BIO334 Theory

David Parker + wikipedia

Table of Contents

Mering

Multiple sequence alignment

Motivations for sequence alignment

1. Find genes that are related by common descent
2. to identify and check the state of "active sites"
3. to identify and characterize "protein domains"
4. to make phylogenetic inferences ("trees")

How it's done

- Substitution matrix
  - A substitution matrix describes the frequency at which a character in a nucleotide sequence or a protein sequence changes to other character states over evolutionary time.
  - Each amino acid is more or less likely to mutate into various other amino acids. For instance, a hydrophilic residue such as arginine is more likely to be replaced by another hydrophilic residue such as glutamine, than it is to be mutated into a hydrophobic residue such as leucine.
- Pairwise Alignment
  - BLAST: quick and dirty
    - basic local alignment search tool
    - Using a heuristic method, BLAST finds similar sequences, by locating short matches between the two sequences. This process of finding similar sequences is called seeding. It is after this first match that BLAST begins to make local alignments.
  - Dynamic Programming: correct and slow
    - The Smith–Waterman algorithm performs local sequence alignment; that is, for determining similar regions between two strings of nucleic acid sequences or protein sequences. Instead of looking at the entire sequence, the Smith–Waterman algorithm compares segments of all possible lengths and optimizes the similarity measure.
- Multiple Alignment
  - Combinatorial Explosion: very many possible solutions
  - Complexity: O(alignment_length^number of sequences)
  - => an NP-complete problem !
    - Although a solution to an NP-complete problem can be verified "quickly", there is no known way to find a solution quickly. That is, the time required to solve the problem using any currently known algorithm increases rapidly as the size of the problem grows. As a consequence, determining whether it is possible to solve these problems quickly, called the P versus NP problem, is one of the fundamental unsolved problems in computer science today.
    - While a method for computing the solutions to NP-complete problems quickly remains undiscovered, computer scientists and programmers still frequently encounter NP-complete problems. NP-complete problems are often addressed by using heuristic methods and approximation algorithms.

Phylogenetic trees

Generating phylogenetic trees

- Phenetic: trees are constructed based on observed characteristics directly, not on evolutionary history. (Distance based methods)
- Cladistic: trees are constructed based on fitting observed characteristics to some model of evolutionary history (Maximum Likelihood methods)

Which genes to use:

suitable marker genes ...

- should occur in every organism
- should rarely undergo horizontal transfer
- should be evolving 'slowly'
- should only occur in one copy per genome
- should function in a process that sees no change

For old events:

- Ribosomes and polymerases.

For recent events:

- Fast evolving genes.

Distance based methods (phenetic technique)

UPGMA: unweighted pair group method with arithmetic mean

1. Alignment
2. Distance matrix
3. Choose the smallest distance
4. Join them
5. Calculate the new distances to every other node in the tree
   1. We do this with a simple average of distances
   2. Dist[Spinach, MonHum]
      1. = (Dist[Spinach, Monkey] + Dist[Spinach, Human])/2
      2. = (90.8 + 86.3)/2 = 88.55
6. Repeat until all nodes are joined

Maximum Likelihood (cladistic technique)

- The likelihood is the probability of the data given the model
- The probability of observing the data under the assumed model will change depending on the parameter values of the model.
- The aim of maximum likelihood is to choose the value of the parameter that maximizes the probability of finding the data.

Typically, the model has additional free 'parameters':

- The rate of evolution can vary across parts of the tree
- The rate of evolution can vary from site to site in the protein

maximum likelihood is computed like so:

1. Image all ancestral possibilities and evolutionary paths.
2. Compute the likelihood of each path
   - L(path) = L(root) x Π L(branches)
   - =P(G→T)P(G→G) P(G→A)P(G→G) [...]
3. Multiply all likelihoods over all possible paths
4. Throughout, do not forget to optimize all free parameters
5. Repeat for each tree topology, identify the one with best Likelihood

How do we verify a tree?

- Simulation
- Bootstrapping
  - Bootstrap involves resampling with replacement from one's molecular data with to create fictional datasets, called bootstrap replicates, of the same size. Specifically, the molecular data is typically organized as a multiple sequence alignment (MSA) of s species ×n characters. Since most models assume independent characters, we generate a replicate by sampling n characters, with replacement, from the original MSA and do this B times. Krishna Kumar Ojha et al. (2022)

Shimizu

Genome-wide polymorphisms

In genomics, a genome-wide association study (GWA study, or GWAS), also known as whole genome association study (WGA study, or WGAS), is an observational study of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait. GWA studies typically focus on associations between single-nucleotide polymorphisms (SNPs) and traits like major human diseases, but can equally be applied to any other genetic variants and any other organisms.

Nucleotide diversity

Nucleotide diversity is a measure of genetic variation. It is usually associated with other statistical measures of population diversity, and is similar to expected heterozygosity. This statistic may be used to monitor diversity within or between ecological populations, to examine the genetic variation in crops and related species, or to determine evolutionary relationships.

Selective sweep / hitchhiking

In genetics, a selective sweep is the process through which a new beneficial mutation that increases its frequency and becomes fixed (i.e., reaches a frequency of 1) in the population leads to the reduction or elimination of genetic variation among nucleotide sequences that are near the mutation. In selective sweep, positive selection causes the new mutation to reach fixation so quickly that linked alleles can "hitchhike" and also become fixed.

Balancing selection

Balancing selection refers to a number of selective processes by which multiple alleles are actively maintained in the gene pool of a population at frequencies larger than expected from genetic drift alone. This can happen by various mechanisms, in particular, when the heterozygotes for the alleles under consideration have a higher fitness than the homozygote. In this way genetic polymorphism is conserved.

Tajima's D

Tajima's D is computed as the difference between two measures of genetic diversity: the mean number of pairwise differences and the number of segregating sites, each scaled so that they are expected to be the same in a neutrally evolving population of constant size.

The purpose of Tajima's D test is to distinguish between a DNA sequence evolving randomly ("neutrally") and one evolving under a non-random process, including directional selection or balancing selection, demographic expansion or contraction, genetic hitchhiking, or introgression. A randomly evolving DNA sequence contains mutations with no effect on the fitness and survival of an organism. The randomly evolving mutations are called "neutral", while mutations under selection are "non-neutral". For example, a mutation that causes prenatal death or severe disease would be expected to be under selection. In the population as a whole, the frequency of a neutral mutation fluctuates randomly (i.e. the percentage of individuals in the population with the mutation changes from one generation to the next, and this percentage is equally likely to go up or down) through genetic drift.

Tajimas D > 0

$\pi > s / \sum_{i=1}^{n-1} \frac{1}{i}$

- Many higher heterozygous sites
  - Population size decreases rapidly (bottleneck effect, founder effect)
    - singleton may be removed from the population
  - Balancing selection (heterozygote advantage)

Tajimas D < 0

$\pi < s / \sum_{i=1}^{n-1} \frac{1}{i}$

- Many lower heterozygous sites
  - Population size increases rapidly
    - Singleton may be introduced in the population
  - Positive/Negative selection (purifying selection, selective sweep, directional selection)

positive selection on replacement mutations

??

Wagner

Metabolic networks

A metabolic network is a network of chemical reactions whose two main functions are to produce

- chemical energy (for maintenance of cell functions and for biosynthesis)
- molecular building blocks for biosynthesis

Metabolic genotype

The part of a genome that encodes metabolic enzymes

Metabolic phenotype

Most general: All the molecules that a metabolism can synthesize in a given chemical environment

The most important of these molecules are biomass precursors (amino acids, DNA and RNA building blocks etc.).

A metabolism is viable if it can synthesize all of them.

More specific: the spectrum of nutrients on which a metabolism is viable

Metabolic flux

The rate at which an enzyme converts substrate into product per unit time.

Metabolic models

Metabolic models include artificial reactions that allow the flux of metabolites in and out of the system: external aka exchange reactions. They act on external metabolites, that is, a metabolite that is outside a cell that can be transported in or out of a cell. By convention, the role of an external reaction is to remove metabolites from the metabolic network, so an influx of a given metabolite into the network carries a negative flux. You may have already noticed that external reactions are necessary for FBA to work – metabolites have to be introduced somewhere and taken out of the system somewhere else if there is to be flux with unchanging metabolite concentrations.

For that very reason, external reactions are a convenient way to specify the kind of environment in which we want to simulate cell growth. By changing the lower bounds of an exchange reaction (remember influx carries a negative flux!) you can define what metabolites are available to a cell. Note that external reactions are different from transport reactions as the latter allow import/export of metabolites into the cell. As you get to more complex metabolic networks, you will see that cells often have multiple ways of taking up or excreting metabolites. For example, in the toy model (Figure 1) you may have noticed that there are two pathways for the uptake of metabolite B_e. If you wanted to change how much of B_e the cell takes up, you don't need to change the bounds of both transport reactions R7 and R16; you can simply change the bounds of the external reaction and leave it to FBA to decide which pathway is used for the uptake of B_e.

Flux balance analysis

FBA is a method to predict the flux of material through a metabolic network that maximizes the flux through a target reaction given two kinds of constraints:

- The first one arises from the relative proportions of reactants and products in chemical reactions (given in the stoichiometric matrix)
- the other is how much flux a reaction is allowed to carry (reaction bounds).

FBA is a very powerful tool in the analysis of metabolic networks, and also computationally efficient even for genome-scale models, as it assumes the network is in steady state – metabolites do not accumulate, they are produced as fast as they are consumed and consumed as fast as they are produced.

Flux balance analysis needs

1. a list of chemical reactions known to be catalyzed by enzymes in a given organism
2. Information about nutrients in the chemical environment of a cell and their uptake rate (usually in mol/g dry weight [DW] and hour)

Flux balance analysis computes

1. allowable metabolic fluxes through a metabolic network (fluxes that do not violate the law of mass conservation)
2. within the set of allowable fluxes, those that have desirable properties (e.g., maximal rate of biomass production, maximal biomass yield per unit of carbon source.

Example questions for flux balance analysis

1. Can a given organism (metabolism) survive in environment X?
2. How fast could it grow in this environment?
3. Why are many enzymatic reactions dispensable in any one environment?
4. Why do some metabolisms have many reactions, while others have few?
5. Does network function and flux influence network evolution?
6. Is it possible to design "resistance-proof" antimetabolic drugs?

Stoichiometric matrix

Stoichiometric coefficients represent the relative molar amounts of reactants (educts, substrates) and products participating in a chemical reaction. For example, in the reaction A + 2 B -> C, the stoichiometric coefficient of A is 1, that of B is 2 and that of C is 1. To denote whether a molecule is a reactant or a product of a reaction, we add a sign to its stoichiometric coefficient. Thus, the stoichiometry of A is -1, the stoichiometry of B is -2 and that of C is 1.

Essential reactions: Finding reactions that must be there

A reaction is essential for the synthesis of a molecule from a specific substrate if its removal makes the synthesis of that molecule from the substrate impossible. As an example, consider the reactions involved in the synthesis of metabolite O from substrate B_e in figure 1. Synthesis of O from B_e requires reactions R8 and R9, but not R7 and R15, as these can be replaced by R16. If we were to delete either R8 or R9, it would be impossible to synthesize O from B_e. We therefore say that reactions R8 and R9 are essential for the production of O from substrate B_e.

Active reactions: Finding reactions in flux

Active reactions are reactions that have a non-zero metabolic flux, that is, reactions proceeding at a rate different from zero. It is NOT the upper and/or lower bound of a reaction that determines whether it is active, but the effective flux of the reaction in the FBA solution. Therefore, you can only know whether a reaction is active or not after performing FBA.

Flux Variability Analysis (FVA): Finding alternative pathways

FBA solutions are rarely unique. For example, think about the active pathway you found in exercise 1.6 that produces X_e from A_e. Is this the only possible solution that maximizes production of X_e? Pay particular attention to reactions R3 and R4. In general, we find that at least some reactions can take a whole range of fluxes without affecting the flux through the objective function. Metabolic networks are thus "flux variable", and FVA can provide some insights into the extent of the network's flux variability. FVA computes the minimum and maximum value of the flux through a reaction while keeping a given objective, such as biomass synthesis, unchanged, thus estimating the range of possible fluxes through a reaction (the reaction's flux variability).

Robinson

It's just data

High-throughput sequencing

Gene Expression Profiling: questions of interest

- What genes have changed in expression? (e.g. between disease/normal, affected by treatment)
  - Gene discovery, differential expression
- Is a specified group of genes all up-regulated in a particular condition?
  - Gene set differential expression
- Can the expression profile predict outcome?
  - Class prediction, classification
- Are there tumour sub-types not previously identified? Do my genes group into previously undiscovered pathways?
  - Class discovery, clustering
- (with single cell gene expression): What changes in cell type composition are observed? What genes have changed in expression in a given subtype of cells?

Single cell

Bulk vs Single Cell RNA Sequencing (scRNA-seq):

- Bulk RNA-seq: average expression level
  - comparative transcriptomics
  - disease biomarker
  - homogenous systems
- scRNA-seq: expression level of each single cell
  - define heterogeneity
  - identify rare cell population
  - cell population dynamics
  - finding cell subpopulation-specific changes in state

Cytometry

Flow cytometry

In this process, a sample containing cells or particles is suspended in a fluid and injected into the flow cytometer instrument. The sample is focused to ideally flow one cell at a time through a laser beam, where the light scattered is characteristic to the cells and their components. Cells are often labeled with fluorescent markers so light is absorbed and then emitted in a band of wavelengths. Tens of thousands of cells can be quickly examined and the data gathered are processed by a computer.

- Cell counting
- Cell sorting
- Determining cell characteristics and function
- Detecting microorganisms
- Biomarker detection
- Protein engineering detection
- Diagnosis of health disorders such as blood cancers

Mass cytometry

Mass cytometry is a mass spectrometry technique based on inductively coupled plasma mass spectrometry and time of flight mass spectrometry used for the determination of the properties of cells (cytometry). In this approach, antibodies are conjugated with isotopically pure elements, and these antibodies are used to label cellular proteins. Cells are nebulized and sent through an argon plasma, which ionizes the metal-conjugated antibodies. The metal signals are then analyzed by a time-of-flight mass spectrometer. The approach overcomes limitations of spectral overlap in flow cytometry by utilizing discrete isotopes as a reporter system instead of traditional fluorophores which have broad emission spectra.

- Finding molecular biomarkers associated with drug response
- Differential abundance of cell populations

Hierarchical Clustering

- Divisive: This is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.
- Agglomerative: This is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

Metric: The choice of an appropriate metric will influence the shape of the clusters, as some elements may be relatively closer to one another under one metric than another. For example, in two dimensions, under the Manhattan distance metric, the distance between the origin (0,0) and (0.5, 0.5) is the same as the distance between the origin and (0, 1), while under the Euclidean distance metric the latter is strictly greater.

Linkage: The linkage criterion determines the distance between sets of observations as a function of the pairwise distances between observations.

Dimension reduction

- Many types of data come as a matrix of N samples (e.g., cells, patients) x G features (e.g., genes, proteins)
- Each sample is a point in G-dimensional space
- Goal: represent the data in 2-3 dimensions, but preserve structure as best as possible (i.e., points that are close in G dimensions should be close in 2 or 3 dimensions)

PCA: Principal Component Analysis (PCA) is a statistical method for dimensionality reduction.

- Form successive linear combinations of the features that are: orthogonal, ordered by variance