BINF 8211/6211
Design and Implementation of Bioinformatics Databases
Lecture #5

Dr. D. Andrew Carr
Dept. Bioinformatics and Genomics UNCC
Spring 2016

Keys

- Keys:
  - one or more attributes that define another attribute
  - Attribute part of a key is called the **key attribute**
  - **Composite keys** have more than 1 attribute
  - **Super key**
    - Key that uniquely identifies each row
  - **Candidate key**
    - Minimal super key
  - **PRIMARY KEY and Foreign key**
    - **Very important**

Primary and Foreign Keys

- Primary keys
  - Attributes
  - Super keys
  - **Unique and not null**
  - Should (Must) provide **entity integrity**.
- Foreign Keys
  - Reference primary keys in associate tables.
  - **Referential integrity** is predicated on the foreign key being present as a primary key in another table or being null.

Eight fundamental operators of relational algebra pt 1.

- SELECT
  - Restrict
  - Horizontal subset of the data
- PROJECT
  - All values in a given attribute
  - Vertical subset of the data
- INTERSECT
  - From set theory, overlap between tables
- DIFFERENCE
  - The non-overlapping entries between tables
- UNION
  - Combination of tables with shared attribute spaces

Eight fundamental operators of relational algebra pt 2.

- PRODUCT
  - Cartesian product from mathematics
    - All combine with all
- DIVIDE
  - Complicated and requires 2d to 1d relationship
  - Returns the intersection between specific attribute, entity pairs
- JOIN
  - This is the fundamental tool of the RDBMS system
  - Natural Join
    - Rows and columns that have common attributes
    - PRODUCT → SELECT → PROJECT
  - LEFT OUTER, RIGHT OUTER
    - Keeps all the values from 1 table and merges with the other table where it can

Data Dictionary

- AKA: System Catalog
- By definition: “ detailed accounting of all of the tables.”
- For each table:
  - Attribute names
  - Types
  - Constraints
  - Key type
  - Range

ER modeling topics

- Documenting and graphing relationships using IE and Chen conventions
- Specifying and graphically communicating constraints
  - The Chen convention allows more complete conceptual modeling
  - The IE convention translates more directly to implementation

ER Modeling tools

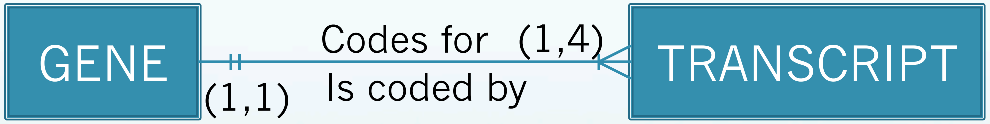
- For this class
  - MySQL Workbench
  - http://dev.mysql.com/downloads/workbench/5.1.html
    - Open source
    - Free
    - Easy to use
- Other acceptable tools
  - Microsoft Visio
  - Oracle

Modeling Behaviors

- Entity information to communicate
  - The attributes describing the entity
  - The relationships between entities
  - Whether one entity is existence – dependent on another
- Relationship information to communicate
  - The Cardinality (is it 1:1, 1:m, M:N)
  - Attributes may belong to the relationship rather than to either entity
  - Relationships may be weak or strong
- Attribute Behavior to watch out for
  - Attributes combining two types of information that you might want to use separately are composite attributes
  - Required attributes must have a value, optional attributes can be left empty
  - One or a combination of attribute values must uniquely identify each attribute instance (so these cannot be optional).

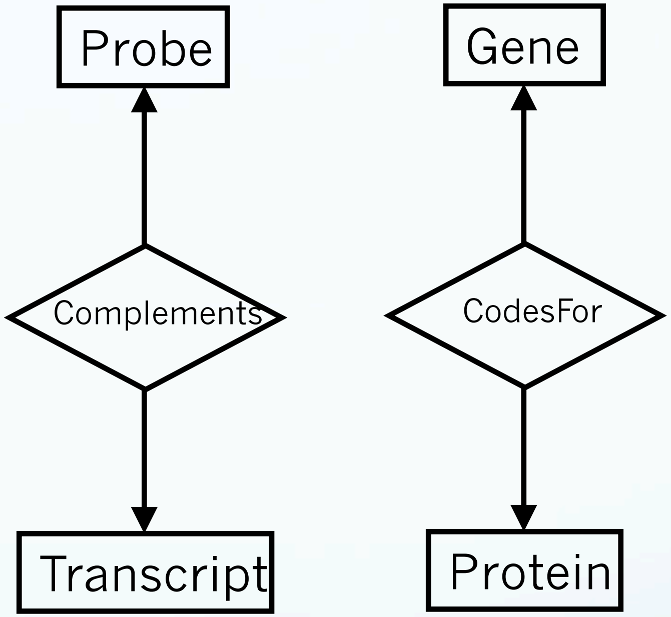
Relationships: Connectivity and Cardinality

- Relationships must be expressed as one of three types.
  - Connectivity is what that type is (1:1, 1:M, N:M)
  - Cardinality is more specific – is 1:M really 1:3? Or 0:3?
    - This is described using a range, e.g. (0,3)
      - Some dictionaries are not inclusive of one or both ends (greater than 0 but not 0, for example).
      - Graphically the range is place adjacent to the connector.



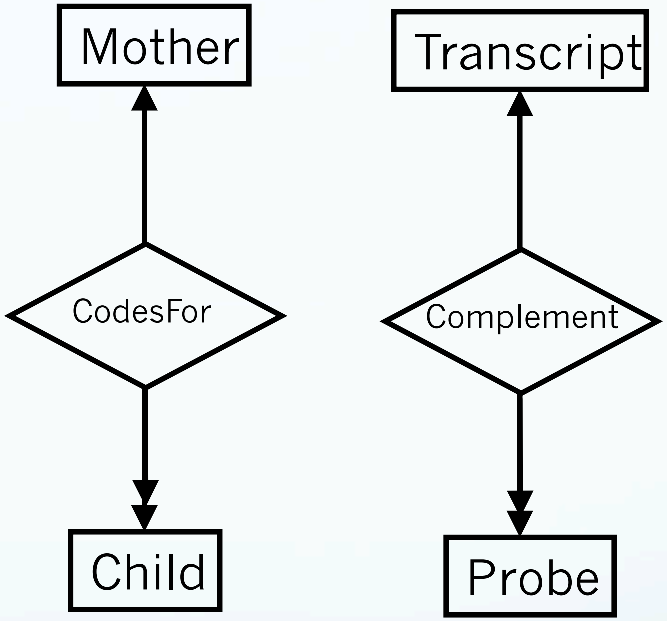
One-to-one Relationships

- Formally: if there can only be zero or one *instance* of entity A for zero or one *instance* of entity B then there is a one-to-one relationship
  - A predicted gene coding sequence makes zero or one protein in prokaryotes
  - Probe complementarity to zero or one transcripts and a given transcript is complementary to zero or one probe



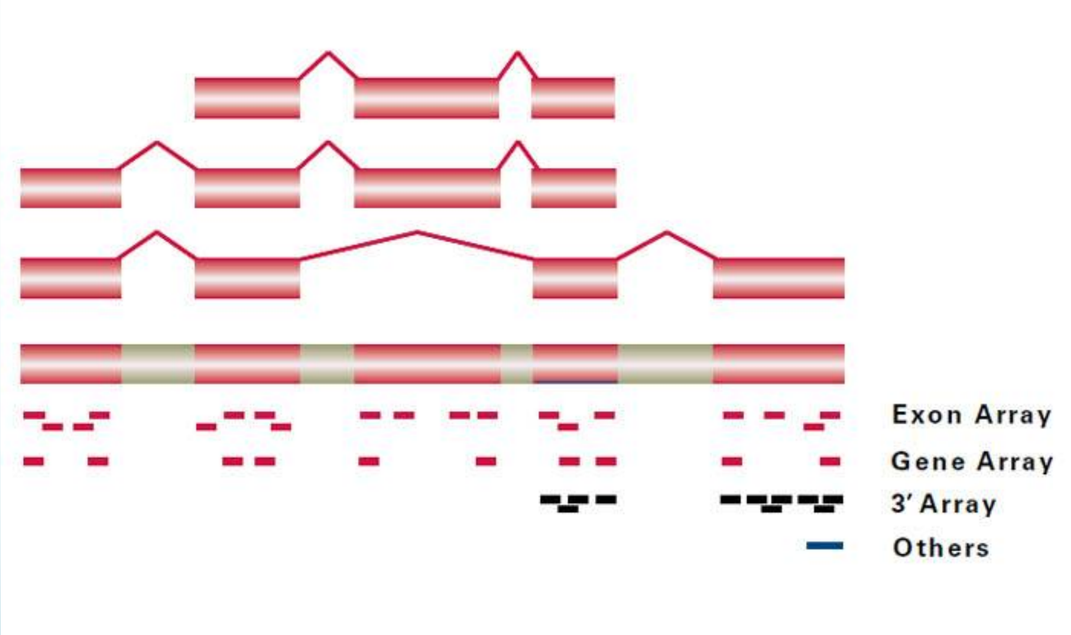
One-to-Many Relationships

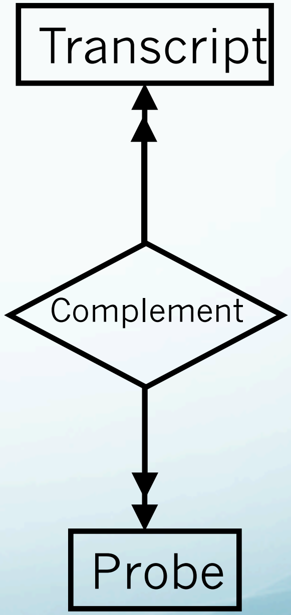
- Formally: if there are two entities, A and B and if A_i is related to zero, one, or more instances of entity B and B_i is related to zero or one instance of entity A then this is a one-to-many relationship.
- A [human]mother may have zero, one or many birth children, but a child may have only one biological mother



Many-to-Many relationships

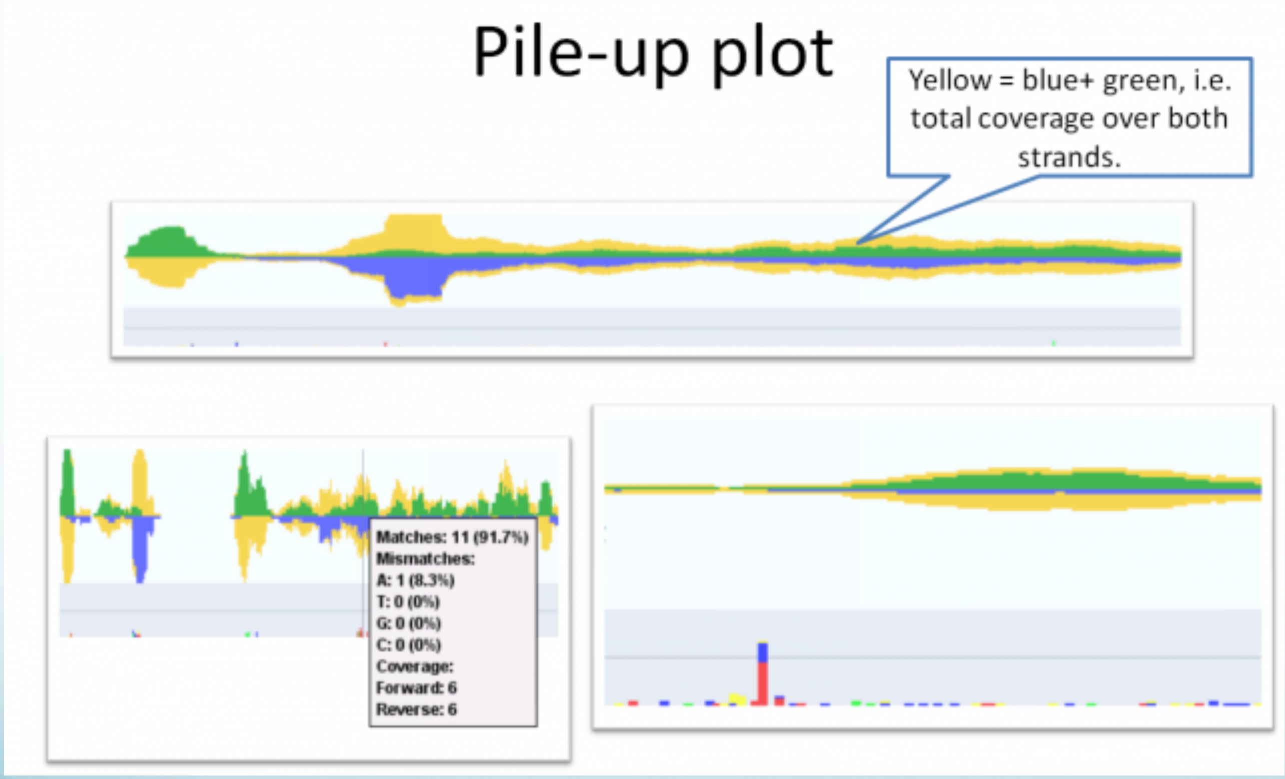
- Formally, this relationship exists if, for two entities A and B, for an instance A_i there can be zero, one, or many instances of B_i and, for instance B_i, there can be zero, one, or many instances of A_i.





Many-to-Many relationships

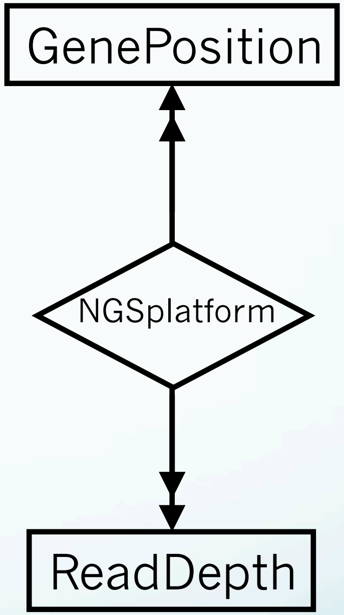
- NGS example



Pile-up plot

Yellow = blue+ green, i.e. total coverage over both strands.

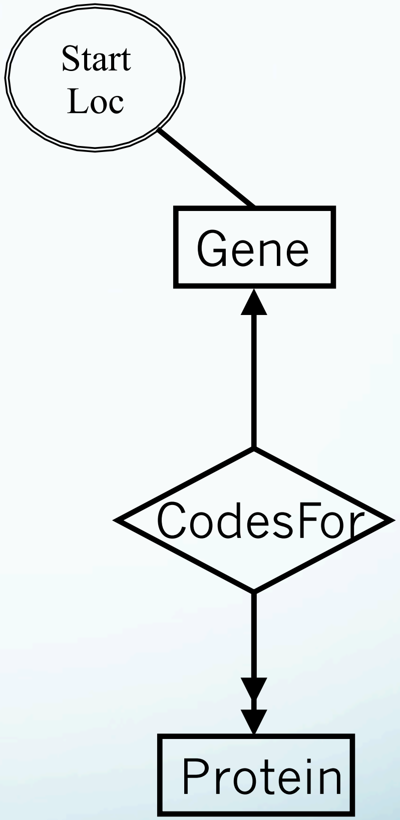
Matches: 11 (91.7%)
Mismatches:
A: 1 (8.3%)
T: 0 (0%)
G: 0 (0%)
C: 0 (0%)
Coverage:
Forward: 6
Reverse: 6



Documenting ER Models: Chen Style

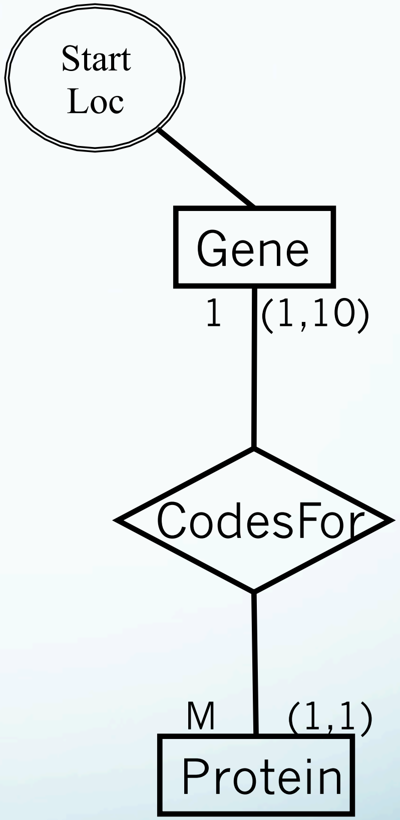
- The ER conceptual model paradigm is by Chen, there are several variants of the graphical representation style.

- Chen style
  - Entities are in rectangles
  - Attributes are in ovals connected to the entities
  - Relationships are in diamonds
  - Uses arrows to show the cardinality of the relationship
    - The single arrow pointing to Gene means a Protein belongs to (maps to) at most one Gene(zero or one)
    - The double arrow pointing to the Protein means that a Gene can code for more than one Protein (zero, one or many – presumably this is in Eukaryotes)



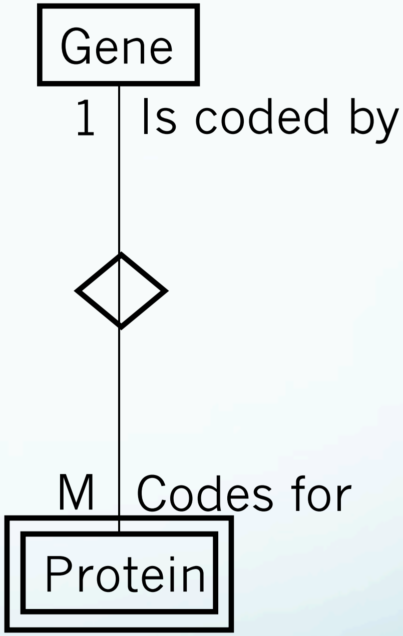
Chen Alternative Styles - 1

- You may replace the arrows with numbers and letters that express the *cardinality* of the relationship.
  - '1' indicates a protein comes from one gene
  - 'M' indicates 'many'
  - You can set more specific constraints, such as
    - there must be a gene before a protein can exist (1,1)
    - a gene might be restricted to no more than 10 protein isomers and there must be at least one isomer (1,10).



Chen Alternative Styles -2

- To make the relationship sensible reading in both directions
  - remove the relationship name from the diamond (some versions remove the diamond also)
  - add the relationship description and its inverse on the correct entity edges
- Indicate a *weak* entity (existence-dependent) by putting a double box around it
  - But note that if there are many ‘parents’ (entities with similar 1:M relationships) you won’t know which is meant to impose the mandatory relationship.

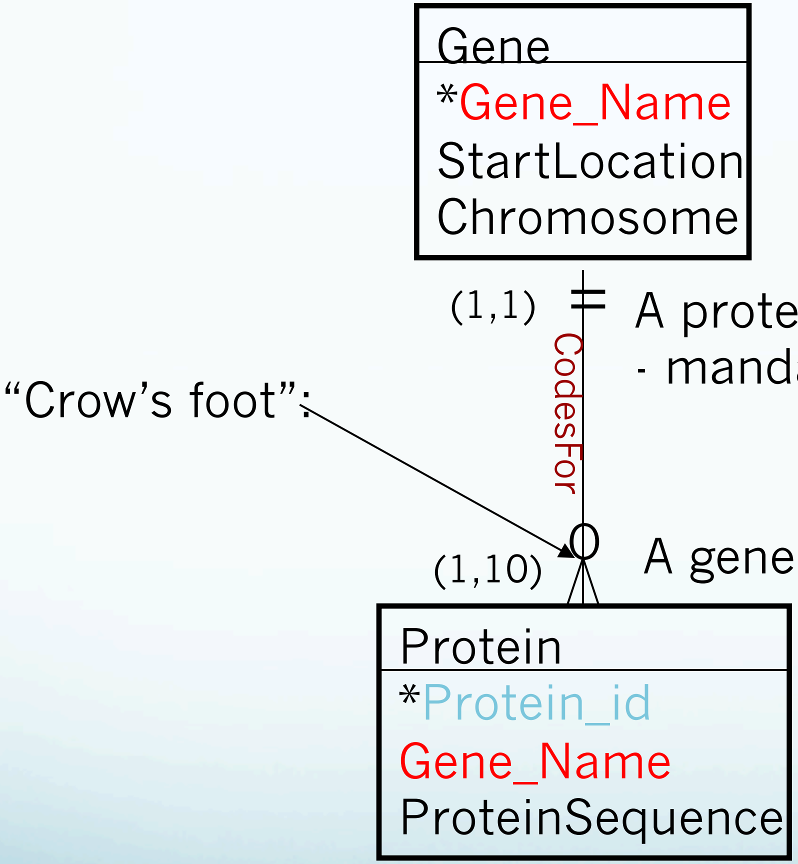


The Information Engineering Style

- Entities are represented by rectangles with a list of attributes – the entity name is the first attribute and is set off in some way.
- Connectors carry more information
  - || means one and only one (mandatory relationship)
  - 0| means zero or one
  - >| means one or more (mandatory relationship)
  - >0 means zero, one or more

Note: The symbols are written 90° to the connector line.

IE Relationship Diagram



Gene

*Gene_Name

StartLocation

Chromosome

(1,1)

CodesFor

“Crow’s foot”:

(1,10)

Protein

*Protein_id

Gene_Name

ProteinSequence

A protein belongs to one and only one gene - mandatory

A gene may have zero, one or more proteins

ER Modeling tools

- For this class
  - MySQL Workbench
  - http://dev.mysql.com/downloads/workbench/5.1.html
    - Open source
    - Free
    - Easy to use
- Other acceptable tools
  - Microsoft Visio
  - Oracle