EuGene'Hom web interface

April 2, 2003

Contents

1 Filling in the form

In the standard query form, it is only required to fill the sequence and homology frames.

1.1 Sequence frame

The query sequence in FASTA format or raw DNA can be specified :

- by giving the local path and filename of the sequence file (the 'Browse' button can be helpful to locate it),
- or by pasting the sequence in the proposed frame.

1.2 Homology frame

The set of homologous sequence(s) in (multi-)FASTA format can be specified :

- by giving the local path and filename of the sequence(s) file (the 'Browse' button can be helpful to locate it),
- or by pasting the sequence(s) in the proposed frame.

The sequence(s) are preprocessed using `formatdb` and the query sequence searched against the database obtained. The proteic substitution matrix used for the TblastX search must be chosen in the scrolling list : BLOSUM80 (the default matrix), BLOSUM45, PAM30, PAM70. If no sequence is provided, EuGene'Hom will run using the proteic model alone which is not advised for good prediction.

1.3 Options frame (available on the expert form only)

1. The first option allows to specify if the proteic coding model is used (default) or not.
2. The second option allows to force non partial gene prediction (not set by default). This forces predictions to start and end in intergenic mode and therefore prevents the occurrence of partial gene structures. Useful if you know a single (or only complete) gene appears in the sequence.

1.4 User information frame (available on the expert form only)

User can provide information integrated to perform the prediction. They are provided in a small specifically designed langage. The language can contain two types of statements. Statements on signals (translation start, splice sites) and on the sequence itself (coding, non coding...). Information on signals (stop/start/splice sites) use the following syntax :

```
<type> <strand> <position> <value>
```

where :

- `<type>` can be start, stop, acceptor, donor
- `<strand>` can be f for forward and r for reverse
- `<position>` is a position on the sequence
- `<value>` is the cost associated with the signal. It can range from 0.0 to 1.0. A value of 0.0 means that no site of this type exists at this position. The cost will override any existing cost.

Informations on regions have the following syntax :

```
<type> <strand>{<phase>} [<start>..<end>] <value>
```

where

- `<type>` can be exon, intron, utr3, utr5, intergenic
- `<strand>` if either f for forward or r for reverse
- `<phase>` is only used for exons and introns. The phase of an exon indicates the position of the codons wrt. to the start/end of the sequence (modulo 3, same as reported by BlastX). The phase of an intron specifies how the codon is split by the splice site : between 2 codons, after 1 nucleotide or after 2. Basically, if you don't care of the intron "phase" just put the same information for all 3 phases.
- `<start>` and `<end>` are positions on the sequence that indicate the scope of the information.
- `<value>` is a score that can take any floating point value plus the extra value "infinity" and "-infinity". This cost will be added to the existing cost. A positive score helps predicting something consistent with the information, a negative one prevents from predicting something consistent with the information. Thus a cost of "-infinity" effectively forbids the prediction.

1.5 GFF input frame (available on the expert form only)

If desired, an existing annotation for the sequence can also be plotted on the graphical output (cf. graphical output section) with the EuGene'Hom prediction (on the same graph). An annotation in a GFF format can be specified :

- by giving the local path and filename of a file to upload (the "Browse" button can be helpful to locate it),
- or by directly pasting the annotation in the proposed frame.

1.6 Output parameters frame (available on the expert form only)

This frame configures the output format and contents (does not affect prediction, only text or graphical output) :

- *Verbose* : display additional running information (release number, modules loading, readed files, length of the optimal path and name of graph file) before output prediction when set and 'Text' format is asked.
- *Format* : define the content of the output 'Text' for textual output, 'Image' for graphical output or 'Both' for both.
- *Text format* : define the format for the textual output 'Standard' for EuGene'Hom format or 'GFF'.

- *Number of nucleotides per image* : number of nucleotides presented per image (> 1000). The default value is 6,000 nuc. (or the size of the sequence if it is shorter).
- *Number of overlapping nucleotides between images* : number of nucleotides ($< 15\%$ of the number of nucleotides per image) that overlap between successives images (for long sequences). The default value is heuristically determined.
- *Position of the first plotted nucleotide* : specify the nucleotide from which the graphical representation is started.
- *Position of the last plotted nucleotide* : specify the nucleotide from which the graphical representation is stopped.
- *Horizontal resolution (pixels)* : define the horizontal definition in pixel of the graphical representation. The larger the resolution, the finer the image but the larger the files. The default value is 900. A maximum of 1,200 and a minimum of 100 are enforced on the Web interface.
- *Vertical resolution (pixels)* : define the horizontal definition in pixel of the graphical representation. The larger the resolution, the finer the image but the larger the files. The default value is 300. A maximum of 500 and a minimum of 100 are enforced on the Web interface.
- *Coding score smoothing window width* : all the statistics (coding/non coding curve) presented on graphical representation are smoothed and normalized using a sliding window. You can sepcify the half-size of the window here. By default the window used is $97 = 1 + 2 * 48$ nucleotides wide. Modify this here.
- *Offset* : allows to offset the nucleotide position. That is, the prediction for nucleotide at position *i* of the given sequence is noted as nucleotide *i+offset*. Useful to perform prediction on an extracted sequence without loosing the original position.
- *Score normalisation* : EuGene'Hom always compares the relative likelihood of being non coding or coding in the 6 phases. By default, the graphical output normalizes these 7 scores but you can ask for non normalization or to normalize independantly each coding phase w.r.t. the non coding hypothesis. This does not affect the prediction, only the graphical output.

2 Reading the output

2.1 Textual output

As an illustration, here is an example of the standard textual output of the example provided on the web interface :

Gene number	Element number	Feature type name	Strand	Left end	Right end	Length	Phase	Frame
1	0	Utr5	-	1	199	199	NA	NA
2	0	Utr5	+	340	359	20	NA	NA
2	1	Initial	+	360	393	34	+1	+3
2	2	Internal	+	596	732	137	+2	+1
2	3	Internal	+	830	876	47	+1	+2
2	4	Internal	+	961	1286	326	+3	+2
2	5	Internal	+	1396	1478	83	+2	+3
2	6	Internal	+	1573	1648	76	+1	+1
2	7	Internal	+	1757	1818	62	+2	+1
2	8	Internal	+	1962	2057	96	+1	+3
2	9	Internal	+	2145	2306	162	+1	+3
2	10	Terminal	+	2491	2607	117	+1	+1
2	0	Utr3	+	2608	2626	19	NA	NA

The text output simply presents the list of exons and coding and transcribed unstranslated regions predicted. From left to right, this gives :

- The gene number and the exon number. UTR are always numbered 0 and initial exon 1.

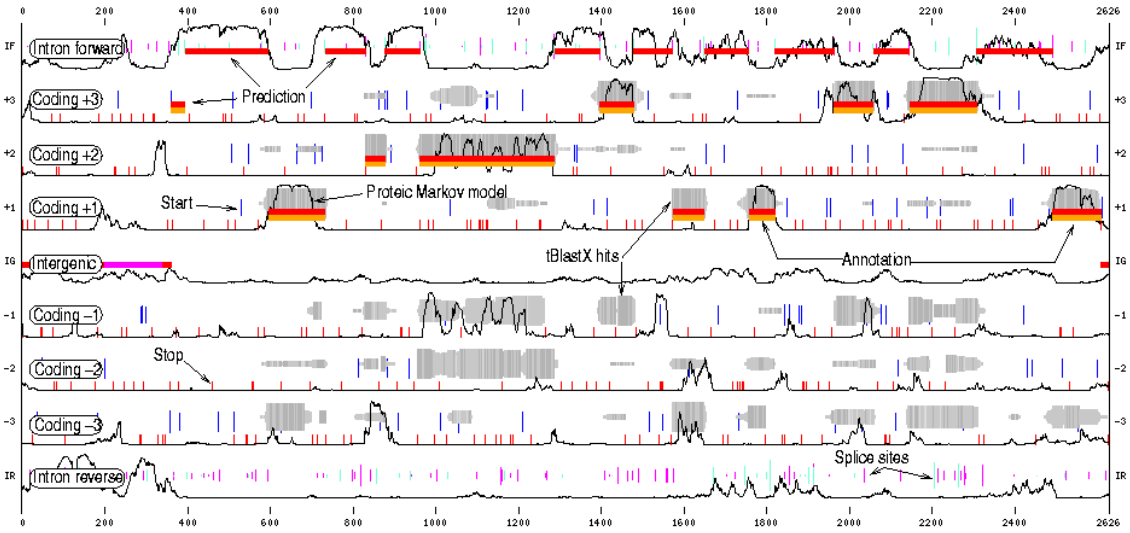
- The feature type name of the region : Utr5, Utr3, Initial exon (ATG to splice site), Internal exon, Terminal exon (splice site to stop) or Single-exon gene (ATG to stop).
- The strand for the region : ”+” for the forward strand and ”-” for the reverse one.
- The left and right extremities of the region with its length.
- The phase and frame of the region (for exons only). The phase indicates the relative position of the codons in the region with respect to the beginning of the region. The frame indicates the relative position of the codons with respect to the extremity of the sequence itself (beginning or end according to the strand).

Alternatively, a GFF output can be produced. In this case we get :

```
31482   EuGene  Utr5    1       199     0       -       .
31482   EuGene  Utr5    340     359     0       +       .
31482   EuGene  Init    360     393     0       +       2
31482   EuGene  Intr    596     732     0       +       0
31482   EuGene  Intr    830     876     0       +       1
31482   EuGene  Intr    961     1286    0       +       1
31482   EuGene  Intr    1396    1478    0       +       2
31482   EuGene  Intr    1573    1648    0       +       0
31482   EuGene  Intr    1757    1818    0       +       0
31482   EuGene  Intr    1962    2057    0       +       2
31482   EuGene  Intr    2145    2306    0       +       2
31482   EuGene  Term    2491    2607    0       +       0
31482   EuGene  Utr3    2608    2626    0       +       .
```

2.2 Graphical output

As an illustration, here is the graphical output of the example provided on the web interface.



- The horizontal axis is the sequence (here 2,626 nuc. long).
- The vertical axis represent possible predictions :
  - 6 possible coding phases (+1, +2, +3 for the direct strand, -1, -2, -3 for the reverse one).
  - introns on the forward (IF) and reverse (IR) strands.
  - other non coding regions (IG for intergenic or UTR regions).

Each of these possible prediction is called a track in the sequel.

- On the 6 coding tracks, in-frame START codons are represented as blue vertical lines. The longer the line, the better the score of the START (according to the Start window array model). In-frame STOP codons are represented as small red vertical lines. Thin black lines represent the smoothed normalized proteic coding/non coding score. HSP clusters are represented as grey blocks whose thickness is proportional to the number of hits at a given position.
- On the IF and IR tracks, splice sites are visible as green and magenta vertical lines whose length indicates the site score (according to the donor/acceptor window array model).
- The large red and magenta blocks indicate the prediction of EuGene'Hom : red for exon, intron, UTR and magenta for intergenic region.
- The GFF annotation provided by the user is visible as orange blocks. In this example, EuGene'Hom predicts a gene structure that matches the annotation.