Database search 2: Blast

Peter Sestoft
sestoft@dina.kvl.dk
Department of Natural Sciences, KVL

2004-09-03

Literature:

Claverie and Notredame: *Bioinformatics for Dummies*, Wiley 2003, chapter 7

Blast: Basic Local Alignment Search Tool

Problem: You have a sequence (DNA or protein) and want to find other sequences similar to it.

Solution: Use the Blast *program* to search sequence *databases* such as Genbank+EMBL+DDJB, or SWISSPROT.

You can run Blast against the databases at several *servers*, including

- `http://www.ncbi.nlm.nih.gov/BLAST/`
- `http://www.ch.embnet.org/`
- `http://www.ddbj.nig.ac.jp/`

Choose a server not in your own timezone — then it will be less loaded, and faster (e.g. 10 sec versus 28 min).

Or make your own Blast server: download and install Blast and the databases.

The program and the databases are freely available for download (but require lots of disk space and RAM).

Some search programs:

- PubMed: search literature databases by keyword (Escherichia), author, etc.
- Entrez: search literature and sequence databases by keyword (Escherichia), accession number (J01636), etc.
- Blast: search sequence databases by similarity to given sequence (SGCALILAVL)
- FastA: search sequence databases by similarity to given sequence (SGCALILAVL)
- SRS = Sequence Retrieval System: highly integrated search of literature and sequence databases

Some databases:

- Non-redundant (nr) nucleotides from GenBank+EMBL+DDBJ
- Non-redundant (nr) proteins from SwissProt+GenPept(translated GenBank)+PDB+PIR
- SwissProt – proteins with annotations
- PDB – proteins with 3D structure

Some server locations:

- NCBI = National Center for Biotechnology Information in Maryland, USA (NCBI Blast)
- Swiss EMBnet node (European Molecular Biology net)
- SIB = Swiss Institute of Bioinformatics, Geneve, Switzerland (Expasy server)
- . . . or create your own server

BlastN: Search a nucleotide database with a nucleotide query sequence

We want to search for sequences similar to this one (a mouse nucleolin mRNA fragment):

```
  1 gctcttccga gctgctcgct ctccacacgc gccgccgccg taatccgcca ccatggtgaa
 61 gctcgcaaag gctggcaaaa cccacggtga ggccaagaaa atggctcctc ctccaaagga
121 ggtggaagag gatagtgaag atgaagaaat gtcagaagat gaagatgaca gcagtggaga
181 agaggaggtt gtcatccctc agaaaaaagg caaaaaggct accacaaccc cagcaaagaa
241 ggtggttgtt tcacaaacaa aaaaggctgc agttcccaca ccagctaaga aagcagctgt
301 gaccccaggc aaaaaggcag tagccacacc agctaagaaa aacattacac cagccaaagt
361 cattccaaca ccgggtaaga agggagctgc acaagcaaaa gcgttggtac caactcctgg
421 taaaaaggga gctgccactc cagctaaggg ggctaagaac ggtaagaatg ccaagaagga
481 agacagtgat gaggatgaag atgaagagga tgaagaagat agcgatgagg atgaagatga
541 tgaggaagag gatgagtttg agccaccaat agtaaaagga gtgaagccag caaaagcagc
601 tcctgctgct cctgcctccg aggatgagga agatgatgag gatgaagatg atgaggaaga
661 ggaagatgaa gaagaggaag atgactctga ggaagaagtt atggagatca caacagccaa
721 aggaaagaaa actcctgcaa aagttgttcc tatgaaagcc aagagtgtgg ctgaggagga
781 ggatgaggag gaagaggatg aagatgacga ggatgaggat gatgaggaag aggatgacga
841 agatgatgat gaggaagaag aggaggaaga acctgttaaa gcagcacctg gaaaacggaa
901 gaaggagatg accaagcaga aagaagcccc tgaagccaag aaacagaaag tagaaggctc
961 agaaccaact acacctttca atctgttcat tggaaacctt aatccaaaca agtctgttaa
...
```

Reading BlastN output

The output is usually very long and has several sections:

- The search program used: e.g. BLASTN version 2.2.9.
- The database searched in: All GenBank+EMBL+DDBJ+PDB sequences
- A summary graphic display of the structure of the hits (with gaps), best hit first
- A summary list of the hits with their score and E-value, best hit first
- A detailed list of alignments, one for each hit:

  An alignment between a subsequence of the query and a subsequence of the subject (from the database)
- Some internal Blast information. (Will be explained next week).

The significance of a match: Score and E-value

A higher *score* indicates a better match between query sequence and a database sequence.

But in a large database there may be sequences that get a high score by pure chance.

Therefore an *E-value* (expectation value) is computed from the score, taking the database size into account.

For a given score, the larger the database, the larger the E-value.

The E-value is the number of hits with at least this score that one must expect to obtain purely by chance.

Rules of thumb:

- When E≥ 1 the hit is very unreliable, probably not biologically relevant.
- When E≤ 0.0001, that is `1e-4`, the hit is very reliable, probably biologically relevant.

  That may indicate: common origin, and for protein: common structure, common function.

A better, more statistical, explanation of the E-value is given in two weeks' time.

A closer look at an alignment in Blast output

```
>gi|20196543|emb|AL355987.31|  Human DNA sequence from clone RP11-216L13 ...
          Length = 182003
 Score =  121 bits (61), Expect = 3e-24
 Identities = 187/225 (83%), Gaps = 8/225 (3%)
 Strand = Plus / Minus

Query: 93    ccaagaaaatggctcctcctccaaaggaggtggaag-aggatagtgaagatgaagaaatg 151
              ||||||||||||||| |||| ||||||||||  |||| |||| |||||||||| ||||||
Sbjct: 93984 ccaagaaaatggctgctcccccaaaggaggcagaaggagga-agtgaagatgaggaaatg 93926

Query: 152   tcagaagatgaaga---tgacagcagtggagaagaggaggttgtcatccctcagaaaaaa 208
              |||||||||||||   ||||||| ||||||||||     |||||  |||||| | |||
Sbjct: 93925 tcagaagatgaagaagatgacagctgtggagaagagac---tgtcacacctcaggagaaa 93869

Query: 209   ggcaaaaaggctaccacaaccccagcaaagaaggtggttgtttcacaaacaaaaaaggct 268
              ||||| |||||| |  |||||||||| |||||||||   ||||| | ||||||||||||||
Sbjct: 93868 ggcaagaaggctgctgcaaccccagccaagaaggtgacagtttcccgaacaaaaaaggct 93809

Query: 269   gcagttcccacaccagctaagaaagcagctgtgaccccaggcaaa 313
              |||| | |||||||  | ||||||||| |||| || |||||||||
Sbjct: 93808 gcaggtgccacacccaccaagaaagcaactgtcactccaggcaaa 93764
```

Note:

- Accession number, description, score, E-value, percent identities, strand orientation (plus/plus, plus/minus)
- There can be gaps (`---`) in the alignment. Usually only subsequences are aligned: *local* alignment
- Often the alignment consists of several non-contiguous segments.

BlastP: Searching a protein database with a protein query

BlastP is very similar to BlastN in use and output.

An amino acid query sequence (protein):

```
VKLAKAGKTHGEAKKMAPPPKEVEEDSEDEEMSEDEDDSSGEEEVVIPQKKGKKATTTPA
KKVVVSQTKKAAVPTPAKKAAVTPGKKAVATPAKKNITPAKVIPTPGKKGAAQAKALVPT
PGKKGAATPAKGAKNGKNAKKEDSDEDEDEEDEDDSDEDEDDEEEDEFEPPIVKGVKPAK
AAPAAPASEDEEDDEDEDDEEDDDEEEEDDSEEEVMEITTAKGKKTPAKVVPMKAKSVAE
EEDDEEEDEDDEDEDDEEEDDEDDDEEEEEEEPVKAAPGKRKKEMTKQKEAPEAKKQKVE
```

BlastP parameters

The Matrix and Gap costs parameters determine the score of an alignment (and hence the E-value).

Each pair of amino acids, e.g. V/V, or I/V, get a score.

The score is positive if they match well, negative if they match badly.

A *substitution matrix* gives a score for each pair of amino acids.

The score of a gap (`-`) against an amino acid is the *gap cost*.

The BLOSUM50 substitution matrix (symmetric)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0
R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	2	-3
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	-1
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5

TBlast, BlastX, TBlastX

DNA and mRNA can be automatically translated into protein.

This can be done in three forward reading frames:

```
DNA     atggtgaagctcgcaaaggctggcaaaacccacggtgaggccaagaaaatggctcctc

RF1     M  V  K  L  A  K  A  G  K  T  H  G  E  A  K  K  M  A  P
RF2      W  *  S  S  Q  R  L  A  K  P  T  V  R  P  R  K  W  L  L
RF3       G  E  A  R  K  G  W  Q  N  P  R  *  G  Q  E  N  G  S
```

Why do this?

The DNA-sequences TTT and TTC are distinct.

But they code for the same amino acid F.

The 'translated Blast' programs TBlast, BlastX, and TBlastX will consider TTT and TTC the same.

Therefore a TBlastX search with nucleotide query on a nucleotide database will produce more results than a BlastN search with nucleotide query on a nucleotide database.

Namely, TTT matches TTC according to TBlastX but not according to BlastN.

The five Blast search programs

Given a query sequence, Blast searches a database of sequences.

There are different Blast programs for different types of database and query sequence:

	Database sequences	
Query sequence	Nucleotide	Protein
Nucleotide	`blastn` `tblastx` (translate query and database)	`blastx` (translate query)
Protein	`tblastn` (translate database)	`blastp`

Use `blastn` to find DNA — coding or non-coding — similar to the query sequence.

Use `blastp` to find proteins similar to the query protein.

Use `tblastn` to find DNA (a gene) that possibly code for a protein similar to the query protein.

Use `blastx` to find proteins possibly coded for by (parts of) the query DNA sequence.

Use `blastx` to find sequencing errors (frameshifts) in the query DNA sequence — if it codes for a protein.

Use `tblastx` to find genes that code for a protein similar to that coded for by the query DNA sequence.

Low complexity masking

The nucleolin protein alignments found by BlastP contains many X's:

```
 gi|31543315|ref|NP_035010.2| nucleolin [Mus musculus]
 gi|26327461|dbj|BAC27474.1| unnamed protein product [Mus musculus]
 Score =  105 bits (263), Expect = 7e-22
 Identities = 102/286 (35%), Positives = 102/286 (35%)

Query: 1   VKLAKAGKTHGEAKKMAPPPKXXXXXXXXXXXXXXXXXXXXXXXXXIPQXXXXXXXXXX 60
           VKLAKAGKTHGEAKKMAPPPK                          IPQ
Sbjct: 2   VKLAKAGKTHGEAKKMAPPPKEVEEESEDENMSEDEDEKIPEEEVVIPQKKGKKATTTPA 61

Query: 61  XXXXXSQXXXXXXXXXXXXXXXXXXXXXXXXPAKKNITPAKVIPTPGKKGAAQAKALVPX 120
                SQ                        PAKKNITPAKVIPTPGKKGAAQAKALVP
Sbjct: 62  KKVVVSQTKKAAVPTPAKKAAVTPGKKAVATPAKKNITPAKVIPTPGKKGAAQAKALVPT 121

Query: 121 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXFEPPIVKGVXXXX 180
                                                            FEPPIVKGV
Sbjct: 122 PGKKGAATPAKGAKNGKNAKKEDSDEDEDEEDEDDSDEDEDDEEEDEFEPPIVKGVKPAK 181

Query: 181 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXVMEITTAKGKKTPAKVVPMKAKSVAX 240
                                              VMEITTAKGKKTPAKVVPMKAKSVA
Sbjct: 182 AAPAAPASEDEEDDEDEDDEEDDDEEEEHDSEEEVMEITTAKGKKTPAKVVPMKAKSVAE 241
```

BlastP automatically X-es out subsequences that have *low complexity*.

Low complexity means that the subsequence has too many 'too common' amino acids.

Such a subsequence is likely to match other sequences by pure coincidence, with no biological significance.

The main substitution matrix families

- The PAM matrices (Dayhoff et al. 1978) are created from sequences with $\geq$85% identity, and thus evolutionarily close. The matrix PAM1 represents 1 % accepted mutations, and PAM250 = PAM1^{250} etc. are derived by 'extrapolation'. Higher numbers represent higher evolutionary distance.
- The BLOSUM matrices (Henikoff and Henikoff 1992) are created from BLOCKS, manual multiple alignments of evolutionarily more distant sequences. Thus BLOSUM62 is built from such multiple alignments with $\geq$62 % identity, and BLOSUM45 from multiple alignments with $\geq$45 % identity. Thus lower numbers represent higher evolutionary distance. The BLOSUM matrices perform better than PAM when aligning distant sequences.
- The GONNET matrices (Gonnet et al. 1992) are created like the PAM matrices, but from a broader material, and are better for computing evolutionary distance. The GONNET matrices are used by default in the ClustalW multiple alignment program (since version 1.8).

According to the Blast documentation,

> [...] the BLOSUM62 matrix is among the best for detecting most weak protein similarities. For particularly long and weak alignments, the BLOSUM45 matrix may prove superior. [...] The BLOSUM series does not include any matrices [...] suitable for the shortest queries, so the older PAM matrices may be used instead.

Blast uses BLOSUM62 with gap opening cost $d = -11$ and gap extension cost $e = -1$ by default.

Next week we shall see what that means.