Synteny data in ParameciumDB

Olivier Arnaiz, Mark Gibson and Linda Sperling

(ParameciumDB working document heavily inspired by the FlyBase comparative_implementation_standard by Peili Zhang & coll.)

6 June 2006

A data implementation standard for comparative genomics in Chado has been described in a FlyBase document[1]. That data implementation was formulated by the developers at FlyBase (http://www.flybase.org) and adopted by FlyBase for production. TIGR is also representing comparative data (for gene families) in Chado[2].

In order to present syntenic releations resulting from Whole Genome Duplication in ParameciumDB (http://paramecium.cgm.cnrs-gif.fr), we have attempted to generalize and adapt the FlyBase and TIGR implementations.

In this document, we first show a generalized model for comparative genomics in Chado. In a second section, we illustrate the way we can apply this general model to the representation of (1) paralogous regions, (2) paralogous genes termed "ohnologs" (paralogs generated by whole genome duplication) and (3) syntenic regions.
Chado standards for genomic sequences and annotation are used.

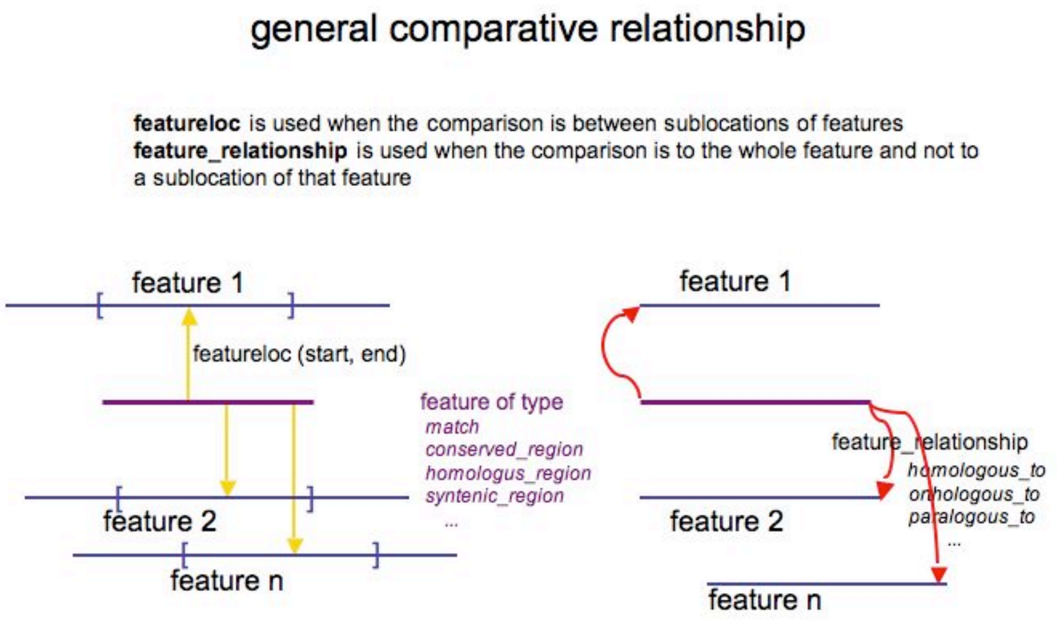
General model for comparative genomic data in Chado

The general model can involve 2 or more sequence features hence can capture pairwise comparisons as well as gene families. A feature of type

[1] P. Zhang, D. Emmert, P. Zhou, B. Gelbert and the FlyBase consortium. compartive_implementation_standard.doc.

[2] Posted to the gmod-schema mailing list 5-Jun-2006 by Samuel Angiuoli, angiuoli@tigr.ORG.

'match', 'syntenic_region', 'orthologous_region' or 'paralogous_region' is created. The comparison can be represented through 2 to n featurelocs, or alternatively through 2 to n feature_relationships, as shown in Figure 1. We use featurelocs when sublocations of features are being related (see examples of paralogous and syntenic regions below), and feature_relationships of type 'homologous_to', 'orthologous_to' or 'paralogous_to' when entire features are being related (see example of paralogous genes below).



general comparative relationship

featureloc is used when the comparison is between sublocations of features
feature_relationship is used when the comparison is to the whole feature and not to a sublocation of that feature

Figure 1. General model of implementation of comparative genomics in Chado.

We propose this general model as a way to create multiple-feature relationships in Chado, i.e. as a mechanism for grouping features, in much the same way as Sam Angiuoli of TIGR is capturing gene familes in Chado.

Specific examples of synteny data in ParameciumDB

Paralogous regions

Paralogous regions are implemented similarly to alignment match features exactly as in the FlyBase comparative genomics implementation, the only difference being that the paralogous regions are stored as features of type 'paralogous_region' whereas alignment match features are of type 'match'**.** Unlike FlyBase, we do not use a fake organism of genus 'Computational' and species 'result' since we are working within a single species. The paralogous_region features are linked to the appropriate analysis through analysisfeature links. Each paralogous region has two featureloc records that localize the paralogous region on two different chromosomes, with different featureloc.rank. Fig. 2 illustrates this implementation.

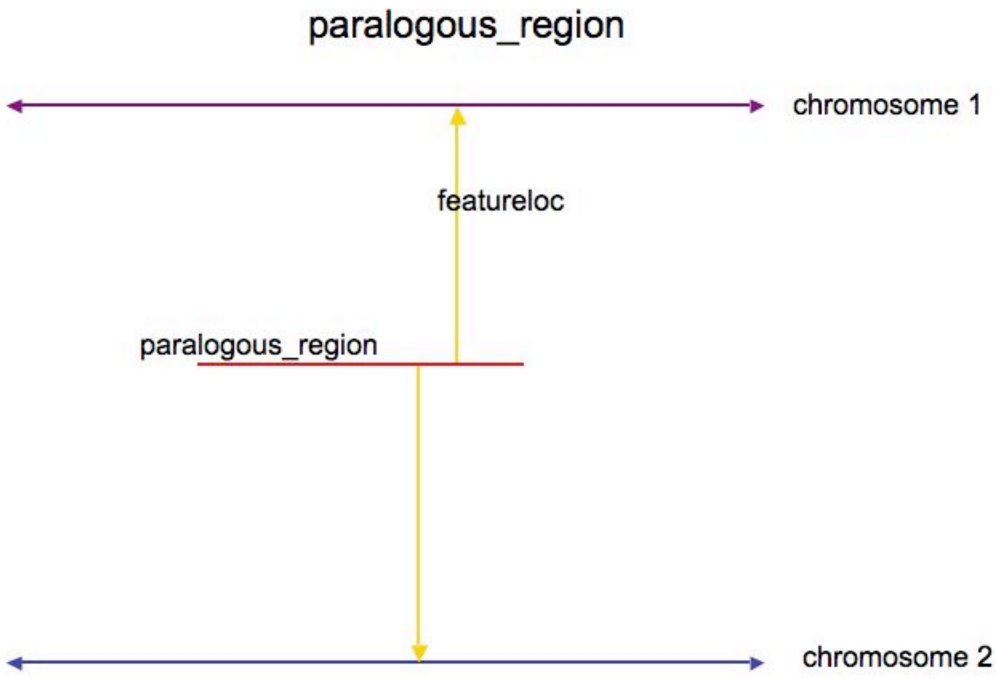


Figure 2. Implementation of paralogous region features in Paramecium.

Paralogous genes

A relationship between paralogous genes is represented in ParameciumDB Chado using a feature_relationship of type "paralogous_to". The data storage is depicted in Figure 3. Although the simplest way to do this would be to directly create a feature_relationship between the two genes (or two relationships since "paralogous_to" is symmetric), this solution would not scale up if more genes were added. We therefore prefer to first create a feature of type 'paralogous_region'. We then create feature_relationships between the paralogous_region and each of the paralogous genes. The annotations (transcript, exons, polypeptide) are standard and are not shown.

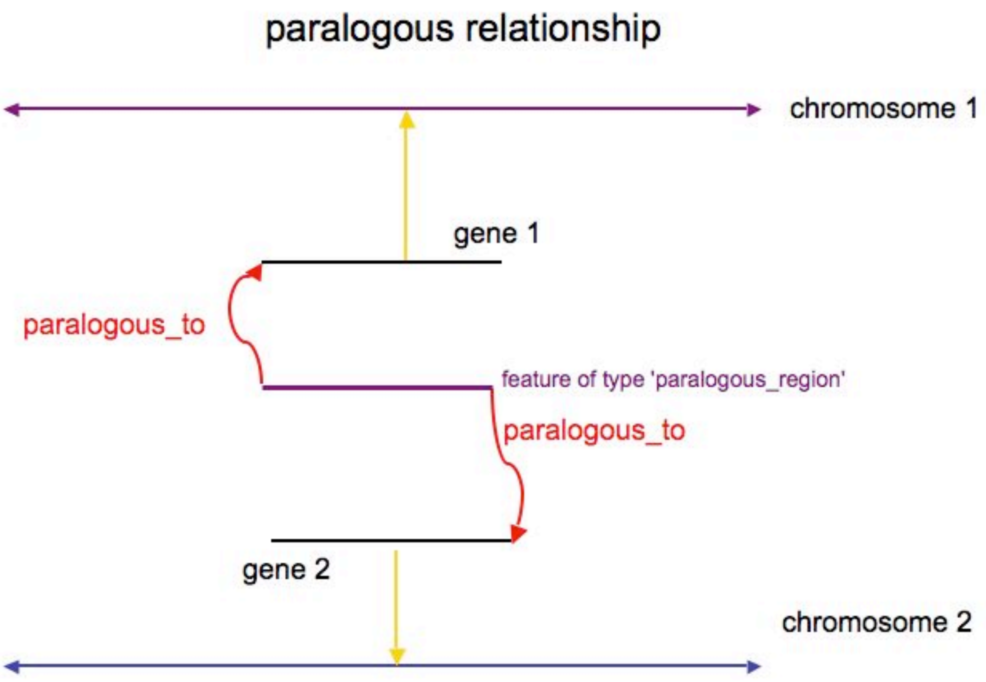


Figure 3. ParameciumDB Chado representation of a paralogy relationship between genes.

Synteny

Syntenic regions between chromosomes are stored as features of type 'syntenic_region', essentially as in the FlyBase implementation. The spans of these syntenic regions on each of the chromosomes are stored as featureloc records, with different ranks. A feature of uniquename 'syntenic_block:1', type 'syntenic_region' is created. The feature 'syntenic_block:1' has two featureloc records, one to each chromosome exactly as for "paralogous_regions". The implementation of the syntenic relationship between chromosomes is shown in Figure 4.

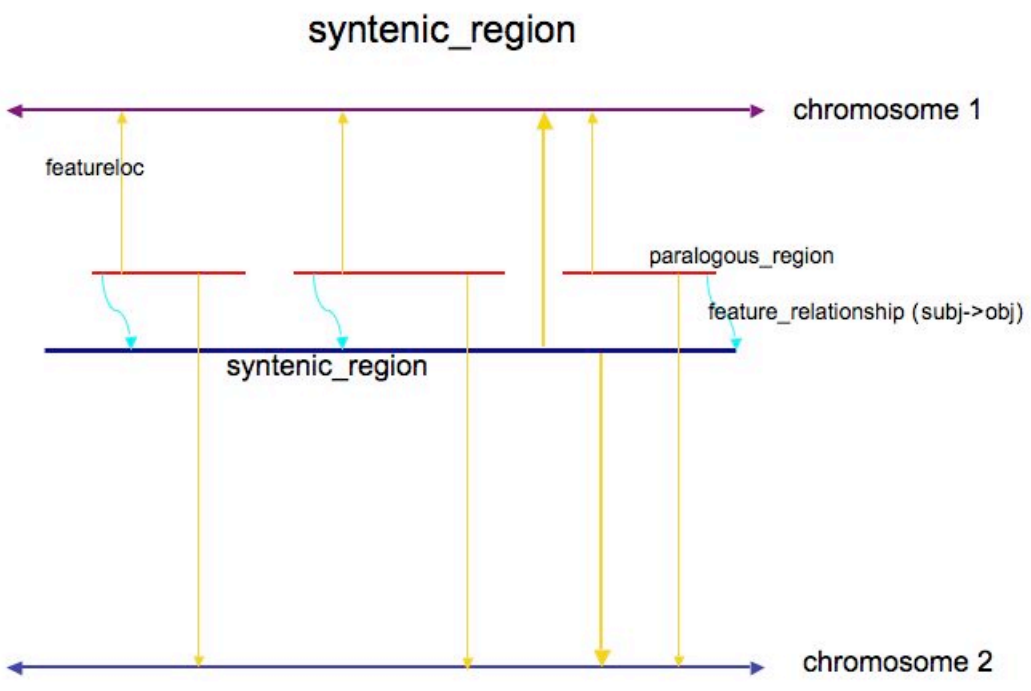


Figure 4. Chado representation of syntenic relationship between a pair of chromosmes, based on paralogous_regions as shown in Figure 2.

In addition, the component paralogous_regions of a syntenic region are recorded through feature_relationship records between the syntenic region feature and the relevant paralogous region features which are implemented with a feature_relationship type 'part_of' and feature_relationship.rank 0 (the default value) i.e. the different paralogous regions within each syntenic block are not ranked with respect to each other.