THIS IS A DRAFT

Bioinformatics: advancing the scientific understanding of living systems through computation

Final Revision for ISCB September 5, 2014 based on feedback from the GEP Community, NIBLSE Community, Others

Topic	Learning Goals	Sample Learning Objectives
Computation in the life sciences	What is the role of computation in hypothesis-driven discovery processes within the life sciences?	Describe the role of bioinformatics in the scientific research method. Explain the necessity for computation in life sciences research. Explain the role of wet-bench techniques in verifying computational results in life science research. Compare and contrast computer-based research with wet-lab research. Read a scientific article and evaluate how bioinformatics methods were employed by the authors to explore a particular hypothesis. Given a scientific question, develop a hypothesis and define computational approaches that could be used to explore the hypothesis. Evaluate the social, legal, and ethical implications of computational approaches to understanding biology.
	What computational concepts are important in bioinformatics?	Define the term *algorithm*. Explain the difference between a *heuristic* (approximate) algorithm and an *exact* algorithm. Describe the three basic programming structures: sequential, repetition (e.g., **`while`**, **`for`**) and selection (e.g., **`if`**). Use variables and data structures (e.g., lists, arrays, scalars, hash functions). Describe what a regular expression is. Explain the concept of cloud computing. Describe the importance of "big data" in bioinformatics. Describe the means by which "big data" are managed and stored (e.g. dmptool.org).
	What statistical concepts are important in bioinformatics?	Calculate average, median, mode, range, standard deviations for a given data set. Calculate p-values using a t-test for discrete data. Calculate p-values using a z-test for continuous data. Calculate an e-value statistic. Describe the importance of statistical analysis of big data sets. Create a network to illustrate gene interactions.

DNA - Information Storage **[GENOMICS]**	Where are data about the genome found (e.g., nucleotide sequence, epigenomics) and how are they stored and accessed?	Describe how nucleotide sequence data are represented (FASTA, FASTQ, GenBank). Describe the nucleotide databases available at NCBI. Describe how the NCBI nucleotide databases intersect with other nucleotide databases (EBI, DDBJ, UniProt, etc.). Compare and contrast the data contained in different nucleotide databases. Search for a sequence record in a nucleotide database with a given accession number. Create a collection of nucleotide sequence records that meet a specified criterion (e.g., gene name or symbol). Determine the DNA methylation state of a particular region of a genome. Describe the types of metadata that accompany sequence data to make for useful biological interpretation (e.g. biological source, accession number, GeneID, journal articles, etc.).
	How can bioinformatics tools be employed to analyze genetic information?	Calculate the alignment score between two DNA sequences using a provided scoring matrix. Perform a BLASTN search and interpret the results. Explain the BLASTN algorithm for nucleotide sequence information. Interpret the biological significance of an e-value. Annotate a prokaryotic gene (derive a model). Annotate a eukaryotic gene (derive a model). Create and interpret a multiple sequence alignment (e.g., T-COFFEE, MUSCLE, etc.). For a genomic region of interest (e.g., the neighborhood of a particular gene), use a genome browser to view nearby genes, transcription factor binding regions, epigenetic information, etc. Describe Hidden Markov Models and how they can be used to assess motifs.
RNA - Information Transfer **[TRANSCRIPTOMICS]**	Where are data about the transcriptome found (e.g., expression, epigenomics and structure) and how are they stored and accessed?	Identify the euchromatin/heterochromatin boundaries, histone states in a given sequence, and the nucleosome modifications Compare and contrast DNA structure at telomeres and centromeres. Describe the RNA databases available at NCBI (e.g., ESTs, UniGene). Describe the types of metadata that accompany sequence data to make for useful biological interpretation (e.g. biological source, accession number, GeneID, journal articles, etc.).
	How can bioinformatics tools be employed to examine *transfer* of genetic information?	Given a microarray or RNA-seq data file, find the set of significantly differentially expressed genes. Perform motif discovery on the promoter regions of a set of genes identified by a ChIP-seq experiment. Use RNA structure prediction programs (e.g., RNASoft, RNAFold, RNAStructure) to evaluate possible structures for an RNA sequence.

		Identify the possible different splice isoforms possible from a given gene sequence.
Protein - Information in Action [PROTEOMICS]	Where are data about the proteome found (e.g., amino acid sequence and structure) and how are they stored and accessed?	Describe how protein sequence data are represented (e.g., FASTA, GenBank, etc.) Describe the different protein databases available at NCBI (sequence, structure, function). Describe how the NCBI databases intersect with other databases (e.g., EBI, DDBJ, UniProt, etc.). Compare and contrast data contained in different databases. Search for a protein record in a database with a given accession number. Create a collection of records that meet a specified criterion (e.g., gene name or symbol). Describe the types of metadata that accompany sequence, structure, and function data to make for useful biological interpretation (e.g. biological source, accession number, UniProt number, journal articles, etc.).
	How can bioinformatics tools be employed to examine protein structure and function?	Explain the BLASTP, BLASTX, tBLASTn, tBLASTx algorithms for protein sequence information Interpret the biological significance of an e-value. Describe Hidden Markov Models and how they can be used to assess motifs. Query a dataset with a specific protein sequence to learn about potential functions (e.g. Pfam, CDD, SwissProt, UniProt, etc.). View and interpret the structure output from Protein Data Bank (e.g. Cn3D, Jmol, etc.). Propose potential functions for a give protein structure. Explain the outputs from protein-folding algorithms to predict structure from sequence. Understand how protein structures are determined (e.g. NMR, crystallography). Compare and contrast the output from 2-D gel experiments. Analyze the output from mass spectrometry analysis (e.g., use the MASCOT package).
Small Molecules - Cellular Homeostasis [METABOLOMICS & SYSTEMS BIOLOGY]	Where are data about metabolomics and systems biology found and how are they stored and accessed?	Describe how metabolomics data are represented (e.g. Human Metabolome Database, METLIN Database, etc.) Describe the different metabolomic databases that are available. Describe the types of metadata that accompany metabolomic data to make for useful biological interpretation.
	How can bioinformatics tools be employed to examine flow of molecules within pathways?	Perform a GO analysis to identify the pathways relevant to a set of genes (e.g., identified by a transcriptomic study or a proteomic experiment). Use the KEGG pathway database to look up the interaction network of a pathway. Interpret the data from experiments (e.g., mass spectrometry, nuclear magnetic resonance, etc.) to determine levels of small molecule metabolites.

Ecology and Evolution [METAGENOMICS]	How can bioinformatics tools be employed to examine ecological niches?	Create and interpret a multiple sequence alignment (e.g., T-COFFEE, MUSCLE, etc.). Describe the components of a phylogenetic tree (e.g., root, node, leaf). Explain the various types of phylogenetic trees (e.g., rooted, unrooted). Interpret a phylogenetic tree (e.g., which organism is most closely related to a given organism in the tree) Sketch a phylogenetic tree from its Newick representation. Use bootstrapping to assess the quality of a phylogenetic tree. Create a phylogenetic tree for a set of related sequences (nucleotide or amino acid) (e.g., MEGA). Use pre-existing tools to analyze a metagenomic data set to determine the set of organisms present in a metagenomic sample (e.g., 16s rRNA, Greengenes, mothur, etc.)
Computational Skills	How do biologists employ software development as part of the scientific discovery process?	Write a script to calculate the reverse complement of a nucleotide sequence Write a script to determine reading frames of a nucleotide sequence. Write a script to calculate melting point of double-stranded DNA. Write a script to retrieve the promoter regions for a set of related genes. Write a script to find the longest open reading frame in a given nucleotide sequence Write a script to calculate the reverse complement of a nucleotide sequence. Write a script to convert an RNA sequence to cDNA and to amino acids Write a script to calculate molecular weight and isoelectric point. Write a script to count amino acid frequency. Write a script that compares the relative hydrophilicity/hydrophobicity of two protein sequences.
	What higher-level computational skills can be used in bioinformatics research?	Use a spreadsheet to perform simple data analysis. Use a spreadsheet to open, read, parse, modify and output comma-separate (.csv) files that will be ready to use in subsequent tools. Perform elementary statistical analysis on an "omics" dataset (e.g. using Excel or Weka). Perform Input/Output with data files. Interact with remote servers. Construct a bioinformatics pipeline. Use open source libraries and packages (e.g., BioPerl, Biopython, R, BioConductor). Use programs at the Unix/Linux command line to analyze bioinformatics data. Use graph theory to represent data networks.