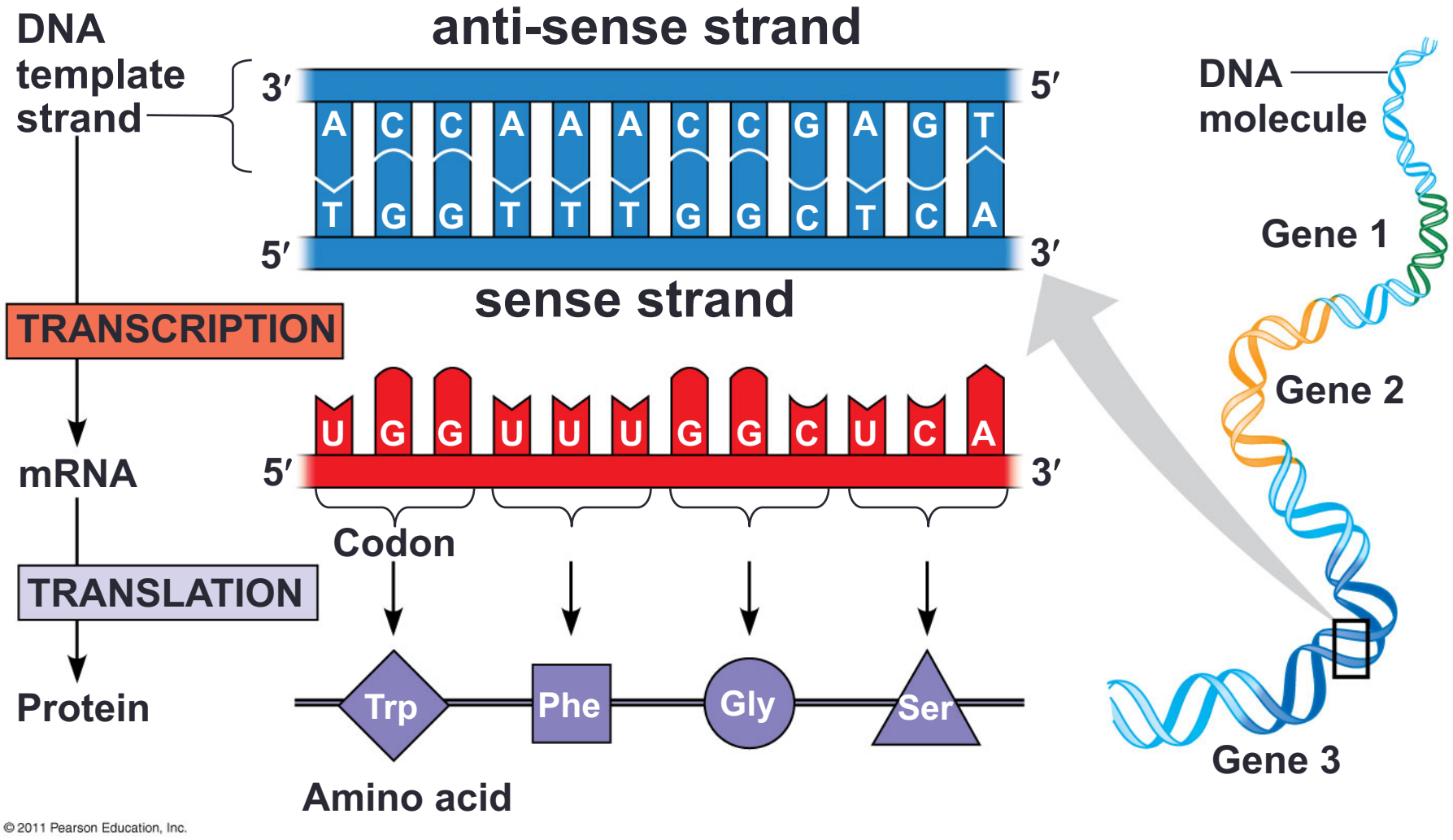
CHAPTER 3: DEALING WITH DATABASES

Dr. Garrett Dancik



- If we know the **DNA** sequence (and know the gene structure)
  - Then we can predict the **mRNA** sequence
    - Then we can predict the **protein** sequence
- If we know the protein sequence, then we can (possibly) predict
  - **secondary and tertiary structure, domains**
  - **protein function**

What is a database?

- A **database** is a repository of information with a specific structure, that enables entering of and extraction of data
- Many databases today are electronic, which enables efficient searching
- It is useful to think of a database as a table (whether or not the data are displayed that way or not
  - A row in the table corresponds to each entry, or **record**, of the database
  - Each column of the table corresponds to a **field,** and each record has the same set of fields (which may be blank for some records)
  - In general, a record consists of information for one or more fields

Examples of flat-files databases

- A **flat-file** database is the simplest database where collections of data are stored as single text files or a collection of different text files

(A)

NAME	TELEPHONE	ADDRESS
S. Claus	0203 450	The North Pole, Lapland
M. Mouse	0202 453	Disneyworld, Florida
A. Moonman	0104 459	Craterland, The Moon

(B) GenBank Flat-File Format http://www.ncbi.nlm.nih.gov

```
LOCUS       SCU49845     5028 bp    DNA
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and
            Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina;
           Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
```

Relational databases

- A relational database is the most common database for biological information
- A relational database stores data in multiple tables
- Each table is linked to at least one other table through a shared field called a **key**, which must be unique

Let's convert the following database to a relational database

Single table

Name	Gender	CityState	Item	Description	Purchase Date
Bob Smith	Male	Willimantic, CT	Jeans	Description of jeans	1/19/2014
Bob Smith	Male	Willimantic, CT	Dog Food	Description of dog food	1/19/2014
Jane Doe	Female	Hartford, CT	Shoes	Description of shoes	2/02/2014
Jane Doe	Female	Hartford, CT	Dog Food	Description of dog food	2/02/2014

Rather than store information in a single table, we want to spread information across multiple tables in order to avoid redundant data.

Relational Database example

Customer table

CustomerID	Name	Gender	CityState
1	Bob Smith	Male	Willimantic, CT
2	Jane Doe	Female	Hartford, CT

Item table

ItemID	Name	Description
1	Jeans	Description of jeans
2	Dog Food	Description of dog food
3	Shoes	Description of shoes

Transaction table

PurchaseID	Customer ID	Item ID	PurchaseDate
1	1	1	1/19/2014
2	1	2	1/19/2014
3	2	3	2/02/2014
4	2	2	2/02/2014

CustomerID, ItemID, and PurchaseID are **keys** that uniquely identify each customer, item, and purchase, respectively.

Relational databases in bioinformatics

protab1

Protein-code	Protein-name	Length	Species-origin
P1001	Hemoglobin	145	Bovine
P1002	Hemoglobin	136	Ovine
P1003	Eye Lens Protein	234	Human
.....			

protab2

Protein-code	Protein-sequence
P1001	MDRTTHGFDLKLLSPRTVNQWLMLALFFGHS...
P1002	MDKTSHGFEIKLLTPKKLQQWLMIAIYFGHT...
P1003	SRTHEEEGKLMQWPPRPLYIALFTEPPYP...
.....	

Overview of databases

- Many (thousands of) bioinformatics databases are accessible via the internet
- Although databases are generally built around one biological aspect (e.g., DNA sequences), they will often link to external relevant information (e.g., protein sequences)
- The underlying biological data (usually experimentally determined), is refered to as the **data**
- Additional information (research citations, links to other databases, interpretation of the data) is referred to as **annotation**
- **Primary databases** contain data that is experimentally derived
- **Secondary** databases contain data that is predicted from primary data.

How many databases are there?

- Every year the journal *Nucleic Acids Research* has an issue devoted to new and updated database
  - They currently list >1600 databases
  - http://www.oxfordjournals.org/nar/database/c

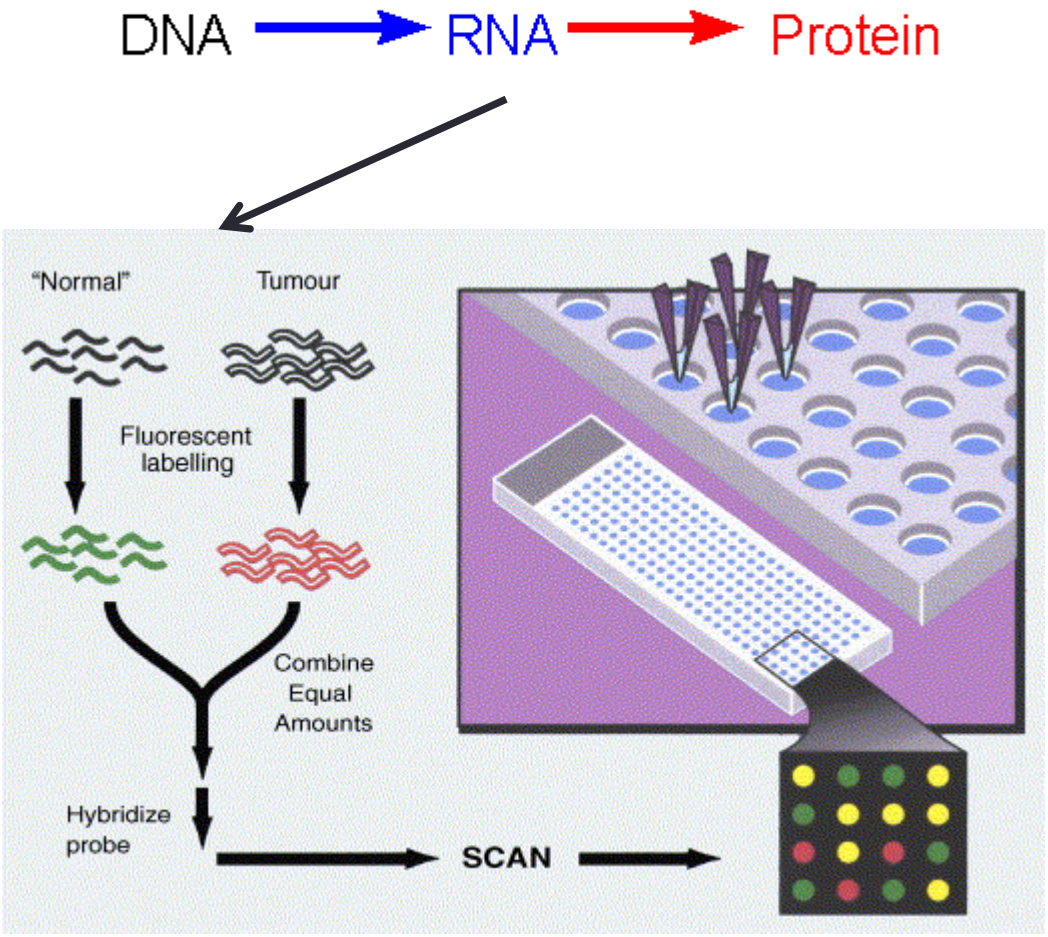
- Selected database types
  - Sequence databases
    - DNA
    - RNA
    - Protein
  - Microarray and gene expression databases
  - Oganelle databases
  - Plant databases
  - Immunological databases

Sequence databases – generating data

- A DNA sequence is a sequence of chromosomal DNA and contains introns, exons, and untranslated regions.
- Complementary DNA, or **cDNA**, are DNA sequences synthesized from reverse transcription of a mRNA.
  - cDNA sequences correspond to genes that are expressed at the time of sampling
  - cDNA will not contain any introns or control sequences (such as the promoter) that are not transcribed
- An **expressed sequence tag** (**EST**) is a partial cDNA sequence
- Protein sequences are generated experimentally or are translated from nucleotide sequence data

Microarrays and gene expression databases

- Microarrays measure gene expression
- Databases include Gene Expression Omnibus (GEO), Array Express, and others



Bowden and Brennan. Br J Oral Maxillofac Surg 2004; 42(6): 42-545

Additional database types

- Protein interaction databases
  - Proteins must interact with other proteins or DNA/RNA to carry out their function
  - **Systems** biology involves studying these interactions (e.g., biological networks) in order to understand the dynamic behavior of a cell or organism
- Structural databases
  - Structure of DNA, RNA, or protein

Understanding Data Quality

- Always remember that Garbage in = Garbage out!
- If multiple individuals sequence the same gene, then the data produced will be **redundant** (identical or nearly identical)
- A **non-redundant** database finds a consensus sequence which summarizes the redundant entries
- Data consistency can be checked automatically
  - DNA sequences should consist of only the letters A,C,G, and T, though experimental uncertainty can be recorded
  - Protein structure has physical limitations (bonding geometry), and errors can be identified, and either corrected or annotated
- Ontologies were developed so consistent naming conventions are used
  - http://www.genenames.org
  - http://www.geneontology.org

Understanding data quality

- If a gene or protein is labeled as "hypothetical", "putative", or "predicted", then it has been predicted using computational methods
- Database entries generally have version numbers that allow for tracking of changes