Bioinformatics

Lecture 1

Introduction 16/1/2021

Indian Institute of information technology

Course outline

Refer to the pdf document posted online

- 13 lectures
- 2h sessions

Contact details: SAILATHA RAVI, PhD
Sailatha.r@iiits.in

Course policy and requirement

- Attendance is compulsory
- You must **write up assignment solutions on your own.**
- Late assignments will lose 10% per day, up to 5 days, after which they will not be accepted

Goals of this Course

- Introduce some biological terminology – Present the algorithms pursued by natural process
- Present some areas of bionformatics
- Provide an overview
- Show that there are interesting algorithmic & computational problems
- Provide you the knowledge you need to work on research projects

Requirements

Be curious

Ask questions

Discussion among peers

Some basic knowledge on statistics and

algorithms could help

Text Books

Bioinformatics- Sequence and Genome analysis, David. W. Mount

An Introduction to Bioinformatics Algorithms, Neil C. Jones and Pavel A. Pevzner

Additional reading materials and work sheets will be provided.

Course Structure

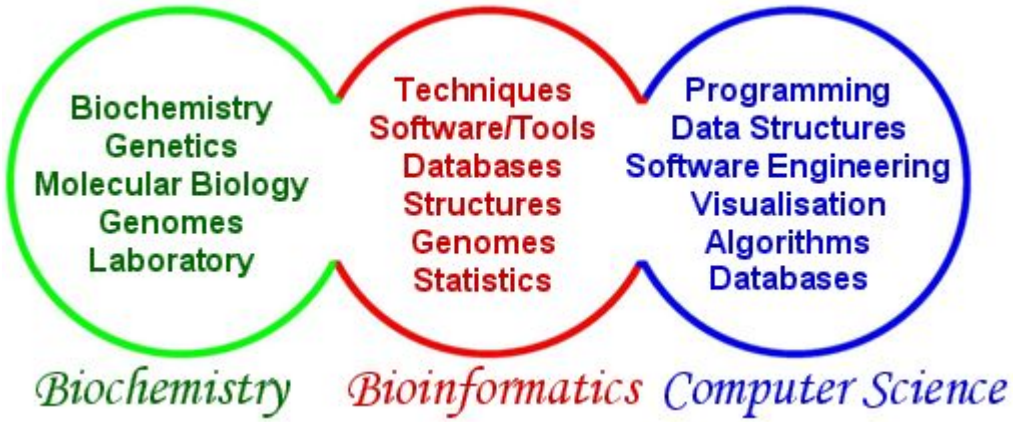
- Introduction & biological Terminology (2 lectures)
  - Genome
  - Central dogma, DNA protein and RNA
- Genetic code and protein code- Database search [1/2 lectures]- Disease predictions, SNPs
- Sequence Analysis (3 lectures)
  - Pair-wise Sequence alignment
  - Searches on strings
  - Multiple Sequence Alignment
- Phylogenetics (2 lectures)
  - Parsimony
  - Likelihood
  - Discrete operations on trees
- Structure prediction [2 lectures]
  - RNA
  - DNA
  - Protein folding problem
- Special topics and current trend in Bioinformatics [1-2 lectures]
  - Cancer, Pharmacogenomics [1 lecture],
  - Population Genetics (1 lecture), Interactome
  - Class Group assignment presentation [1 lecture]
- Course revision (1 lecture)

Bioinformatics

- Original definition (1979 by Paulien Hogeweg):

  "application of information technology and computer science to the field of molecular biology"
- Using information technology, designing novel algorithms and methods of analyses *(computational biology)*
- Establishing innovative software and databases of information, allowing open access to the records held within them *(bioinformatics)*

Bioinformatics- Prospective future



What does biology have to do with computers?

- Huge amount of data: too much to analyze by hand, lots of mystery left about how life works.
- Requires clever algorithms to:
  - find interesting patterns • store / search / compare • visualize vast collections of data • predict missing or hard-to-observe features (like protein structure or evolutionary relationships)
- Nearly all molecular biology is now “computational biology”: biologists depend on computer scientists every day.

Algorithms

Algorithms- A set of specific instructions to define and solve a problem

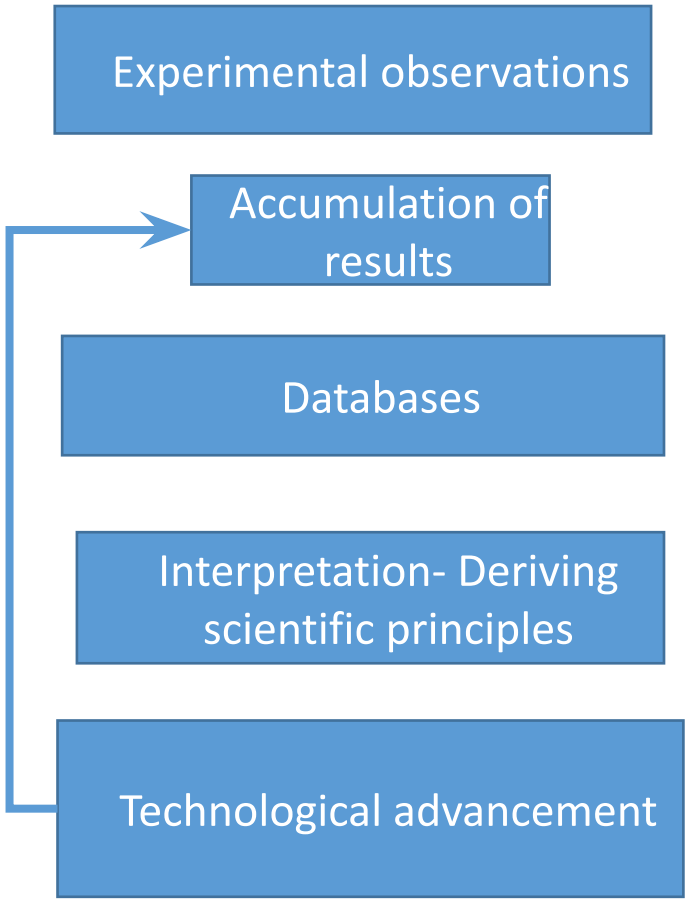
Programming- Using programming languages to code/implement the specific set of instructions

“The benefit of using a computer is not to solve an insolvable problem but to arrive at a solution more quickly and accurately than a human can”.

What is Bioinformatics?

- *The use of techniques from mathematics, statistics, and computer science to solve biological problems*
- Many activities of the cell can be interpreted as manipulation of strings from a small alphabet
- Things we will **not** be studying
  - How to use cells to perform computation
  - How the cells perform the computation
- Instead, we will be studying computations that can help us
  - Find genes that are similar - Pattern matching
  - Finding Transcription Factors – Motif discovery
  - Aligning Genes from different organisms – Multiple Alignment
  - Retracing evolutionary history - Phylogenetic trees
  - How reactions are facilitated - Protein folding

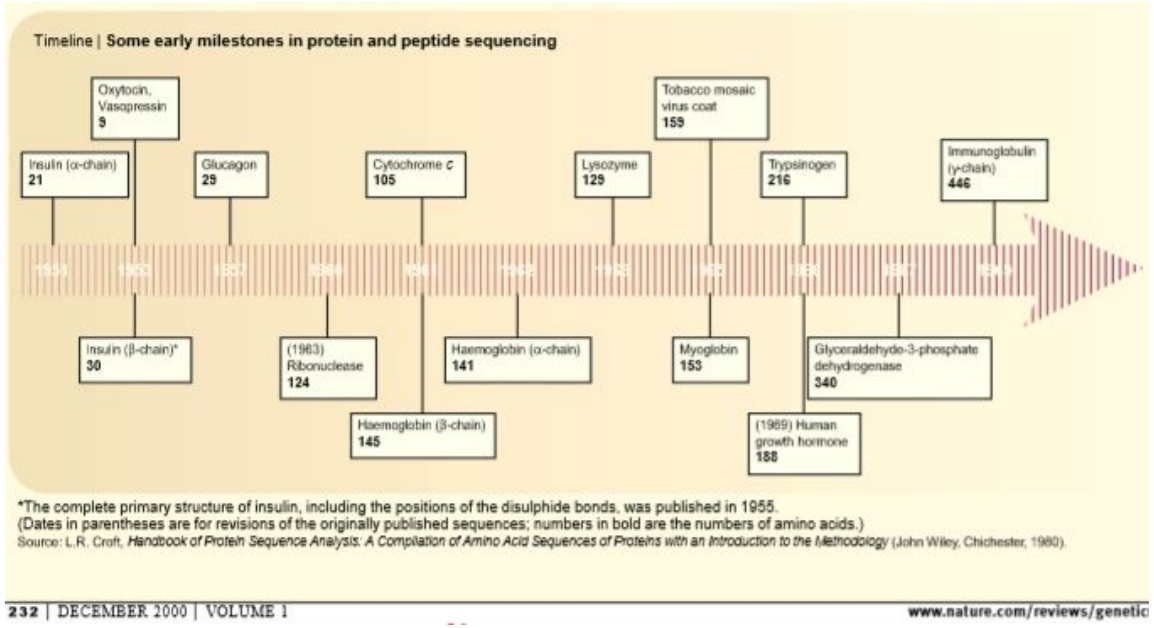
Workflow



- Identifying patterns or mismatches with speed and accuracy
- Building models
- 3D representations
- Predicting missing or hard to observe features that aligns with known scientific principles

History of Bioinformatics

- Computers role in molecular biology recognized as early as 1960s- Data from Protein chemistry
- The scientists contributed to the conceptual and technical foundation
-



1955 By Sanger et al., at Cambridge University- 1958 Nobel prize in Chemistry
Semi-automated machine by Stanford Moore and William Stein- reduced the time to half.
Late 1960s- Automated ‘Sequenator’ by Edman

History of Bioinformatics

Computer technology + Mathematics + Molecular Biology-
 Life sciences
    1) amino acid sequences
    2) Macromolecular information- conceptual link between molecular biology and computer science led to Information theory
    3) Availability of high speed digital computers developed during the second world war became available

Margaret Dayoff- used FORTRAN to deduce the protein sequence using overlapping peptides

Figure 4 | **The IBM 7090 computer, which Margaret Dayhoff used for her early work.** This famous computer was one of the first solid-state machines and was used widely in business and defence settings, as well as scientific applications.
(Photograph courtesy of IBM archives.)

good-natured, she thought: still
when it saw Alice. It looked
ought to be treated
good-natured, she thought, still
Cat only
a greet many
It looked good-
The Cat only grinned when it saw Alice.
be treated with respect.
still it had very long claws
claws and a great many teeth, so she
so she felt that it ought

```
The Cat only grinned when it saw Alice.
    Cat only         when it saw Alice. It looked
                                        It looked good-


good-natured, she thought: still
good-natured, she thought, still
                           still it had very long claws


claws and a great many teeth, so she
          a greet many        so she felt that it ought


ought to be treated
         be treated with respect.
```

Genome

- Genome of the Cow
- a sequence of 2.86 billion letters enough letters to fill a million

```
TATGGAGCCAGGTGCCTGGGGCAACAAGACTGTGGTCACTGAATTCATCCTTCTTGGTCTAACAGAGAACATAG
AACTGCAATCCATCCTTTTTGCCATCTTCCTCTTTGCCTATGTGATCACAGTCGGGGGCAACTTGAGTATCCTG
GCCGCCATCTTTGTGGAGCCCAAACTCCACACCCCCATGTACTACTTCCTGGGGAACCTTTCTCTGCTGGACAT
TGGGTGCATCACTGTCACCATTCCTCCCATGCTGGCCTGTCTCCTGACCCACCAATGCCGGGTTCCCTATGCAG
CCTGCATCTCACAGCTCTTCTTTTTCCACCTCCTGGCTGGAGTGGACTGTCACCTCCTGACAGCCATGGCCTAC
GACCGCTACCTGGCCATTTGCCAGCCCCTCACCTATAGCATCCGCATGAGCCGTGACGTCCAGGGAGCCCTGGT
GGCCGTCTGCTGCTCCATCTCCTTCATCAATGCTCTGACCCACACAGTGGCTGTGTCTGTGCTGGACTTCTGCG
GCCCTAACGTGGTCAACCACTTCTACTGTGACCTCCCGCCCCTTTTCCAGCTCTCCTGCTCCAGCATCCACCTC
AACGGGCAGCTACTTTTCGTGGGGGCCACCTTCATGGGGGTGGTCCCCATGGTCTTCATCTCGGTATCCTATGC
CCACGTGGCAGCCGCAGTCCTGCGGATCCGCTCGGCAGAGGGCAGGAAGAAAGCCTTCTCCACGTGTGGCTCCC
ACCTCACCGTGGTCTGCATCTTTTATGGAACCGGCTTCTTCAGCTACATGCGCCTGGGCTCCGTGTCCGCCTCA
GACAAGGACAAGGGCATTGGCATCCTCAACACTGTCATCAGCCCCATGCTGAACCCACTCATCTACAGCCTCCG
GAACCCTGATGTGCAGGGCGCCCTGAAGAGGTTGCTGACAGGGAAGCGGCCCCCGGAGTG ...
```

What does it mean?
Identify patterns.

Computational Successes

1. Genome assembly
2. Gene discovery
3. Understanding the origin of swine flu
4. Predicting protein structure

https://fold.it/portal/info/about

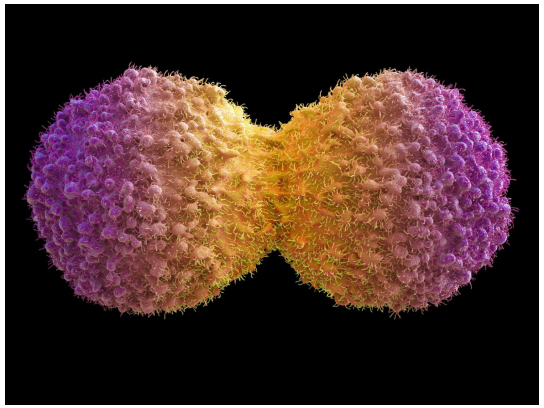
First genome sequenced in 1995 (the bacteria H. influenzae with a genome of 1,830,140 letters). • 1st draft of human genome finished in 2001 (~ 3 billion letters)

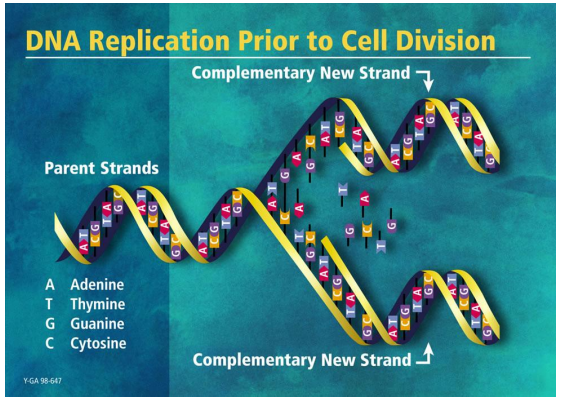
- Now: Over 1100 bacterial genomes
- Hundreds of higher-order genomes done or in progress.
- Several complete human genomes finished.

Topics of interest

Cancer	https://www.youtube.com/watch?v=0PX5UBuvk_w
Genetic Disorders	https://www.youtube.com/watch?v=hOfRN0KihOU
Evolution	https://www.youtube.com/watch?v=wvTv8TqWC48
Protein Engineering and Drug Interactions	

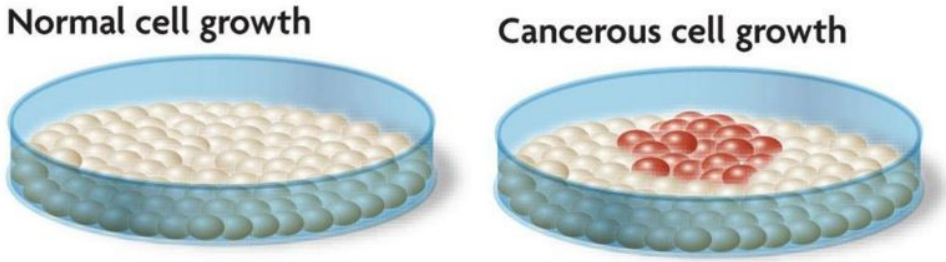
Cancer cells



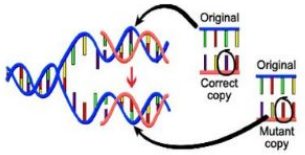


1. Checkpoint
2. Chemical signals tell it when to stop replicating
3. Space constraints

Apoptosis
Cell death

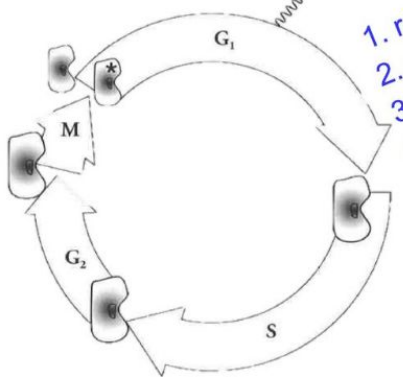


Mutations cause a disruption in the cell cycle.

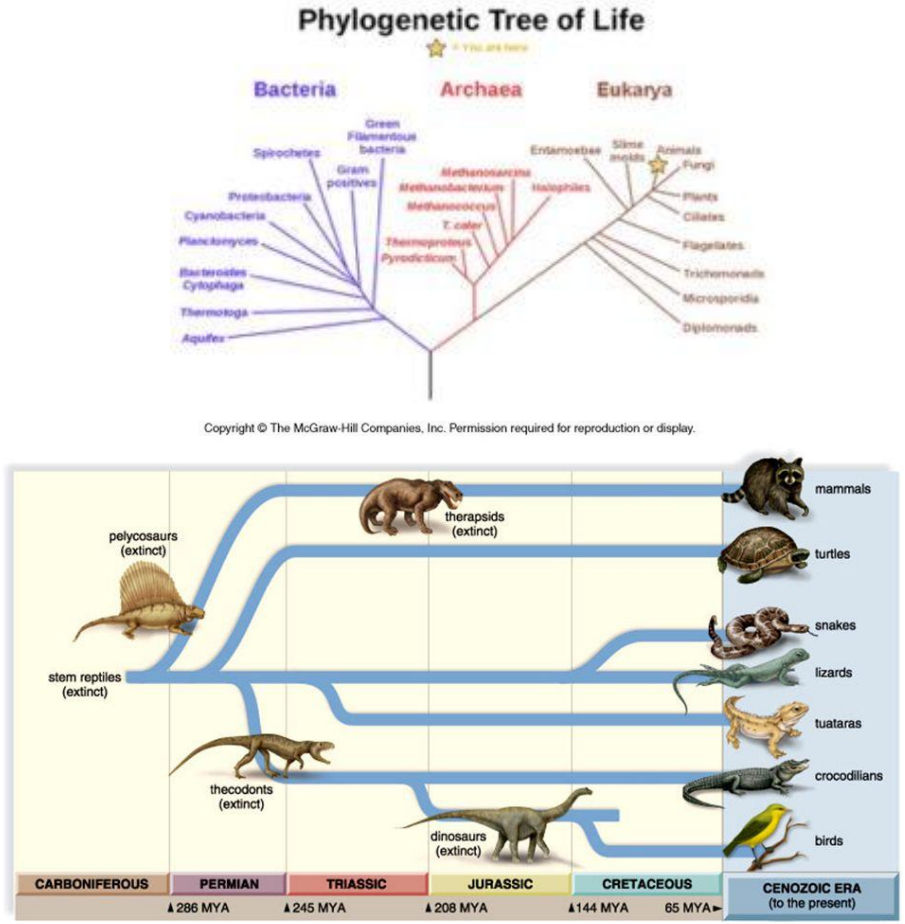


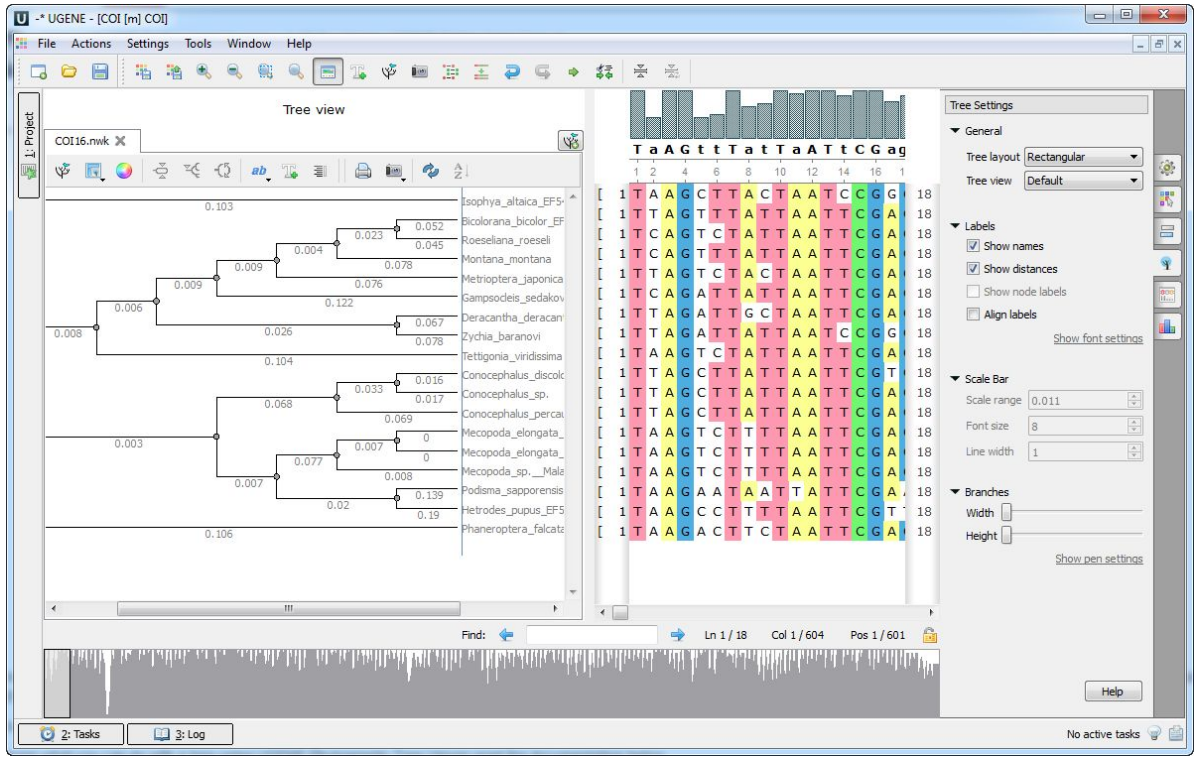
Mutations may be caused by:

1. radiation (UV rays)
2. smoking
3. Pollutants
4. chemicals - cleaning
5. viruses
6. genetic

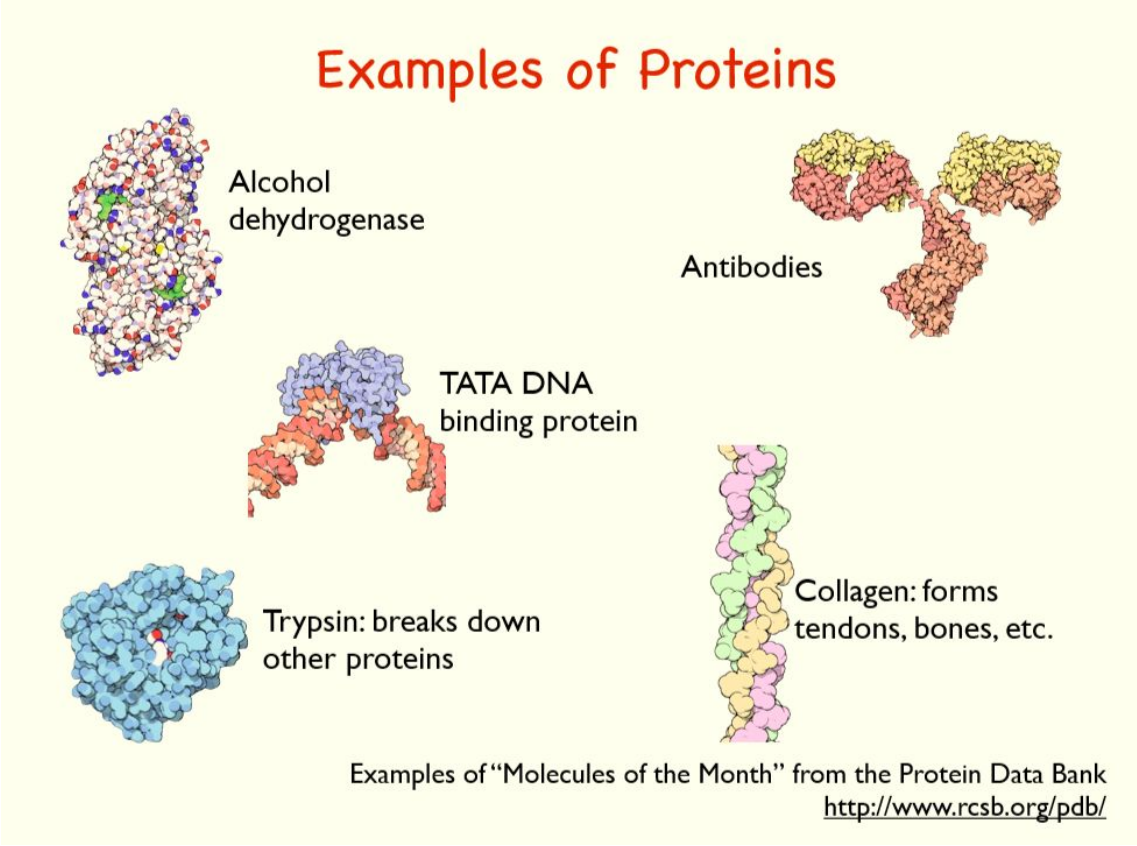


Phylogenetic tree





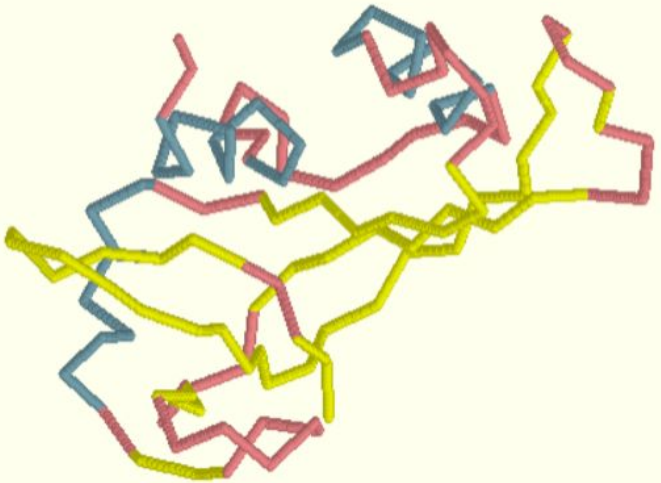
Protein structure prediction



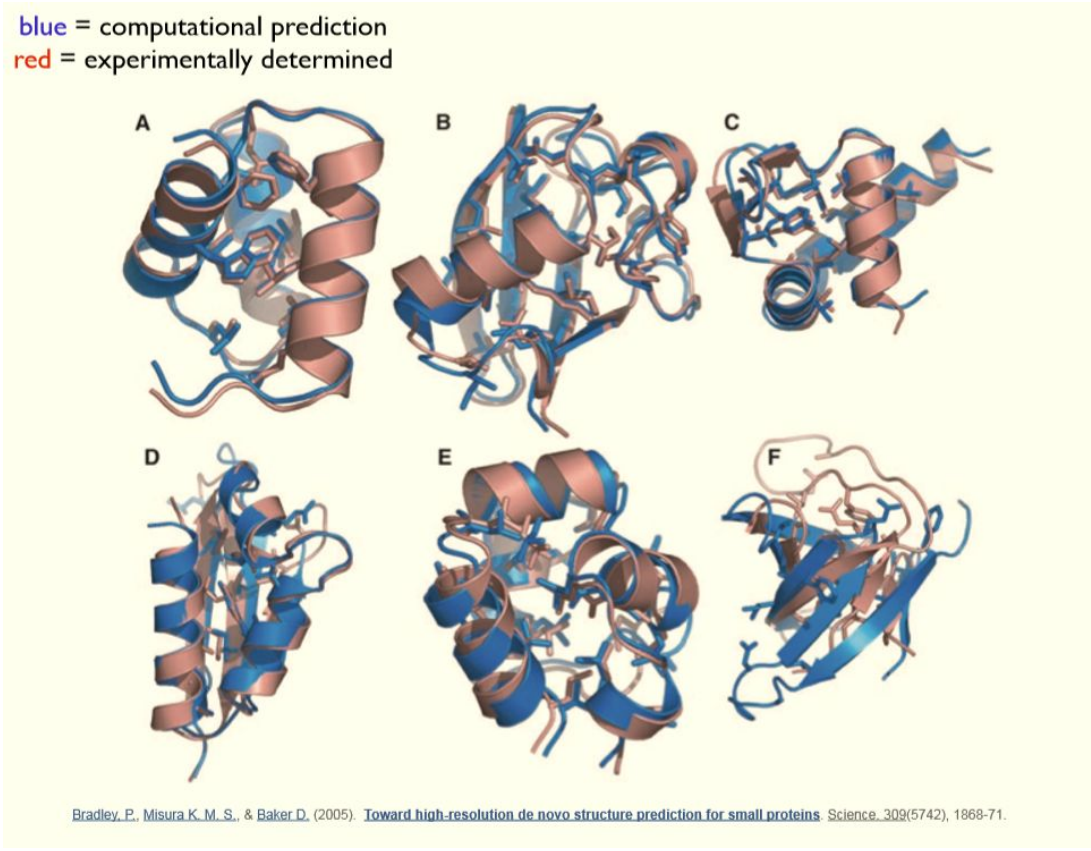
Predicting Structure

Given: ATGAAGATGATAGATGGGGCCCGACAG…

Determine:



Protein folding



Protein folding online platforms

Rosetta@home

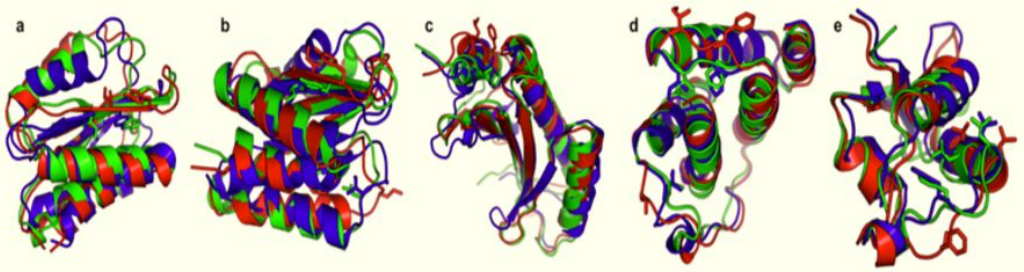
Protein Folding, Design, and Docking

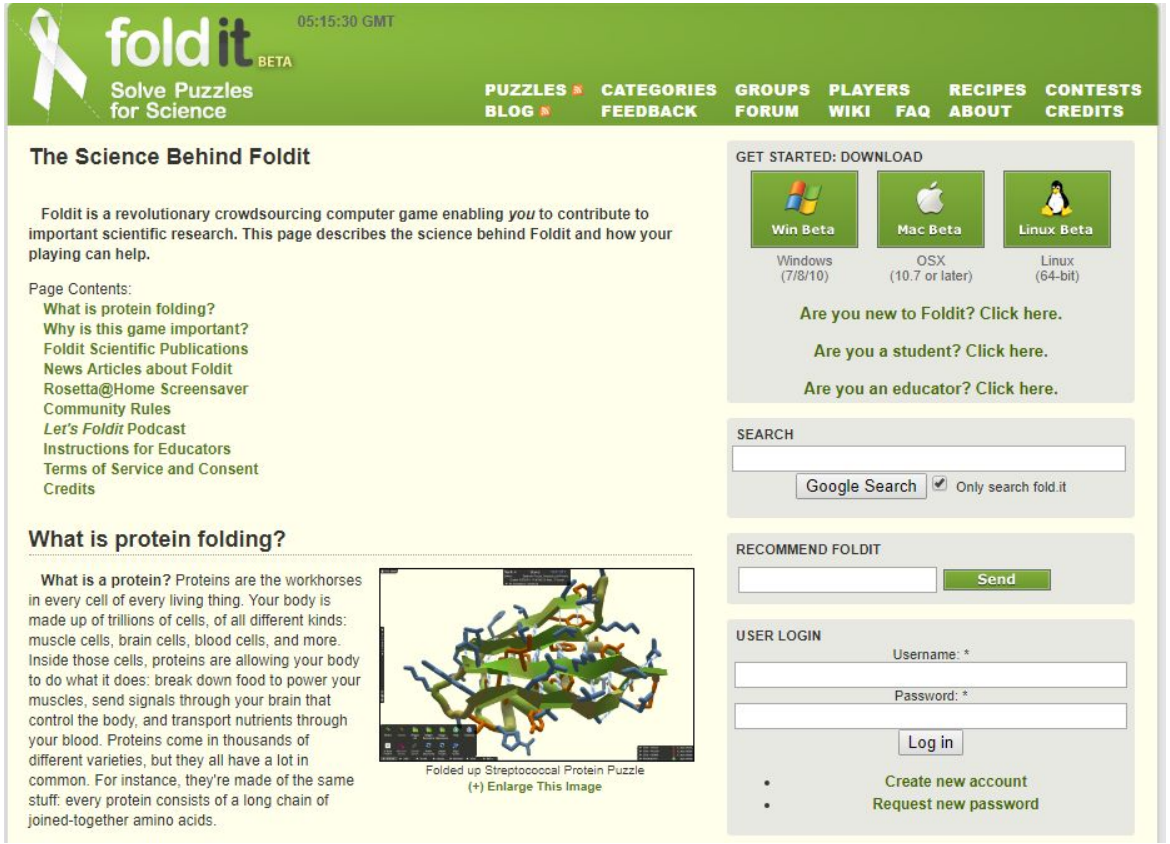
http://boinc.bakerlab.org/rosetta/

fold it BETA
Solve Puzzles for Science

http://fold.it/portal/

Cooper, Khatib et al, Nature, 466, 756-60. (2010)





https://fold.it/portal/info/about