BINF 8211/6211
Design and Implementation of Bioinformatics Databases
Lecture #3

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

Class Information

- Class website
- http://ponden.github.io/databases2016
- User: student :: $tUd3Nt
- Homework 1
  - Due Thursday

Database

- A database is an electronic system for managing data and the meta-data describing that data: the system is self-describing.
- The management software handles the complex structures, the allocation of files and bits, identification and retrieval tasks.
- The data dictionary holds the meta-data
  - What type of entity is described
  - What are the attributes of each entity?
  - What are the relationships between the entities?
  - What are the constraints that need to be applied to attribute values and relationships?

Data Model Levels

- External
  - End user view
  - Basic representation
- Conceptual.
  - Linkage to schema
  - Greater detail
- Internal model
  - Code, Script, Implementation
- Physical Model
  - How the data is actually stored
    - Binary

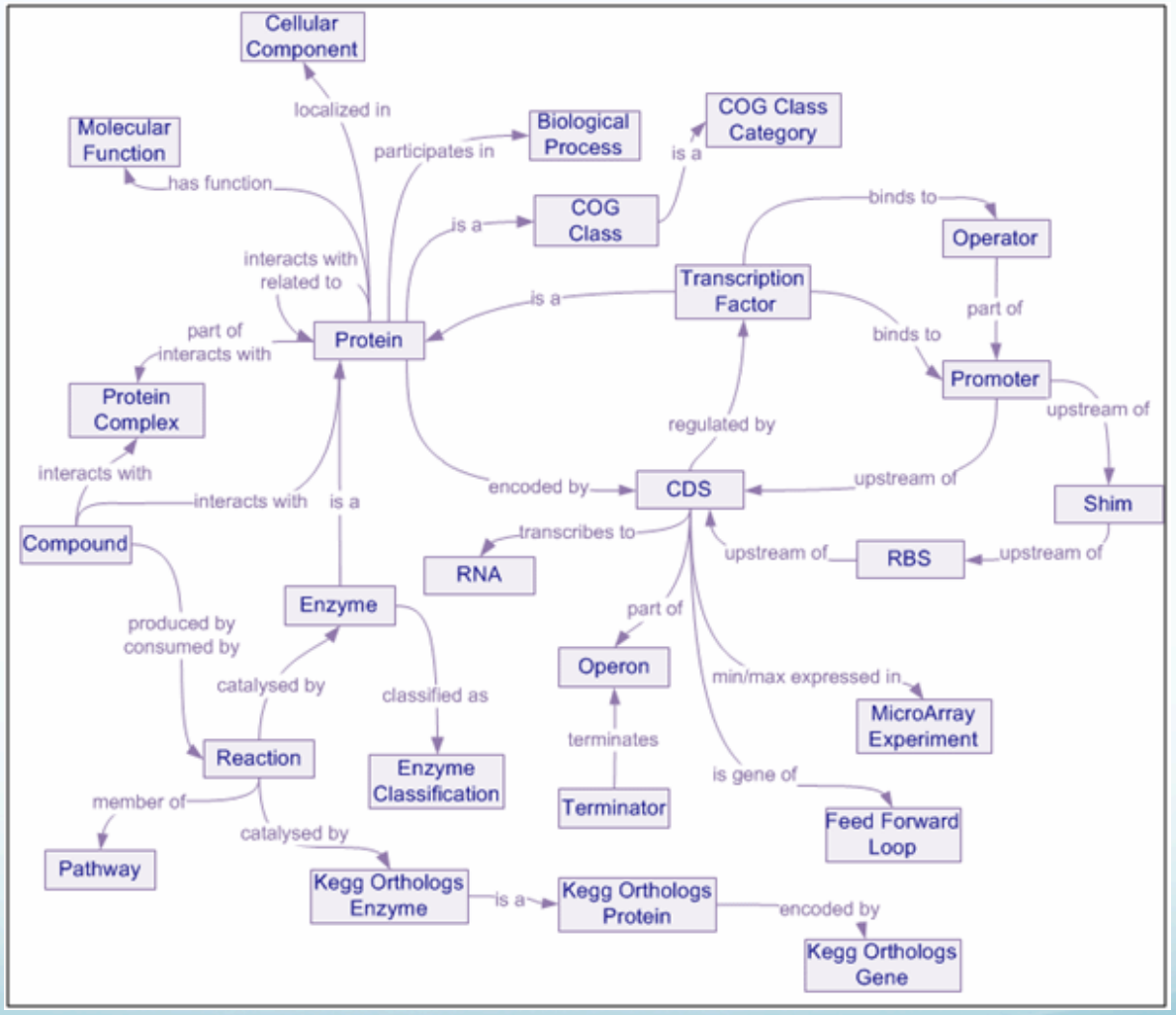
Data models

- Models are representations that capture complex systems so as to create a blueprint (abstraction) for determining rules and patterns.
- Data elements, their properties, and the relationships between them are defined.
  - Relationships can be PART-OF or HAS-A.
    - An operon HAS multiple genes whose expression is required to be balanced.
    - Biological relationships are often due to processes whose steps we want to study: a spliceosome is part of a process that creates splice forms from precursor mRNA.

Producing a data model

- A data model is an abstraction (graphical or lexical) that specifies data properties, structures (organizes) and relates elements of information that belong to a single collection.
  - Symbols (graphical) and controlled terms (lexical) are used to explain how the information is modeled.
    - A formal modeling framework is used to make sure your model is clear and does not violate required conditions.
  - It explicitly determines the structure of data elements, including data types, formats and relationships and metadata (like units).
    - In a programming language you would refer to a data structure, however a lot of biological data (images, sound, descriptions of experiments and journal articles and embedded graphs) would be considered unstructured in that context.

Conceptual Model High-Level Representation



http://intbio.ncl.ac.uk/?projects=bacillondex

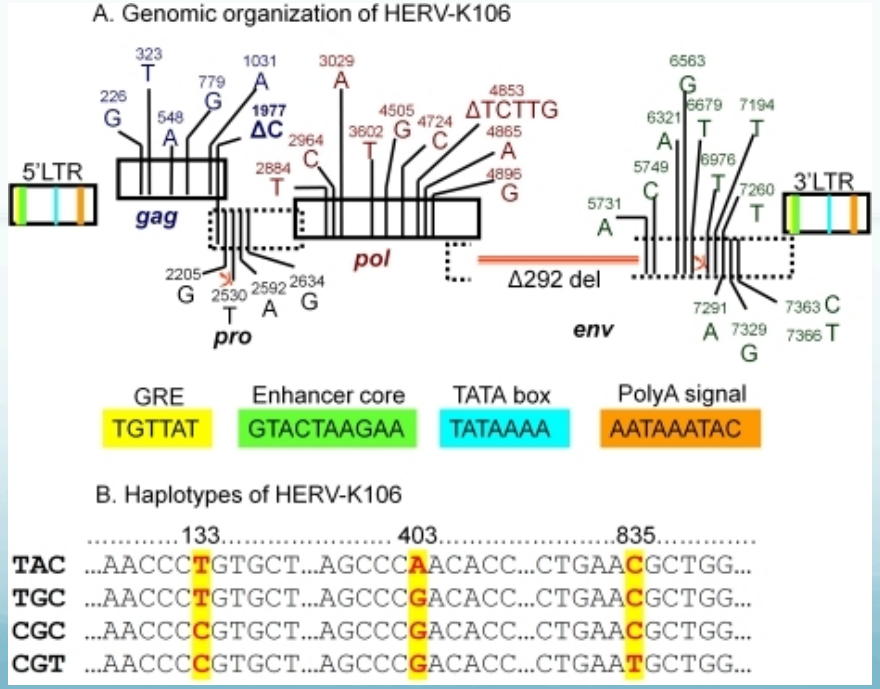
ER -Model

- Developed in 1976
  - Most common model (standard)
  - Used to describe the data and relationships.
    - Abstract manner used help draft and design databases
  - Peter Chen – published first graphical method

- More common Crows Foot model

Entity = Noun

* Is a DNA sequence a thing, or a description (attribute) of a thing?
* Does the sequence have properties that we want to describe that are not necessarily of interest in describing a gene?
* Do you want to filter the sequence independently from filters that make sense for filtering a gene?



Attribute = Adjective

* Think of these as categories of descriptions, or column headers in a data table.
  * PeakID is a category, each specific peak has a value in a field of the column 'PeakID'.
  * Chromosome is a category, the value in each field per row is the chr# value.
  * Start and End are categories delimiting the peak in the context of the chromosome.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	PeakID	Chr	Start	End	Strand	Peak Sco	Focus Ra	Annotation	Detailed Anno	Distance to T	Nearest Pror	PromoterID	Nearest Unig	Nearest Refs	Nearest Ense	Gene Name	Gene Alias	Gene Descrip
2	chr18-1	chr18	69007968	69008268	+	593	0.939	intron (NR_03	intron (NR_03	74595	NR_034133	400655	Hs.579378	NR_034133		LOC400655	-	hypothetical
3	chr9-1	chr9	88209966	88210266	+	531.9	0.946	Intergenic	Intergenic	-50894	NM_001185	79670	Hs.597057	NM_001185	ENSG000000	ZCCHC6	DKFZp666B1	zinc finger, C
4	chr14-1	chr14	62337073	62337373	+	505.4	0.918	intron (NM_17	intron (NM_17	244485	NM_172375	27133	Hs.27043	NM_139318	ENSG000001	KCNH5	EAG2|H-EAG	potassium vo
5	chr17-1	chr17	5076243	5076543	+	492.1	0.936	intron (NR_03	intron (NR_03	2414	NM_207103	388325	Hs.462080	NM_207103	ENSG000001	C17orf87	FLJ32580|M	chromosome
6	chr17-2	chr17	47851714	47852014	+	476.2	0.824	Intergenic	Intergenic	-259488	NM_001082	56934	Hs.463466	NM_001082	ENSG000001	CA10	CA-RPX|CAR	carbonic anh
7	chr10-1	chr10	98420680	98420980	+	474.9	0.967	intron (NM_15	intron (NM_15	49439	NM_152309	118788	Hs.310456	NM_152309	ENSG000001	PIK3AP1	BCAP|RP11-3	phosphoinos
8	chr9-2	chr9	81294389	81294689	+	456.3	0.957	Intergenic	Intergenic	-82159	NM_007005	7091	Hs.444213	NM_007005	ENSG000001	TLE4	BCE-1|BCE1|	transducin-li
9	chr14-2	chr14	36817736	36818036	+	452.3	0.757	intron (NM_13	intron (NM_13	81017	NM_001195	145282	Hs.660396	NM_001195	ENSG000001	MIPOL1	DKFZp313M	mirror-image
10	chr18-2	chr18	20049825	20050125	+	449.7	0.853	intron (NM_08	intron (NM_08	56219	NM_018030	114876	Hs.370725	NM_018030	ENSG000001	OSBPL1A	FLJ10217|OF	oxysterol bin
11	chr7-1	chr7	12226829	12227129	+	445.7	0.901	intron (NM_01	intron (NM_01	9606	NM_001134	54664	Hs.396358	NM_001134	ENSG000001	TMEM106B	FLJ11273|M	transmembr
12	chr14-3	chr14	88712188	88712488	+	443.1	0.844	intron (NM_00	intron (NM_00	240869	NM_005197	1112	Hs.621371	NM_001085	ENSG000000	FOXN3	C14orf116|C	forkhead box
13	chr18-3	chr18	62951924	62952224	+	443.1	0.947	Intergenic	Intergenic	-382689	NR_033921	643542	Hs.652901	NR_033921		LOC643542	-	hypothetical
14	chr3-1	chr3	32196769	32197069	+	443.1	0.87	Intergenic	Intergenic	-58256	NM_178868	152189	Hs.154986	NM_178868	ENSG000001	CMTM8	CKLFSF8|CKL	CKLF-like MA
15	chr11-1	chr11	110685448	110685748	+	425.8	0.907	Intergenic	Intergenic	-9849	NR_034154	399948	Hs.729225	NR_034154		C11orf92	DKFZp781P1	chromosome
16	chr4-1	chr4	81755366	81755666	+	423.2	0.908	intron (NM_15	intron (NM_15	279618	NM_152770	255119	Hs.527104	NM_152770	ENSG000001	C4orf22	MGC35043	chromosome

Relationship = Verb

- Between any two entities (a subject and an object) there is some type of relationship (even none is a type).
- When a relationship does exist we put it in one of three general categories:
  - 1:1, 1:M, M:N (many to many where M≠N)
- When does it make sense to use 1:1?
  - Would you annotate the DNA and protein of a gene with similar labels?
- 1 Gene : Many mRNA alternative transcripts
  - Each form of mRNA comes from only 1 gene
- One transcription factor can bind to many operator sites and each operator site can bind many proteins → M:N

Constraints reflect natural limits on properties and processes.

* Many of the attributes have easy to define limits – we can build rules and limit our use of storage if these are defined.
* What rules could you state constraining peaks appearing on chromosome 18?
  * Start and end cannot exceed the length
  * The values must be positive integers
  * Strand must have a value of either '+' or '-'
  * Descriptions could be limited to keywords in listed ontologies.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	PeakID	Chr	Start	End	Strand	Peak Sco	Focus Ra	Annotation	Detailed Anno	Distance to T	Nearest Pror	PromoterID	Nearest Unig	Nearest Refs	Nearest Ense	Gene Name	Gene Alias	Gene Descrip
2	chr18-1	chr18	69007968	69008268	+	593	0.939	intron (NR_03	intron (NR_03	74595	NR_034133	400655	Hs.579378	NR_034133		LOC400655	-	hypothetical
3	chr9-1	chr9	88209966	88210266	+	531.9	0.946	Intergenic	Intergenic	-50894	NM_001185(	79670	Hs.597057	NM_001185(	ENSG000000	ZCCHC6	DKFZp666B1	zinc finger, C
4	chr14-1	chr14	62337073	62337373	+	505.4	0.918	intron (NM_17	intron (NM_17	244485	NM_172375	27133	Hs.27043	NM_139318	ENSG000001	KCNH5	EAG2|H-EAG	potassium vo
5	chr17-1	chr17	5076243	5076543	+	492.1	0.936	intron (NR_03	intron (NR_03	2414	NM_207103	388325	Hs.462080	NM_207103	ENSG000001	C17orf87	FLJ32580|M(	chromosome
6	chr17-2	chr17	47851714	47852014	+	476.2	0.824	Intergenic	Intergenic	-259488	NM_001082!	56934	Hs.463466	NM_001082!	ENSG000001	CA10	CA-RPX|CAR	carbonic anh
7	chr10-1	chr10	98420680	98420980	+	474.9	0.967	intron (NM_15	intron (NM_15	49439	NM_152309	118788	Hs.310456	NM_152309	ENSG000001	PIK3AP1	BCAP|RP11-3	phosphoinos
8	chr9-2	chr9	81294389	81294689	+	456.3	0.957	Intergenic	Intergenic	-82159	NM_007005	7091	Hs.444213	NM_007005	ENSG000001	TLE4	BCE-1|BCE1|	transducin-lil
9	chr14-2	chr14	36817736	36818036	+	452.3	0.757	intron (NM_13	intron (NM_13	81017	NM_001195!	145282	Hs.660396	NM_001195!	ENSG000001	MIPOL1	DKFZp313M2	mirror-image
10	chr18-2	chr18	20049825	20050125	+	449.7	0.853	intron (NM_08	intron (NM_08	56219	NM_018030	114876	Hs.370725	NM_018030	ENSG000001	OSBPL1A	FLJ10217|OF	oxysterol bin
11	chr7-1	chr7	12226829	12227129	+	445.7	0.901	intron (NM_01	intron (NM_01	9606	NM_001134!	54664	Hs.396358	NM_001134!	ENSG000001	TMEM106B	FLJ11273|M(	transmembr
12	chr14-3	chr14	88712188	88712488	+	443.1	0.844	intron (NM_00	intron (NM_00	240869	NM_005197	1112	Hs.621371	NM_001085(	ENSG000000	FOXN3	C14orf116|C	forkhead box
13	chr18-3	chr18	62951924	62952224	+	443.1	0.947	Intergenic	Intergenic	-382689	NR_033921	643542	Hs.652901	NR_033921		LOC643542	-	hypothetical
14	chr3-1	chr3	32196769	32197069	+	443.1	0.87	Intergenic	Intergenic	-58256	NM_178868	152189	Hs.154986	NM_178868	ENSG000001	CMTM8	CKLFSF8|CKL	CKLF-like MA
15	chr11-1	chr11	110685448	110685748	+	425.8	0.907	Intergenic	Intergenic	-9849	NR_034154	399948	Hs.729225	NR_034154		C11orf92	DKFZp781P1	chromosome
16	chr4-1	chr4	81755366	81755666	+	423.2	0.908	intron (NM_15	intron (NM_15	279618	NM_152770	255119	Hs.527104	NM_152770	ENSG000001	C4orf22	MGC35043	chromosome

ER – Model cont'd

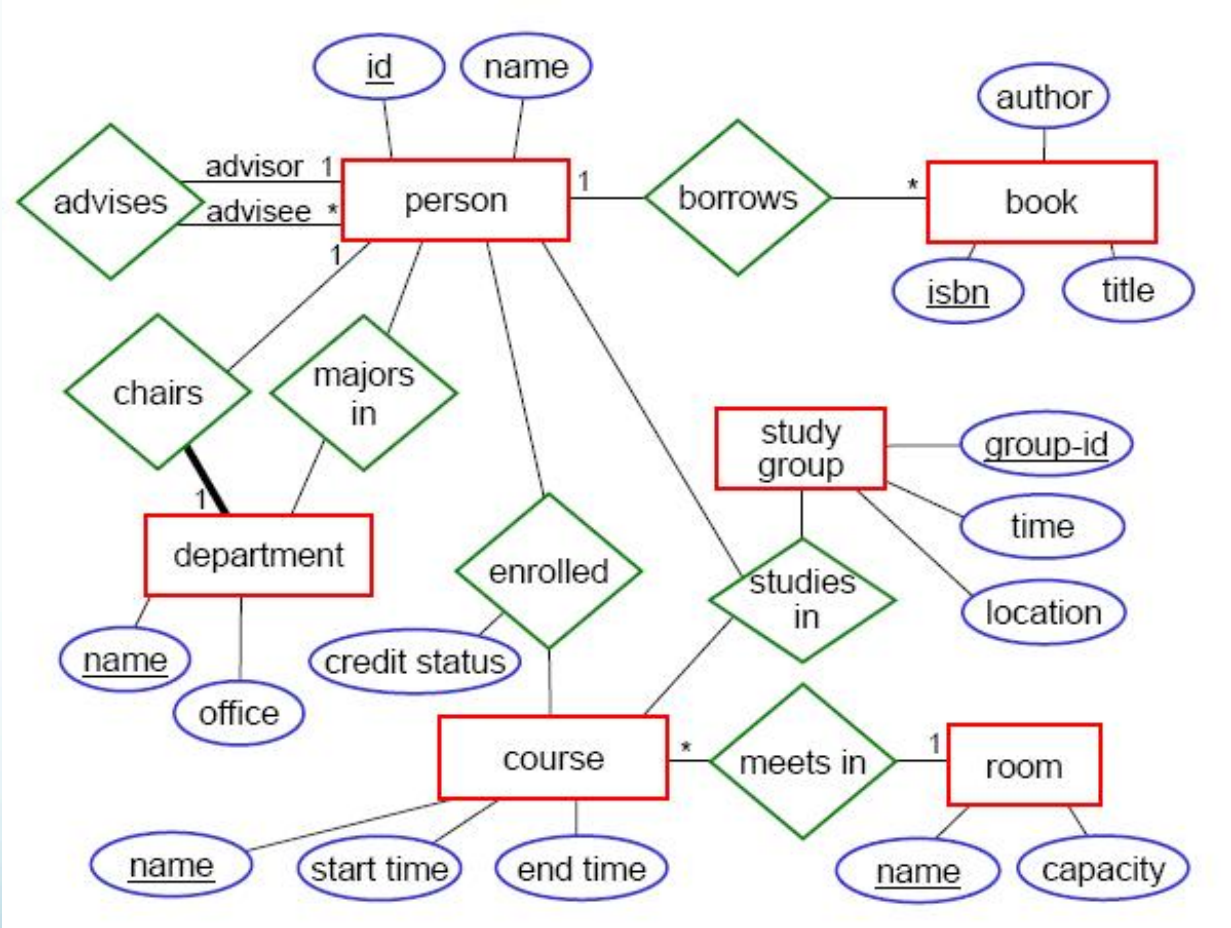


Image Taken from Web:http://www.snipview.com/q/Entity%E2%80%93relationship_model

Object Oriented Data Model

- An object is a abstraction of a real world object.
  - Object
    - Similar to an entity
    - Has attributes
    - Example: “Person” has attributes
      - Name
      - SSN
      - DOB
    - Methods describe OO behavior

OO Model continued

- Classes
  - Hierarchical collections of objects
    - Each object has 1 and only 1 parent
  - Inheritance:
    - Each object inherits the attributes and methods of the parent object

Setting up a database is time consuming – do a cost-benefit analysis first.

- Does the model contain many entities?
- Are there 1:M or M:N relationships between *any* of the entities?
- Do you have really large sets of data?
- Do you need security?
- Will several people need access to the data with an assurance that neither can change it without reference to the other?
- File systems usually link information by repeating various columns, and often entire sets of files – there is a lot of redundancy that must be managed. Updates to one file MUST be replicated in all other files but often are not
  - Data inconsistency
  - Data anomalies when insertions and deletions occur in one file but not others

ER – Model cont'd

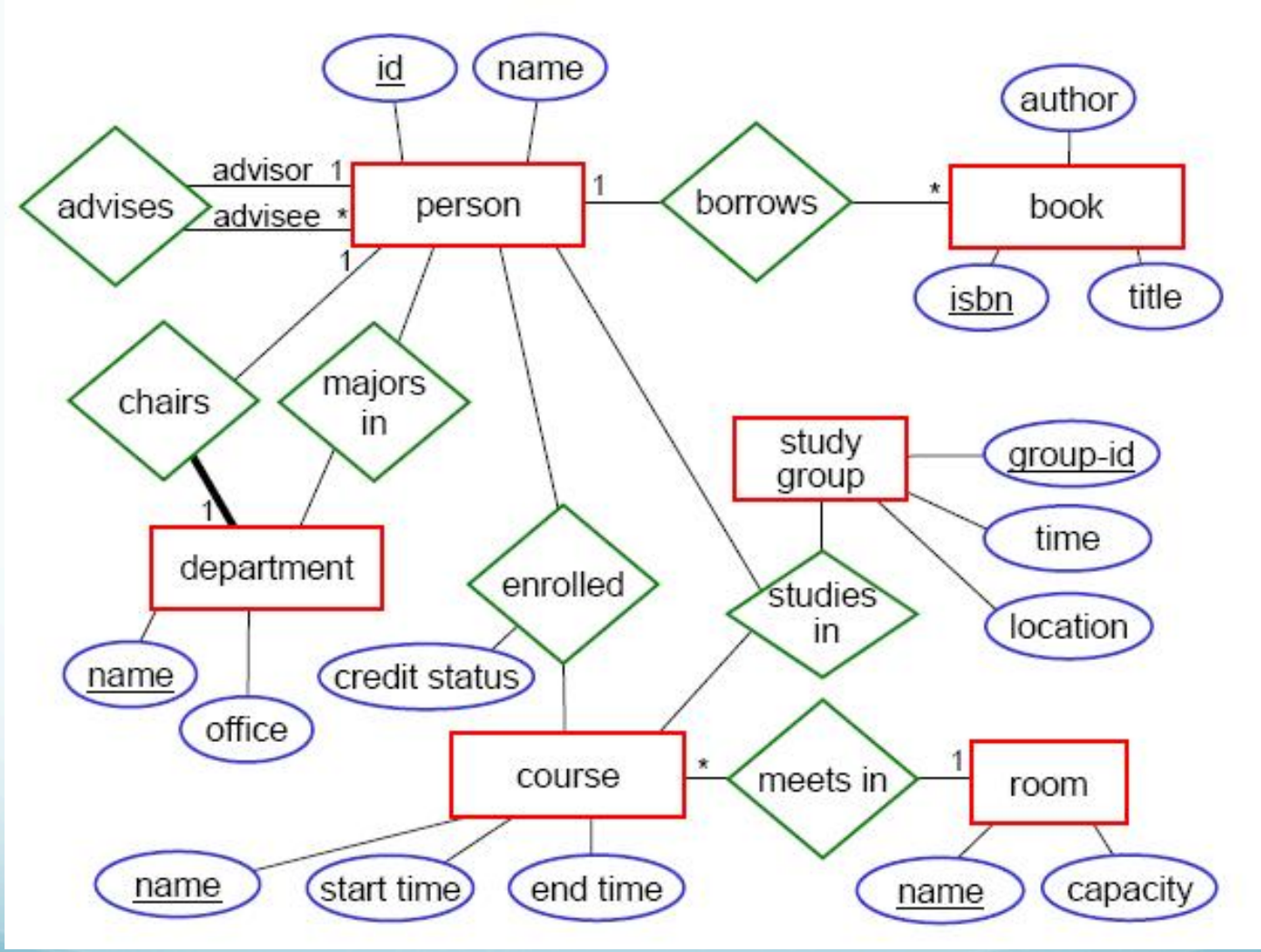


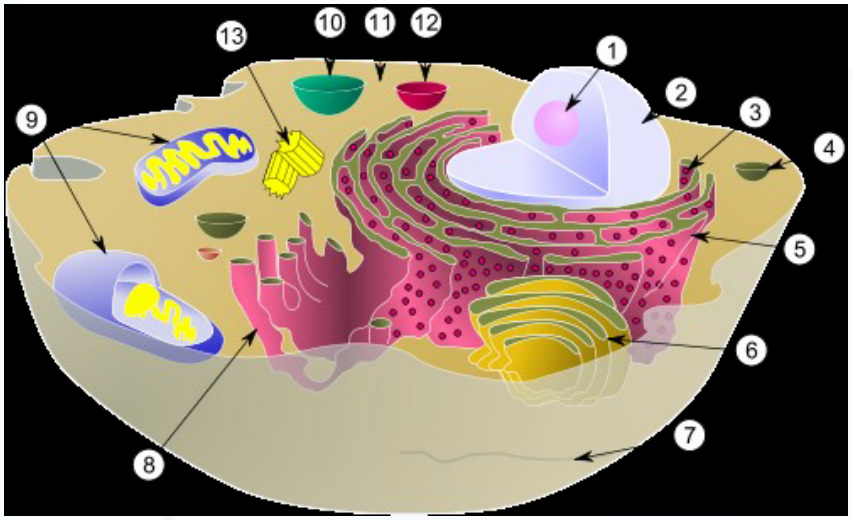
Image Taken from Web:http://www.snipview.com/q/Entity%E2%80%93relationship_model

Biological Example

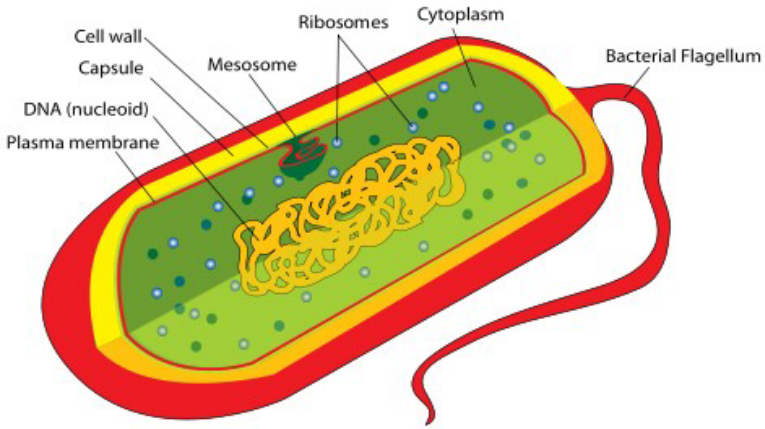
- To illustrate the concepts, vocabulary and approach we will examine am a use case in which we want to predict the presence of bacterial operons by determining whether genes that are close together also have expression levels that are very similar under nearly all conditions.

Example from J. Weller 2015

Conceptual biological representations (models) of cells.



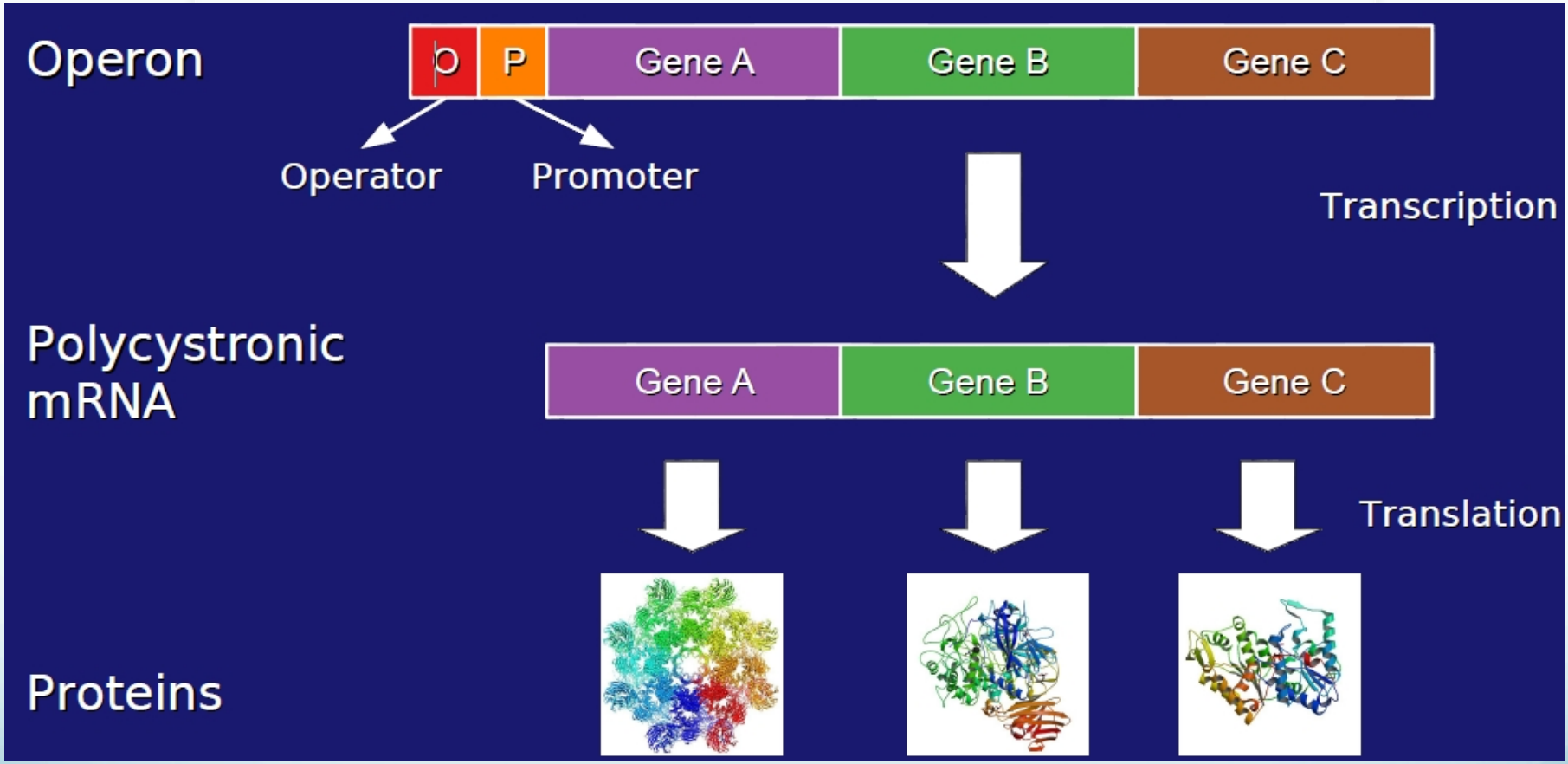
Type: Eukaryote
Diameter: 10-100um
HAS Nucleus
HAS Organelles
IS Unicellular OR multicellular



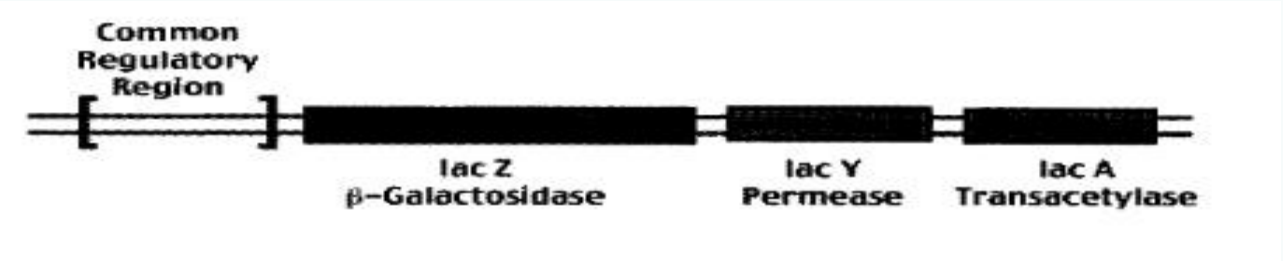
Type: Prokaryote
Diameter: 1-2um
HAS Nucleoid
HAS NO Organelles
IS Unicellular

Rules: A cell cannot be both a eukaryote and a prokaryote.
A cell must be one or the other.

Another model: mRNA in prokaryotes



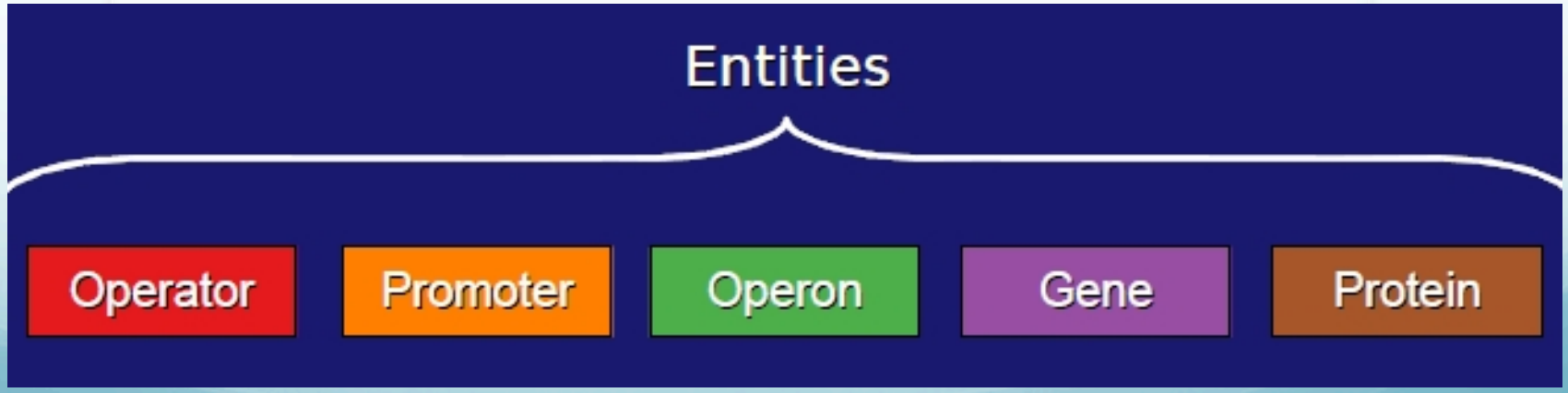
The 3 proteins are translated at the same time.



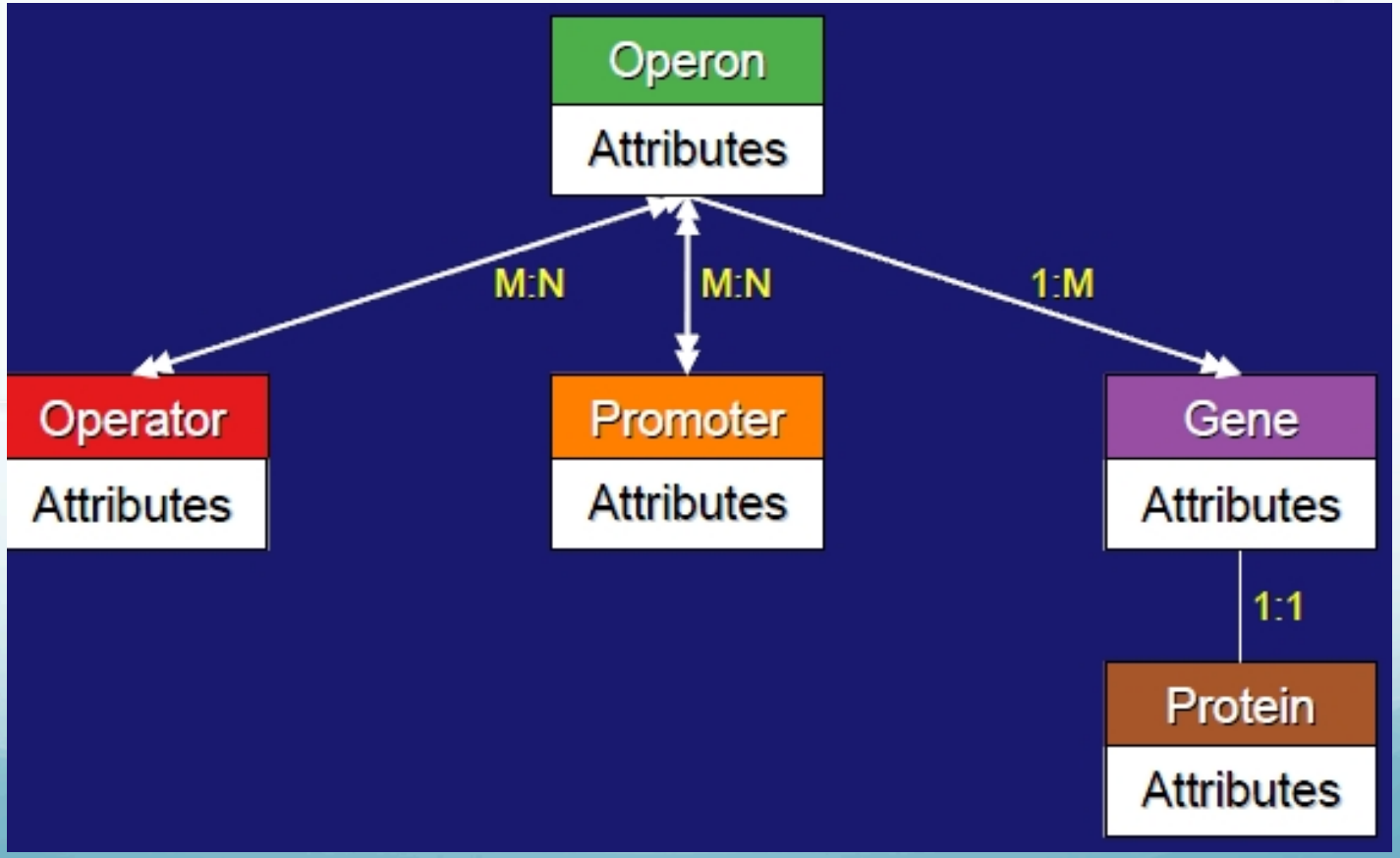
- Operon
  - HAS-A Regulatory region
    - HAS operator
    - HAS promotor
  - HAS gene (≥ 3?)
    - Has polycistronic transcript
    - HAS coding sequence
  - HAS ribosome binding site
  - HAS a terminator (≥ 1?)
  - Process: Transcription initiation Machinery IS-A complex of [sigma + polymerase]
  - Process: Operator binds activator/repressor
  - Process: Repressor Mechanism IS steric hindrance

Defined Relationships

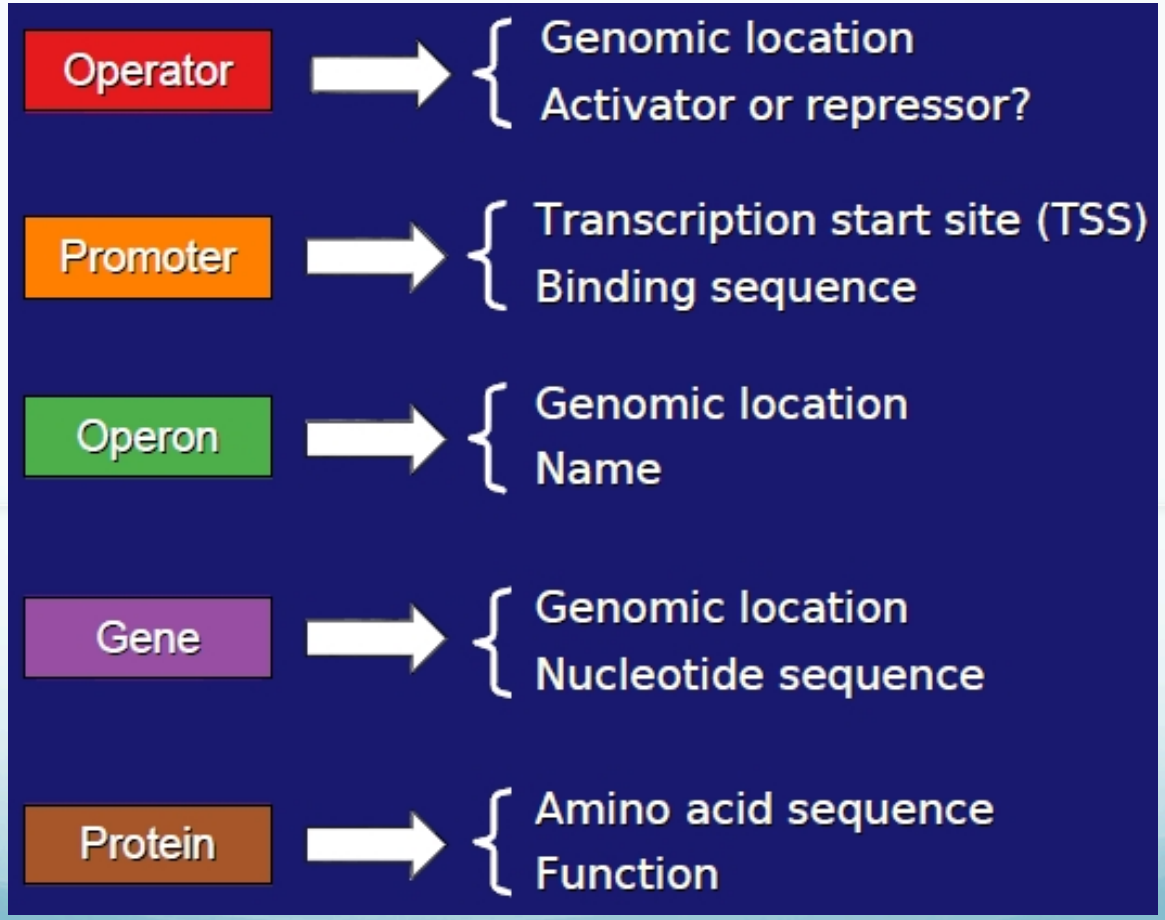
Using operon biological construct, what are the entities?



An Entity Relationship Diagram gives a visual representation of the elements as you have defined them – laying them out this way let's you spot problems and share the model with others.

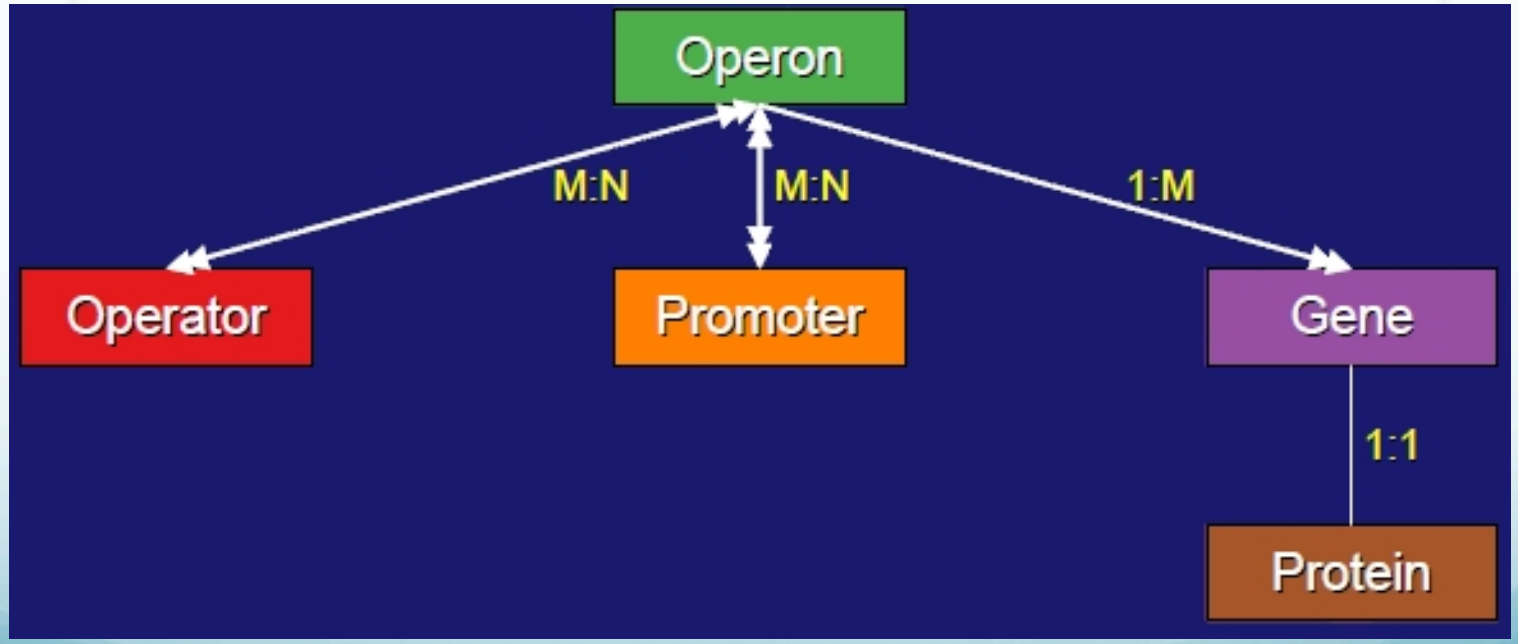


The terms used to describe the entity are called attributes (adjectives) each type of property is an attribute.

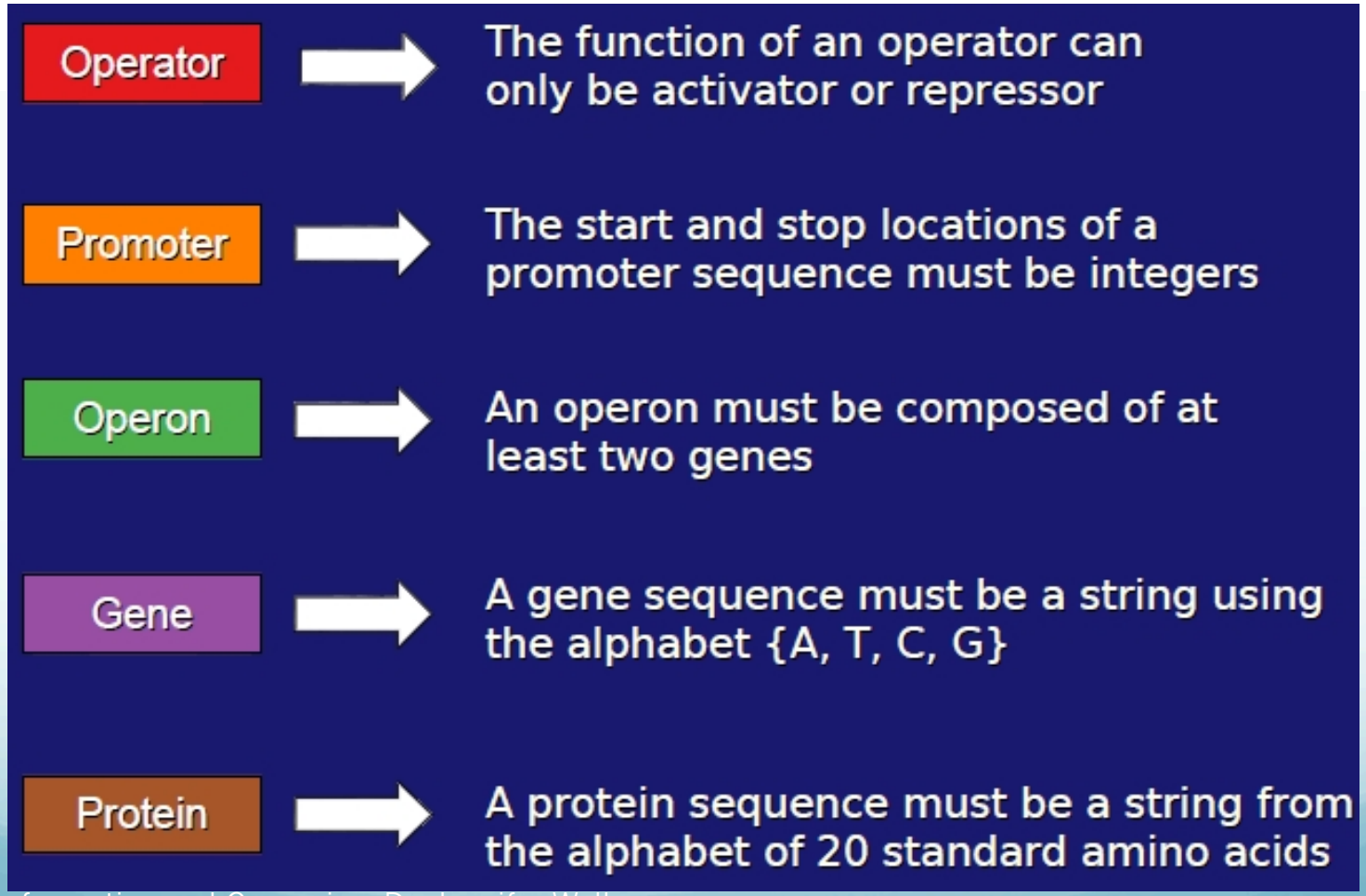


In complex systems like cells one part often acts upon other parts: the action is a relationship (verb).

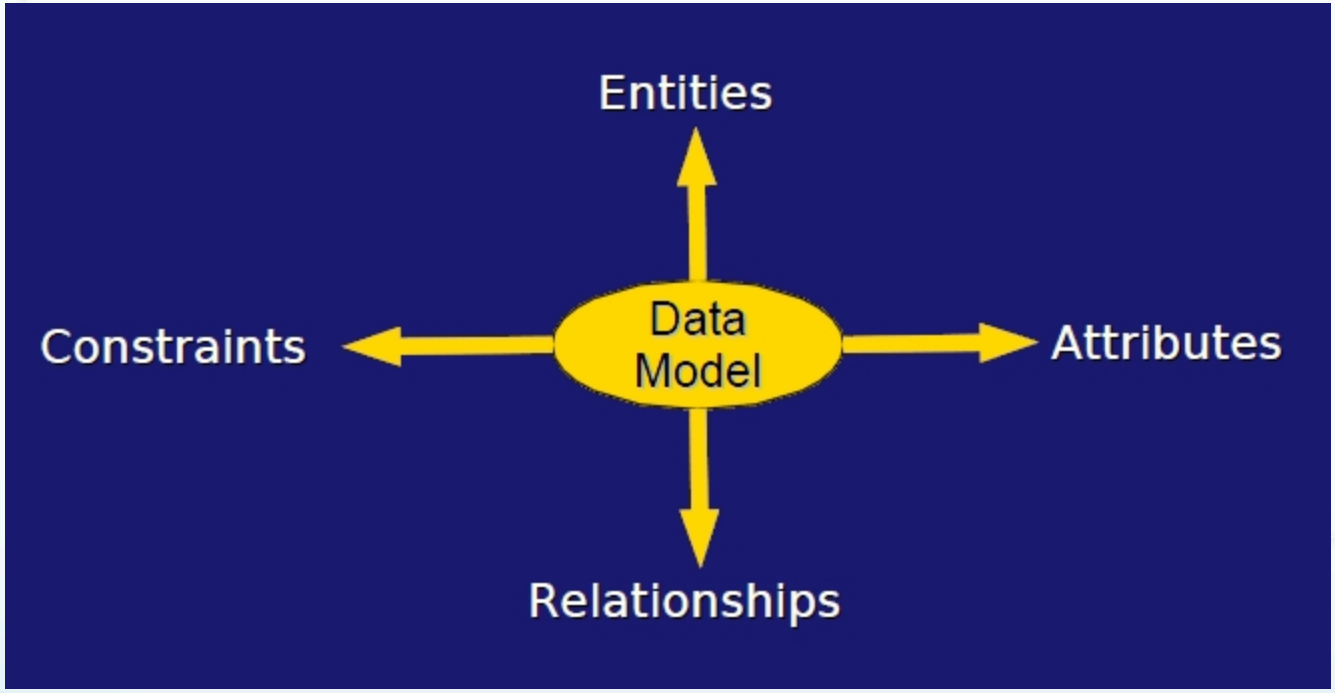
The actions have been divided into 3 types: one-to-one (1:1), one-to-many (1:M) and many-to-many (M:N).



Most attributes are not infinite or continuous – there is a range across which they apply. The range is called a *constraint*.

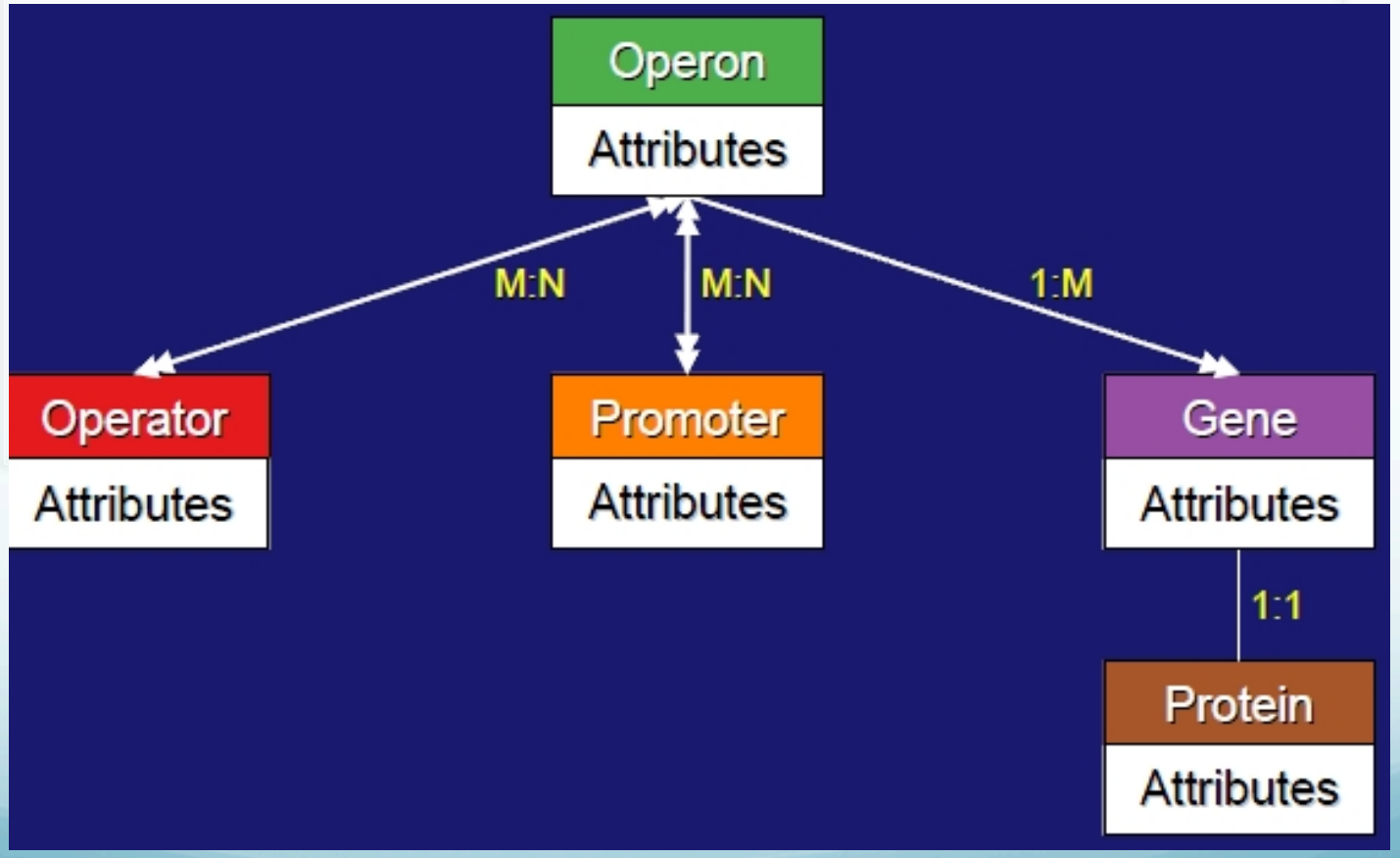


Getting started : looking at the operon ‘rules’ allows us to propose a data model for storing observations about those properties in specific cases.



*Business Rules

An ER Diagram – so far.



A spread sheet is a useful way to lay out the attributes for an entity. Notice that this spread sheet includes 3 of the entities we separated in the diagram. What effect does this have?

Operon	Promoter	TSS (Absolute)	TSS (Relative)	Gene	Gene Start Loc	Gene End Loc	Protein GI	Protein Description
moaABCDE	moaAp1	816050	-217	moaA	816267	817256	16128749	molybdopterin biosynthesis protein A
moaABCDE	moaAp1	816050	-217	moaB	817278	817790	16128750	molybdopterin biosynthesis protein B
moaABCDE	moaAp1	816050	-217	moaC	817793	818278	16128751	molybdopterin biosynthesis, protein C
moaABCDE	moaAp1	816050	-217	moaD	818271	818516	16128752	molybdopterin synthase, small subunit
moaABCDE	moaAp1	816050	-217	moaE	818518	818970	16128753	molybdopterin synthase, large subunit
flgAMN	flgAp	1130108	-22	flgA	1129427	1130086	16129035	assembly protein for flagellar basal-body periplasmic P ring
flgAMN	flgAp	1130108	-22	flgM	1129058	1129351	16129034	anti-sigma factor for FliA (sigma 28)
flgAMN	flgAp	1130108	-22	flgN	1128637	1129053	16129033	export chaperone for FlgK and FlgL

The column headers help make the placement of data consistent.

Operon: name
Promoter: name
TSS absolute: a coordinate
TSS relative: a relative coordinate
Gene: name
Gene Start Loc: a coordinate
Gene Stop Loc: a coordinate
Protein GI: a cross-reference key
Protein Description: controlled vocab.

Note: a spreadsheet row will be referred to as an ‘instance’ , reflecting the view that this is a member of a set of things sharing the same types of attributes.

We can split up the data to reduce the repeats:

Operon	Promoter	Promoter TSS (Absolute)	Promoter TSS (Relative)
moaABCDE	moaAp1	816050	-217
flgAMN	flgAp	1130108	-22

Sheet 1:
Operon_Promoter

Sheet 2:
Gene_Protein

Operon	Gene	Gene Start Loc	Gene End Loc	Protein GI	Protein Description
moaABCDE	moaA	816267	817256	16128749	molybdopterin biosynthesis protein A
moaABCDE	moaB	817278	817790	16128750	molybdopterin biosynthesis protein B
moaABCDE	moaC	817793	818278	16128751	molybdopterin biosynthesis, protein C
moaABCDE	moaD	818271	818516	16128752	molybdopterin synthase, small subunit
moaABCDE	moaE	818518	818970	16128753	molybdopterin synthase, large subunit
flgAMN	flgA	1129427	1130086	16129035	assembly protein for flagellar basal-body periplasmic P ring
flgAMN	flgM	1129058	1129351	16129034	anti-sigma factor for FliA (sigma 28)
flgAMN	flgN	1128637	1129053	16129033	export chaperone for FlgK and FlgL

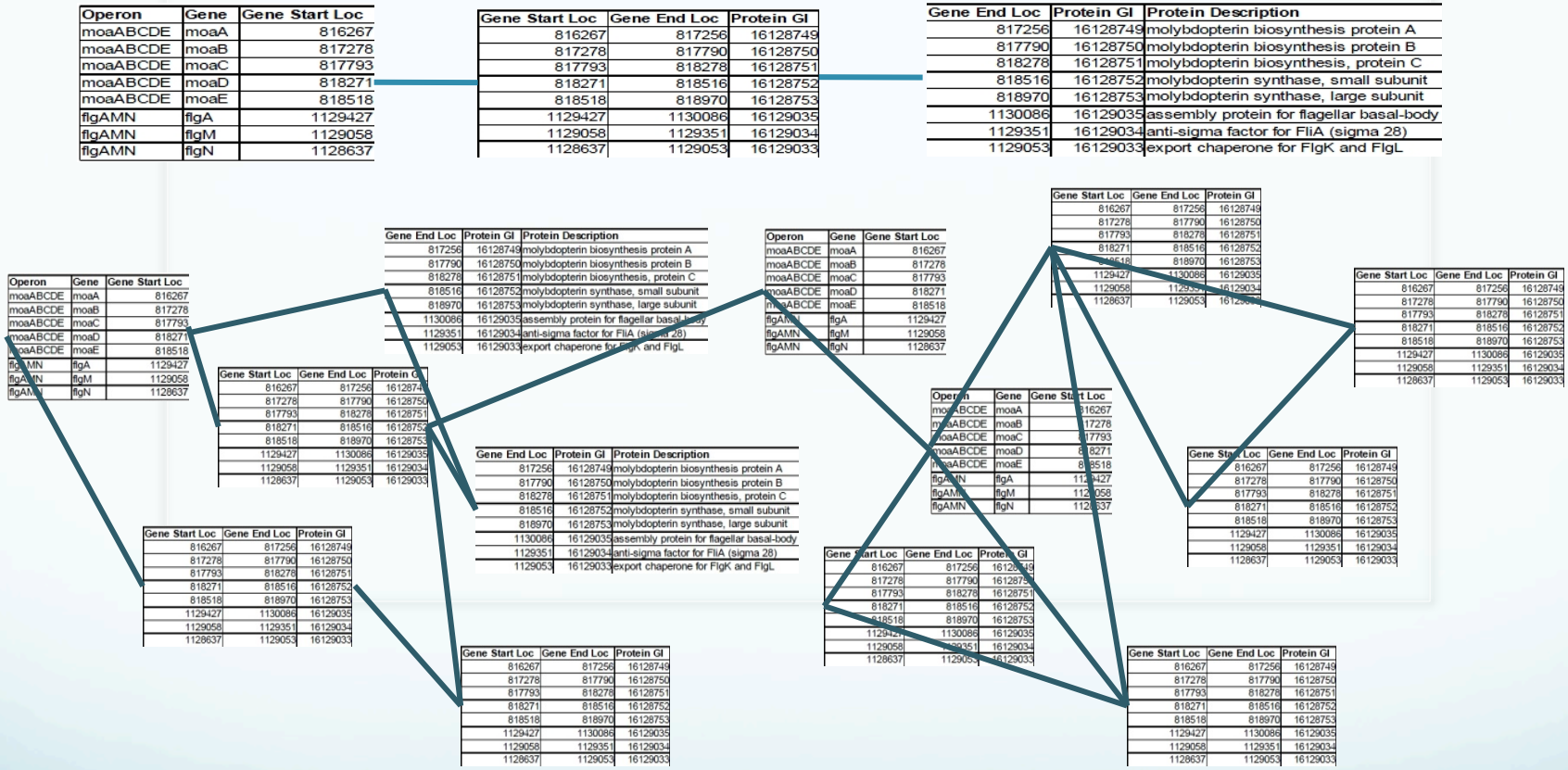
The sheets are linked to each other through the Operon column and the values for each instance. Spreadsheet applications may let you document this linkage, otherwise it is up to you to infer its presence.

The information in the spreadsheet is repetitive and redundant – why?

Operon	Promoter	TSS (Absolute)	TSS (Relative)	Gene	Gene Start Loc	Gene End Loc	Protein GI	Protein Description
moaABCDE	moaAp1	816050	-217	moaA	816267	817256	16128749	molybdopterin biosynthesis protein A
moaABCDE	moaAp1	816050	-217	moaB	817278	817790	16128750	molybdopterin biosynthesis protein B
moaABCDE	moaAp1	816050	-217	moaC	817793	818278	16128751	molybdopterin biosynthesis, protein C
moaABCDE	moaAp1	816050	-217	moaD	818271	818516	16128752	molybdopterin synthase, small subunit
moaABCDE	moaAp1	816050	-217	moaE	818518	818970	16128753	molybdopterin synthase, large subunit
flgAMN	flgAp	1130108	-22	flgA	1129427	1130086	16129035	assembly protein for flagellar basal-body periplasmic P ring
flgAMN	flgAp	1130108	-22	flgM	1129058	1129351	16129034	anti-sigma factor for FliA (sigma 28)
flgAMN	flgAp	1130108	-22	flgN	1128637	1129053	16129033	export chaperone for FlgK and FlgL

The 1:M relations require that we enter the same value many times.

If you want to further simplify any given sheet, the number of sheets will increase – a management challenge ensues.



Operon	Gene	Gene Start Loc
moaABCDE	moaA	816267
moaABCDE	moaB	817278
moaABCDE	moaC	817793
moaABCDE	moaD	818271
moaABCDE	moaE	818518
flgAMN	flgA	1129427
flgAMN	flgM	1129058
flgAMN	flgN	1128637

Gene Start Loc	Gene End Loc	Protein GI
816267	817256	16128749
817278	817790	16128750
817793	818278	16128751
818271	818516	16128752
818518	818970	16128753
1129427	1130086	16129035
1129058	1129351	16129034
1128637	1129053	16129033

Gene End Loc	Protein GI	Protein Description
817256	16128749	molybdopterin biosynthesis protein A
817790	16128750	molybdopterin biosynthesis protein B
818278	16128751	molybdopterin biosynthesis, protein C
818516	16128752	molybdopterin synthase, small subunit
818970	16128753	molybdopterin synthase, large subunit
1130086	16129035	assembly protein for flagellar basal-body
1129351	16129034	anti-sigma factor for FliA (sigma 28)
1129053	16129033	export chaperone for FlgK and FlgL

As the number of sheets grows you tend to add one more: a Master Plan reminding yourself what the individual related sheets include.

Linking sheets as shown can simplify data entry, but now finding and retrieving data may be very complicated (even with linked tables).

The output of a microarray platform produces 5000 measurements, and the experiment of interest had 21 arrays, triplicates of 7 conditions.

Rows: 21*5000 = 105,000
Storage Feasibility: how many rows does Excel 2007 support?
Retrieval Feasibility:
part 1: locate all raw expression values for one probe across all arrays.
Part 2: How many of the probes produce valid measurements [log10(m) >2.4] across all of the arrays?

What you get: a database is an electronic system for managing data and the meta-data describing that data: properly done, the system is self-describing.

- The management software handles the complex structures, the allocation of files and bits, identification and retrieval tasks.
- The data dictionary holds the meta-data
  - What type of entity is described?
  - What are the attributes of each entity?
  - What are the relationships between the entities?
  - What are the constraints that need to be applied to attribute values and relationships?

Data Records – some vocabulary

Probe_Name	Intensity at 680nm	Intensity at 650 nm	Background at 680nm	Background at 650nm
At_100912	1256	2580	250	126
At_008973	2690	1387	127	95
At_205439	158	128	127	101
At_900196	12876	25908	278	309

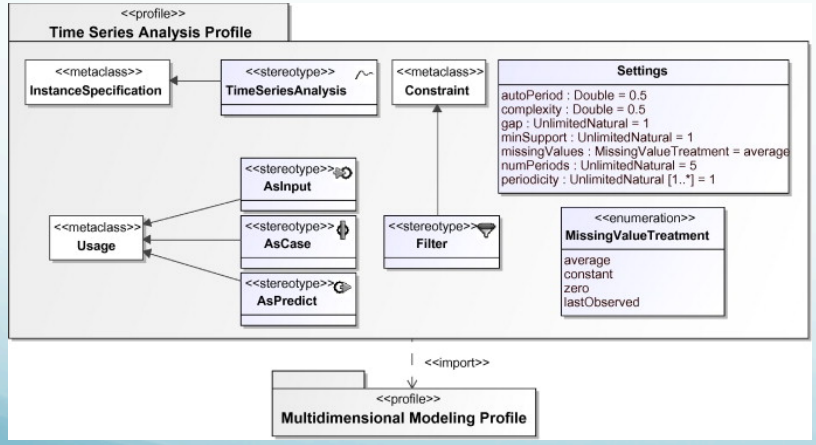
- The table format is convenient as an organizing principle
  - Each *field* is a (row x column) intersection with a particular type of information that has meaning, in context
  - A column is a consistent type of information – a type of attribute
  - A row is a record, or an entity instance, every row in this table will show the same number of fields, for different instances, each position in the set of fields will have an attribute of the same type
  - A table is a file containing the collection (entity) of similar records (set of entity instances, or all the rows).

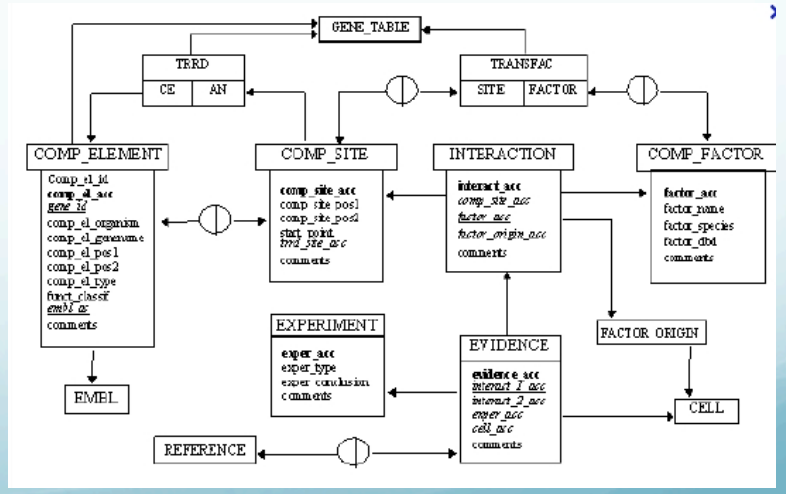
Basic Data Relationships

- Determine the *what* – the entities
- List the most important attributes
- Link the entities together with one of the 3 relationship types:
  - One-to-one
  - One-to-many
  - Many-to-Many
- Make sure there are attributes that allow the linking to happen: relationships are stored in the database between entity instances (rows in two tables).
  - Controlled redundancy: the two tables share one type of attribute (one column) in common – this permits logical linkages between tables
    - The columns are declared to contain **keys** in the data dictionary – this tells the application that the tables are linked.

Data Structure Diagram

- A data structure diagram is a representation of a conceptual model that uses standardized graphical notations and terms to document entities and their relationships and transformations.
  - Constraints are also noted
- For database models the entity-relationship (ER) model is common
  - Chen notation or Information Engineering (IE or Crow's foot) notation is most common.
  - UML is also used but can get very complicated





If you repeat information you waste space and possibly cause entry errors, leading to inconsistencies.

Operon	Promoter	TSS (Absolute)	TSS (Relative)	Gene	Gene Start Loc	Gene End Loc	Protein GI	Protein Description
moaABCDE	moaAp1	816050	-217	moaA	816267	817256	16128749	molybdopterin biosynthesis protein A
moaABCDE	moaAp1	816050	-217	moaB	817278	817790	16128750	molybdopterin biosynthesis protein B
moaABCDE	moaAp1	816050	-217	moaC	817793	818278	16128751	molybdopterin biosynthesis, protein C
moaABCDE	moaAp1	816050	-217	moaD	818271	818516	16128752	molybdopterin synthase, small subunit
moaABCDE	moaAp1	816050	-217	moaE	818518	818970	16128753	molybdopterin synthase, large subunit
flgAMN	flgAp	1130108	-22	flgA	1129427	1130086	16129035	assembly protein for flagellar basal-body periplasmic P ring
flgAMN	flgAp	1130108	-22	flgM	1129058	1129351	16129034	anti-sigma factor for FliA (sigma 28)
flgAMN	flgAp	1130108	-22	flgN	1128637	1129053	16129033	export chaperone for FlgK and FlgL

Memory is cheap, but it is a waste to store the same information multiple times.

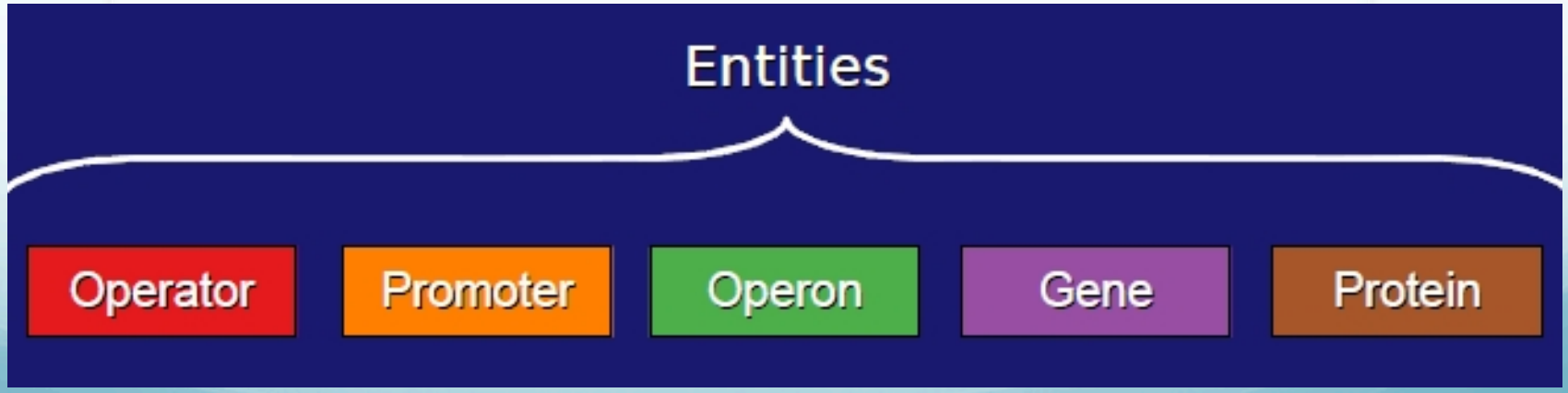
If you need to change something you must *change every instance* of it, so you have to find every instance of it.

If you lose track of one of 5 instances of moaABCDE and change the other 4, now attributes that should be identical will vary.

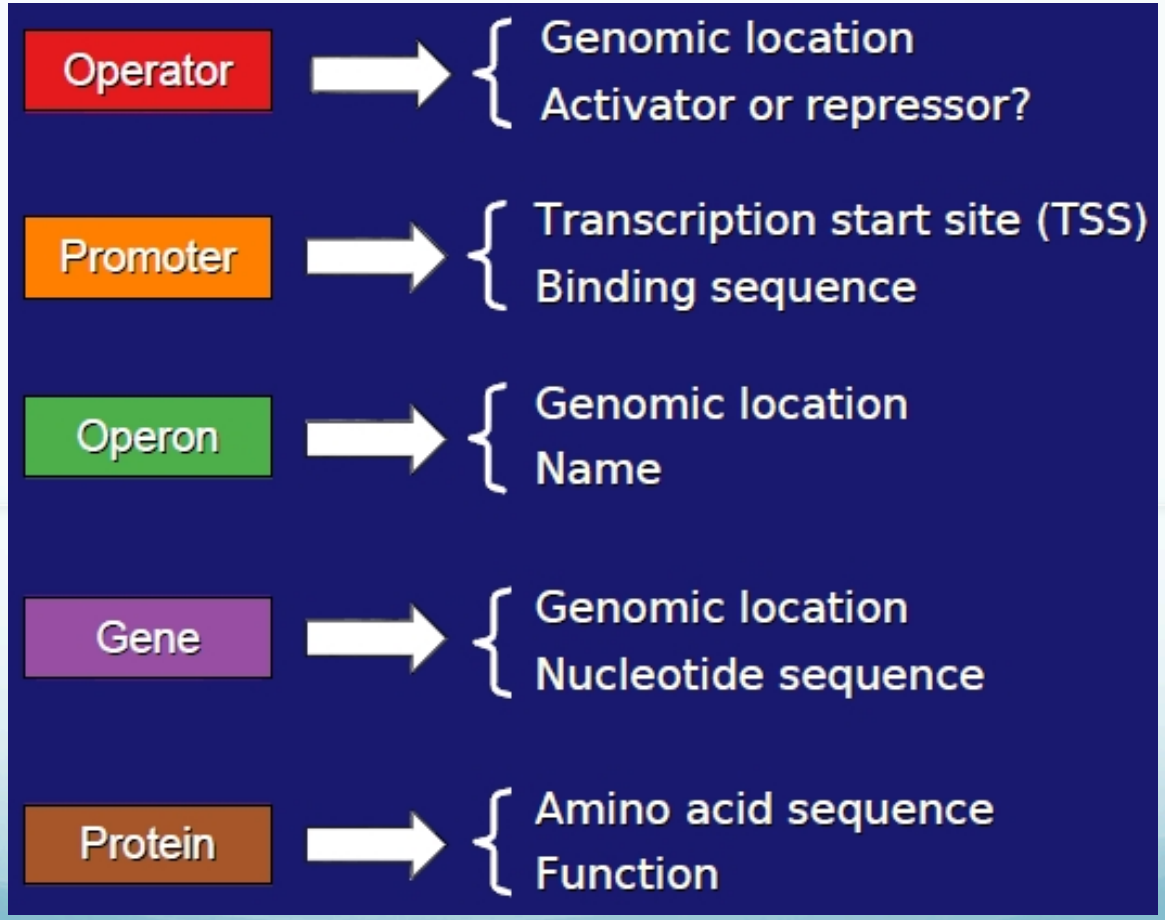
An entity is a thing you can name (a noun) that has properties whose descriptions you want to store.

Nouns: person, place, object, event, idea

Going to our operon biological construct, what are the entities?

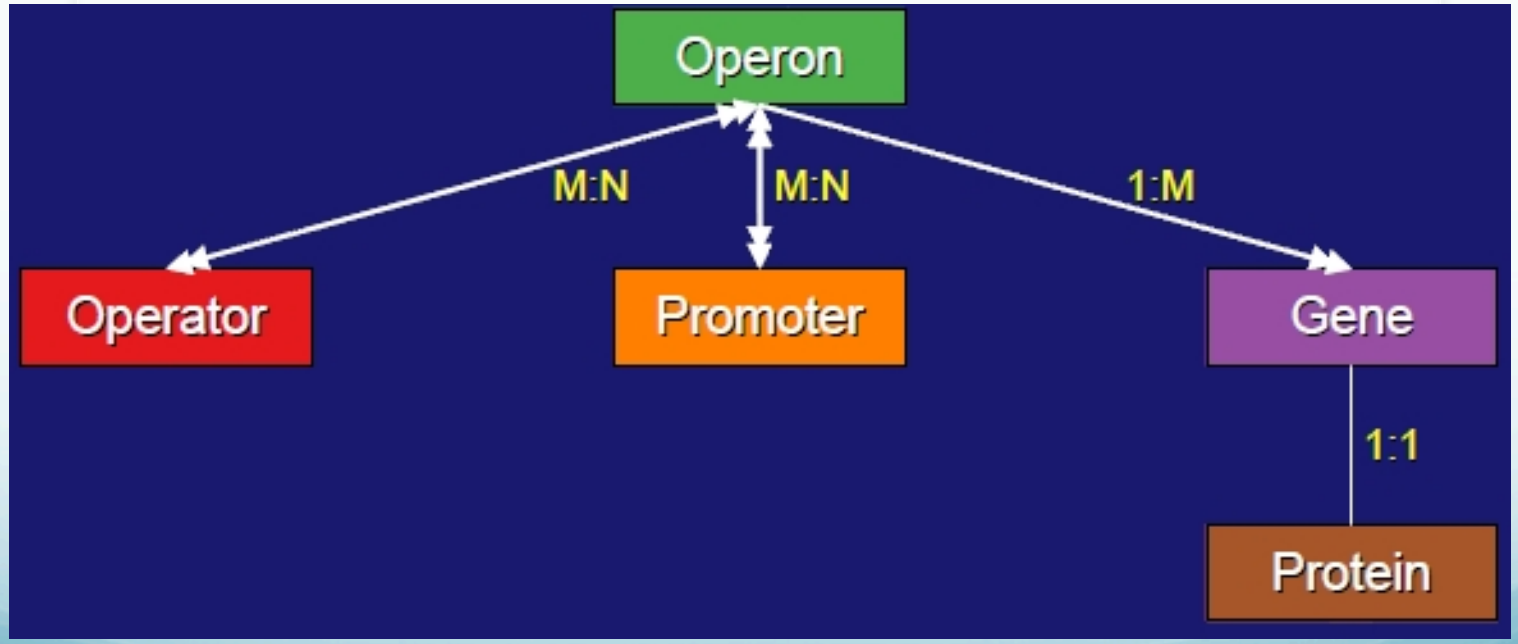


The terms used to describe the entity are called attributes (adjectives) each type of property is an attribute.

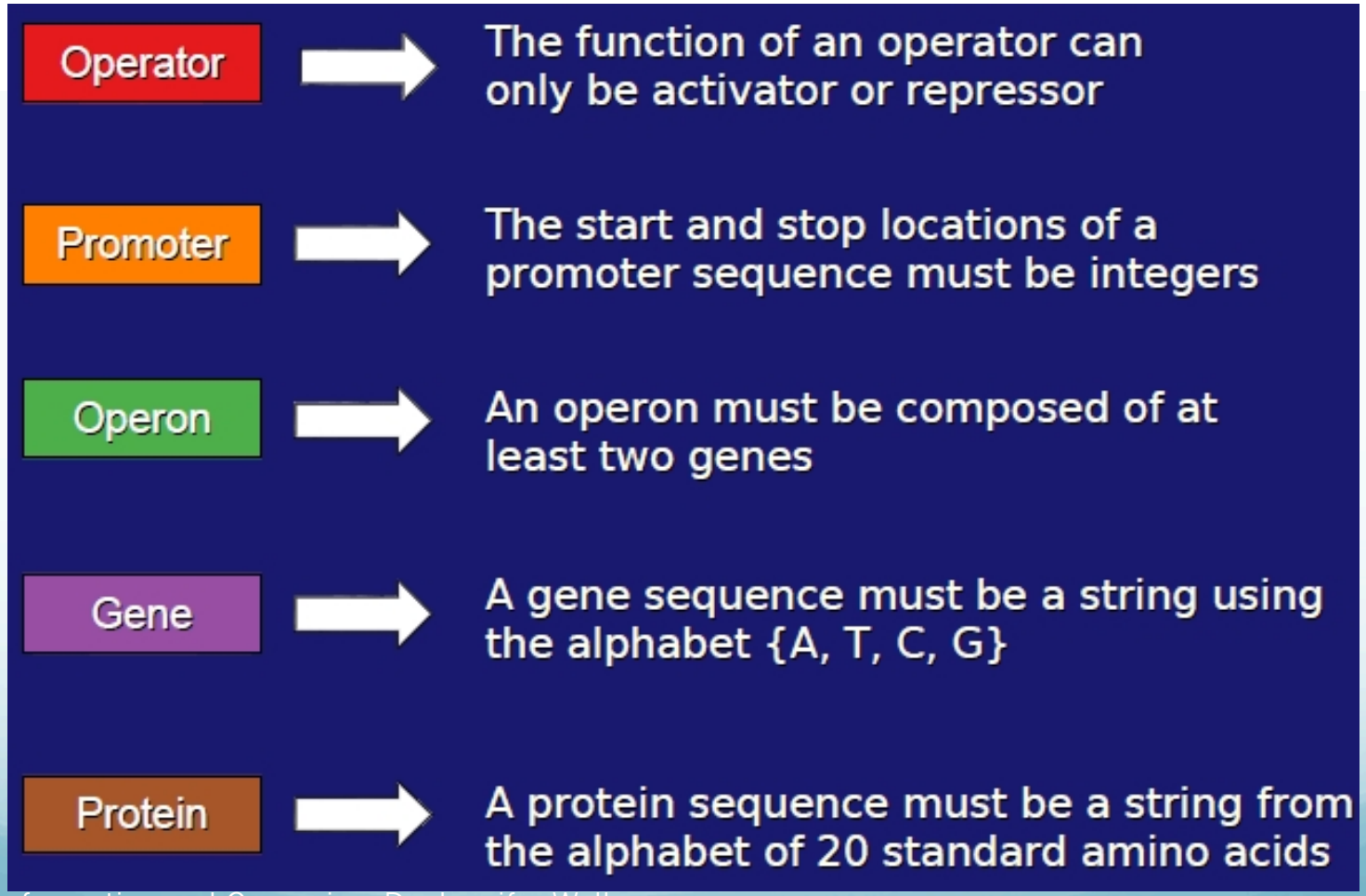


In complex systems like cells one part often acts upon other parts: the action is a relationship (verb).

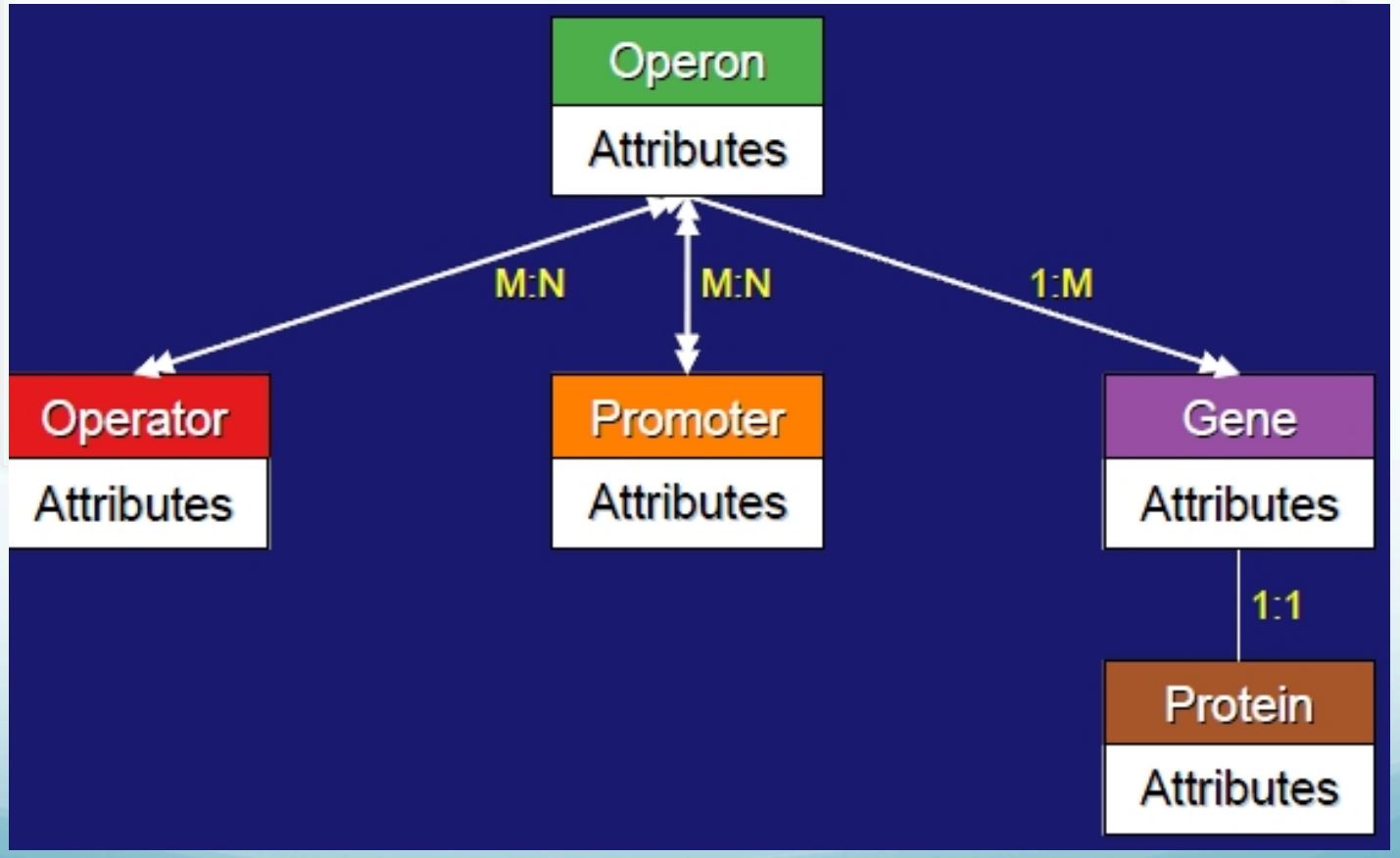
The actions have been divided into 3 types: one-to-one (1:1), one-to-many (1:M) and many-to-many (M:N).



Most attributes are not infinite or continuous – there is a range across which they apply. The range is called a *constraint*.



An ER Diagram – so far.



It is possible to improve spreadsheets, of course, as shown below.

Operon	Promoter	Promoter TSS (Absolute)	Promoter TSS (Relative)
moaABCDE	moaAp1	816050	-217
flgAMN	flgAp	1130108	-22

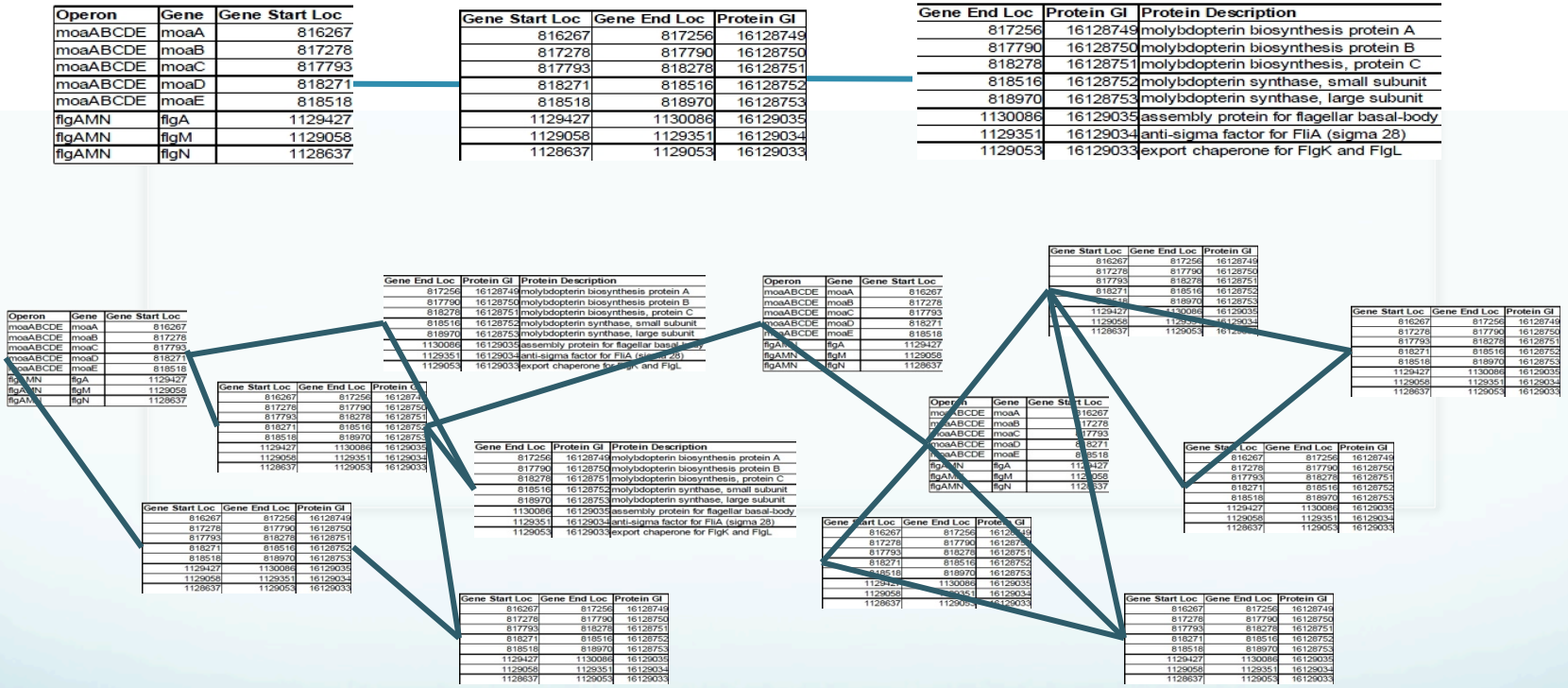
Sheet 1: Operon_Promoter

Sheet 2: Gene_Protein

Operon	Gene	Gene Start Loc	Gene End Loc	Protein GI	Protein Description
moaABCDE	moaA	816267	817256	16128749	molybdopterin biosynthesis protein A
moaABCDE	moaB	817278	817790	16128750	molybdopterin biosynthesis protein B
moaABCDE	moaC	817793	818278	16128751	molybdopterin biosynthesis, protein C
moaABCDE	moaD	818271	818516	16128752	molybdopterin synthase, small subunit
moaABCDE	moaE	818518	818970	16128753	molybdopterin synthase, large subunit
flgAMN	flgA	1129427	1130086	16129035	assembly protein for flagellar basal-body periplasmic P ring
flgAMN	flgM	1129058	1129351	16129034	anti-sigma factor for FliA (sigma 28)
flgAMN	flgN	1128637	1129053	16129033	export chaperone for FlgK and FlgL

For this to work you must know (implicit) that the two sheets are related by the 'Operon' column – you find the correct row by using the value in the field in that column.

If you want to further simplify any given sheet, the number of sheets will increase – a management challenge ensues.



Operon	Gene	Gene Start Loc
moaABCDE	moaA	816267
moaABCDE	moaB	817278
moaABCDE	moaC	817793
moaABCDE	moaD	818271
moaABCDE	moaE	818518
flgAMN	flgA	1129427
flgAMN	flgM	1129058
flgAMN	flgN	1128637

Gene Start Loc	Gene End Loc	Protein GI
816267	817256	16128749
817278	817790	16128750
817793	818278	16128751
818271	818516	16128752
818518	818970	16128753
1130086	16129035	assembly protein for flagellar basal-body
1129351	16129034	anti-sigma factor for FliA (sigma 28)
1129053	16129033	export chaperone for FlgK and FlgL

Gene End Loc	Protein GI	Protein Description
817256	16128749	molybdopterin biosynthesis protein A
817790	16128750	molybdopterin biosynthesis protein B
818278	16128751	molybdopterin biosynthesis, protein C
818516	16128752	molybdopterin synthase, small subunit
818970	16128753	molybdopterin synthase, large subunit
1130086	16129035	assembly protein for flagellar basal-body
1129351	16129034	anti-sigma factor for FliA (sigma 28)
1129053	16129033	export chaperone for FlgK and FlgL

In the second example, you will end up with a master sheet that manages the individual data sheets.

Data model levels

- Conceptual data modeling, AKA the Use Case – highest level relationships and rules. It includes
  - The scope of the model – WHAT is stored
  - Used for communicating – standard symbols and texts are employed (ER modeling, UML) to developers and users.
  - The processes, such as inputs (and sources), outputs (applications that produce these), transformation steps (for example a reference location based on creating a SAM file)
- Logical data modeling - defines the elements and their characteristics and the relationships that interconnect them.
- Physical modeling - application of the logical data model using database management software (DBMS) – how data are stored.

- Note: the logical data model explicitly determines the structure of data and limits the type of DBMS that will be effective.

Data Model Levels

- External
  - End user view
  - Basic representation
- Conceptual.
  - Linkage to schema
  - Greater detail
- Internal model
  - Code, Script, Implementation
- Physical Model
  - How the data is actually stored
    - Binary

If you are sharing the data with other scientists, how do you organize access? How do you prevent others from changing the data?

Security: it is possible to lock fields and sheets to limit accidental changes but this is a very limited approach.

Concurrent access: there is no way for multiple people to use the same sheet at the same time without making separate copies.

Relational Table Characteristics

- Two dimensional (rows and columns)
- Each row (tuple) is a single entity set.
- Every column has distinct name (attribute)
- Row+column = single data value
- All values in the same format within a column
- Columns have specific ranges (attribute domain)
- Order of rows and columns is permutable
- There exists at least one attribute(s) that uniquely identifies each row