

# *Pathway Tools Schema and Semantic Inference Layer*

## *Genes, Operons, and Replicons*

**Peter D. Karp**

**SRI International**

# References

- **Pathway Tools User's Guide, Volume I**
  - Appendix A: Guide to the Pathway Tools Schema
- **Ontology Papers section of <http://biocyc.org/publications.shtml>**
  - "An Evidence Ontology for use in Pathway/Genome Databases,"
  - "An ontology for biological function based on molecular interactions,"
  - "Representations of metabolic knowledge: Pathways,"
  - "Representations of metabolic knowledge,"

# *Frame Data Model*

- **Frame Data Model -- organizational structure for a PGDB**
- **Knowledge base (KB, Database, DB)**
- **Frames**
- **Slots**

# *Knowledge Base*

- **Collection of frames and their associated slots, values, facets, and annotations**
- **AKA: Database, PGDB**
- **Can be stored within**
  - An Oracle or MySQL DB
  - A disk file
  - Pathway Tools binary program

# Frames

- Entities with which facts are associated

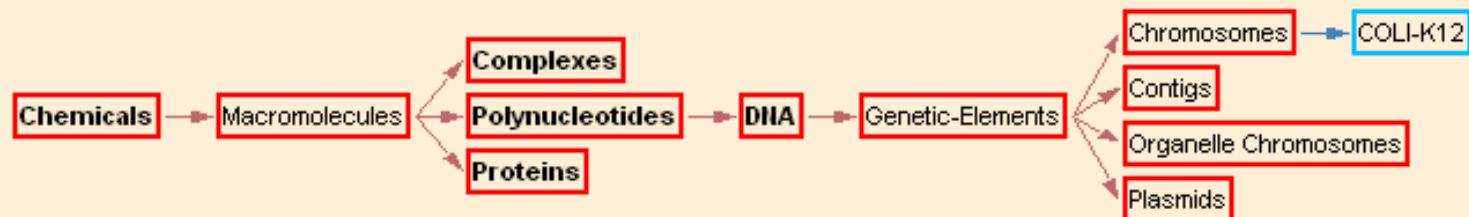
- Kinds of frames:

- Classes: Genes, Pathways, Biosynthetic Pathways
- Instances (objects): trpA, TCA cycle

- Classes:

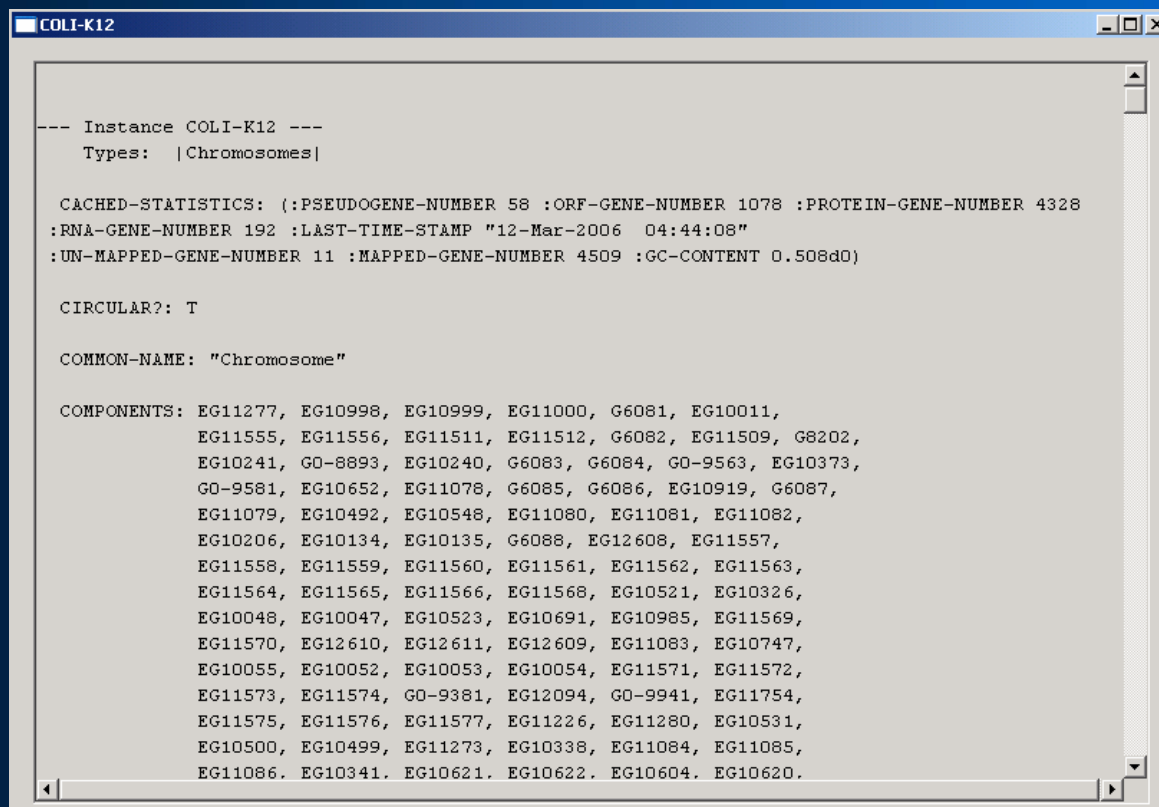
- Superclass(es)
- Subclass(es)
- Instance(s)

- A symbolic frame name (id, key) uniquely identifies each frame



# Slots

- Encode attributes and properties of a frame
- Represent relationships between frames
  - The value of a slot is the identifier of another frame



```
COLI-K12
--- Instance COLI-K12 ---
Types: |Chromosomes|

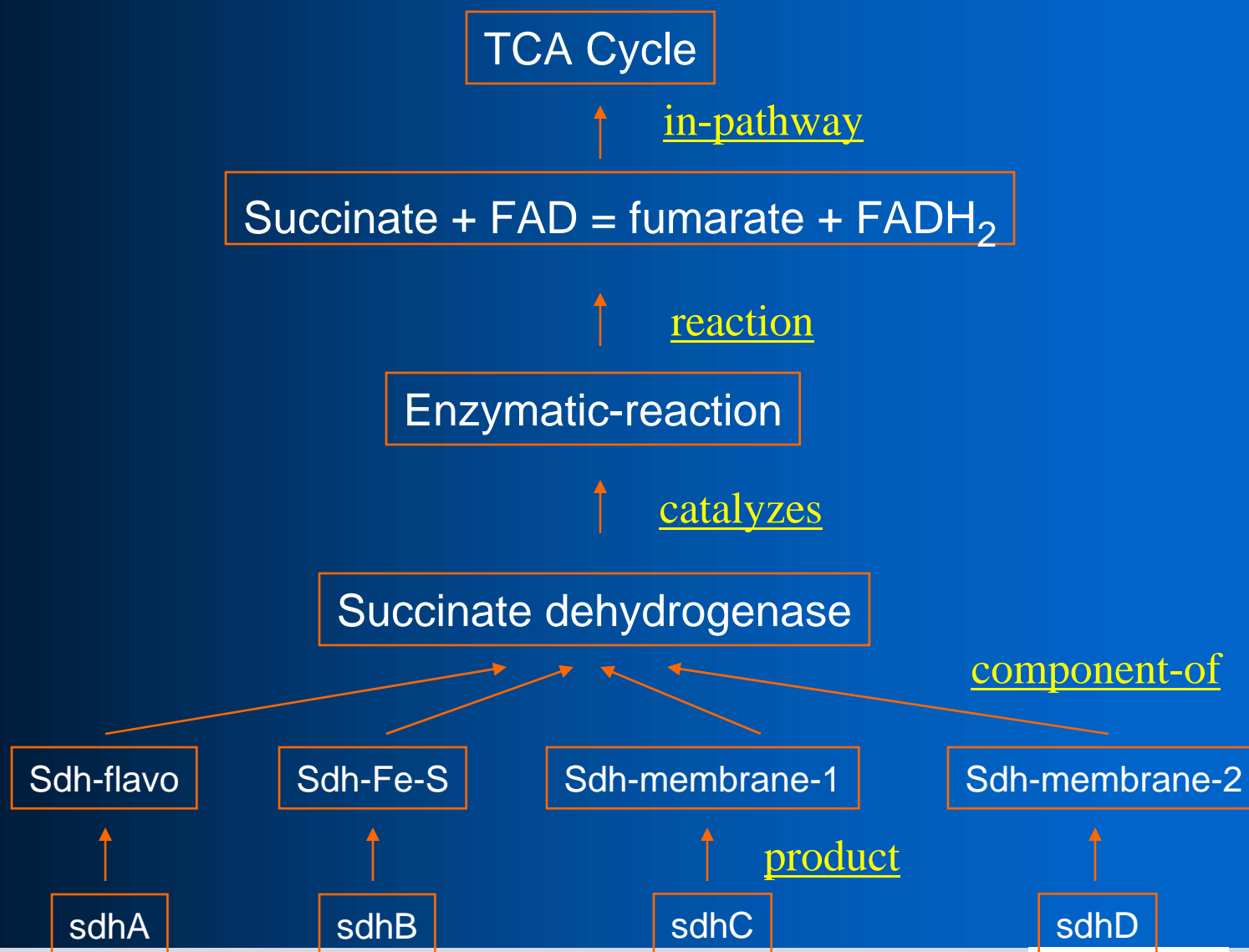
  CACHED-STATISTICS: (:PSEUDOGENE-NUMBER 58 :ORF-GENE-NUMBER 1078 :PROTEIN-GENE-NUMBER 4328
:RNA-GENE-NUMBER 192 :LAST-TIME-STAMP "12-Mar-2006 04:44:08"
:UN-MAPPED-GENE-NUMBER 11 :MAPPED-GENE-NUMBER 4509 :GC-CONTENT 0.508d0)

CIRCULAR?: T

COMMON-NAME: "Chromosome"

COMPONENTS: EG11277, EG10998, EG10999, EG11000, G6081, EG10011,
EG11555, EG11556, EG11511, EG11512, G6082, EG11509, G8202,
EG10241, GO-8893, EG10240, G6083, G6084, GO-9563, EG10373,
GO-9581, EG10652, EG11078, G6085, G6086, EG10919, G6087,
EG11079, EG10492, EG10548, EG11080, EG11081, EG11082,
EG10206, EG10134, EG10135, G6088, EG12608, EG11557,
EG11558, EG11559, EG11560, EG11561, EG11562, EG11563,
EG11564, EG11565, EG11566, EG11568, EG10521, EG10326,
EG10048, EG10047, EG10523, EG10691, EG10985, EG11569,
EG11570, EG12610, EG12611, EG12609, EG11083, EG10747,
EG10055, EG10052, EG10053, EG10054, EG11571, EG11572,
EG11573, EG11574, GO-9381, EG12094, GO-9941, EG11754,
EG11575, EG11576, EG11577, EG11226, EG11280, EG10531,
EG10500, EG10499, EG11273, EG10338, EG11084, EG11085,
EG11086. EG10341. EG10621. EG10622. EG10604. EG10620.
```

# Slot Links



# Slots

- **Number of values**
  - Single valued
  - Multivalued: sets, bags
- **Slot values**
  - Any LISP object: Integer, real, string, symbol (frame name)
- **Every slot is described by a “slot frame” in a KB that defines meta information about that slot**
  - Datatype, classes it pertains to, constraints
  - Two slots are inverses if they encode opposite relationships
    - ◆ Slot Product in class Genes
    - ◆ Slot Gene in class Polypeptides



# *Pathway Tools Ontology / Schema*

- **Ontology classes: 1621**
  - Many datatypes from genomes to pathways
  - Classification schemes for pathways, chemical compounds, enzymatic reactions (EC system)
  - Cell Component Ontology
  - Protein Feature ontology
- **Comprehensive set of 221 attributes and relationships**
- **Evidence codes, supporting citations**

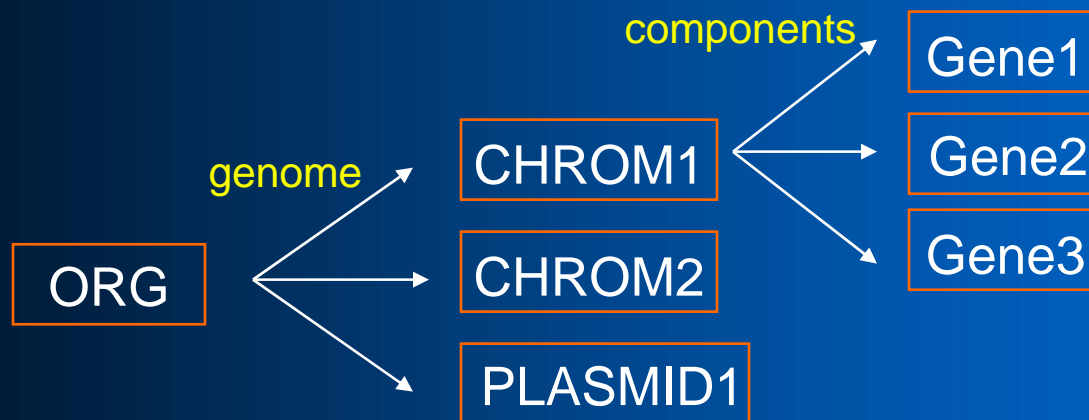
# *Root Classes in the Pathway Tools Ontology*

- **Chemicals** -- All molecules
- **Polymer-Segments** -- Regions of polymers
- **Protein-Features** -- Features on proteins
- **Paralogous-Gene-Groups**
  
- **Organisms**
  
- **Enzymatic-Reactions** -- Link enzymes to reactions they catalyze
- **Generalized-Reactions** -- Reactions and pathways
  
- **CCO** -- Cell Component Ontology
- **Evidence** -- Evidence ontology
  
- **Notes** -- Timestamped, person-stamped notes
- **Organizations**
- **People**
- **Publications**

# *Use GKB Editor to Inspect the Pathway Tools Ontology*

- **GKB Editor = Generic Knowledge Base Editor**
- **Type in Navigator window: (GKB) or**
- **[Right-Click] Edit->Ontology Editor**
  
- **View->Browse Class Hierarchy**
- **[Middle-Click] to expand hierarchy**
- **To view classes or instances, select them and:**
  - Frame -> List Frame Contents
  - Frame -> Edit Frame

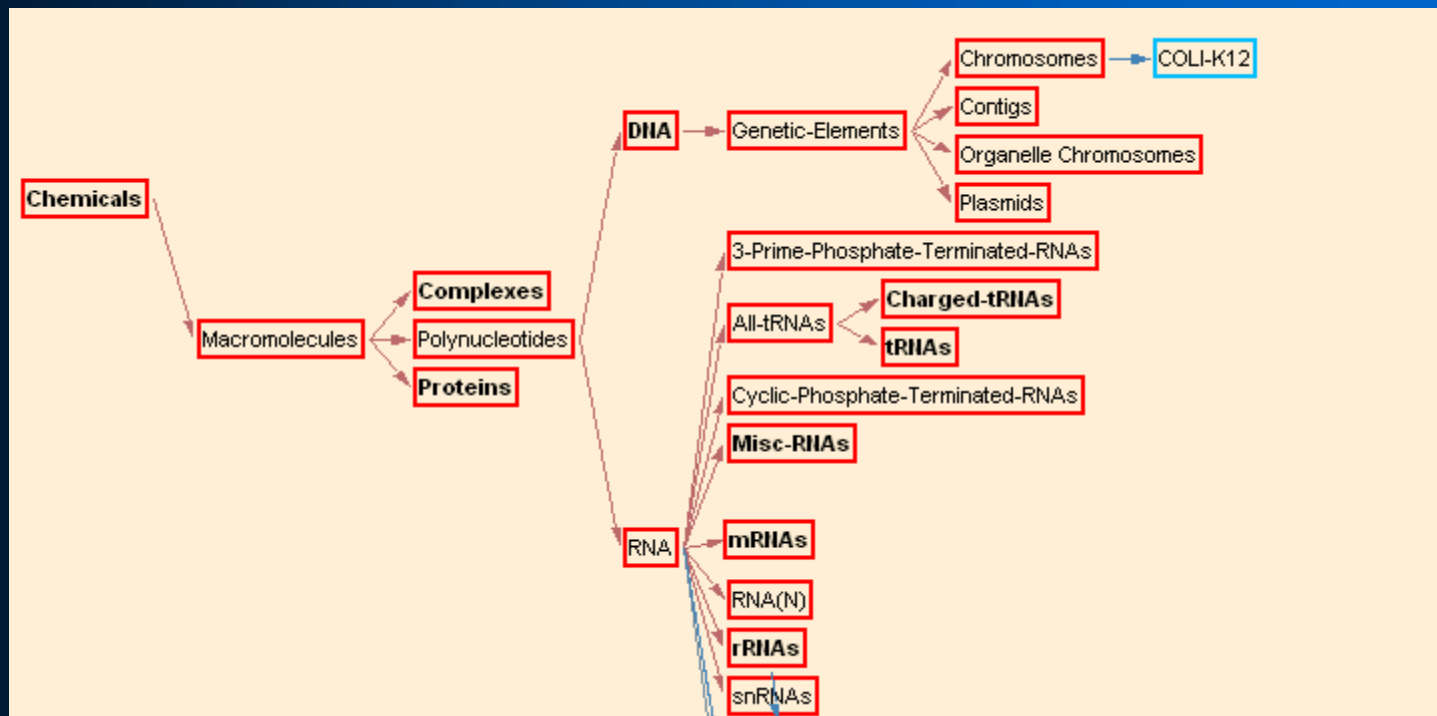
# Representing a Genome



- **Classes:**

- ORG is of class Organisms
- CHROM1 is of class Chromosomes
- PLASMID1 is of class Plasmids
- Gene1 is of class Genes

# Polynucleotides

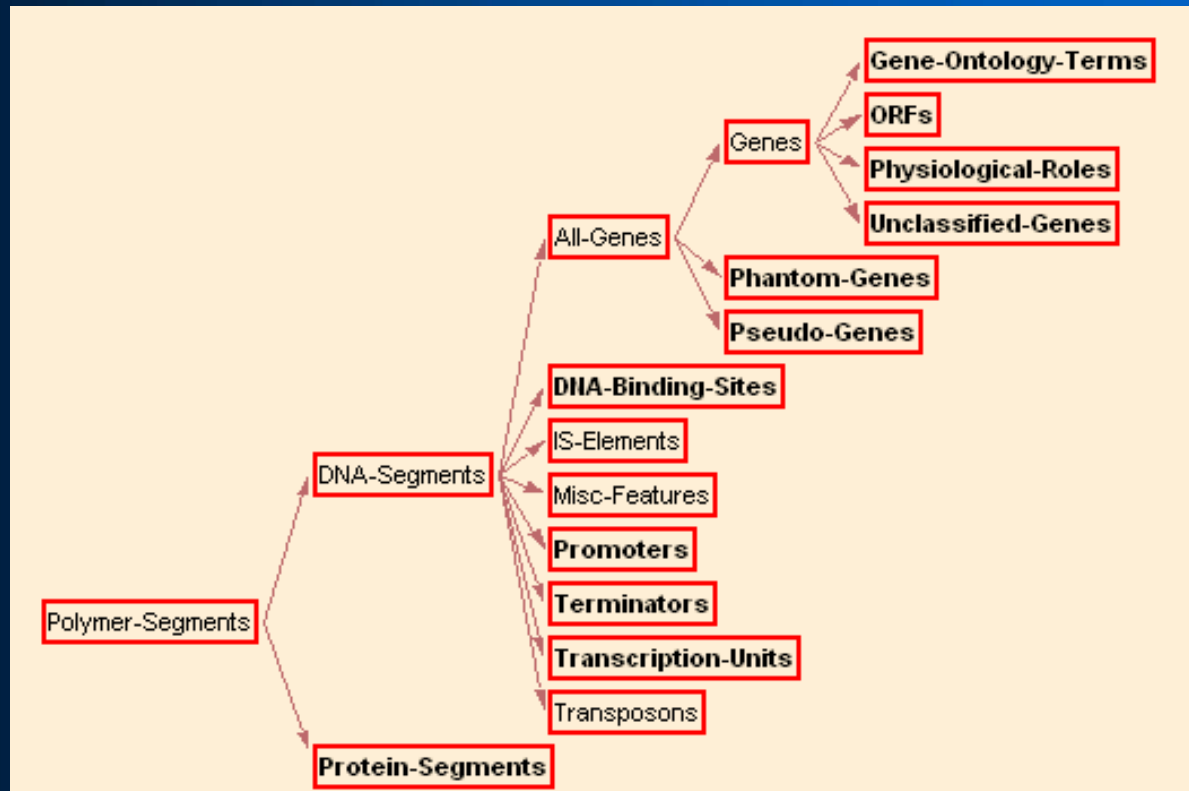


Review slots of COLI and of COLI-K12

# *Genetic-Elements*

- **Sequence is stored in a separate file**

# Polymer-Segments



Review slots of Genes

# *Complexities of Gene / Gene-Product Relationships*

- **The Product of a gene can be an instance of Polypeptides or RNAs**
- **An instance of Polypeptides can have more than one gene encoding it**
- **Sequence position:**
  - Nucleotide positions of starting and ending codons specified in Left-End-Position and Right-End-Position (usually greater, except at origin)
  - Transcription-Direction + / -
- **Alternative splicing:**
  - Nucleotide positions of starting and ending codons specified in Left-End-Position and Right-End-Position
  - Intron positions specified in Splice-Form-Introns of gene product
    - ◆ (200 300) (350 400)



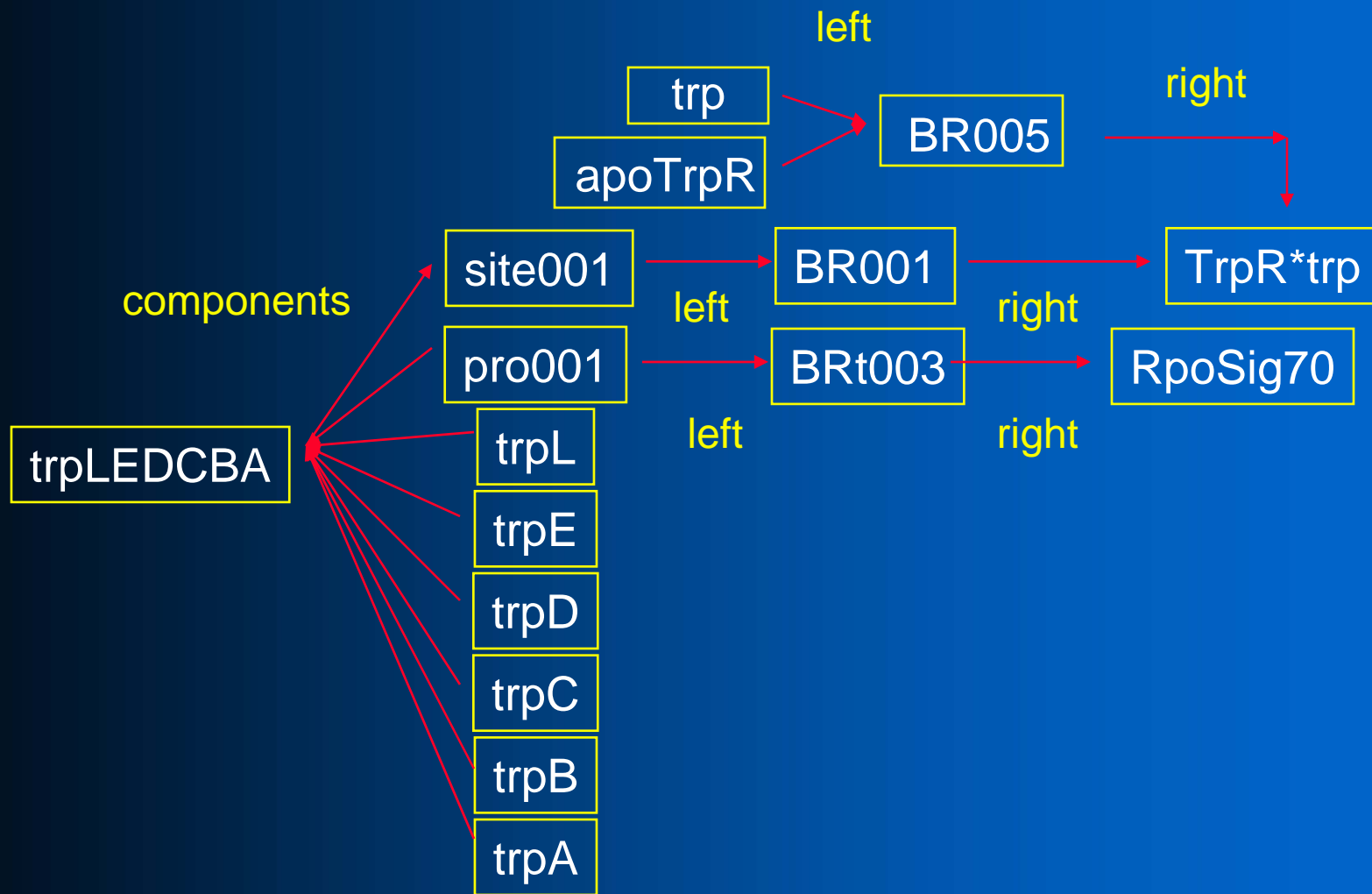
## *Semantic Inference Layer*

- **Chromosome-of-gene(gene)**
- **Adjacent-genes?(g1 g2)**
- **Neighboring-genes-p(g1 g2 n)**
- **does-gene-code-for-protein-?(gene)**

# *Operons and Transcription Units*

- **Operon:** A set of two or more genes that are transcribed as a unit. May include multiple promoters.
- **Transcription Unit:** A set of one or more genes that are transcribe as a unit from a single promoter.

# Ontology for Transcriptional Regulation



# *Transcriptional Regulation*

- **Transcription-Units**

- Its Components include genes, promoter, terminator, TF-binding sites

- **Binding-Reactions are defined for:**

- Each TF to its binding site
- RNA polymerase to each promoter

# *Semantic Inference Layer*

- **Operon-of-gene(gene)**
- **Genes-in-same-operon(gene)**

# *Semantic Inference Layer*

- **regulators-of-gene-transcription(gene) => protein-list**
- **transcription-unit-promoter(tu) => promoter**
- **transcription-unit-genes(tu) => gene-list**
- **transcription-unit-binding-sites(tu) => binding-site-list**
- **transcription-unit-transcription-factors(tu) => TF-list**
  - All TFs that bind to binding sites in TU
- **transcription-unit-terminators**
- **binding-sites-affecting-gene(gene)**
- **binding-site-transcription-factors(site)**
- **binding-site-promoters**
- **binding-site-transcription-units**
- **promoter-binding-sites**

## *Example Computations*

- **Given TF, find all metabolic reactions it regulates**
- **Find all TFs that bind a given ligand**
- **Find all TFs that bind more than one ligand and enumerate them**