



# Sequence Retrieval

Lukas Käll  
lukask@kth.se





## Sequence retrieval from databases

Task: You have a sequence at hand (a query sequence) and want to find its closest sequence homologs in a database. You will need:

- ▶ Speed
- ▶ Accuracy
- ▶ Statistical evaluation



# FASTA

A popular format to store sequence information is the so called FASTA format. It is a textfile format, and specifies that each sequence entry should be begun with a greater than sign (“>”), followed by rows of sequence. The format does not specify any conventions of the formatting of the name or sequence.



# FASTA

A popular format to store sequence information is the so called FASTA format. It is a textfile format, and specifies that each sequence entry should be begun with a greater than sign (“>”), followed by rows of sequence. The format does not specify any conventions of the formatting of the name or sequence.

```
>sp|Q66LE6|2ABD_HUMAN Serine/threonine-protein phosphatase 2A 55 kD
subunit B delta isoform OS=Homo sapiens OX=9606 GN=PPP2R2D PE=1 SV=
MAGAGGGGCPAGGNDFQWCFQSQVKGAIDEDVAEADIISTVEFNYSGLLATGDKGGRVVI
FQREQENKSRPHSRGEYNVYSTFQSHPEFDYLSLEIEEKINKIRWLPQQNAHFLLST
NDKTIKLWKISERDKRAEGYNLKDEDGRLRDPFRITALRVPILKPMDLMVEASPRRIFAN
AHTYHINSISVNSDHETYLSADDLRINLWHLEITDRSFNIVDIKANMEELTEVITAAEF
HPHQCNVVFVSSSKGTIRLCDMRSSALCDRHSKFFEEPEDPSSRSFFSEIISISDVKFS
HSGRYMTRDYLSVKVWDLNMEsrpVETHQVHEYLRSKLSLYENDCIFDKFECCWNGSD
SAIMTGSYNNFFRMFDRDTRRDVTLEASRESSKPRASLKP RKVCTGGKRRKDEISVDSL
FNKKILHTAWHPVDNVIAVAATNNLYIFQDKIN
>sp|P28221|5HT1D_HUMAN 5-hydroxytryptamine receptor 1D OS=Homo sapi
N=HTR1D PE=1 SV=1
MSPLNQAEGLPQEASNRSLNATETSEAWDPRTLQALKISLAVVLSVITLATVLSNAFVL
TTILLTRKLHTPANYLIGSLATTDLLVLSILVMPISIAYTITHWNFGQILCDIWLSSDIT
CCTASILHLCVIALDRYWAITDALEYSKRRTAGAHATMIAIVWAISICISIPPLFRQAK
AQEMS DCLVNTSQISYTIYSTCGAFYIPSVLLIILYGRYRAARNRILNPPSYGKRFT
TAHLITGSAGSSLCSLNSLHEGHSHSAGSPLFFNHVKIKLADSALERKRISAARERKAT
KILGIILGAFIICWLPFFVSVLVLPCIRDSCWIHPALDFDFTWLGYNLSLNIPIYTVFN
EEFRQAFQKIVPFRKAS
>sp|Q2UXF7|6FEH_WHEAT Fructan 6-exohydrolase OS=Triticum aestivum O
EH PE=1 SV=1
MAARLPLAACVVAFLHCLLLSSLVRSPTALRRLSEAESSLVRHGHGVGIRPAYHFLPAK
NQNDPNPGMYHNGVYHMFYQYNPLGAMWQPGNLSWGHVSRSRLVNW DALD TALDPTAF
DYNGCWSGSATILPGGIPALLYTGRIDADKEVQVQNVAFPKNPADPLLREWVKPAYNPVI
PLPADVPGDNFRDPTTAWVGRDGLWRIA VAAKVGGPNGIASTLIYRSKDFRHWKRNASPL
YTSRAAGMVECPDLFPVAEPGVEEGRGLGYASGPASGAVRHVLLKLSVMNTTQDYAVGRYD
DVADTFVPEVDVERNADD CRTWRRFDYGHVYASKSFFDSSKNRRVLWAWANESDSQDNDI
ARGWSGVQTVPRKVWLD EDEGKQVRQWPIEEIETLRSKR VVGLLGAQV NAGGVNKITGVGA
QADVEAIFEIPSLEEAETFPQNWLLDPQKLCEENGASVPGKVGPPGGLLVMASSNMQEHTA
LFRVYFRNQYVFLMCTDLTRSTGRNVYKPYGGFVQIDYGGQHTISLRTLIDSGVV
ERFGGQRTCTTARVYPERADERSHVYVFNSTGLVYVSLAWLAAASVSVYDIA
```



## $k$ -tuples

We can divide any sequence into shorter stretches of length  $k$ . Such pieces are known as  $k$ -tuples (a.k.a.  $k$ -words or  $k$ -mers).



## $k$ -tuples

We can divide any sequence into shorter stretches of length  $k$ . Such pieces are known as  $k$ -tuples (a.k.a.  $k$ -words or  $k$ -mers).

For example, the sequence  
ASEQUENCE renders 7 3-tuples.

ASEQUENCE

\*\*\*

\*\*\*

\*\*\*

\*\*\*

\*\*\*

\*\*\*

\*\*\*



# Indexing

```
>sp|Q66LE6|2ABD_HUMAN Serine/threonine-protein phosphatase 2A 55 kD
subunit B delta isoform OS=Homo sapiens OX=9606 GN=PPP2R2D PE=1 SV=
MAGAGGGGCPAGGNDFQWCFSQVKGAIDEDVAEADIISTVEFNYSGLLATGDKGGRVVI
FQREQENKSRPHSRGEYNVYSTFQ SHEPEFDYLSLEIEEKINKIRWLPQQNAAHLLST
NDKTIKLWKISERDKRAEGYNLKDDEGRLRDPFRITALRVPILKPMDLMEASPRRIFAN
AHTYHINSISVNSDHETYSADDLRINLWHELEITDRSFNIVDIKPANMEELTEVITAAEF
HPHQCNVVFVYSSSKGTIRLCDMRSSALCDRHSKFEEPEPSSRSFFSEIISSISDVKFS
HSGRYMTRDYLSVKVWDLNMRPVETHQVHEYLRSKLSLYENDCIFDKFECWNGSD
SAIMTGSYNNFFRMFRDRDTRRDVTLSEASSKPRASLKRKRVCTGGKRRKDEISVDSLD
FNKKILHTAHPVDNVIAVAATNNLYIFQDKIN
>sp|P28221|5HT1D_HUMAN 5-hydroxytryptamine receptor 1D OS=Homo sapi
N=HTR1D PE=1 SV=1
MSPLNQSAEGLPQEASNRSLNATETSEAWDPRTLQALKISLAVVLSVITLATVLSNAFVL
TTILLTRKLHTPANYLIGSLATDILLVLSILVMPISIAYTITHWNFGQILCDIWLSSDIT
CCTASILHLCVIALDRYWAITDALEYSKRRTAGAHATMIAIIVWAISICISIPPLFWRQAK
AQEEMSDCLVNTSQISYTIYSTCGAFYIPSVLLIILYGRYRAARNRILNPPSLYGRKFT
TAHLITGSAGSSLCSLNSSLHEGSHSAGSPLFFNHVKIKLADSALERKRISAARERKAT
KILGILGAFIICWLPFFVSVLVLPICRDSCWIHPALDFDFTLWGLYNSLNIPIYTVFN
EEFRQAFQKIVPFRKAS
>sp|Q2UXF7|6FEH_WHEAT Fructan 6-exohydrolase OS=Triticum aestivum O
EH PE=1 SV=1
MAARLPLAACVAFHLCLLSSLVRSPTALRRLSEAESSLVRHGHGVGIRPAYHFLPAK
NWQNDPNPMPYHNGVYHMFYQYNPLGAMWQPGNLSWGHVSRSRLVNDALDALTALDPTAPF
DYNGCWSGSATILPGGIPALLYTGRIDADKEVQVQNVAFPKNPADPLLREWVKPAYNPVI
PLPADVPDGNFRDPTTAWVGRDGLWRIAVAANKVGGPNGIASTLIYRSKDFRHWKRNASPL
YTSRAAGMVECPDLFPVAEPGVEEGRLYASGPASGAVRHVHLKLSVMNTTQDYYAVGRYD
DVADTFVPEVDVERNADDRCRTWRRFDYGHVYASKSFFDSSKRRRVLWAWANESDSQDNDI
ARGWSGVQTVPRKVVWLEDEGKQVRQWPIEEIETLRSKRIVGLLGAQVNAVGGVNKITGVGA
QADWEATFEIPSLREAEFTFQNWLLDDPKLCEENGASVPGKVGDFGLVWAGSNWQETA
IYPRVFRNGKYKVLWTDLHSTGSDRYKPSYSGFVDTDEGQATISLTLIDRSVV
EHPGGDSRTGITARVYPRADRNDRVYVFNQGLVYVSKLAWLAAQVYVYVHCR
```



# Indexing

...  
...  
IIS  
...  
MAA  
MAG  
MAS  
...  
VIT

```
>sp|Q66LE6|2ABD_HUMAN Serine/threonine-protein phosphatase 2A 55 kD
subunit B delta isoform OS=Homo sapiens OX=9606 GN=PPP2R2D PE=1 SV=
MAGAGGGGCPAGGNDFQWCFSQVKGAIDEDVAEADIISTVEFNYSGDLLATGDKGGRVVI
FQREQENKSRPHSRGEYNVYSTFQSHPEFDYLSLEIEEKINKIRWLPQQNAHFLLST
NDKTIKLWKISERDKRAEYGNLKDEDGRLRDPFRITALRVPILKPMDLMVEASPRRIFAN
AHTYHINSISVNSDHETYLSADDLRINLWHLEITDRSFNIVDIKANMEELTEVITAAEF
HPHQCNVVFVSSSKGTIRLCDMRSSALCDRHSKFFEEPDPSSRSFFSEIISISDVKFS
HSGRYMMTRDYL SVKVDLNMESRPVETHQVHEYLRSKLSLYENDCIFDKFECCWNGSD
SAIMTGSYNNFFRMRDRDRDVTLEASRESSKPRASLKRKVKCTGGKRRRDEISVDSLD
FNKKILHTAWHPVDNVIAVAATNNLYIFQDKIN
>sp|P28221|5HT1D_HUMAN 5-hydroxytryptamine receptor 1D OS=Homo sapi
N=HTR1D PE=1 SV=1
MSPLNQAEGLPQEASNRSLNATETSEAWDPRTLQALKISLAVVLSVITLATVLSNAFVL
TTILLTRKLHTPANYLIGSLATTDLLVSLVMPISIAYTITHWNFGQILCDIWLSSDIT
CCTASILHLCVIALDRYWAITDALEYSKRRTAGAHATMIAIVWAISICISIPPLFWRQAK
AQEEMSDCLVNTSQISYTIYSTCGAFYIPSVLLIILYGRYRAARNRILNPPSLYGRFT
TAHLITGSAGSSLCSLNSLHEGHSAGSPLFFNHVKIKLADSALERKRISAARERKAT
KILGIILGAFIICWLPPFVSVLVLPICRDSCWIHPALDFFTWLGYNLSLNIPIYTVFN
EEFRQAFQKIVPFRKAS
>sp|Q2UXF7|6FEH_WHEAT Fructan 6-exohydrolase OS=Triticum aestivum O
EH PE=1 SV=1
MAARLPLAACVVAFLHCLLLSSLVRSPTALRRLSEAESSLVRHGHGVGIRPAYHFLPAK
NWQNDPNGPMYHNGVYHMFYQYNPLGAMWQPGNLSWGHVSVRDLVNWVDALDALTDPAPF
DYNGCWSGSATILPGGIPALLYTGRIDADKEVQVQNVAFPKNPADPLLREWVKPAYNPVI
PLPADVPGDNFRDPTTAWVGRDGLWRIA VAAKVGGPNGIASTLIYRSKDFRHWKRNASPL
YTSRAAGMVECPDLFPVAEPGVEEGRGLGYASGPASGAVRHVLLKLSVMNTTQDYAVGRYD
DVADTFVPEVDVERNADD CRTWRRFDYGHVYASKSFFDSSKNRRVLWAWANESDSQDNDI
ARGWSGVQTVPRKQVWLEDEGKQVRQWPIEEIETLSKRVRVGLLGAQVNAAGVKNITGVGA
QADVEAIFEIPSLAEAETFPQNWLLDPQKLCEENGASVPGKVGPPGGLLVMASSNMQEHTA
LFRVYFRNQKYYVLMVTLTRSTGRDNYKPYGGFVQVYRGGQRTISLRTLDSGVV
ERPGGQRTCTTARVYFTRADERSHVYFPRNRTLVEYSLELAWLANASVSVYRKA
```





## Basic Local Alignment Search Tool (BLAST)

- ▶ Heuristic to create local alignments, to find the closest homologs of all sequences in a database.



## Basic Local Alignment Search Tool (BLAST)

- ▶ Heuristic to create local alignments, to find the closest homologs of all sequences in a database.
- ▶ Depends on indexes of  $k$ -tuples of the sequence database. Typically one select  $k = 3-6$  for proteins and  $k = 6-20$  for DNA



## Basic Local Alignment Search Tool (BLAST)

- ▶ Heuristic to create local alignments, to find the closest homologs of all sequences in a database.
- ▶ Depends on indexes of  $k$ -tuples of the sequence database. Typically one select  $k = 3-6$  for proteins and  $k = 6-20$  for DNA
- ▶ Algorithm:
  1. Divide the query sequence into  $k$ -tuples.
  2. Expand the list of  $k$ -tuples to any  $k$ -tuples that match with an alignment score  $>$  a threshold  $T$ .
  3. Scan the index for all sequence matching the expanded list.
  4. For each matching sequence, extend the matches. These extended matches are known as high-scoring segment pairs (HSPs).
  5. Calculate the significance of each HSP.



## Extending $k$ -tuple alignments



## Significance (Karlin-Altschul statistics)

The number of sequences expected to get a local alignment score greater or equal to  $S$  can be estimated as

$$E = Kmne^{-\lambda S}$$

Here,  $m$ , is the length of the query sequence,  $n$  is the summed length of all sequences in the database, and  $\lambda, K$  are constants estimated for each scoring matrix and database.

$E$  is known as the expected value or  $E$ -value of a HSP. I.e. "How many sequence would I expect by chance to match as well as this HSP?"



Thanks!