Introduction to Molecular Biology

Part 2

DNA & RNA: Flow of Information

aka “The Central Dogma”!!
DNA to RNA to Protein

A gene is expressed in two steps

– **Transcription**: RNA Synthesis
– **Translation**: Protein Synthesis

The Code Book

- DNA, RNA, and Proteins are examples of strings written in either the four-letter nucleotide of DNA and RNA (A C G T/U)
- Or the twenty-letter amino acid sequences that make up proteins. Each amino acid is coded by 3 nucleotides called **codons**
DNA & RNA

- DNA = Deoxyribonucleic acid
- RNA = Ribonucleic acid
- They are almost the same...
- There is no T base in RNA
- A similar base U takes its place
- An oxygen atom is added to the sugar component of RNA

How are proteins made...

DNA: TACCGCGGCTATTACTGCCAGGAAGGAACCT
How are proteins made...

DNA: TAC CGC GGC TAT TAC TGC CAG GAA GGA ACT

How are proteins made: Transcription

DNA: TAC CGC GGC TAT TAC TGC CAG GAA GGA ACT
RNA: AUG GCG CCG AUA AUG ACG GUC CUU CCU UGA
How are proteins made: Transcription

DNA: TAC CGC GGC TAT TAC TGC CAG GAA GGA ACT
RNA: AUG GCG CCG AUA AUG ACG GUC CUU CCU UGA

How are proteins made: Translation

DNA: TAC CGC GGC TAT TAC TGC CAG GAA GGA ACT
RNA: AUG GCG CCG AUA AUG ACG GUC CUU CCU UGA
Pro: Met Ala Pro Ile Met Thr Val Leu Pro Stop
DNA to RNA to Protein

A gene is expressed in two steps

– **Transcription:** RNA Synthesis
– **Translation:** Protein Synthesis
Splicing

Terminology

- **Codon**: The sequence of 3 nucleotides in DNA/RNA that encodes for a specific amino acid.

- **mRNA (messenger RNA)**: A ribonucleic acid whose sequence is complementary to that of a protein-coding gene in DNA.

- **Ribosome**: The organelle that synthesizes polypeptides under the direction of mRNA.

- **rRNA (ribosomal RNA)**: The RNA molecules that constitute the bulk of the ribosome and provides structural scaffolding for the ribosome and catalyzes peptide bond formation.

- **tRNA (transfer RNA)**: The small L-shaped RNAs that deliver specific amino acids to ribosomes according to the sequence of a bound mRNA.
Revisiting the Central Dogma

• In going from DNA to proteins, there is an intermediate step where mRNA is made from DNA, which then makes protein

• Why the intermediate step?
  – DNA is kept in the nucleus, while protein synthesis happens in the cytoplasm, with the help of ribosomes

Proteins

• Proteins do all essential work for the cell
  – build cellular structures
  – digest nutrients
  – execute metabolic functions
  – Mediate information flow within a cell and among cellular communities.

• Proteins are often enzymes that catalyze reactions.
• Also called “poly-peptides”
Polypeptide vs Protein

• A protein is a polypeptide, however to understand the function of a protein given only the polypeptide sequence is a very difficult problem.

• **Protein folding** is an open problem. The 3D structure depends on many variables.

• Current approaches often work by looking at the structure of homologous (similar) proteins.

  > Improper folding of a protein is believed to be the cause of mad cow disease.

Protein Folding

• Proteins are not linear structures, though they are built that way

• The amino acids have very different chemical properties; they interact with each other after the protein is built
  – This causes the protein to fold and adopt it’s functional structure
  – Proteins may fold in reaction to some ions, and several separate chains of peptides may join together through their hydrophobic and hydrophilic amino acids to form a polymer
Protein Folding

• The structure that a protein adopts is vital to its chemistry

• Its structure determines which of its amino acids are exposed to carry out the protein’s function

• Its structure also determines what substrates it can react with

How are proteins made...

DNA: TAC CGC GGC TAT TAC TGC CAG GAA GGA ACT
RNA: AUG GCG CCG AUA AUG ACG GUC CUU CCU UGA
Pro: Met Ala Pro Ile Met Thr Val Leu Pro Stop
DNA Sequencing: Shotgun Sequencing

From gel to reads...
DNA Sequencing: Shotgun Sequencing

<table>
<thead>
<tr>
<th>Target Genome</th>
<th>ATTTGCGCAGAGACCTAAGGCTATTAGCTTTGCCCCTAAAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reads</td>
<td>ATTTGC AGAGACCTAAG TTAGCTTGCC AAG TGGCCCTAA</td>
</tr>
<tr>
<td>Overlapping</td>
<td>ATTTGC AGAGACCTAAG TTAGCTTGCC AAG TGGCCCTAA</td>
</tr>
<tr>
<td>Contigs</td>
<td>ATTTGCGCAGAGACCTAAG TTAGCTTGCCCCTAAAG</td>
</tr>
</tbody>
</table>

DNA Sequencing Problem

- Given a set of sequences, find the minimal length string containing all members of the set as substrings.

Assume that you take many copies of a book, pass each of them through a shredder with a different cutter, and then you try to make the text of the book back together just by gluing together the shredded pieces. It is obvious that this task is pretty difficult. Furthermore, there are some extra practical issues as well. The original copy may have many repeated paragraphs, and some shreds may be modified during shredding to have typos. Parts from another book may have also been added in, and some shreds may be completely unrecognizable.
Sequencing Process

- **Sanger Sequencing, 1980s**
  - uses gel electrophoresis
  - process simplified over years (fluorescent method)
- **Drawbacks**
  - expensive and time consuming
  - mostly good for shorter genomes (viruses and bacterial DNA)

Next Generation Sequencing

- **Next Generation Sequencing**
  - inexpensive and produce a huge amount of data (reads) quickly
  - reads are 100 to 300 base pairs long
  - Current equipment can produce reads that are 100 to 500 long
  - # reads produced varies between 500 million to 20 billion
    - (i.e. This data is 100GB to 6000GB – BIG DATA!)
- **Drawbacks**
  - produce short sequences that limit the sequence assembly process
Third Generation Sequencing

- Can provide longer reads (50 to 60 K base pairs)
  exploits GC-rich regions of the genome
  ensures uniform coverage
- Drawbacks
  higher error rates (15 to 20%)
  but randomness of reads, with sufficient coverage can help
  reduce errors

Information Theory of DNA Sequencing

- What is the efficiency of constructing a complete genome sequence, given millions of short reads?
- What is the minimum number of reads required for reliable reconstruction?
- What are the fundamental limits of any sequence assembly algorithm?
Some parameters

• Throughput
  How much of the genome can be sequenced (in a single pass, or requires repetitions)

• Speed
  How fast is the process from DNA to genome assembly?

• Sequencing Quality - $Q = -10 \log_{10}(e)$
  $e$ is the estimated probability that the base is incorrect
  Higher values of $Q$ indicate smaller probability of error
  Lower values of $Q$ can render reads unusable in assembly

• Coverage - $C = \frac{N \times L}{G}$
  $N$ - # of reads, $L$ – reads length, $G$ – length of genome
  A measure of redundancy with which a sequence assembly process begins

Sequencing Capacity

• Human genome is $3 \times 10^9$ bases (G)
• Individual reads are 100-1000 base pairs (L)
• There are 10s’ to 100’s million reads (N) depending on process

  e.g. For $G = 3 \times 10^9$, $L = 300$, $N = 80$ million, Coverage is

  $$C = \frac{80,000,000 \times 300}{3,000,000,000} = 80\%$$

  • In fact, 80% coverage implies there is sufficient data to cover 100% of the genome.
Optimal Assembly

- What are the necessary and sufficient conditions for genome assembly?
- Is there a lower bound on the read length (L) to provide a complete reconstruction of a genome?
- What are the optimal assembly algorithms? If any?
- I.e. say with 80% coverage, what is the combination of minimal N and L for a given genome of size G?
Sequence Analysis

- **Sequence Databases** (e.g. GenBank)
  Primary (raw sequence data), secondary (biological knowledge)

- **Sequence Alignment** (global, local, multiple)
  Needed for structural, functional, and evolutionary inferences. Motifs, domains...

- **Gene & Promoter Prediction**
  open reading frames, exons, introns, ...

- **Molecular Phylogenetics**
  Evolutionary history of living organisms, phylogenetic tree construction, ...

Structural Bioinformatics

- **Protein Structure**
  Protein functions are determined by their structure
  Databases, Visualization, Classification

- **Protein Structure Prediction**

- **Protein Structure Comparison**

- **RNA Structure Prediction**
Genomics & Proteomics

- **Structural Genomics**
  Genome mapping, sequence, assembly, annotation, comparison

- **Functional Genomics**
  Gene expression

- **Proteomics**
  Entire set of expressed proteins in a cell

Computation

- **Exhaustive Search**
  regulatory motifs in DNA, profiles

- **Greedy Algorithms**
  genome rearrangements, motif search

- **Dynamic Programming Algorithms**
  DNA sequence comparison, alignment, gene prediction

- **Divide-and-Conquer Algorithms**
  sequence alignment

- **Graph Algorithms**
  DNA sequencing, fragment assembly, peptide sequencing

- **Combinatorial Pattern Matching**
  similarity search, database searches

- **Clustering and Trees**
  gene expression analysis, tree construction

- **Hidden Markov Models**
  profile alignment

- **Randomized Algorithms**
- **Machine Learning**
- **Genome Compression & Search**
Molecular Biology: Challenges

How does the structure and function at the molecular level account for the hierarchy?

- Molecular
- Intracellular
- Intercellular
- Tissue
- Organism
- Communities

More –omics (but no C!!)
More –omics (but no C!!)

...and interactomics!
References

• Adapted from slides posted at the web site of the above book.
• What is Metabolomics? https://www.ebi.ac.uk/training/online/course/introduction-metabolomics/what-metabolomics [2019]