Module 6

BIOINFORMATICS

Jérome Gouzy and Daniel Kahn

Local organiser: Peter Mergaert

This document will present some tools and databases that can be used for *in silico* sequence analysis, including tools which will be used during the practical exercise. The aim is to provide a primer on how to interpret DNA and protein sequences and how to exploit large scale EST data, with applications to *M. truncatula* sequences. A very useful introduction to bioinformatics can be found in the *Trends guide to bioinformatics*[1] published in 1998.

A more thorough inventory of tools and databases can be accessed from the following two well organised web sites:

DEAMBULUM Infobiogen	`http://www.infobiogen.fr/services/deambulum/english/menu.html`
GENOMEWEB HGMP	`http://www.hgmp.mrc.ac.uk/GenomeWeb/`

1. Gene detection in genomic sequences

One of the most difficult yet relevant tasks for sequence analysis is the accurate recognition of coding sequences from genomic sequences. This is a complex task primarily because of gene splicing. In this field the best programs use species specific statistical models to discriminate exons from introns, and combine these models with specialised tools tailored to detect gene starts, splice donor and splice acceptor sites. In addition it becomes more and more useful to take into account the wealth of heterogeneous information available. In particular it appears useful to exploit homology with genes already characterised, as well as Expressed Sequence Tags (ESTs) from *M. truncatula* or from other higher plants.

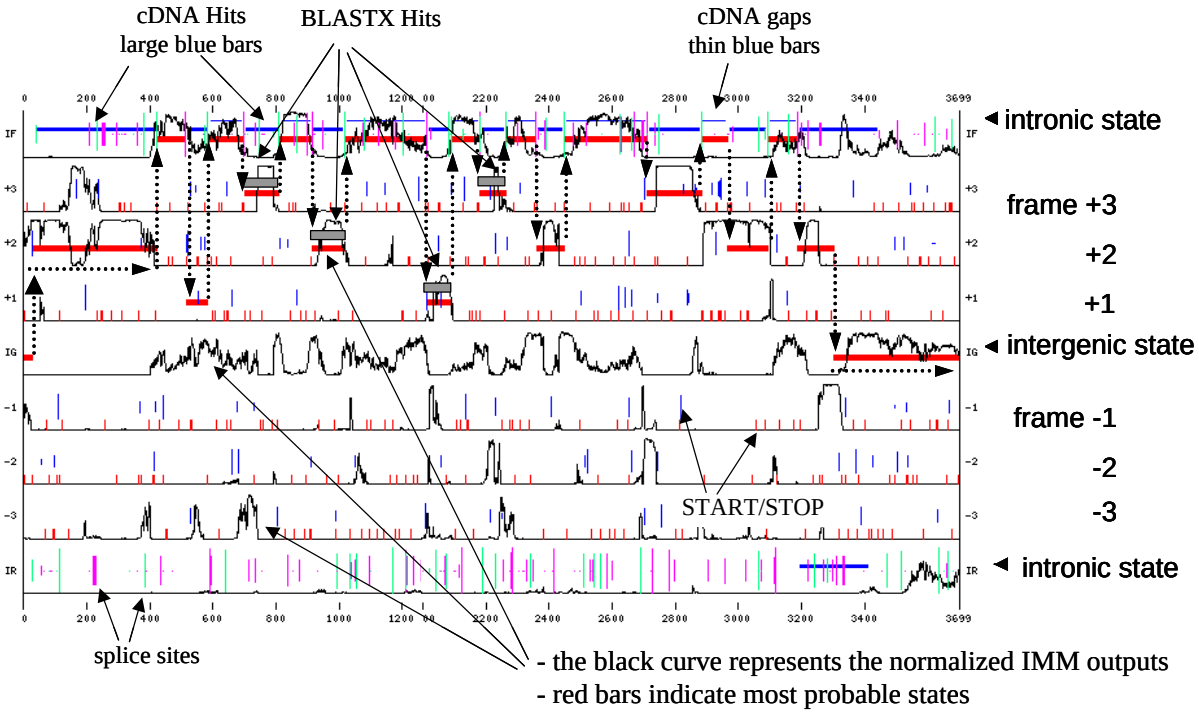
EuGène : A Eukaryotic Gene finder that combines several sources of evidence

For plant gene detection, we recommend the use of the EuGène program[2] developed by Thomas Schiex. This program uses Interpolated Markov Models (IMMs) up to the eighth order to describe the various states for genomic DNA: exons on frames 1, 2, 3, introns and intergenic sequences. It incorporates the output of signal prediction software for gene starts and splice sites. And it exploits homology with cDNA or EST databases using BLASTN, and with protein databases using BLASTX.

During the course we will use EuGène for gene detection in *M. truncatula* genomic sequences. A typical EuGène output is shown on the figure below.

[1] `http://journals.bmn.com/supp/browse/issue?jcode=supp&supcode=1998%4005`

[2] Schiex, T., Moisan, A. & P. Rouzé, 2001, *Lecture Notes Comput. Sci.* **2066**, 111-125 `http://www.inra.fr/bia/T/schiex/Export/LNCS-EuGene.pdf`



Such an output can be readily interpreted in terms of potential gene start, exon coding potential, exon homology and matches with existing ESTs. The influence of external information (homology and ESTs) can also be probed in order to assess the robustness of the prediction.

Useful links

GENSCAN	`http://genes.mit.edu/GENSCAN.html`
EuGène	`http://www.inra.fr/bia/T/EuGene/`
GeneMark	`http://opal.biology.gatech.edu/GeneMark/`

2. Protein sequence annotation

Assigning protein function on the basis of sequence analysis is possible to some extent and can prove extremely useful in order to generate testable hypotheses. The general chart for protein sequence analysis can be delineated as follows:

- identify proteins sharing extensive homology, i.e. that can be aligned over their entire length;
- identify protein domains;
- look for functional motifs;
- identify other structural features such as signal sequences, transmembrane segments, internal repeats and stretches with low complexity.

Homology search

'Homology is the relationship of two characters that have descended, usually with divergence, from a common ancestral character'[3]. Homology is termed *orthology* when divergence follows speciation, *paralogy* when divergence follows duplication.

It is important not to confuse homology and similarity. Similarity between any two protein sequences can be defined as the percentage of identical aminoacids after optimal alignment. Such an optimal alignment can always be obtained, and similarity computed, for any pair of proteins, whether homologous or not. In favourable cases the homology relationship can be deduced from sequence similarity. Beware that reciprocally low sequence similarity is no sufficient proof of non-homology.

Interpreting sequence similarity requires a critical view about :

- statistical significance: with what probability would a similar or better alignment be obtained by chance alone?
- the quality of previous annotation of homologous sequences; is experimental evidence available to support function? Beware that numerous erroneous functional predictions have crept into databases as a consequence of high throughput automated annotation. These errors tend to propagate!

The most popular tool for similarity search is the BLAST program which comes in five different flavours:

- BLASTP compares an amino acid query sequence against a protein sequence database;
- BLASTN compares a nucleotide query sequence against a nucleotide sequence database;
- BLASTX compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database;
- TBLASTN compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands);
- TBLASTX compares the six-frame translation of a nucleotide query sequence against the six-frame translation of a nucleotide sequence database (usually not recommended).

BLAST scores are sorted by increasing expectation (E-value). E-values correspond to the expected number of matches with a score better or equal to the observed score. They depend on the size of the database and on the size and composition of the query sequence. E-values are estimated using an asymptotic approximation for the statistics of high score occurrences. Although some underlying approximations are questionable (like the assumed independence of amino-acid frequencies), E-values are usually a good indicator of statistical significance.

One notable exception occurs when the query sequence contains segments with low sequence complexity (very short repeats or compositionally biased segments). These low complexity segments will match with any segment in the database with a similar bias in a statistically significant way which however is usually not related to homology. It is therefore useful to filter out these segments using for instance the SEG program.

[3] Fitch, W.M., 2000, *Trends Genet.* **16**, 227-231

It can also be useful to iterate BLASTP searches with the best matches in order to capture more distantly related sequences. An efficient program for doing so is PSI-BLAST[4]. In this program the best matches are aligned to generate a Position Specific Scoring Matrix (PSSM) which is used as a query in the next iteration. Beware however that PSSMs are extremely sensitive tools which can in some instances confuse homology search if used indiscriminately. Indeed any false positive arising during one iteration will be amplified during the next iterations.

Finally it should be kept in mind that because of protein modularity, homology will frequently concern one or a few domains, not necessarily the entire protein. To interpret a protein sequence correctly it is therefore necessary to consider its domain arrangement and to compare it with the domain arrangements of putative homologues.

Protein domains and motifs

Protein domains can be identified using specialised databases such as PROSITE profiles, ProDom, PFAM, SMART and TIGRFAMS.

The PROSITE database uses PSSMs to capture the diversity of sequences within domain families. It can be searched using the ProfileScan utility[5].

The ProDom database automatically clusters homologous domains. These can be searched for sequence similarity using BLASTP[6]. Domain clustering provides for a non redundant output of domains matching the query sequence.

PFAM, SMART and TIGRFAMS use Hidden Markov Models (HMMs) to capture sequence diversity. HMMs are extremely sensitive tools well suited to the detection of remote homologues. They are however costly in terms of CPU consumption.

Protein 'motifs' or 'signatures' are localised features which can be described with patterns such as for example `[AC]-x-V-x(4)-{ED}`, which translates into:

```
[Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}.
```

Protein motifs can be associated to active sites, modification sites or be diagnostic of protein families or subfamilies. They are used in the PROSITE and PRINTS database. Motifs however cannot capture the full range of sequence variation and are therefore less sensitive than PSSMs or HMMs.

Combination of methods : InterProScan

The above protein motif and family databases are federated in the InterPro database which includes a shared documentation for each family. A convenient procedure for searching all the above databases is to use the InterProScan server maintained at the European Bioinformatics Institute[7].

[4] http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/psi1.html
[5] http://hits.isb-sib.ch/cgi-bin/PFSCAN
[6] http://www.toulouse.inra.fr/prodom/doc/blast_form.html
[7] http://www.ebi.ac.uk/interpro/scan.html

Signal peptides, targeting peptides and transmembrane helices

Signal peptides, some targeting peptides and transmembrane helices can be detected efficiently.

For the detection of signal peptides we recommend PSORT or SignalP[8]. Chloroplast transit peptides can be efficiently detected using ChloroP[9]. Both SignalP and ChloroP are based on appropriately trained neural networks.

For the detection of transmembrane helices, a recent benchmark showed that the HMM based TMHMM[10] program outperforms other available methods[11]

Links

BLAST	Http://www.ncbi.nlm.nih.gov/BLAST/
Course	Http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html
Tutorial	Http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html
FAQ	Http://www.ncbi.nlm.nih.gov/BLAST/blast_FAQs.html
PROSITE	Http://www.expasy.org/prosite/
PFAM	Http://www.sanger.ac.uk/Software/Pfam/
ProDom	Http://www.toulouse.inra.fr/prodom.html
InterPro	Http://www.ebi.ac.uk/interpro
MetaFam	Http://metafam.ahc.umn.edu/
PSORT	Http://psort.nibb.ac.jp
TMHMM, ChloroP, SignalP	Http://www.cbs.dtu.dk/services

References

R.Apweiler, T.K.Attwood, A.Bairoch, A.Bateman, E.Birney, M.Biswas, P.Bucher, L.Cerutti, F.Corpet, M.D.R.Croning, R.Durbin, L.Falquet, W.Fleischmann, J.Gouzy, H.Hermjakob, N.Hulo, I.Jonassen, D.Kahn, A.Kanapin, Y.Karavidopoulou, R.Lopez, B.Marx, N.J.Mulder, T.M.Oinn, M.Pagni, F.Servant, C.J.A.Sigrist, E.M.Zdobnov. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites, Nucleic Acids Research vol 29(1):37-40

Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL (2000): The Pfam Protein Families Database, Nucleic Acids Research 28:263-266

Corpet F, Servant F, Gouzy J, Kahn D (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. Nucleic Acids Res. 28:267-269.

Hofmann K., Bucher P., Falquet L., Bairoch A. (1999) *The PROSITE database, its status in 1999* Nucleic Acids Res. 27:215-219

[8] Http://www.cbs.dtu.dk/services/SignalP/
[9] Http://www.cbs.dtu.dk/services/ChloroP/
[10] Http://www.cbs.dtu.dk/services/TMHMM/
[11] Moller, S., Croning, M.D.R. and R. Apweiler, 2001, *Bioinformatics* **17**:646-653

Junker V.L., Apweiler R., Bairoch A. Representation of functional information in the SWISS-PROT data bank. Bioinformatics 15:1066-1067(1999).

Silverstein, K.A.T., E. Shoop, J.E. Johnson, A. Kilian, J.L. Freeman, T.M. Kunau, I.A. Awad, M. Mayer and E.F. Retzel. (2001) "The MetaFam Server: a comprehensive protein family resource," Nucleic Acids Research, **29**:49-51.

➔ the NAR Database Issue (first issue of each year) contains a wealth of information about biological databases.

3. EST clustering and annotation

EST projects generate numerous partial sequences which can be potentially very informative. Moreover when a sufficient number of overlapping ESTs is available it becomes possible to reconstitute cDNA sequences for entire genes. This in turn allows to predict the full extent of the gene product. It is also invaluable for interpreting the exon structure in genomic DNA (see chapter 1). However in order to exploit the data it is necessary to assign each EST to a specific gene, which relies on an adequate EST clustering methodology.

Clustering methodology

Classically EST clustering follows four successive steps:

1. Pre-processing
   - Deletion of vector sequences and polyA tails.
   - Masking of DNA repeats.
   - Detection and removal of chimeric ESTs.

2. Overlap detection.
Overlaps are systematically computed using programs such as BLASTN on the entire set. This generates a set of relationships between individual ESTs.

3. Clustering.
ESTs are grouped into clusters by transitive closure. At this stage clusters may correspond to more than one gene. Conversely one gene may be split into more than one cluster when ESTs do not overlap. Therefore clusters should be interpreted with caution.

4. Cluster resolution and contig assembly.
Programs such as `cap3` or `phrap` are used to extract homogeneous subsets and assemble the corresponding sequences into putative spliced gene sequences.

Difficulties of EST clustering

EST clustering remains a difficult task for bio-informatics. Current automated cluster resolution procedures are not completely reliable. Results are usually quite sensitive to the choice of parameter values. Also EST clustering is inevitably confused by the presence of multigene families, by alternative splicing, by DNA repeats (particularly in UTRs) or by artefacts such as chimeric cDNA clones. Therefore the input of an expert is required in order to validate or modify the clusters resulting from automated analysis, before the clusters can be usefully annotated.

Links

TIGR Gene Indices	http://www.tigr.org/tdb/mtgi/
MtDB (CCGB)	http://www.medicago.org/
MtC	http://medicago.toulouse.inra.fr/MtC
Unigene (not yet available for Mt)	http://www.ncbi.nlm.nih.gov/UniGene/
cap3	http://genome.cs.mtu.edu/cap/cap3.html
Phrap	http://www.phrap.org/

References

Burke J, Davison D, Hide W. (1999) d2_cluster: a validated method for clustering EST and full-length cDNAsequences. Genome Res. 1999 Nov;9(11):1135-42.

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J. (2000) An optimized protocol for analysis of EST sequences. Nucleic Acids Res. 2000 Sep 15;28(18):3657-65.

Parsons JD, Rodriguez-Tome P. (2000) JESAM: CORBA software components to create and publish EST alignments and clusters. Bioinformatics. Apr;16(4):313-25.

*Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J. (*2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res. Jan 1;29(1):159-64.

4. Medicago specific resources

Here we provide a list of the major Web sites dedicated to *Medicago truncatula*, which are both complementary and partly redundant.

```
http://www.medicago.org/
http://chrysie.tamu.edu/medicago/mtdb/
http://www.genome.ou.edu/medicago.html
http://www.ncgr.org/mgi/index.html
http://www.tigr.org/tdb/mtgi/
http://www.genome.clemson.edu/affiliated_cugi/medicago/
http://medicago.toulouse.inra.fr/
```

5. Database Appendix

General DNA databases

EMBL	EBI EMBL	http://www.ebi.ac.uk/embl
GenBank	NCBI	http://www.ncbi.nlm.nih.gov/
DDBJ	DNA Databank of Japan	http://www.ddbj.nig.ac.jp/

Specialised DNA databases

DbEST Expressed Sequence Tags	NCBI	http://www.ncbi.nlm.nih.gov/dbEST/ dbEST (Nature Genetics 4:332-3;1993) is a division of GenBank that contains sequence data and other information on "single-pass" cDNA sequences, or Expressed Sequence Tags
DbSTS Sequence Tagged Sites	NCBI	http://www.ncbi.nlm.nih.gov/dbSTS/
dbSNP Single Nucleotide Polymorphism	NCBI	http://www.ncbi.nlm.nih.gov/SNP/
GSS Genome Survey Sequence	NCBI	http://www.ncbi.nlm.nih.gov/dbGSS Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences
HTGS High Throughput Genomic Sequences	NCBI	http://www.ncbi.nlm.nih.gov/HTGS/ Unfinished High Throughput Genomic Sequences: phases 0, 1 and 2 (finished, phase 3 HTG sequences are in nr)
nt/nr	NCBI	All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences). No longer "non-redundant".

Protein Sequence Databases

SWISS-PROT	SIB/EBI	http://www.expasy.org/sprot/ High quality annotated protein sequence database
TrEMBL	EBI/SIB	Http://www.expasy.org/sprot/ Translations of EMBL sequences, not yet in SWISS-PROT.
PIR	PIR-International	Http://www.mips.biochem.mpg.de/proj/protseqdb/
nr	NCBI	All non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF

Database query systems

Entrez	UNIX WWW X-WINDOWS WIN9X MACOSWW	`http://www.ncbi.nlm.nih.gov/Entrez/`
SRS Sequence Retrieval System	UNIX WWW	`Http://www.lionbio.co.uk/publicsrs.html` `http://srs.ebi.ac.uk`
ACNUC	UNIX WWW X-WINDOWS WIN9X MACOS	`Http://pbil.univ-lyon1.fr/databases/acnuc.html`

EMBOSS	UNIX
	WWW
		`http://www.hgmp.mrc.ac.uk/Software/EMBOSS/`
		`http://www.pasteur.fr/cgi-bin/biology/bnb_s.pl?query=EMBOSS`
		`http://www.genopole-lille.fr/fr/plateaux_techniques/bioinformatique/softs/`

GCG	UNIX WWW	`http://www.accelrys.com/about/gcg.html`

6. Time schedule

Entry point into the exercises:

`http://medicago.toulouse.inra.fr/Mt/Embo/index.html`

Gene detection in genomic sequences (1h30)

Practising with EuGène
`http://medicago.toulouse.inra.fr/Mt/Embo/Genomic.html`

Protein sequence annotation (1h)

EST clustering and annotation (1h)

`http://medicago.toulouse.inra.fr/Mt/Embo/ESTClusters.html`

Practise on student examples (1h)