CS 5984
Algorithms in Bioinformatics

An Introduction

Strings, Trees, and Sequences

Lenwood S. Heath
Computer Science
Virginia Tech
heath@cs.vt.edu

August 27, 2001

Overview

- Disclaimers and Observations
- Molecular Biology
- Evolution and Genetics
- Algorithms
- Applications
- Biology and Computer Science

Disclaimers and Observations

- I am not a biologist or chemist!
- To every general rule in biology there are exceptions.
- Biology is much broader than what computer scientists need to know to do good bioinformatics.
- Biologists speak a different language than computer scientists.
- Bioinformatics is about bringing biological and computational scientists together to accomplish some goal in a life science.

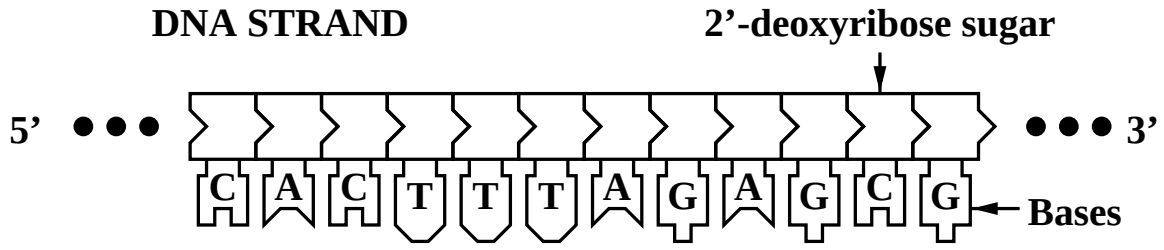
Molecular Biology

- Cell function
- Nucleic acids, DNA, RNA, chromosomes, genes
- Amino acids, proteins

The Cell's Fetch-Execute Cycle

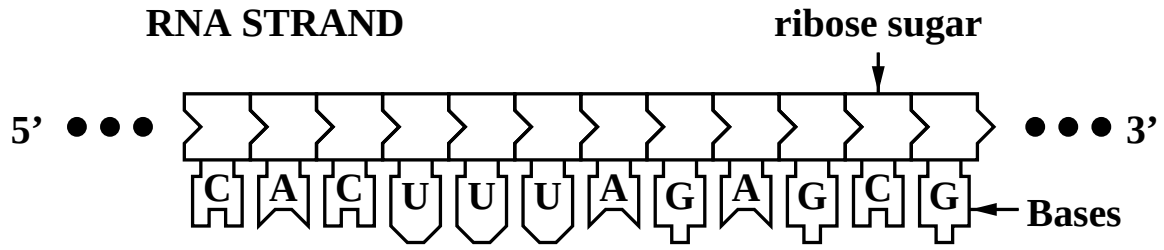
- **Stored Program:** DNA, chromosomes, genes
- **Fetch/Decode:** RNA, ribosomes
- **Execute Functions:** Proteins — oxygen transport, cell structures, enzymes
- **Inputs:** Nutrients, environmental signals, external proteins
- **Outputs:** Waste, response proteins, enzymes

Nucleic acids



A = **adenine** complements **T** = **thymine.**

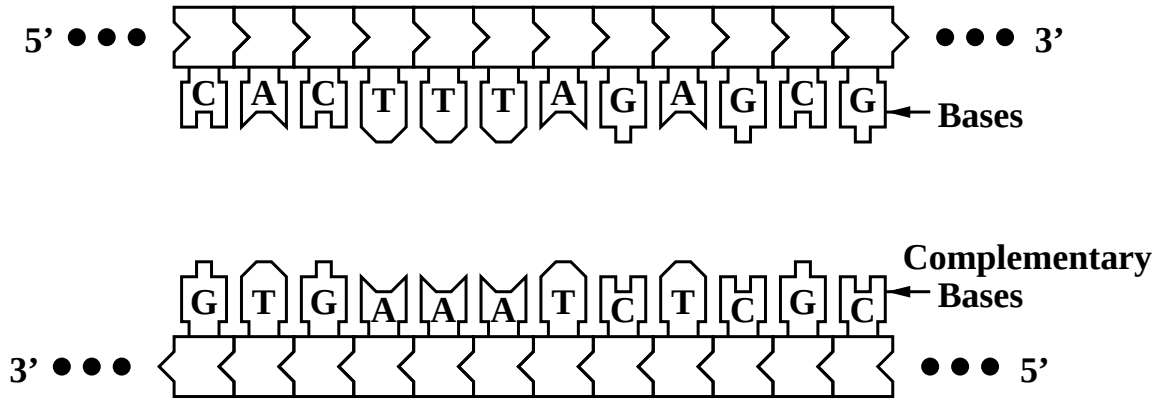
C = **cytosine** complements **G** = **guanine.**



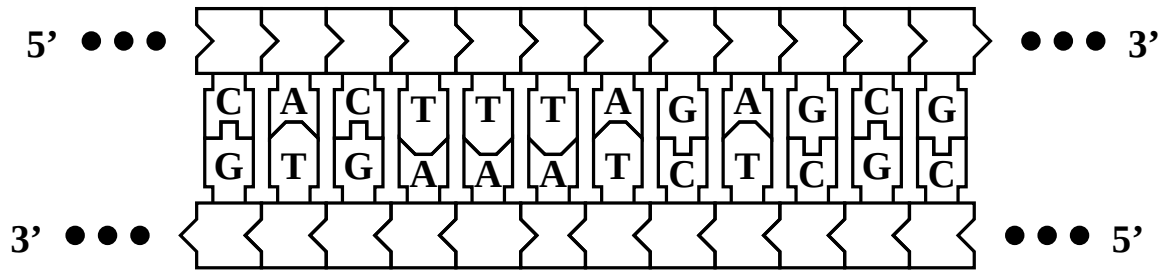
U = **uracil** replaces **T** = **thymine**.

DNA

DNA Strand and Its Complement

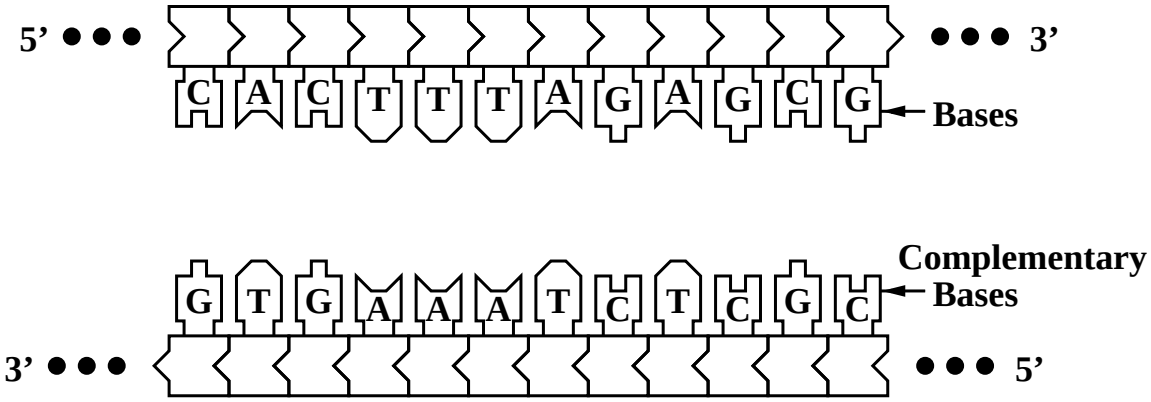


Double-Stranded DNA

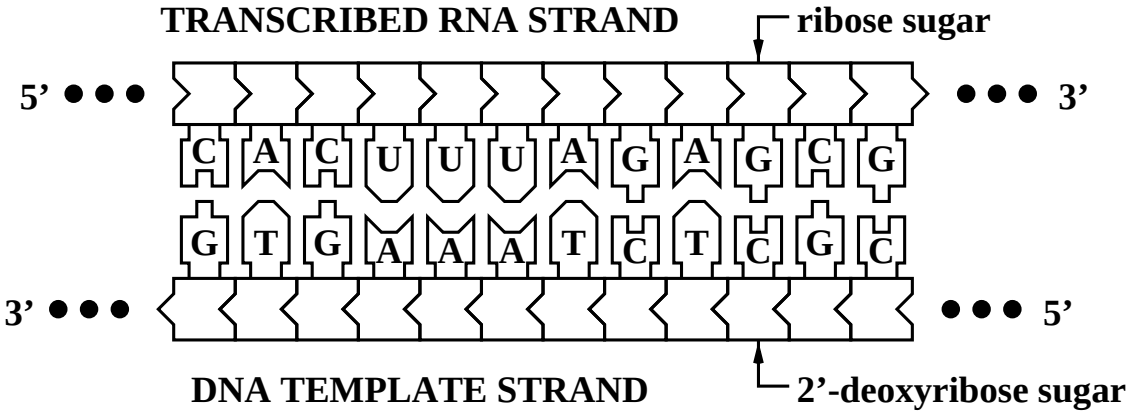


Transcription

DNA strand and its complement are unzipped



RNA is transcribed from the DNA template or noncoding strand



Chromosomes

- Long molecules of DNA — 10^4 to 10^8 base pairs
- In higher organisms, come in matched pairs — humans have 26 pairs.
- A **gene** is a subsequence of chromosome that encodes a protein.
- Remainder (up to 90%) of chromosome is **intergenic DNA.**
- While only a fraction of the genes are **expressed** in a cell at any time, **all genes are present in every cell.**

Amino Acids

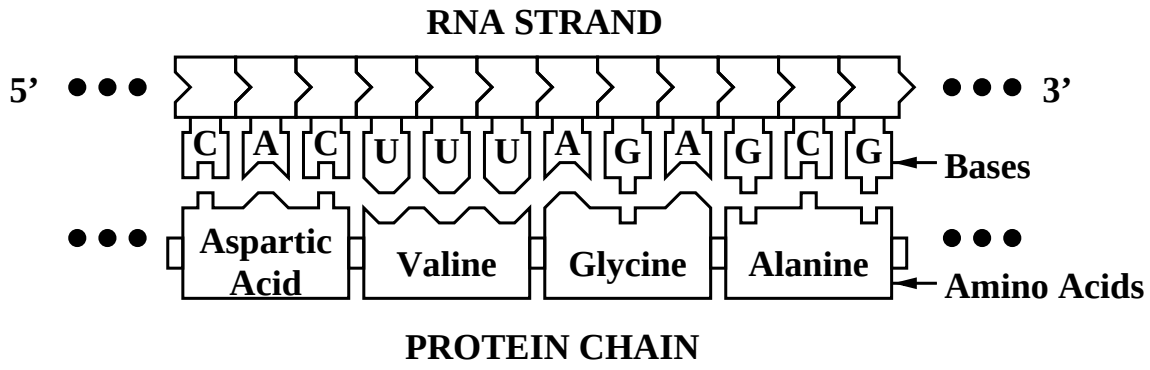
- **Protein** is a large molecule that is a chain of **amino acids** (100 to 5000).

- There are 20 common amino acids (Alanine, Cysteine, ..., Tyrosine)

- Three bases — called a **codon** — suffice to encode an amino acid. There are also START and STOP codons.

- The encoding of amino acids via 3-base codons is the **genetic code.**

Proteins



Unlike DNA, proteins have three dimensional structure.

- Controlled by genetic sequence, physics, chemistry.
- Protein folds in three dimensions and assumes a **configuration** (shape) that allows it to perform its **function.**

Evolution and Genetics

- **Genetics of a species:** the complement of genes shared by all members of a species
- **Genotype:** genetic information in an individual organism
- **Phenotype:** observable characteristics of an organism determined by its genotype and environment
- **Mutation:** change in the genotype of an organism that can be inherited by its organism
- **Evolution:** the accumulation of mutations over generations; new species can result

Mutations

- Insertion, deletion, change in bases
- Genes can change relative positions on chromosome by DNA recombination
- **Phylogeny:** Evolutionary tree of species derived from observations of mutations

Algorithms and Problems

A **problem** is a relationship between inputs — **instances** — and outputs — **solutions.**

EXAMPLES.

- Sorting
- Searching
- Convex hull of a set of points
- Shortest paths in graphs
- Minimum spanning tree (MST)
- Traveling salesman problem

Algorithms Solve Problems

An **algorithm** is a step-by-step procedure for **solving** a problem.

EXAMPLES.

- Bubble sort; heap sort; quick sort
- Sequential search; binary search
- Graham's algorithm (for convex hull)
- Dijkstra's algorithm (for shortest path)
- Kruskal's algorithm (for MST)

Pseudocode

Kruskal-$W(G, w)$
▷ Find a MST for $G = (V, E)$ with edge weights w,
▷ where $w : E \to \{1, 2, \ldots, W\}$.

for $i \leftarrow 1$ **to** W ▷ Initialize bucket array.
    do bucket[i] $\leftarrow$ **nil** ▷ Empty linked list

▷ Put edges in buckets by weight.
for $(u, v) \in E$
    do bucket[$w(u, v)$] $\leftarrow$ bucket[$w(u, v)$], (u, v)

$E' \leftarrow \emptyset$

▷ Process edges by increasing weight.
for $i \leftarrow 1$ **to** W
    do $E' \leftarrow E' \cup$ bucket[i]
        Do a modified DFS on $G' = (V, E')$,
            resulting in the spanning forest (V, E'_i)
        $E' \leftarrow E'_i$
return $T = (V, E')$ ▷ T is an MST.

Kruskal's algorithm for bounded-weight graphs

Resource Consumption

Algorithms consume resources.
What resources?

- **Time** — Why is it taking so long?
- **Memory** — RAM or disk utilization
- **Communications** — in a distributed system
- **Processors** — in a parallel computer

For hard problems, TIME usually dominates.

Computational Complexity

Instances come in all sizes. Let N be the size of a typical instance.

Computational complexity describes how execution time scales with N.

EXAMPLES.

- $O(\log N)$ — binary search
- $O(N)$ — sequential search
- $O(N \log N)$ — heap sort
- $O(N^2)$ — bubble sort
- $O(2^N)$ — examine all substrings of a string of length N
- $O(N!)$ — traveling salesman problem

Applications

A Sample of Computational Applications in Biology

- Sequence Comparison
- Fragment Assembly
- Database Searching
- Phylogenies
- Genome Rearrangement — John Paul Vergara
- Protein Folding

Sequence Comparison

Compare two DNA sequences and give a measure of similarity.

Example: ACATCGGAATAG CACGGAATACGG

A good alignment:

```
A C A T C G G A A T A - G -
  | |   | | | | | | |   |
- C A - C G G A A T A C G G
```

- Matches: +5 points
- Insertions/deletions: −3 points
- Substitutions: ?
- **Similarity score:** 38

Algorithms for Comparison

Dynamic Programming — Optimal

- Develop a table of optimal scores for longer and longer substrings
- Optimal solution pops out "at the end of the table"
- Generally polynomial time but **not linear** — $O(N)$ — time

Heuristics — Suboptimal — e.g., BLAST

Fragment Assembly

Aim: Sequence a DNA strand of thousands of bp's.

- Replicate a single DNA strand of thousands of bp's millions of times.
- Chop up into random fragments (substrands).
- Sequence each fragment.
- Compute how to best piece these fragments together to find original sequence.

Example

Use binary strings:

1 1 0 1 1 1 0 0

Random fragments:

1 1 0 1 0 1 1 1 0 0

Possible fragment assemblies:

1 1 0 1 1 1 0 0
1 1 0 1 0 0 1 1
1 0 1 1 0 0
0 1 1 1 0 0 1 1 0 1
others?

What if data contains errors?

Database Searching

Aim: Find a plausible function for a protein whose gene has been sequenced.

- Sequence a DNA strand and locate a gene within it.

- Use a web based search tool accessing a database of genes and their functions to find gene matches that are close enough. E.g., BLAST and PSI-BLAST.

- Use matches to guess or confirm gene function.

- Annotate and add your new gene to database.

Phylogenies

Aim: Use genetic information about a group of species to determine phylogenetic tree.

- Generally have to postulate missing ancestors at internal nodes of tree.
- Can use some specific traits from different species; or can estimate a genetic distance between species.
- Problem is generally NP-complete.

Genome Rearrangement

Given two gene sequences, how many mutations (movements) might have lead from one to the other?

Example:

$$Q \quad X \quad J \quad Z \quad B \quad M \quad Y \quad E$$

$$X \quad J \quad E \quad B \quad M \quad Y \quad Q \quad Z$$

If mutations are only movements of blocks of genes, then here is a plausible sequence of three mutations:

$$Q \quad X \quad J \quad \underline{Z \quad B \quad M \quad Y} \quad E$$

$$Q \quad Z \quad B \quad M \quad Y \quad \underline{X \quad J \quad E}$$

$$\underline{Q \quad Z} \quad X \quad J \quad E \quad B \quad M \quad Y$$

$$X \quad J \quad E \quad B \quad M \quad Y \quad Q \quad Z$$

Reversals

Assume that each mutation is a reversal of a gene sequence

Previous example:

$Q \quad X \quad J \quad Z \quad B \quad M \quad Y \quad E$

$X \quad J \quad E \quad B \quad M \quad Y \quad Q \quad Z$

Here is a sequence of four reversals:

$$\underline{Q \quad X \quad J \quad Z \quad B \quad M \quad Y \quad E}$$

$$\underline{Q \quad X \quad J \quad E \quad Y \quad M \quad B} \quad Z$$

$$\underline{B \quad M \quad Y \quad E \quad J \quad X} \quad Q \quad Z$$

$$X \quad J \quad E \quad \underline{Y \quad M \quad B} \quad Q \quad Z$$

$$X \quad J \quad E \quad B \quad M \quad Y \quad Q \quad Z$$

Sorting by Reversals

Really a restricted sorting problem

Recent dissertation:

> "Sorting by Bounded Permutations," John Paul C. Vergara, Department of Computer Science, Virginia Tech 1997.

John Paul modeled the problem using

- Permutations, and
- Graphs

and developed polynomial-time algorithms for special cases.

Protein Folding

Aim: Find the shape and function of a protein.

- Protein shape is complex and three-dimensional; it partially determines protein function.
- How does shape arise from the genetic sequence and the environment?
- Rich continuous/discrete problem.

What Computer Science Offers Biology

- Expertise in modeling problems with sequences, graphs, etc.
- Expertise in solving problems algorithmically
- Analysis of problems
  - Tractable or intractable
  - Efficiency
  - Randomness
  - Approximation
- Direction in implementation

Biology and Computer Science

- Discrete models for DNA, RNA, proteins, and phylogeny
- Efficient algorithms for tractable problems
- Approximation algorithms for intractable problems
- The challenge of large search problems