Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology* 48, 443-453.

GLOBALNO poravna dve zaporedji (primerja celo zaporedje z drugim celim zaporedjem).

Maksimalno ujemanje je definirano kot najvecje število elementov enega zaporedja, ki jih lahko poravnamo z elementi drugega zaporedja.
Dopustimo delecije, uvedena je kazen za vrzeli.

Identicnim elementom pripiše vrednost 1, razlicnim 0. Neujemanja in vrzeli so enako obteženi (0).

Primer: kako poravnati naslednji zaporedji

KLSCVLREDYWEDKAGVD

KSCILRNDDYEKAVGVD

```
KLSCVLRE-DYWEDKA--GVD
: :: ::  :: : ::  :::
K-SCILRNDDY-E-KA-VGVD
```

```
KLSCVLRED-YWEDKAGV--D
: :: :: : : : :: :  :
K-SCILRNDDY-E-KA-VGVD
```

	K	L	S	C	V	L	R	E	D	Y	W	E	D	K	A	G	V	D
K	1													1				
S			1															
C				1														
I																		
L		1				1												
R							1											
N																		
D									1				1					1
D									1				1					1
Y										1								
E								1				1						
K	1													1				
A															1			
V					1												1	
G																1		
V					1												1	
D									1				1					1

1. KORAK

Pripraviš matriko, poiščeš vse identične pozicije. Identično ovrednotiš z 1...

	K	L	S	C	V	L	R	E	D	Y	W	E	D	K	A	G	V	D
K	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
S	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1
D	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1
Y	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0
K	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
V	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
V	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
D	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1

1. KORAK

… različno z 0

Matriko modificiraš tako, da slediš različnim potem skozi matriko. Poiščeš samo tisto, ki da največji rezultat.

Skozi matriko greš dvakrat:

Začneš spodaj desno in greš proti zgornjemu levemu kotu po vseh možnih poteh tako, da jim določiš neko vrednost.

Ko si določil vse vrednosti za vse poti poiščeš tisto, ki je najvišje ocenjena (nekje zgoraj levo) in ji slediš nazaj proti spodnjemu desnemu kotu.

	K	L	S	C	V	L	R	E	D	Y	W	E	D	K	A	G	V	D
K	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
S	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
C	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
L	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
R	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
D	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	1
D	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	1
Y	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
E	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0
K	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
V	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
V	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	2	0
D	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1

2. KORAK

ostalim celicah določiš vrednost z naslednjo operacijo:

$$M_{i,j}=M_{i,j}+\max(M_{k,j+1}, M_{i+1,l})$$

kjer je k katerokoli celo število večje od i in l katerokoli celo število večje od j.

Torej se pomakneš v naslednji stolpec in naslednjo vrstico. V naslednji vrstici paroma **V-V** dodaš vrednost D-D. Ostalim dodaš vrednost D-D. Enako postopaš v stolpcu, kjer **V-V** dodaš vrednost D-D, ostalim pripišeš samo D-D.

	K	L	S	C	V	L	R	E	D	Y	W	E	D	K	A	G	V	D
K	1	0	0	0	0	0	0	0	0	6	6	5	5	1	3	2	1	0
S	0	0	1	0	0	0	0	0	0	6	6	5	5	4	3	2	1	0
C	0	0	0	1	0	0	0	0	0	6	6	5	5	4	3	2	1	0
I	0	0	0	0	0	0	0	0	0	6	6	5	5	4	3	2	1	0
L	0	1	0	0	0	1	0	0	7	6	6	5	5	4	3	2	1	0
R	0	0	0	0	0	0	1	0	0	6	6	5	5	4	3	2	1	0
N	0	0	0	0	0	0	0	0	0	6	6	5	5	4	3	2	1	0
D	0	0	0	0	0	0	0	0	1	6	6	5	6	4	3	2	1	1
D	0	0	0	0	0	0	0	0	1	6	6	5	6	4	3	2	1	1
Y	0	0	0	0	0	0	0	0	0	7	6	5	5	4	3	2	1	0
E	0	0	0	0	0	0	0	1	0	5	5	6	5	4	3	2	1	0
K	5	4	4	4	4	4	4	4	4	4	4	4	4	5	3	2	1	0
A	4	4	4	4	3	3	3	3	3	3	3	3	3	3	4	2	1	0
V	3	3	3	3	4	3	3	3	3	3	3	3	3	3	3	2	2	0
G	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	1	0
V	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	2	0
D	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1

NADALJUJEŠ...
Ponavljaš postopek.

Celici prišteješ navišjo vrednost **vrstice** tik zraven in pod celico ali **stolpca** takoj desno in spodaj.

	K	L	S	C	V	L	R	E	D	Y	W	E	D	K	A	G	V	D
K	13	12	11	10	10	9	8	8	7	6	6	5	5	1	3	2	1	0
S	11	11	12	10	10	9	8	8	7	6	6	5	5	4	3	2	1	0
C	10	10	10	11	10	9	8	8	7	6	6	5	5	4	3	2	1	0
I	10	10	10	10	10	9	8	8	7	6	6	5	5	4	3	2	1	0
L	9	10	9	9	9	10	8	8	7	6	6	5	5	4	3	2	1	0
R	8	8	8	8	8	8	9	8	7	6	6	5	5	4	3	2	1	0
N	8	8	8	8	8	8	8	8	7	6	6	5	5	4	3	2	1	0
D	8	8	8	8	8	8	8	8	8	6	6	5	6	4	3	2	1	1
D	7	7	7	7	7	7	7	7	8	6	6	5	6	4	3	2	1	1
Y	6	6	6	6	6	6	6	6	6	7	6	5	5	4	3	2	1	0
E	5	5	5	5	5	5	5	6	5	5	5	6	5	4	3	2	1	0
K	5	4	4	4	4	4	4	4	4	4	4	4	4	5	3	2	1	0
A	4	4	4	4	3	3	3	3	3	3	3	3	3	3	4	2	1	0
V	3	3	3	3	4	3	3	3	3	3	3	3	3	3	3	2	2	0
G	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	1	0
V	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	2	0
D	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1

… DO KONCA

Število v vsaki celici matrike je največje število identičnih parov, ki jih najdeš na poti od spodaj navzgor.

	K	L	S	C	V	L	R	E	D	Y	W	E	D	K	A	G	V	D
K	13	12	11	10	10	9	8	8	7	6	6	5	5	1	3	2	1	0
S	11	11	12	10	10	9	8	8	7	6	6	5	5	4	3	2	1	0
C	10	10	10	11	10	9	8	8	7	6	6	5	5	4	3	2	1	0
I	10	10	10	10	10	9	8	8	7	6	6	5	5	4	3	2	1	0
L	9	10	9	9	9	10	8	8	7	6	6	5	5	4	3	2	1	0
R	8	8	8	8	8	8	9	8	7	6	6	5	5	4	3	2	1	0
N	8	8	8	8	8	8	8	8	7	6	6	5	5	4	3	2	1	0
D	8	8	8	8	8	8	8	8	8	6	6	5	6	4	3	2	1	1
D	7	7	7	7	7	7	7	7	8	6	6	5	6	4	3	2	1	1
Y	6	6	6	6	6	6	6	6	6	7	6	5	5	4	3	2	1	0
E	5	5	5	5	5	5	5	6	5	5	5	6	5	4	3	2	1	0
K	5	4	4	4	4	4	4	4	4	4	4	4	4	5	3	2	1	0
A	4	4	4	4	3	3	3	3	3	3	3	3	3	3	4	2	1	0
V	3	3	3	3	4	3	3	3	3	3	3	3	3	3	3	2	2	0
G	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	1	0
V	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	2	0
D	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1

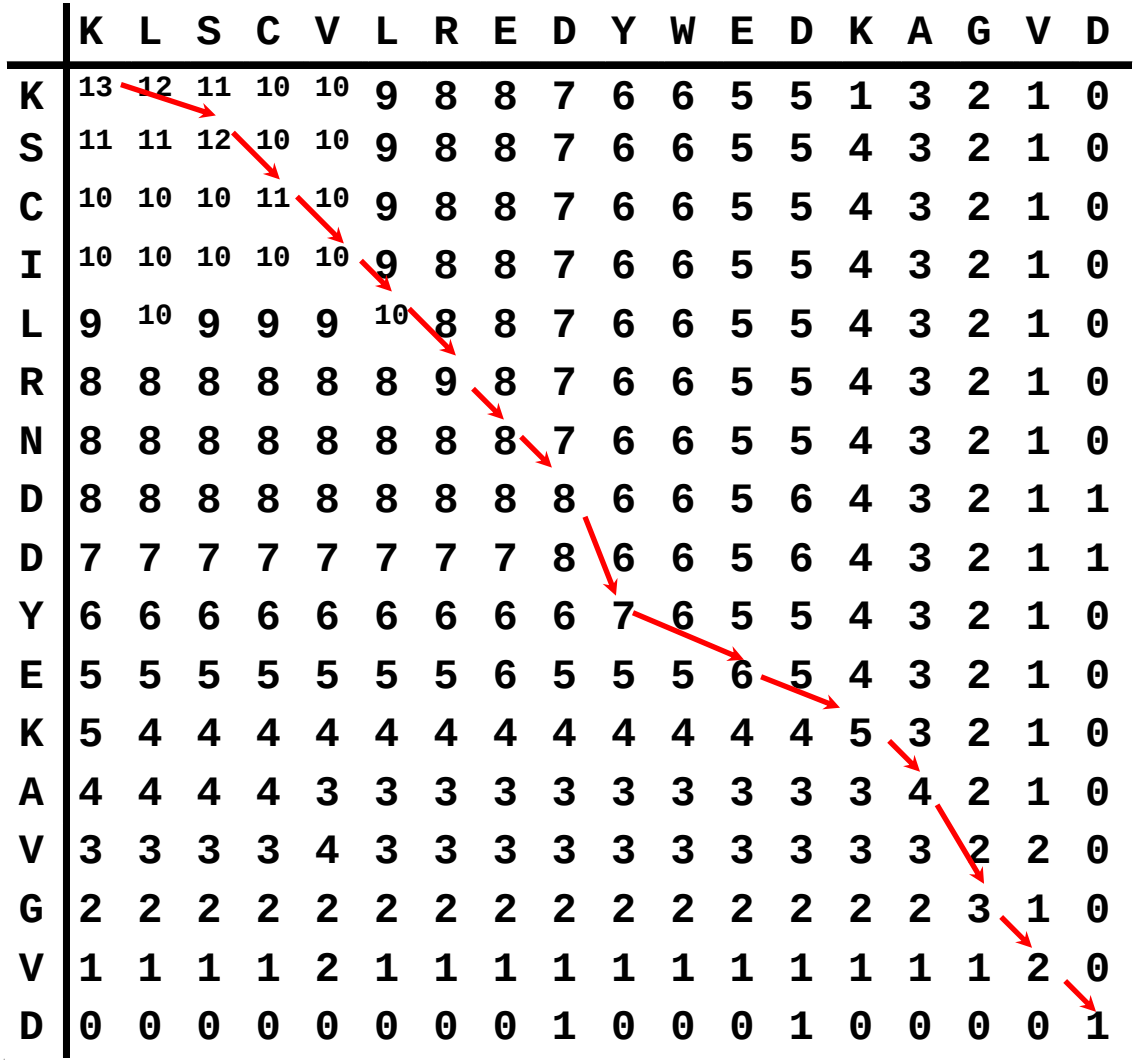
POIŠČEŠ NAJBOLJŠO PORAVNAVO OD ZAČETKA...

Rad bi najbolj optimalno poravnavo. Poiščeš **celico** z najvišjo vrednostjo nekje v zgornji vrstici ali levem stolpcu. Poravnava se tukaj začne, nadaljuješ pa navzdol in desno. Vedno se pomakneš eno **vrstico** in en **stolpec** desno in poiščeš **najvišjo vrednost** v tej vrstici ali stolpcu. Poravnava mora skozi to točko.

	K	L	S	C	V	L	R	E	D	Y	W	E	D	K	A	G	V	D
K	13	12	11	10	10	9	8	8	7	6	6	5	5	1	3	2	1	0
S	11	11	12	10	10	9	8	8	7	6	6	5	5	4	3	2	1	0
C	10	10	10	11	10	9	8	8	7	6	6	5	5	4	3	2	1	0
I	10	10	10	10	10	9	8	8	7	6	6	5	5	4	3	2	1	0
L	9	10	9	9	9	10	8	8	7	6	6	5	5	4	3	2	1	0
R	8	8	8	8	8	8	9	8	7	6	6	5	5	4	3	2	1	0
N	8	8	8	8	8	8	8	8	7	6	6	5	5	4	3	2	1	0
D	8	8	8	8	8	8	8	8	8	6	6	5	6	4	3	2	1	1
D	7	7	7	7	7	7	7	7	8	6	6	5	6	4	3	2	1	1
Y	6	6	6	6	6	6	6	6	6	7	6	5	5	4	3	2	1	0
E	5	5	5	5	5	5	5	6	5	5	5	6	5	4	3	2	1	0
K	5	4	4	4	4	4	4	4	4	4	4	4	4	5	3	2	1	0
A	4	4	4	4	3	3	3	3	3	3	3	3	3	3	4	2	1	0
V	3	3	3	3	4	3	3	3	3	3	3	3	3	3	3	2	2	0
G	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	1	0
V	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	2	0
D	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1

… DO KONCA

Ko tako nadaljuješ prideš do spodnje vrstice ali desnega stolpca.



	K	L	S	C	V	L	R	E	D	Y	W	E	D	K	A	G	V	D
K	13	12	11	10	10	9	8	8	7	6	6	5	5	1	3	2	1	0
S	11	11	12	10	10	9	8	8	7	6	6	5	5	4	3	2	1	0
C	10	10	10	11	10	9	8	8	7	6	6	5	5	4	3	2	1	0
I	10	10	10	10	10	9	8	8	7	6	6	5	5	4	3	2	1	0
L	9	10	9	9	9	10	8	8	7	6	6	5	5	4	3	2	1	0
R	8	8	8	8	8	8	9	8	7	6	6	5	5	4	3	2	1	0
N	8	8	8	8	8	8	8	8	7	6	6	5	5	4	3	2	1	0
D	8	8	8	8	8	8	8	8	8	6	6	5	6	4	3	2	1	1
D	7	7	7	7	7	7	7	7	8	6	6	5	6	4	3	2	1	1
Y	6	6	6	6	6	6	6	6	6	7	6	5	5	4	3	2	1	0
E	5	5	5	5	5	5	5	6	5	5	5	6	5	4	3	2	1	0
K	5	4	4	4	4	4	4	4	4	4	4	4	4	5	3	2	1	0
A	4	4	4	4	3	3	3	3	3	3	3	3	3	3	4	2	1	0
V	3	3	3	3	4	3	3	3	3	3	3	3	3	3	3	2	2	0
G	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	1	0
V	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	2	0
D	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1

REZULTAT

Optimalna poravnava, kombinacija obeh zacetnih, dolocenih “na oko”.

```
KLSCVLRED-YWEDKA-GVD
* ** ** * * * ** ***
K-SCILRNDDY-E-KAVGVD
```

```
KLSCVLRED-YWEDKAGV--D
* ** ** * * * ** *   *
K-SCILRNDDY-E-KA-VGVD
```

```
KLSCVLRE-DYWEDKA-GVD
* ** **  ** * ** ***
K-SCILRNDDY-E-KAVGVD
```

Startna točka poti z največjo vrednostjo je vedno na N-terminalnem delu proteina. Izračunana je s pomočjo procesa točkovanja, ki se začne na C-terminalnem delu- rezultat je **GLOBALNA** poravnava.

SMITH-WATERMAN-ov ALGORITEM

Smith TF and Waterman SM (1981) Identification of common molecular subsequences. *Journal of Molecular Biology* 147, 195-197.

Poravna dve zaporedji **LOKALNO**. Kadar so zaporedja podobna samo na enem delu in bi se podobnost izgubila, če bi jih primerjali globalno, npr. pri evolucijsko oddaljenih proteinih, kjer so ohranjene samo lokalne homologije.

Spremenjen Needleman-Wunschev algoritem:

- **negativna vrednost (utež) se določi neujemanjem**

Vrednost celic bo padla, ko pridemo v območja neujemanja in narasla v podobnih delih zaporedij.

- **nič mora biti najmanjša vrednost v matriki**

Zato, da se vsak kratek del podobnih zaporedij začne z nič. Če je rezultat v celici manjši kot nič mu navadno pripiše vrednost 0.

- **začetek ali konec najbolj optimalne poti lahko najdemo kjerkoli v matriki in ne samo v zadnji vrstici ali stolpcu**

Preiskati moramo celo matriko za dele zaporedij z visoko lokalno podobnostjo.

	K	L	S	C	V	L	R	E	D	Y	W	E	D	K	A	G	V	D
K	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
S	0	0.5	2	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
C	0	0.5	0.5	3	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5
I	0	0.5	0.5	1.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5
L	0	2	0.5	1.5	2.5	3.5	2	2	2	2	2	2	2	2	2	2	2	2
R	0	0.5	1.5	1.5	2.5	2.5	4.5	3	3	3	3	3	3	3	3	3	3	3
N	0	0.5	1.5	1.5	2.5	2	3	4	4	4	4	4	4	4	4	4	4	4
D	0	0.5	1.5	1.5	2.5	2	3	4	5	3.5	3.5	3.5	5	3.5	3.5	3.5	3.5	5
D	0	0.5	1.5	1.5	2.5	2	3	4	5	4.5	4.5	4.5	5	4.5	4.5	4.5	4.5	6
Y	0	0.5	1.5	1.5	2.5	2	3	4	3.5	6	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5
E	0	0.5	1.5	1.5	2.5	2	3	5.5	3.5	4.5	5.5	7	5.5	5.5	5.5	5.5	5.5	5.5
K	1	0.5	1.5	1.5	2.5	2	3	4	5	5	5.5	5	6.5	8	6.5	6.5	6.5	6.5
A	0	0.5	1.5	1.5	2.5	2	3	4	5	4.5	5.5	5	6.5	6	9	7.5	7.5	7.5
V	0	0.5	1.5	1.5	4	2	3	4	5	4.5	5.5	5	6.5	6	7.5	8.5	10	8.5
G	0	0.5	1.5	1.5	2.5	3.5	3.5	4	5	4.5	5.5	5	6.5	6	7.5	10	8	9.5
V	0	0.5	1.5	1.5	4	3.5	3	4	5	4.5	5.5	5	6.5	6	7.5	8.5	11	9.5
D	0	0.5	1.5	1.5	2.5	3.5	3.5	4	6.5	4.5	5.5	5	8	6	7.5	8.5	10.5	12

Začneš v zgornjem levem kotu. Vsakemu elementu matrike določiš vrednost po enačbi:

$$M_{i,j}=M_{i,j}+\max(M_{k,j-1}, M_{i-1,l})$$

kjer je k katerokoli celo število manjše od i in l katerokoli celo število manjše od j.

Vsako neujemanje smo kaznovali z -0.5.

Rezultat je ista poravnava kot pri Needleman-Wunsch-ovem algoritmu.

```
KLSCVLRED-YWEDKA-GVD
*  **  **  *  *  *  **  ***
K-SCILRNDDY-E-KAVGVD
```

PRIMERJAVA OBEH ALGORITMOV

NEEDLEMAN-WUNSCH

GLOBALNA poravnava- optimalno poravna vse elemente zaporedij

Ne zahteva dodatne teže za vrzel

Rezultat se ne more zmanjšati med dvema celicama na poti

```
TTGACACCCTCC-CAATTGTA
    **   **    **   *
ACCCCAGGCTTTACACAT---
```

SMITH-WATERMAN

LOKALNA poravnava- optimalno poravna samo del ali dele zaporedij

Zahteva dodatno težo za vrzel, da učinkovito deluje

Rezultat se lahko tudi zmanjša med dvema celicama na poti

```
---------TTGACACCCTCCCAATTGTA
         **  ****
ACCCCAGGCTTTACACAT----------
```

OCENJEVANJE

SCORING

Kako določiti uteži za neujemanje?
Uporaba ocenjevalnih matrik:
Vseh 20 aminokislin primerjamo z ostalimi 20 aminokislinami in jim določimo uteži.

MATRIKE ZAMENJAV

1. Identična matrika (*Identity matrix*):
Ujemanje je vredno 1 točko, neujemanje 0

2. Ocenjevanje na podlagi genetskega koda (*Genetic code scoring*):
Minimalno število zamenjav nukleotidov potrebnih za zamenjavo kodona ene aminokisline s kodonom druge aminokisline

3. Ocenjevanje na podlagi kemijske podobnosti (*Chemical similarity scoring*):
Ocenjevanje na podlagi kemijske podobnosti stranskih ostankov aminokislin (npr. polarni, nepolarni, veliki, majhni, nabiti).

4. Določene na podlagi opaženih zamenjav (*Observed substitutions*):
Na podlagi poravnah homolognih proteinov so določili uteži za zamenjave posameznih aminokislin. Več matrik:

PAM (***P****ercentage of* ***A****ccepted* ***M****utations*)
BLOSUM (***BL****ocks* ***SU****bstitution* ***M****atrix*)
GONNET
Na podlagi 3D struktur

IDENTIČNA MATRIKA

IDENTITY MATRIX

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	1																			
S	0	1																		
T	0	0	1																	
P	0	0	0	1																
A	0	0	0	0	1															
G	0	0	0	0	0	1														
N	0	0	0	0	0	0	1													
D	0	0	0	0	0	0	0	1												
E	0	0	0	0	0	0	0	0	1											
Q	0	0	0	0	0	0	0	0	0	1										
H	0	0	0	0	0	0	0	0	0	0	1									
R	0	0	0	0	0	0	0	0	0	0	0	1								
K	0	0	0	0	0	0	0	0	0	0	0	0	1							
M	0	0	0	0	0	0	0	0	0	0	0	0	0	1						
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1					
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1				
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1			
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

	A	C	G	T
A	1			
C	0	1		
G	0	0	1	
T	0	0	0	1

Najbolj osnovna ocenjevalna matrika.

Identičnemu elementu pripiše vrednost 1, različnemu vrednost 0.

Manj efektivna pri zaporedjih, kjer so prisotne le šibke homologije.

Bolj primerna za primerjavo DNA ali RNA zaporedij

OCENJEVANJE NA PODLAGI GENETSKEGA KODA

GENETIC CODE SCORING

Fitch WM (1966) An improved method for testing for evolutionary homology. *Journal of Molecular Biology* 16, 9-16.

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	0																			
S	1	0																		
T	2	1	0																	
P	2	2	1	0																
A	2	1	1	1	0															
G	1	1	2	2	1	0														
N	2	2	1	2	2	2	0													
D	2	2	2	2	1	1	1	0												
E	2	2	2	2	1	1	2	1	0											
Q	2	2	2	1	2	2	2	2	1	0										
H	2	2	2	1	2	2	1	1	2	1	0									
R	1	1	1	1	2	1	2	2	2	1	1	0								
K	2	2	1	2	2	2	1	2	1	1	2	1	0							
M	2	2	1	2	2	2	2	2	2	2	2	1	1	0						
I	2	1	1	2	2	2	1	2	2	2	2	1	1	1	0					
L	2	1	2	1	2	2	2	2	2	1	1	1	2	1	1	0				
V	2	2	2	2	1	1	2	1	1	2	2	2	2	1	1	1	0			
F	1	1	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	0		
Y	1	1	2	2	2	2	1	1	2	2	1	2	2	3	2	2	2	2	0	
W	1	1	2	2	2	1	2	2	2	2	2	1	2	2	2	1	2	2	2́	0

Redko uporabljana matrika, za konstrukcijo filogenetskih dreves.

Aminokislinam pripiše vrednost, ki ustreza številu nukleotidnih zamenjav (0, 1, 2 ali 3), potrebnih za zamenjavo kodonov dveh aminokislin.

PAM MATRIKA

P*ercentage of* A*ccepted* M*utations or* P*oint* A*ccepted* M*utations*

Dayhoff MO *et al.* (1978) A model of evolutionary change in proteins. In Atlas of proteins and sequence structure (ed. M. Dayhoff), Vol. 5, Suppl 3, 345-352. National biomedical research foundation, Silver Spring, MD, USA.

Matrika uteži na podlagi pogostosti zamenjav aminokislin z drugimi aminokislinami. Predlagane na podlagi stotin poravnav zelo sorodnih proteinov. Iz poravnav so izračunali frekvence zamenjav aminokislin pri zaporedjih majhne evolucijske razdalje, pod 1% divergence (v povprečju 1 zamenjava na 100 aminokislin). Rezultat je matrika zamenjav **PAM1**.

*Evolucijska razdalja **PAM1** označuje verjetnost, da se je zamenjala 1 aminokislino na 100 aminokislin. Npr. primerjamo proteine, ki so se spremenili za 1%.*

Predpostavijo, da proteinska zaporedja divergirajo zaradi kumulativnih nepovezanih sprememb.

S pomočjo PAM1 matrike lahko generiramo PAM-k matrike s k-kratno pomnožitvijo same sebe. Uporabljamo jih za primerjavo zaporedij, ki so se spremenile za k % ali so k evolucijskih enot narazen. Npr. **PAM250** poda frekvenco mutacij proteinov, ki so se spremenili za 250% (250 mutacij na 100 aminokislin).

Evolucijska razdalja 1 PAM

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E	10	0	7	**56**	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Matrika verjetnosti mutacij za evolucijsko razdaljo 1 PAM (1 mutacija na 100 aminokislin) (verjetnosti so pomnožene z 10 000). Element M_{ij} pomeni verjetnost da je aminokislina v koloni (j) zamenjana z aminokislino v vrstici (i) pri določenem evolucijskem intervalu (1 PAM). Npr. obstaja **0.56 %** verjetnosti, da bo Asp zamenjan z Glu (Atlas of Protein Sequence and Structure, Suppl 3, 1978, M.O. Dayhoff, ed. National Biomedical Research Foundation, 1979).

Evolucijska razdalja 250 PAM

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17

Matrika verjetnosti mutacij za evolucijsko razdaljo 250 PAM (250 mutacij na 100 aminokislin) (verjetnosti so pomnožene s 100).

Npr. obstaja **13 %** verjetnosti, da bo Ala v prvem zaporedju ohranjen tudi v drugem. 3 % verjetnosti je, da bo v drugem zaporedju na tem mestu Arg.

Programi za iskanje in primerjavo zaporedij uporabljajo **matriko logaritemskih obetov (*Log Odds Matrix, Dayhoff matrix*)**, ki temelji na matriki verjetnosti mutacij za evolucijsko razdaljo 250PAM. Izračunajo jo tako, da izračunajo logaritem razmerja verjetnosti mutacij in relativne frekvence, da določena aminokislina mutira naključno:

$$S_{ij} = \log(M_{ij} / f_i)$$

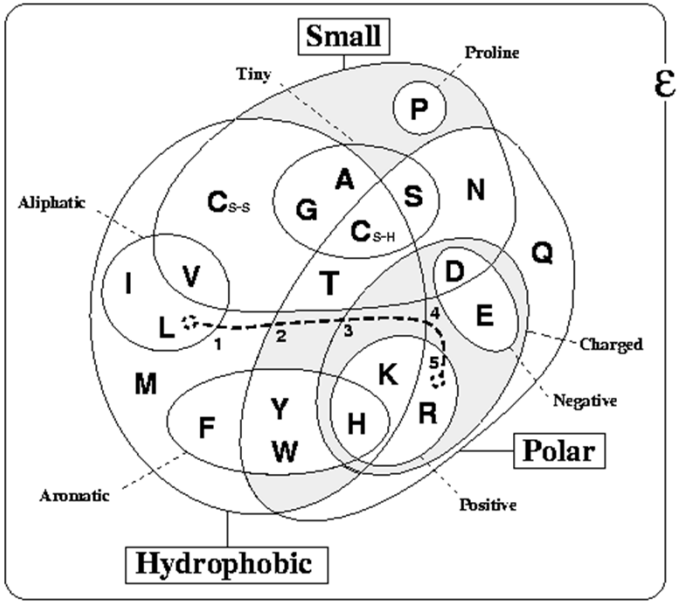
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12																			
S	0	2																		
T	-2	1	3																	
P	-3	1	0	6																
A	-2	1	1	1	2															
G	-3	1	0	-1	1	5														
N	-4	1	0	-1	0	0	2													
D	-5	0	0	-1	0	1	2	4												
E	-5	0	0	-1	0	0	1	3	4											
Q	-5	-1	-1	0	0	-1	1	2	2	4										
H	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	8								
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	8				
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

Vrednosti večje od 0 označujejo bolj verjetne mutacije (se pojavijo bolj verjetno med našima dvema zaporedjima kot med dvema naključnima) vrednosti 0 so nevtralne mutacije (naključne) vrednosti manjše od 0 predstavljajo manj verjetne mutacije.

Vrednost –10 pomeni, da pričakujemo, da bomo našli določen par aminokislin desetkrat manj kot če bi ga našli naključno.

Pri iskanju zaporedij po bazah uporabljamo najpogosteje matriko PAM250.
Ali je takšna visoka divergenca (250 mutacij na 100 aminokislin) ustrezna za primerjave?
Ko evolucijska razdalja narašča, narašča tudi verjetnost, da dobimo po mutaciji nazaj isto aminokislino.

Procent različnosti	Evolucijska razdalja v PAM
1	1
5	5
10	11
20	23
30	38
40	56
50	80
60	112
70	159
80	246



Pri PAM250 ostane še vedno identičnih 20 % aminokislin. Vendar aminokisline mutirajo različno: identičnih ostane 55 % Trp, 52 Cys, 27 % Gly in samo 7 % Ser, 6 % Asn.

Aminokisline so v matriki združene po velikosti, obliki, naboju, sposobnostjo tvorbe ionskih, hidrofobnih in vodikovih vezi. Ta združevanja so rezultat naravne selekcije in so dosti manj omejena z genetskim kodom (npr. matrika ocen na podlagi genetskega koda).

BLOSUM MATRIKE

BLO*cks* SU*bstitution* M*atrix*

Henikoff S and Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences USA 89, 10915-10919.

PAM matrika je omejena: določena je bila iz omejenega seta podatkov (samo majhni globularni proteini), z opazovanjem zelo sorodnih proteinov so bile poudarjene aminokisline, ki bolj mutirajo in ob predpostavki, da vsa mesta enako mutirajo.

BLOSUM matrike so bile izračunane iz lokalnih poravnav, blokov (blocks), sorodnih proteinov pri različnih stopnjah divergence. Ti bloki so predstavljali seznam dovoljenih zamenjav. Matrika logaritemskih obetov je bila izračunana tako, da so izračunali logaritem razmerja opaženih frekvenc posameznih parov aminokislin (q_{ab}) in pričakovanih frekvenc istih parov določenih iz populacije vseh parov (p_{ab}):

$$s_{ab} = \log(q_{ab}/e_{ab})$$

Zato, da so izničili vpliv identičnih parov aminokislin pri zelo sorodnih zaporedjih, so zaporedja združili v gruče na podlagi minimalne identičnosti (npr. 6 2%). Za vsako gručo so izračunali povprečni prispevek vsakega položaja, torej so gruče efektivno obravnavali kot eno zaporedje. Na tak način so lahko izračunali različne matrike za različno identičnost gruč (30-100 %). Npr. zaporedja, ki so več kot 80 % identična so uporabili za izdelavo BLOSUM80 matrike, tiste, ki so več kot 62 % identična za pripravo BLOSUM62 matrike itn.

BLOSUM62 matrika

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

Če je opaženo število razlik med dvema aminokislinama enako pričakovanim potem je $s_{ab}=0$, če je manjše od pričakovanega potem je $s_{ab}<0$ in če je večje od pričakovanega je $s_{ab}>0$.

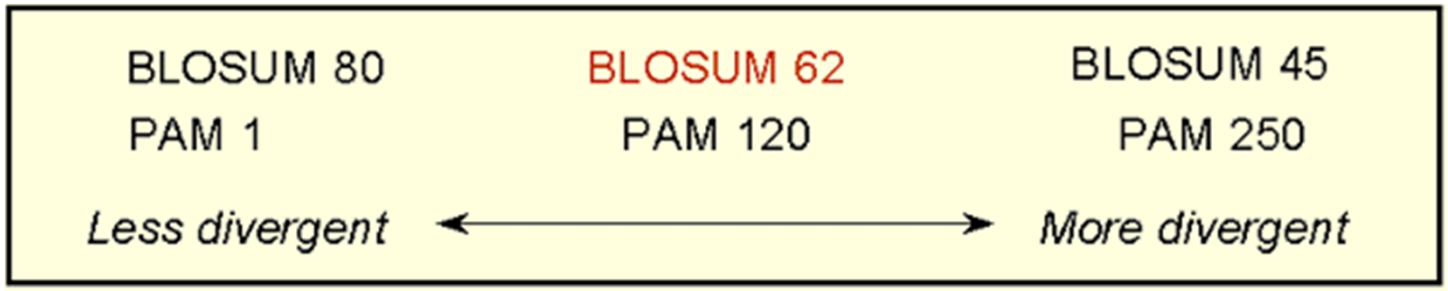
BLOSUM matrike dopuščajo manj zamenjav hidrofilnih aminokislin, medtem ko dopuščajo več hidrofobnih zamenjav in neujemanj cisteinov in triptofanov.

Katero matriko uporabimo?

Najbolje je primerjati zaporedji z matriko, ki ustreza njuni evolucijski razdalji:

PAM40, PAM120	za zelo podobna kratka zaporedja
PAM250	dolga zaporedja z nizko podobnostjo
BLOSUM50, BLOSUM62	za detekcijo šibkih homologij

V praksi ne poznaš evolucijske razdalje med obema zaporedjima, zato moraš uporabiti več različnih matrik.



http://www.ncbi.nlm.nih.gov/Education/BLASTinfo

PRIMER

```
>gi|122615|sp|P02023|HBB_HUMAN HEMOGLOBIN BETA CHAIN
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPVKAHGKKVLGAF
SDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALA
HKYH

>gi|70497|pir||HBAK hemoglobin beta chain - Nile crocodile
ASFDPHEKQLIGDLWHKVDVAHCGGEALSRMLIVYPWKRRYFENFGDISNAQAIMHNEKVQAHGKKVLASF
GEAVCHLDGIRAHFANLSKLHCEKLHVDPENFKLLGDIIIIVLAAHYPKDFGLECHAAYQKLVRQVAAALA
AEYH
```

BLOSUM62

```
Score = 168 bits (425), Expect = 7e-42
Identities = 79/142 (55%), Positives = 105/142 (73%), Gaps = 1/142 (0%)

Query: 6   PEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP-VKAHG 64
           P EK  +  LW KV+V   GGEAL R+L+VYPW +R+FE+FGD+S   A+M N  V+AHG
Sbjct: 5   PHEKQLIGDLWHKVDVAHCGGEALSRMLIVYPWKRRYFENFGDISNAQAIMHNEKVQAHG 64

Query: 65  KKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTP 124
           KKVL +F + + HLD ++  FA LS+LHC+KLHVDPENF+LLG++++ VLA H+ K+F
Sbjct: 65  KKVLASFGEAVCHLDGIRAHFANLSKLHCEKLHVDPENFKLLGDIIIIVLAAHYPKDFGL 124

Query: 125 PVQAAYQKVVAGVANALAHKYH 146
              AAYQK+V  VA ALA +YH
Sbjct: 125 ECHAAYQKLVRQVAAALAAEYH 146
```

PAM250

```
Score = 141 bits (474), Expect = 3e-34
Identities = 78/144 (54%), Positives = 104/144 (72%), Gaps = 1/144 (0%)

Query: 4   LTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVM-GNPVKA 62
           + P+EK  +  LW KV+V + GGEAL+R+L+VYPW +R+FE+FGD+S ++A+M  + V+A
Sbjct: 3   FDPHEKQLIGDLWHKVDVAHCGGEALSRMLIVYPWKRRYFENFGDISNAQAIMHNEKVQA 62

Query: 63  HGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEF 122
           HGKKVL++F++++ HLD ++  FA LS LHC+KLHVDPENF+LLG++++ VLA H+ K+F
Sbjct: 63  HGKKVLASFGEAVCHLDGIRAHFANLSKLHCEKLHVDPENFKLLGDIIIIVLAAHYPKDF 122

Query: 123 TPPVQAAYQKVVAGVANALAHKYH 146
              +AAYQK+V  VA ALA  YH
Sbjct: 123 GLECHAAYQKLVRQVAAALAAEYH 146
```

BLOSUM50

Score = 157 bits (550), Expect = 4e-39
Identities = 79/142 (55%), Positives = 105/142 (73%), Gaps = 1/142 (0%)

```
Query: 6   PEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP-VKAHG 64
           P EK  +  LW KV+V   GGEAL R+L+VYPW +R+FE+FGD+S   A+M N  V+AHG
Sbjct: 5   PHEKQLIGDLWHKVDVAHCGGEALSRMLIVYPWKRRYFENFGDISNAQAIMHNEKVQAHG 64

Query: 65  KKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTP 124
           KKVL +F + + HLD ++  FA LS+LHC+KLHVDPENF+LLG++++ VLA H+ K+F
Sbjct: 65  KKVLASFGEAVCHLDGIRAHFANLSKLHCEKLHVDPENFKLLGDIIIIVLAAHYPKDFGL 124

Query: 125 PVQAAYQKVVAGVANALAHKYH 146
             +AAYQK+V  VA ALA +YH
Sbjct: 125 ECHAAYQKLVRQVAAALAAEYH 146
```

BLOSUM62

Score = 168 bits (425), Expect = 7e-42
Identities = 79/142 (55%), Positives = 105/142 (73%), Gaps = 1/142 (0%)

```
Query: 6   PEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP-VKAHG 64
           P EK  +  LW KV+V   GGEAL R+L+VYPW +R+FE+FGD+S   A+M N  V+AHG
Sbjct: 5   PHEKQLIGDLWHKVDVAHCGGEALSRMLIVYPWKRRYFENFGDISNAQAIMHNEKVQAHG 64

Query: 65  KKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTP 124
           KKVL +F + + HLD ++  FA LS+LHC+KLHVDPENF+LLG++++ VLA H+ K+F
Sbjct: 65  KKVLASFGEAVCHLDGIRAHFANLSKLHCEKLHVDPENFKLLGDIIIIVLAAHYPKDFGL 124

Query: 125 PVQAAYQKVVAGVANALAHKYH 146
              AAYQK+V  VA ALA +YH
Sbjct: 125 ECHAAYQKLVRQVAAALAAEYH 146
```

BLOSUM90

Score = 195 bits (459), Expect = 2e-49
Identities = 79/142 (55%), Positives = 105/142 (73%), Gaps = 1/142 (0%)

```
Query: 6   PEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP-VKAHG 64
           P EK  +  LW KV+V   GGEAL R+L+VYPW +R+FE FGD+S   A+M N  V+AHG
Sbjct: 5   PHEKQLIGDLWHKVDVAHCGGEALSRMLIVYPWKRRYFENFGDISNAQAIMHNEKVQAHG 64

Query: 65  KKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTP 124
           KKVL +F +   HLD ++  FA LS LHC+KLHVDPENF+LLG++++ VLA H+ K+F
Sbjct: 65  KKVLASFGEAVCHLDGIRAHFANLSKLHCEKLHVDPENFKLLGDIIIIVLAAHYPKDFGL 124

Query: 125 PVQAAYQKVVAGVANALAHKYH 146
             +AAYQK V  VA ALA  YH
Sbjct: 125 ECHAAYQKLVRQVAAALAAEYH 146
```

TEŽA VRZELI

GAP PENALTY

Ni matematičnega modela, ki bi opisal razvoj vrzeli.
Splošno sprejeto je teža vrzeli podana s:

$$t = v + u*k$$

Kjer je v teža za vstavitev vrzeli, u teža za vsako nadaljno vzrel in k dolžina vrzeli.

Bolj biološki pomen- večja teža za vstavitev vrzeli, manjša za podaljševanje (običajno poteka več delecij ali insercij skupaj).

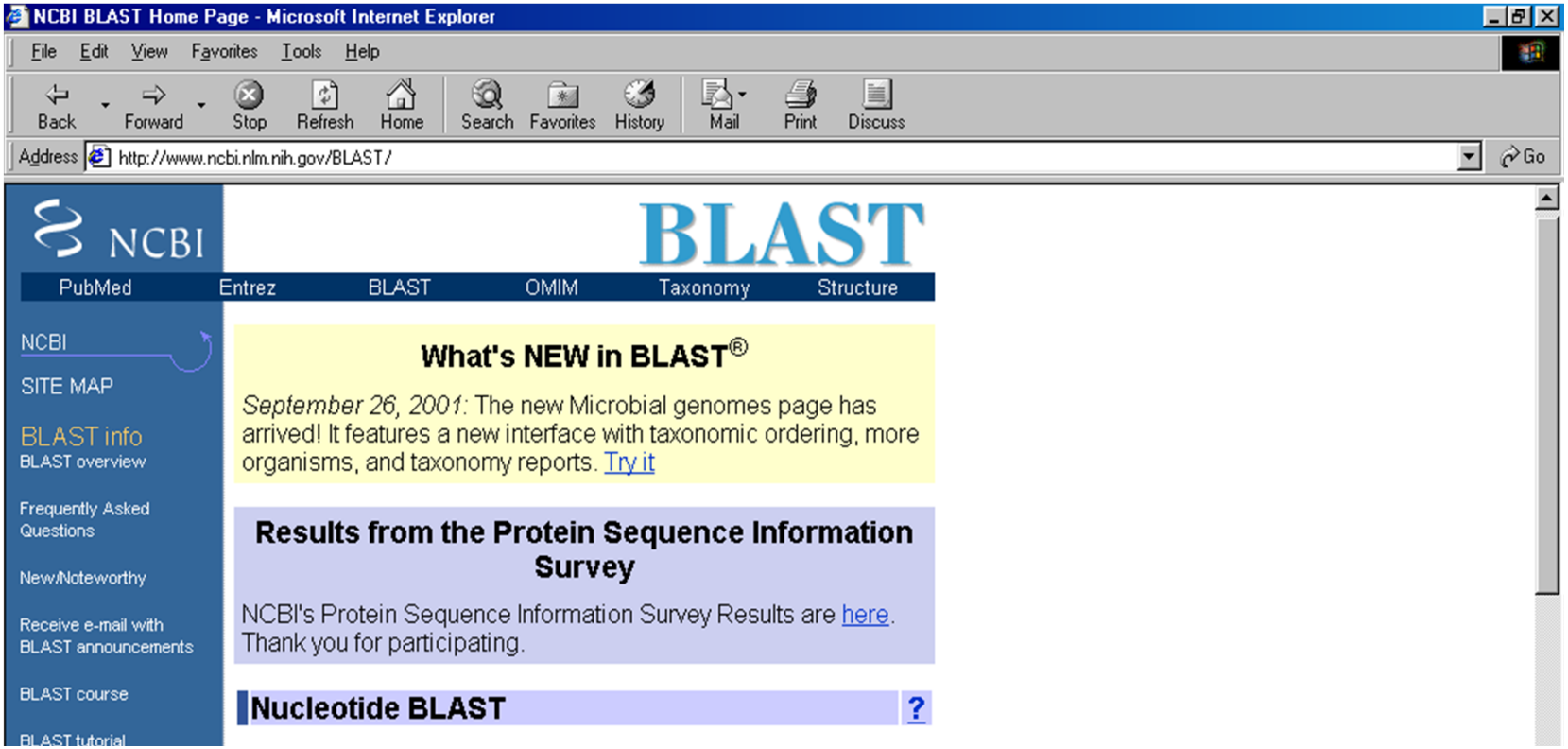
BLAST ALGORITEM

*(**B**asic **L**ocal **A**lignment **S**earch **T**ools)*

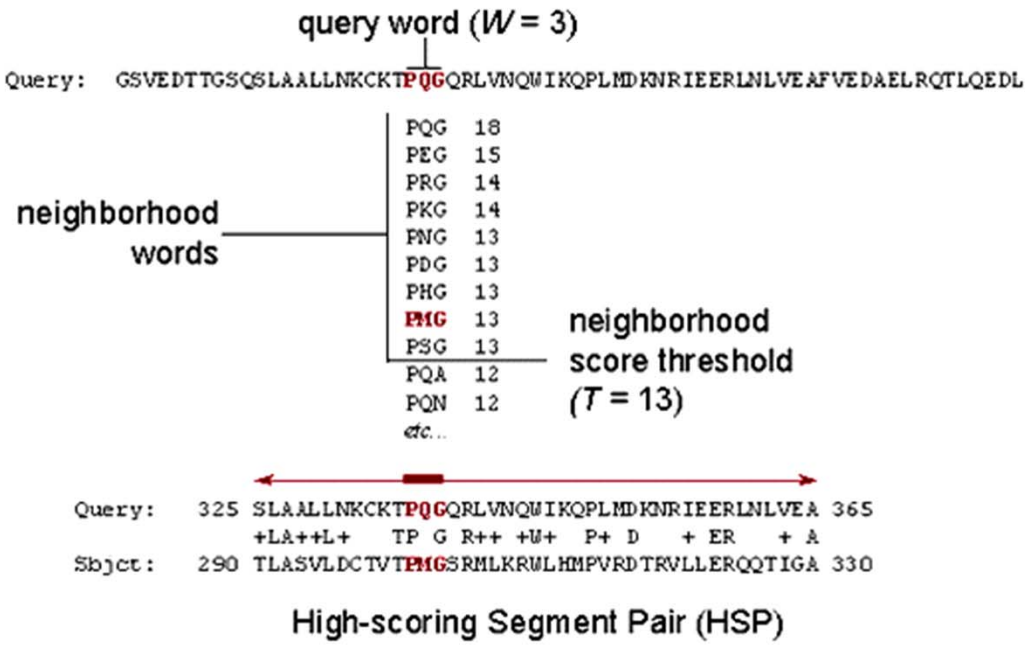
Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215, 403-410.

Hitro iskanje po podatkovnih zbirkah.

BLAST na internetu:
http://www.ncbi.nlm.nih.gov/BLAST/



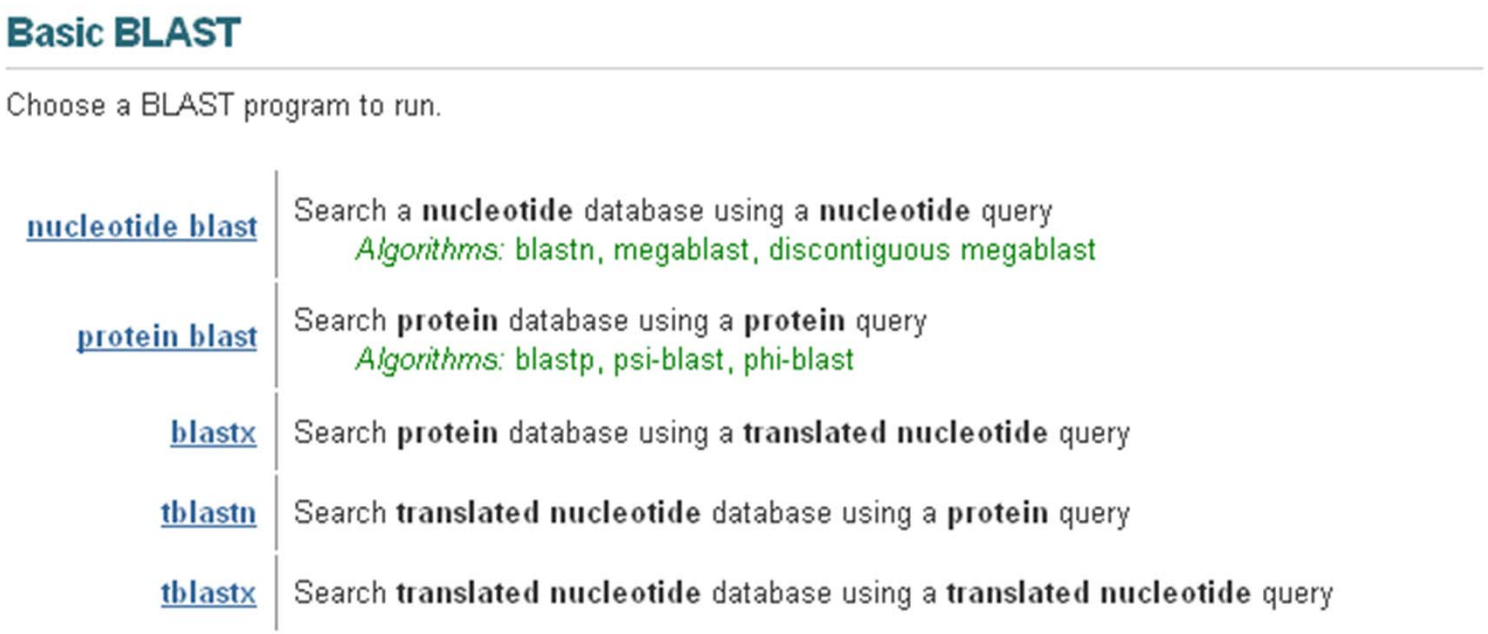
The BLAST Search Algorithm



http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/BLAST_algorithm.html

BLAST ALGORITEM

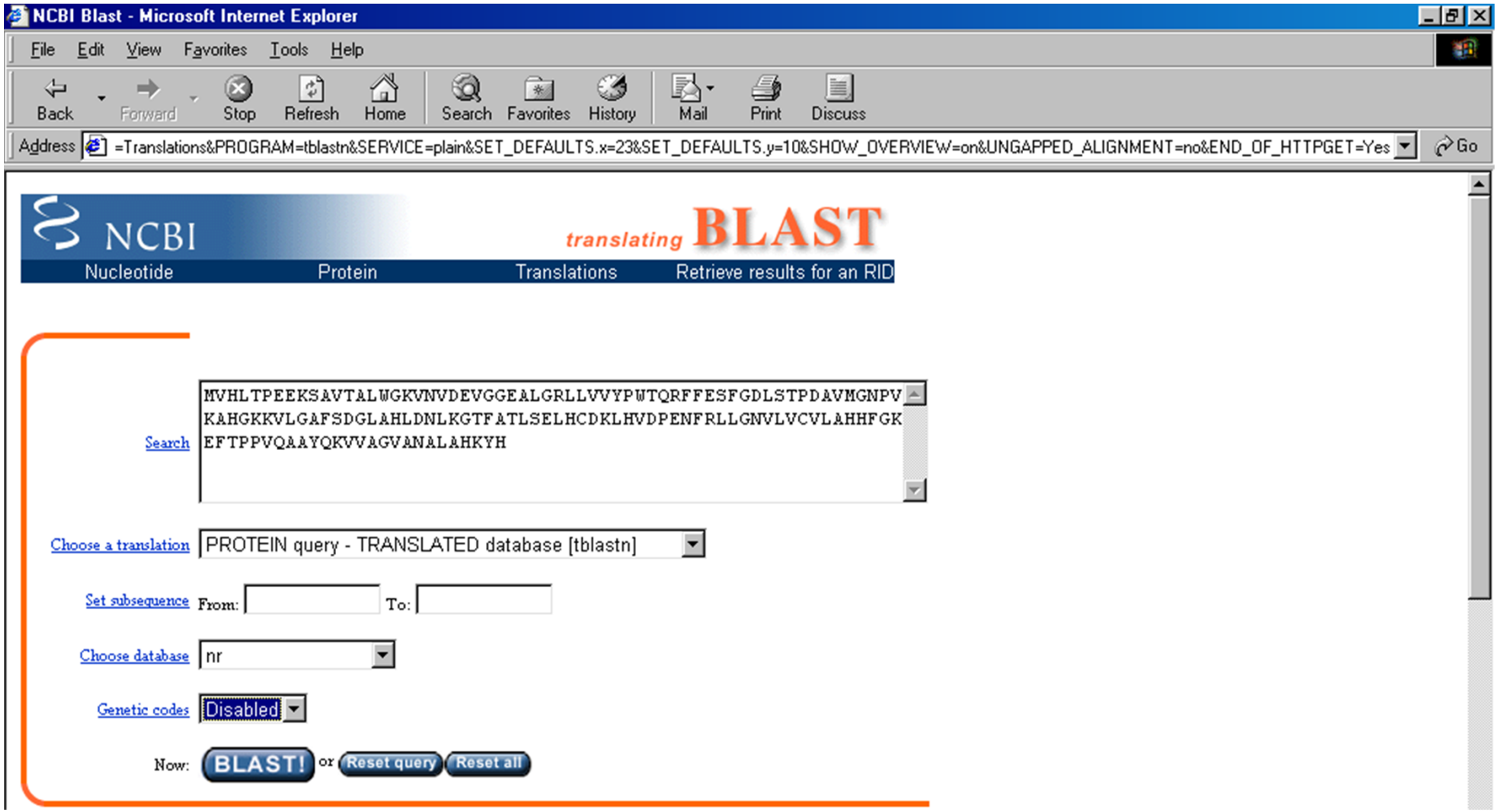
BLAST algoritem je hevristična metoda, ki razbije vnešeno zaporedje in zaporedja v bazi na krajša zaporedja. Išče zaporedja dolžine W, ki dajo rezultat vsaj T, ko jih primerjamo z zaporedji v bazi. Ovrednotimo jih z eno od matrik zamenjav. Zaporedja v bazi, ki imajo rezultat vsaj T ali večji, podaljšamo na vsako stran tako, da najdemo lokalno najbolj optimalno poravnavo ali par z visokim rezultatom (*High Scoring Pair, HSP*). Ta ima vrednosti S ali E nižje kot določen prag. Poravnave, ki ustrezajo tem pogojem bo BLAST podal kot rezultat.

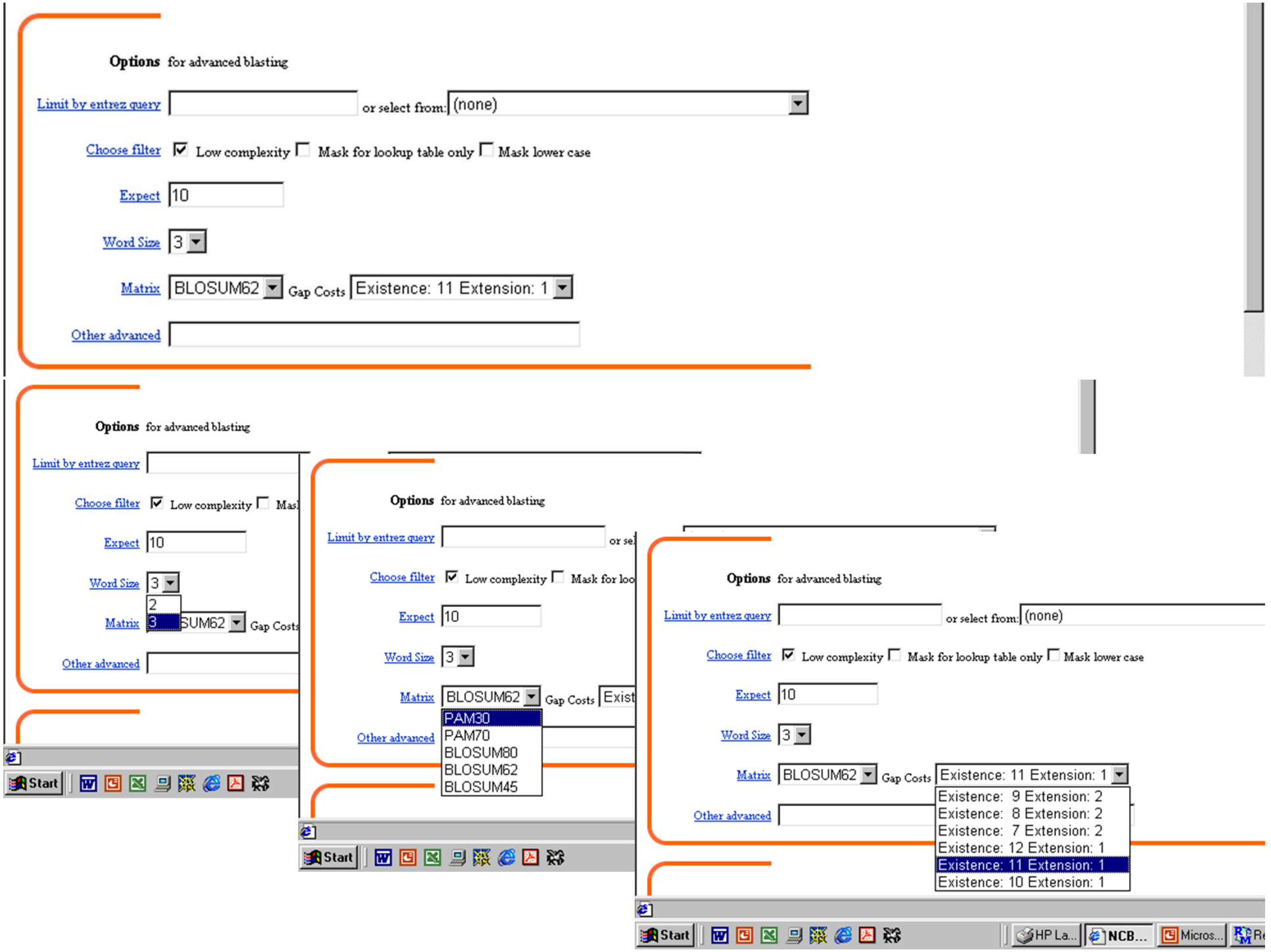


Program	Tarča	Podatkovna zbirka	Primerjava na nivoju	Uporaba
blastn	DNA	DNA	DNA	Išče identična DNA zaporedja
blastp	protein	protein	proteinskem	Išče homologne proteine
blastx	DNA	protein	proteinskem	Analiziraš novo DNA ko iščeš gene in homologne proteine
tblastn	protein	DNA	proteinskem	Išče gene v DNA zaporedjih, ki še niso identificirana
tblastx	DNA	DNA	proteinskem	Struktura genov

PRIMER

>gi|122615|sp|P02023|HBB_HUMAN HEMOGLOBIN BETA CHAIN
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH


NCBI Blast - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Discuss
Address =Translations&PROGRAM=tblastn&SERVICE=plain&SET_DEFAULTS.x=23&SET_DEFAULTS.y=10&SHOW_OVERVIEW=on&UNGAPPED_ALIGNMENT=no&END_OF_HTTPGET=Yes
Go
NCBI
translating BLAST
Nucleotide Protein Translations Retrieve results for an RID
Search
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPV
KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPPVQAAYQKVVAGVANALAHKYH
Choose a translation
PROTEIN query - TRANSLATED database [tblastn]
Set subsequence
From:
To:
Choose database
nr
Genetic codes
Disabled
Now:
BLAST!
or
Reset query
Reset all


Options for advanced blasting
Limit by entrez query
or select from: (none)
Choose filter
Low complexity
Mask for lookup table only
Mask lower case
Expect
10
Word Size
3
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
Other advanced
2
3
PAM30
PAM70
BLOSUM80
BLOSUM62
BLOSUM45
Existence: 9 Extension: 2
Existence: 8 Extension: 2
Existence: 7 Extension: 2
Existence: 12 Extension: 1
Existence: 11 Extension: 1
Existence: 10 Extension: 1
Start
HP La...
NCB...
Micros...

FILTRIRANJE

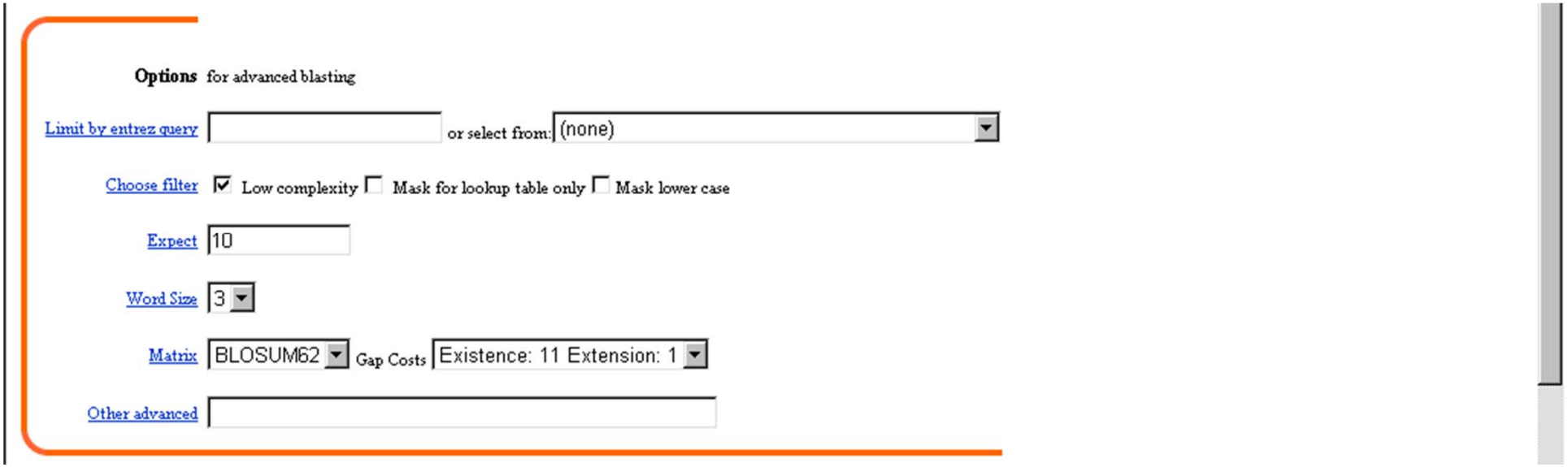
Ponavljajoče regije v zaporedjih (npr. poliA v DNA ali ponovitve aminokislin pri proteinih) ne upošteva pri analizi, ker bi močno zvišale rezultat.

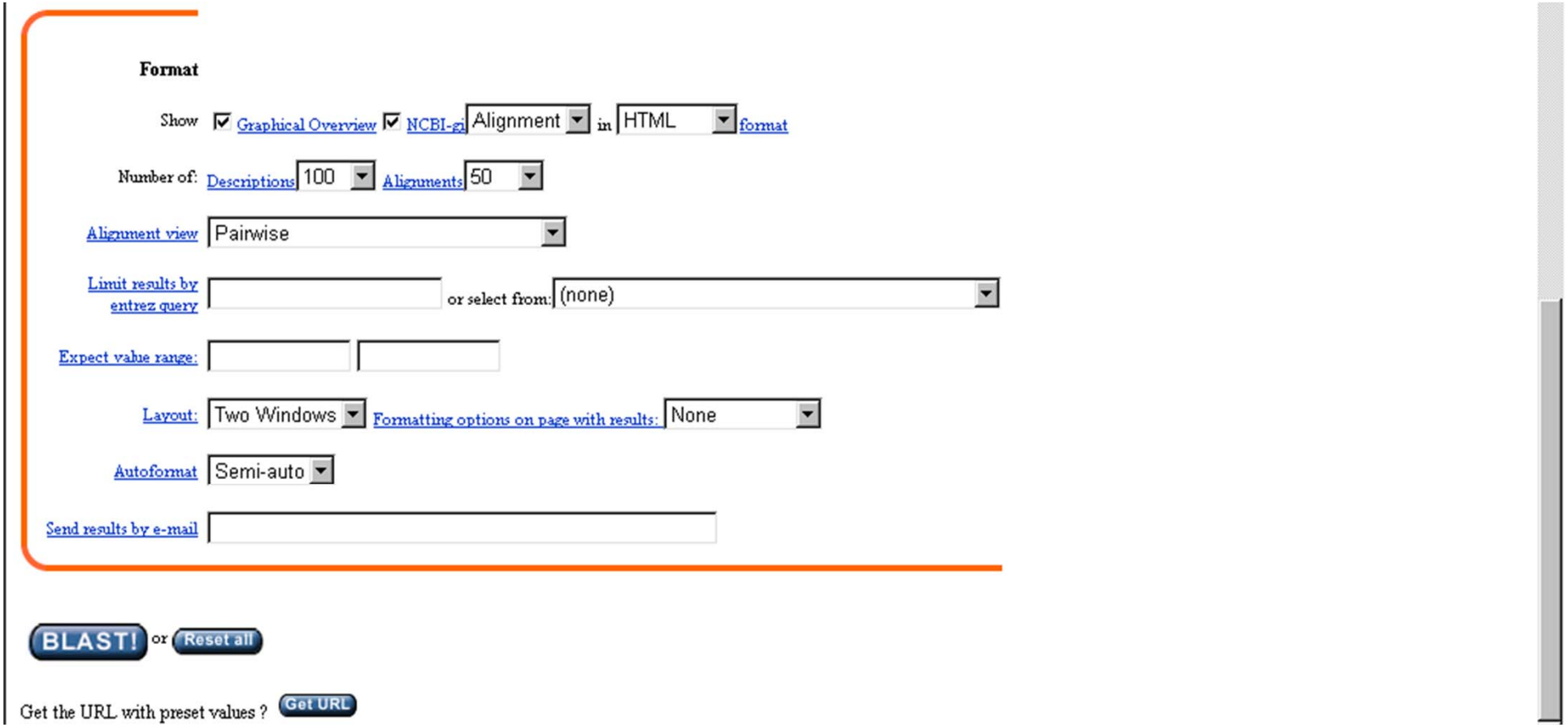
Pri poravnavah takšne regije označi z X.

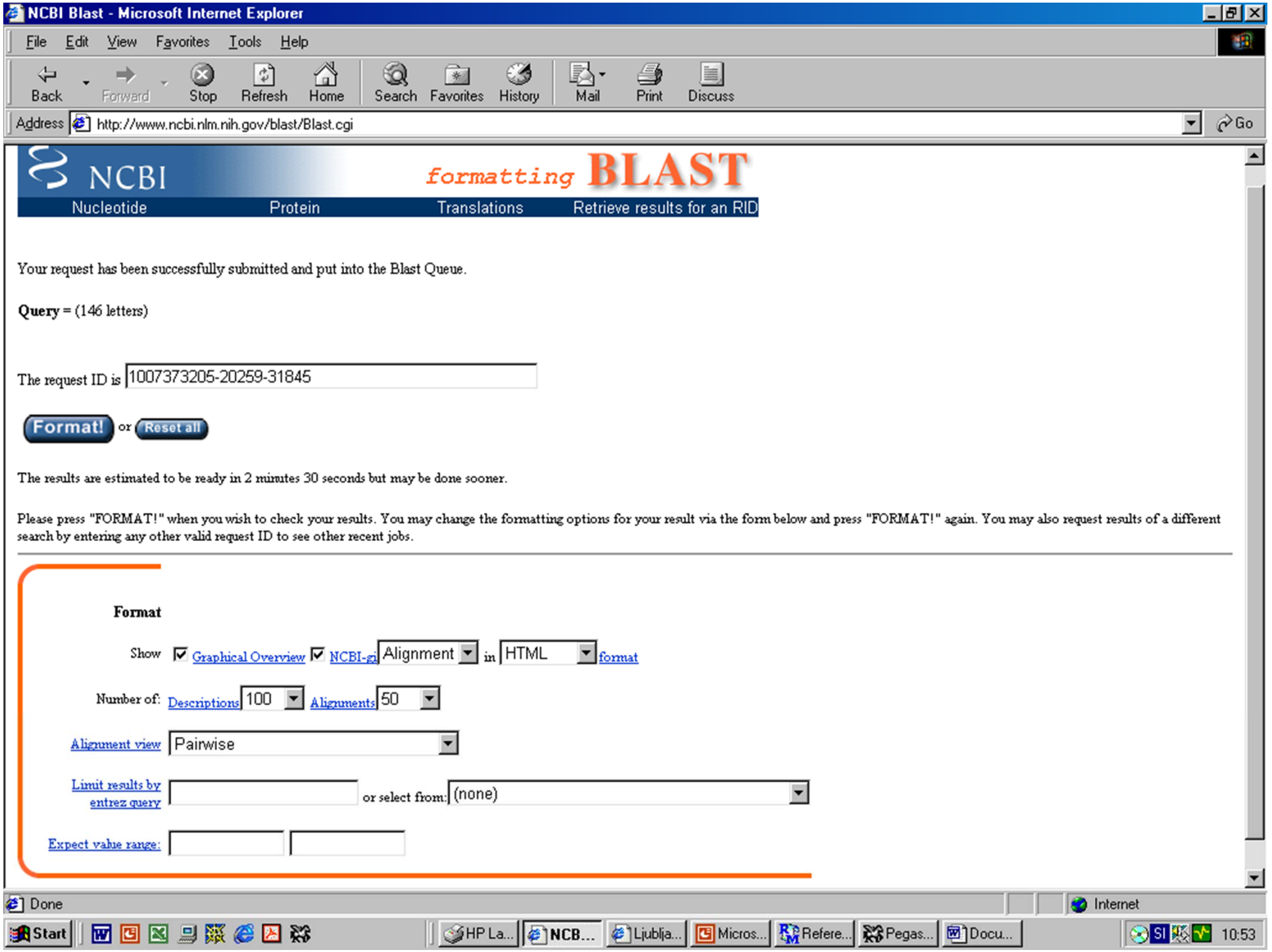
SEG program, ki je del BLAST-a

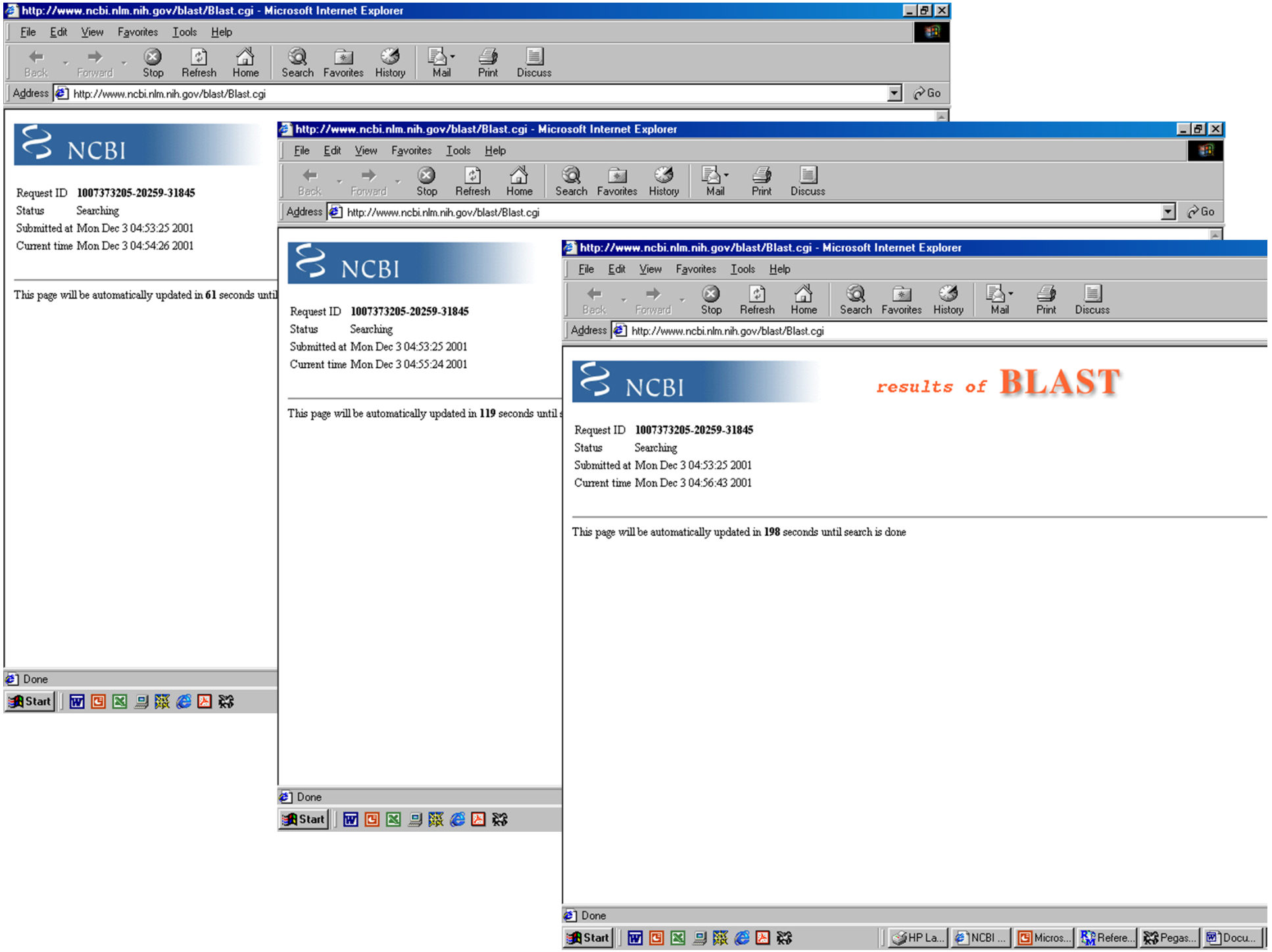
Wootton JC (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. Computational Chemistry 18, 269-285.

```
>SERA_PLAFG (P13823)
MKSYISLFFILCVIFNKNVIKCTGESQTGNTGGGQAGNTVGDQAGSTGGSPQGSTGASQPGSSEPSNPVSSGHSVSTVSVSQTSTSSEKQDTI
QVKSALLKDYMGLKVTGPCNENFIMFLVPHIYIDVDTEDTNIELRTTLKETNNAISFESNSGSLEKKKYVKLPSNGTTGEQGSSTGTVRGDTE
PISDSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSESLPANGPDSPTVKPPRNLQNICETGKNFKLVVYIKENTLIIKWKVYGETKDTT
ENNKVDVRKYLINEKETPFTSILIHAYKEHNGTNLIESKNYALGSDIPEKCDTLASNCFLSGNFNIEKCFQCALLVEKENKNDVCYKYLSEDI
VSNFKEIKAETEDDDEDDYTEYKLTESIDNILVKMFKTNENNDKSELIKLEEVDDSLKLELMNYCSLLKDVDTTGTLDNYGMGNEMDIFNNLK
```




Format
Show
Graphical Overview
NCBI-gi
Alignment
in
HTML
format
Number of:
Descriptions
100
Alignments
50
Alignment view
Pairwise
Limit results by entrez query
or select from:
(none)
Expect value range:
Layout:
Two Windows
Formatting options on page with results:
None
Autoformat
Semi-auto
Send results by e-mail
BLAST!
or
Reset all
Get the URL with preset values ?
Get URL


NCBI Blast - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Discuss
Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Go
NCBI
formatting BLAST
Nucleotide Protein Translations Retrieve results for an RID
Your request has been successfully submitted and put into the Blast Queue.
Query = (146 letters)
The request ID is 1007373205-20259-31845
Format! or Reset all
The results are estimated to be ready in 2 minutes 30 seconds but may be done sooner.
Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.
Format
Show Graphical Overview NCBI-gi Alignment in HTML format
Number of: Descriptions 100 Alignments 50
Alignment view Pairwise
Limit results by entrez query or select from: (none)
Expect value range:
Done
Internet
Start
HP La... NCB... Ljublja... Micros... Refere... Pegas... Docu...
SI 10:53


http://www.ncbi.nlm.nih.gov/blast/Blast.cgi - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Discuss
Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Go
NCBI
Request ID 1007373205-20259-31845
Status Searching
Submitted at Mon Dec 3 04:53:25 2001
Current time Mon Dec 3 04:54:26 2001
This page will be automatically updated in 61 seconds until
Done
Start
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Discuss
Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Go
NCBI
Request ID 1007373205-20259-31845
Status Searching
Submitted at Mon Dec 3 04:53:25 2001
Current time Mon Dec 3 04:55:24 2001
This page will be automatically updated in 119 seconds until
Done
Start
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Discuss
Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
NCBI
results of BLAST
Request ID 1007373205-20259-31845
Status Searching
Submitted at Mon Dec 3 04:53:25 2001
Current time Mon Dec 3 04:56:43 2001
This page will be automatically updated in 198 seconds until search is done
Done
Start
HP La...
NCBI ...
Micros...
Refere...
Pegas...
Docu...

NCBI

results of BLAST

TBLASTN 2.2.1 [Apr-13-2001]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.

RID: 1007373205-20259-31845

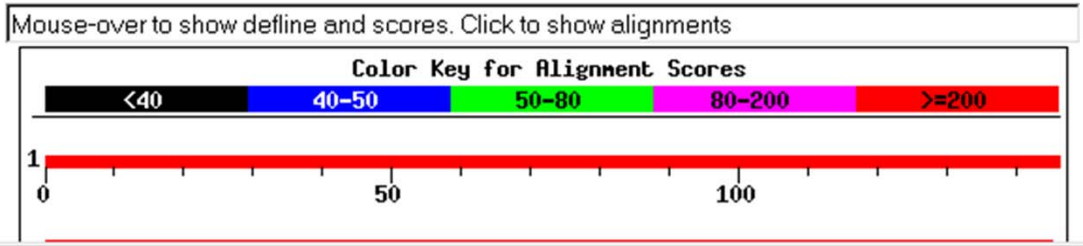
Query=
(146 letters)

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,
GSS, or phase 0, 1 or 2 HTGS sequences).;
1,035,966 sequences; 4,683,384,655 total letters

If you have any problems or questions with the results of this search
please refer to the **BLAST FAQs**

Taxonomy reports

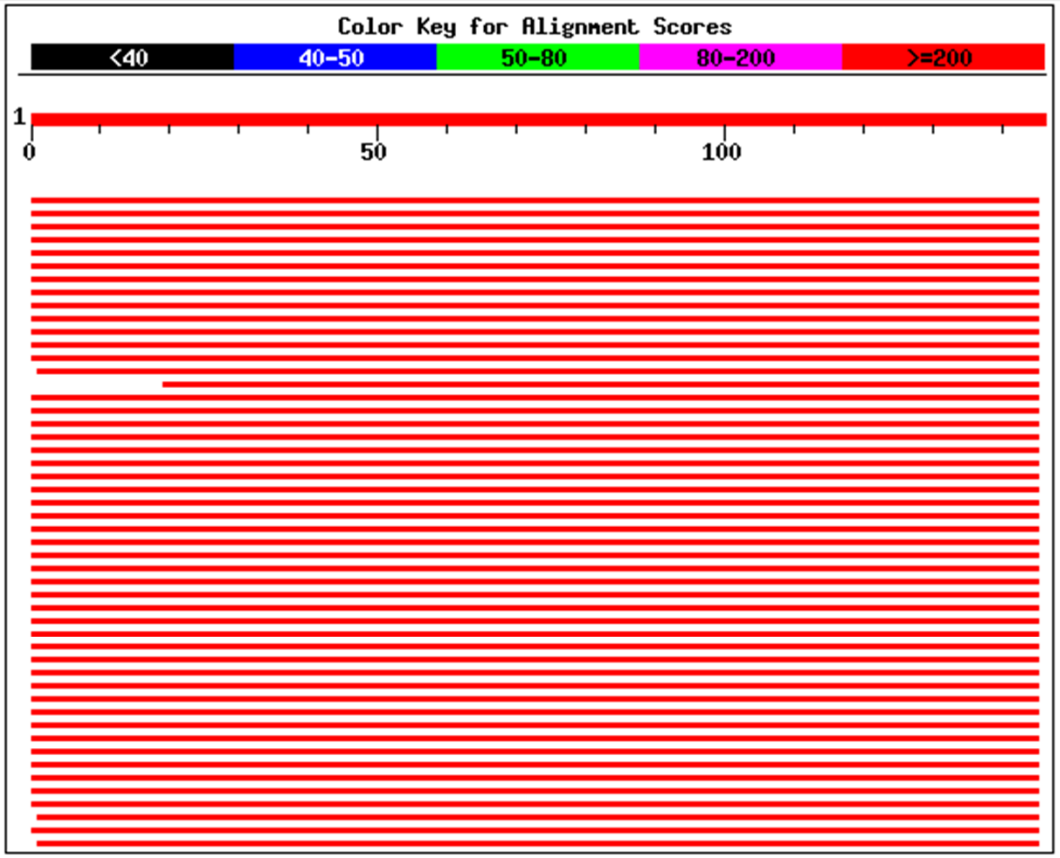
Distribution of 122 Blast Hits on the Query Sequence

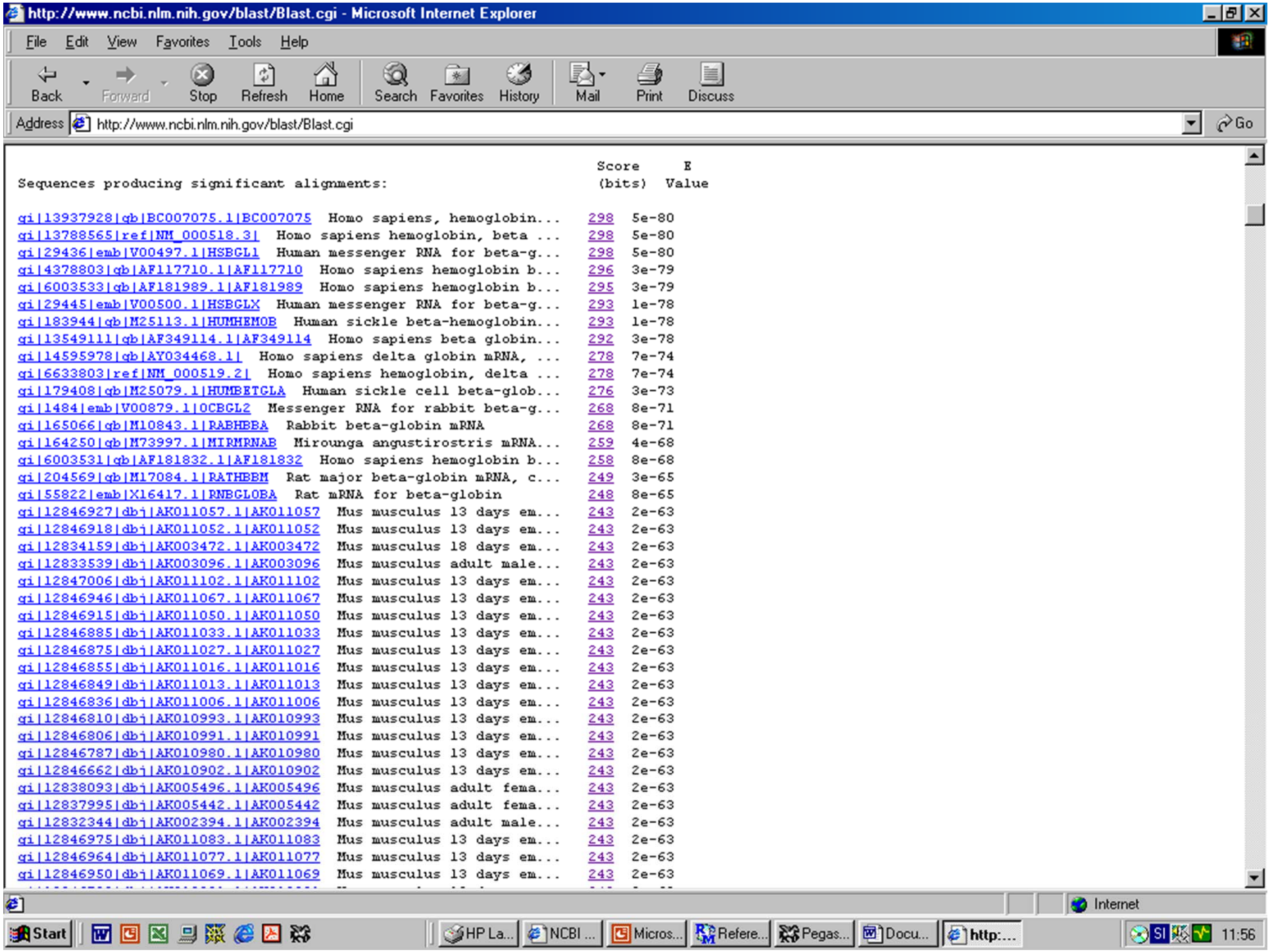


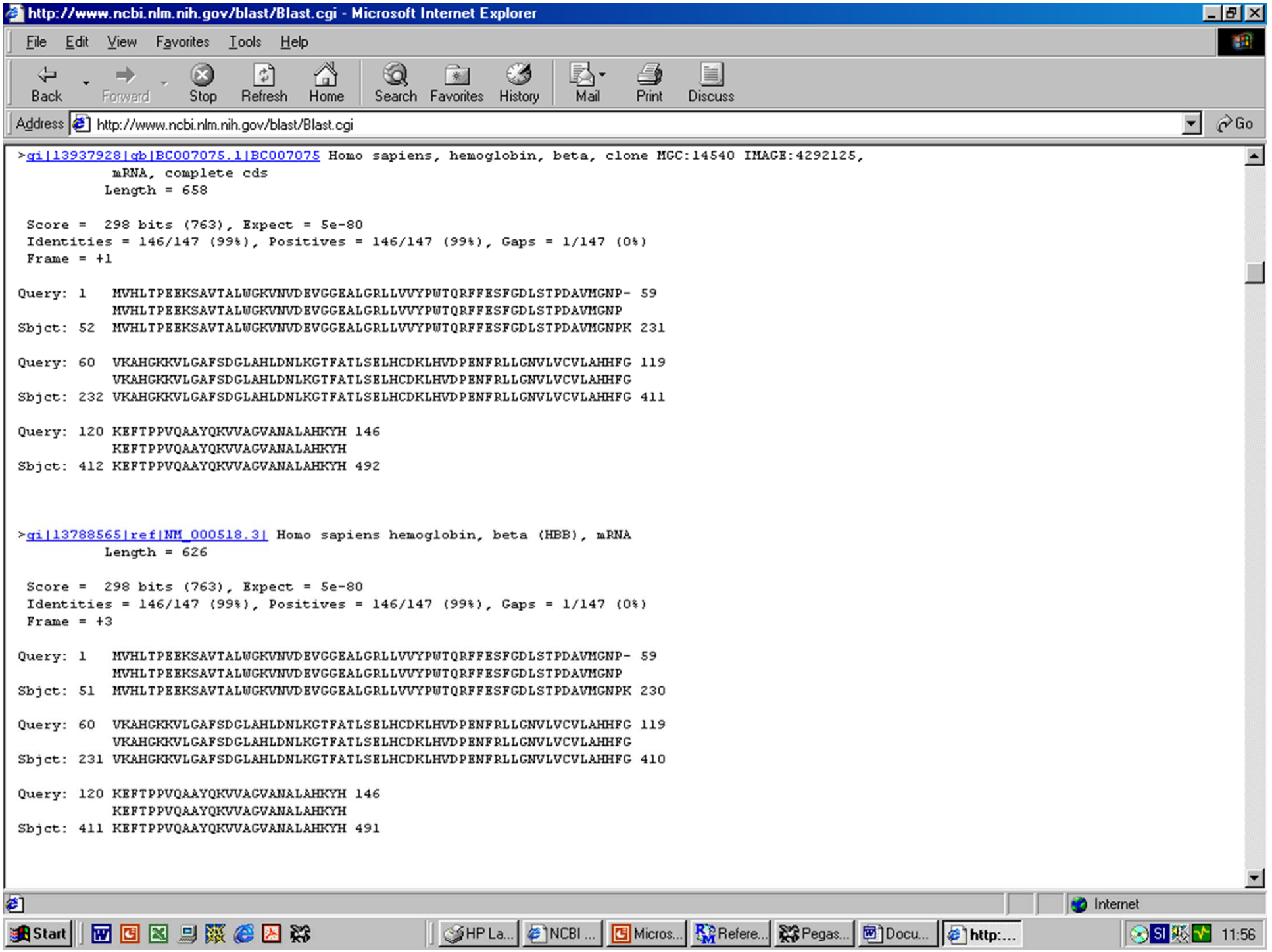
Taxonomy reports

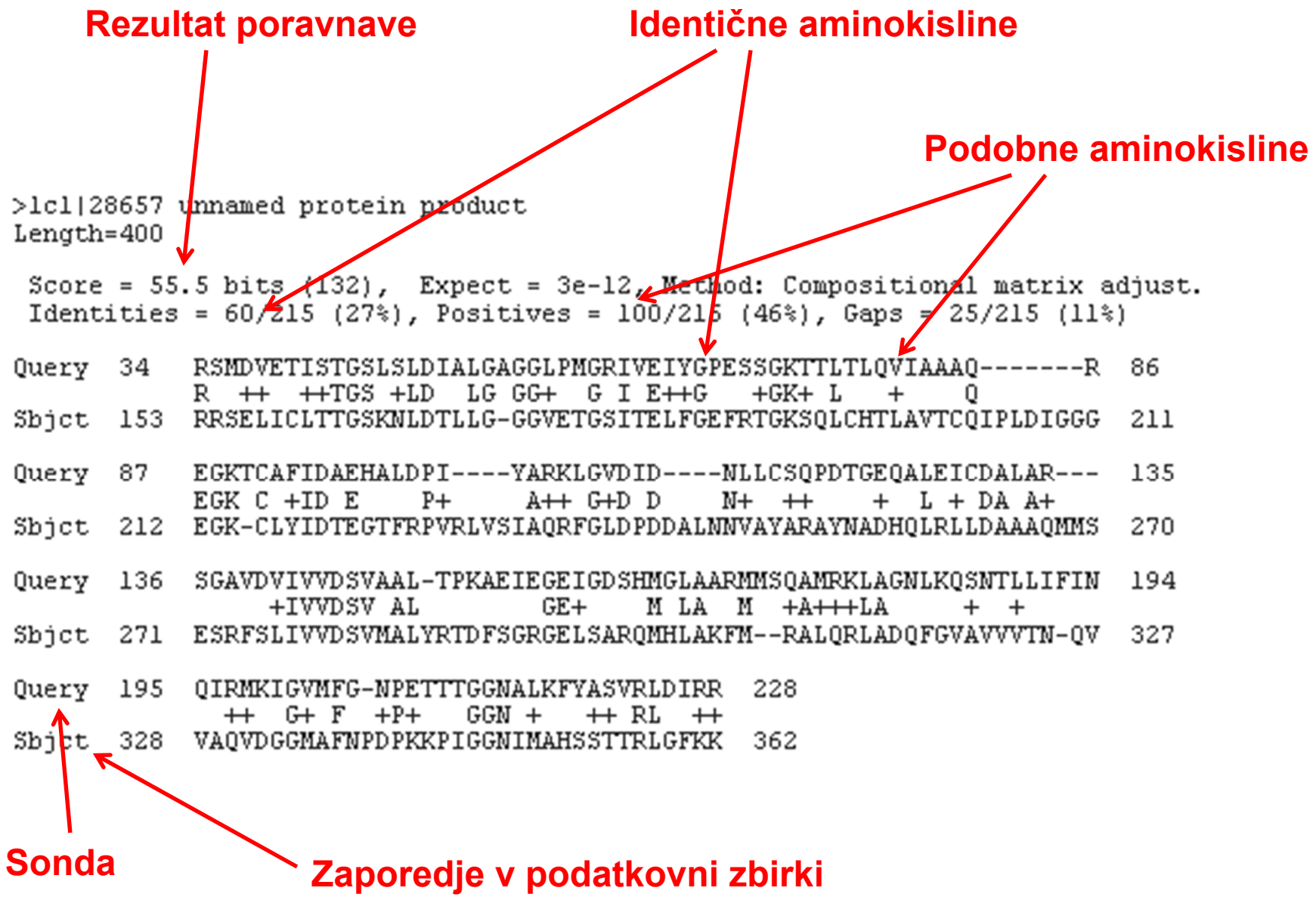
Distribution of 122 Blast Hits on the Query Sequence

M10843 Rabbit beta-globin mRNA..S= 268 E=8e-71




http://www.ncbi.nlm.nih.gov/blast/Blast.cgi - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Discuss
Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Go
Sequences producing significant alignments: Score (bits) E Value
gi|13937928|gb|BC007075.1|BC007075 Homo sapiens, hemoglobin... 298 5e-80
gi|13788565|ref|NM_000518.3| Homo sapiens hemoglobin, beta ... 298 5e-80
gi|29436|emb|V00497.1|HSBGL1 Human messenger RNA for beta-g... 298 5e-80
gi|4378803|gb|AF117710.1|AF117710 Homo sapiens hemoglobin b... 296 3e-79
gi|6003533|gb|AF181989.1|AF181989 Homo sapiens hemoglobin b... 295 3e-79
gi|29445|emb|V00500.1|HSBGLX Human messenger RNA for beta-g... 293 1e-78
gi|183944|gb|M25113.1|HUMHEMOB Human sickle beta-hemoglobin... 293 1e-78
gi|13549111|gb|AF349114.1|AF349114 Homo sapiens beta globin... 292 3e-78
gi|14595978|gb|AY034468.1| Homo sapiens delta globin mRNA, ... 278 7e-74
gi|6633803|ref|NM_000519.2| Homo sapiens hemoglobin, delta ... 278 7e-74
gi|179408|gb|M25079.1|HUMBETGLA Human sickle cell beta-glob... 276 3e-73
gi|1484|emb|V00879.1|OCBGL2 Messenger RNA for rabbit beta-g... 268 8e-71
gi|165066|gb|M10843.1|RABHBBA Rabbit beta-globin mRNA 268 8e-71
gi|164250|gb|M73997.1|MIRMRNAB Mirounga angustirostris mRNA... 259 4e-68
gi|6003531|gb|AF181832.1|AF181832 Homo sapiens hemoglobin b... 258 8e-68
gi|204569|gb|M17084.1|RATHBBM Rat major beta-globin mRNA, c... 249 3e-65
gi|55822|emb|X16417.1|RNBGLOBA Rat mRNA for beta-globin 248 8e-65
gi|12846927|dbj|AK011057.1|AK011057 Mus musculus 13 days em... 243 2e-63
gi|12846918|dbj|AK011052.1|AK011052 Mus musculus 13 days em... 243 2e-63
gi|12834159|dbj|AK003472.1|AK003472 Mus musculus 18 days em... 243 2e-63
gi|12833539|dbj|AK003096.1|AK003096 Mus musculus adult male... 243 2e-63
gi|12847006|dbj|AK011102.1|AK011102 Mus musculus 13 days em... 243 2e-63
gi|12846946|dbj|AK011067.1|AK011067 Mus musculus 13 days em... 243 2e-63
gi|12846915|dbj|AK011050.1|AK011050 Mus musculus 13 days em... 243 2e-63
gi|12846885|dbj|AK011033.1|AK011033 Mus musculus 13 days em... 243 2e-63
gi|12846875|dbj|AK011027.1|AK011027 Mus musculus 13 days em... 243 2e-63
gi|12846855|dbj|AK011016.1|AK011016 Mus musculus 13 days em... 243 2e-63
gi|12846849|dbj|AK011013.1|AK011013 Mus musculus 13 days em... 243 2e-63
gi|12846836|dbj|AK011006.1|AK011006 Mus musculus 13 days em... 243 2e-63
gi|12846810|dbj|AK010993.1|AK010993 Mus musculus 13 days em... 243 2e-63
gi|12846806|dbj|AK010991.1|AK010991 Mus musculus 13 days em... 243 2e-63
gi|12846787|dbj|AK010980.1|AK010980 Mus musculus 13 days em... 243 2e-63
gi|12846662|dbj|AK010902.1|AK010902 Mus musculus 13 days em... 243 2e-63
gi|12838093|dbj|AK005496.1|AK005496 Mus musculus adult fema... 243 2e-63
gi|12837995|dbj|AK005442.1|AK005442 Mus musculus adult fema... 243 2e-63
gi|12832344|dbj|AK002394.1|AK002394 Mus musculus adult male... 243 2e-63
gi|12846975|dbj|AK011083.1|AK011083 Mus musculus 13 days em... 243 2e-63
gi|12846964|dbj|AK011077.1|AK011077 Mus musculus 13 days em... 243 2e-63
gi|12846950|dbj|AK011069.1|AK011069 Mus musculus 13 days em... 243 2e-63
Internet
Start
HP La... NCBI ... Micros... Refere... Pegas... Docu... http:...
11:56


http://www.ncbi.nlm.nih.gov/blast/Blast.cgi - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Discuss
Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Go
>gi|13937928|qb|BC007075.1|BC007075 Homo sapiens, hemoglobin, beta, clone MGC:14540 IMAGE:4292125,
mRNA, complete cds
Length = 658
Score = 298 bits (763), Expect = 5e-80
Identities = 146/147 (99%), Positives = 146/147 (99%), Gaps = 1/147 (0%)
Frame = +1
Query: 1 MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP- 59
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP
Sbjct: 52 MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK 231
Query: 60 VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG 119
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
Sbjct: 232 VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG 411
Query: 120 KEFTPPVQAAYQKVVAGVANALAHKYH 146
KEFTPPVQAAYQKVVAGVANALAHKYH
Sbjct: 412 KEFTPPVQAAYQKVVAGVANALAHKYH 492
>gi|13788565|ref|NM_000518.3| Homo sapiens hemoglobin, beta (HBB), mRNA
Length = 626
Score = 298 bits (763), Expect = 5e-80
Identities = 146/147 (99%), Positives = 146/147 (99%), Gaps = 1/147 (0%)
Frame = +3
Query: 1 MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP- 59
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP
Sbjct: 51 MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK 230
Query: 60 VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG 119
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
Sbjct: 231 VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG 410
Query: 120 KEFTPPVQAAYQKVVAGVANALAHKYH 146
KEFTPPVQAAYQKVVAGVANALAHKYH
Sbjct: 411 KEFTPPVQAAYQKVVAGVANALAHKYH 491
Internet
Start
HP La... NCBI ... Micros... Refere... Pegas... Docu... http:...
11:56


Rezultat poravnave
Identične aminokisline
Podobne aminokisline
>lcl|28657 unnamed protein product
Length=400
Score = 55.5 bits (132), Expect = 3e-12, Method: Compositional matrix adjust.
Identities = 60/215 (27%), Positives = 100/215 (46%), Gaps = 25/215 (11%)
Query 34 RSMDVETISTGSLSLDIALGAGGLPMGRIVEIYGPESSGKTTLTLQVIAAAQ-------R 86
R ++ ++TGS +LD LG GG+ G I E++G +GK+ L + Q
Sbjct 153 RRSELICLTTGSKNLDTLLG-GGVETGSITELFGEFRTGKSQLCHTLAVTCQIPLDIGGG 211
Query 87 EGKTCAFIDAEHALDPI----YARKLGVDID----NLLCSQPDTGEQALEICDALAR--- 135
EGK C +ID E P+ A++ G+D D N+ ++ + L + DA A+
Sbjct 212 EGK-CLYIDTEGTFRPVRLVSIAQRFGLDPDDALNNVAYARAYNADHQLRLLDAAAQMMS 270
Query 136 SGAVDVIVVDSVAAL-TPKAEIEGEIGDSHMGLAARMMSQAMRKLAGNLKQSNTLLIFIN 194
+IVVDSV AL GE+ M LA M +A+++LA + +
Sbjct 271 ESRFSLIVVDSVMALYRTDFSGRGELSARQMHLAKFM--RALQRLADQFGVAVVVTN-QV 327
Query 195 QIRMKIGVMFG-NPETTTGGNALKFYASVRLDIRR 228
++ G+ F +P+ GGN + ++ RL ++
Sbjct 328 VAQVDGGMAFNPDPKKPIGGNIMAHSSTTRLGFKK 362
Sonda
Zaporedje v podatkovni zbirki

FASTA ALGORITEM

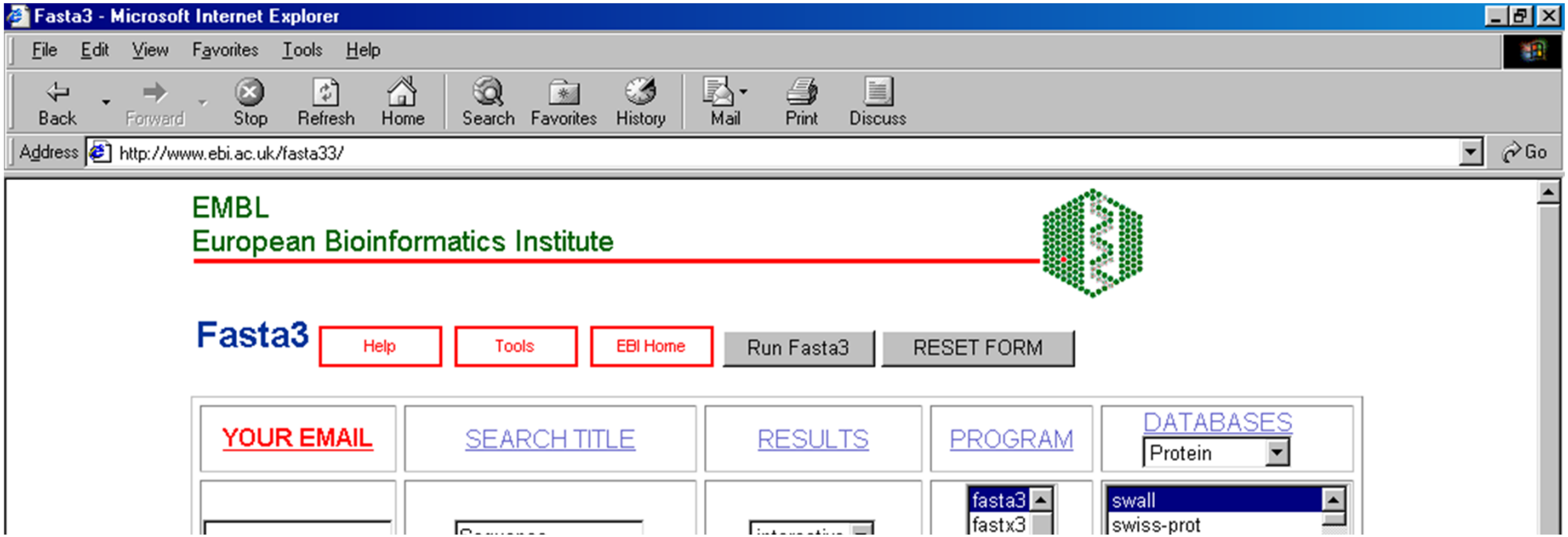
Lipman DJ and Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227, 1435-1441.

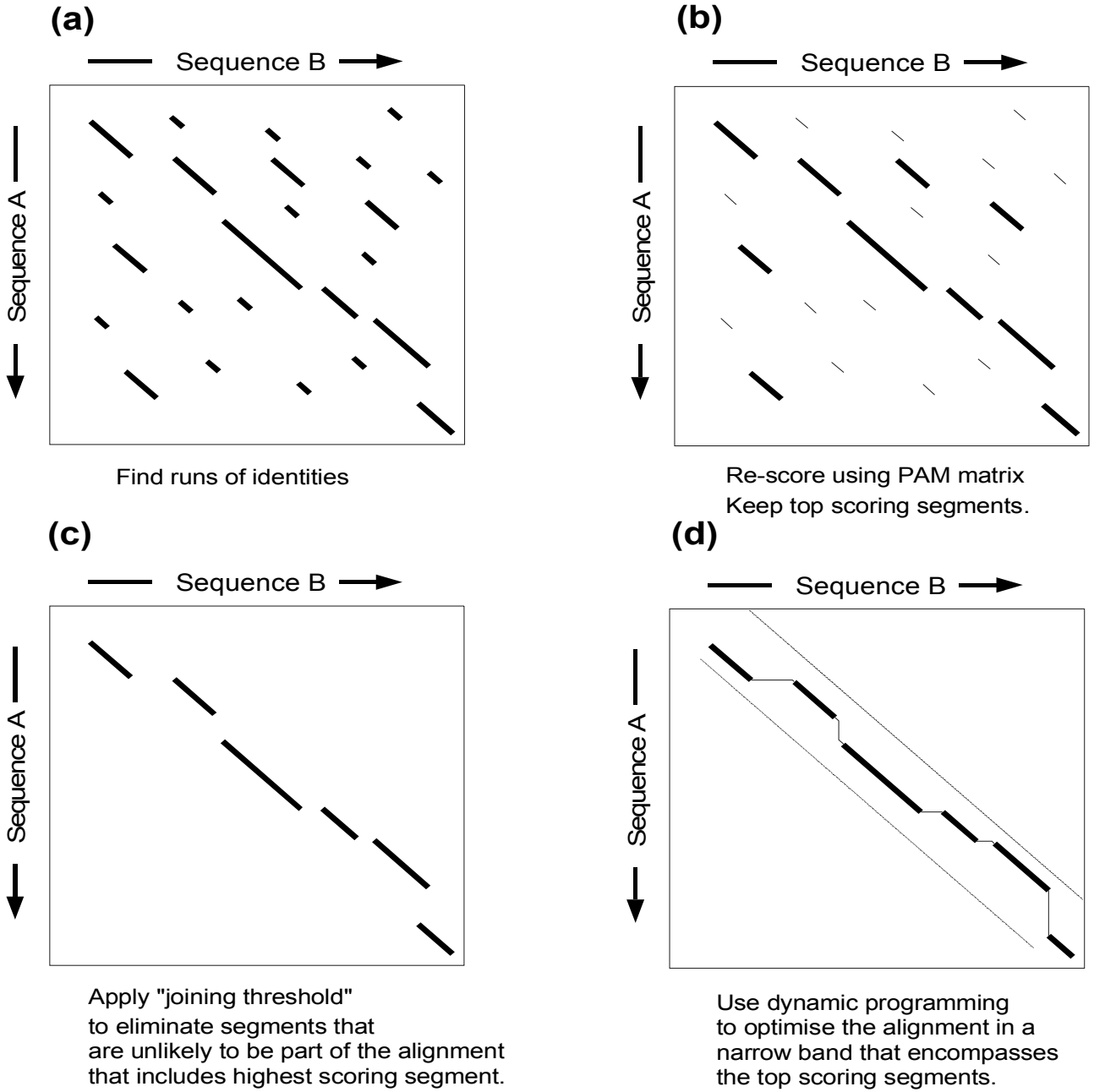
Zaporedja razbije na krajše dele, besede ali k-tuple (parameter, ki se ga lahko poljubno nastavi):

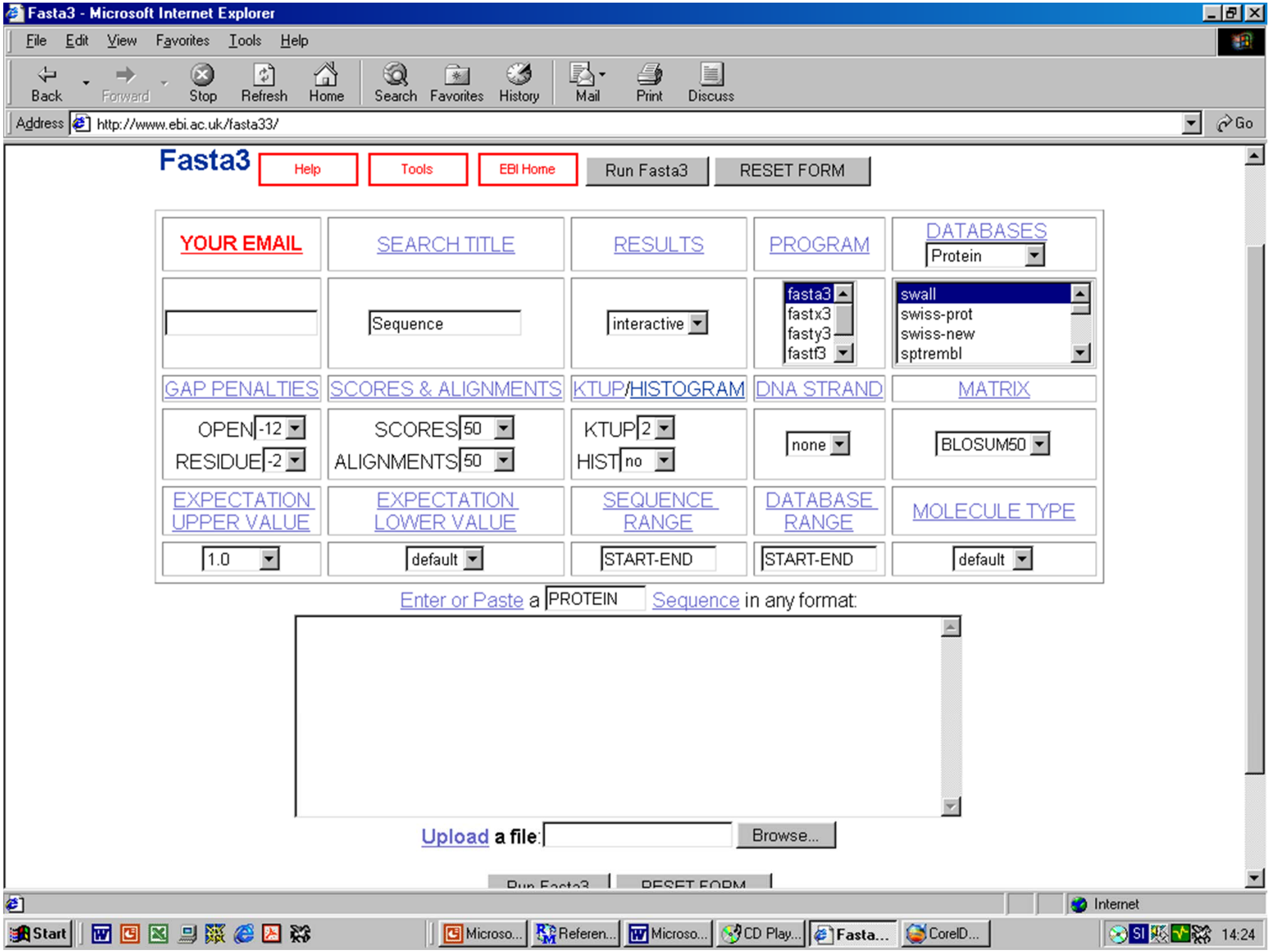
k-tuple 1 počasno iskanje, bolj občutljivo

k-tuple 2 hitrejše iskanje, manj občutljivo

FASTA na internetu:
http://www.ebi.ac.uk/fasta33/




(a)
Sequence B
Sequence A
Find runs of identities
(b)
Sequence B
Sequence A
Re-score using PAM matrix
Keep top scoring segments.
(c)
Sequence B
Sequence A
Apply "joining threshold"
to eliminate segments that
are unlikely to be part of the alignment
that includes highest scoring segment.
(d)
Sequence B
Sequence A
Use dynamic programming
to optimise the alignment in a
narrow band that encompasses
the top scoring segments.


Fasta3 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Discuss
Address http://www.ebi.ac.uk/fasta33/
Go
Fasta3
Help
Tools
EBI Home
Run Fasta3
RESET FORM
YOUR EMAIL
SEARCH TITLE
RESULTS
PROGRAM
DATABASES
Protein
Sequence
interactive
fasta3
fastx3
fasty3
fastf3
swall
swiss-prot
swiss-new
sptrembl
GAP PENALTIES
SCORES & ALIGNMENTS
KTUP/HISTOGRAM
DNA STRAND
MATRIX
OPEN -12
RESIDUE -2
SCORES 50
ALIGNMENTS 50
KTUP 2
HIST no
none
BLOSUM50
EXPECTATION UPPER VALUE
EXPECTATION LOWER VALUE
SEQUENCE RANGE
DATABASE RANGE
MOLECULE TYPE
1.0
default
START-END
START-END
default
Enter or Paste a PROTEIN Sequence in any format:
Upload a file:
Browse...
Run Fasta3
RESET FORM
Internet
Start
Microso...
Referen...
Microso...
CD Play...
Fasta...
CorelD...
14:24

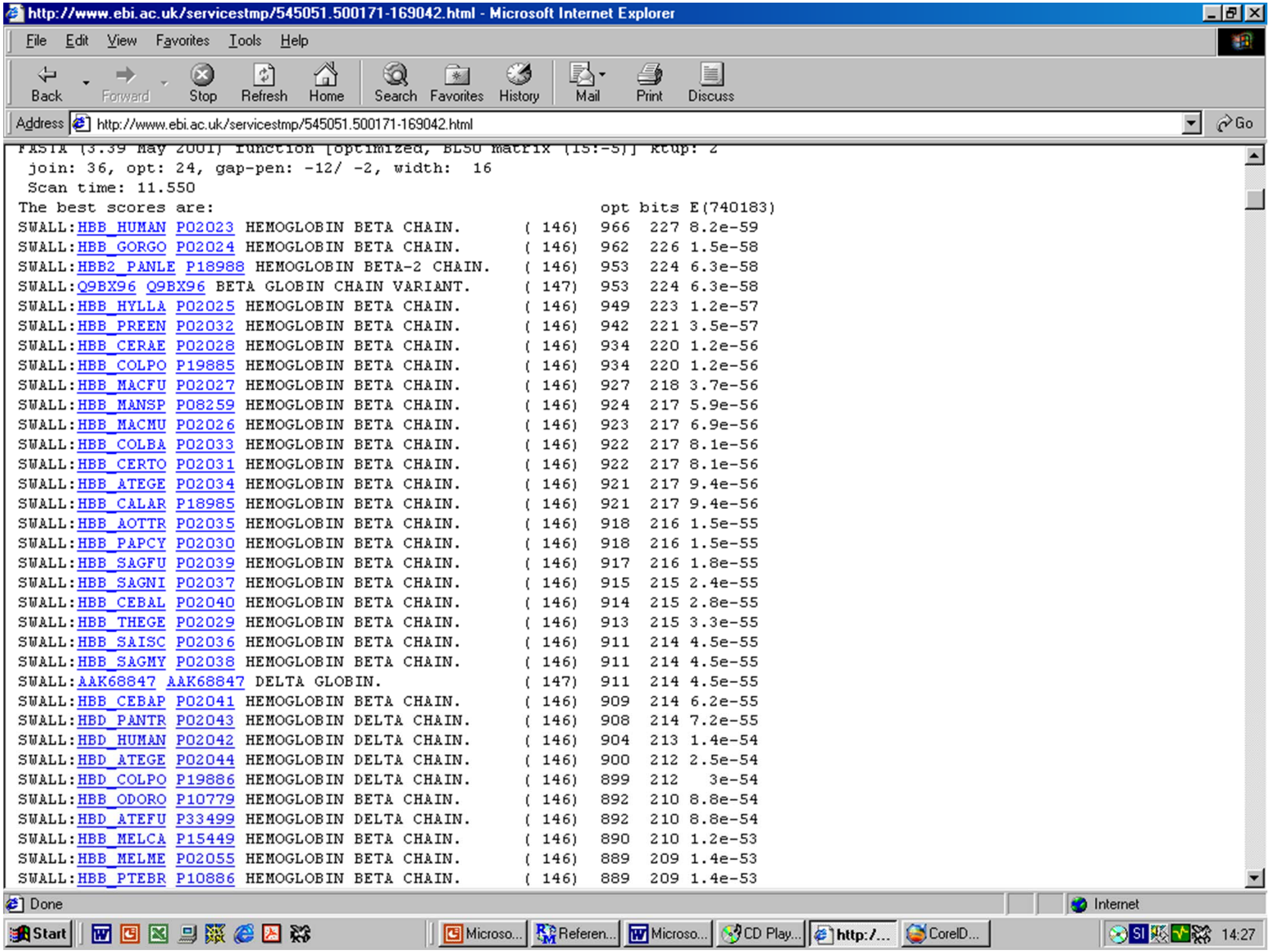
EMBL
European Bioinformatics Institute

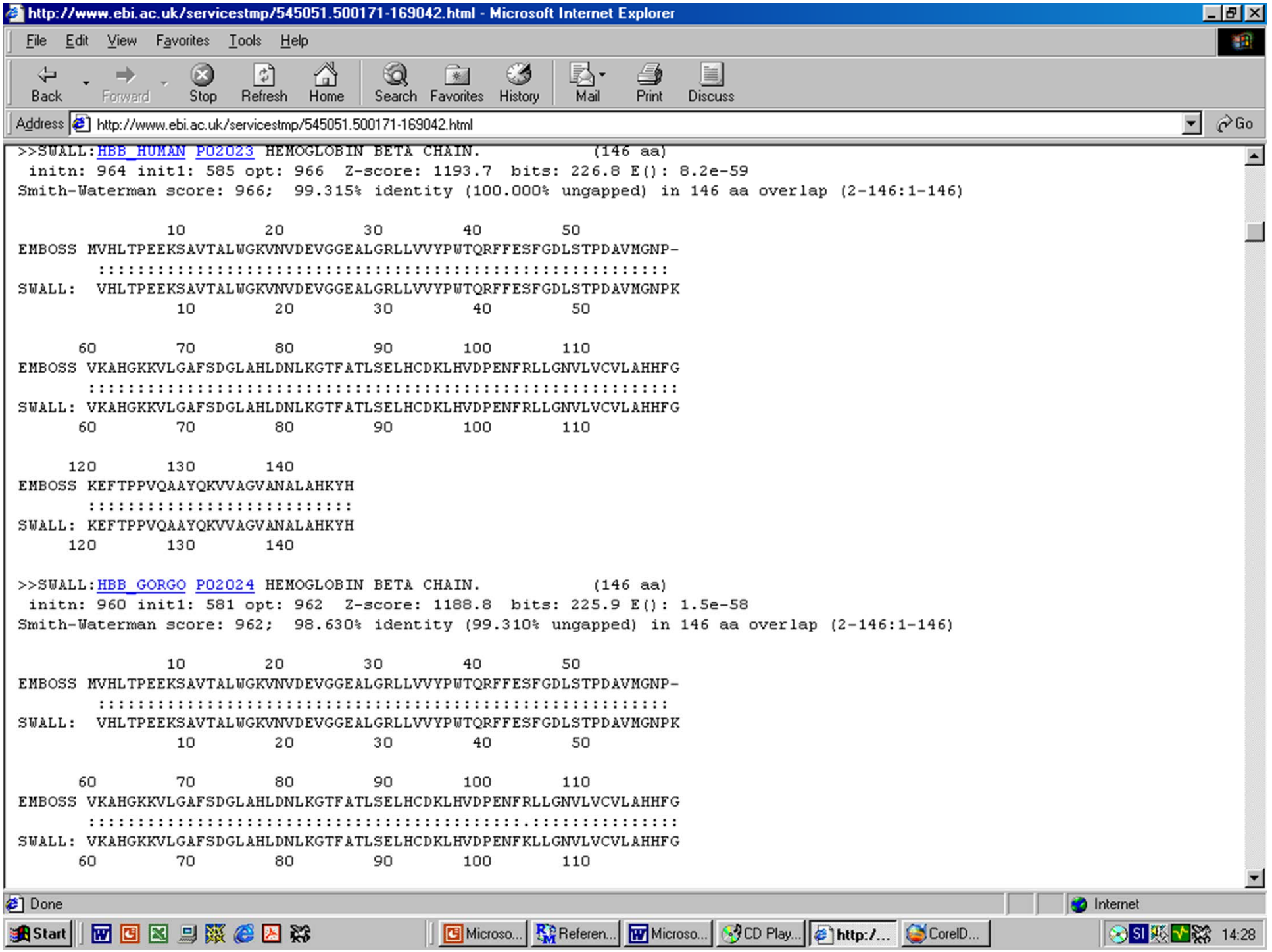
Results of Search:

Program: fasta33_t
Database: +swall+
Title: Sequence
SeqLen: 146

View using Mview | VisualFasta | SUBMIT ANOTHER JOB

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

@:1-: 146 aa
 EMBOSS_001
 vs  SWISS-PROT All library
searching /ebi/services/idata/fastadb/swall library

235139372 residues in 741331 sequences
 statistics extrapolated from 60000 to 740183 sequences
  Expectation_n fit: rho(ln(x))= 5.6371+/-0.000183; mu= 1.3266+/- 0.010
 mean var=67.0586+/-13.734, 0's: 136 Z-trim: 648  B-trim: 1914 in 1/64
```


http://www.ebi.ac.uk/servicestmp/545051.500171-169042.html - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Discuss
Address http://www.ebi.ac.uk/servicestmp/545051.500171-169042.html
Go
FASTA (3.39 May 2001) function [optimized, BL50 matrix (15:-5)] ktup: 2
join: 36, opt: 24, gap-pen: -12/ -2, width: 16
Scan time: 11.550
The best scores are: opt bits E(740183)
SWALL:HBB_HUMAN P02023 HEMOGLOBIN BETA CHAIN. (146) 966 227 8.2e-59
SWALL:HBB_GORGO P02024 HEMOGLOBIN BETA CHAIN. (146) 962 226 1.5e-58
SWALL:HBB2_PANLE P18988 HEMOGLOBIN BETA-2 CHAIN. (146) 953 224 6.3e-58
SWALL:Q9BX96 Q9BX96 BETA GLOBIN CHAIN VARIANT. (147) 953 224 6.3e-58
SWALL:HBB_HYLLA P02025 HEMOGLOBIN BETA CHAIN. (146) 949 223 1.2e-57
SWALL:HBB_PREEN P02032 HEMOGLOBIN BETA CHAIN. (146) 942 221 3.5e-57
SWALL:HBB_CERAE P02028 HEMOGLOBIN BETA CHAIN. (146) 934 220 1.2e-56
SWALL:HBB_COLPO P19885 HEMOGLOBIN BETA CHAIN. (146) 934 220 1.2e-56
SWALL:HBB_MACFU P02027 HEMOGLOBIN BETA CHAIN. (146) 927 218 3.7e-56
SWALL:HBB_MANSP P08259 HEMOGLOBIN BETA CHAIN. (146) 924 217 5.9e-56
SWALL:HBB_MACMU P02026 HEMOGLOBIN BETA CHAIN. (146) 923 217 6.9e-56
SWALL:HBB_COLBA P02033 HEMOGLOBIN BETA CHAIN. (146) 922 217 8.1e-56
SWALL:HBB_CERTO P02031 HEMOGLOBIN BETA CHAIN. (146) 922 217 8.1e-56
SWALL:HBB_ATEGE P02034 HEMOGLOBIN BETA CHAIN. (146) 921 217 9.4e-56
SWALL:HBB_CALAR P18985 HEMOGLOBIN BETA CHAIN. (146) 921 217 9.4e-56
SWALL:HBB_AOTTR P02035 HEMOGLOBIN BETA CHAIN. (146) 918 216 1.5e-55
SWALL:HBB_PAPCY P02030 HEMOGLOBIN BETA CHAIN. (146) 918 216 1.5e-55
SWALL:HBB_SAGFU P02039 HEMOGLOBIN BETA CHAIN. (146) 917 216 1.8e-55
SWALL:HBB_SAGNI P02037 HEMOGLOBIN BETA CHAIN. (146) 915 215 2.4e-55
SWALL:HBB_CEBAL P02040 HEMOGLOBIN BETA CHAIN. (146) 914 215 2.8e-55
SWALL:HBB_THEGE P02029 HEMOGLOBIN BETA CHAIN. (146) 913 215 3.3e-55
SWALL:HBB_SAISC P02036 HEMOGLOBIN BETA CHAIN. (146) 911 214 4.5e-55
SWALL:HBB_SAGMY P02038 HEMOGLOBIN BETA CHAIN. (146) 911 214 4.5e-55
SWALL:AAK68847 AAK68847 DELTA GLOBIN. (147) 911 214 4.5e-55
SWALL:HBB_CEBAP P02041 HEMOGLOBIN BETA CHAIN. (146) 909 214 6.2e-55
SWALL:HBD_PANTR P02043 HEMOGLOBIN DELTA CHAIN. (146) 908 214 7.2e-55
SWALL:HBD_HUMAN P02042 HEMOGLOBIN DELTA CHAIN. (146) 904 213 1.4e-54
SWALL:HBD_ATEGE P02044 HEMOGLOBIN DELTA CHAIN. (146) 900 212 2.5e-54
SWALL:HBD_COLPO P19886 HEMOGLOBIN DELTA CHAIN. (146) 899 212 3e-54
SWALL:HBB_ODORO P10779 HEMOGLOBIN BETA CHAIN. (146) 892 210 8.8e-54
SWALL:HBD_ATEFU P33499 HEMOGLOBIN DELTA CHAIN. (146) 892 210 8.8e-54
SWALL:HBB_MELCA P15449 HEMOGLOBIN BETA CHAIN. (146) 890 210 1.2e-53
SWALL:HBB_MELME P02055 HEMOGLOBIN BETA CHAIN. (146) 889 209 1.4e-53
SWALL:HBB_PTEBR P10886 HEMOGLOBIN BETA CHAIN. (146) 889 209 1.4e-53
Done
Internet
Start
Microso... Referen... Microso... CD Play... http:/... CorelD...
14:27


http://www.ebi.ac.uk/servicestmp/545051.500171-169042.html - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Discuss
Address http://www.ebi.ac.uk/servicestmp/545051.500171-169042.html
Go
>>SWALL:HBB_HUMAN P02023 HEMOGLOBIN BETA CHAIN. (146 aa)
initn: 964 init1: 585 opt: 966 Z-score: 1193.7 bits: 226.8 E(): 8.2e-59
Smith-Waterman score: 966; 99.315% identity (100.000% ungapped) in 146 aa overlap (2-146:1-146)
10 20 30 40 50
EMBOSS MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP-
SWALL: VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
10 20 30 40 50
60 70 80 90 100 110
EMBOSS VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
SWALL: VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
60 70 80 90 100 110
120 130 140
EMBOSS KEFTPPVQAAYQKVVAGVANALAHKYH
SWALL: KEFTPPVQAAYQKVVAGVANALAHKYH
120 130 140
>>SWALL:HBB_GORGO P02024 HEMOGLOBIN BETA CHAIN. (146 aa)
initn: 960 init1: 581 opt: 962 Z-score: 1188.8 bits: 225.9 E(): 1.5e-58
Smith-Waterman score: 962; 98.630% identity (99.310% ungapped) in 146 aa overlap (2-146:1-146)
10 20 30 40 50
EMBOSS MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP-
SWALL: VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
10 20 30 40 50
60 70 80 90 100 110
EMBOSS VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
SWALL: VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFKLLGNVLVCVLAHHFG
60 70 80 90 100 110
Done
Internet
Start
Microso...
Referen...
Microso...
CD Play...
http:/...
CorelD...
14:28

INTERPRETACIJA REZULTATOV

Kako opišemo poravnavo?
Kako vemo kako kvalitetna je naša poravnava?

Procent identičnosti
Kakšen delež aminokislin ali nukleotidov je v poravnavi identičnih.

Procent podobnosti
Poveš še koliko je aminokislin, ki so si podobne po fizikalno-kemijskih lastnostih.

Rezultat poravnave *(Raw scores)*
Rezultat poravnave glede na matriko zamenjav in težo vrzeli, ki si jo določil.

Z-vrednosti *(Z-values)*
Število standardnih deviacij od povprečja. Oceni se za vrednost poravnave iz poravnav mnogo naključnih zaporedij enake dolžine kot zaporedji, ki ju proučujemo.

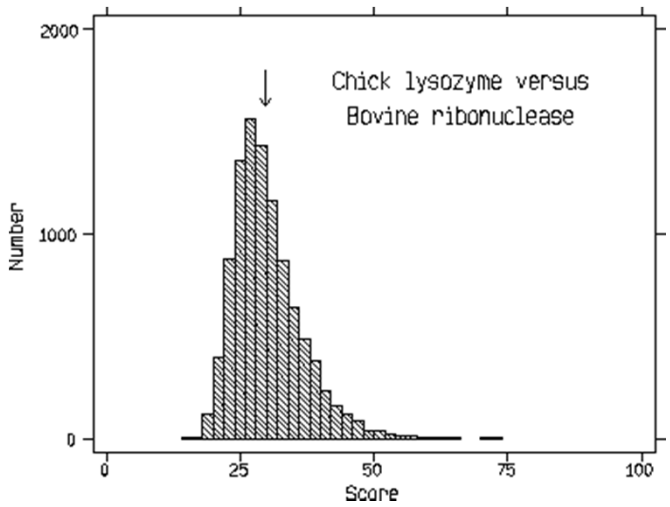
PRIMER: poravnave dveh homolognih in dveh nehomolognih proteinov

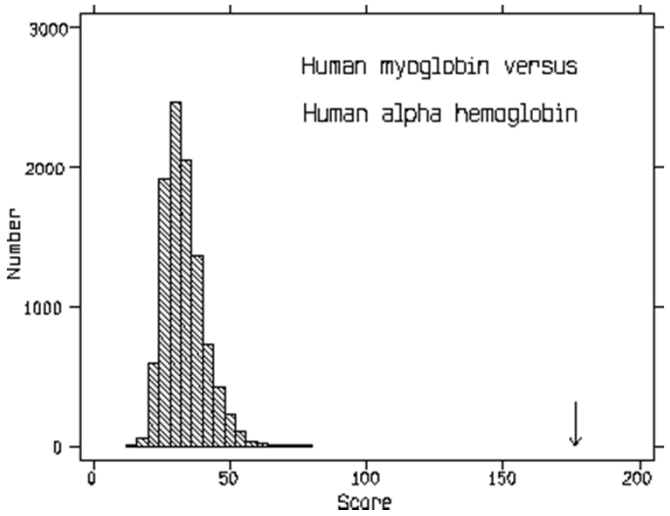
```
Human alpha haemoglobin (141 aa) vs. Human myoglobin (153 aa)

VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSAQ
 :: ..   : ..:::::.:.  ..:.:.: :.: . :.: . : .: .:.     ..:..
GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASED

VKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLP
.: :: .: .::.. . . ..  .....:.. :: : ..    ....:.:.. .:... :
LKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHP

AEFTPAVHASLDKFLASVSTVLTSKYR------
..:.........: :.  .. ..:.:.
GDFGADAQGAMNKALELFRKDMASNYKELGFQG


Chicken lysozyme (129 aa) vs. Bovine ribonuclease (124 aa)

KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINS
:  .    :: ..:. .:.  . . ..  :.....:.  :..  . ... .. .. ....
KETA----AAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHESLADVQA

RWWCNDGRTP--GSRNLCNIPCSALLSSDITASVNCAKKIVSDGDGMNAWVAWRNRCKGT
   :.. ...  .... :  ..:..  .:  .. ...:   .. .. .:      :.:.
V--CSQKNVACKNGQTNCYQSYSTMSITDCRET-GSSKYPNCAYKTTQANKHIIVACEGN

DVQAWIRGCRL
   .   . ..
PYVPVHFDASV
```

Vzameš eno zaporedje in :
1.Naključno pomešaš aminokisline v enem zaporedju
2.Ga poravnaš z drugim zaporedjem
3.Določiš rezultat
4.Ponavljaš

Na koncu dobiš distribucijo rezultatov zaporedij, ki imajo podobno kemično zgradbo. Če je rezultat tvojega proteina veliko večji kot je rezultat večine permutiranih zaporedij potem lahko sklepaš, da sta homologa.

Zlato pravilo: *če je identičnost višja od 25 % potem sta homologa, če je manj kot 15 % potem je homologija zelo dvomljiva. Med 15-25 % moraš imeti trditev da sta homologa podprto s statistično analizo (ki je v splošnem priporočljiva za katerokoli primerjavo zaporedij).*





Vendar:
Takšna analiza bi bila za vsako iskanje zamudna

STATISTIČNA ANALIZA LOKALNIH PORAVNAV

(INTERPRETACIJA REZULTATOV ISKANJA Z BLAST PROGRAMOM)

Karlin S and Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. 87, 2264-2268.
Dembo A et al. (1994) Limit Distribution of maximal non-aligned two-sequence segmental score. Ann. Prob. 22, 2022-2039.

Zelo na kratko in na hitro:

$$E = kmn\, e^{-\lambda S}$$

S	rezultat poravnave
m, n	dolžina zaporedij, ki jih primerjamo
k, λ	parametra odvisna od aminokislinske sestave podatkovne zbirke in ocenjevalne matrike (se jih da oceniti)

Bitne vrednosti (Bit scores)

Preračunani rezultati poravnav upoštevajoč statistične lastnosti (parametra k, λ) ocenjevalnih matrik poravnave. Uporabni, ker lahko primerjamo rezultate različnih iskanj

P-vrednosti

Verjetnost, da se pojavi poravnava z rezultatom, ki je enak ali boljši. Izračunan iz primerjave rezultata poravnave s pričakovano distribucijo HSP iz preimerjav naključnih zaporedij iste dolžine in sestave kot zaporedje s katerim iščemo. Najbolj signifikantne vrednosti bodo blizu 0.

E-vrednosti *(E-values)*

Pričakovano število zaporedij z enakim rezultatom poravnave v enako veliki bazi, ki bi se pojavile naključno

E=10	10 zadetkov- le- ta ni signifikanten. Če ni drugega včasih pomembno, lahko usmeri v nadaljne eksperimente
E<0.01	naključno pojavljanje je zelo redko. Takšen zadetek je signifikanten
E~1E-50	zelo velika verjetnost, da sta obe zaporedji evolucijsko povezani.