COMP 4980: Introduction to Bioinformatics

Examples of project topics

Below are some options of topic choice [1]. The purpose of this list is to give you some detailed idea and guidance to choose your topic of interest. You could pick one from this list for your project, but please don't be limited by the list. Feel free to come up with other topics of your choice. Make sure to send your topic to the instructor for approval before Oct. 28. Topics that are not approved may result in redoing the project on other topics.

Topic 1: A good resource for information on disease-related molecules is in the Online Mendelian Inheritance of Man (OMIM). OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily. The full-text, referenced overviews in OMIM contain information on all known mendelian disorders and over 15,000 genes. OMIM focuses on the relationship between phenotype and genotype. It is updated daily, and the entries contain copious links to other genetics resources, such as the National Center of Biotechnology Information (NCBI) website.

If you are interested in a project relevant to inherit human disease, this is a very good resource to find the genetic reason of the disease, and go further to find relevant literature from there.

Topic 2: One important exercise a bioinformatician performs is to compare amino acid sequences. One reason to make comparisons is to determine the parts of the proteins that are critical for function. These regions are generally conserved within proteins that perform the same duties. Conserved regions are those that have nearly the same amino acid sequences. Proteins that perform the same duties are called homologs and can be found in different species. For example, *p*53 from humans and *p*53 from frogs perform the same functions. There are some regions within these proteins that will be similar in both humans and frogs. We call these regions conserved sequences. A **multiple sequence alignment** allows the bioinformatician to readily line up amino acid sequences of related proteins. The conserved regions are identified in the alignment.

For example, you can perform a multiple sequence alignment of three homologs of cytochrome C from human, yeast, and dog (or other organisms). Based on your multiple sequence alignment result, you could explain how the alignment gives you a clue as to which parts of the cytochrome C protein you would hypothesize are most important to its function. (The function is the same in all three organisms.) Then, find some evidences to support your results.

Topic 3: **Paralogs** are homologs within a single species that arose by gene duplication. For example, the human *p*53 has paralogous genes *p*63 and *p*73. By using global sequence alignment, one can answer questions such as: Which amino acid sequences within *p*53 are conserved in *p*63 and *p*73? Is the particular domain common within *p*53, *p*63, and *p*73 associated with a specific function? etc..

Topic 4: Phylogeny is the study about the history of the evolution of a species or group, especially in reference to lines of descent and relationships among broad groups of organisms. One can construct a phylogenetic tree using the mitochondrial gene that codes for 16S ribosomal RNA from different organisms, such as modern *H. sapiens*, chimpanzee, pygmy chimpanzee, Neanderthal, Denisovan, gorilla, orangutan, mitred leaf monkey, and hanuman langur.

[1] Some ideas are from the textbook: Concepts in Bioinformatics and Genomics, by Jamil Momand and Alison McCurdy, published by Oxford University Press.

Find the two sequences that have the shortest and longest evolutionary distance between them, and confirm their the evolutionary distance between them (distance matrix and length of tree branch).

If you assume that modern *H. sapiens* and chimpanzee diverged from a common ancestor 5 MYA, what is the amount of time in which *H. sapiens* and Neanderthal diverged from a common ancestor? Is this a realistic estimate for the number of years when modern *Homo sapiens* and Neanderthal evolved from a common ancestor?

What is the substitution rate per site per year in the 16S ribosomal gene assuming that modern *H. sapiens* and chimpanzee shared a common ancestor 5 MYA. Is this a realistic substitution rate?

Topic 5: Develop a bioinformatics tool. There could be so many project ideas under this topic. For example, you can develop a sequence alignment program to generate the scoring matrix (M), the traceback matrix (TB), the alignment score, and the pairwise alignment of a given alignment problem, such as

- a global sequence alignment using the PAM150 substitution matrix and a linear gap penalty with $w = -4$,
- a local alignment using the BLOSUM62 substitution matrix and a linear gap penalty with $w = -4$.

Note that the first step in program development is to understand the problem, its inputs, and its outputs. A good way to do that is to generate a solution by hand. In addition to helping you understand how to solve the problem, the solution can also be used to test your program.

Topic 6: Single-cell analysis offer exciting opportunities to uncover many secrets about life. Single-cell RNA-seq (scRNA-seq) data has been used mostly in currently analysis. Typically, an scRNA-seq dataset contains the expression profiles of a large number of genes, where each gene corresponds to a dimension, and the expression profile of each cell corresponds to a data point in the high dimensional cell state space. Therefore, a key step in scRNA-seq data analysis is dimensionality reduction.

Principal component analysis (PCA) is an essential method of dimensionality reduction for single-cell RNA-seq data. Since there are different packages and variations of PCA, how do different methods/packages of PCA performs on varied datasets, and how to adjust parameters? You can investigate this by examine a few different PCA methods on several scRNA-seq datasets and report your findings. To help you get started, you may read the paper Benchmarking principal component analysis for large-scale single-cell RNA-sequencing and reproduce part of their work.

Topic 7: Genome-wide association studies (GWAS) have been widely used in identifying associations between genetic variants and traits or disease. The most used genetic variants are single nucleotide polymorphisms (SNPs), which are changes of single DNA base-pairs. GWAS perform statistical hypothesis tests for each SNP, with the null hypothesis being no association between the SNP and the phenotype. Many software packages have been developed for GWAS analysis based on varied statistical models, such as PLINK, TASSEL, EMMA, and FaST-LMM. However, different packages often produce dissimilar association results and have varied performance in terms of speed. You may examine a few of different GWAS packages using same input data, and compare their output and computational speed. What can you find? How are their results similar, or dissimilar? What do you think of the reasons behind? How are the computational speed vary among packages?

Topic 8: Healthcare institutions need modern and powerful technology to provide high-quality, cost-effective care to patients. However, despite the considerable progress in the computerization and digitization of medicine, efficient and robust management tools have yet to materialize. One important reason for this is the extreme complexity and variability of healthcare operations, the needs of which have outgrown conventional management. You can conduct a literature review on the current development of healthcare management technologies. What are the up-to-date applications? What are their achievements and limitation of current solutions? How can we improve the management technologies? To help you get started, you can read this paper Improving healthcare operations management with machine learning.