

BLAST as a microcosm of all that is wrong with computational biology

Titus Brown

Nov 7, 2013

(NCBI) BLAST

- You've all used it?
- Very popular! Fast, sensitive way to find sequence similarity => putative homology.
- Primary sequence comparison tool used by biologists, computational biologists.

(NCBI) BLAST

- One or more query sequences...
- against a “subject” database.
- Finds core strong match, extends outwards.

```
SIRGGGVVDHGISDDESQHSGDAGIS  
S+RGGG++ G+S++          D+G  
SVRGGGIEIGLSEE-----DSGAE
```



Consideration #1

- BLAST *only* cares about sequence similarity.
- No positional information taken into account, for e.g. protein domains.

Query: 1628 AHLLVNSQKC-KQTSSECIDTTDNAASVISARAS----TGTLEAEFPINTVASTTNPTTP 1682
N+Q C K C + S++ +A+ + TL F I++ ST T

Sbjct: 1537 MTFHANTQMCVKLDLQSCPTNVASVKSILGEKAAEFSTSSTLSRVFRIDSEGSTQTGT-- 1710

Query: 1683 PQDYTYXXXXXXXXXXXXXXXXXXXXXTHRKRKRETSTLWAPGFGFNVTKKQRREPIGQDDLNG 1742
Y +KRKRE LW PEGF + KK+R+E ++LN

Sbjct: 1711 TNYLVYIIAGGGIMVLIIVIAGVIVSQRKRRENGNLWVPEGFQLFKRRKE----NELNL 1878

Query: 1743 LNGSIHPGELTQLDT-AGTPFLNRWENTSLPQKSNHYHVQYTPENITFLPNNGTVPXXXX 1801
N L++ D A TPFL + + Q S + +L

Sbjct: 1879 NN-----LSKADMNAQTPFL---PHATEAQASKYSASSSDTPETDYL----- 1995

Query: 1802 XXXXXXXXXXXXXEPTDNRKWTPQHLEAADLSRAGSACTPVTDLTPPPHIDVDEDDVNAR 1861
D R+WTP HLEAA+ S C + TPP + DD+NAR

Sbjct: 1996 -----HGSCASKEDKRQWTPHHLEAANNSNVN--CQIMN--TPPQSECPESSDINAR 2139

Query: 1862 GPDGVTPLMVASIRGGGVHDHGISDDESQHSGDAGISGEGSDSMIXXXXXXXXXXXXXXTDR 1921
GPDG TPLM+AS+RGGG++ G+S++ D+G GEGSD+MI TDR

Sbjct: 2140 GPDGYTPLMIASVRGGGIEIGLSEE-----DSGAEGEGSDNMIAGLILQGASLSATTDR 2301

Query: 1922 SGETXXXXXXXXXXXXXXXXXXXXXXXXXNMKDHSGRTPPLHSAVAADAQGVFQILLRNRAT 1981
+GET NMKD +GRTPPLH++VAADAQGVFQILLRNRAT

Sbjct: 2302 TGETALHLAARYARADAAKRLLDAGADANMKDQTGRTPPLHNSVAADAQGVFQILLRNRAT 2481

Query: 1982 DLDARTNDGTTMILASRLAVEGMVEELISANADVNAVDDHGKSALHWAAVNNVDAVST 2041
DLDA+TNDGTTTP+ILASRLAVEGMVE+LI+A+ADVNAV+HGKS+LHWAAVNN DA+

Sbjct: 2482 DLDAKTNDGTTPLILASRLAVEGMVEDLITAHADVNAVDNHGKSSLHWAAVNNNDAIRA 2661

Query: 862 TCVCTPGFQGP TCANDINECMSPCKNGGKCRNREPGYFCECLDGYSGVNCEENVDDCAS 921
TC QG T AN C G C N + C C +G++G CE ++ C

Sbjct: 34 TCEVQAASQGT TVAN-----VCNGQGTCINS GNSHTCTCAEGFTGSYCETIINHCDP 189

Query: 922 DPCMNGGTCLDDVNSYKCLCKRGFDGNQCQNDVNECENEPCCKNGATCTDYVNSYACTCPP 981
+PC+N C +N Y+C C+ GF G+QCQ D++EC + PC NG TC + +N + C+CP

Sbjct: 190 NPCINAVKCTSGINGYEC DCEAGFQGSQCQLDIDECTSNPCMNGGTCFNAINGFQCSCPR 369

Query: 982 GFRGTTCMENIDE CNIGSCLNGGTCVDGINSYSCNCMAGFTGANCERDIDECVSSPC--K 1039
G G C C+ C N G C GI S++C C G+ G C DI+EC S+PC +

Sbjct: 370 GTLGVLCEVVSSLCDPNPCQNNGHCTSGIGSFTCQCKPGYGGYLCNGDINECASNPCSTE 549

Query: 1040 NGAPCIHGINTFTCQCLTGYTGPTCAQMVDLCQNNPCRNGGQCSQTGTTSK---CLCTSS 1096
C+ GIN F+C C GY G TC+ C NNPC NG C+ C CT+

Sbjct: 550 GSLDCVQGINEFSLCKDGYGDTCSNQASSCSNPNCLNGATCTDNSLEPLRYFCSCCTND 729

Query: 1097 YSGVYCDVPRLSCSAAATWQGEETSLCQHGGQCINS GSTHYCSCRAGYVGSYCETD--- 1153
Y G C++ +C + +C + G+C++ GS YC C GY G+ C ++

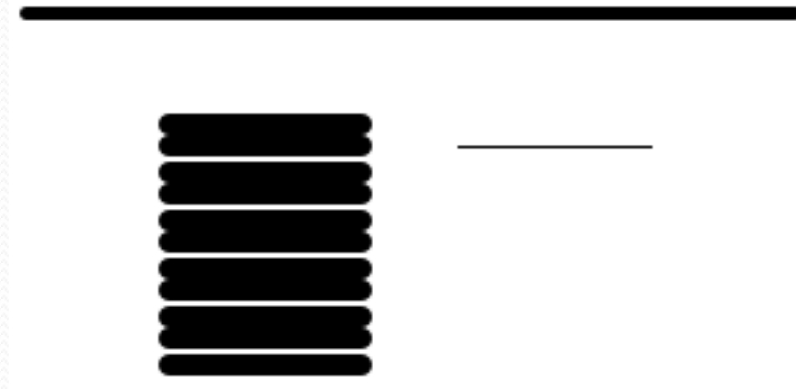
Sbjct: 730 YRGKNCEMEFSTCPSLDM-----ICYNDGKCVD-GSAPYCKCPFGYTGTQCMSNTNT 882

Query: 1154 EDDCASY 1160
E C+SY

Sbjct: 883 EKQCSSY 903

Consideration #2

- BLAST is a *local* alignment algorithm.
- Strong matches are reported first; multiple matches may be out of order between query, subject.



Consideration #3

- BLAST creates gapped alignments.

```
SIRGGGVVDHGISDDESQHSGDAGIS
S+RGGG++ G+S++           D+G
SVRGGGIEIGLSEE-----DSGAE
```

- This means it's totally inappropriate for (for example) primer matching, unless you change the parameters.
- (Who here has actually changed BLAST parameters?)

Consideration #4

- BLAST e-values are database-size dependent.
- BLAST bit scores are not.

Score = 87.0 bits (214), Expect = 5e-16

- You can't technically compare e-values from BLASTs against different databases!

Consideration #5

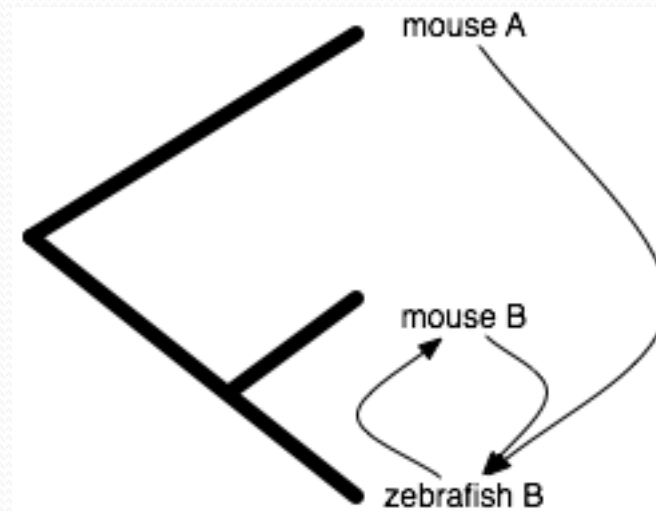
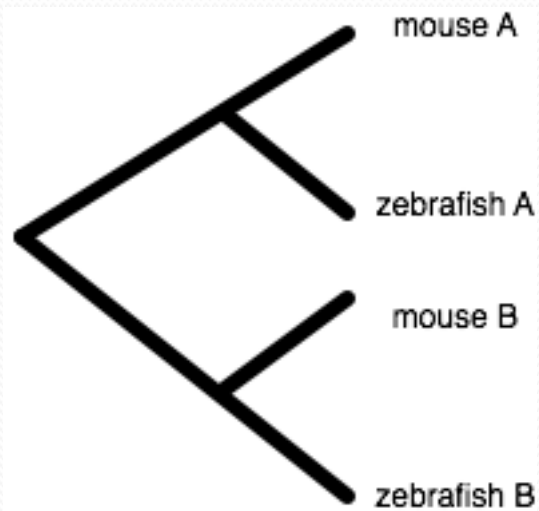
- BLAST uses an heuristic to speed things up: *requires* an **exact** match between 11 bases (DNA) or 3 amino acids in order to start an alignment.

```
Query: 862   TCVCTPGFQGPTCANDINECMSPPCKNGGKCRNREPGYFCECLDGYSGVNCEENVDDCAS 921
            TC      QG T AN          C   G C N   + C C +G++G  CE  ++ C
Sbjct: 34    TCEVQAASQGTTVAN-----VCNGQGTCINSGNSHTCTCAEGFTGSYCETIINHCDP 189

Query: 922   DPCMNGGTCLDDVNSYKCLCKRGFDGNQCQNDVNECENEPCCKNGATCTDYVNSYACTCPP 981
            +PC+N   C   +N Y+C C+ GF G+QCQ D++EC + PC NG TC + +N + C+CP
Sbjct: 190   NPCINAVKCTSGINGYECDCEAGFQGSQCQLDIDECTSNPCMNGGTCFNAINGFQCSCPR 369
```

Consideration #6

- Reciprocal BLAST is a *horrible* (but frequently used) heuristic for “orthology”. Intended for:



...but local alignments cause trouble here!



Considerations #7+

- BLAST implementation is (was?) impenetrable: completely inextensible, very optimized, built on a huge library.
 - Does it have bugs? Nobody knows...
 - V. difficult to embed => difficult to reuse
- BLAST text output format changes frequently and is designed for humans only to read; very hard for computers to parse.



BLAST is also kind of inconvenient

- No good Web interface for uploading your own databases (that I know of).



So, nobody uses BLAST, right?

- Absolutely wrong!
- Biologists love it: it's fast, sensitive, and has a nice Web interface at NCBI.
- Bioinformaticians love/hate it:
 - Biologists => programmers use it by default, and then spend a lot of time correcting for its problems.
 - Computer scientists => biologists often can't escape:
 - Lots of biology behind BLAST; tough to write your own.
 - Biologists *believe* in BLAST, and not your own dinky algorithm.

Digression: it's not BLAST's fault, really.

- Most of the “considerations” I presented are completely obvious and stated clearly all over the place.
- Everybody uses BLAST because it's there, it (mostly) works, and it's trusted by (almost) everyone.
- BLAST use may be starting to break down, though:
 - Doesn't scale to volume of data
 - Default gapping model is inappropriate for short-read mapping
 - Has significant false positive rate on very divergent proteins (metagenomics, “evolutionarily interesting” organisms)



This workshop & BLAST

- We'll be (mis)using BLAST just like everyone else.
- We'll show you how to run BLAST at the command line:
 - Make your own BLAST databases
 - Do searches with many sequences
- We'll show you BLAST output “parsing”
 - Make your *own* spreadsheet of matches
 - Your very own reciprocal BLAST script...



The UNIX command line

- Many computer folk, and most bioinformaticians, work with a text interface to their computers: “the command line”.
- Sort of the granddaddy of all interfaces... think back to teletypes.
- Why?
 - Writing *new* programs is much easier if you write them for the command line (text, no graphics)
 - Simple & flexible (not nec. *good*) user interface design: **none**
 - Simple “pipelining” ability
- Almost all bioinformatics programs work at the command line, or via a Web interface.