Mathematical Genomics
Génomique mathématique

(Org: **David Sankoff** (University of Ottawa))

TOM BOOTHBY, Simon Fraser University
Graphs on Surfaces and Computational Genomics

In topological graph theory, we consider ways of drawing graphs on surfaces without crossing lines. Attempting to draw permutation chord diagrams on orientable surfaces, we find a surprise: the 'genus' of a permutation is equal to the 'block interchange distance' from that permutation to the identity. Further study reveals that a similar notion of nonorientable genus for signed permutations, this time the 'double cut and join' (DCJ) distance appears. We discuss this correspondence, and how some familar results on DCJ factorizations appear from a topological point of view.

KELLY BURKETT, University of Ottawa
Gene genealogies for finding disease-predisposing genetic variants

The gene genealogy describes relationships among sequences sampled from a population. The gene genealogy has been incorporated in approaches to estimate population genetic parameters, like mutation or recombination rates, based on sampled DNA sequences. Knowledge of the underlying genealogy also has potential application in the discovery of disease-predisposing genetic variants. In particular, individuals inheriting the same phenotype-influencing genetic variant are more closely related to each other than individuals not carrying the same variant and they should also share a similar phenotype. We are therefore interested in quantifying the degree to which individuals sharing the same phenotype are clustered in the ancestral tree at the location of the variant that influences the phenotype. In this presentation, I discuss a number of possible statistics for measuring the proximity in the tree of individuals sharing similar phenotype values. I use simulation to show how the proposed statistics can be used to find regions harbouring disease-predisposing variants, with an emphasis on rare genetic variants. Since the true ancestral trees are unknown, I also discuss an approach to sample trees that are compatible with a sample of observed sequences.

NADIA EL-MABROUK, University of Montreal
Resolving Gene Trees with Polytomies

Accurate gene tree reconstruction is a fundamental problem in phylogenetics with many important applications. By reconciling a gene tree with a specie tree, we infer the history of duplications and losses that have shaped the gene family, which reveals the orthology/paralogy relationship between gene copies. However, due to various limitations such as insufficient differentiation between gene sequences, alignment ambiguity or inconsistency with gene order or other genome-level information, it is often difficult to support a single gene tree topology with high confidence. In this case, it may be more appropriate to collapse weakly supported internal nodes or remove dubious nodes, leading to a non-binary gene tree, with polytomies representing non-resolved parts of the tree. The question is then to resolve polytomies, based on appropriate criteria. In this presentation, I will review various optimization criteria for inferring a most parsimonious resolution of a polytomy in term of duplications and losses induced by the reconciliation with a given binary gene tree. I will more specifically deal with the problem of finding a resolution leading to a minimum number of non-apparent duplications, which are those annotated as dubious in the Ensembl database. I will present algorithmic and complexity results for this problem, based on interesting properties on the graph representing the speciation and duplication relationships between the leafs of the polytomy.

PEDRO FEIJAO, Bielefeld University
An Algebraic Theory for Genome Rearrangements

Genome rearrangements are evolutionary events where large, continuous pieces of the genome shuffle around, changing the order of genes in the genome of a species. Gene order data may be useful in estimating the evolutionary distance among genomes, and also in reconstructing the gene order of ancestral genomes.

In 2000, Meidanis and Dias proposed a framework for studying rearrangement problems, called *Algebraic Formalism*, based on permutation groups to model genomes and rearrangement operations. In its original formulation, it focuses on representing the order in which genes appear in chromosomes, and applies to circular chromosomes only.

Recently, Feijão and Meidanis introduced an extension of this formalism, called the *Adjacency Algebraic Theory*, where permutations represent the adjacencies between genes in a genome. This allowed the algebraic theory to model linear chromosomes and the use of the original algebraic distance formula in the general multichromosomal case, with both linear and circular chromosomes. It was also shown that there is a direct relationship between the original model (called *chromosomal*) and the adjacency model.

This resulting algebraic rearrangement distance is very similar, but not quite the same, to the Double-Cut-and-Join distance, a well known comprehensive model of genome rearrangents.

This talk is intended as an introduction to algebraic concepts used in genome rearrangement problems, where I will present the main ideas of the Algebraic Theory and some of its most recent developments.

RICHARD FRIEDBERG, Columbia University
Diffusion on the DCJ Lattice

In the DCJ (double cut and join) model of genome rearrangement, the segment ends can be treated as independent entities so that a rearrangement is defined by how these 2N entities are paired at synapses. A DCJ step is made by cutting any two pairs and reconnecting the 4 cut ends. Each possible rearrangement corresponds to one of the (2N-1)!! complete pairings. Treating each pairing as a graph vertex and the DCJ operation as specifying the graph edges, one can study diffusion on the graph by random DCJ operations. The diffusion equation is

$$\frac{d}{dt}P_\alpha = -P_\alpha + M^{-1}\Sigma_{\beta\wedge\alpha}P_\beta \tag{1}$$

where P_α is the probability of being at site α, $\wedge$ denotes DCJ neighbors, and $M = N(N-1)$ is the number of neighbors. Starting with all probability at site 0, we seek the probability of being at any site α at time t.

Symmetries of the DCJ lattice are inherited from S_{2N} acting on the segment ends. Eigenfunctions of $-d/dt$ belong to a set A of the irreps of S_{2N}. The eigenvalue corresponding to the irrep R is $\Lambda_R = (1-\lambda_R(2))(2N-1)/(2N-2)$ where $\lambda_R(2)$ is the character of single pair exchange in R. Thus

$$P_\alpha(t) = \Sigma_R^A e^{-\Lambda_R t} P_\alpha^R(0) \tag{2}$$

where $\Sigma_R^A P_\alpha^R(0) = \delta_{0\alpha}$. The bulk of the calculation consists in finding the numbers $P_\alpha(0)$.

ARASH JAMSHIDPEY, University of Ottawa
Stochastic DCJ jump process on signed permutation groups and the validity of the median as an approximation to the true ancestor

A genome can be represented by its set of gene adjacencies. Each gene is denoted by a signed number where the sign indicates the orientation. A double-cut-and-join (DCJ) operation acts on two adjacencies of the genome, xy and wz, where x, y, w, z are heads or tails of some genes, cuts these adjacencies and rejoins the four points in two possible ways, either xw, yz or xz, yw. The dcj distance between two genomes is the minimum number of DCJs necessary to convert one into the other. A DCJ jump process on the space of all multichromosomal genomes is a continuous time version of a DCJ random walk where at random Poisson times we uniformly randomly choose one of all possible DCJs and let it act on the current genome state.

Suppose we have k independent DCJ jump processes, $X_t^1, ..., X_t^k$, all starting at the same genome X_0. Our goal is to study the median value of $G_t := \{X_t^1, ..., X_t^k\}$, namely $\min_M \sum_i dcj(M, X_t^i)$. We prove that as the number of genes n goes to ∞, the median value approximates the total divergence time kt as well as the total dcj distance of the true ancestor to G_t, namely $\sum_i dcj(X_0, X_t^i)$, if the number of rearrangements $t < n/4$. Furthermore, we investigate when this approximation does not hold for $t > n/4$. We make use of a new algebraic representation of DCJ which enables us to show that the state space of the process has a group structure.

MEGAN OWEN, University of Waterloo
Statistics in tree space

We introduce new notions of mean and variance for a set or distribution of phylogenetic trees. These definitions of mean and variance are analogous to those for a weighted set of points in Euclidean space, but with the underlying space being the space of phylogenetic trees constructed by Billera, Holmes, and Vogtmann (2001). A property of this space (non-positive curvature) ensures there is a unique shortest path between any two trees. Furthermore, this path can be computed in polynomial time, leading to a practical algorithm for computing the mean and variance. I will compare the mean and variance to existing consensus tree and summary methods, as well as present applications to such biological problems as the reconstruction of phylogenetic trees and the classification of lung airway scans.

LAXMI PARIDA, IBM Research
Random Graphs in Population Genomics

The modeling of the evolutionary dynamics of evolving populations as random graphs offers a new methodology for analysis. This exploration begins as a quest for understanding the reconstructability of common evolutionary history of populations. It provides new insights including a purely topological (or graph theoretic definition) of traditional population genomic entity like the GMRCA (Grand Most Common Ancestor) of individuals under mutations as well as recombinations. Apart from giving interesting characterizations of another important structure called the ARG (Ancestral Recombinations Graph), it provides the basis for identifying a mathematical minimal nonredundant structure in the ARG and for adapting the coalescence theory (a well-studied notion in population genetics) very naturally in designing ARG sampling algorithms. This connection also opens the door for many interesting questions ranging from human migration paths, to genetic diversity study in plant (cacao) cultivars.

MARIBEL HERNANDEZ ROSALES, University of Leipzig
Cographs: A Mathematical Characterization for Valid Orthology Relations

The divergence of all the genes that descended from a single gene in an ancestral species can be represented as a tree, a gene tree that takes into account both speciation and duplication events. Orthology refers specifically to the relationship between two genes that arose by a speciation event, recent or remote. Comparing orthologous genes is essential to the correct reconstruction of species trees, so that detecting and identifying orthologous genes is an important problem in comparative and evolutionary genomics as well as phylogenetics. In this work, we look at the connection of trees and orthology by trying to answer the following question: How much information about the gene tree, the species tree, and their reconciliation is already contained in the orthology relation among genes? A solution to the first part of this question has already been given by Boecker and Dress in 1998 in a different context. In particular, they completely characterized certain maps which they called symbolic ultrametrics. Semple and Steel [2003] then presented an algorithm that can be used to reconstruct a phylogenetic tree from any given symbolic ultrametric. In this work we investigate a new characterization of orthology relations, based on symbolic ultramterics for recovering the gene tree. Surprisingly, symbolic ultrametrics are very closely related to cographs, graphs that do not contain an induced path on any subset of four vertices. We will show that the tree corresponding to a symbolic ultrametric can also be recovered using cotrees, trees that can be canonically associated to cographs.