Second Course on Introduction to Sequence Analysis

Nucleotide Analysis Part I

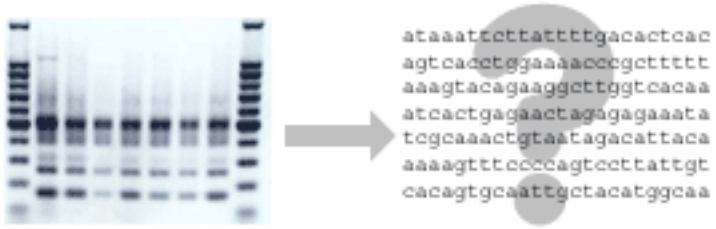
Osvaldo Graña
ograna@cnio.es

CNIO Bioinformatics Unit

07/April/2009

Introduction

Suppose you has discovered an unknown fragment of DNA deduced from a gel, which you have had sequenced. You will want to find out as much as you can about it.



Introduction

- Is it contaminated with vector sequences?
- Is it an already known gene?
- Is it related to any other genes either by having a common ancestor?
- Is it similar in function to other genes via convergent evolution?
- What could the protein sequence be for this nucleotide fragment if it is translated and what might this be like?

Checking for vector contamination

- DNA/RNA from a biological source are usually inserted into a cloning vector (e.g. Plasmid or phage) so that they can be cloned.
- Sequencing of such constructs frequently produces raw sequences that include segments derived from vectors.
- Also transposable elements from the cloning host (generally bacteria or yeast) can insert into the cloned DNA/RNA while the clone is being propagated and will then be sequenced, as can DNA/RNA contaminants.

Checking for vector contamination

- Vector sequences can be found in the sequence databanks for various reasons.
- The most common cause is that the submitters forget to remove them.
- Another is that they are submitted to the databanks because they are, after all, vectors which others might find useful.

Checking for vector contamination

- The EBI provides a vector screening service called BLAST2_EVEC.
- It is based on NCBI BLAST2 and uses the latest implementation of the BLAST algorithm and a special sequence databank known as EMVEC to check your sequences for vector contamination.
- EMVEC is a sequence collection that contains complete and fragmented sequences of vectors. The main use of this library is to provide support to sequencing teams with the process of removing vector sequences from their data prior to submission to the EMBL, GenBank or DDBJ databanks.

EMVEC: http://www.ebi.ac.uk/Tools/blastall/vectors.html

How to check for vector contamination

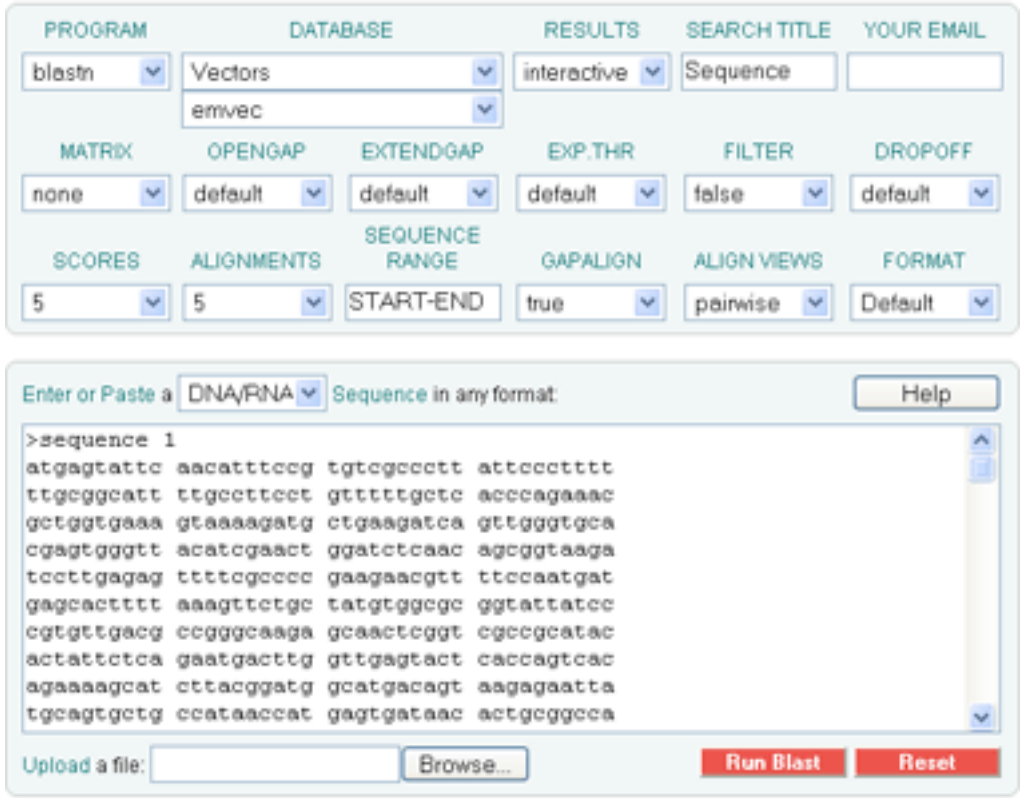
- Using BLAST2 EVEC you can check unknown sequences for contamination.
- We are going to check two sequences that are in the link below:

http://ubio.bioinfo.cnio.es/people/ograna/public_html/introductionToSequenceAnalysis/checkingForVectorContamination/

- Open the EMVEC database query web page to start the check

http://www.ebi.ac.uk/Tools/blastall/vectors.html

How to run a check for vector contamination



Running a job with BLAST2 EVEC

•The sequence is entered into the textbox in FASTA format, which consists of a one-line header starting with a “>” symbol, followed by the sequence name

•“Interactive” is chosen so that I will have the results delivered to the browser as soon as they are available. Alternatively you can chose “email” and fill in your email and have the results delivered via email

•It is possible to change the default title

•The BLASTN program is used, which is designed to search a nucleotide query sequence against a DNA databank, in this case the emvec database

•The number of scores (hits to the database) and the number of alignments of these against the query sequence is returned

Running a job with BLAST2 EVEC

- With the “alignments” option we can select the number of returned alignments

- Other options are just left on “default”

- So ….. which sequence is contaminated?

Interpreting the results of BLAST2 EVEC

• Click the button “BLAST result” to see the results

• You should have noticed that sequence 1 and sequence 2 appear to be contaminated with vectors

• A closer look is required to see if these are genuine hits, or a result of finding a loose match to a similar functional domain to one found in the vectors database. This is due to the contents of the database, and the e-values allowing these hits to pass through

• In the case of sequence 1, you will see that the parameter “Value” for all these hits **is zero**, so all these hits are exact matches and are definitly a result of vector contamination (no gaps in the alignment)

• In the case of sequence 2, you will see that the parameter “Value” for all these hits is always positive. As a rule of thumb, for a first look at these, if the value is not “to the power of” (e.g. 1.2x10-7) it is not likely to be significant

Interpreting the results of BLAST2 EVEC

• Here you will see that the alignments are very short, and also contain gaps in the alignment (No “I” symbol joining the upper and lower sequence). This is obviously not a good match and can be ignored which is probably a result of similar functional domains being present in the query sequence and in the entry in the vector database.

• You should now be convinced that sequence 1 is contaminated with vectors and sequence 2 is clean of vectors.

• Sequence 1 will need to have the sequence code trimmed before it becomes useful.

At the bottom of the page are statistical results as follows:

•The name of the database that was searched
•The last update date of the database used in the search
•The size of the database

Interpreting the results of BLAST2 EVEC

•Lambda and *K* are statistical parameters used in calculating the significance of the gapped alignment score. They are determined on the basis of the alignment scores of random sequences using a combination of scoring matrix and gap penalties (Bioinformatics, D. W Mount, page 251).

•H is the relative entropy that shows the ability of the entire substitution matrix to discriminate related from unrelated sequences.

•The scoring matrix that was used

•The gap creation/gap extension costs

•Values for various steps in the search process leading to the identification of HSPs (High-scoring element pair).

•Calculated values for database size, query size, and so on used in the search.

Interpreting the results of BLAST2 EVEC

•Underneath the score list, you can view the pairwise alignments of sequences from the score list against your query sequence. Note that matching sequences are connected with a "|" symbol. As this is a perfect match, all of the nucleotides are connected with a "|" symbol, mismatches would be connected with a blank space. A gap would be represented with a "-" symbol. Thus a sequence alignment can be represented in the format:

Score= normalized bits (raw bits)

The score of the alignment, called *score* or *bit score*, is the sum of log odds scores of each matching amino acid pair in the alignment less gap penalites; the raw score in bits is also shown in parentheses. The expect value E (E-value) of chance matches of unrelated sequences from a database of this size and the percent identities in the alignment is shown.

```
>EM_VEC:CVPJB8 X98612.1 Artificial cloning vector pjb8
          Length = 5410

 Score = 1707 bits (861), Expect = 0.0
 Identities = 861/861 (100%)
 Strand = Plus / Minus


Query: 1    atgagtattcaacatttccgtgtcgcccttattccctttttgcggcattttgccttcct 60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 5197 atgagtattcaacatttccgtgtcgcccttattccctttttgcggcattttgccttcct 5138


Query: 61   gtttttgctcacccagaaacgctggtgaaagtaaaagatgctgaagatcagttgggtgca 120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 5137 gtttttgctcacccagaaacgctggtgaaagtaaaagatgctgaagatcagttgggtgca 5078


Query: 121  cgagtgggttacatcgaactggatctcaacagcggtaagatccttgagagttttcgcccc 180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 5077 cgagtgggttacatcgaactggatctcaacagcggtaagatccttgagagttttcgcccc 5018
.
.
.
etc
```

Understanding an EMBL entry

EMBL entries (as below) in the database are structured to be usable by human readers as well as by computer programs. Each entry in the database is composed of lines. Different types of lines, each with its own format, are used to record the various types of data which make up the entry. Some entries will not contain all of the line types, and some line types occur many times in a single entry. As noted, each entry begins with an identification line (ID) and ends with a terminator line (//).

```
ID   X98612; SV 1; circular; other DNA; STD; SYN; 5410 BP.
XX
AC   X98612;
XX
DT   04-JUL-1996 (Rel. 48, Created)
DT   08-JUL-1998 (Rel. 56, Last updated, Version 3)
XX
DE   Artificial cloning vector pjb8
XX
KW   cloning vector.
XX
OS   unidentified cloning vector
OC   other sequences; artificial sequences; vectors.
XX
RN   [1]
RA   Matthews A.P.;
RT   ;
RL   Unpublished.
XX
RN   [2]
RP   1-5410
RA   Matthews A.P.;
RT   ;
RL   Submitted (21-JUN-1996) to the EMBL/GenBank/DDBJ databases.
RL   A.P. Matthews, The Sanger Centre, Informatics, Wellcome Trust Genome
RL   Campus, Hinxton, Cambs CB10 1SA, UK
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..5410
FT                   /organism="unidentified cloning vector"
FT                   /mol_type="other DNA"
FT                   /db_xref="taxon:45196"
FT   misc_feature    1..5410
FT                   /note="cloning vector pjb8"
XX
SQ   Sequence 5410 BP; 1261 A; 1444 C; 1351 G; 1354 T; 0 other;
     gatccggaat tctcatgttt gacagcttat catcgataag ctttaatgcg gtagtttatc        60
     acagttaaat tgctaacgca gtcaggcacc gtgtatgaaa tctaacaatg cgctcatcgt       120
     catcctcggc accgtcaccc tggatgctgt aggcataggc ttggttatgc cggtactgcc       180
     gggcctcttg cgggatatcg tccattccga cagcatcgcc agtcactatg gcgtgctgct       240
```

Understanding an EMBL entry

- The ID (IDentification line) line is always the first line of an entry. The general form of the ID line is:

Term	Primary accession number	Sequence Version Number	dataclass	molecule	Data Class	Taxonomic division	sequencelen (Base Pair
e.g.	X98612	SV 1	circular	other DNA	STD	SYN	5410 BP

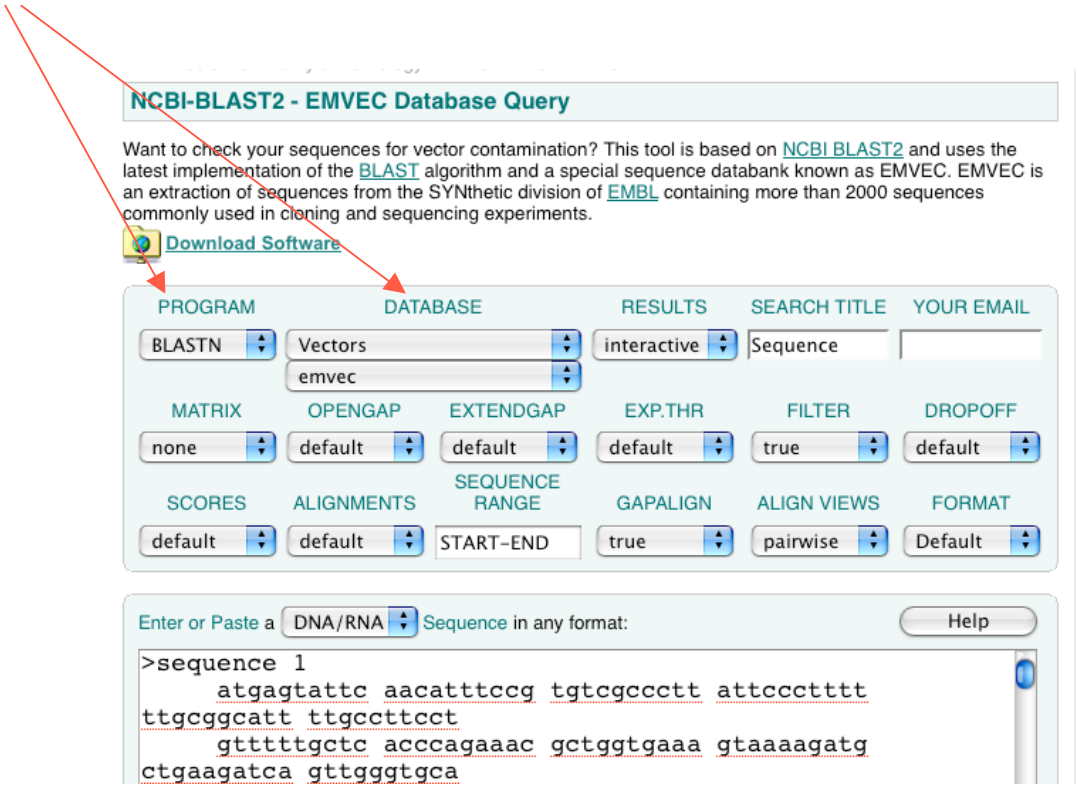
- The XX line contains no data or comments. It is used instead of blank lines to avoid confusion with the sequence data lines.
- The AC (Accession Number) line lists the accession numbers associated with this entry.
- The DT (DaTe) line shows the date/release number of creation, date/release number of the last modification of the entry and the version number.
- The DE (DEscription) lines contain general descriptive information about the sequence stored.
- The KW (KeyWord) lines provide information which can be used to generate cross-reference indexes of the sequence entries based on functional, structural, or other categories deemed important. The keywords chosen for each entry serve as a subject reference for the sequence, and will be expanded as work with the database continues. Often several KW lines are necessary for a single entry.
- The OS (Organism Species) line specifies the preferred scientific name of the organism which was the source of the stored sequence.
- The OC (Organism Classification) lines contain the taxonomic classification of the source organism.
- The RN (Reference Number) line gives a unique number to each reference citation within an entry.
- The RC (**R**eference **C**omment) line type is an optional line type which appears if the reference has a comment.
- The RP (Reference Position) line type is an optional line type which appears if one or more contiguous base spans of the presented sequence can be attributed to the reference in question.
- The RX (**R**eference **Cross**-reference) line type is an optional line type which contains a cross-reference to an external citation or abstract database.
- The RA (Reference Author) lines list the authors of the paper (or other work) cited.
- The RT (Reference Title) lines give the title of the paper (or other work).
- The RL (Reference Location) line contains the conventional citation information for the reference.
- The PR (**PR**oject) line shows the International Nucleotide Sequence Database Collaboration (INSDC) Project Identifier that has been assigned to the entry.
- The CC lines are free text comments about the entry, and may be used to convey any sort of information thought to be useful.

- The FH (Feature Header) lines are present only to improve readability of an entry when it is printed or displayed on a terminal screen. The lines contain no data and may be ignored by computer programs.
- The FT (Feature Table) lines provide a mechanism for the annotation of the sequence data. Regions or sites in the sequence which are of interest are listed in the table. A complete and definitive description of the feature table is given here.
- The SQ (SeQuence header) line marks the beginning of the sequence data and gives a summary of its content.
- The sequence data lines has lines of code starting with two blanks. The sequence is written 60 bases per line, in groups of 10 bases separated by a blank character, beginning in position 6 of the line. The direction listed is always 5' to 3'
- The // (terminator) line also contains no data or comments. It designates the end of an entry.

```
ID   X98612; SV 1; circular; other DNA; STD; SYN; 5410 BP.
XX
AC   X98612;
XX
DT   04-JUL-1996 (Rel. 48, Created)
DT   08-JUL-1998 (Rel. 56, Last updated, Version 3)
XX
DE   Artificial cloning vector pjb8
XX
KW   cloning vector.
XX
OS   unidentified cloning vector
OC   other sequences; artificial sequences; vectors.
XX
RN   [1]
RA   Matthews A.P.;
RT   ;
RL   Unpublished.
XX
RN   [2]
RP   1-5410
RA   Matthews A.P.;
RT   ;
RL   Submitted (21-JUN-1996) to the EMBL/GenBank/DDBJ databases.
RL   A.P. Matthews, The Sanger Centre, Informatics, Wellcome Trust Genome
RL   Campus, Hinxton, Cambs CB10 1SA, UK
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..5410
FT                   /organism="unidentified cloning vector"
FT                   /mol_type="other DNA"
FT                   /db_xref="taxon:45196"
FT   misc_feature    1..5410
FT                   /note="cloning vector pjb8"
XX
SQ   Sequence 5410 BP; 1261 A; 1444 C; 1351 G; 1354 T; 0 other;
     gatccggaat tctcatgttt gacagcttat catcgataag ctttaatgcg gtagtttatc        60
     acagttaaat tgctaacgca gtcaggcacc gtgtatgaaa tctaacaatg cgctcatcgt       120
     catcctcggc accgtcaccc tggatgctgt aggcataggc ttggttatgc cggtactgcc       180
     gggcctcttg cgggatatcg tccattccga cagcatcgcc agtcactatg gcgtgctgct       240
     agcgctatat gcgttgatgc aatttctatg cgcacccgtt ctcggagcac tgtccgaccg       300
     ctttggccgc cgcccagtcc tgctcgcttc gctacttgga gccactatcg actacgcgat       360
     catggcgacc acacccgtcc tgtggatctg cctcgctggc ctgccgcagt tcttcaacct       420
     cccggcgcag cttttcgttc tcaatttcag catcccttte ggcataccat tttatgacgg       480
     cggcagagtc ataaagcacc tcattaccct tgccaccgcc tcgcagaacg ggcattccct       540
     gttcctgcca gttctgaatg gtacggatac tcgcaccgaa aatgtcagcc agctgctttt       600
     tgttgacttc cattgttcat tccacggaca aaaacagaga aaggaaacga cagaggccaa       660
```

Detailed information about BLAST2 EVEC options

•To get a detailed explanation of the different options click directly over the option name.



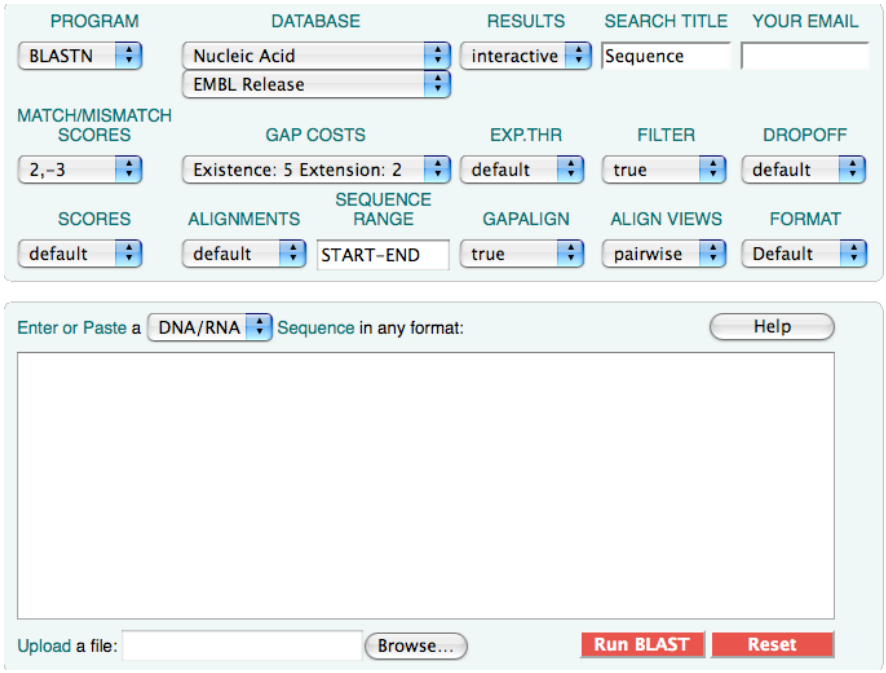
Conclusion

- We now know that sequence 2 is clear of vector contamination.
- As NCBI-BLAST2 has basically the same options as BLAST2 EVEC we can now perform a BLAST search with this tool, to look for regions of sequence similarity in this query sequence with known sequences in the EMBL Nucleotide Sequence Database.

Running a NCBI-BLAST2 similarity search

•BLAST stands for Basic Local Alignment Search tool.
•It is aimed to find regions of sequence similarity, which will yield functional and evolutionary cluess about the structure and function of your query sequence.
•We will now be able to find out what sequence Sequence 2 is, as this sequence is a real entry in the EMBL Nucleotide Sequence Database [EMBL release], and see what sequences it is in fact similar to.

We are going to use the form:
http://www.ebi.ac.uk/Tools/blastall/nucleotide.html



Sequence 2:
http://ubio.bioinfo.cnio.es/people/ograna/public_html/introductionToSequenceAnalysis/checkingForVectorContamination/

Running a NCBI-BLAST2 similarity search

So… what entry in the EMBL Nucleotide Sequence Database is sequence 2?

Results of the NCBI-BLAST2 similarity search

We can see several hits:

- Some of the high scores reflect associations of our query sequence to several uses in different pattents
- *The first one is supposed to be our hit*

NCBI-BLAST2 Results

SUBMISSION PARAMETERS			
Title	Sequence	Database	em_rel
Sequence length	4145	Sequence type	n
Program	NCBI-blastn	Version	2.2.19 [Nov-02-2008]
Matrix	blastn matrix:2 -3	Gap extension penalty	2
Open gap penalty	5		

Show Annotation | BLAST Result | MView | XML | SUBMIT ANOTHER JOB

Clear all | Check all | Invert selection | Reset | Show Alignments | Download | fasta

Alignment	DB:ID	Source	Length	Score	Identity%	Positives%	E()
1	EM_MUS:X14897	Mouse fosB mRNA	4145	7240	98		0.0
2	EM_PAT:I15766	Sequence 1 from patent US 5470736.	4144	7240	98		0.0
3	EM_PAT:AR096729	Sequence 1 from patent US 6008323.	4144	7240	98		0.0
4	EM_PAT:DL235420	NOVEL COMPOSITIONS AND METHODS FOR CANCER.	4145	7240	98		0.0
5	EM_PAT:DL205759	NOVEL COMPOSITIONS AND METHODS FOR CANCER.	4145	7240	98		0.0
6	EM_PAT:DD014118	METHOD FOR EXAMINING ISCHEMIC CONDITIONS.	4145	7240	98		0.0
7	EM_PAT:AX695339	Sequence 966 from Patent WO03008583.	4145	7240	98		0.0
8	EM_PAT:AX306253	Sequence 1004 from Patent WO0188188.	4145	7240	98		0.0
9	EM_MUS:AC148988	Mus musculus BAC clone RP23-85B15 from chromosome 7, complete sequence.	154177	3762	97		0.0
10	EM_MUS:AC118017	Mus musculus chromosome 7, clone RP23-457C1, complete sequence.	196427	3762	97		0.0
11	EM_HTG:AC073784	Mus musculus clone RP23-412G23, WORKING DRAFT SEQUENCE, 46 unordered pieces.	235302	3762	97		0.0
12	EM_PAT:DL235419	NOVEL COMPOSITIONS AND METHODS FOR CANCER.	26993	3732	96		0.0
13	EM_PAT:DL205758	NOVEL COMPOSITIONS AND METHODS FOR CANCER.	26993	3732	96		0.0

Why don't we get 100% of sequence identity?

We find the answer looking at the alignments

BLAST algorithm

•BLAST stands for Basic Local Alignment Search Tool. It is used to compare a novel sequence with those contained in nucleotide and protein databases. The emphasis of this tool is to find regions of sequence similarity. These can yield clues about the structure and function of this novel sequence, and about its evolutionary history and homology with other sequences in the database.

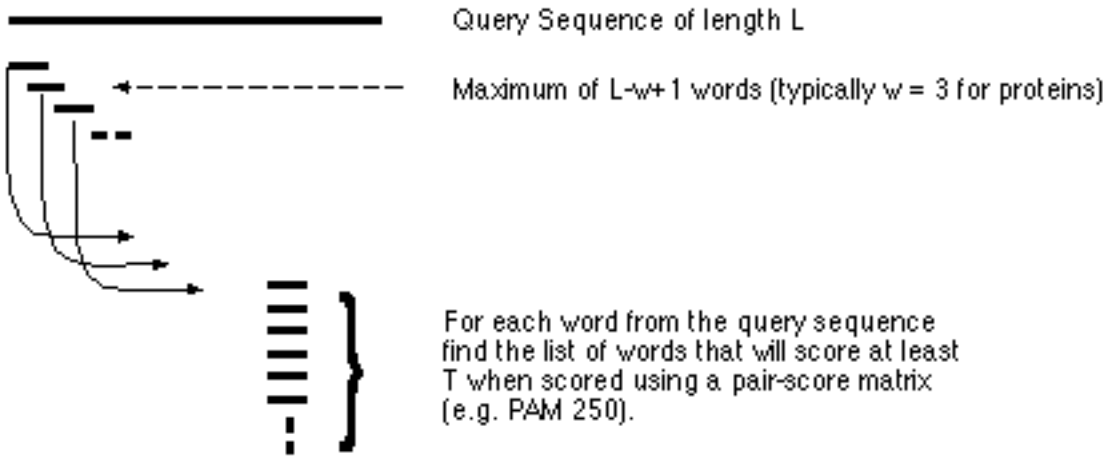
•The fundamental unit of the BLAST algorithm output is the High-scoring Segment Pair (HSP). An HSP consists of two sequence fragments of arbitrary but equal length whose alignment is locally maximal and for which the alignment score meets or exceeds a threshold or cutoff score. A set of HSPs is defined by two sequences, a scoring system, and a cutoff score; this set may be empty if the cutoff score is sufficiently high. In the programmatic implementations of the BLAST algorithm described here, each HSP consists of a segment from the query sequence and one from a database sequence.

•The approach to similarity searching taken by the BLAST programs is first to look for similar segments (HSPs) between the query sequence and a database sequence, then to evaluate the statistical significance of any matches that were found, and finally to report only those matches that satisfy a user-selectable threshold of significance.

•There are 2 main versions of BLAST available at the EBI, namely WU-BLAST2 and NCBI-BLAST2. These are distinctly different software packages, although they have a common lineage for some portions of their code, so the two packages do their work differently and obtain different results and offer different features.

BLAST algorithm

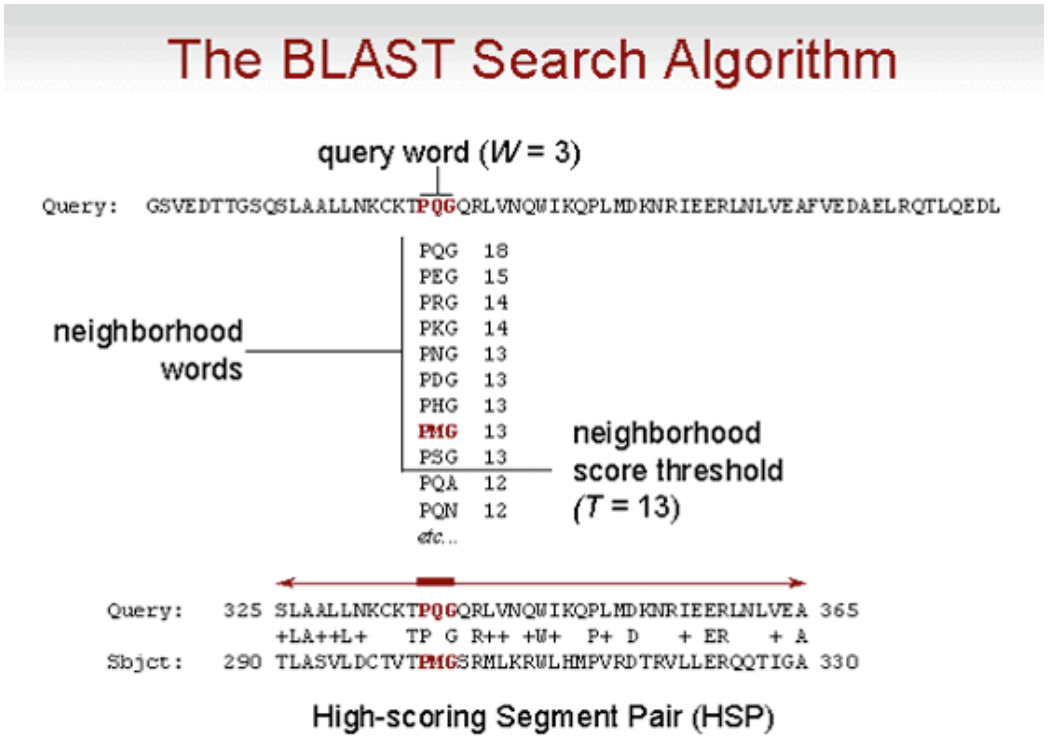
(1) For the query, find the list of high scoring words of length w



BLAST algorithm

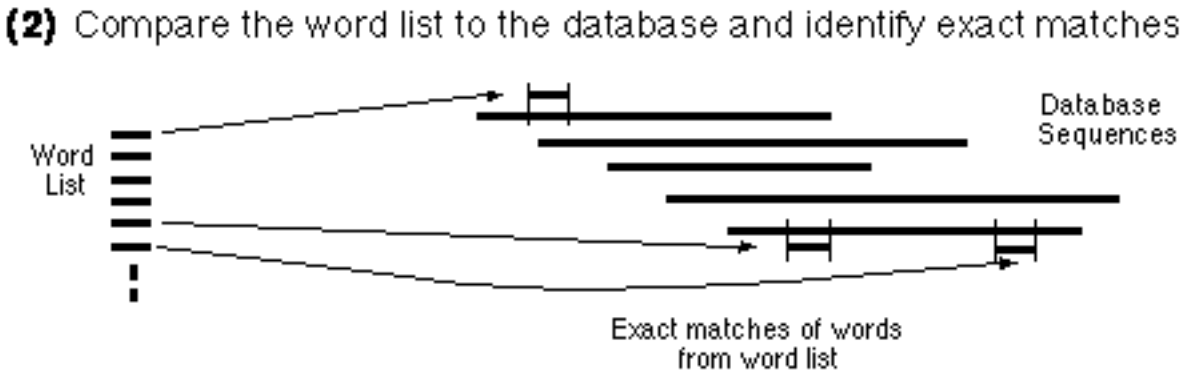
Selected words for W=3:

GSV
SVE
VED
EDT
..
…
…..
PQG



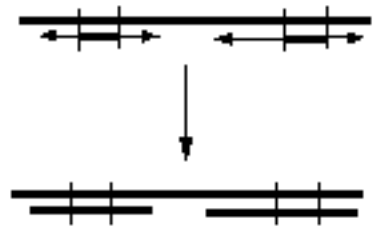
Source: NCBI

BLAST algorithm



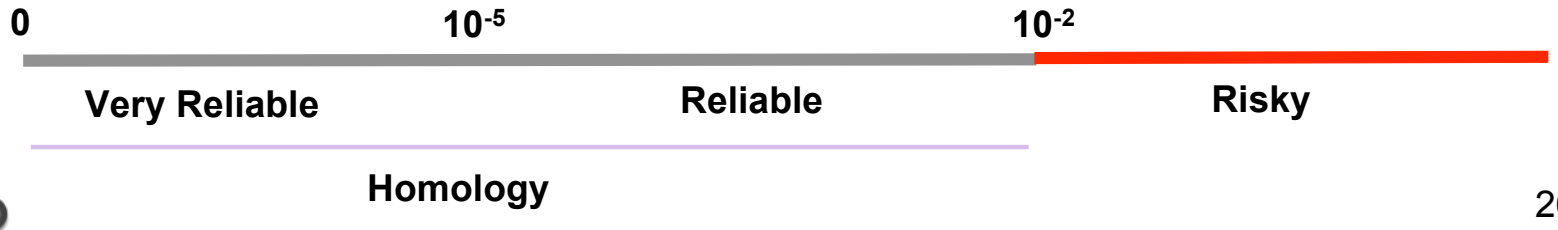
BLAST algorithm

(3) For each word match, extend the alignment in both directions to find alignments that score greater than a threshold of value S



Maximal Segment Pairs (MSPs)

Figure from Barton, G.J. Protein Sequence Alignment and Database Scanning (University of Oxford, Laboratory of Molecular Biophysics)



FASTA similarity search - Introduction

- FASTA provides a rapid way to find short stretches of similar sequence and any sequence in a database.

- This program is much more sensitive than BLAST programs, which is reflected by the length of time required to produce results. FASTA produces optimal local alignment scores for the comparison of the query sequence to every sequence in the database.

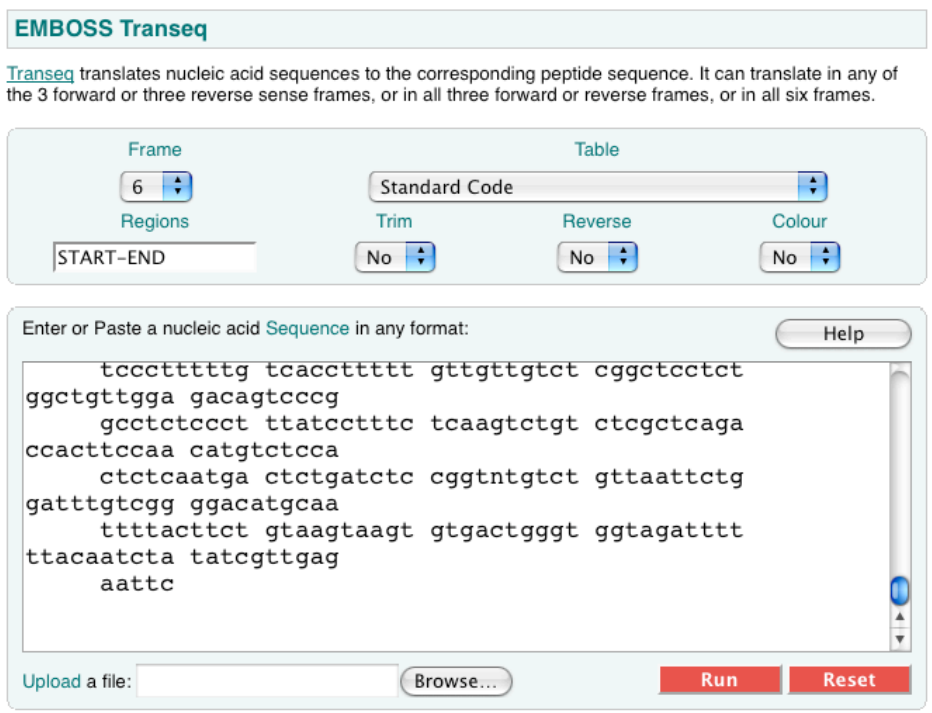
http://www.ebi.ac.uk/Tools/fasta/

Translating our sequence with EMBOSS Transeq

Transeq translates nucleic acid sequences to the corresponding peptide sequence. It can translate in any of the 3 forward or three reverse sense frames, or in all three forward or reverse frames, or in all six frames.

We will use as an example the previous sequence 2 (http://ubio.bioinfo.cnio.es/people/ograna/public_html/introductionToSequenceAnalysis/checkingForVectorContamination/sequence2.txt)

The Transeq form here:
http://www.ebi.ac.uk/Tools/emboss/transeq/



EMBOSS Transeq results

Transeq Results	
Frame	6
Translation table	Standard (0)
Regions	START-END
Trim	no
Reverse	no
Color	no
Transeq output	transeq-20090403-1046312999.output
SUBMIT ANOTHER JOB	

```
>sequence_1 2
INSYFDTHQNSHLENPLFVTKYRRLGHI* ITEN* REILSQTVIDITSIKVSPVLIVILHS
AIATWQTSVA* KSKQKQTKERSHKSKTVQQLIVQTKPLNLSLGSLK* IFLHLAVYDLTFT
EHKPSLKSAVEILK* KGLLIE* H* IP* RKKNLNIKITD* NSLAN* HTNMQLGNHAVFYLR
KHKTKLLK* F* RG* NPGPLPGC* N* TSGEF* SLQF* NLLKSP* LQLIKSSARNSDTPLES
ITVYEHVGCYQPQSI* QGCSVMNLIQREHA* AASTACWATFLPVVTQSIRVLGVSEAHAA
PRHCPAGFWAGSDPRVAPRETDRAWTFRRYSGGLKGIWDLAEGTCIETWAVLRTGD* ASP
SSALWRRVRSTPDSHLIPLGHSLGFPATSAWSQGPPCAQGNVSSFSRRLRLRLPV* LITL
RRVSVPVFGGLLRQSTHRRRLPGVRRSRGNARLLRANGHRNHNQPGSSVARATHPHLFHG
PVPGAATGLPASSC* PL* HARNQLLNPRPECLQHWRGKRKWWAFNQHNHQWTCVCPSSQS
QA* KTPRRDTYPRRRRKAKGSQRAEQAGCS* VQEPSEGADRSTSGGN* SA* RGKGRAGVG
DRRAAKREGTPGVCPGGPQTGLQDPLRRGAGARPAGRGERFARVNIR* GRRLRLAAAAPS
TTPPALPEQPRRTPQPDGFSLYTQ* SSSPRRPLPRC* PFVHFLVCPHLPGGLRVRRRPTH
QRQRAAVRPAELALPSCSVNSLDKQNKQTRKEQGGGR* GGEGRKQSGGVCVDPLTLLSDH
LPPLPSDMTEGPPLCFVLRLWFSVPRRDRRAGDFGDRGWGGDGHPSCISLSCYFNPTSGD
RWLAGWVGWGATPTFGVLREAGGERVLSVGCRVG* GRAGMHLQRDPTRK* QHRPVLLFPH
PPIHPQGCRVTKIALFCSLGP* LINLTFPRGYNLLLDELSPRLREVDAPFGSLLTPLPAD
SKM* TPI* LLSLSLLGKLAQVGFFSSSATEPPPNSGPLPPLCSIMLCPSHPHPHPRRPWP
SSLGLTGFGQQGALRRPSCWSALYCE* VVGLLGAPDGIDPQPSKTFPGPPLLPLASSLPL
TGS* TRKDDHDASRWPSCSGPRLFLFKSFAFPSLGRQLLPTLGAPHPLTEVEAIFREVFR
AEALAPLSSIFESPNIFGLAYLRGG* VPTIPLHPIPSVPKTSSVPSLQLSPRENPTSQIS
IF* YWGDGPYRPSPVLHGTFHTLSWALGSKPNPKPHPQLFIPFLVPKKHLYLLCINKYII
YECACVCVCVRACVRASFLVFKCAVEFKIASGDLSQTFWLSLFVTFLLLSRLLWLLETVP
ASPFILSQVCLAQTTSNMSPLSMTLISGXSVNSGFVGDMQFYFCK* V* LGGRFFTIYIVE
NS
>sequence_2 2
* ILILTLTKIVTWKTRFL* QSTEGLVTFKSLRTREKYYRKL* * TLHP* KFPQSLL* YCTV
QLLHGKLV* HRSQSKNKPKKGATRVKLFNS* * FKLSH* IYHWDR* NESSYTLQCMI* LLQ
NTSQV* NQQ* RY* NEKVC* * SNIKYPEGKKT* ISK* LIKIHLQISTRICNLEIMQCFI* E
NIKQNY* NSFRGGKIQVLCQDAKIRLQGNFEVFNFETY* KAHDYS* LRAVHATVTRL* RA
LLCMNMLAATSHSQFNKAAQS* T* YRESTPRQQAQLAGPLSSLS* HNQSVYLVYLKRTLH
RGTARRVSGRGAIPASPPVKPTEPGLSGGTAAV* RGSGILQRELASKLGQFSEPETKLPR
AAHFGDVSGLLRTRISFHSAIALASRRPQRGHRGPPVPREMFQAFPGDYDSGSRCSSSPS
AESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTAITTSQDLQWLVQPTLISSMA
QSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGSGGPSTSTTTSGPVSARPARA
```

Looking at these results, are we able to say what are the coding frames?

EMBOSS Transeq results

Not yet….. we will be able to check this out latter!

Help - EMBOSS-Transeq

Transeq help: http://www.ebi.ac.uk/Tools/emboss/transeq/transeq_frame.html

TABLE

Which Genetic Code table to use. These are kept synchronised with those maintained at the NCBI's Taxonomy Browser.

REGIONS

Which regions of the user's DNA molecule are to be translated.

TRIM

Remove "*" and "X" (stop and ambiguity) symbols from the end of the translation.

REVERSE

Choose this option if you wish to reverse and compliment your sequence.

COLOUR

Choose this option if you wish to colour your translation.

Help - EMBOSS-Transeq

SEQUENCE INPUT WINDOW

You can cut and paste or type a nucleotide sequence into the large text window. A free text (raw) nucleotide sequence is simply a block of characters representing a DNA/RNA sequence. You may also paste a sequence in GCG, FASTA, EMBL, GenBank, PIR, NBRF or Phylip format. Partially formatted sequences will not be accepted.

Copying and Pasting directly from word processors may yield unpredictable results as hidden/control characters may be present. Adding a return to the end of the sequence may help certain applications understand the input. Some examples of common sequence formats may be seen here.

UPLOAD A FILE

You may upload a file from your computer which containing a valid nucleotide sequence in any format (GCG, FASTA, EMBL, GenBank, PIR, NBRF or Phylip) using this option. Please note that this option only works with Netscape Browsers or Internet Explorer version 5 or later. Some word processors may yield unpredictable results as hidden/control characters may be present in the files. It is best to save files with the Unix format option to avoid hidden windows characters.

Thanks for your attention !

I would like to thank also the effort done by the **2Can** initiative at the EBI. Some of the slides shown in this tutorial were selected from the **2Can Support Portal**.