Awesomeness
Design Document

Back to Databasics

Alissa Lawrence  aklawren@calpoly.edu
Devlin Cronin    rdcronin@calpoly.edu
Matthew Werner   mwerner@calpoly.edu
Michael Gage     mgage@calpoly.edu
Spencer Woodworth swoodwor@calpoly.edu
Entity Sets

Experiment
The representation of the studies performed on cells of a given species in two different stages of development. The attribute ID represents the date and identifier of the experiment conducted. The attribute Comparison indicates which two stages of development were examined; for instance, this might be the comparison between lactating and non-lactating animals. Species refers to the specie from which the cell was taken.

Gene
The gene entity is the gene that has been tested by the experiment. The attribute Species indicates from which species the cell has been taken; currently, this will be limited to bovine, but may expand in the future. Other attributes are Name, Abbreviation, Chromosome, which give information about the gene. Beginning Site and Ending Site show where the gene is located. Promoter is the unique 2,000 character base set.

Regulatory Element
This represents the regulatory element that is found. Factor is the factor to which the regulatory element binds. Sequence represents the character set of the base set that comprises the regulatory element. Beginning, Sense, and Length give information about the location of the regulatory element. Model refers to the source of the data which indicates the relationship. La, La/, Lq, Lpv, Sc, Sm, Spv, Ppv are numerical data which show the probability of the sequence being correct.

PSG1
This represents the data of P-Value, N, Rate, La, MAC, MIC, FIDs from the PSG worksheet in the data.

FACs
This is the set of FAC values from the PSG sheet; since this is optional, we will represent this as a separate table to avoid populating excessively with NULL. The FAC-Number is the FAC value of the table. Note: We are operating under the assumption that the FAC values and the T-Numbers refer to the same data.

Aggregation
Gene and Experiment form an aggregation. This will be used by the relationship between these two entities and Regulatory Elements and PSG1.

Relationships
Experiment is connected to Gene via the relationship of Performed On; this shows about which gene the experiment provides data. This is a one-to-many relationship, as one experiment could provide data about multiple different genes. There is a participation constraint on Experiment; each Experiment is required to link to at least one Gene. This relationship will have a single attribute, which is Regulation. Because Gene does not inherently enough information to uniquely distinguish each entry, this is a weak entity set connecting to Experiment.

Regulatory Elements connect to the aggregation of the Experiment and the Gene via Predicts. Each regulatory element is the result of the data of one experiment in regards to a single gene. Thus, this aggregation of gene and experiment serves as a unique identifier. Since one experiment-gene aggregate can supply information about multiple regulatory

Back to Databasics
Design Document2
elements, this is a one-to-many relationship. There is a participation constraint on the part of the Regulatory Element; each Regulatory Element is required to link to at least one member of the aggregation between Gene and Experiment. Because each Regulatory Element entry can only be uniquely identified by the aggregation, this is a weak entity set.

PSG1 describes a set of Regulatory Elements; therefore, this is a one-to-many ratio. The PSG1 has a participation constraint; each PSG1 must describe at least one Regulatory Element. PSG1 is also connected to the aggregation between Gene and Experiment, because we will need to access information stored in PSG1 through the aggregation. Similar to Regulatory Elements, PSG1 can only be uniquely identified with its connection to the aggregation. Thus, this is also a weak entity set.

FACs are connected to the PSG1 entity by linking out to the PSG1. FAC has a participation constraint; each FAC must link out to a minimum of one PSG1 entity. This is a many-to-many relationship, as multiple FACs could potentially link to the same PSG1, and each FAC could link to multiple PSG1 entities.

Primary Keys

Experiment: Experiment
Gene (weak)
  Discriminators: Name
Regulatory Element (weak)
  Discriminators: Beginning, Sense, Length, Model
PSG1 (weak)
  Discriminators: FID, MAC (weak)
FAC: FAC-Number

Other

As mentioned above, both PSG1 and Regulatory Element are weak entity sets. These are each identified by the relationship to the aggregation of Gene and Experiment.

We had the option to create the Experiment attributes of Comparison and Species as their own entities; however, we decided against this course of action. Our rationale is that we have only a single part of each. For species, all we have is the name. For comparison, all we have is a string describing the comparison. As such, to place these in separate tables would be inefficient and bloated.

Constraints

- All primary keys and discriminators must be non-null.
- A regulatory sequence string must match the substring of the promoter sequence for the same gene that starts at the position indicated, and goes in the direction specified by the sense attribute.
- Length of the regulatory element must be less than the number of possible base pairs from its Beginning location.

Back to Databasics
Design Document3
Change Log
The following changes were made from the original version of this document:
January 27th, 2012
- *Species* attribute moved from *Gene* to *Experiment*, this results in *Gene* becoming a weak entity set.
- *Regulation* attribute moved from *Gene* to *Performed On*, the relationship between *Gene* and *Experiment*.
- Made *Regulatory Elements* and *PSG1* each weak entity sets, identified by the aggregation.
- Added a relation between *PSG1* and the aggregation, because we will need to access PSG1 information directly from the database.

E-R Model

Back to Databasics
Design Document4
SQL Create Table Statements

```sql
CREATE TABLE Experiment(
    Species VARCHAR2(30),
    ExID VARCHAR2(30),
    Comparison VARCHAR2(100),
    TessJob VARCHAR2(50),
    StoreTime VARCHAR2(50),
    Email VARCHAR2(50),
    TRANSFACStrings INT, --possibly boolean
    MySite INT, --possibly boolean
    Selected1 INT, --possibly boolean
    TRANSFACMatrix INT, --possibly boolean
    IMD INT, --possibly boolean
    CBIL INT, --possibly boolean
    JASPAR INT, --possibly boolean
    MyWeight INT, --possibly boolean
    Selected2 INT, --possibly boolean
    Combine VARCHAR2(100), --unsure
    Attr VARCHAR2(100), --unsure
    Matches VARCHAR2(100), --unsure
    UseCore INT, --possibly boolean
    Mismatch FLOAT,
    TSA FLOAT,
    TW FLOAT,
    TA FLOAT,
    GroupSelect1 VARCHAR2(100), --unsure
    TD FLOAT,
    TC FLOAT,
    TM FLOAT,
    Deficit FLOAT,
    Threshold FLOAT,
    Selected3 INT, --possibly boolean
    GroupSelect2 VARCHAR2(100), --unsure
    PsuedoCounts INT, --unsure
    ATContent FLOAT,
    Distribution VARCHAR2(100), --unsure
    HandleAmbig VARCHAR2(100), --unsure
    PRIMARY KEY (Species, ExID)
);

CREATE TABLE Gene(
    GeneName VARCHAR2(100),
    GeneAbbre VARCHAR2(20),
    Chromosome INT,
    BeginSite FLOAT,
    EndSite FLOAT,
    Promoter VARCHAR2(2050),
    Species VARCHAR2(30),
    ExID VARCHAR2(30),
    Regulation VARCHAR2(10),
    PRIMARY KEY (GeneAbbre, Species, ExID),
    FOREIGN KEY(Species, ExID) REFERENCES Experiment,
    UNIQUE (GeneName, Species, ExID)
);

-- there will not be any rows in the Summary table
CREATE TABLE Summary(
    ModID VARCHAR2(30), -- single values
    Fld VARCHAR2(30), -- comma separated list, not sure we can even have this here
    Mac VARCHAR2(30),
    La FLOAT,
    PRIMARY KEY (ModID)
);

```

*Back to Databasics*

*Design Document 7*
Rate FLOAT,
N INT,
Species VARCHAR2(30),
ExID VARCHAR2(30),
GeneAbbre VARCHAR2(20),
FOREIGN KEY(Species, ExID, GeneAbbre) REFERENCES Gene,
PRIMARY KEY (ModID, Mac, Species, ExID, GeneAbbre)
);

--there will not be any rows in the SummaryFAC table for any models starting with I
CREATE TABLE SummaryFAC(
   Fac VARCHAR2(100),
   ModID VARCHAR2(30),
   Mac VARCHAR2(30),
   FOREIGN KEY(ModID, Mac) REFERENCES Summary,
   PRIMARY KEY (Fac, ModID, Mac)
);

CREATE TABLE RegElms(
   Seq VARCHAR2(30),
   Fid VARCHAR2(30),
   ModID VARCHAR2(30), --will be null for R values, this is the factor in parenthesis after the model
   Beg INT,
   Leng INT,
   Sns VARCHAR2(30),
   Mac VARCHAR2(30),
   La FLOAT,
   Las FLOAT,
   Lq FLOAT,
   Ld FLOAT,
   Lpv FLOAT,
   Sc FLOAT,
   Sm FLOAT,
   Spv FLOAT,
   Ppv FLOAT,
   Species VARCHAR2(30),
   ExID VARCHAR2(30),
   GeneAbbre VARCHAR2(20),
   FOREIGN KEY(Species, ExID, GeneAbbre) REFERENCES Gene,
   PRIMARY KEY (Beg, Leng, Sns, Mac, Species, ExID, GeneAbbre)
);

--there will not be any rows in the RegElmsFac for any models starting with I
CREATE TABLE RegElmsTNums(
   TNum VARCHAR2(100),
   Fid VARCHAR2(30) REFERENCES Hits (Fid),
   Mac VARCHAR2(30) REFERENCES Hits (Mac),
   PRIMARY KEY (TNum, Fid, Mac)
);

**SQL DB-Cleanup Statements**

-- Back to Databasics CSC366 Cleanup
DROP TABLE RegElmsTNums;
DROP TABLE RegElms;
DROP TABLE SummaryFAC;
DROP TABLE Summary;
DROP TABLE Gene;
DROP TABLE Experiment;

Back to Databasics
Design Document