

Bioinformatics: Life Sciences on Your Computer

Week 2 Practice Problems - Part 2

These are practice problems for BLAST, and for the related algorithms FASTA and SSEARCH.

1. Do a blastp search of the Swiss-Prot database using gi|10764643.
 - a. What is the E value and accession number of the best match?
 - b. Use FASTA to perform the same search (query = 10764643), limited to Swiss-Prot. What is the e-value of the same result from part A from FASTA?
 - c. Why is that E value different from BLAST?
 - d. Do an SSEARCH (<http://www.ebi.ac.uk/Tools/sss/fasta/>) with the same protein (query = 10764643). Be sure to flip to SSEARCH. Search only the UniProtKB/SwissProt database and use the default matrices and gap penalties. You may need to retrieve the actual sequence from NCBI. What is the E-value of the best match?
2. Run a BLAST search with the *C. elegans* Sma-4 protein (P45897). Limit to the Swiss-Prot database and exclude *Caenorhabditis elegans* (use organism line and check Exclude) from the results.
 - a. Use Taxonomy Reports to find the best hit in chicken (*Gallus gallus*). What is the score and E value?
 - b. Hit Edit and Resubmit. This time limit your results to include only *Gallus gallus* (use organism line and uncheck Exclude). Record the score and E value.
 - c. Explain any discrepancy.

Discussion topic: The best match was to a "mothers against decapentaplegic homolog." How did that protein get its name?

3. If a database size is 36 bits and a query protein sequence that is 256 amino acids in length is used to search that database, then.....
 - a. What is the size of the search space (in bits)?
 - b. If a match is found with a score of 62 bits, then calculate the E Value based on your search space calculated in part A.
4. The *E. coli* RecA protein sequence has the gi number is 67471977. Go to the University of Virginia FASTA server (<http://fasta.bioch.virginia.edu>), call up the FASTA program, and search the database called "Human/Refseq Proteins." Use the default parameters.
 - a. What scoring matrix and gap penalties does the FASTA program use?
 - b. What is the name and accession number of the best scoring sequence (best matches are on top)?
 - c. What do the dots and double-dots (colons) in the alignment mean?
 - d. What is the lowest (worst) reported bits score of the listed matches? Its corresponding E-value?
 - e. Is that a homolog of the query protein?
5. Use the following mouse ITPR1 isoform to find all homologs in cow (*Bos taurus*). Limit to the RefSeq database to avoid redundant matches.

```
>mouse ITPR1 protein isoform 1
MSDKMSSFLHIGDICSLEYAEGSTNGFISTLGLVDDRCVVQPEAGDLNNPPKKFRDCLFKL
CPMNRYSAQKQFWKAAKPGANSTTDAVLLNKLHHAADLEKKQNETENRKLGLTVIQYGNV
```

IQLLHLKSNKYLTV NKRLPALLEKNAMRVTLDEAGNEGSWFYIQPFYKLRSIGDSVVIGD
KVVLNPVNAGQPLHASSHQLVDNPGCNEVNSVNCNTSWKIVLFMKWSDNKDDILKGGDVV
RLFHAEQEKFLTCDEHRKKQHVFRLRTTGRQSATSATSSKALWEVEVVQHDPGRGGAGYWN
SLFRFKHLATGHYLA AEVDPDFEEECLEFQPSVDPDQDASRSRLRNAQEKMVYSLVSVPE
GNDISSIFELDPTTLRGGDSLVRNSYVRLRHLCTNTWVHSTNIPIDKEEEKPVMLKIGT
SPLKEDKEAFAIVPVSPA EVRDLDFANDASKVLGSIAGKLEKGTITQNERRSVTKLLEDL
VYFVTGGTNSGQDVLEVVFSSKPNRERQKLMREQNILKQIFKLLQAPFTDCGDGPMLRLEE
LGDQRHAPFRHICRLCYRVLRRHSQQDYRKNQEYIAKQFGFMQKQIGYDVLAE DTITALLH
NNRKLLEKHITAAEIDTFVSLVRKNREPRFLDYLSDL CVSMNKSIPVTQELICKAVLNPT
NADIL IETKLVLRSRFEFEGVSTGENALEAGEDEEEVWLFWRDSNKEIRSKSVRELAQDAK
EGQKEDRDILSYRYQLNLFARMCLDRQYLAIN EISGQLDVDLILRCMSDENLPYDLRAS
FCRLMLHMHVDRDPQE QVTPVKYARLWSEIPSEIAIDDDYDSSGTSKDEIKERFAQTMEFV
EEYL RDVVCQRFPFSDKEKNKLTFEVVNLARNLIYFGFYNFSDLLRLTKILLAILDCVHV
TTIFPISKMTKGEENKGSNMRSIHG VGELMTQVVLRGGGFLPMT PMAAAPEGNVKQAE P
EKEDIMVMDTKLKIIEILQFILNVRLDYRISCLLCIFKREFDESNSQSSETSSGNSSQEG
PSNVPGALDFEHIEEQAE GIFGGSEENTPLDLDHGGRTFLRVLLH LTMHDYPPLVSGAL
QLLFRHFSQRQEV LQAFKQVQLLVTSQDVDNYKQIKQDLDQLRSIVEKSELWVYKGQGP D
EPM DGASGENEHKKTEEGTSKPLKHESTSSYNRVVKEILIRLSKLCVQESASVRKSRKQ
QQRLLRNMGAHAVVLELLQIPYEKAEDTKMQEIMRLAHEFLQNF CAGNQQNQALLHKHIN
LFLNPGILEAVTMQHI FMNNFQLCSEINERVVQH FVHCIETHGRNVQYIKFLQTIVKAEG
KFIKKCQDMVMAELVNSGEDVLV FYNDRASFQTLIQMMRSE RDRMDENSPLMYHIHLVEL
LAVCTEGKNVYTEIKCN SLLPLDDIVRVVTHEDCIPEVKIAYINFLNHCYVDTEVEMKEI
YTSNHMWKLFENFLVDICRACNNTSDRKHAD SILEKYVTEIVMSIVTTFFSSPFS DQSTT
LQTRQPVFVQLLQG VFRVYHCNWLMP SQKASVESCIRVLSDVAKSRAIAIPVDLDSQVNN
LFLKSHNIVQKTALNWRLSARNAARRDSVLAASRDYRNIIERLQDIVSALEDRLRPLVQA
ELSVLVDVLHRPELLFPENTDARRKCESGGFICKLIKHTKQLLEENEEKLCIKVLQTLRE
MMTKDRGYGEKQIS IDESENAELPQAPEAENSTEQELEPSPPLRQLEDHKRGEALRQILV
NRYYGNI RPSGRRESLTSFGNGPLSPGGPSKPGGGGGGPGSSSTS RGEMSLAEVQCHLDK
EGASNLVIDLIMNASSDRVFHESILLAI ALLEGNTTIQHSFFCRLTEDKKSEKFFKV FY
DRMKVAQQEIKATVTVNTSDLGNKKK DDEVDRDAPSRKKAKEPTTQITEEV RDQLLEASA
ATRKAF TTFRREADPDDHYQSGEGTQATTDKAKDDLEMSAVITIMQPILRFLQLLCENHN
RDLQNFLRCQNNKTNYNLVCETLQFLDCICGSTTGGLGLLGLYINEKNVALINQTLES LT
EYCQGPCHENQNCIATHESNGIDIITALILNDINPLGKKRMDLVLELKN NASKLLLAIME
SRHDSENAERILYNMRPKELVEVIKKAYMQGEVEFEDGENGEDGAASPRNVGHNIYILAH
QLARHNKELQTMLKPGGQVDGDEALEFYAKHTAQIEIVRLDRTMEQIVFPVPSICEFLTK
ESKLRIYYTTERDEQGSKINDFFLRSEDLFNE MNWQKKLRAQPVL YWCARNMSFWSSISF
NLAVLMNLLVAFFYPFKGVRGGTLEPHWSGLLWTAMLISLAIVIALPKPHGIRALIASTI
LRLIFSVGLQPTLFLLGAFNVCNKIIFLMSFVGNCGTFTTRGYRAMVLDVEFLYHLLYLLI
CAMGLFVHEFFY SLLLFDLVYREETLLNVIKSVTRNGRSIILTAVLALILVYLFSIVGYL
FFKDDFILEVDRLPNETAVPETGESLANDFLYSDVCRVETGENCTSPAPKEELLPAEETE
QDKEHTCETLLMCIVTVLSHGLRSGGGVG DVLRKPSKEEPLFAARVIYDLLFFFMVIIIV
LNLI FGVIIDTFADLRSEKQKKEEILKTTCFICGLERDKFDNKTVTFEEHIKEEHN MWHY
LCFIVLVKVKDSTEYTG PESYVAEMIRERNLDWFFPRMRAMSLVSSDSEGEQNELRN LQEK
LESTMKLVTNL SGQLSELKDQMTEQRKQKQRIGLLGHPPHMNVNPQQPA

- How many definite homologs?
- How many of those have inositol 1,4,5-trisphosphate in the description?
- How many are marked PREDICTED?

Answers

1.a. $2e-69$, P75830

b. $1.9e-34$

c. FASTA and BLAST have the same goals but are completely different algorithms.

d. $1.8e-39$

2.a. 184, $4e-50$

b. 184, $2e-52$

c. The score is the same, but the E value changed because the search space of the second search was smaller.

3.a. 44 bits

b. 2^{-18} or $3.8e-6$ (0.0000038)

4.a. Blosom 50, -10/-2

b. DMC1 dosage suppressor of mck1 homolog, NP_008999.2

c. Double-dots for identity, dots for conservative substitutions

d. 29,3, E value = 9.8

e. No conclusion. The E value is too high to infer homology, but non-homology cannot be inferred.

5a. 12 (note that when the E value is shown as 0, that means the E value $< 1e-100$)

b. 3

c. 8