How blast works

http://faculty.washington.edu/jht/GS559_2017/

Genome 559: Introduction to Statistical and Computational Genomics

Prof. James H. Thomas

Fast alignment searches

- Most common method is the BLAST search (Basic Local Alignment Search Tool). Initial step is different from dynamic programming alignment.

- Search sequence broken into small **words** (usually 3 residues long for proteins). 20 * 20 * 20 = 8,000 protein words. These act as **seeds** for searches.

- The target dataset is pre-indexed for all positions that have an ungapped match for each word above some score threshold (using a score matrix, by default BLOSUM62).

BLAST searches

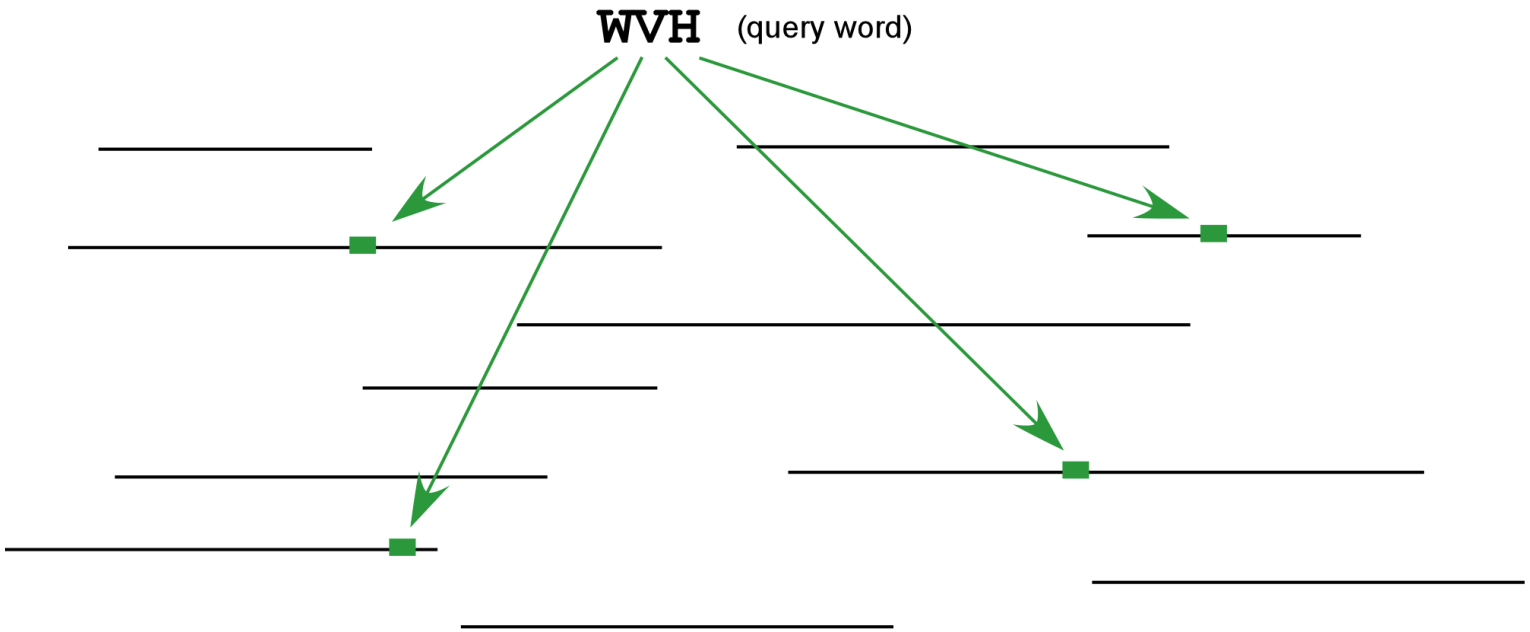
- For example, the search sequence word "WVH" might score above threshold with these indexed sequences:

Indexed word	Score
WVH	**23**
WIH	**22**
WVY	**17**
WIY	**16**

- Target sequences around each indexed word hit are retrieved and match is extended in both directions:

```
...VFEWVHLLP...      your sequence
   ⟵ WIY ⟶           database (many sites)
```

Schematic of indexed matches



Result – instead of aligning these 3 amino acids to <u>everything</u>, they are aligned only with the tiny fraction of sequence regions that are good candidates for a valid alignment.

Extension and scoring

	Match Score:	Total Score:
...QSVFEWVHLLPGA... ..WIY..	16	16
...QSVFEWVHLLPGA... ..WIY**Q**..	-3	13
...QSVFEWVHLLPGA... ..WIYQ**K**..	-2	11
...QSVFEWVHLLPGA... ..WIYQK**A**..	-1	10

Extension termination and Reporting

- Extension continued until alignment score drops below some threshold.

- Extensions whose maximal score is above some threshold are kept for reporting. Traceback starts at maximal score.

- For web interfaces, various formatting, links, and overviews are added.

- It is easy to set up blast on your local computer; useful for custom databases and automation.

Key to speed: word matching and prior indexing

- Only a very small part of total search space is analyzed.

- Word positions are indexed prior to the search, so the relevant parts of search space are reached quickly.

- Tradeoff is sensitivity – occasionally matches will be missed (e.g. when differences are common and dispersed enough that no local words match above threshold).

Blast match statistics

- E-value (expect value) reports number of matches of this score (or higher) expected if the database were composed of random sequences.

- Scores (aka bit scores) are independent of database size. They simply measure the quality of the specific alignment found.

- E-values are DEPENDENT on database size (in a random dataset, the more data, the more likely you are to find a match of a given score or higher.)

link

score

E-value

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Links
YP_001206898.1	voltage-dependent potassium channel [Bradyrhizobium sp. ORS278]	45.1	45.1	97%	2e-04	G
YP_001203097.1	hypothetical protein BRADO0944 [Bradyrhizobium sp. ORS278]	42.7	42.7	90%	6e-04	G
YP_422725.1	Kef-type K+ transporter NAD-binding component [Magnetospirillum magneticum AMB-1]	42.7	42.7	90%	7e-04	G
NP_774496.1	hypothetical protein blr7856 [Bradyrhizobium japonicum USDA 110]	42.7	42.7	97%	8e-04	G
ZP_00054971.2	COG1226: Kef-type K+ transport systems, predicted NAD-binding component [Magnetospi	41.6	41.6	90%	0.001	
ZP_01903404.1	Potassium channel protein [Roseobacter sp. AzwK-3b]	41.2	41.2	90%	0.002	
YP_001241318.1	voltage-dependent potassium channel [Bradyrhizobium sp. BTAi1]	40.8	40.8	97%	0.003	G
ZP_01056912.1	potassium channel protein [Roseobacter sp. MED193]	40.0	40.0	90%	0.005	
ZP_00053231.2	COG1226: Kef-type K+ transport systems, predicted NAD-binding component [Magnetospi	39.7	39.7	97%	0.006	
YP_421023.1	ATP-sensitive inward rectifier potassium channel 10 [Magnetospirillum magneticum AMB-1]	39.7	39.7	92%	0.006	G
YP_423196.1	Kef-type K+ transporter NAD-binding component [Magnetospirillum magneticum AMB-1]	39.7	39.7	97%	0.006	G
YP_759166.1	cation channel family protein [Hyphomonas neptunium ATCC 15444]	39.3	39.3	90%	0.008	G
ZP_00055625.2	COG1226: Kef-type K+ transport systems, predicted NAD-binding component [Magnetospi	38.9	38.9	92%	0.010	
YP_001832925.1	Ion transport 2 domain-containing protein [Beijerinckia indica subsp. indica ATCC 9039]	38.5	38.5	92%	0.015	G
ZP_05085139.1	Ion channel family protein [Pseudovibrio sp. JE062]	38.1	38.1	100%	0.018	
ZP_01753431.1	Potassium channel protein [Roseobacter sp. SK209-2-6]	38.1	38.1	87%	0.019	
ZP_01546037.1	cyclic nucleotide-binding domain (cNMP-BD) protein [Stappia aggregata IAM 12614]	37.7	37.7	87%	0.021	
NP_772389.1	hypothetical protein bll5749 [Bradyrhizobium japonicum USDA 110]	37.4	37.4	82%	0.028	G
YP_001419445.1	cyclic nucleotide-binding protein [Xanthobacter autotrophicus Py2]	37.4	37.4	65%	0.033	G
YP_568497.1	Ion transport protein [Rhodopseudomonas palustris BisB5]	37.4	37.4	75%	0.034	G
ZP_05786154.1	potassium channel protein [Silicibacter lacuscaerulensis ITI-1157]	37.0	37.0	90%	0.035	
ZP_05083809.1	Ion transport protein [Pseudovibrio sp. JE062]	37.0	37.0	97%	0.037	
ZP_01747053.1	Potassium channel protein [Sagittula stellata E-37]	37.0	37.0	90%	0.041	
YP_002362527.1	Ion transport 2 domain protein [Methylocella silvestris BL2]	36.6	36.6	92%	0.046	G
NP_949569.1	cyclic nucleotide regulated K+ channel [Rhodopseudomonas palustris CGA009]	36.2	36.2	75%	0.070	G
YP_001993678.1	cyclic nucleotide-binding protein [Rhodopseudomonas palustris TIE-1]	36.2	36.2	75%	0.070	G
ZP_02151093.1	potassium channel protein, putative [Phaeobacter gallaeciensis 2.10]	35.8	35.8	75%	0.080	
ZP_05114270.1	transporter, cation channel family [Labrenzia alexandrii DFL-11]	35.8	35.8	87%	0.094	
ZP_05738744.1	Ion transport protein [Silicibacter sp. TrichCH4B]	35.4	35.4	97%	0.11	
YP_780671.1	cyclic nucleotide-binding protein [Rhodopseudomonas palustris BisA53]	35.4	35.4	75%	0.12	G
YP_533883.1	cyclic nucleotide-binding domain-containing protein [Rhodopseudomonas palustris BisB18]	35.4	35.4	75%	0.12	G
ZP_05052091.1	transporter, cation channel family [Octadecabacter antarcticus 307]	35.4	35.4	87%	0.13	
ZP_01546940.1	potassium channel related protein [Stappia aggregata IAM 12614]	35.0	35.0	87%	0.15	
ZP_05067149.1	Ion transport protein [Octadecabacter antarcticus 238]	35.0	35.0	90%	0.15	
ZP_01437479.1	extracellular solute-binding protein, family 3 [Fulvimarina pelagi HTCC2506]	34.7	34.7	92%	0.17	
YP_484999.1	cyclic nucleotide-binding domain-containing protein [Rhodopseudomonas palustris HaA2]	34.3	34.3	75%	0.25	G
ZP_02189219.1	hypothetical protein BAL199_06269 [alpha proteobacterium BAL199]	33.9	33.9	77%	0.38	
YP_003447642.1	hypothetical protein AZL_004600 [Azospirillum sp. B510]	33.5	33.5	97%	0.45	
YP_571140.1	cyclic nucleotide-binding [Rhodopseudomonas palustris BisB5]	33.1	33.1	75%	0.52	G
ZP_01003918.1	putative potassium channel protein [Loktanella vestfoldensis SKA53]	33.1	33.1	90%	0.56	
YP_001264879.1	TrkA domain-containing protein [Sphingomonas wittichii RW1]	33.1	33.1	62%	0.62	G
YP_001233535.1	voltage-gated potassium channel [Acidiphilium cryptum JF-5]	33.1	33.1	90%	0.63	G
YP_002973442.1	Ion transport 2 domain protein [Rhizobium leguminosarum bv. trifolii WSM1325]	32.7	32.7	85%	0.66	G
ZP_01745766.1	Ion transport protein [Sagittula stellata E-37]	32.7	32.7	92%	0.68	
ZP_05113853.1	Ion channel family [Labrenzia alexandrii DFL-11]	32.7	32.7	90%	0.77	
YP_001524780.1	cyclic nucleotide-binding protein [Azorhizobium caulinodans ORS 571]	32.7	32.7	65%	0.81	G

You should know…

- How blast speeds up pair alignments.
- A blast alignment is essentially the same as a local DP alignment.
- What an E-value tells you.