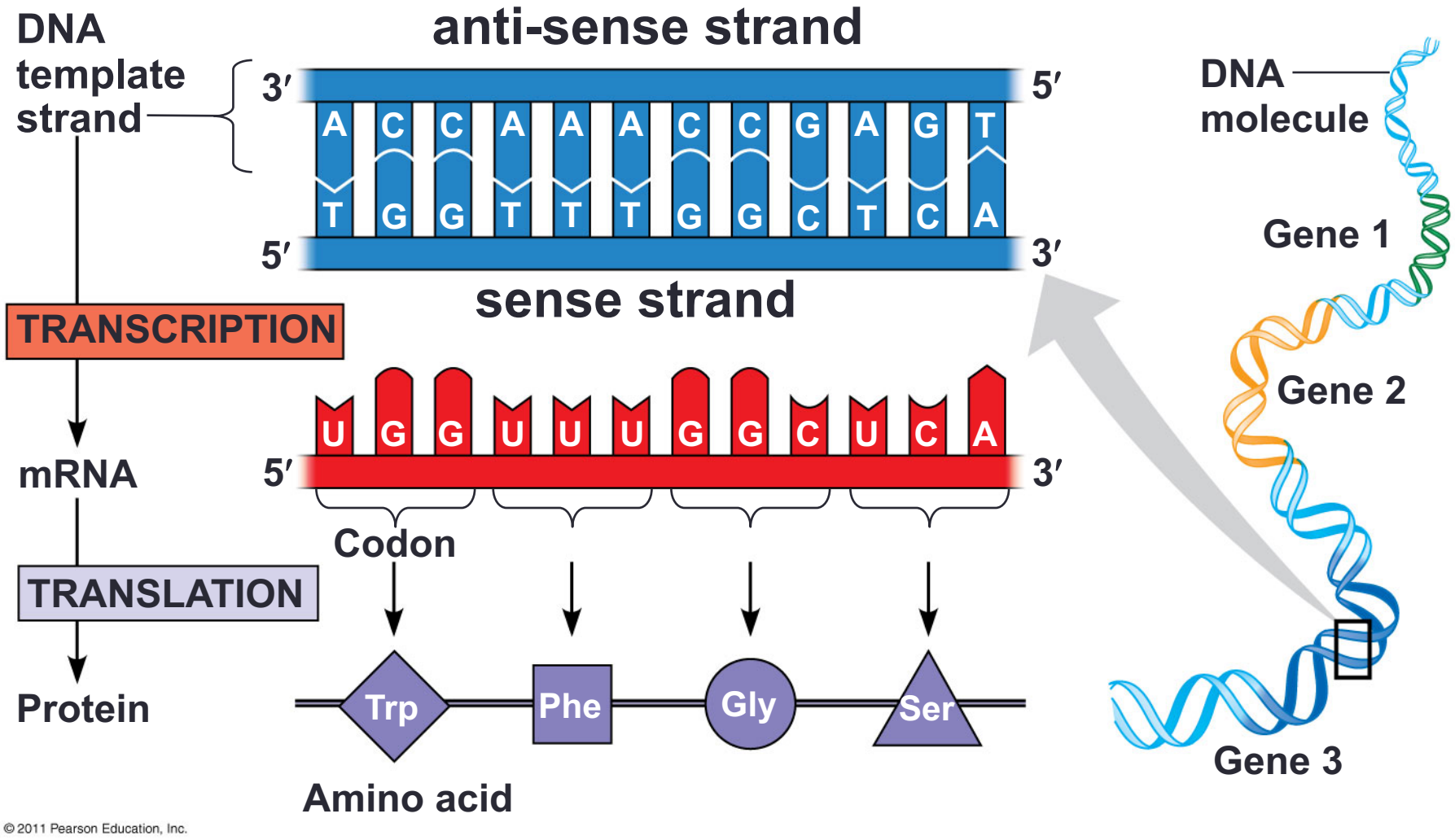
BIOINFORMATICS DATABASES

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**

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 referred 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
  - Organelle 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

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
- In GenBank and GenPept, we will generally look at RefSeq entries, which are non-redundant and well annotated
  - RefSeq accession numbers have meaningful prefixes:
    - See table 1 from https://www.ncbi.nlm.nih.gov/books/NBK21091/