Bioinformatics

Robert Holland
Jon Reckner
Jason Shields

What Is Bioinformatics?

- Bioinformatics is the unified discipline formed from the combination of biology, computer science, and information technology.
- "The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information.“ –Frank Tekaia

A Molecular Alphabet

- Most large biological molecules are *polymers*, ordered chains of simple molecules called *monomers*
- All monomers belong to the same general class, but there are several types with distinct and well-defined characteristics
- Many monomers can be joined to form a single, large *macromolecule*; the ordering of monomers in the macromolecule encodes information, just like the letters of an alphabet

Related Fields: Computational Biology

- The study and application of computing methods for classical biology
- Primarily concerned with evolutionary, population and theoretical biology, rather than the cellular or molecular level

Related Fields: Medical Informatics

- The study and application of computing methods to improve communication, understanding, and management of medical data
- Generally concerned with how the data is manipulated rather than the data itself

Related Fields: Cheminformatics

- The study and application of computing methods, along with chemical and biological technology, for drug design and development

Related Fields: Genomics

- Analysis and comparison of the entire *genome* of a single species or of multiple species
- A genome is the set of all genes possessed by an organism
- Genomics existed before any genomes were completely sequenced, but in a very primitive state

Related Fields: Proteomics

- Study of how the genome is expressed in proteins, and of how these proteins function and interact
- Concerned with the actual states of specific cells, rather than the potential states described by the genome

Related Fields: Pharmacogenomics

- The application of genomic methods to identify drug targets
- For example, searching entire genomes for potential drug receptors, or by studying gene expression patterns in tumors

Related Fields: Pharmacogenetics

- The use of genomic methods to determine what causes variations in individual response to drug treatments
- The goal is to identify drugs that may be only be effective for subsets of patients, or to tailor drugs for specific individuals or groups

History of Bioinformatics

- Genetics
- Computers and Computer Science
- Bioinformatics

History of Genetics

- Gregor Mendel
- Chromosomes
- DNA

Gregor Mendel (1822-1884)

- Credited with the theories of Heredity
- Developed his theories through the study of pea pods.
- Studied them “for the fun of the thing”

Mendel's Experiments

- Cross-bred two different types of pea seads
  - Sperical
  - Wrinkled
- After the 2^{nd} generation of pea seeds were cross-bred, Mendel noticed that, although all of the 2^{nd} generation seeds were spherical,   about $1/4^{th}$ of the 3^{rd} generation seeds were wrinkled.

Mendel’s Experiments (cont.)

- Through this, Mendel developed the concept of “discrete units of inheritance,” and that each individual pea plant had two versions, or alleles, of a trait determining gene.
- This concept was later fully developed into the concept of chromosomes

History of Chromosomes

- Walter Flemming
- August Weissman
- Theodor Boveri
- Walter S. Sutton
- Thomas Hunt Morgan

Walther Flemming (1843-1905)

- Studied the cells of salamanders and developing improved fixing and staining methods
- Developed the concept of mitosis cell reproduction (1882).

August Weismann (1834-1914)

- Studied plant and animal germ cells
- distinguished between body cells and germ cells and proposed the theory of the continuity of germ plasm from generation to generation (1885)
- Developed the concept of meiosis

Theodor Boveri (1862-1915)

- Studied the eggs of exotic animals
- Used a light microscope to examine chromosomes more closely
- Established individuality and continuity in chromosomes
- Flemming, Boveri, and Weismann together are given credit for the discovery of chromosomes although they did not work together.

Walter S. Sutton (1877-1916)

- Also studied germ cells specifically those of the *Brachystola magna* (grasshopper)
- Discovered that chromosomes carried the cell’s unit’s of inheritance

Thomas Hunt Morgan (1866-1945)

- Born in Lexington, KY
- Studied the *Drosophilae* fruit fly to determine whether heredity determined Darwinist evolution
- Found that genes could be mapped in order along the length of a chromosome

History of DNA

- Griffith
- Avery, MacLeod, and McCarty
- Hershey and Chase
- Watson and Crick

Frederick Griffith

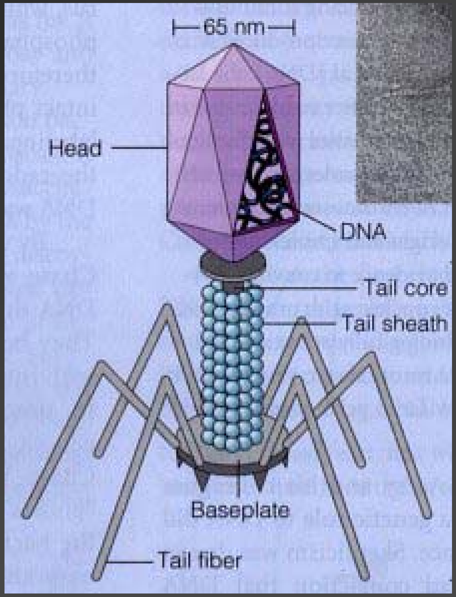
- British microbiologist
- In 1928, Studied the effects of bacteria on mice
  - Determined that some kind of “transforming factor” existed in the heredity of cells

Colin MacLeod
Maclyn McCarty

- 1944 - Through their work in bacteria, showed that Deoxyribonucleic Acid (DNA) was the agent responsible for transferring genetic information
  - Previously thought to be a protein

Alfred Hershey (1908-1997) Martha Chase (1930-)

- 1952 - Studied the bacteriophage T2 and its host bacterium, *Escherichia coli*
- Found that DNA actually is the genetic material that is transferred


65 nm
Head
DNA
Tail core
Tail sheath
Baseplate
Tail fiber

James Watson (1928-)
Francis Crick (1916-)

- 1951 – Collaborated to gather all available data about DNA in order to determine its structure
- 1953 Developed
  - The double helix model for DNA structure
  - The AT-CG strands that the helix is consisted of

"The structure was too pretty not to be true."

-- JAMES D. WATSON

History of Computers

Computer Timeline

- ~1000BC The abacus
- 1621 The slide rule invented
- 1625 Wilhelm Schickard's mechanical calculator
- 1822 Charles Babbage's Difference Engine
- 1926 First patent for a semiconductor transistor
- 1937 Alan Turing invents the Turing Machine
- 1939 Atanasoff-Berry Computer created at Iowa State
  - the world's first electronic digital computer
- 1939 to 1944 Howard Aiken's Harvard Mark I (the IBM ASCC)
- 1940 Konrad Zuse -Z2 uses telephone relays instead of mechanical logical circuits
- 1943 Collossus - British vacuum tube computer
- 1944 Grace Hopper, Mark I Programmer (Harvard Mark I)
- 1945 First Computer "Bug", Vannevar Bush "As we may think"

Computer Timeline (cont.)

- 1948 to 1951 The first commercial computer – UNIVAC
- 1952 G.W.A. Dummer conceives integrated circuits
- 1954 FORTRAN language developed by John Backus (IBM)
- 1955 First disk storage (IBM)
- 1958 First integrated circuit
- 1963 Mouse invented by Douglas Englebart
- 1963 BASIC (standing for **B**eginner's **A**ll Purpose **S**ymbolic **I**nstruction **C**ode) was written (invented) at Dartmouth College, by mathematicians John George Kemeny and Tom Kurtzas as a teaching tool for undergraduates
- 1969 UNIX OS developed by Kenneth Thompson
- 1970 First static and dynamic RAMs
- 1971 First microprocessor: the 4004
- 1972 C language created by Dennis Ritchie
- 1975 Microsoft founded by Bill Gates and Paul Allen
- 1976 Apple I and Apple II microcomputers released
- 1981 First IBM PC with DOS
- 1985 Microsoft Windows introduced
- 1985 C++ language introduced
- 1992 Pentium processor
- 1993 First PDA
- 1994 JAVA introduced by James Gosling
- 1994 Csharp language introduced

Putting it all Together

- Bioinformatics is basically where the findings in genetics and the advancement in technology meet in that computers can be helpful to the advancement of genetics.
- Depending on the definition of Bioinformatics used, or the source , it can be anywhere between 13 to 40 years old
  - Bioinformatics like studies were being performed in the ’60s long before it was given a name
    - Sometimes called “molecular evolution”
  - The term Bioinformatics was first published in 1991

Genomics

- Classic Genomics
- Post Genomic era
  - Comparative Genomics
  - Functional Genomics
  - Structural Genomics

What is Genomics?

- Genome
  - complete set of genetic instructions for making an organism
- Genomics
  - any attempt to analyze or compare the entire genetic complement of a species
  - Early genomics was mostly recording genome sequences

History of Genomics

- 1980
  - First complete genome sequence for an organism is published
    - FX174 - 5,386 base pairs coding nine proteins.
    - ~5Kb
- 1995
  - *Haemophilus influenzea* genome sequenced (flu bacteria, 1.8 Mb)
- 1996
  - *Saccharomyces cerevisiae* (baker's yeast, 12.1 Mb)
- 1997
  - *E. coli* (4.7 Mbp)
- 2000
  - *Pseudomonas aeruginosa* (6.3 Mbp)
  - A. thaliana genome (100 Mb)
  - D. melanogaster genome (180Mb)

2001 The Big One

- The Human Genome sequence is published
  - 3 Gb
  - And the peasants rejoice!

What next?

- Post Genomic era
  - Comparative Genomics
  - Functional Genomics
  - Structural Genomics

Comparative Genomics

- the management and analysis of the millions of data points that result from Genomics
  - Sorting out the mess

Functional Genomics

- Other, more direct, large-scale ways of identifying gene functions and associations
  - (for example yeast two-hybrid methods

Structural Genomics

- emphasizes high-throughput, whole-genome analysis.
  - outlines the current state
  - future plans of structural genomics efforts around the world and describes the possible benefits of this research

Proteomics

What Is Proteomics?

- Proteomics is the study of the proteome—the “PROTEin complement of the genOME”
- More specifically, "the qualitative and quantitative comparison of proteomes under different conditions to further unravel biological processes"

What Makes Proteomics Important?

- A cell's DNA—its genome—describes a blueprint for the cell's potential, all the possible forms that it could conceivably take. It does *not* describe the cell's actual, current form, in the same way that the source code of a computer program does not tell us what input a particular user is currently giving his copy of that program.

What Makes Proteomics Important?

- All cells in an organism contain the same DNA.
- This DNA encodes every possible cell type in that organism—muscle, bone, nerve, skin, etc.
- If we want to know about the type and state of a particular cell, the DNA does not help us, in the same way that knowing what language a computer program was written in tells us nothing about what the program does.

What Makes Proteomics Important?

- There are more than 160,000 genes in each cell, only a handful of which actually determine that cell's structure.
- Many of the interesting things about a given cell's current state can be deduced from the type and structure of the proteins it expresses.
- Changes in, for example, tissue types, carbon sources, temperature, and stage in life of the cell can be observed in its proteins.

Proteomics In Disease Treatment

- Nearly all major diseases—more than 98% of all hospital admissions—are caused by an particular pattern in a group of genes.
- Isolating this group by comparing the hundreds of thousands of genes in each of many genomes would be very impractical.
- Looking at the proteomes of the cells associated with the disease is much more efficient.

Proteomics In Disease Treatment

- Many human diseases are caused by a normal protein being modified improperly. This also can only be detected in the proteome, not the genome.
- The targets of almost all medical drugs are proteins. By identifying these proteins, proteomics aids the progress of pharmacogenetics.

Examples

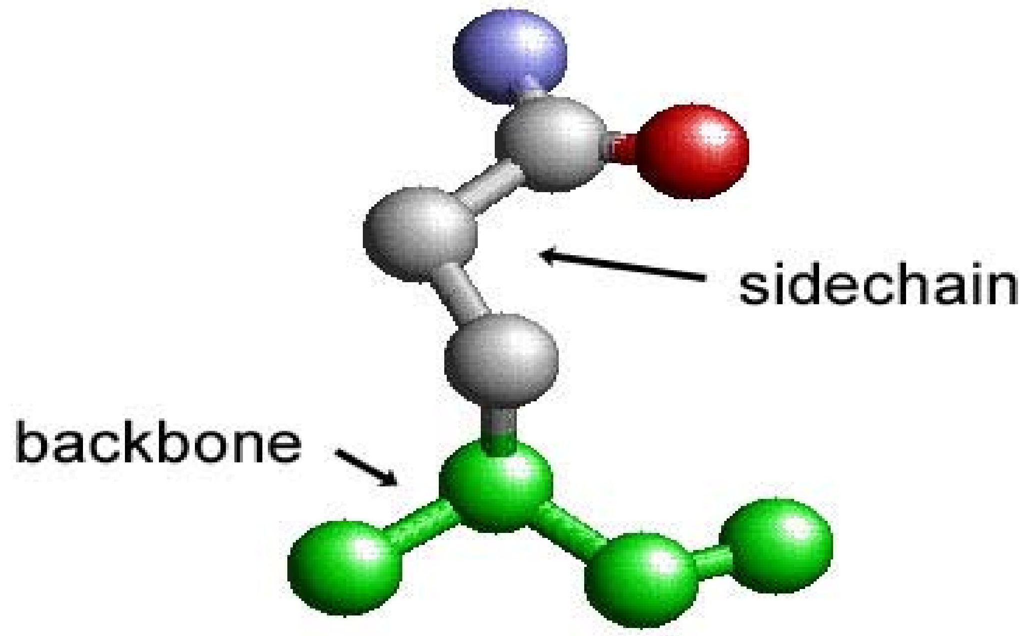
- What do these have in common?
  - Alzheimer's disease
  - Cystic fibrosis
  - Mad Cow disease
  - An inherited form of emphysema
  - Even many cancers

Protein Folding

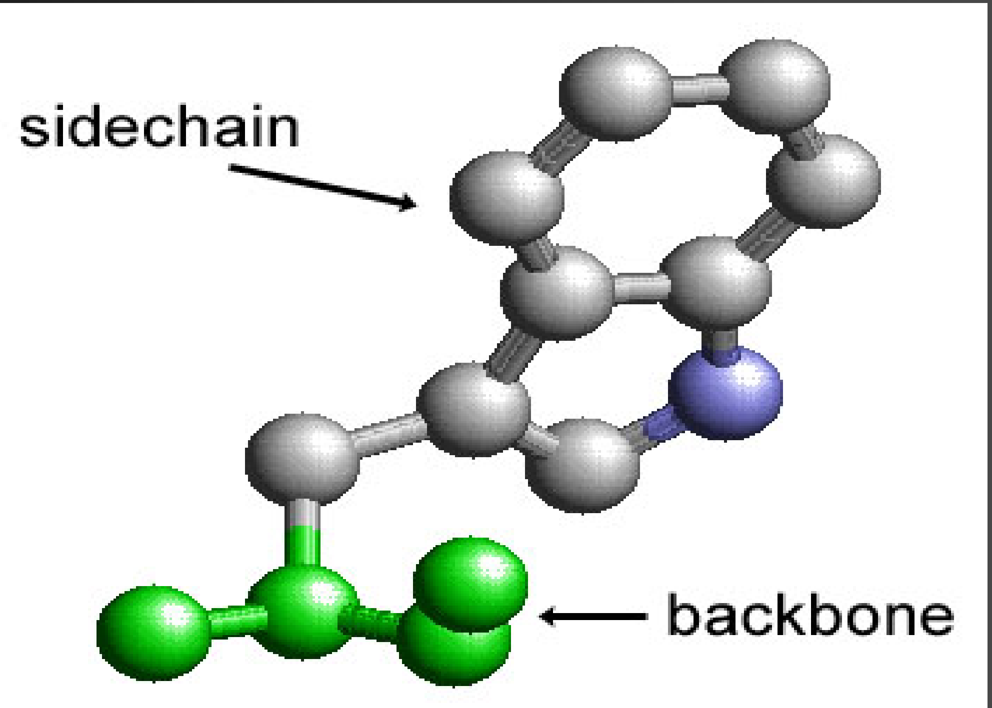
What is it?

- Fundamental components
  - Proteins

- Ribosome's string together long linear chains of amino acids.
  - Called Proteins
  - Loop about each other in a variety of ways
    - Known as folding
    - Determines whether or not the protein functions



The amino acid **Glutamine**



The amino acid **Tryptophan**

Dangers

- Folding determines function
- Of the many ways of folding one means correct functionality
  - Misfolded proteins can mean the protein will have a lack of functionality
    - Even worse can be damaging or dangerous to other proteins
    - Too much of a misfolded protein can be worse then too little of a normal folded one
    - Can poison the cells around it

History

- Linus Pauling – half a century ago
  - Discovered
    - A-helix
    - B-sheets
      - These are found in almost every protein
- Christian Anfinsen – early 1960's
  - Discovered
    - Proteins tie themselves
    - If separated fold back into their own proper form
      - No folder or shaper needed

Expansion to Anfinsen

- Sometime the protein will fold into the WRONG shape
- Chaperones
  - Proteins who’s job is to keep their target proteins from getting off the right folding path
  - These two key elements help us understand keys to protein folding diseases

What is Protein Folding

- Primary Structure
  - 3-D conformation of a protein depends only on its linear amino acid sequence
  - In theory can be computed explicitly with only this information
  - One of the driving forces that is thought to cause protein folding is called the hydrophobic effect

Hydrophobic effect

- Certain side chains do not like to be exposed to water
  - Tend to be found at the core of most proteins
  - Minimize surface area in contact with water

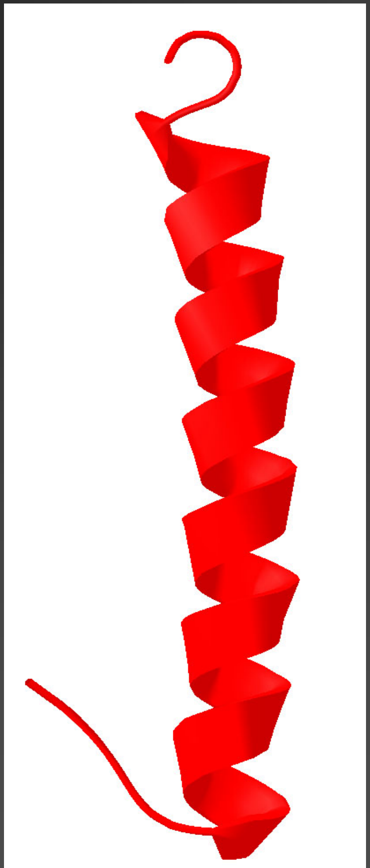
Proteins

- Two Repetitive features of a protein
  - Alpha-helix
  - Beta-sheet

Alpha-helix

- consecutive residues
  - Arranged in spiral staircase

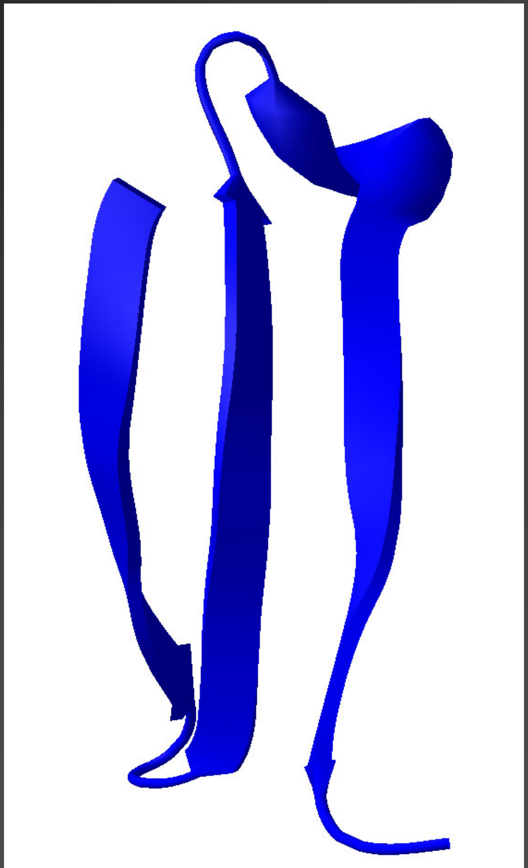
Alpha-helix



Beta-Sheets

- Comprised of two or more extended strands of amino-acids joined by inter strand hydrogen bonds

Beta-sheet



Hydrogen Bonds

- In both secondary structures
  - Alpha-helix
  - Beta-Sheets

- Responsible for stabilization
- Greatly effect the final fold of the protein

Fold Calculation

- Of all the possible ways the protein could fold, which one is
  - Most stable structure
  - Lowest energy
- Calculation of protein energy is only approximate
  - Thus compounding the complexity of such a calculation
  - Requiring enormous computational power

Why Fold Proteins

- Many genetic diseases are caused by dysfunctional proteins
  - By learning the structures we can learn the functions of each protein
  - Build better cures
  - Understand mutation
  - Assign structures functions to every protein
    - Thus understand the human genome
    - Decode the Human DNA

Resources

- http://www.faseb.org/opar/protfold/protein.html
- http://bioinformatics.org/faq/
- http://www.hhmi.org/news/baker2.html
- http://bioinfo.mshri.on.ca/trades/
- http://www.ncbi.nlm.nih.gov/Education/
- http://bioinformatics.org/faq/
- http://www.toplab.de/proteomics.htm
- http://www.wiley.co.uk/wileychi/genomics/proteomics.html
- http://everything2.com/?node=proteome
- http://us.expasy.org/proteomics_def.html
- http://www.sdu.dk/Nat/CPA/proteomics.html
- http://www.accessexcellence.org/AB/BC/Gregor_Mendel.html
- http://www.laskerfoundation.org/news/gnn/timeline/1888.html
- http://www.webref.org/scientists/
- http://dmoz.org/Science/Biology/Genetics/History/
- http://www.cshl.org/
- http://bioinformatics.org/faq/
- http://www.netsci.org/Science/Bioinform/feature06.html
- http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookDNAMOLGEN.html
- http://www.accessexcellence.org/AE/AEPC/WWC/1994/geneticstln.html
- http://www.mun.ca/biology/scarr/4241/TKAMgenetics.html
- http://www.cs.iastate.edu/jva/jva-archive.shtml
- http://www-sop.inria.fr/acacia/personnel/Fabien.Gandon/lecture/uk1999/history/
- http://inventors.about.com/library/inventors/blsoftware.htm
- http://www.nature.com/genomics/