BLAST workshop

Max Carter-Brown
Anglia Ruskin University,
Wellcome Sanger Institute
max.carter-brown@aru.ac.uk

Mark Blaxter
Wellcome Sanger Institute,
University of Edinburgh

Garry Blakey
University of Edinburgh

Contents

1. Outcomes

By the end of this document, you should be able to:

1. ...

2. Introduction

The protein sequences you will be working on are from the genomes of bacteriophage. BPT4 derives from bacteriophage T4 and BPT5 derives from bacteriophage T5. Here are the sequences of the bacteriophage proteins, in single-letter amino-acid code, in FASTA format.

Shell 1: Protein sequences of bacteriophage proteins BPT4 and BPT5

```
>BPT4
MKILNLGDWHLGVKADDEWIRGIQIDGIKQAIEYSKKNGITTWIQYGDIFDVRKAITHKTMEFAREIVQT
LDDAGITLHTIVGNHDLHYKNVMHPNASTELLAKYPNVKVYDKPTTVDFDGCLIDLIPWMCEENTGEILE
HIKTSSASFCVGHWELNGFYFYKGMKSHGLEPDFLKTYKEVWSGHFHTISEAANVRYIGTPWTLTAGDEN
DPRGFWMFDTETERTEFIPNNTTWHRRIHYPFKGKIDYKDFTNLSVRVIVTEVDKNLTKFESELEKVVHS
LRVVSKIDNSVESDDSEEVEVQSLQTLMEEYINAIPDITDSDREALIQYANQLYVEATQ

>BPT5
MRILFSADHHIKLGQDKVPKEWQKRRFLMLGERLNDIFHNHNCDLHIAGGDILDVADPSSEEIELLEQFM
SRLDHPGKIFTGNHEMLTKTISCLYHYAGVINKVTSGKWEVITKPYRSPEFDIVPYDEIHKAKWKPPVSK
LCFTHVRGEIPPHVKPEIDLTKYNCYDTVIAGDLHSYTNSQTIGSTRLLYPGSPLTTSFHRERTKGTNGC
FIIDTDTLKVEWIELGDLPQLIRKTIGAGEEMEPSDYDRVVYEVTGDVVQLKSIKDSDLLDKKINHRVTK
DAKLNLVDLDMLGELELYFREVEKLSQGDIDRILARAAKYVKDYN
```

We want to know what these bacteriophage proteins do - what their functions are, and what role they might play in the biology of bacteriophage T4. One way of doing this is to compare these

proteins to others in public databases, and infer “using electronic annotation” that they have the same function(s) as the sequences they are closely similar to. We rely on the fact that careful biochemical and genetic work on one representative of a family of proteins, generating experimental evidence of function, can be transferred over to other similar proteins. This is a very reasonable prior assumption, as it simply assumes that similar proteins have retained similar functions through evolution.

But what happens when there are no other proteins similar to our sequence?

Further reading...

In this case, previous work has established that BPT4 is distantly related to a family of proteins found in many bacteria, including *Escherichia coli*. The family is usually named **SbcD** after the gene name of the *E. coli* representative, *sbcD*. Another family related to both the bacteriophage proteins and to SbcD is **MRE11**.

BTP4 itself is an exonuclease (removing nucleotides from the end of a nucleotide chain) that plays a role in viral genome replication, DNA recombination, and host DNA degradation.

3. BLAST and psi-BLAST

BLAST (the Basic Local Alignment Search Tool) is a program for comparing a sequence with every member of a database of sequences and presenting you with a list of the most similar. It is very a sophisticated program with a number of parameters that can be altered to optimize the sensitivity of the search it performs.

You will carry out experiments using BLAST, and a derivative of BLAST called psi-BLAST that is useful for sensitive searches to find distant similarities, to find proteins similar to BPT4 and BPT5 in the well-annotated, comprehensive sequence database “SwissProt”.

See [1], [2], and [3] for further information on BLAST and psi-BLAST.

4. The databases at EBML-EBI

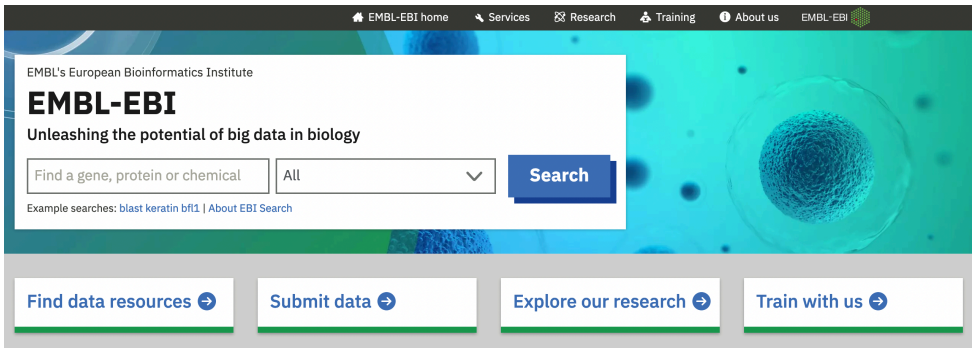

Figure 1: The homepage of the European Bioinformatics Institute (EBI)

The European Bioinformatics Institute (EBI) is a key bioinformatics institute in Europe, based near Cambridge, UK. It is an outstation of the European Molecular Biology Organisation, an international body that supports research across Europe (and beyond).

The EBI support a range of useful bioinformatics resources that we will make use of during this practical. EBI is home to the European Nucleotide Archive (ENA, part of the global database consortium that includes GenBank), SwissProt (a large database of curated protein sequences and

metadata), the protein families database InterPro, the genome databases in the ENSEMBL family, and many other resources.

It is a great place to start on a career in bioinformatics.

5. Getting started

Open up your favourte text processor for making notes (I recommend **Word** if you have it, or if not **Google Docs**). Copy the sequence of the BPT4 and BPT5 proteins (above) into your document.

Now go to the EBI sequence search pages at:

https://www.ebi.ac.uk/jdispatcher/sss

Select **NCBI BLAST "Protein"** (the other option is nucleotide, which we don't want)

The protein similarity search page allows you to enter data (your sequence), choose from a range of parameters (which database to compare to, what sort of search), and then submit this analysis job to a cluster of high-performance computers at EBI. The computers will process your job, and the website returns the result in a variety of formats that, for example, link out to other parts of the services at EBI.

6. Your first search

6.1. Select your database

We will use UniProtKB/Swiss-Prot. This database is strongly curated from the research literature, and we can therefore believe the functional annotations that are attached to the sequences in the database. Deselect the UniProt Knowledgebase if it is selected.

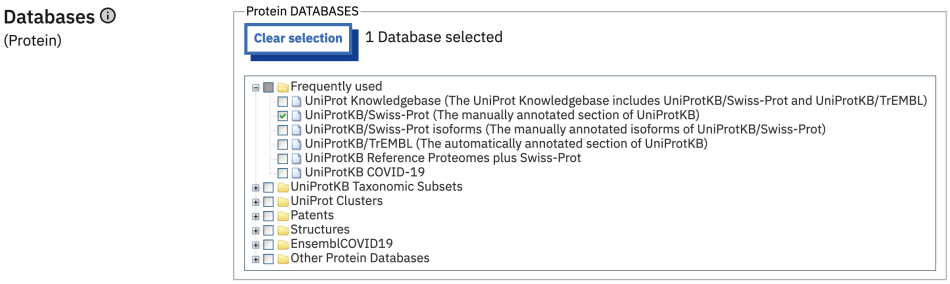

Figure 2: Select the UniProtKB/Swiss-Prot database

6.2. Enter your input sequence

The sequence you search with is known as the **Query**. Copy and paste the BPT4 sequence into the search window.

6.3. Set your parameters

Check that "**blastp**" is selected under "PROGRAM". blastp is the version of BLAST that compares a protein query to a protein database. Click the "**More options**" button. You are presented with a wide range of parameters you can change to customise your search. For the first search, select the **BLOSUM62** matrix.

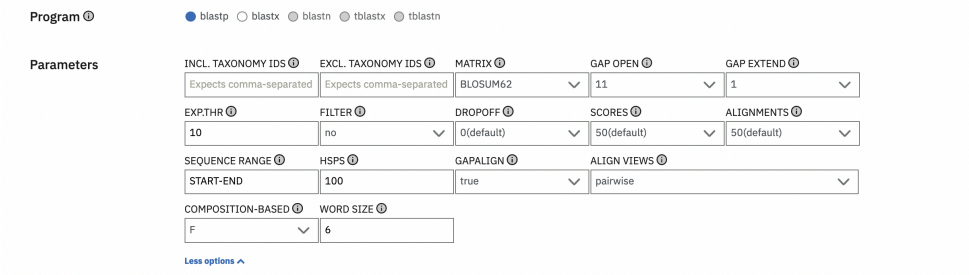

Figure 3: Setting the BLASTP parameters

Further reading...

When aligning two sequences, we must define criteria for the *best* alignment, so we must analyse many/all possible alignments and compute a score to reflect alignment quality. There are usually three criteria: frequency of the amino acid, substitution likelihood, and penalties for opening gaps. People have devised scoring matrices to help programs decide the best alignment. More specifically, scoring matrices (e.g. BLOSUM62) are used by BLAST to assess the quality of each match found in an alignment, assign it a score, and, ultimately, assess the likelihood of that score being a chance similarity or a biologically meaningful match.

6.4. Submit your job

Click “Submit” and wait for the results to appear. The EBI service will first check your sequence and program settings, then return a “holding” page. Wait for your job to complete.

7. Results...

The EBI service formats the results of the search in a number of different ways. From the menu on the Results page, click on “Tool Output” - this is the simple text version of the output generated directly by BLAST.

```
BLASTP 2.12.0+
<cut text>

Database: uniprotkb_swissprot
           571,609 sequences; 206,878,625 total letters

Query= EMBOSS_001

Length=339
                                                      Score     E
Sequences producing significant alignments:          (Bits)  Value

SP:P04521 EXO1_BPT4 Exonuclease eria phage ...  706     0.0
SP:Q96YR6 MRE11_SULTO DNA double-strand bre...  51.6    4e-06
SP:Q97WG9 MRE11_SACS2 DNA double-strand bre...  39.3    0.040
```

The first line shows which version of the program was run (in the example above version 2.12.0+ of BLASTP). (In the output, a reference for the method is then given - we have cut it from the text above.)

The program then tells us which database was searched (in this case “uniprotkb_swissprot”) and how big the database is (0.5 million sequences, 206 million amino acid residues). The name and length of the Query sequence is then given, followed by the hits table.

For each hit, the name of the database sequence or subject is given, followed by the Score that the best alignment between the query and the subject achieved, given the similarity matrix used, and then the E-value of the match.

The Score is derived from the similarity matrix - basically the alignment gets given points for matches, and gets points deducted for mismatches and gaps.

The E-value, or expect value, is an estimate, based on some fearsome statistics, of the number of times one would expect to find a bit score that good, in a database as big as the one that was searched, using the particular query sequence. An E-value of 1 means that about one match as good as the one found would be expected by chance. An E-value of 0.001 (or 1e-03) means that one would expect a match that good in about 1 in 1000 random trials. Very small E-values mean it is less and less likely that the match is due to chance, and more likely that it is due to true biological similarity. The lowest E-value possible is 0.0: basically an identical hit over a long protein. The program does not report tiny E-values between 1e-199 and 0.0 - it just reports 0.0.

Below this table section are the individual pairwise **alignments** between the query and each subject. These are sometimes interesting to look at to visualise how similar your sequence is in the alignment.

The report finishes by summarizing the parameters used, and the core numbers used in the calculation of the Score and E-value. The parameters used can also be reviewed in the EBI web page by clicking on the **Submission Details** tab.

8. Interpreting BLAST results

The sequence you searched with (BPT4) appears at the top of the list, because it is in SwissProt, under the name EXO1_BPT4.

The following 19 proteins contain the strongest scoring local alignments found by BLAST, in descending order of E-value.

We can judge how sensitive and specific the search is by noting:

- The ‘E-value’ for the first hit to a protein with an annotation as MRE11 or SBCD – in this case 4e-06
- How many false positives (proteins that aren’t involved in recombination) there are before the first ‘true’ hit, which is MRE11_SULTO. In this case the answer is zero, and there are no “false positives” until the 7th hit (“hydroxyacylglutathione hydrolase” is not involved in recombination - obviously you have to know what each annotation means to make this decision, but you can find out by looking up this activity in a database of biochemical pathways such as Expasy http://web.expasy.org/pathways/).
- The number of SBCD and MRE11 hits (apart from BPT4 itself) in the top 20. (20 is an arbitrary choice – we could choose the top 50.)

The top hit (other than BPT4 to itself) is to MRE11_SULTO, with an E-value of 4e-06. This is a small number, and reinforced by the appearance of several MRE11 proteins in the top 20, would probably give you confidence (if you were doing this search ‘for real’) that BPT4 is related in function to MRE11 proteins. You will also see that five of the 20 top hits are annotated as ‘‘DNA double-strand break repair protein”, and have a gene name “MRE11”.

The E-value for SBCD_ECOLI is 0.30. This is not a particularly small value, and if you were doing this search 'blind', would not indicate a significant similarity. You would be much less certain about the reality of the apparent similarity to SbcD. There are two additional hits to RAD50 DNA repair proteins, again suggesting some role in DNA repair.

There are other sequences in the list, some with matches as good as those to the MRE11 and SBCD sequences that we know are not truly related to BPT4 (e.g. GLO2_SYNY3).

9. The effect of the substitution matrix

How do different substitution matrices affect the sensitivity and specificity of BLAST searches?

Go to the EBI sequence search pages at **https://www.ebi.ac.uk/jdispatcher/sss**

Select **NCBI BLAST "Protein"**.

Repeat the search with BPT4 twice more, but each time, choose a ***different*** amino-acid scoring matrix from the list in the box called 'MATRIX'.

You should choose one other BLOSUM matrix, and one PAM matrix. If you click on the underlined word MATRIX you will get some background information on the choice of scoring matrix.

Questions

When the search completes, record the same three measures of sensitivity as before:

- What is the E-value of the first hit to a protein with an SBCD or MRE11 identifier?
- What is the number of false positive hits above the first SBCD or MRE11 identifier?
- What is the number of SBCD or MRE11 identifiers in the top 20? (in some cases, it is quite possible that no SbcD or MRE11 sequences will appear in the output lists).

The important thing to note about these results is that the choice of scoring matrix can make a big difference to the output of a search. The BLOSUM62 matrix is the default choice because experience has shown that its use often gives the most reliable results across a wide range of different kinds of proteins – but this is not always the case, and in a critical case it is always important to run searches using several different scoring tables.

10. Are BTP4 and BTP5 related?

Repeat the same three BLASTP analyses above using **BTP5** as the query sequence. Record the results.

Questions

- What is the E-value of the first hit to a protein with an SBCD or MRE11 identifier?
- What is the number of false positive hits above the first SBCD or MRE11 identifier?

- What is the number of SBCD or MRE11 identifiers in the top 20? (in some cases, it is quite possible that no SbcD or MRE11 sequences will appear in the output lists).
- Is BPT5 likely to have the same function as BPT4?
- Does the BLAST algorithm ever find a significant alignment between BPT5 and BPT4?

Bibliography

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.

[2] S. F. Altschul and T. L. Madden, "Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSIBLAST: a new generation of protein database search programs," *Nucleic Acids Res*, vol. 25, no. 17, pp. 3389–3402, 1997.

[3] S. F. Altschul and E. V. Koonin, "Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases," *Trends in biochemical sciences*, vol. 23, no. 11, pp. 444–447, 1998.