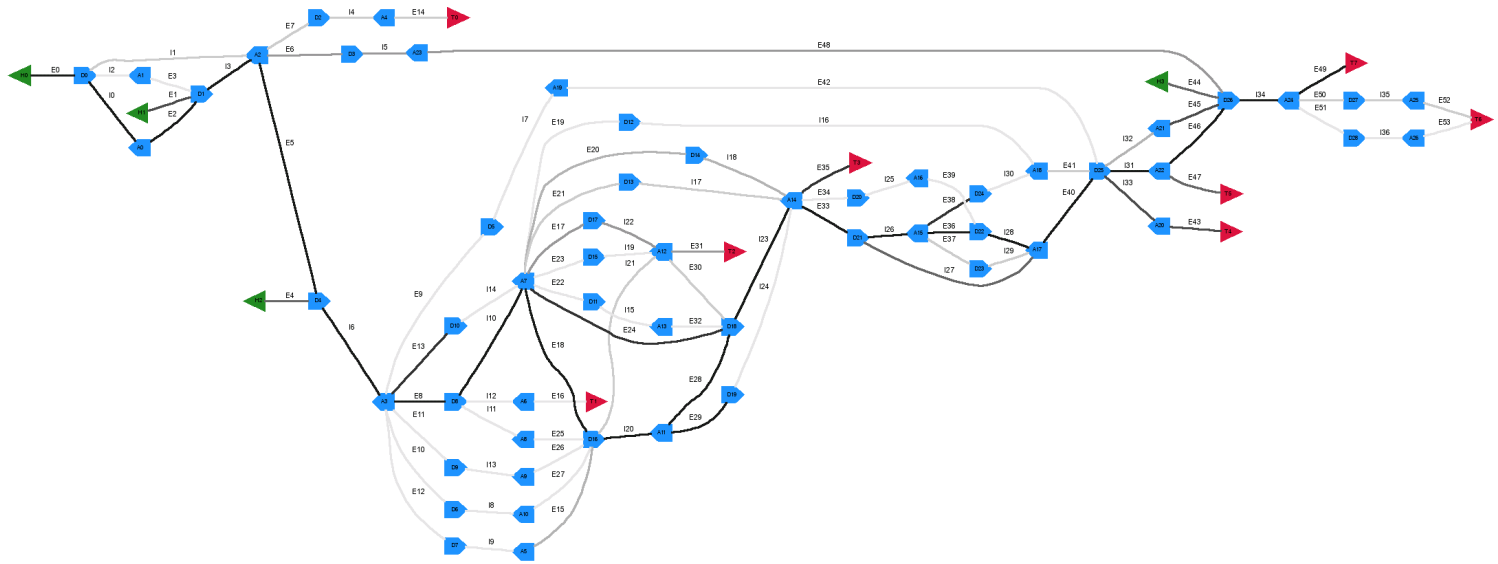
# Representing human genes with graphs :

# a graphic web interface



Michel Crausaz
17.12.2002

DEA in Bioinformatics
2001 - 2002

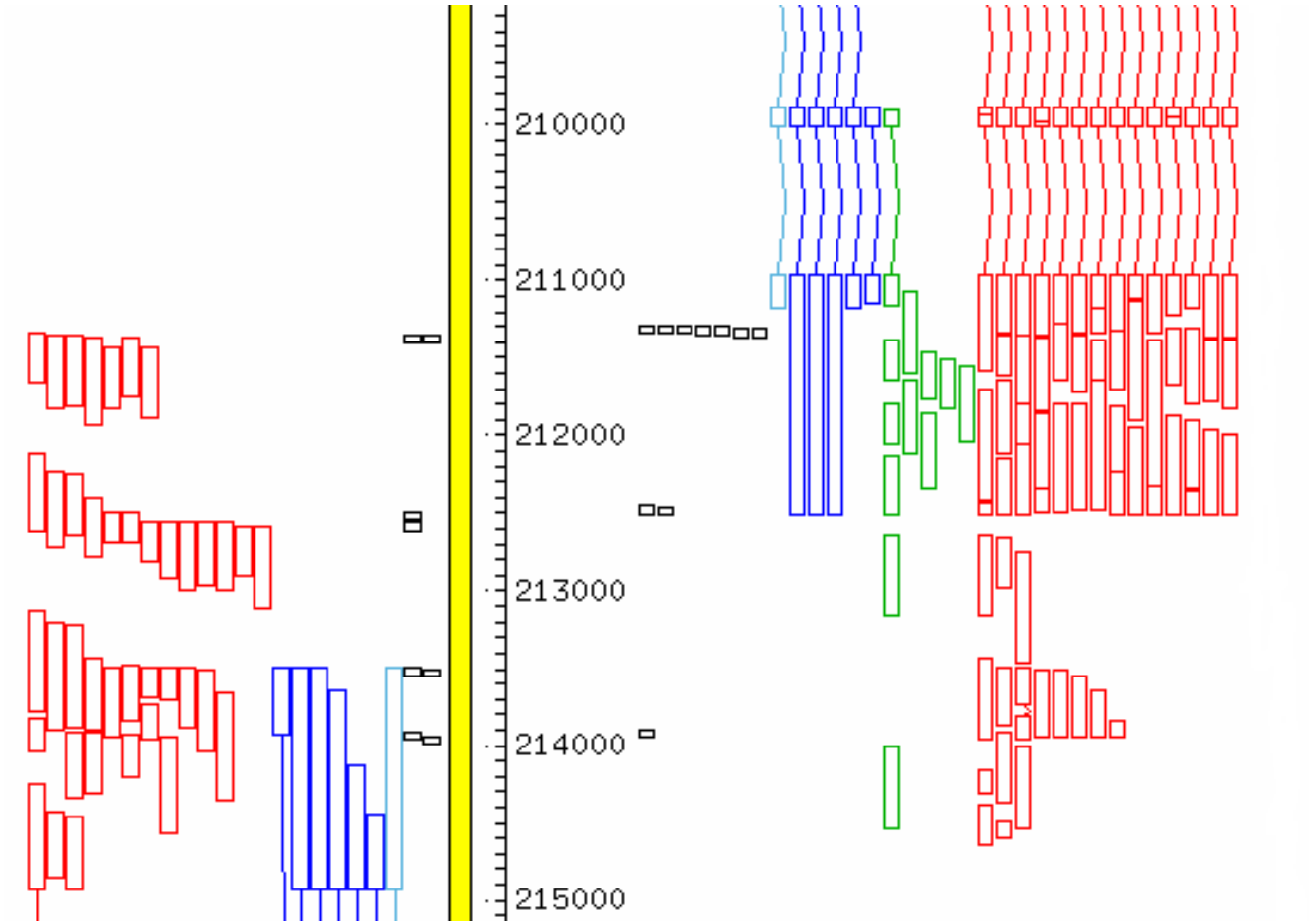
# Contents

- **Introduction**
  *Transcriptome and Tromer graph output*
- **Graph drawing**
  *DOT algorithm and DOT language*
- **Trome2Map Concept**
  *PERL scripts and web interface*
- **Conclusion and future prospects**

# Introduction

*Transcriptome and Tromer graph output*

# AceDB representation

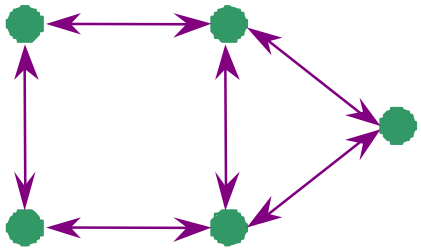



Alignments of transcripts to the genome (yellow bar) were visualized using AceDB. The direction of transcription is bottom to top on the left and top to the bottom on the right. Light-blue: RefSeq sequences; dark-blue: full-length cDNA sequences; green: ORESTES sequences; red: EST sequences. 3'tags are represented by black boxes, with one box per cluster member.

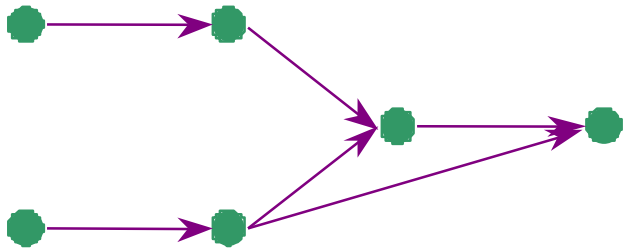
# Data

- NCBI human chromosome contigs NT_*
- Human EST section of EMBL
- Human HTC section of EMBL
- human mRNA documented in the human section of EMBL
- ORESESTES sequences from the LICR/FAPESP human Cancer Genome project
- Human mRNA documented in the NCBI curated RefSeq database
- published gene list of human chromosome 21
- SEREX sequences
- 3' tags

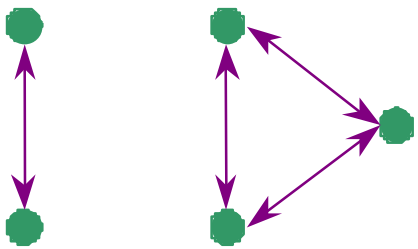
# Graph nomenclature



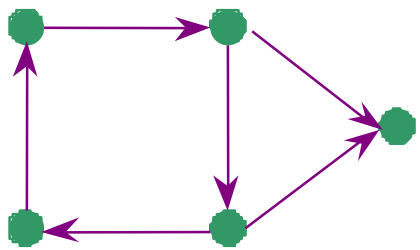
undirected and connexe



directed (digraph)



undirected and unconnexe



digraph fully connected

# Tromer graph output (graph description)

```
>map|NT_026281_10|NT_026281.7|- Chromosome 5 347746..383506
H0 [383506] 0,1
 =E0=> D0 1,-30
D0 [383386] 1,1
 H0 1,-30 =E0=>
 -I0-> A0 1,-10
A0 [375227] 1,1
 D0 1,-10 -I0->
 =E1=> D1 2,-60
D1 [374983] 1,1
 A0 2,-60 =E1=>
 -I1-> A1 2,-20
A1 [373531] 1,1
 D1 2,-20 -I1->
 =E2=> D2 2,-60
D2 [373393] 1,2
 A1 2,-60 =E2=>
 -I3-> A3 1,-10
 -I2-> A2 1,-10
A2 [362874] 1,1
 D2 1,-10 -I2->
 =E3=> T0 2,-30
T0 [361968] 1,0
 A2 2,-30 =E3=>
A3 [355204] 1,1
 D2 1,-10 -I3->
 =E4=> D3 1,-30
...
```

# Tromer graph output (detailed list)

```
E0 383506..383386 0,1
 E:BG192901 1..121 (383506..383386)
E1 375227..374983 0,2
 E:AW938686 31..213 (375167..374983)
 E:BG192901 122..366 (375227..374983)
E2 373531..373393 0,2
 E:AW938686 214..352 (373531..373393)
 E:BG192901 367..505 (373531..373393)
E3 362874..361968 0,2 I3
 E:BG192359 680..11 (362635..361968)
 E:BG192901 506..774 (362874..362602)
E4 355204..355078 0,1
 E:AW938686 353..479 (355204..355078)
E5 350898..350705 0,1 I4
 E:BI829756 1..194 (350898..350705)
E6 347956..347746 0,2
 E:AW938686 480..676 (347956..347762)
 E:BI829756 195..405 (347956..347746)
E7 336319..335999 0,2
 E:BF329257 1..259 (336277..336020)
 E:BI829756 406..724 (336319..335999)
I0 383385..375228 1
 E:BG192901 121..122 GT/AG -10
I1 374982..373532 2
 E:AW938686 213..214 GT/AG -10
 E:BG192901 366..367 GT/AG -10
I2 373392..362875 1
 E:BG192901 505..506 GT/AG -10
...
```

# Graph drawing

*DOT algorithm and DOT language*

# Aesthetic criteria

- **Expose hierarchical structure in the graph**. In particular, aim edges in the same general direction if possible. This aids finding directed paths and highlights source and sink nodes.
- **Avoid visual anomalies** that do not convey information about the underlying graph (*edge crossings, sharp bends*).
- **Keep edges short**.
- **Favour symmetry and balance**.

# DOT algorithm

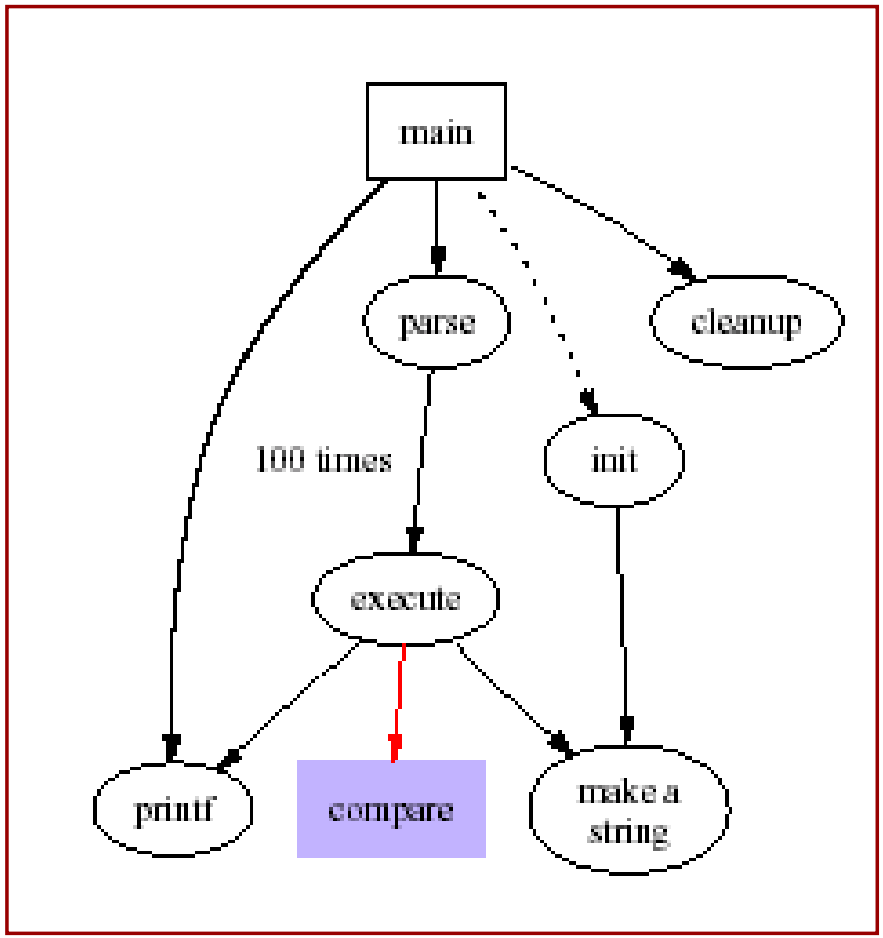
***AT&T Bell Laboratories***

This graph drawing algorithm has four steps :

- place nodes in discrete ranks,
- set the order of nodes within ranks to avoid edge crossing,
- set the actual layout coordinates of nodes,
- find the spline control points for edges.

# DOT language

```
1: digraph G {
2: size ="4,4";
3: main [shape=box];
4: main -> parse [weight=8];
5: parse -> execute;
6: main -> init [style=dotted];
7: main -> cleanup;
8: execute -> { make_string; printf}
9: init -> make_string;
10: edge [color=red];
11: main -> printf [style=bold,label="100 times"];
12: make_string [label="make a\nstring"];
13: node [shape=box,style=filled,color=".7 .3 1.0"];
14: execute -> compare;
15: }
```



# DOT in command line

```
$ dot –Tps graph1.dot –o graph1.ps

$ dot –Tpng graph1.dot –o graph1.png

$ dot –Tjpg graph1.dot –o graph1.jpg

$ dot –Tismap graph1.dot –o graph1.ismap
```

# Trome2Map Concept

*PERL scripts and web interface*

# Tromer2Map concept

***http://ludwig-sun2.unil.ch/~mcrausaz/form2.html***

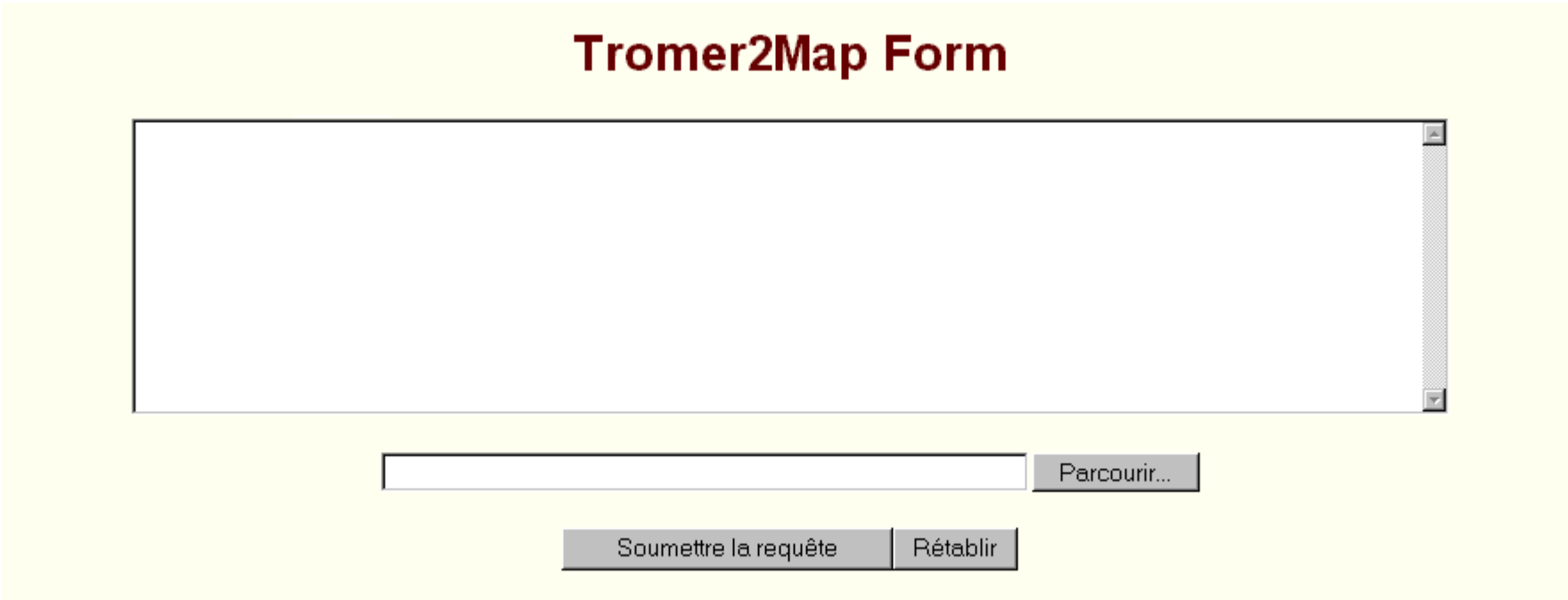
- **Trome2Dot** : transforms a Tromer graph file in a graph in DOT language, takes as argument a library text file;
- **Tromap2.pl** : returns the resulting web page, calls DOT program, Trome2dot and itself;
- **Info** : returns a list of associated RNA fragments depending on the exon/intron clicked on the map;
- **Fetch_web** : searches a clicked entry and returns it in raw text format;
- **Libraries** : the formatting files, for header and image map.

# Libraries

| | |
|---|---|
| `>default lib` | library name |
| `g_bgcolor = white` | graph background color |
| `g_nodesep = 0.4` | node separation |
| `g_rankdir = LR` | landscape orientation |
| `n_fontname = Arial` | node font name |
| `n_fontsize = 12` | node font size |
| `n_style = filled` | node shape style |
| `n_A = dodgerblue` | acceptor color |
| `n_D = dodgerblue` | donor color |
| `n_H = forestgreen` | start color |
| `n_T = crimson` | stop color |
| `e_fontname = Arial` | edge font name |
| `e_fontsize = 16` | edge font name |
| `e_style = bold` | edge style |
| `e_arrowsize = 1.0` | size of arrows |
| `e_E = black` | exon label color |
| `e_I = grey` | intron label color |

```
special
E8 = red
I8 = red
E21 = green
I14 = blue
E23 = green
```

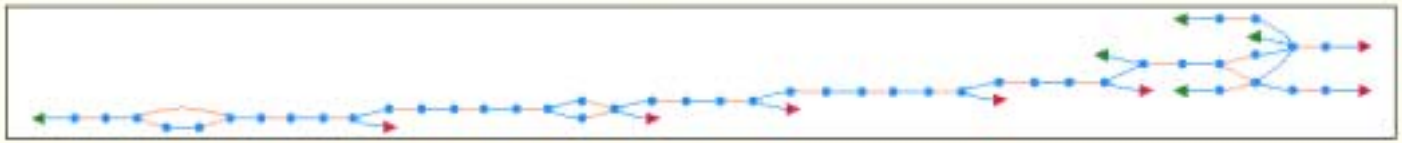
# Web submission form



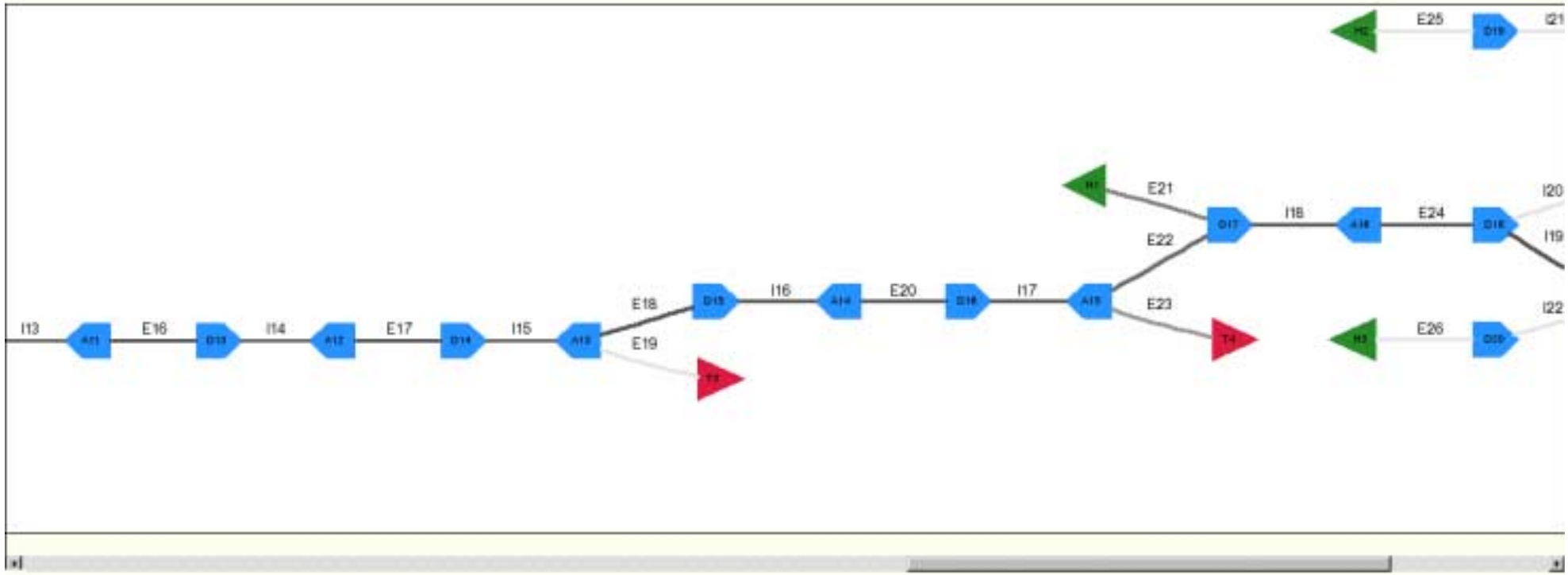

***http://ludwig-sun2.unil.ch/~mcrausaz/form2.html***

**Graph accession number :**
NT_026281_2

**Overview :**



Contig : NT_026281.7 Chromosome : 8
Start : **199929** Stop : **228011**
Nbr. of exons : **34** Nbr. of introns : **25**
Size : **28082 bp's**




**Exon / Intron :** 1 EST 2 EST's 3 - 5 EST's 5 - 10 EST's 10 - 20 EST's 20 - 50 EST's 50 - 100 EST's 100 - 500 EST's **more than 500 EST's**

For more information : email

**Nodes :** H = Start A = Acceptor D = Donor T = Stop

# GRAPH INFO

Graph accession number : **NT_023132_30** Contig : **NT_023132.8** Chromosome : **5** Start : **982075** Stop : **1005508**

---

# EXON / INTRON : E40

Position (start - stop) : **987910 -- 987809** Size : **-101 bp's** Nbr. EST : **766** Rest : **I5**

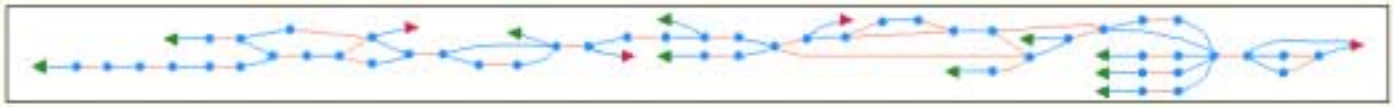
---

## REFERENCES :

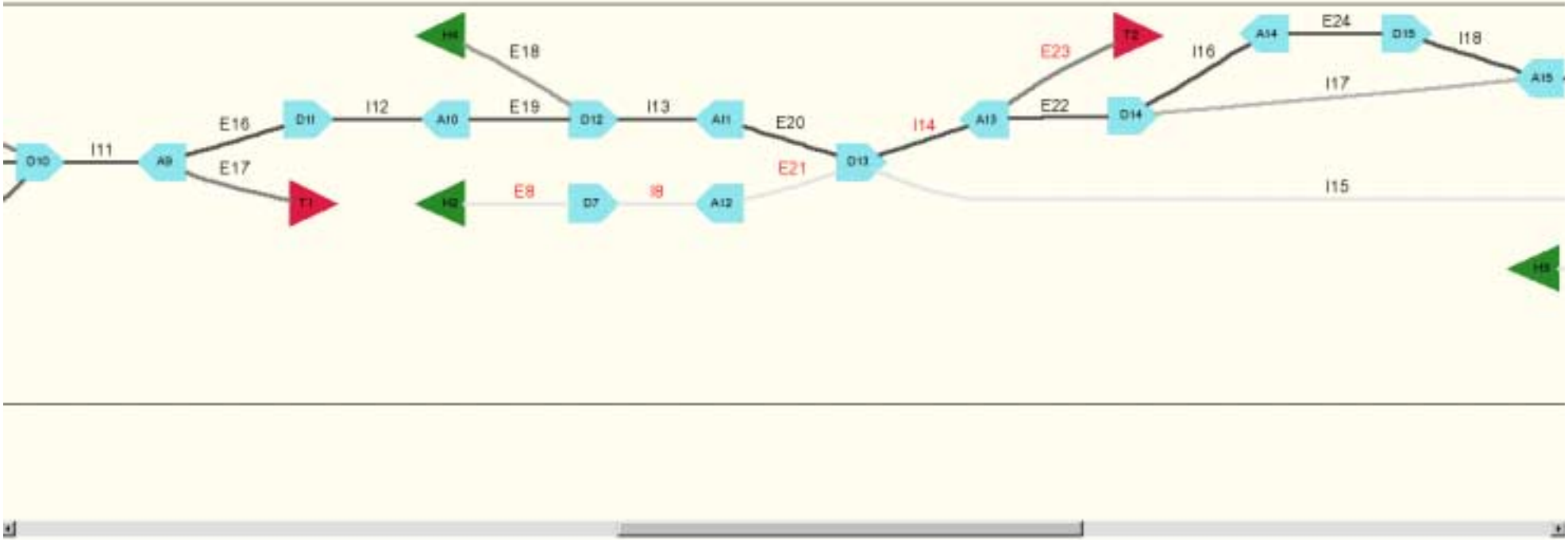
| | | | |
|---|---|---|---|
| **2** EST - EMBL | AA035121 | 43..144 | (987910..987809) |
| **3** EST - EMBL | AA037740 | 358..467 | (987910..987809) |
| **4** EST - EMBL | AA074067 | 458..433 | (987834..987809) |
| **5** EST - EMBL | AA075046 | 410..302 | (987914..987809) |
| **6** EST - EMBL | AA102448 | 513..436 | (987884..987809) |
| **7** EST - EMBL | AA126842 | 1..32 | (987840..987809) |
| **8** EST - EMBL | AA146783 | 359..460 | (987910..987809) |
| **9** EST - EMBL | AA148103 | 537..438 | (987907..987809) |
| **10** EST - EMBL | AA157802 | 437..335 | (987910..987809) |
| **11** EST - EMBL | AA166616 | 428..331 | (987910..987809) |
| **12** EST - EMBL | AA173861 | 107..208 | (987910..987809) |
| **13** EST - EMBL | AA173870 | 157..258 | (987910..987809) |
| **14** EST - EMBL | AA176703 | 525..432 | (987904..987809) |
| **15** EST - EMBL | AA187914 | 537..435 | (987910..987809) |
| **16** EST - EMBL | AA191576 | 430..329 | (987914..987809) |
| **17** EST - EMBL | AA224055 | 212..313 | (987910..987809) |
| **18** EST - EMBL | AA227437 | 523..441 | (987897..987809) |
| **19** EST - EMBL | AA307641 | 182..283 | (987910..987809) |
| **20** EST - EMBL | AA309956 | 365..408 | (987910..987867) |
| **21** EST - EMBL | AA316208 | 1..33 | (987841..987809) |
| **22** EST - EMBL | AA329394 | 165..268 | (987910..987808) |
| **23** EST - EMBL | AA363621 | 1..91 | (987899..987809) |
| **24** EST - EMBL | AA366934 | 1..67 | (987875..987809) |
| **25** EST - EMBL | AA367013 | 75..176 | (987910..987809) |
| **26** EST - EMBL | AA393177 | 251..352 | (987910..987809) |
| **27** EST - EMBL | AA469347 | 454..407 | (987856..987809) |
| **28** EST - EMBL | AA469360 | 471..433 | (987847..987809) |
| **29** EST - EMBL | AA482747 | 45..146 | (987910..987809) |

**Graph accession number :**
NT_025133_2

**Overview :**



Contig : NT_025133.7 Chromosome : 19
Start : **82076** Stop : **126090**
Nbr. of exons : **42** Nbr. of introns : **30**
Size : **44014 bp's**




**Exon / Intron :** 1 EST 2 EST's 3 - 5 EST's 5 - 10 EST's 10 - 20 EST's 20 - 50 EST's 50 - 100 EST's 100 - 500 EST's **more than 500 EST's**

**Nodes :** H = Start A = Acceptor D = Donor T = Stop

# Conclusion

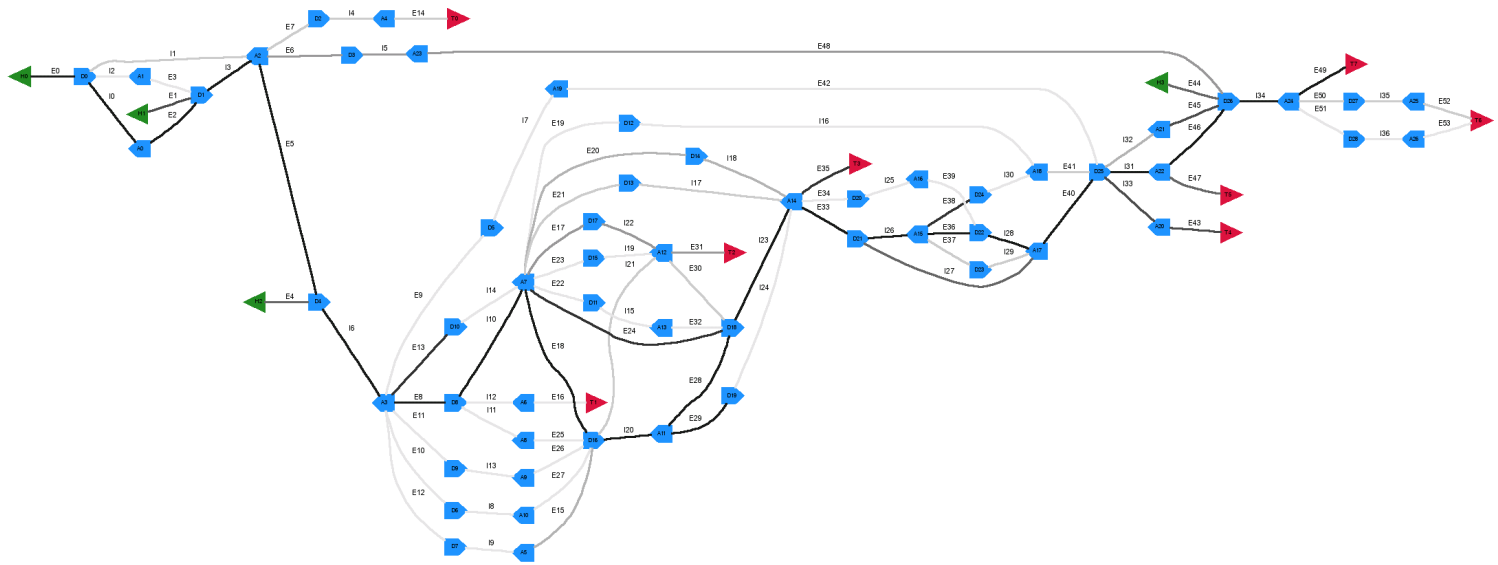
- universality - readability - comprehensibility ;
- user-friendly graph representation ;
- open and evolutive ;
- possible link to a transcriptome database.

# Further developments

- alignment on genomic contigs ;
- 3' tags ;
- develop libraries ;
- zoom in and out.

# Representing human genes with graphs :

# a graphic web interface



**http://ludwig-sun2.unil.ch/~mcrausaz/form2.html**

Michel Crausaz
17.12.2002

DEA in Bioinformatics
2001 - 2002