The Sequence Ontology

Suzanna Lewis

2003

This talk…

- Why is there a SO
- What is the SO
- SO and GFF3
- A bit about mereology
- Some examples using the SO to describe Drosophila and other examples of things the SO is useful for…

Ontologies help with decision making

Where should I eat…?

handy ontology tells us what's there…

Type of cuisine

(Presumable) country of origin

Ontologies don't just organize data; they also facilitate inference, and that creates new knowledge, often unconsciously in the user.

What a 5 year old child will likely infer about the world from this helpful ontology:

Fresh Juice is a national cuisine…

Flag of fresh juice

Where delicatessen food hails from from…

'Frozen Yogurt' cuisine in search of a national identity?

- Bio-medical knowledge and sequence data have grown to such proportions that ontologies and knowledge bases have simply become necessities.

- We need to get this right, otherwise we won't—
  - *know what we know, or*
  - *where to find it, or*
  - *what to infer from it.*

obo principles

1. Be Open Source.
2. Use common syntax - GO, OWL.
3. Work together for a consensus.
4. Share name/id space - domain:string.
5. Define your concepts.
6. Involve the community.

The aims of SO

1. Develop a shared set of terms and concepts to annotate biological sequences.

2. Apply these in our separate projects to provide consistent query capabilities between them.

3. Provide a software resource to assist in the application and distribution of SO.

4. Meet the OBO criteria.

This is useful if you want to:

- Annotate sequence using consistent descriptions.
- Share semantics between model organism databases and thus enable practical querying.
- Describe alterations and mutations at the sequence level and higher.

e.g. What is a pseudogene?

- Human
  - Sequence similar to known protein but contains frameshift(s) and/or stop codons which disrupts the ORF.
- Neisseria
  - A gene that is inactive - but may be activated by translocation (e.g. by gene conversion) to a new chromosome site.
  - - note such a gene would be called a "cassette" in yeast.

Or, for example, give me all the dicistronic genes

- Define a dicistronic gene in terms of the cardinality of the transcript to open-reading-frame relationship and the spatial arrangement of open-reading frames.

First steps

1. Use in an existing exchange format
2. Freezing a pertinent (and useful) part of the ontology
3. Making inferences from some real data.

GENERIC FEATURE FORMAT VERSION 3

- Author: *Lincoln Stein*
- Not the most expressive way of representing genomic features but…
  - It is simple
  - Can be modified with just a text editor
  - Can be processed with shell tools like grep.
- Yet it has fragmented into multiple incompatible dialects, mostly because people wanted to extend it.

…A conundrum

GFF3—having it both ways

- Addresses the most common extensions to GFF

and still

- Preserves backward compatibility with previous formats.

GFF3 extensions

- Adds a mechanism for representing hierarchical grouping of features and subfeatures.
- Distinguishes group membership from feature name/id
- Allows a single feature, such as an exon, to belong to more than one group at a time.
- Describes an explicit convention for pairwise alignments
- Describes an explicit convention for features that occupy disjoint regions

GFF3 extensions today

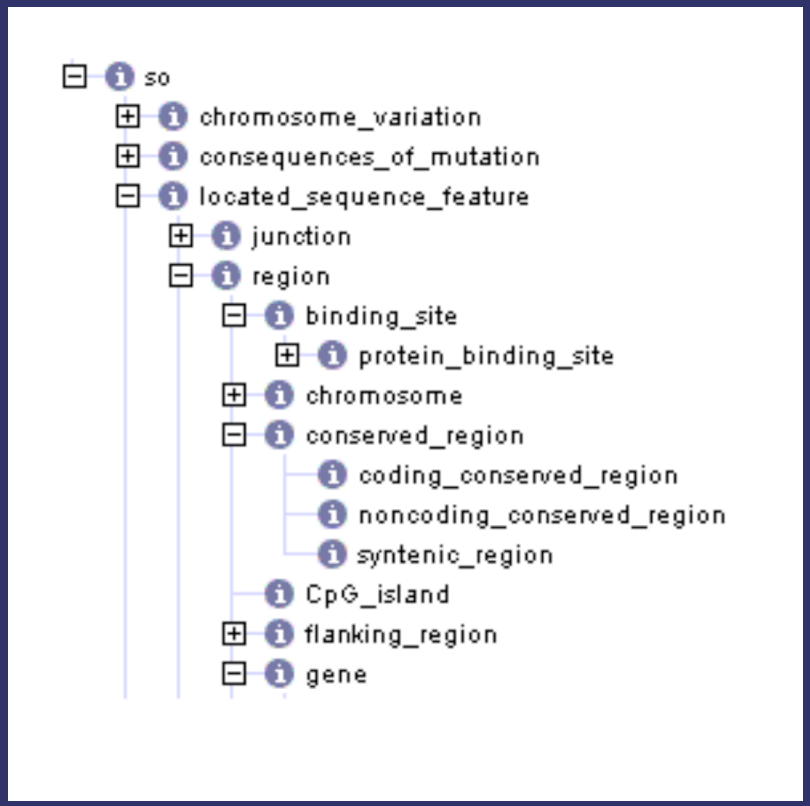
- Constrains the feature type field to the *SO*
- Will be committed in July

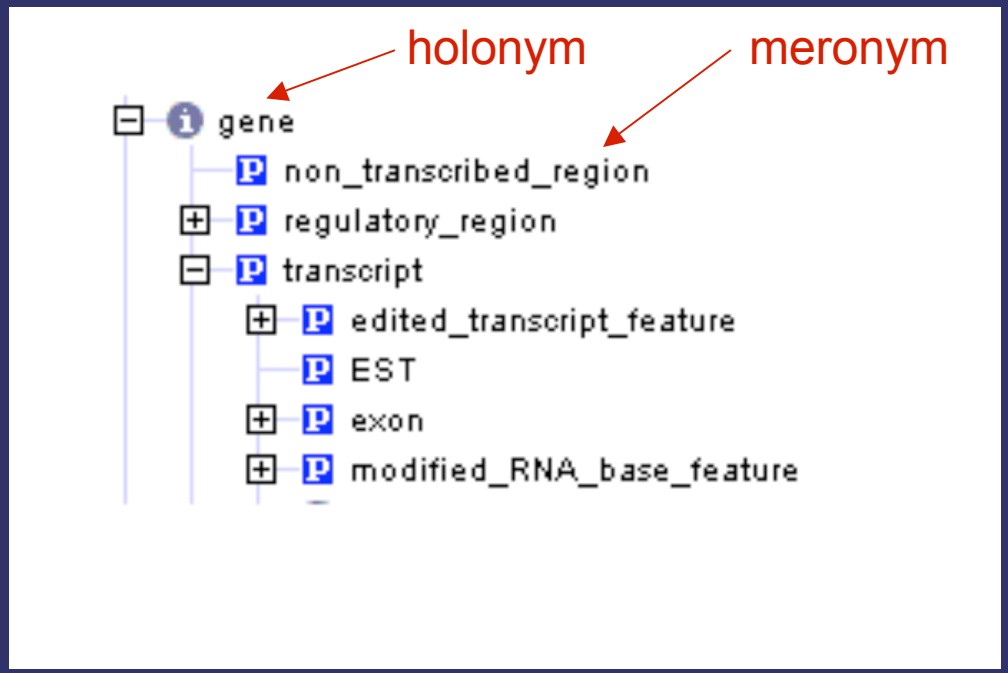
Sequence Ontology for Feature Annotation—SOFA (aka SO alpha)

- Includes only locatable features
- Designed for data exchange, e.g. in GFF3
- Will be frozen for 12 months

What are the relationships among the 913 (currently) concepts?

- ISA—927 relationships
- PARTOF—186 relationships





How can we use these relationships?

- ISA
  - Children inherit the properties of their parents.
  - Subsumption/ inference
  - Reason over the relationships
  - Description logics
- PART_OF
  - Parts do not inherit the properties of the whole.
  - Classical extensional mereology

Other kinds of 'parts'—piece?

- Parts are not the same as pieces. Consider a body being dissected into constituent parts or hacked to pieces. There are an infinite number of pieces.
- A part has:
  - Autonomy
  - Non-arbitrary boundaries
  - Determinate function with respect to the whole

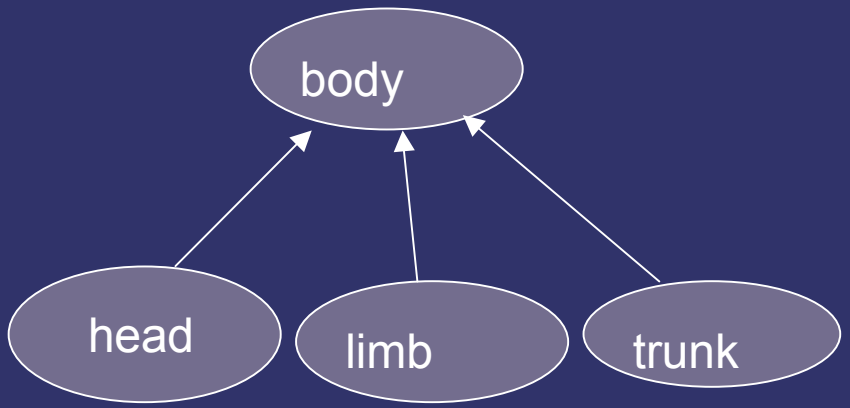
Other kinds of 'parts'

- Collections (lion/pride)
  - Not homomerous but separable.
- Mass (slice/cake)
  - homomerous and separable
- Place/area (England/Europe)
  - not separable, but homomerous.

(homomerous = same kind as whole)

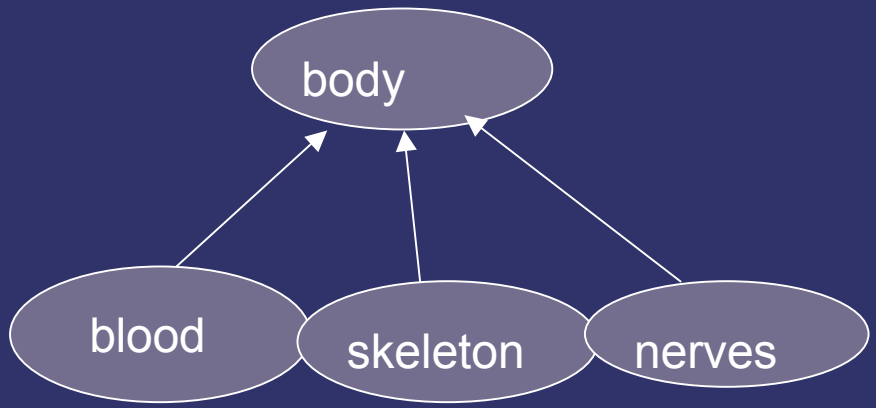
A cohesive organizational principle is required throughout the meronomy

- Segmental parts



- Spatially cohesive
- Encountered sequentially.

- Systemic parts



- Spatially interpenetrating
- Greater functional unity

D.A.Cruse, 1986

There is not one all inclusive meronomy to describe the universe.

- A well formed meronomy should consist of elements of the same type:
  - Cohesive physical objects
  - Geographic areas
  - Abstract nouns
- At the top of the hierarchy there is a whole
  - *i.e.* we do not say heart part_of cardiovascular system part_of body part_of population part_of biomass

D.A.Cruse, 1986

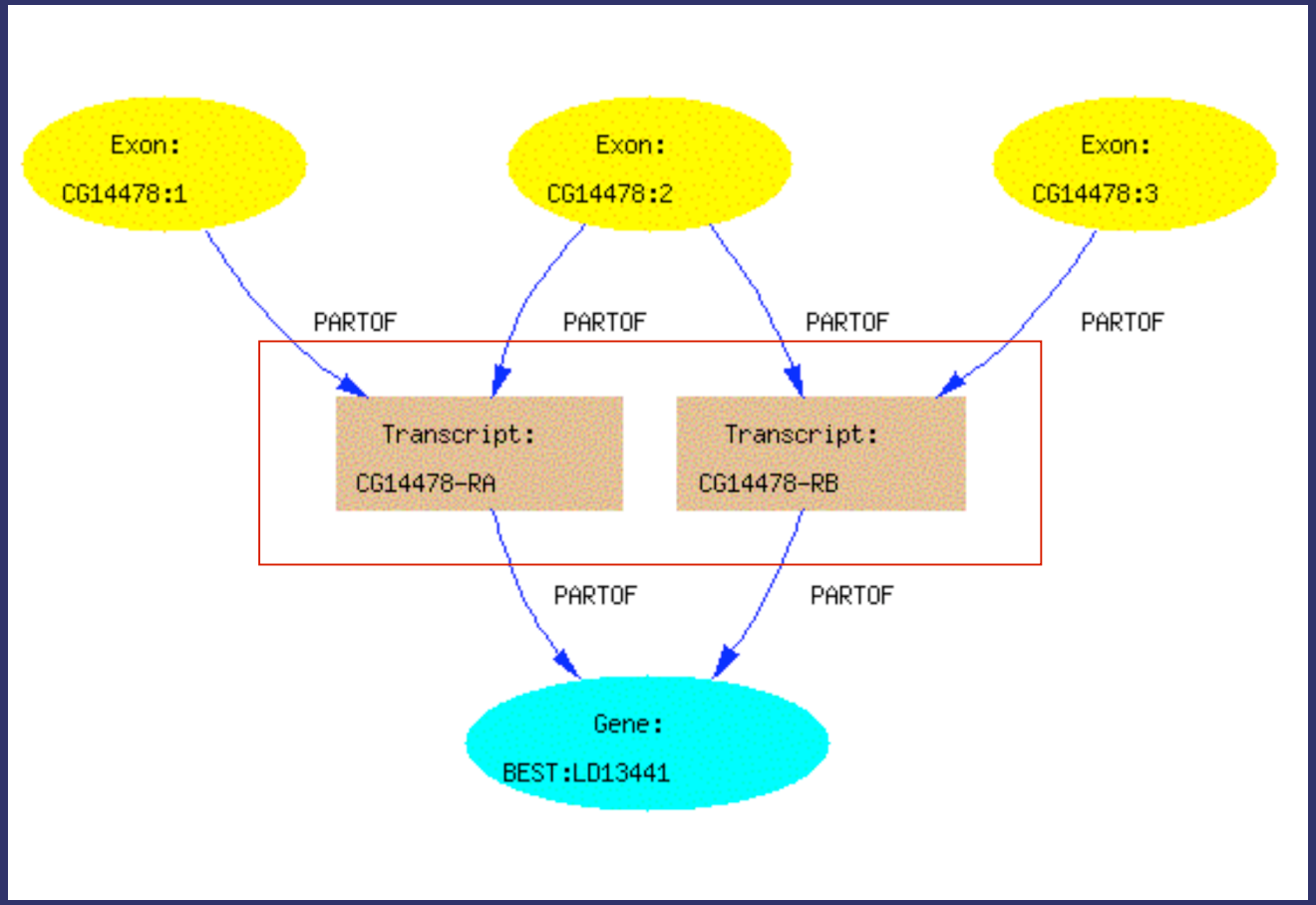
Classical Extensional Mereology

- The formal properties of parts:
  1. If A is a proper part of B then B is not a part of A (nothing is a proper part of itself)
  2. If A is a part of B and B is a part of C then A is a part of C
- Because of these rules, we can apply some functions to parts…

Functions that operate on parts

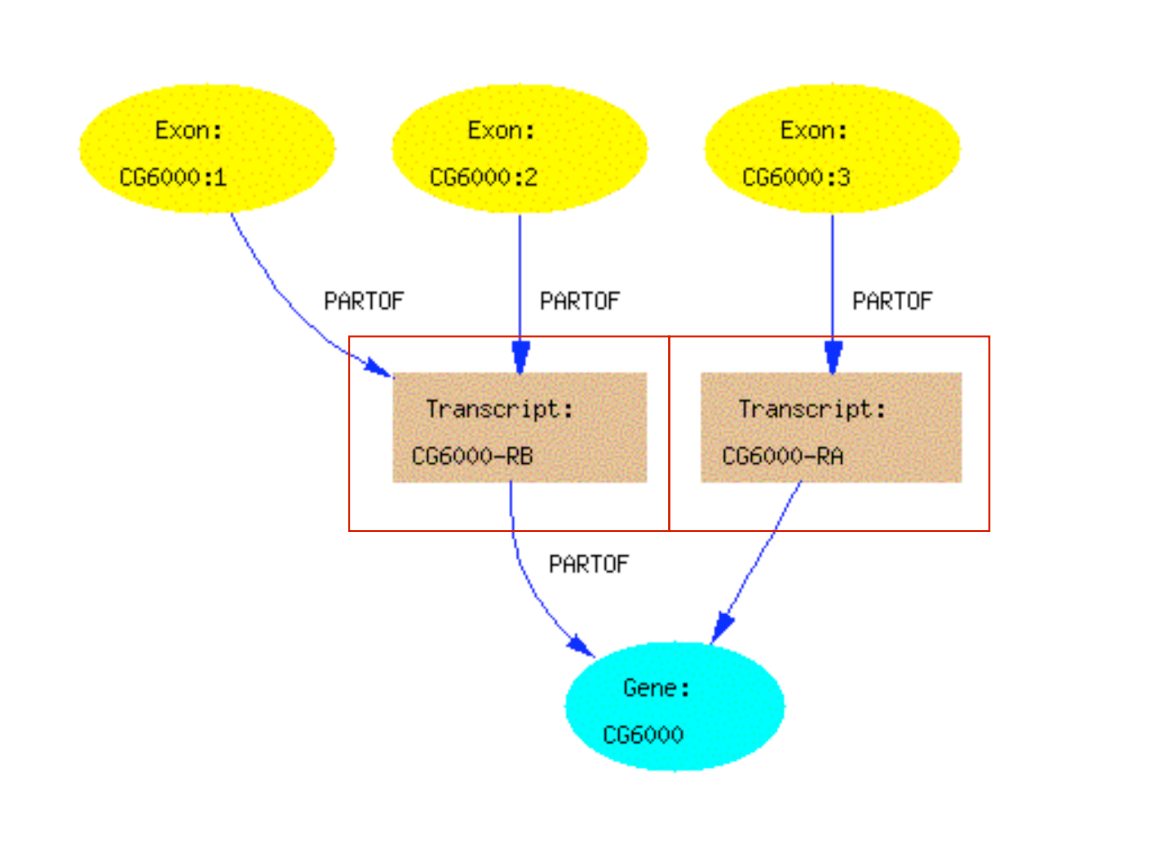
- Overlap
- Disjoint
- Binary product
- Binary sum
- Difference

Individuals overlap if they have a part in common.



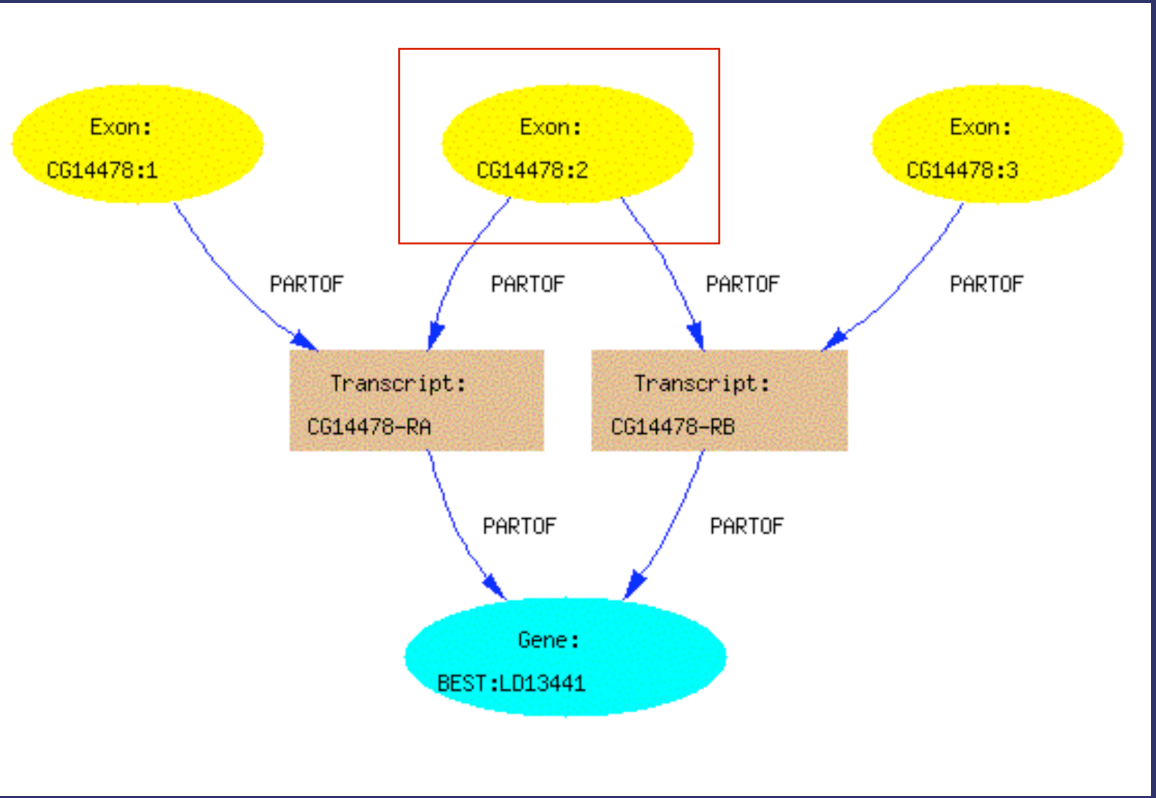
overlap

Individuals are disjoint if they share no parts in common.



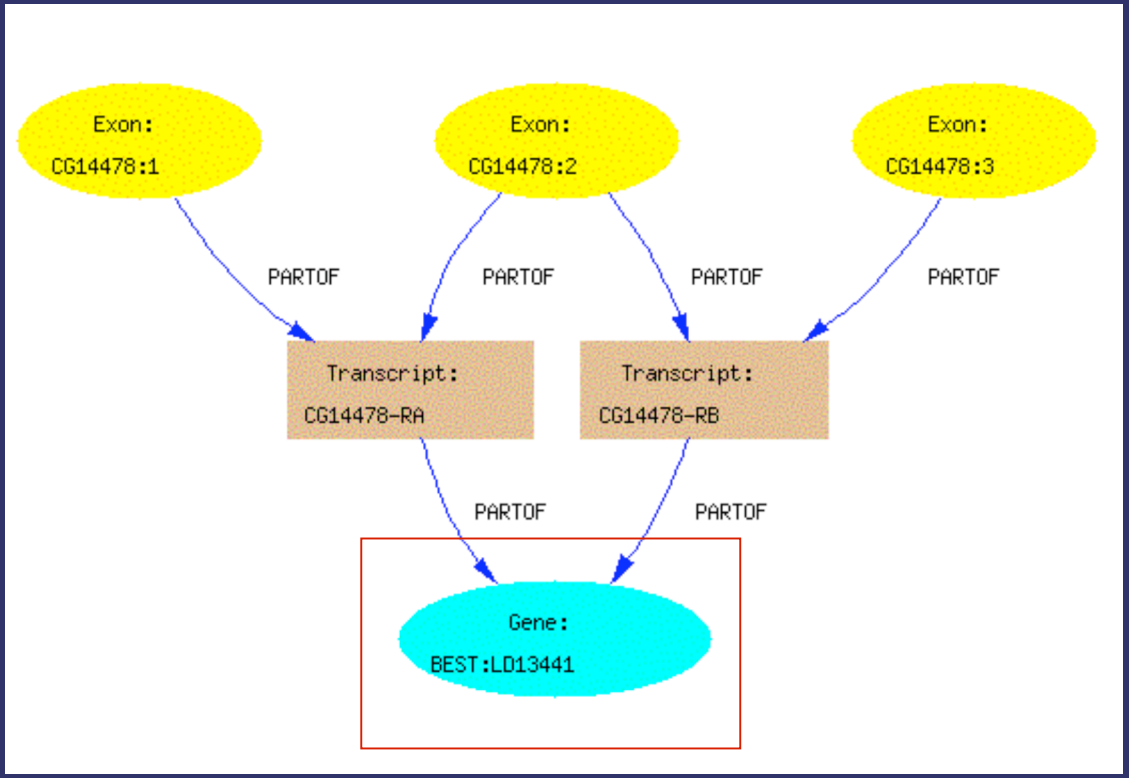
disjoint

When two individuals overlap it is the parts that they share in common.



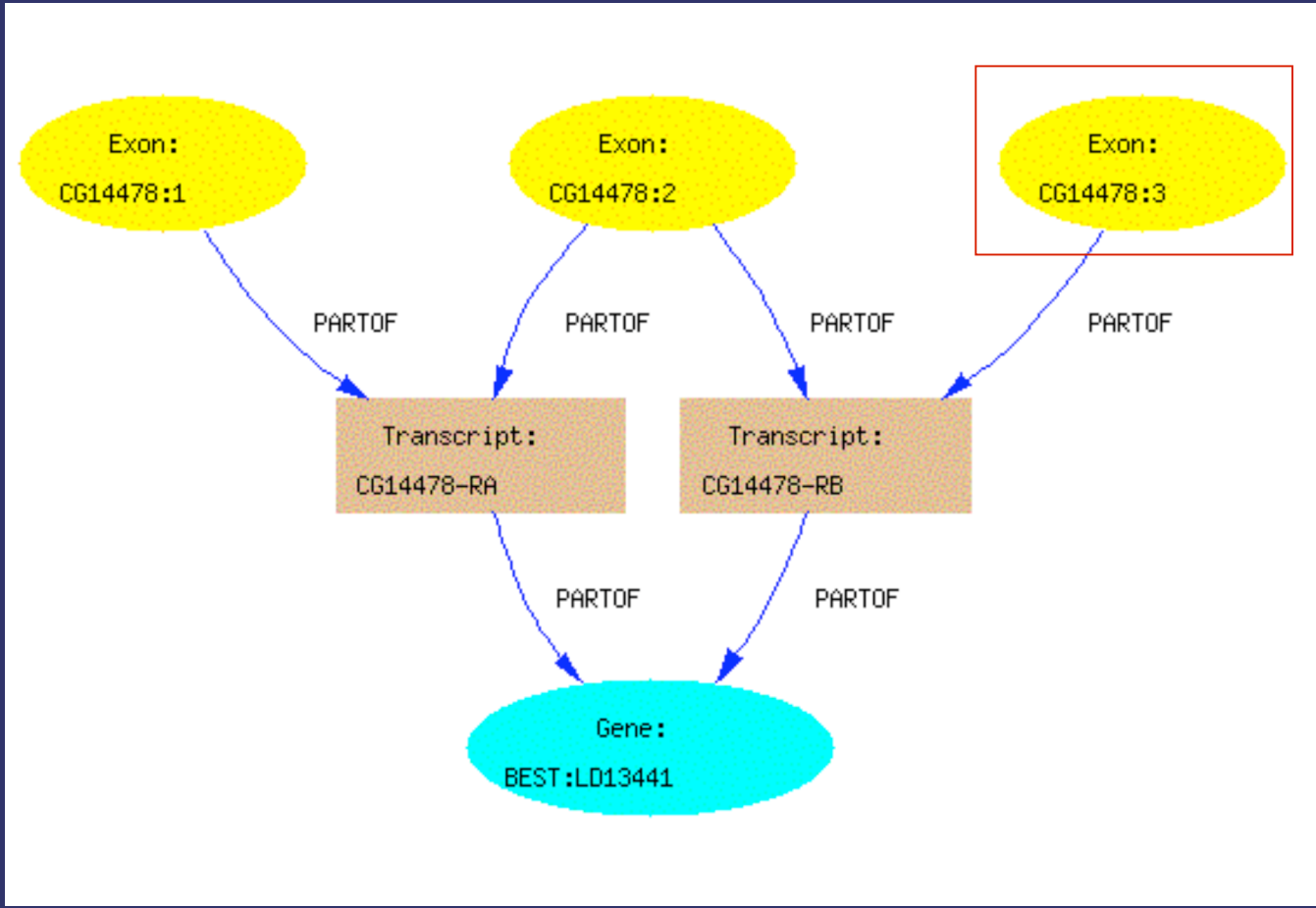
Binary product

The individuals wholly containing at least one of x and y



Binary sum

The parts contained in x which are not parts of y, where x is not itself a part of y.



difference

Given these functions…

(and some sequence marked up with the SO)

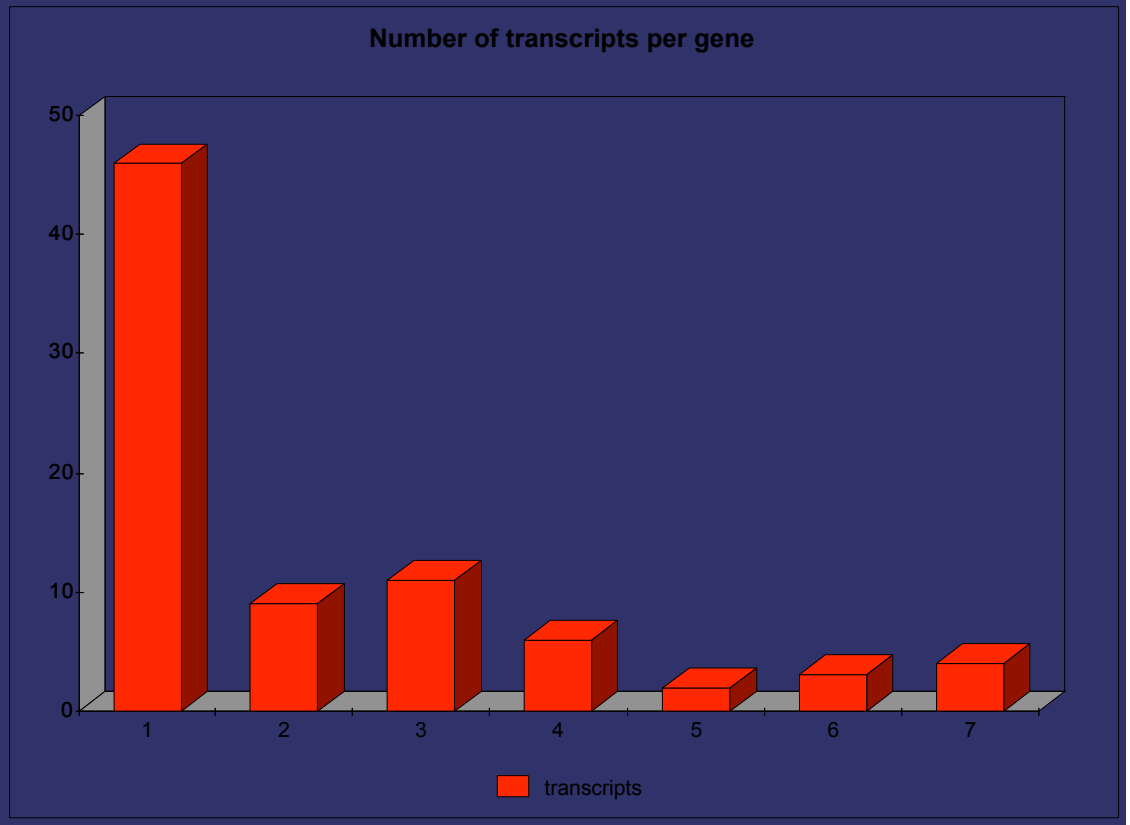
We can ask these questions…

- What are the genes with 'disjoint' transcripts?
- How often are exons unique to a transcript?
- Which exons are in all the transcripts for the gene?

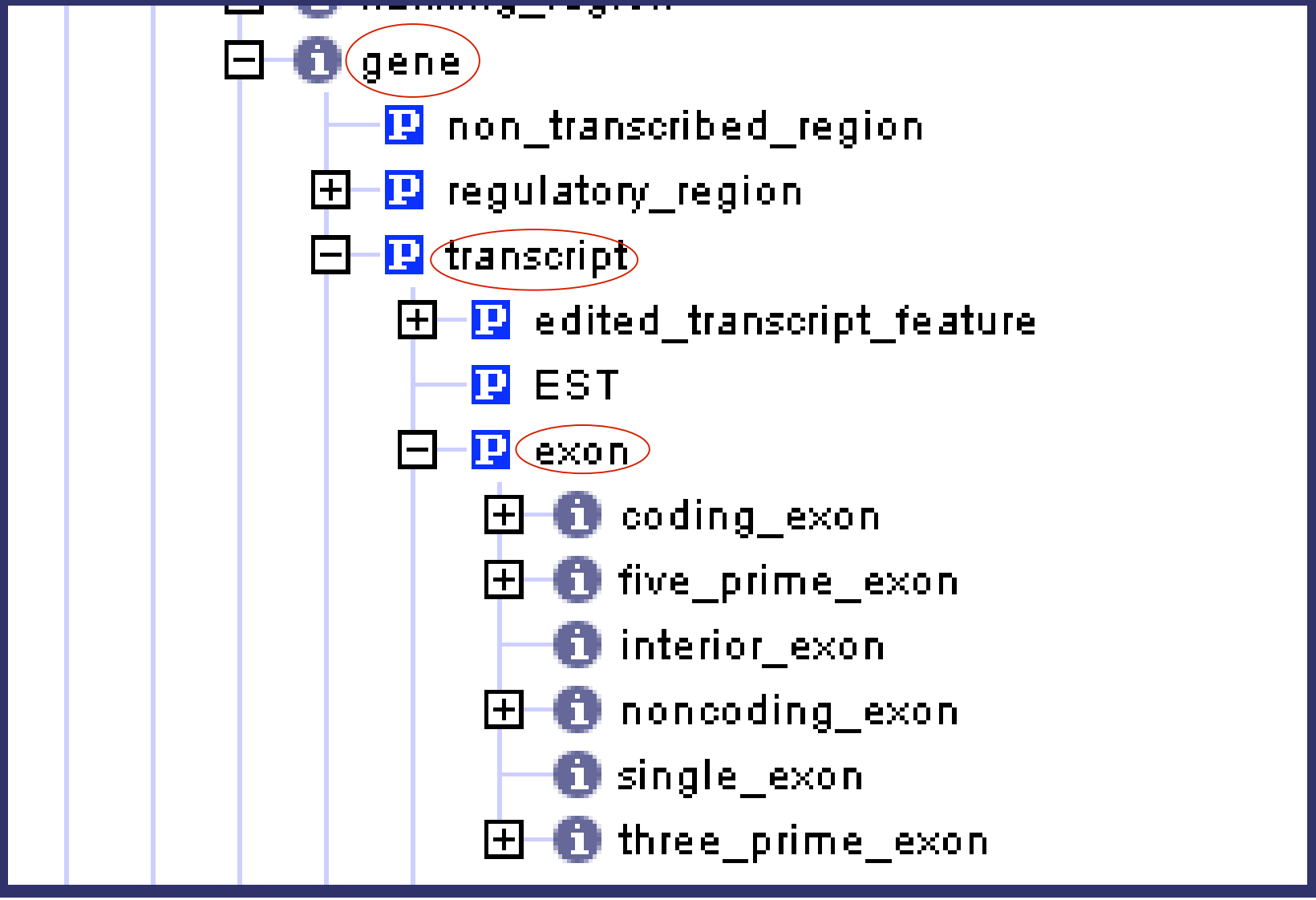
D.mel Chromosome 4

- 82 genes
- 179 transcripts
- 750 exons

- 36 multi transcript genes
- 46 single transcript genes

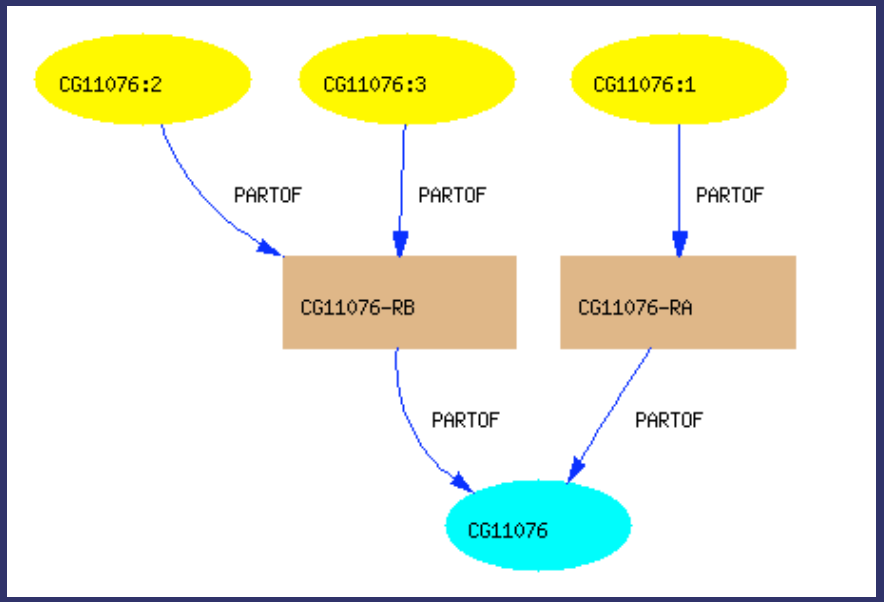


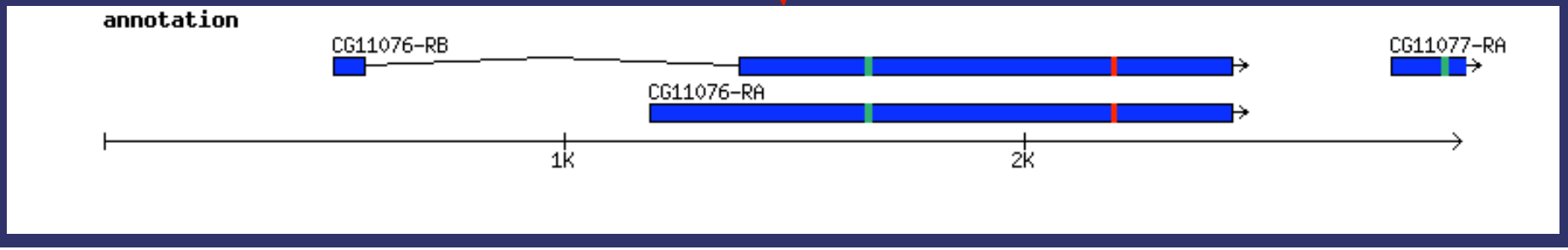
Marked up sequence using these parts of SO….



Which genes on chromosome 4 have 'disjoint' transcripts?

- only one gene out of 82





How often are exons unique to a transcript?
How often does an exon appear in all of the transcripts?

Exon part of single transcript	285
Exon in all transcripts	243 (52%)
Exon in one transcript	148 (32%)
Exon in > 1 but < all	74 (16%)

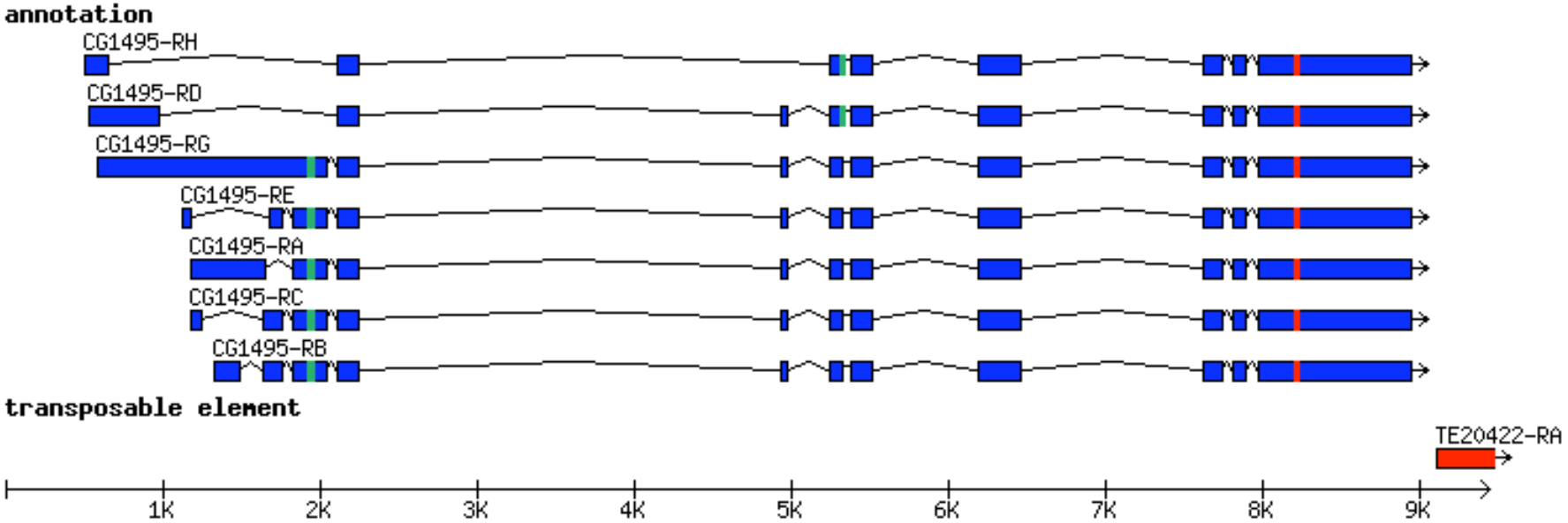
More Questions…

- For exons that occur in all the transcripts, How often are they coding exons?
- For exons that occur in only one of the transcripts, how often are they noncoding?
- Do unique exons contain the stop codon more often than exons in all the transcripts?

	All exons	Single exon	Between 1 and all
coding	221 (91%)	60 (40%)	47 (63%)
Not coding	2 (1%)	88 (60%)	19 (25%)
Both coding and non coding	20 (8%)	N/A	8 (10%)
Contains start	24 (10%)	25 (16%)	20 (27%)
Contains end	26 (11%)	15 (10%)	9 (12%)

Even more questions…

- Are single exons evolving faster than shared exons?
  - Ka/Ks coding exons – compare with pseudoobscura.
- Can we validate alternate transcripts?



Beaucoup Possibilities

- Evidence networks
- Transcription factor & other binding sites
- Intersection graphs
  - precompute cytology
  - insertions + gene features
- Correlate with Yeast 2 hybrid / P-P interactions

Summary

- Achieve a balance between ease of use and richness of expression
- GFF3 and SO(fa) freeze (Michael TBD???)
- PART_OF relationships provide new operations on the data
- Already gaining the benefits of the PART_OF relationships that enable inferences from genomic annotations

Low-down

- Taking longer than we thought to stabilize
- Using "slim" for SOFAing
- Issues with protein motifs and sequence variations
- Phenotype needs are urgent
- Image annotation haunts me

Acknowledgments

- Michael Ashburner
- Lincoln Stein
- Richard Durbin
- J. Michael Cherry
- Judith Blake
- *Karen Eilbeck*
- Christopher J. Mungall
- Mark Yandell
- George Hartzell
- Colin Wiel

- Peter Good and the NIH

http://song.sourceforge.net

References

- D.A. Cruse – Lexical Semantics, Cambridge University Press 1986
- Peter Simons – Parts a Study in Ontology, Oxford University Press 1987