

A few BLAST details

Julin Maloof

April 17, 2018

Slides courtesy of Venkatsean Sundaresan

BLAST (Basic Local Alignment Search Tool)

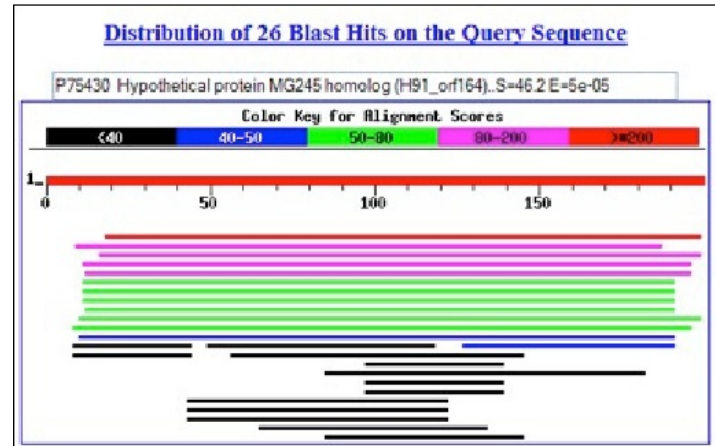
QUERY sequence(s)

```
>gi|15237380|ref|NP_197163.1| myb family transcription factor (MYB43) [Arabidopsis thaliana]
MGRQPCQDKVGLKKGPTIEEDKKLINFILTHGMCWRALPKLSGLRCGKSCRLRWINYLRPDLKRGLL
SEYEEQKVINLHAQLGNRWKXIASHLPGRTDNEIKHWNTHIKKLLKMGIDPLTHKPLSEQEAQQAGG
RKKSLVPHDDKNPKQDQQRDEQHQLEQALEKNNTSVSGDGFCDVPLINPHEILIDISSHHHHSN
DONVNINTSKFTSPSSSSSTSSCISVVVPGDEFKFFDEMEILDKLWSSDDSLGDDISKDGKFNNTV
DTMNLWDINDLSSLDMMFNEHDDGFIGNGNGSRMVLDDQSWTFDLL
```

BLAST program

BLAST
database

BLAST results



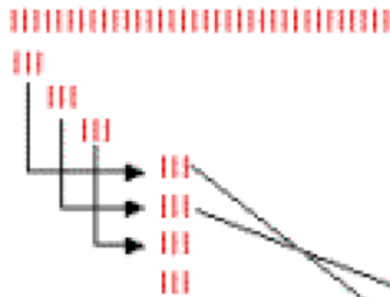
- Search for similarity to infer “homology”
- “mutual best hits” or reciprocal BLAST

BLAST

- BLAST is optimized to search large databases quickly.
- How does it do this?

BLAST: Heuristic algorithm

Query sequence of length L (this is the sequence with which you do a search)



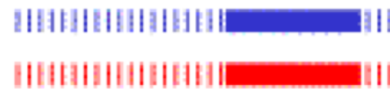
Compile list of words (w) from query
usually $w=3$ for proteins
There are $L-w+1$ words in sequence L
Begin with high scoring words



Compare word list with sequences
in database and identify matches



Extend matches in both directions
until further extension causes the
score to drop by a certain amount



High scoring segment pair HSP

Q : ROBJOEZACANNLIZ

Break this up into 3 letter words

ROB , OBJ , BJO , . . , ZAC , ...ANN , ...NLI , LIZ

Search sequences S1, S2, etc. in database

Find a match with the word ZAC then extend on both sides until no
or weak matches

Q : ROBJOEZACANNLIZ

S1 : TOMZOEZACANNLIA

Q : ROBJOEZACANNLIZ

S2 : TOMZOEZACAMYLEA

Search with high scoring words first for better chance
of high scoring alignments

Q : LVAAVGVCWDILRAAA

In the above example, BLOSUM62 scores for matches to LVA
and CWD are 12 and 26 respectively, so search with CWD

Q : LVAAVGVCWDILRAAA

| | | | | | | |

S : AGGAVVVCWDILKAGG

useful parameters

- Word size: the size of the chunks that the query sequence is chopped into
- Threshold: minimum score for a word match to be considered to seed an extension

Query Word ($W = 3$)

TL SHAWRLSNETDKRPFIE TAERL **RDQ** HKKDYPEYKYQPRRRKNGKPGSSSEADAHSE

Determine neighborhood

RDQ 16	QDQ 12	EDQ 11	RDN 11	RDB 11	BDQ 10	RDP 10
RBQ 14	REQ 12	HDQ 11	RDD 11	ADQ 10	XDQ 10	RDT 10
RDZ 14	RDR 12	ZDQ 11	RDH 11	MDQ 10	RQQ 10	RDY 10
KDQ 13	RDK 12	RNQ 11	RDM 11	SDQ 10	RSQ 10	RDX 10
RDE 13	NDQ 11	RZQ 11	RDS 11	TDQ 10	RDA 10	DDQ 9 ...

Seed using neighborhood words
greater than neighborhood score
threshold ($T = 11$)

Query: 1 TL SHAWRLSNETDKRPFIE TAERL **RDQ** HKKDYPEYKYQPRRRKNGKPGSSSEADAHSE 58
TL WRL N +KRPF+E AERLR+QHKKD+P+YKYQPRRRK+ K G S D +
Sbjct: 140 TLESGWRLNENPGEKRPFVEGAERL **REQ** HKKDHPDYKYQPRRRKSVKNGQSEPEDGSEQ 197

How
BLAST
works

HSP = High-scoring Segment Pair – a segment pair whose score will not increase by further extension or by trimming

Score (S) = measures alignment quality (scoring matrix - gaps)

E value (E) = number of different alignments with score S that are expected to occur by chance in a search of that database

BLAST Summary

BLAST essentially computes regions of high “similarity” in local alignments of 2 proteins

- BLAST breaks search into “chunks” by finding all subsequences (stretches of similarity, or “words”) of length k (e.g., $k=3$) that occur in both seqs
 - build score on matches (scoring matrix, gap cost)
 - extend from subsequences to see if you can increase score
 - HSP (High-scoring Segment Pair whose score cannot be improved by extension or trimming)
 - compute total score (when no more extensions are possible)
- Then compare BLAST score against precomputed expected scores for all proteins in database
- Then rank score