



AltSplice computational pipeline

Alternative Introns/Exons, Events, and Splice Patterns


Jean-Jack Riethoven
ASD Project
Seqdb group seminar



Alternative Splicing in one slide..



- ☞ ~20.000 to 30.000 genes (human)
- ☞ ~100.000 to 120.000 proteins (human)
- ☞ (pre-mRNA) alternative splicing is one of the methods used

 exon
 intron



ASD: Alternative Splicing Database

☞ ASD Project:

- “aims to understand the mechanism of alternative splicing on a genome-wide scale by creating a database of alternatively spliced exons from human, and other model species”.

☞ Three main databases:

- AltExtron: computational / prototype & research
- AltSplice: computational / production
- AEdb: manual curated / production

☞ Various satellite databases coming up



AltSplice Pipeline

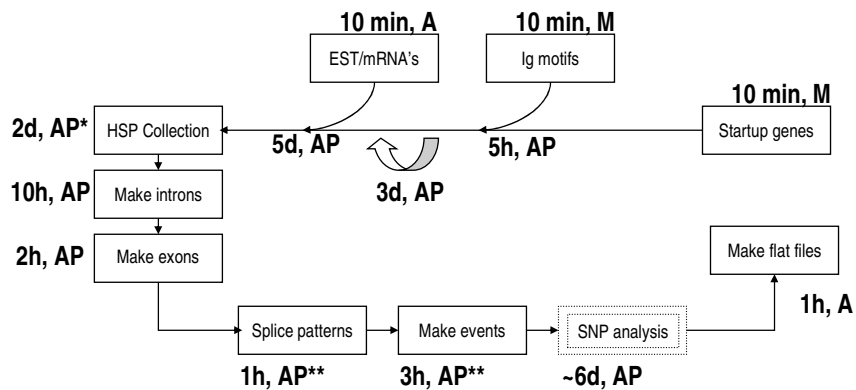
- ### ☞ Philosophy:
- to have an automated pipeline that can deliver a high-quality set of transcript confirmed alternative events, together with appropriate annotation, in under two weeks from the baseline data sets.

☞ Design considerations:

- **Quality control + speed.**
- Each step in pipeline should have own module.
- Every module has its own inherent quality checks and warning/error reporting.
- Where possible, modules are run in parallel.
- Where possible and appropriate, modules should have mile-stone markers.
- As much as possible, independence of external modules.



ASD Pipeline – Time by Step



A = Automatic, P = Parallel
 * = partly parallel
 ** = parallel capable, but at moment sequential



ASD Pipeline

- ☞ Programmed in: Perl (5.6)
- ☞ Uses:
 - DB_File (EST/peptide indices, file indices)
 - DBI::Oracle (db loading)
 - Proc::Simple (child process handling)
 - Bio::Tools::BPlite/BPbl2seq (blast parsing)*
- ☞ Parallel execution via LSF, but handled by own job queue + checking code (100 concurrent jobs)
 - re-submit if final milestone not reached
 - know exactly when job is finished

* Trying to minimise dependency on BioPerl (too much in flux)



Start-up Data Sets

- ☞ Ensembl known genes [19298]
 - Normal gene sequence
 - With flanked sequence at 5' and 3' sides (default 3000 bases)
 - Annotation (dbxrefs, structure, cross-species)
- ☞ EST/mRNA sequences from EMBL [~5.5 million]
- ☞ Ig motifs from NCBI [171]

[] v13.31: July 2003 for human



Ig Blast

- ☞ Purpose: to remove any immuno-globulin coding genes from the data set.
- ☞ Why: highly similar domains

- ☞ Since the blast database for Ig related genes is so small (171 sequences at the moment) the all genes vs. Ig motifs blast quickly done.

- ☞ Pid = 95%, coverage = 95%. Uses unflanked gene sequences.
- ☞ Typical reduction: 30-60 genes.



Redundancy Analysis

- ☞ Purpose: to remove genes that are highly redundant, i.e. large part of its sequence is similar with other gene(s).
- ☞ Typical reduction: 500-1000 genes (1-2%).
- ☞ Blast all vs. all genes –ungapped–, e-value= 10^{*-15} , Pid=99%, coverage=90% uses unflanked gene sequences..
- ☞ For each pair of genes that are indicated to be redundant, we remove the smaller.

[18632 genes left]



Gene vs. EST/mRNA Blast

- ☞ Purpose: to find (parts of) the genes that are confirmed by expressed transcripts.
- ☞ E-value = 10^{*-10} , max. 1000 highest scoring transcripts per gene, max. 1000 HSP's per transcript. This blast uses the flanked sequences (3000 bases on either side).
- ☞ At the moment the longest step in our pipeline – we are investigating how to reduce this time-sink.
 - A trivial but partial solution would be to only blast the new and modified transcripts.



Initial HSP Collection

- ☞ Purpose: to build a collection of high scoring pairs (HSP's) that fulfills our minimum conditions and can serve as a reference for further steps in the pipeline.
- ☞ Stored in GFF (general feature format) files.
 - Select only those HSP's that have minimum length of 25 bases, and have a pid of 95% and up.
 - Per gene-transcript combo, group HSP's by strand. Select HSP's for that strand that has highest coverage and pid, in that order.
 - Remove HSP's that match more than 1 gene.
 - Remove HSP's that are alone on flanking region.
 - Remove HSP's that have more than 20% overlap on transcript.
 - Remove lone HSP's (i.e. only one HSP per EST).

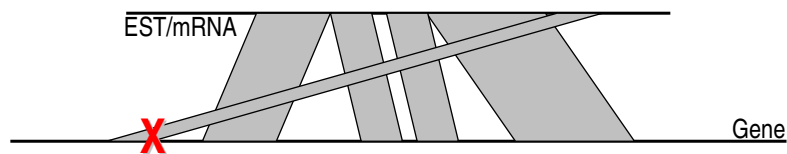


Intron Determination

- ☞ Purpose: determine and validate transcript confirmed introns.
- ☞ By default: use HSP's with pid = 98%.
- ☞ Some more quality control.
 - Remove HSP's that match same gene region.
 - Remove rogue HSP's.
 - Don't use HSP combo's where overlap > 10 or gap > 1 on EST.
- ☞ Determine possible introns between each subsequent HSP.
 - Introns are annotated based upon method.



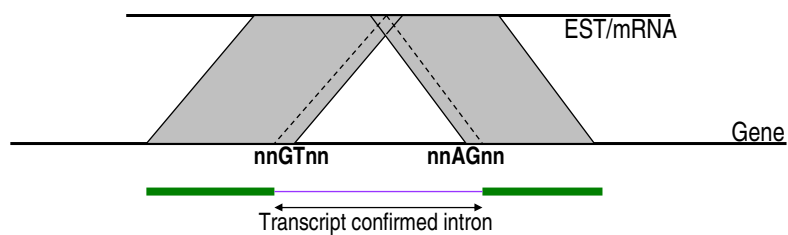
Intron Determination



Removal of rogue HSP's: not in correct order when compared with the other HSP's.



Intron Determination - II



- ☞ Fix overlaps and 0/1 base gaps until we get a group of 'candidate' introns (GT-AG, GC-AG, AT-AC).
- ☞ Validate each intron in group (10b up/downstream exact match between gene & transcript).
 - Exception: allow 1 base mismatch on either side IF confirmed by exact match intron on other transcript. Required for SNP analysis.
- ☞ Choose from the group of 'valid' introns only 1.



Patching / Exons

- ☞ Purpose: to try and fill gaps > 35 bases that are left on both EST and gene side.
 - Terminal and internal.
- ☞ Localised blast (bl2seq) with reduced constraints (e.g. pid=95%).
- ☞ Use the extra HSP's to try and find new introns.
 - Terminal: 80%, of which 30% new.
 - Internal: 20%, of which 40% new.
- ☞ Determine confirmed exons by using the transcript confirmed introns.

[15644 genes left]



Now confirmed features

- ☞ Confirmed intron: a GT-AG, GC-AG, AT-AC intron that has validated, same-sequence up/downstream regions when comparing gene and transcript.
- ☞ Confirmed exon: an HSP match directly flanked on both sides by a confirmed intron.
- ☞ Bias due to methodology:
 - Biased against UTR region.
 - Biased against tiny exons (<25 bases).
 - Biased in favour of GT-AG, GC-AG, AT-AC introns in that order.



Splice patterns

- ☞ Splice patterns form basis for event delineation
 - Not single introns/exons
- ☞ Group transcripts together based on their splice pattern
- ☞ Determine differences between all possible group combinations (using representative patterns)
- ☞ But before grouping: chuck out bad transcripts
 - More than one pair of internal ambiguities
 - Gap ≤ 10 , overlap ≤ 10 bases on EST
 - Non-canonical intron
 - Gap > 10 bases on EST, or gap < 40 bases on gene



Splice Patterns – Grouping Rules

- ☞ Sort transcript on decreasing number of introns (+secondary)
- ☞ Basic grouping rules – transcript 2 can be grouped with transcript 1 if:
 - All introns of 2 match exactly with 1, leaving no 'extra' introns
 - In case of discontinuity an intron on other transcript fits with 10 bases tolerance on either side (gap/non-canon intron) or 2*overlap bases (overlap)
 - Terminal hsp does not extend a confirmed feature by more than 25 bases
- ☞ Ambiguity equals ambiguity (terminals)

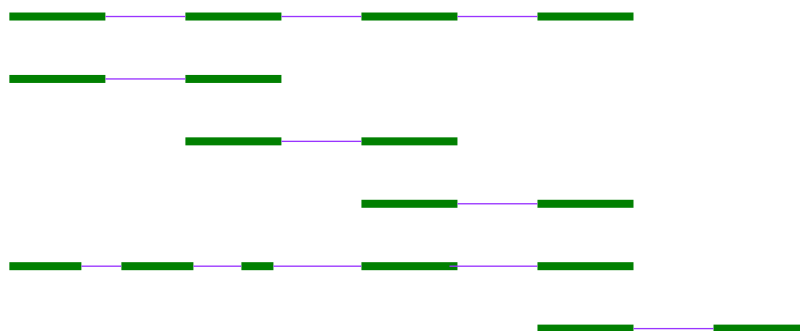


~100..200, 400..700, 1200..1350, 1600..~1690

- ☞ Introns have definite (well-defined) 5' and 3' ends
- ☞ An exon/hsp only has a definite end if it is directly followed by an intron. If not, that end is considered ambiguous (~).

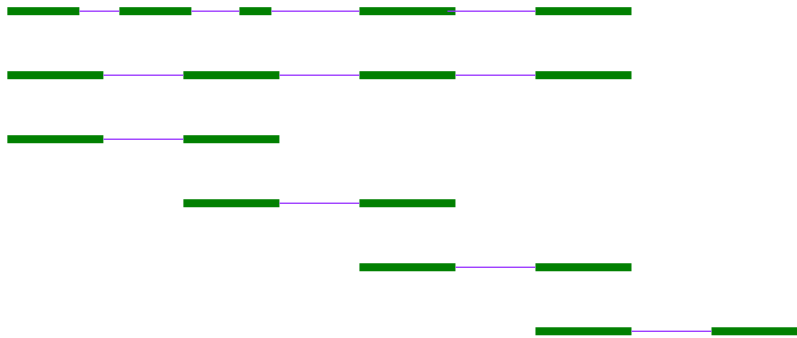


Collection of all transcript matches





Sort them according to number of introns

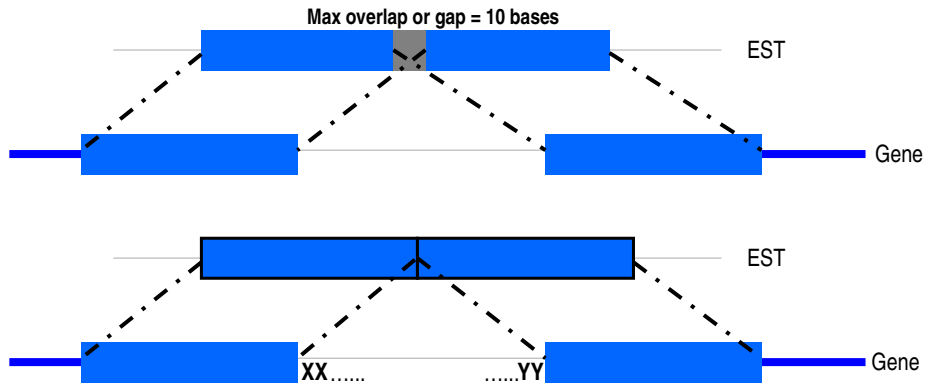


Transcripts are now grouped according to confirmed splice-pattern





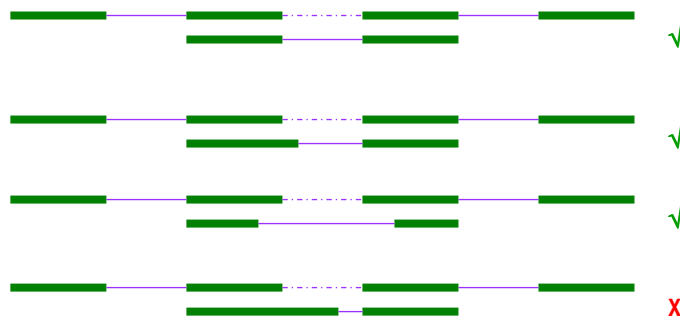
Two types of internal discontinuities



$[XXYY] \notin \{GTAG, GCAG, ATAC\} \Rightarrow$ internal discontinuity
E.g. ~100..200, 400..~500, ~600..700, 900..~1000

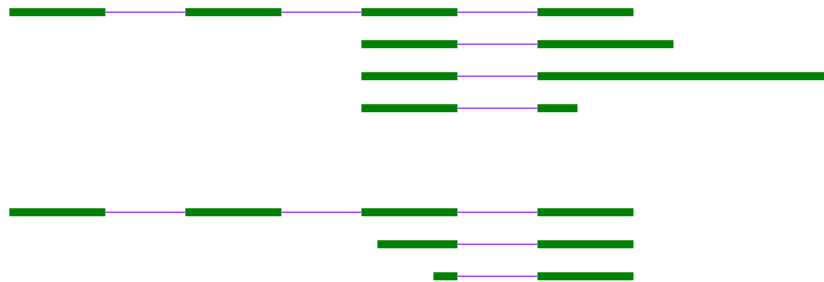


Internal discontinuities



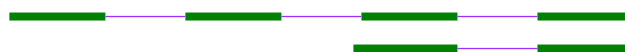


Terminal ambiguities

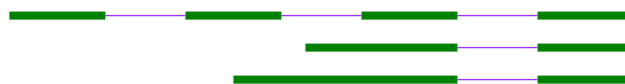


25 bases extension rule

If ambiguous end of an exon (hsp) extends the definite end of an exon in the other transcript by more than 25 bases then the transcripts are dissimilar.



Extension \leq 25 bases: similar



Extension $>$ 25 bases: not similar



Examples of Splice Patterns

```
>ENSG00000179222
CLASS 1
BU845038-1 ~6040..6450,6581..6644,6920..6999,7168..7259,7487..7566,7658..~7701
BX367175-1 ~5732..6450,6581..6644,6920..6999,7168..~7256
BU155558-1,2 ~5780..6450,6581..6644,6920..6999,7168..~7219
BU185730-1,2 ~5780..6450,6581..6644,6920..~6967
BU190480-1,2 ~5780..6450,6581..~6643
CLASS 2
AL557885-2 ~5027..5133,5782..6450,6581..6644,6920..6999,7168..~7258
BG831556-2 ~5083..5133,5782..6450,6581..6644,6920..~6995
BQ709714-2 ~4957..5133,5782..6450,6581..~6631
BG334111-2,5 ~4519..5133,5782..~5885
CLASS 4
BQ062811-4 ~7656..7700,7954..8016,9555..9669,10926..11311,11519..~11725
BI195994-4 ~9592..9669,10926..11311,11519..~11727
Classes with staggered overlap + same structure : (2 & 1)
Classes with staggered overlap only : (4 & 1)
```

EST id has an indication of the class(es) number(s) into which it is (or can) be grouped. Multiple classes means the transcript is ambiguously added to a class.

[Note: various classes removed, a lot of transcripts removed for example purposes]



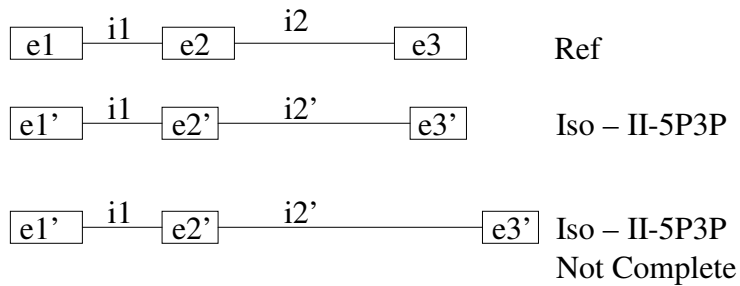
Delineation of Events

- ☞ Each group of transcripts is called a class, representing a unique splice pattern.
- ☞ From each class, pick the first transcript (the representative).
- ☞ Compare this representative against all other representatives.
 - For intron-based events look at overlaps between introns.
 - 'Grow' event until all overlaps have been taken care of.
 - For exon-based events look at exons that include a full intron.
 - Exon isoforms are taken from the above two types there were flanking exons to the events differ.
- ☞ Annotate the events.
 - Determine main and subtype.
 - Determine flanking exon/introns changes.
 - Check completeness of events (part of event on first transcript side is fully included in transcript 2 and vice versa).
- ☞ Group events by main type and location of main component.

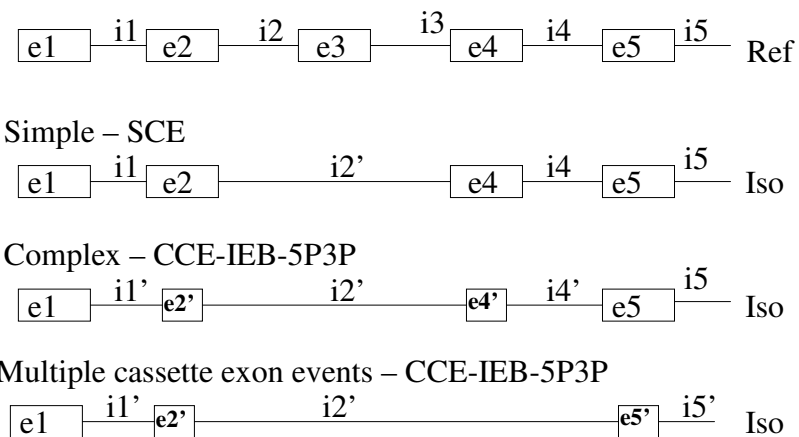


Intron isoforms

☞ The most simple events – basically any intron change where we cannot ‘grow’ the event.

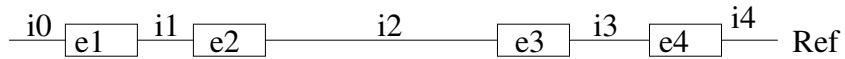


Cassette exons

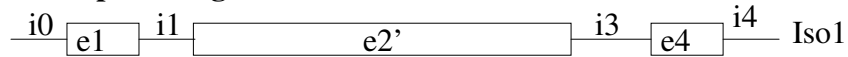




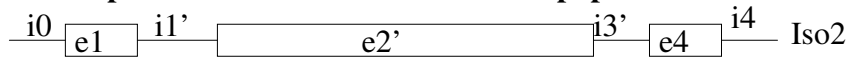
Intron Retention



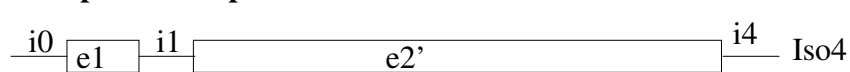
Simple - single - SIR



Complex - exon isoforms - CIR-EB-5p3p



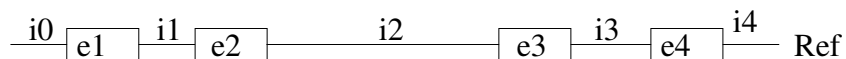
Simple - multiple - SIR



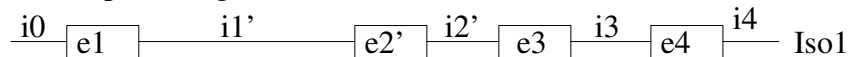
☞ Also have: CIR-CE-5P, etc.



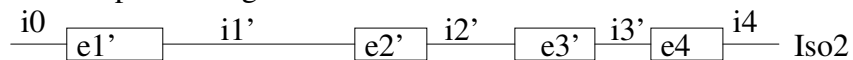
Mutual exclusive event



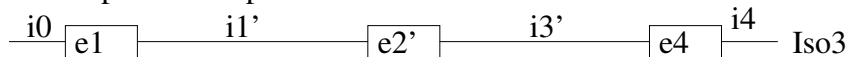
Simple - single - SME



Complex - single - CME-IEB-5P3P



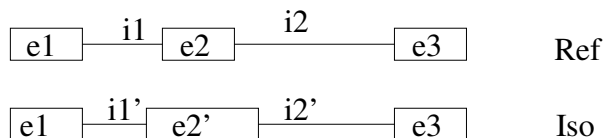
Simple - Multiple - SME





Exon isoform

- Exon isoforms are extracted from all the other events there were we have exon isoforms on the flanks.



Grouping of Event Types

```

Type      :      CASSETTE EXON (SCE,CCE-EB-5P)
Cassette exons: 16749..16845,17123..17189 [164 b]
Occurs in:      (5 & 2) is Complete CCE-EB-5P, (5 & 3) is Complete SCE,
                (5 & 1) is Complete SCE

(5 & 2) :
  14855..19530 (intron) <=> 14855..16748,16846..17122,17190..19530 (introns)
  e1 (14638..14854)      <=> e1' (14596..14854) [42, 0]
  e2 (19531..19647)     <=> e2'' (19531..~19592) [0, -55]
  1 Confirm. EST's      <=> 2 Confirm. EST's

(5 & 3) (5 & 1) :
  14855..19530 (intron) <=> 14855..16748,16846..17122,17190..19530 (introns)
  e1 (14638..14854)      <=> e1 (14638..14854) [0, 0]
  e2 (19531..19647)     <=> e2 (19531..~19647) [0, 0]
  1 Confirm. EST's      <=> 33 Confirm. EST's

TRPT-ISO1: ~13825..13977,14638..14854,19531..19647,20431..20597,21981..~22061
TRPT-ISO2: ~13928..13977,14638..14854,16749..16845,17123..17189,19531..19647,
20431..20597,21981..~22075
TRPT-ISO2: ~10471..10527,13826..13977,14596..14854,16749..16845,17123..17189,19531..~19592
TRPT-ISO2: ~10471..10527,13826..13977,14638..14854,16749..16845,17123..17189,19531..~19647

Type      :      EXON ISOFORM (EI-5P)
Struct    :      14596..14854 (exon)      <=>      14638..14854 (exon)
Length change: -42, 0 (-42)
Occurs in:      (2 & 1) part of (none), (2 & 3) part of (none),
                (2 & 4) part of (CCE-EB-5P), (2 & 5) part of (CCE-EB-5P)

```



Output Files

- ☞ Output files that we generate at the moment:
 - Requires at least one confirmed feature:
 - **.GENE** – “flanked” gene sequence + annotation.
 - **.TRANSCRIPT** – EST/mRNA id’s, description, and concise alignment line.
 - **.INTRON** – confirmed introns + further annotation.
 - **.EXON** – confirmed exons + further annotation.
 - **.CLASSES** – the classes (grouped transcripts) indicating distinct splice patterns for each class.
 - Requires an alternative event:
 - **.GROUPS** – file with alternative events grouped by main type and main component.
 - Further files: various SNP analysis output files.

[8314 genes left]



ASD Pipeline - Issues

- ☞ Trying to keep start to end within two weeks per species
- ☞ Ever increasing size of EST/mRNA collection
 - Longer blast times per gene, esp. If blast index doesn’t fit in memory (+file cache) anymore
 - Possible: resolve this by splitting blast db + parallel blast by gene (dblast/MPI)
- ☞ I/O bound -> 100 processes accessing disk
- ☞ (Disk space – looks trivial but gene vs. gene/est blast fills up disk very fast)



Sources

- ASD Project Homepage
 - <http://www.ebi.ac.uk/asd/>
 - Supported by EC grant QLRT-CT-2001-02062

- ASD Production Pipeline
 - <http://www.ebi.ac.uk/asd/asd/>