Basics of Molecular Biology

Tandy Warnow

December 29, 2016

Introduction to CS 466 Tandy Warnow

Introduction to CS 466

This course is about:

- Understanding the tools used in biological sequence analysis
- Understanding the mathematical models underlying these tools
- Designing better methods for biological sequence analysis

Fundamentally this requires computer science and statistics, but not too much biology!

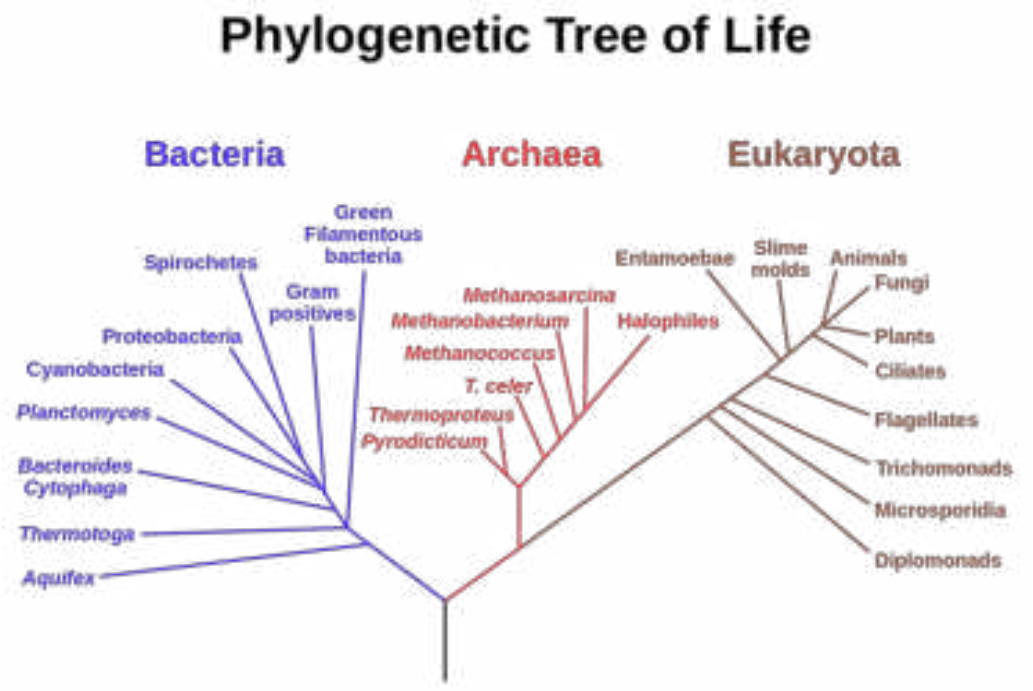
Biological topics we'll cover

- Genome assembly and search
- Multiple sequence alignment and applications
- Phylogenetics
- Protein sequence analysis
- Metagenomics
- Systems biology

A multiple sequence alignment

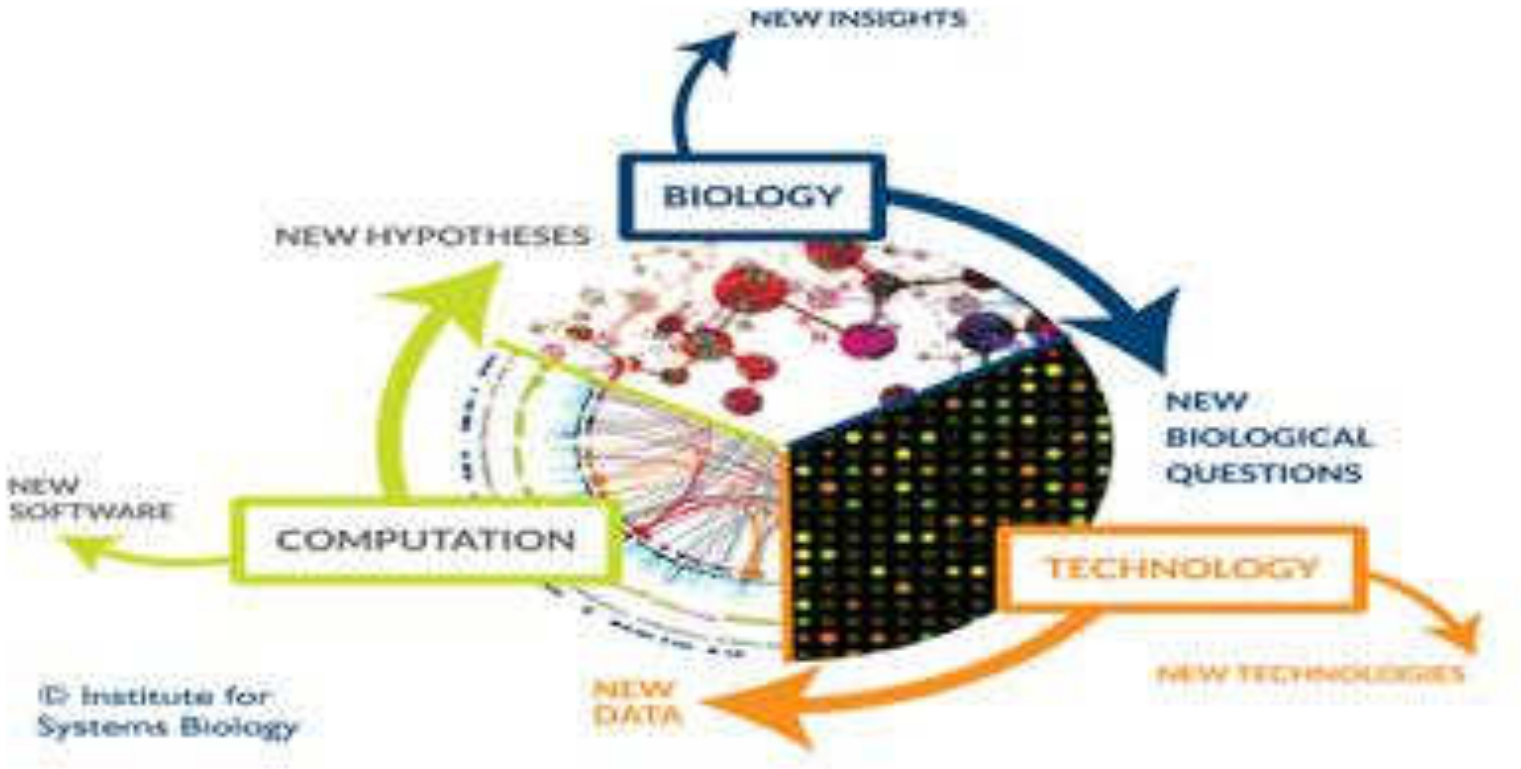
s_1	-	-	-	T	A	C
s_2	-	-	A	T	A	C
s_3	C	-	A	-	-	G
s_4	C	-	A	A	T	G
s_5	C	-	-	T	-	G
s_6	C	T	-	-	A	C
s_7	C	-	A	T	A	C
s_8	G	-	A	-	A	T

A phylogeny



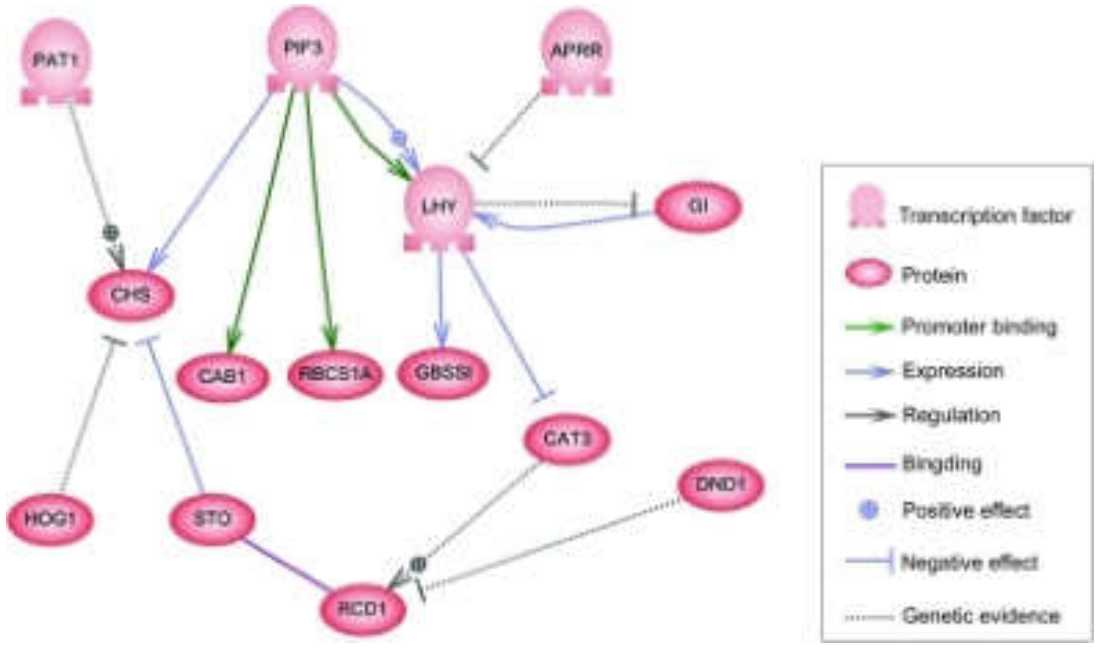
From https://en.wikipedia.org/wiki/File:Phylogenetic_tree.svg

Systems Biology



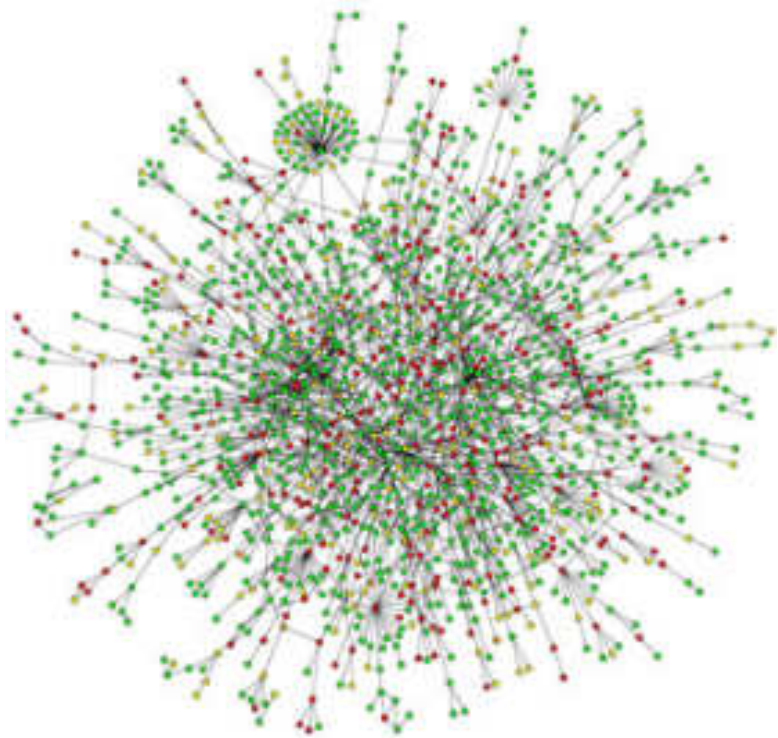
From http://www.assignmentpoint.com/science/biology/systems-biology.html

Regulatory Network



From http://en.wikipedia.org/wiki/Gene_regulatory_network

Protein-Protein Interaction Network



From https://ocw.mit.edu/courses/biology/7-343-network-medicine-using-systems-biology-and-signaling-networks-to-create-novel-cancer-therapeutics-fall-2012/

Steps in a phylogenomic analysis

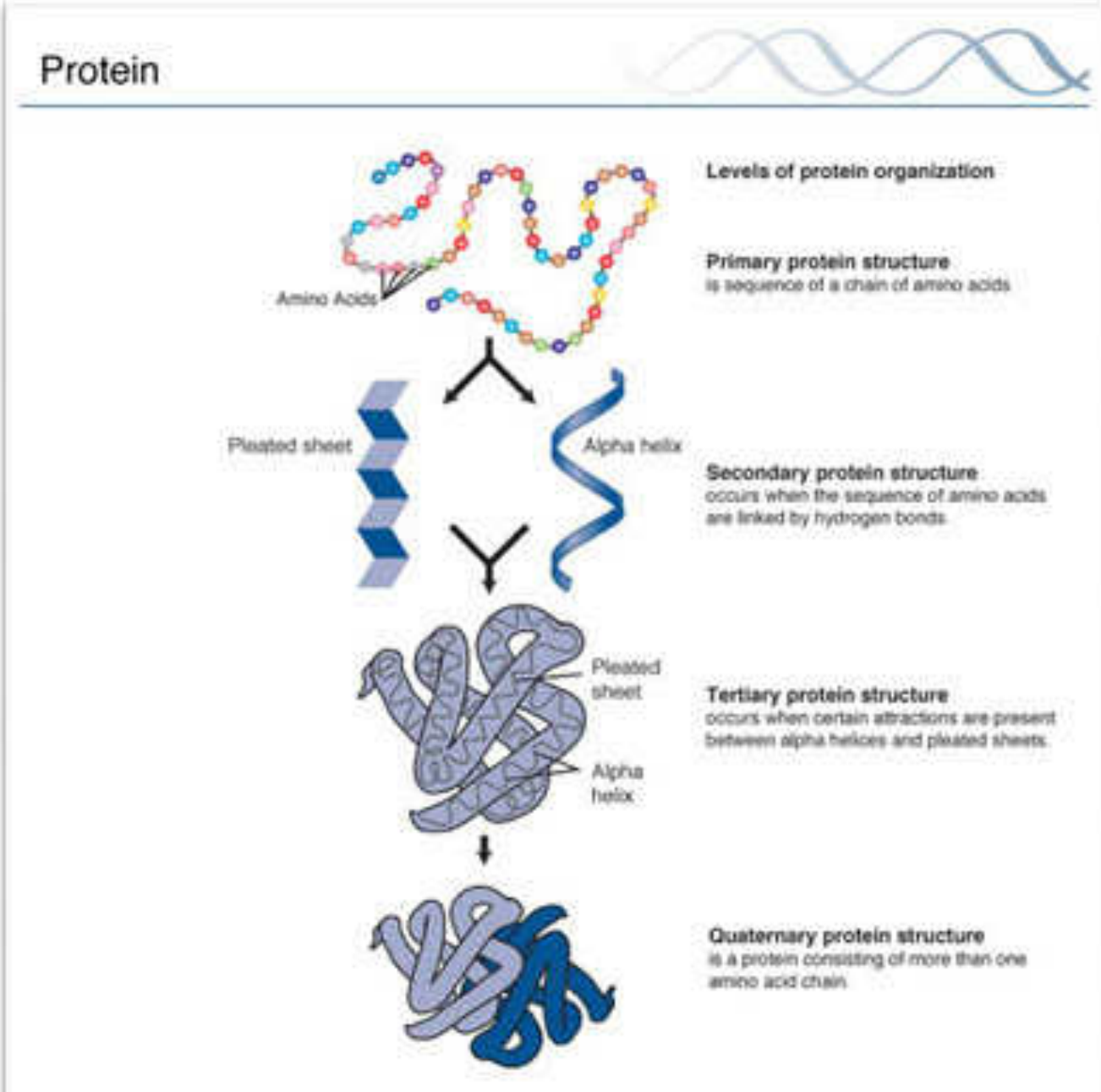
1. Decide which species and genes are needed
2. Collect specimens and obtain sequence data
3. Compute multiple sequence alignments and phylogenetic trees for each gene
4. Estimate the species tree[1]
5. Estimate branch support and dates
6. Answer biological questions

[1] or network

Proteins

- Proteins can be seen as strings of amino acids, but they are also fundamental to function.
- Proteins form structures that enable them to interact with other proteins and with DNA molecules. The inference of protein structure is closely related to the inference of protein function.

Protein structure



Inferring protein structure and/or function

A simple and common technique to infer the structure or function of a newly discovered protein sequence:

- Examine protein sequences in a public database (perhaps with known structure and/or function) and find the ones that are closest (in terms of its AA sequence), using methods like BLAST. These are the "homologous" sequences, and the set of homologous sequences forms a protein "family".
- Then compute a pairwise sequence alignment for the new sequence and the closest homolog, and transfer the information.

Inferring protein structure and/or function

Or you could try one of these:

- Better: Compute a multiple sequence alignment and tree for the new sequence and its homologs, and see where the new sequence is located in the tree. Infer function and structure at the internal nodes of the tree, and transfer the information to the new sequence.
- Even better (if available): Use a **profile Hidden Markov Model** or other statistical model for the family that is annotated with structural features, and align the new sequence to the model. Transfer structural information.

The same could be done with RNAs, which also form structures and have functions.

Repeating themes

- Genome sequencing data
- Multiple sequence alignment
- Phylogenetic tree
- Protein and RNA structure and function
- BLAST (or other database search tools)
- profile Hidden Markov Models (HMMs)

This course

This course is about:

- designing better algorithms – ones that are more accurate and are scalable to large datasets.
- proving theorems about methods under statistical models (especially for phylogenies)
- dataset analysis

You don't need to know any biology for this; you'll learn as you go. Most of the work is really a combination of computer science and statistics.

Basic vocabulary, page 1

- Nucleotide sequences (RNA and DNA): think of them as strings over a four-letter alphabet (ACTG for DNA, ACUG for RNA)
- Protein sequences: composed of amino acids, of which there are 20
- Coding sequences: only some nucleotide sequences are used to create proteins. Those sequences are called **exons**.
- Codons: three nucleotides in a row, that are used to create amino acids. Every codon makes a single amino acid. Hence, this is a many-to-one mapping, called the **Genetic code**.
- Gene: used to mean something specific, and it's no longer clear what it means. However, it's safe to think of this as a collection of regions of a genome that may have some function.

Basic vocabulary, page 2

- Genome: the collection of chromosomes that carry genetic information (in the form of DNA), which is inherited.
- Alleles: in diploid organisms, we have two copies of each chromosome, and so two copies of the gene at a given locus (position) in the genome; each of these copies is called an allele. The collection of alleles is the **genotype**.
- Transcription: the action of changing a DNA sequence into messenger RNA.
- Translation: the action of changing a string of messenger RNA into a string of amino acids. (Think of this as DNA makes RNA, and RNA makes proteins...)
- Exons: the portion of the gene that remains after the **introns** are removed during transcription.

Basic vocabular, page 3

- Mutations: substitutions of nucleotides by other nucleotides in the genome.
- Synonymous mutations (or silent mutations): Mutations that occur in coding regions (i.e., exons) and that do not change the resultant amino acids, due to the many-to-one mapping of the **genetic code**.
- Non-synonymous mutations: mutations that do change the resultant amino acid.
- Natural selection: preferential survival and reproduction or preferential elimination of individuals with certain genotypes
- Neutral evolution: evolutionary changes that are not impacted by selection

Basic vocabulary, page 4

- Phylogeny: an evolutionary tree or possibly evolutionary network
- Gene tree: a phylogeny based on a single gene
- Species tree: a phylogeny that represents how species evolved
- Hybridization: when two different species mate and have offspring, called hybrids
- Horizontal gene transfer: when genetic material is transferred from one organism into another
- Homology: related by descent from a common ancestor
- Species: too complicated to answer
- Genus, Family, Order, Class, Phylum, Kingdom, Domain - the hierarchies in a taxonomy

Basic vocabulary, page 5

- Exome: the part of the genome that is comprised of exons
- Transcriptome: all the messenger RNA that is expressed by the genes in an organism
- Genome assembly: putting together a genome from lots of short DNA strings, generated by a sequencing project
- Reads: the short DNA strings produced by the sequencing technology
- Contigs: somewhat longer DNA strings produced by combining reads