

Basic Local Alignment Search Tool

Jie Zhang 张洁

School of Life Sciences, Peking University

Why BLAST?

"Homology is the central concept for all of biology."

——David Wake. *Science*, 1994

BLAST is the tool most frequently used for calculating sequence similarity, by searching the databases.

If you work with one or a few proteins or genes, it can tell you about their conservation, active sites, structure and regulation in other organisms, etc.

What BLAST does?

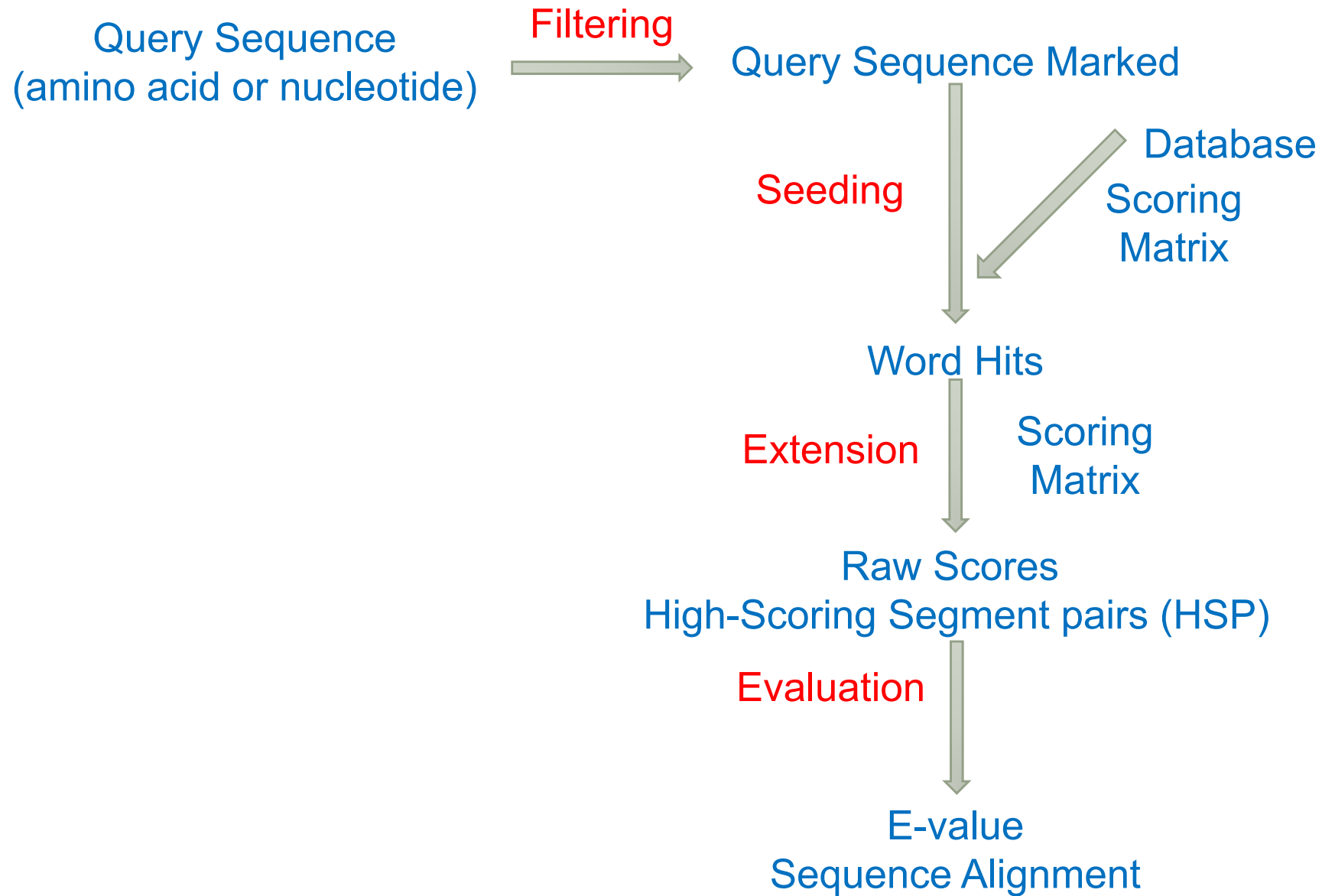
- **Identity**: the occurrence of exactly the same nucleotide or amino acid in the same position in aligned sequences.
- **Similarity**: measure the sameness or difference of the sequences
- **Homology**: is defined in terms of shared ancestors. Homologous sequences are often similar. Sequence regions that are homologous are also called **conserved** regions.

Pertsemlidis, et al. Genome Biol, 2001

Different alignment algorithms

Algorithms	Strategies	Accuracy	Speed
Dynamic Programming	Exhaustive	★★★★★	★★★
FASTA, BLAST	Heuristic	★★★★	★★★★★

How BLAST works



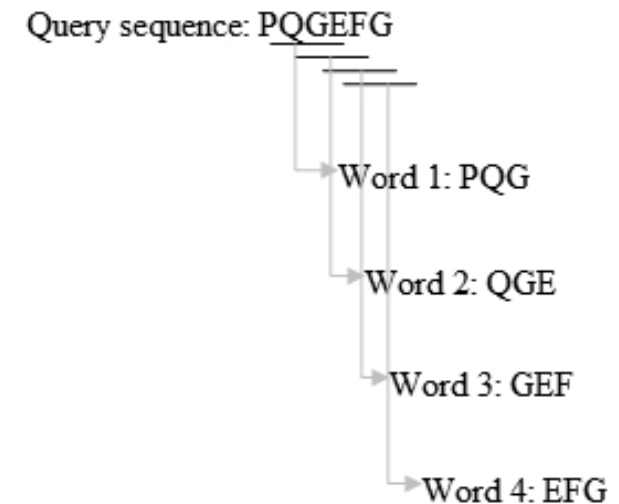
Step 0 — Filtering

- To prevent the production of large numbers of statistically significant but biologically uninteresting results.
- **Low complexity and repeats**, i.e.
 - (CA)_n
 - KLKCLKLKCLKL
- Cover these regions with the following letters
 - Ns (for nucleotide residues)
 - Xs (for amino acid residues)
- - *F* flag: filter query sequence

Altschul, S. F., et al. J Mol Biol, 1990

Step 1 — Seeding

- Make a **w**-letter word list of the query sequence
 - Usually 3 for amino acid sequences, and
 - 11 for nucleotide sequences
- For a query sequence with n letters, the number of words is $n - w + 1$
- - W flag: word size



http://en.wikipedia.org/wiki/BLAST#cite_note-6

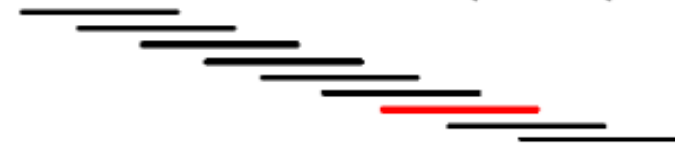
Altschul, S. F., et al. J Mol Biol, 1990

Step 2 — Search word hits

- Scoring matrix
 - for amino acids, BLOSUM or PAM
 - For DNA words, a match is scored as +5 and a mismatch as -4, or as +2 and -3
- No gaps are allowed
- The words whose scores are greater than the threshold ***T*** will remain in the possible matching words list

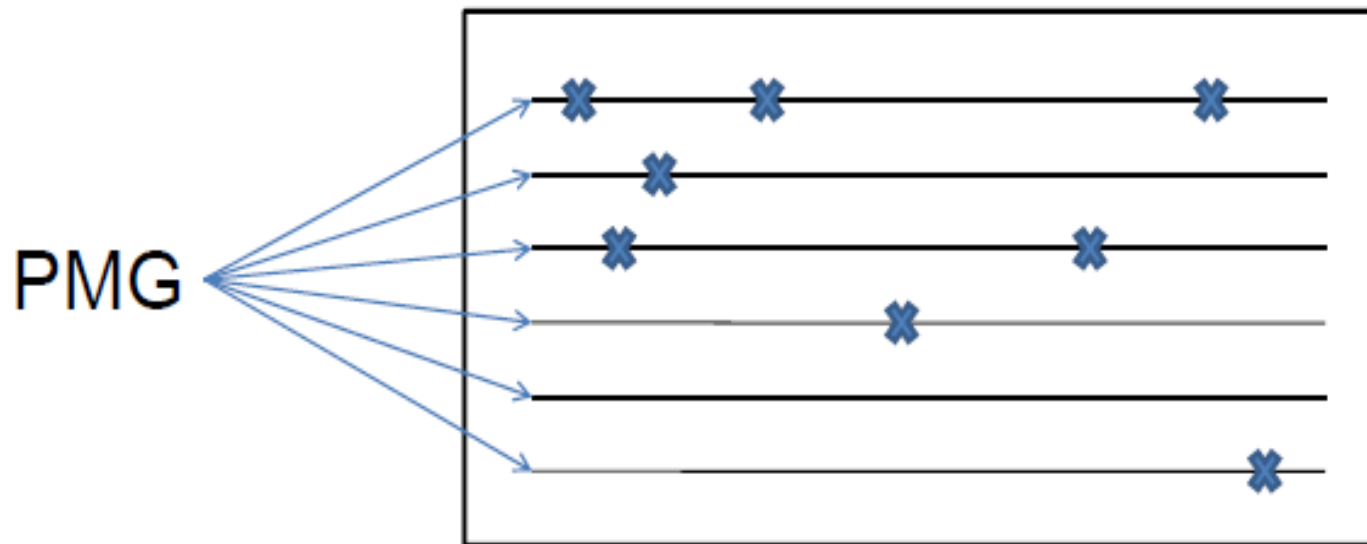
Altschul, S. F., et al. J Mol Biol, 1990

Step 2 — Search word hits

L N K C K T P Q G Q R				Query sequence
				
P	Q	G	$7+5+6=18$	Word
P	E	G	$7+2+6=15$	Neighborhood words
P	R	G	$7+1+6=14$	
P	K	G	$7+1+6=14$	
P	N	G	$7+0+6=13$	
Threshold T=13	P	M	G	$7+0+6=13$
<hr/>				
P	Q	A	$7+5+0=12$	
P	Q	N	$7+5+0=12$	
etc.				

Step 3 — Scanning

- HashTable: direct addressing method
- Deterministic finite automaton/finite state machine: much faster



Altschul, S. F., et al. J Mol Biol, 1990

Step 4 — Extending → HSP

➤ Cutoff score **S**

Query sequence: R P P Q G L F

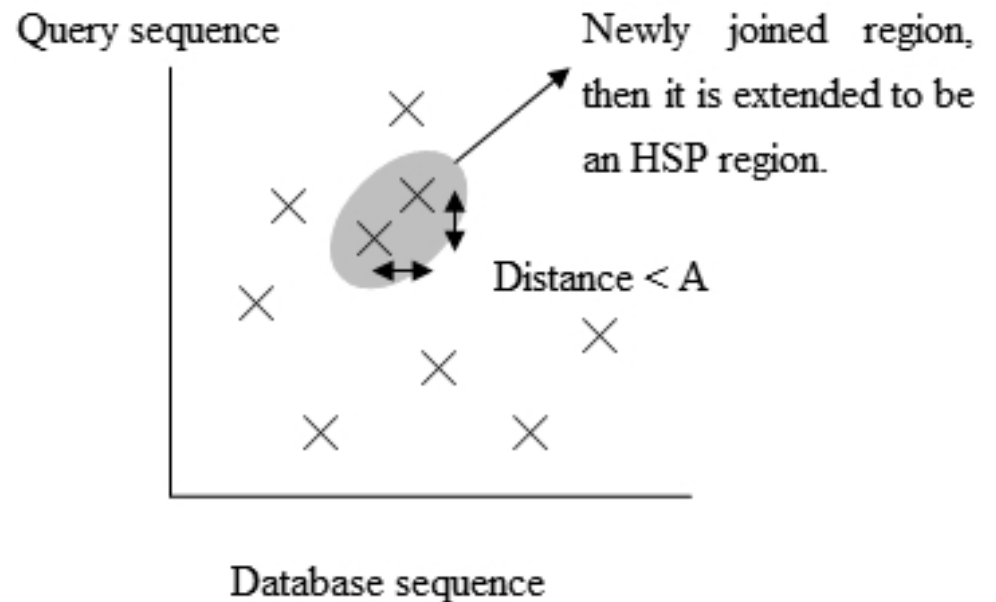
Database sequence: D P P E G V V

└→ Exact match is scanned.

Score: -2 7 7 2 6 1 -1

└→ HSP

Optimal accumulated score = $7+7+2+6+1 = 23$



http://en.wikipedia.org/wiki/BLAST#cite_note-6

Step 5 — Significance evaluation

➤ **Raw scores**: have little meaning without detailed knowledge of the scoring system used.

➤ **Bit scores**: normalizing a raw score using the formula

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

➤ **E values**: corresponding to a given bit score

$$E = mn2^{-S'}$$

Pertsemlidis, et al. Genome Biol, 2001

Step 5 — Significance evaluation

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

m: length of the query sequence
n: length of the database
S: score of HSP

$E > 1$, the alignment occurred by chance
 $E < 0.1$ or 0.05 , statistically significant
 $E < 10^{-5}$, high similarity

Pertsemlidis, et al. Genome Biol, 2001

BLAST programs

nucleotide blast

Search a **nucleotide** database using a **nucleotide** query

Algorithms: blastn, megablast, discontinuous megablast

protein blast

Search **protein** database using a **protein** query

Algorithms: blastp, psi-blast, phi-blast, delta-blast

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

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

BLAST programs

Program	Query Sequence	Database
blastp	Amino acid sequence	Protein sequence database
blastn	Nucleotide sequence	Nucleotide sequence database
blastx	Nucleotide sequence translated in all reading frames	Protein sequence database
tblastn	Amino acid sequence	Nucleotide sequence database translated in all reading frames
tblastx	Six-frame translations of a nucleotide sequence	Six-frame translations of a nucleotide sequence database

Gapped BLAST

- Adopts a lower neighborhood word score threshold to maintain the same level of sensitivity for detecting sequence similarity.
- Gaps allowed

Altschul, S. F., et al. Nucleic Acids Res, 1997

PSI-BLAST

- Position-Specific Iterative BLAST (blastpgp)
- Constructs PSSMs (position specific scoring matrix) automatically
- Searches protein database with PSSMs
- Used to find distant relatives of a protein, and is much more sensitive in picking up distant evolutionary relationships than a standard protein-protein BLAST.

Altschul, S. F., et al. Nucleic Acids Res, 1997

Caveat emptor

- BLAST is based on the similarity statistics, but statistical significance doesn't mean biological significance.
- Always compare Protein sequences if the query sequences encode proteins.
- **Remember**
 1. Similarity does not imply homology!
 2. Non-homology cannot from non-similarity.
 3. Do not use the term “**percent homology**”.

Pertsemlidis, et al. Genome Biol, 2001

References

1. Altschul, S. F., et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.
2. Pertsemlidis, et al. (2001) "Having a BLAST with bioinformatics (and avoiding BLASTphemy)" Genome Biol, REVIEWS2002
3. Altschul, S. F., et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.
4. Joseph Bedell, Ian Korf, Mark Yandell. BLAST. O'Reilly Media, Inc, USA, 2003

Further readings

1. NCBI-BLAST web site <http://blast.ncbi.nlm.nih.gov/Blast.cgi>
2. WU-BLAST web site <http://www.ebi.ac.uk/Tools/sss/wublast/>
3. NCBI BLAST course
http://biology.unm.edu/biology/maggiww/Public_Html/444544seqsim.html
4. Applied Bioinformatics Course web site <http://abc.cbi.pku.edu.cn/>

Acknowledgement

Qi Wang

Lu Dong

Ping Zhu

Zi-Tian Chen

Wei Xu

Lei Sun

Fang-Min Tian

Meng Wang

Yang Ding

Yong-Xin Ye

Xiao-Xu Yang

Dr. Ge Gao

Dr. Li-Ping Wei

Thanks for your attention!