

Biology 644

Old Title: Bioinformatics for Molecular Biologists

Potential New Title: Integrated Bioinformatics
Using R for Both Wet and Dry Scientists

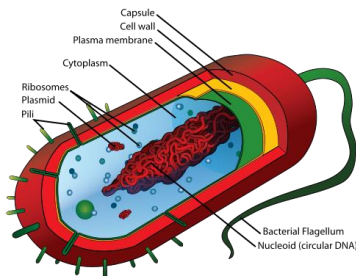
Biology 644: Bioinformatics

Outline

- Temporary Seating Chart
- Review Syllabus
- Brief Molecular Biology Review Lecture
- R Lab
 - Install Packages
 - Chapter 1: Sections 1.1 – 1.5
 - Chapter 1 Supplemental

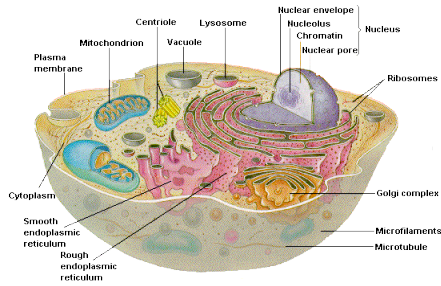
Biology 644: Bioinformatics

Bacterial Cell (Prokaryotic)



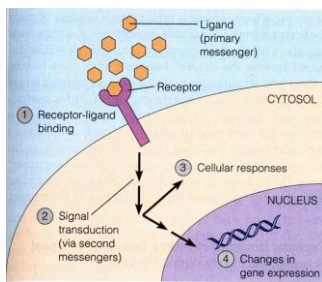
Biology 644: Bioinformatics

Eukaryotic Cell



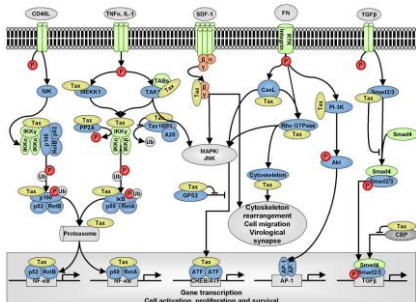
Biology 644: Bioinformatics

Simplest Cell Signaling Diagram You Will Ever See



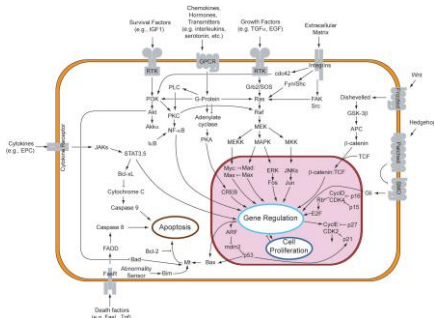
Biology 644: Bioinformatics

Slightly More Complicated Cell Signaling Diagram



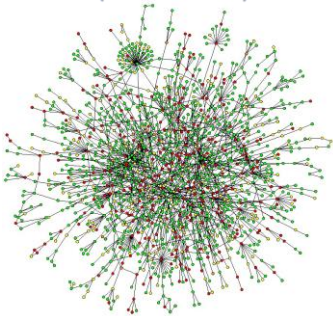
Biology 644: Bioinformatics

Even More Complicated Cell Signaling Diagram



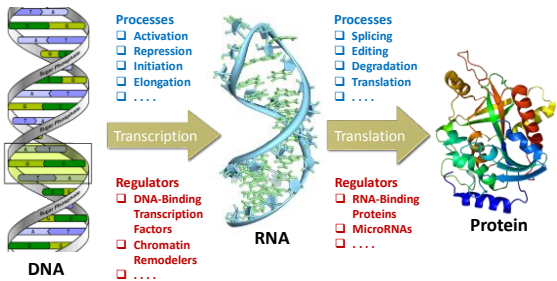
Biology 644: Bioinformatics

Much More Complicated Cell Signaling Diagram (a.k.a. Hairball)



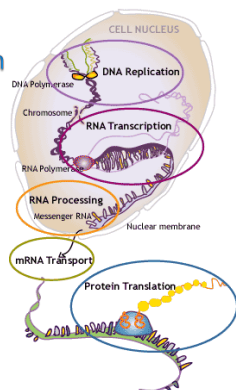
Biology 644: Bioinformatics

The Central Dogma of Molecular Biology



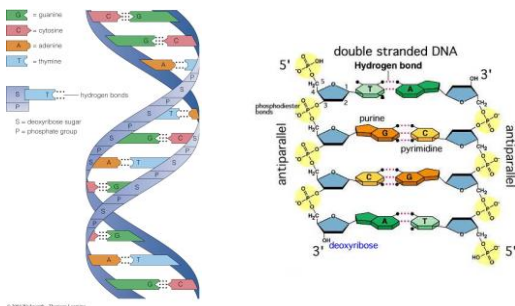
Biology 644: Bioinformatics

DNA → RNA → Protein In the Cell



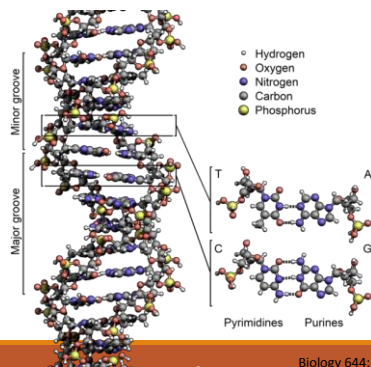
Biology 644: Bioinformatics

DNA Sequence



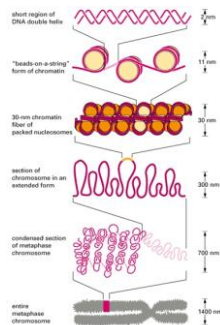
Biology 644: Bioinformatics

DNA Structure



Biology 644: Bioinformatics

The diagram illustrates the process of DNA replication. At the top, a double helix is shown with the overall direction of replication indicated by a red arrow pointing to the right. The origin of replication is marked by a red dot. The leading strand is synthesized continuously towards the origin, while the lagging strand is synthesized discontinuously away from the origin as Okazaki fragments. The diagram labels the DNA polymerase, DNA primase, and DNA ligase. An inset box provides a detailed view of the replication fork, showing the leading and lagging strands, the origin of replication, and the overall direction of replication.



Codon Alphabet

		Second base					
		U	C	A	G		
First base	U	UUU } Phenyl- alanine F UUC } UUA } Leucine L UUG }	UCU } Serine S UCC } UCA } UCG }	UAU } Tyrosine Y UAC } UAA } Stop codon UAG } Stop codon	UGU } Cysteine C UGC } UGA } Stop codon UGG } Tryptophan W	Third base	U C A G
	C	CUU } Leucine L CUC } CUA } CUG }	CCU } Proline P CCC } CCA } CCG }	CAU } Histidine H CAG } CAA } Glutamine Q CAG }	CGU } Arginine R CGC } CGA } CGG }		U C A G
	A	AUU } Isoleucine I AUC } AUA } AUG } Methionine start codon M	ACU } Threonine T ACC } ACA } ACG }	AAU } Asparagine N AAC } AAA } Lysine K AAG }	AGU } Serine S AGC } AGA } Arginine R AGG }		U C A G
	G	GUU } Valine V GUC } GUA } GUG }	GCU } Alanine A GCC } GCA } GCG }	GAU } Aspartic acid D GAC } GAA } Glutamic acid E GAG }	GGU } Glycine G GGC } GGA } GGG }		U C A G

Biology 644: Bioinformatics

Bioinformatics: the analysis, organization, integration, and annotation of biological data

- ❑ **Sequence Bioinformatics:** patterns about the sequence can reveal insight into transcription, translation, and function of synthesized proteins.
 - ❑ **Genome analysis:** Genome assembly, genome annotation, gene finding, alternative splicing, EST analysis and comparative genomics.
 - ❑ **Sequence analysis:** Multiple sequence alignment, sequence search and clustering, function prediction, motif discovery, functional site recognition in protein, RNA and DNA sequences.
 - ❑ **Phylogenetics:** Phylogeny estimation, models of evolution, comparative biological methods, population genetics.
 - ❑ **Analysis of high-throughput biological data:** Microarrays (nucleic acid, protein, array CGH, genome tiling, and other arrays), EST, SAGE, MPSS, proteomics, mass spectrometry.
 - ❑ **Genetics and population analysis:** Linkage analysis, association analysis, population simulation, haplotyping, marker discovery, genotype calling.
 - ❑ **Systems biology:** Systems approaches to molecular biology, multiscale modeling, pathways, gene networks.
 - ❑ **Computational Proteomics:** Filtering and indexing sequence databases, Peptide quantification and identification, Genome annotations via mass spectrometry, Identification of post-translational modifications, Protein-protein interactions, Computational approaches to analysis of large scale Mass spectrometry data, Exploration and visualization of proteomic data, Data models and integration for proteomics and genomics, Querying and retrieval of proteomics and genomics data etc.
- ❑ **Structural Bioinformatics** – 3D structure reveals further information about DNA, protein- DNA binding, RNA, protein-RNA binding, Protein-Protein interactions, etc.
 - ❑ Structural genomics via **mass spectrometry**

Biology 644: Bioinformatics

R

- A free software programming language and software environment for statistical computing and graphics
- Implementation of the S programming language
- Also contains lexical scoping semantics inspired by Scheme
- A GNU project that is freely available under the GNU General Public License
- Written primarily in C, Fortran, and R
- Pre-compiled binary versions are provided for many popular operating systems
- Contains thousands of packages to analyze many different types of biological data (and data from statistics, geology, finance...)

Biology 644: Bioinformatics



- A free, open source and open development software project for the analysis and comprehension of genomic data generated by wet lab experiments in molecular biology
- Provides widespread access to a broad range of powerful statistical and graphical methods for the analysis of genomic data.
- Based primarily on the statistical R programming language, but also contains contributions in other programming languages
- A large number of genome annotation packages
- Most Bioconductor components are distributed as R packages
- Used for the analysis of:
 - Single channel Affymetrix
 - Two or more channel cDNA/Oligo microarrays
 - SAGE, sequence, and SNP data
 - Much more....

Biology 644: Bioinformatics
