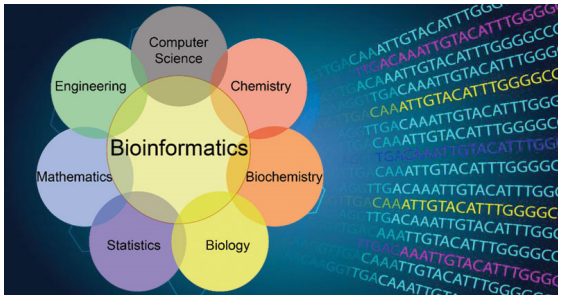
Introduction to Bioinformatics

Daniel Bergey // Great Basin College



1

What is Bioinformatics ??

Bioinformatics is the science dedicated to development and application of computer hardware and software to the acquisition, storage, analysis, and visualization of biological information. → *"Biology" + "Informatics"*

2

Three Primary Components of Bioinformatics

(1) Development of new algorithms and statistics for assessing relationships among large sets of biological data (e.g., DNA sequence data).

(2) Application of these tools for the analysis and interpretation of the various biological data (e.g., nucleotide sequences, amino acid sequences.

(3) The development of databases for efficient storage, access and management of biological information.

3

Bioinformatics vs. Computational Biology

Computational biology: Use and application of different computer programming languages, operating systems to test and verify hypothesis based on available data from sequencing results, image processing, mass spectrometry, etc. Analyze data to find possible scenario/biological process that might be verified with simulations and statistical hypothesis, but also work to model existing processes, determine parameters.

4

Bioinformatics vs. Computational Biology

- **Bioinformatics** includes the study of large sets of biological data, results of scientific studies, and application of biostatistics methods. Examples include analysis and integration of genetic and genomic data, prediction of protein function from data sequence and structural information, and comparisons of proteins to help improve personalized medicine.
- **Computational biology** is generally focused on finding solutions to issues raised by bioinformatics studies. For example, the examination of how proteins interact with each other through the simulation of protein folding, motion, and interaction.

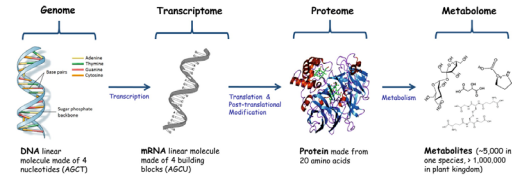
Both disciplines are branches of the rapidly-expanding fields of data science and biotechnology.

5

Bioinformatics vs. Computational Biology

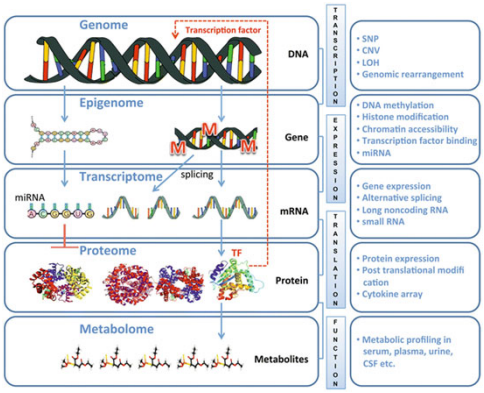
Computational biology: The study of biology using computational techniques. The goal is to learn new biology, knowledge about living systems. *It is about science.*

Bioinformatics: The creation of tools (algorithms, databases) that solve problems. The goal is to build useful tools that work on biological data. *It is about engineering.*



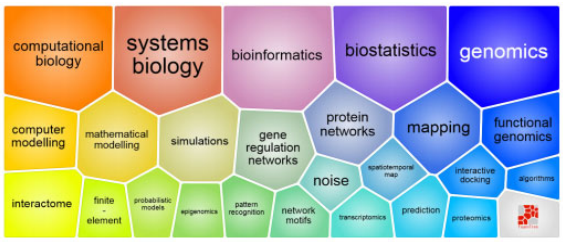
6

Major "Omic" Sub-disciplines



7

Integration of Disciplines



8

- Bioinformatics derives knowledge from computer analysis of biological data.
- These can consist of the information stored in the genetic code, but also experimental results from various sources, patient statistics, and scientific literature.
- Research in bioinformatics includes method development for storage, retrieval, and analysis of the data.
- Bioinformatics is a rapidly developing branch of biology and is highly interdisciplinary, using techniques and concepts from informatics, statistics, mathematics, chemistry, biochemistry, and physics.
- It has numerous practical applications in different areas of biology and medicine.

9

Timeline of Biological Databases

- **1965:** National Biomedical Foundation compiled, published collection of amino acids sequence in the "Atlas of protein sequence and structure", edited by Margaret O. Dayhoff. Amino sequence comparisons made by developing computer software for comparing distantly related sequences to assess their evolutionary relatedness.
- **1980:** European Molecular Biology Lab (EMBL-EBI) established their data library to collect, organize and distribute nucleotide sequence data and related information.
- **1984:** Protein Information Resource (PIR-International) established by he National Biomedical Research Foundation. Most noted for protein sequence database. Extension of Dayhoff's initial work.
- **1982:** GenBank established by NIH. Nucleotide sequence database.
- **1986:** DNA Data Bank of Japan (DDBJ). DNA sequence database.
- **1994:** NCBI established in USA. NCBI serves as primary information databank and provider of information.

10

International Cooperation

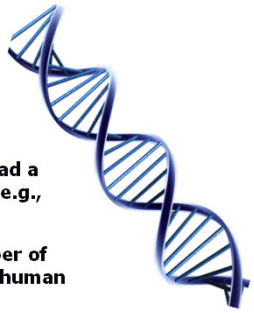
- **DDBJ, EMBL-EBI, GenBank exchange data on a daily basis, so these three databanks contain the same data at any given time.**

11

General Approaches for Acquiring Sequence Information

12

1. DNA Extraction



- Obtain DNA sample from organism (straightforward)
- Sequencers can accurately read a small number of bases pairs (e.g., Illumina NextSeq = 150 bp)
- Organisms have a huge number of base pairs. For example, the human genome has 3,200 Mb

What Next ?

13

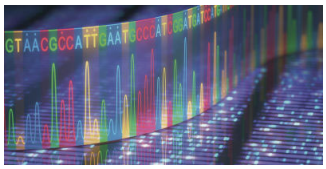
2. Smash the DNA

- Make the DNA fit into the machine by force – i.e., break the entire genome in millions of short fragments.
- For example, a human genome is broken into more than 300M pieces.

14

3. Feed the Short Fragments into a Sequencer

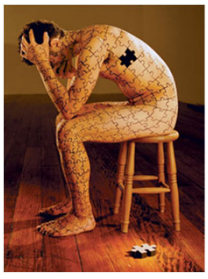
- Sequencers use a variety of technology to individually read chemical components aka base pairs and turn them into digital values.



15

4. Assemble the Pieces

- By using powerful computers and software, the puzzle pieces start to resemble the input product.



http://www.photographycorner.com/product-reviews/services/venus-personalized-jigsaw-puzzles
Puzzle (Biffy Clyro album) - Wikipedia,

16

5. Digital Genome

- **Genome is now converted from chemical to digital, and exploration and hypothesis testing can begin.**
- **For example, a human genome is compressed into ~3.3GB of files.**

https://www.linkedin.com/pulse/personalization-big-data-analytics-personal-genome-peter-b-nichol

17

6. RNA Sequencing

- **<u>Assumption:</u> Gene expression has biological intent**

  Yes! gene expression does NOT necessarily imply, or result in, protein expression – but, we have to start somewhere

- **Simultaneously comparing measurements of all RNAs in a sample.**
- **Multiple samples over time, or across conditions, can be used for gene expression studies, such as differentially expressed genes.**

18

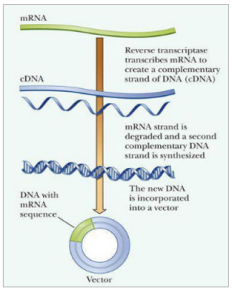
Types of Sequences in Nucleotide Sequence Databases

- **Copy DNA (cDNA) sequences:** A cDNA molecule is obtained by reverse transcription of an RNA molecule. The cDNA sequences therefore represent that part of the genome that is transcribed into RNA. If the cDNA is obtained from mRNA, it will represent only the exon (coding) sequences of the gene expressed in the relevant cell, tissue, organism.
- **Genomic DNA sequences:** These sequences represent the complete genome of the organisms, which includes both transcribed and non-transcribed sequences (e.g., regulatory sequences).

19

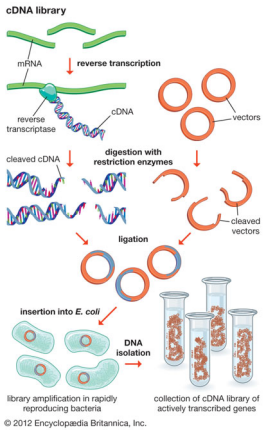
Construction of a cDNA Library (basic steps)

1. Isolate <u>total RNA</u> from cell or tissue.
   - *Note mRNA makes up only about* ***2%*** *of the total RNA in a typical cell.*
2. Purify mRNA fraction
3. Use enzyme (RT) to convert RNA to "copy" DNA (cDNA)
4. Incorporate (fuse) cDNA fragment into a cloning vector.
5. Transform bacterial strain E. coli with engineered plasmid containing cDNA fragments.



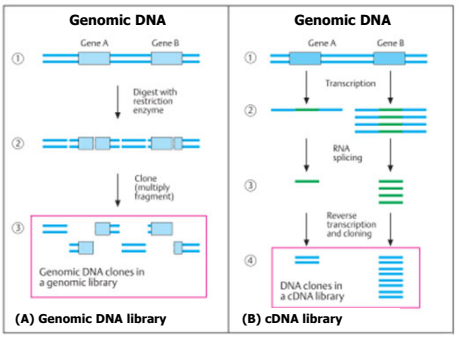
20

Construction of a cDNA Library (more detail)



21

Genomic vs. cDNA Libraries



22

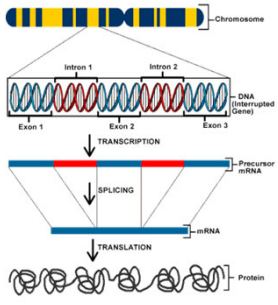
Expressed Sequence Tags (EST)

- ESTs are small pieces of DNA sequence (100 to 800 nucleotides) generated by sequencing randomly selected cDNA clones from a library.
- Expressed Sequence Tags (ESTs) are short, single-pass sequence reads from mRNA (cDNA).
- ESTs are bits of DNA sequence that represent genes expressed in certain cells, tissues, or organs from different organisms and use these to fish a gene out of a portion of chromosomal DNA by matching base pairs

23

Expressed Sequence Tags (ESTs)

- **ESTs** are small pieces of DNA sequence (usually 200 to 500 nucleotides) generated by sequencing either one, or both, ends of an expressed gene.
- The general idea is to sequence bits of DNA that represent genes expressed in certain cells, tissues, or organs from different organisms and use these "**tags**" to fish a gene out of a portion of chromosomal DNA by matching base pairs.



http://www.cyto.purdue.edu/cdroms/cyto6/content/primer/est.htm

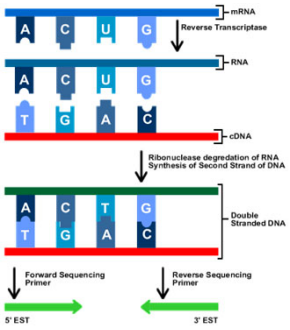
24

- **EST** sequences are obtained by sequencing only a portion of the cDNA molecules produced from mRNA. These relatively short sequences are called "tags" because they can be used as **probes** for the isolation of any gene-of-interest from the genomic DNA.
- This EST approach was used by J. Craig Venter and his group for obtaining the sequence of expressed portion of human genome. The use of ESTs generated enormous amounts of sequence data that enabled construction of a preliminary transcript map of the human genome.

25

From cDNA to ESTs

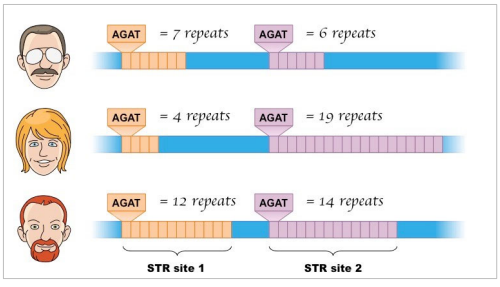
- **5' EST:** Beginning part of cDNA sequence. This part of transcript usually contains protein coding sequences that are usually conserved across species.
- **3' EST:** End part of cDNA that tend to be non-coding (or untranslated regions) that display less cross-species conservation than coding sequences.



26

STRs and VNTRs

- STR: Short Tandem Repeat
- VNTR: Variable Number Tandem Repeat



https://www.youtube.com/watch?v=9bEAJYnVVBA

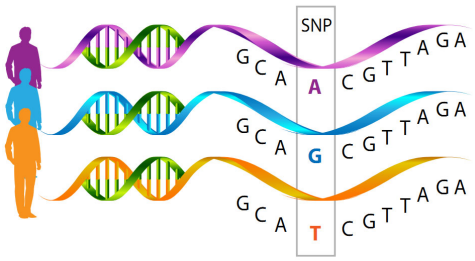
27

Single Nucleotide Polymorphisms (SNP)

- Ongoing research is identifying increasing numbers of SNPs associated with complex diseases such as heart disease, diabetes, and cancer.
- SNPs ("snips") are the most common type of genetic variation among people.
- Each SNP represents a difference in a single nucleotide.
- SNPs occur normally throughout every person's genome about once every **1,000** nucleotides on average, which means there are 4 to 5 million SNPs in every person's genome. *These variations may be unique, or common, in many individuals.*
- More than **100 million** SNPs in populations around the world have been identified, and most SNP variants are found in DNA regions between genes. These SNP variants can be used as biological markers to help locate genes associated with different diseases. When SNPs occur within a gene, or in a regulatory region near a gene, they may play an important role in disease by directly affecting gene function.
- Most SNPs have NO effect on health or development, but some have been shown to be very important in human health. For example, SNPs have been identified that may help predict an individual's response to certain drugs, susceptibility to environmental factors or toxins, and risk of developing particular diseases. SNPs can also be used to track the inheritance of disease genes within families.

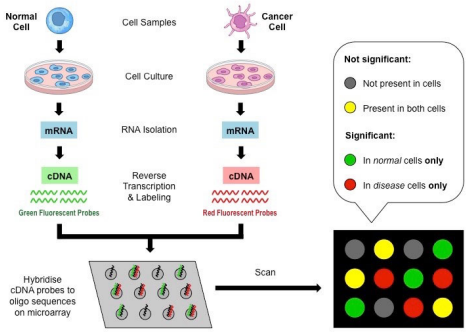
28

Single Nucleotide Polymorphisms (SNP)



29

Microarrays



https://www.illumina.com/science/technology/microarray.html

30

Genome-Wide Association Studies (GWAS)

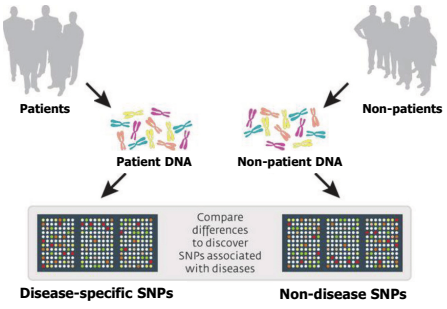
- **GWAS** are used to identify whether common SNPs in the population are associated with disease.
- GWAS look at hundreds of thousands of SNPs across an entire genome to see which ones may be associated with a specific disease.

https://knowgenetics.org/genome-wide-association-studies-gwas/

31

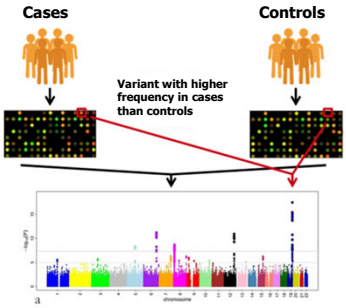
How SNPs are used in GWAS

- **Genome-Wide Association Studies (GWAS)** can be used to identify whether common SNPs in the population are associated with disease. ... GWAS look at hundreds of thousands of SNPs across the whole genome, to see which of them are associated with a specific disease.



32

How SNPs are used in GWAS



https://www.ebi.ac.uk/training/online/courses/gwas-catalogue-exploring-snp-trait-associations/what-is-gwas-catalog/what-are-genome-wide-association-studies-gwas/

33

Genomics

- Genomics uses techniques of molecular biology & bioinformatics identifying and analyzing DNA and RNA sequences of structural genes, regulatory sequences, and non-coding sequences.
- Genomics involves mapping (locating) and sequencing the DNA (e.g., genes) of all the chromosomes of an organism.
- This involves extensive analysis of the nucleic acids using molecular biology techniques before the sequence data are ready for processing by a computer.
- After sequencing techniques became practical, it became clear that it was not reliable to estimate the number of genes in an organism based solely on the number of nucleotide base pairs because of the presence of high numbers of redundant copies of many genes and sequence regions. Genomics has helped to rectify this problem.
- ***Genomics makes Transcriptomics makes Proteomics.***

34

Genomics Timeline Notes

- **1986:** The first automatic DNA sequencer was developed in by Leroy Hood.
- **1995:** Haemophilus influenzae (a bacterium) was the first living organism to be sequenced.
- **2003:** The Human Genome Project (HGP) declared complete (first draft). HGP was an international scientific research project with the goal of determining the base pairs that make up human DNA, and of identifying and mapping all of the genes of the human genome.

Caveat: Even if all the genes in a genome are identified, the genes only indicate that, at some point in time, they may be transcribed to produce functional RNA or protein products. For example, the human genome is estimated to contains around 22,000 protein coding genes, BUT, only a subset of these genes is expressed in any given cell type, at a particular time. → ***There is still LOTS more work left to be done !***

35

Human Genome Project Announced (2001)

36

Metagenomics vs. Genomics

- **Metagenomics** is the study of collective genomes recovered from environmental samples, especially the differentiation of genomes from multiple organisms or individuals. For example, metagenomics studies can involve symbiotic relationships among microbes, analysis of microbial samples on a dead body at a crime scene, etc.

  Metagenomics is focused on studying communities of microorganisms present, without the necessity of obtaining pure cultures.

- In contrast, **genomics** is focused on study of the complete genome of an organism.

37

Transcriptomics

- Transcriptomics is the study of the transcriptome, which includes the **whole set of mRNA** molecules in one cell, or a population of cells.
- Studying the transcriptome provides information about the expression level of genes. High throughput techniques (e.g., microarrays, SAGE) are used that are capable of sampling tens of thousands of different mRNAs at a time.
- These techniques has enabled biologists to routinely monitor gene expression profiles between the control cells vs. treatment cells.

38

Notes on Tool & Database Development

- Laboratories and companies continue to generate massive amounts of data such as DNA sequences, gene expression information, 3D molecular structure, and highly-throughput screening. Consequently, effective databases for storing and quickly accessing data must be development and maintained. An ongoing aim of bioinformatics is developing tools and resources that aid in the analysis of ever-increasing data.
- Data analysis tools are used to analyze the data and interpret the results in a biologically meaningful manner. Efficient analysis requires efficiently deigned databases.
- If efficient queries cannot be performed, or if the performance is too slow, the whole system breaks down since investigators will not be inclined to use the database.

39

Data Integration

- Once information has been analyzed, a researcher will often want to associate or integrate it with the related data from the other databases.
- For example, a scientist may run series of gene expression analysis experiments and observe that a particular set of say 100 genes is more highly expressed in a cancerous lung tissue than in a normal lung tissue. The scientist would then likely ask which of these 100 differentially expressed genes is most likely to contribute to the disease process.

40

Bioinformatics Applications and Molecular Medicine

- The human genome will have profound effects on the fields of biomedical research and clinical medicine. ***Every disease has a genetic component.***

  These components may be inherited, or occur as a result of the body's response to an emotional or environmental stress, or diet, which could cause alterations in the genome (e.g., cancer, heart disease, diabetes, colitis, chronic depression, etc.)

- The completion of the human genome allows investigators to search for the genes directly associated with different diseases, and provides a means for better understanding the molecular basis of disease. Knowledge of the molecular mechanisms of disease will enable better treatments, cures, and even preventative tests to be developed.

41

Personalized Medicine & Pharmacogenetics

- Clinical medicine will become increasingly more personalized with the development of the field of pharmacogenomics – the study of how an individual's genetic inheritance affects the body's response to drugs.

- At present, some beneficial drugs fail to make it to the market because a small percentage of the clinical patient population show adverse affects to a drug due to sequence variants (mutations) in their DNA. As a result, some potentially life saving drugs never make it to the marketplace.

  Today, doctors have to use **trial & error** to find the best drug to treat a particular patient. Progress in pharmacogenetics will minimize this "trial & error" strategy for finding the best drugs for individual treatments.

42

Gene Therapy & Gene Editing

- The potential for using genes themselves to treat disease will likely soon become a reality. Gene therapy is an approach used to treat, cure or even prevent disease by changing the expression patterns of an individual's genes.

- Currently, this field is in a very early stage with clinical trials for many different types of cancer and other diseases ongoing.

- Gene editing seems to hold much promise, but still has significant obstacles to overcome before clinical trials on humans could ever begin. ***Moral & ethical questions? Potential limitations and pitfalls ?***

43

The Reality of Bioweapons

- In **2002**, scientists built the virus poliomyelitis using entirely artificial means.

- This was done using genomic data available on the Internet and materials from a mail-order chemical supply.

- The research was financed by the US Department of Defense as part of a biowarfare response program to prove to the world the reality of bioweapons. The researchers also hoped their work will discourage officials from ever relaxing programs of immunization. *Not surprisingly, his project was met with very mixed feelings.*

https://www.sciencemag.org/news/2002/07/poliovirus-baked-scratch

44

THANK YOU !

- ✓ **INBRE (IDeA Network of Biomedical Research Excellence)**
- ✓ **Nevada Center for Bioinformatics (at UNR)**
- ✓ **Great Basin College**
- ✓ **Dr. Juli Petereit (Director NCB)**
- ✓ **Drs. Tin Nguyen, Lucas Bishop, and Hans Vasquez-Gross**

45