Part IV

Phylogenomics and Population Genomics
MOLECULAR EVOLUTION AND PHYLOGENETICS

Scribed by Albert Wang and Mashaal Sohail (2010)
Guo-Liang Chew and Sara Baldwin (2009)

Figures

22.1 Evolutionary History of Life ........................................................................ 238
   (a) Tree of Life ......................................................................................... 238
   (b) Divergence and Extinction of Dinosaurs .............................................. 238
22.2 Defining tree terminology. A tree of branching nodes is depicted with leaves on one side and the root on the other. ......................................................... 238
22.3 Three types of trees. .................................................................................. 239
22.4 The two steps of distance based phylogenetic reconstruction. .................. 241
22.5 Markov chain accounting for back mutations ............................................ 241
   (a) 4 States ............................................................................................. 241
   (b) 2 States ............................................................................................. 241
22.6 The y axis denotes probability of observing the bases - A(red), others(green). x axis denotes time. ................................................................. 242
22.7 Fraction of altered bases (x axis) versus the Jukes Cantor distance(y axis). Black line denotes the curve, green is the trend line for small values of f while the red line denotes the asymptotic boundary. ......................................................... 243
22.8 Distance models of varying levels of complexity(parameters). .................... 244
22.9 Mapping from a tree to a distance matrix and vice versa ............................. 245
22.10 Ultrametric distances. .............................................................................. 245
22.11 Additive distances. .................................................................................. 246
22.12 UPGMA / Hierarchical Clustering ............................................................ 247
22.13 UPGMA fails to find the correct tree in this case ...................................... 247
22.14 An overview of the character based methods ........................................... 248
22.15 Parsimony scoring: union and intersection ............................................. 249
22.16 Parsimony traceback to find ancestral nucleotides .................................... 250
22.17 Parsimony scoring by dynamic programming ......................................... 250
22.18 A tree to be scored using the peeling algorithm. n=4 ............................... 251
   (a) Filling the matrix .............................................................................. 251
   (b) Initialization and direction ................................................................. 251
22.19 The recurrence ........................................................................................ 251
   (c) Before .............................................................................................. 253
   (d) After .............................................................................................. 253
22.1 Introduction

Phylogenetics is the study of relationships among a set of objects having a common origin, based on the knowledge of the individual traits of the objects. Such objects may be species, genes, or languages, and their corresponding traits may be morphological characteristics, sequences, words etc. In all these examples the objects under study change gradually with time and diverge from common origins to present day objects.

In Biology, phylogenetics is particularly relevant because all biological species happen to be descendants of a single common ancestor which existed approximately 3.5 to 3.8 billion years ago. Throughout the passage of time, genetic variation, isolation and selection have created the great variety of species that we observe today. Not just speciation however, but extinction has also played a key role in shaping the biosphere as we see today. Studying the ancestry between different species is fundamentally important to biology because they shed much light in understanding different biological functions, genetic mechanisms as well as the process of evolution itself.

![Phylogenetic Tree of Life](image)

(a) Tree of Life
(b) Divergence and Extinction of Dinosaurs

Figure 22.1: Evolutionary History of Life

22.2 Basics of Phylogeny

22.2.1 Trees

The ancestry relationships between different objects are represented in the form of a tree, where each node represents a divergence event between two ancestral lineages, the leaves denote the set of present objects and the root represents the common ancestor.

![Defining tree terminology](image)

Figure 22.2: Defining tree terminology. A tree of branching nodes is depicted with leaves on one side and the root on the other.
However, sometimes more information is reflected in the branch lengths, such as time elapsed or the amount of dissimilarity. According to these differences, biological phylogenetic trees may be classified into three categories:

**Cladogram:** gives no meaning to branch lengths; only the sequence and topology of the branching matters.

**Phylogram:** Branch lengths are directly related to the amount of genetic change. The longer the branch of a tree, the greater the amount of phylogenetic change that has taken place. The leaves in this tree may not necessarily end on the same vertical line, due to different rates of mutation.

**Chronogram (ultrametric tree):** Branch lengths are directly related to time. The longer the branches of a tree, the greater the amount of time that has passed. The leaves in this tree necessarily end on the same vertical line (i.e. they are the same distance from the root), since they are all in the present unless extinct species were included in the tree. Although there is a correlation between branch lengths and genetic distance on a chronogram, they are not necessarily exactly proportional because evolution rates / mutation rates were not necessarily the same for different species at different time periods.

![Three types of trees](image)

Figure 22.3: Three types of trees.

### 22.2.2 Occam’s Razor

The principle of Occam’s Razor recommends that among multiple possible hypotheses that may explain an observation, the **simplest**, i.e. the one making the least amount of assumptions should be accepted. This principle is very important in all methods of tree-reconstruction methods because there are often multiple possible ancestries of the same observed data. In that case Occam’s razor may be applied to find the most plausible one. However, this is an approximate principle that do not always work in practice. Sometimes the same trait may emerge independently in two lineages, in which case applying Occam’s razor assigns them to be more closely related than they really are. It will be discussed later how this issue is taken into consideration.
22.2.3 Traits

All methods for tree-reconstruction rely on studying the occurrence of different traits in the given objects. In traditional phylogenetics the morphological data of different species were used for this purpose. In modern methods, genetic sequence data is used instead. While the modern approach approach generates much more amount of data to be analyzed, it has the drawback that back-mutations and independent concurrent mutations are very frequent due to a small alphabet size, consequently defying our reliance on Occam’s razor. Basically the challenges of the modern approach is to reconcile the the signals of a large number of ill-behaved traits as opposed to that of a small number of well-behaved traits in the traditional approach. The rest of the chapter will focus principally on tree-building from gene sequences.

Since this approach deals with comparing between pairs of genes, it is useful to understand the concept of homology: A pair of genes are called paralogues if they diverged from a duplication event, and orthologues if they diverged from a speciation event.

<table>
<thead>
<tr>
<th>FAQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q: Would it be possible to use extinct species’ DNA sequences?</td>
</tr>
<tr>
<td>A: Current technologies only allow for usage of extant sequences. However, there have been a few successes in using extinct species’ DNA. DNA from frozen mammoths have been collected and are being sequences but due to DNA breaking down over time and contamination from the environment, it is very hard to extract correct sequences.</td>
</tr>
</tbody>
</table>

22.2.4 Methods for Tree Reconstruction

The modern methods for inferring ancestry from sequence data can be classified into two approaches, distance based methods and character based methods.

Distance based approaches take two steps to solve the problem, i.e. to quantify the amount of mutation that separates each pair of sequences (which may or may not be proportional to the time since they have been separated) and to fit the most likely tree according to the pair-wise distance matrix. The second step is usually a direct algorithm, based on some assumtions, but may be more complex.

Character based approaches instead try to find the tree that best explains the observed sequences. As opposed to direct reconstruction, these methods rely on tree proposal and scoring techniques to perform a heuristic search over the space of trees.
22.3 Distance Based Methods

The distance based models sequester the sequence data into pairwise distances. This step loses some information, but sets up the platform for direct tree reconstruction. The two steps of this method are hereby discussed in detail.

22.3.1 From alignment to distances

The naive way to interpret the separation between two sequences may be simply the number of mismatches. While this does provide us a distance metric (i.e. \( d(a, b) + d(b, c) \geq d(a, c) \)) this does not quite satisfy our requirements, because we want additive distances, i.e. those that satisfy \( d(a, b) + d(b, c) = d(a, c) \) for a path \( a \rightarrow b \rightarrow c \) of evolving sequence, because the amount of mutations accumulated along a path in the tree should be the sum of that of its individual components. However, the naive mismatch fraction do not always have this property, because this quantity is bounded by 1, while the sum of individual components can easily exceed 1.

The key to resolving this paradox is back-mutations. When a large number of mutations accumulate on a sequence, not all the mutations introduce new mismatches, some of them may occur on already mutated base pair, resulting in the mismatch score remaining the same or even decreasing. For small mismatch-scores however, this effect is statistically insignificant, because there are vastly more identical pairs than mismatching pairs. However, for sequences separated by longer evolutionary distance, we must correct for this effect. The Jukes-Cantor model is one such simple markov model that takes this into account.

Jukes-Cantor distances

To illustrate this concept, consider a nucleotide in state ’A’ at time zero. At each time step, it has a probability 0.7 to retain its previous state and probability 0.1 to transition to each of the other states. The probability \( P(B|t) \) of observing state (base) \( B \) at time \( t \) essentially follows the recursion

\[
P(B|t+1) = 0.7P(B|t) + 0.1 \sum_{b \neq B} P(b|t) = 0.1 + 0.6P(B|t)
\]

(a) 4 States (b) 2 States

Figure 22.5: Markov chain accounting for back mutations
If we plot $P(B|t)$ versus $t$, we observe that the distribution starts off as concentrated at the state ‘A’ and gradually spreads over to the rest of the states, eventually going towards an equilibrium of equal probabilities.

<table>
<thead>
<tr>
<th>time:</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0.7</td>
<td>0.52</td>
<td>0.412</td>
<td>0.3472</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0.1</td>
<td>0.16</td>
<td>0.196</td>
<td>0.2196</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0.1</td>
<td>0.16</td>
<td>0.196</td>
<td>0.2196</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0.1</td>
<td>0.16</td>
<td>0.196</td>
<td>0.2196</td>
</tr>
</tbody>
</table>

The essence of the Jukes Cantor model is to backtrack $t$, the amount of time elapsed from the fraction of altered bases. Conceptually, this is just inverting the x and y axis of the green curve. To model this quantitatively, we consider the following matrix $S(t)$ which denotes the respective probabilities $P(x|y, \Delta t)$ of observing base $x$ given a starting state of base $y$ in time $\Delta t$.

$$S(\Delta t) = \begin{pmatrix}
P(A|A, \Delta t) & P(A|G, \Delta t) & \cdots & P(A|T, \Delta t) \\
P(G|A, \Delta t) & \cdots & & \cdots \\
P(T|A, \Delta t) & \cdots & & P(T|T, \Delta t)
\end{pmatrix}$$

We can assume this is a stationary markov model, implying this matrix is multiplicative, i.e.

$$S(t_1 + t_2) = S(t_1)S(t_2)$$

For a very short time $\epsilon$, we can assume that there is no second order effect, i.e. there isn’t enough time for two mutations to occur at the same nucleotide. So the probabilities of cross transitions are all proportional to $\epsilon$. Further, in Jukes Cantor model, we assume that all the transition rates are same from each nucleotide to another nucleotide. Hence, for a short time $\epsilon$

$$S(\epsilon) = \begin{pmatrix}
1 - 3\alpha\epsilon & \alpha\epsilon & \alpha\epsilon & \alpha\epsilon \\
\alpha\epsilon & 1 - 3\alpha\epsilon & \alpha\epsilon & \alpha\epsilon \\
\alpha\epsilon & \alpha\epsilon & 1 - 3\alpha\epsilon & \alpha\epsilon \\
\alpha\epsilon & \alpha\epsilon & \alpha\epsilon & 1 - 3\alpha\epsilon
\end{pmatrix}$$
At time $t$, the matrix is given by

$$ S(t) = \begin{pmatrix} r(t) & s(t) & s(t) & s(t) \\ s(t) & r(t) & s(t) & s(t) \\ s(t) & s(t) & r(t) & s(t) \\ s(t) & s(t) & s(t) & r(t) \end{pmatrix} $$

From the equation $S(t + \epsilon) = S(t)S(\epsilon)$ we obtain

$$ r(t + \epsilon) = r(t)(1 - 3\alpha\epsilon) + 3\alpha s(t) \quad \text{and} \quad s(t + \epsilon) = s(t)(1 - \alpha\epsilon) + \alpha r(t) $$

Which rearrange as the coupled system of differential equations

$$ r'(t) = 3\alpha(-r(t) + s(t)) \quad \text{and} \quad s'(t) = \alpha(r(t) - s(t)) $$

With the initial conditions $r(0) = 1$ and $s(0) = 0$. The solutions can be obtained as

$$ r(t) = \frac{1}{4}(1 + 3e^{-4\alpha t}) \quad \text{and} \quad s(t) = \frac{1}{4}(1 - e^{-4\alpha t}) $$

Now, in a given alignment, if we have the fraction $f$ of the sites where the bases differ, we have:

$$ f = 3s(t) = \frac{3}{4}(1 - e^{-4\alpha t}) $$

implying

$$ t \propto -\log\left(1 - \frac{4f}{3}\right) $$

To agree asymptotically with $f$, we set the evolutionary distance $d$ to be

$$ d = -\frac{3}{4} \log\left(1 - \frac{4f}{3}\right) $$

Note that distance is approximately proportional to $f$ for small values of $f$ and asymptotically approaches infinity when $f \to 0.75$. Intuitively this happens because after a very long period of time, we would expect the sequence to be completely random and that would imply about three-fourth of the bases mismatching with original. But the uncertainty values of the Jukes-Cantor distance also becomes very large when $f$ approaches 0.75.
Other Models

The Jukes Cantor model is the simplest model that gives us theoretically consistent additive distance model. However, it is a one-parameter model that assumes that the mutations from each base to a different base has the same chance. But, changes between AT or between GC are more likely than changes across them. The first type of substitution is called transitions while the second type is called transversions. The Kimura model has two parameters which take this into account. There are also many other modifications of this distance model that takes into account the different rates of transitions and transversions etc. that are depicted below.

![Distance models of varying levels of complexity](image)

Models also exist for peptides and codons

**FAQ**

Q: Can we use different parameters for different parts of the tree? To account for different mutation rates?

A: It is possible, it is a current area of research.
22.3.2 Distances to Trees

If we have a weighted phylogenetic tree, we can find the total weight (length) of the shortest path between a pair of leaves by summing up the individual branch lengths in the path. Considering all such pairs of leaves, we have a distance matrix representing the data. In distance based methods, the problem is to reconstruct the tree given this distance matrix.

### FAQ

**Q:** In Figure 22.9 The m and r sequence divergence metrics can have some overlap so distance between mouse and rat is not simply m+r. Wouldn’t that only be the case if there was no overlap?

**A:** If you model evolution correctly, then you would get evolutionary distance. Its an inequality rather than an equality and I agree that you cant exactly infer that the given distance is the precise distance. Therefore, the sequences distance between mouse and rat is probably less than m + r because of overlap, convergent evolution, and transversions.

However, note that there is not a one to one correspondence between a distance matrix and a weighted tree. Each tree does correspond to one distance matrix, but the opposite is not always true. A distance matrix has to satisfy additional properties in order to correspond to some weighted tree. In fact there are two models that assume special constraints on the distance matrix:

**Ultrametric:** For all triplets (a, b, c) of leaves, 2 pairs among them have equal distance, and the third distance is smaller. i.e. the triplet can be labelled i,j,k such that

\[ d_{ij} \leq d_{ik} = d_{jk} \]

Conceptually this is because the two leaves that are more closely related (say i, j) have diverged from the third (k) at exactly the same time, and the time separation from the third should be equal, whereas the separation between themselves should be smaller.

**Figure 22.10: Ultrametric distances.**
Additive: Additive distance matrices satisfy the property that all quartet of leaves can be labelled \(i, j, k, l\) such that

\[
d_{ij} + d_{kl} \leq d_{ik} + d_{jl} = d_{il} + d_{jk}
\]

This is in fact true for all positive-weight trees. For any 4 leaves in a tree, there can be exactly one topology, i.e.

![Figure 22.11: Additive distances.](image)

Then the above condition is term by term equivalent to

\[
(a + b) + (c + d) \leq (a + m + c) + (b + m + d) = (a + m + d) + (b + m + c)
\]

This type of redundant equalities must occur while mapping a tree to a distance matrix, because, a tree of \(n\) nodes has \(n - 1\) parameters, one for each branch length, on the other hand, a distance matrix has \(n^2\) parameters. Hence, a tree is essentially a lower dimensional projection of a higher dimensional space. A corollary of this observation is that not all distance matrices have a corresponding tree, but all trees map to an unique distance matrix.

However, real datasets do not exactly satisfy either untrameric or additive constraints. This can be due to noise (when our parameters for our evolutionary models are not precise), stochasticity and randomness (due to small samples), fluctuations, different rates of mutations, gene conversions and horizontal transfer. Because of this, we need tree-building algorithms that are able to handle noisy distance matrices.

Next, two algorithms that directly rely on these assumptions for tree reconstruction will be discussed.
UPGMA - Unweighted Pair Group Method with Arithmetic Mean

This is exactly same as the method of Hierarchal clustering discussed in Lecture 13, Gene Expression Clustering. It forms clusters step by step from the closest related ones to families separated further apart. And forms a branching node for each successive level of coalition. The algorithm can be described properly by the following steps:

**Initialization:**
1. Define one leaf $i$ per sequence $x_i$.
2. Place each leaf $i$ at height 0.
3. Define Clusters $C_i$ each having one leaf $i$.

**Iteration:**
1. Find the pairwise distances $d_{ij}$ between each pairs of clusters $C_i, C_j$ by taking the arithmetic mean of the distances between their member sequences.
2. Find two clusters $C_i, C_j$ s.t. $d_{ij}$ is minimum.
3. Let $C_k = C_i \cup C_j$.
4. Define node $k$ as parent of nodes $i, j$ and place it at height $d_{ij}/2$ above $i, j$.
5. Delete $C_i, C_j$.

**Termination:** When two clusters $C_i, C_j$ remain, place the root at height $d_{ij}/2$ as parent of the nodes $i, j$.

**Weaknesses of UPGMA**

Although this method is guaranteed to find the correct tree if the distance matrix obeys the ultrameric property, it turns out to be an inaccurate algorithm in practice. Apart from lack of robustness, it suffers from the molecular clock assumption that the mutation rate over time is constant for all species. However this is not true as certain species such as rat, mouse evolve much faster than others. The following figure illustrates an example where UPGMA fails:
Neighbor Joining

the neighbor joining method is guaranteed to produce the correct tree if the distance matrix satisfies the additive property. It may also produce a good tree when there is some noise in the data. The algorithm is described below:

Finding the neighboring leaves: Let

\[ D_{ij} = d_{ij} - (r_i + r_j) \] where \( r_i = \frac{1}{n-2} \sum_k d_{ik} \)

Here \( n \) is the number of nodes in the tree hence \( r_i \) is the average distance of a node to the other nodes. It can be proved that the above modification ensures that \( D_{ij} \) is minimal only if \( i, j \) are neighbors. (A proof can be found in page 189 of Durbin’s book).

Initialization: Define \( T \) to be the set of leaf nodes, one per sequence. Let \( L = T \)

Iteration:
1. Pick \( i, j \) such that \( D_{ij} \) is minimal.
2. Define a new node \( k \), and set \( d_{km} = \frac{1}{2}(d_{im} + d_{jm}d_{ij}) \) for all \( m \in L \)
3. Add \( k \) to \( T \), with edges of lengths \( d_{ik} = \frac{1}{2}(d_{ij} + r_i + r_j) \)
4. Remove \( i, j \) from \( L \)
5. Add \( k \) to \( L \)

Termination: When \( L \) consists of two nodes \( i, j \), and the edge between them of length \( d_{ij} \), add the root node as parent of \( i \) and \( j \).

22.4 Character Based Methods

The main framework of the character based model is to search through the tree space while scoring each instance to guide the search. It turns out that there exist fairly good algorithms for tree scoring, while searching is NP-Hard (Due to the large number of possible trees). We’ll first discuss tree scoring techniques.
22.4.1 Scoring

There are two main algorithms for tree scoring. The first one is directly based on Occam’s razor or parsimony. It is rather simple, but works fast. The second method is maximum likelihood method which finds the probability of observing the given data given a tree.

Parsimony

The main idea in this method is very simple. For each base pair, it assigns a reconstruction of it for each ancestral node (or a set of possibilities for the base pair) such that the number of substitutions is minimized. The number of necessary mutations is the score of the score. And the total score of the tree is the sum of scores for each base pair in the sequences. The goal is then to find a tree that minimizes the score.

For the reconstruction step, the basic idea is to assign a set of possible bases at each node. We already know the right ones at the leaves. So the algorithm scans bottom up and assigns a set of bases at each internal node based on its children. The following illustrates this algorithm in detail:

Figure 22.15: Parsimony scoring: union and intersection
6.047/6.878 Lecture 20: Molecular Evolution and Phylogenetics

Figure 22.16: Parsimony traceback to find ancestral nucleotides

Parsimony reconstruction is conceptually very simple and works fast. However, this does not always give us the right reconstruction because it ignores possibilities of back mutations, simultaneous mutations etc. Hence maximum likelihood methods are used more for tree scoring.

Maximum Likelihood - Peeling Algorithm

As with the general Maximum likelihood methods, this algorithm scores a tree according to the (log) probability of observing the data given the tree, i.e. $P(D,T)$. The peeling algorithm again considers individual base pairs and assumes that all sites evolve independently. Then the probability of observing the set of sequences is just the product of individual probabilities of observing the sets of individual base pairs. Another important difference with the parsimony approach is that it takes weighted trees for scoring instead of normal trees.
Here each node has a character $x_i$ and $t_i$ is the corresponding branch length from its parent. Note that we already know the values $x_1, x_2, \ldots, x_n$, so they are constants, but $x_{n+1}, \ldots, x_{n-1}$ are unknown characters at ancestral nodes which are variables. We want to compute $P(x_1, x_2, \ldots, x_n | T)$. For this we sum over all possible combinations of values at the ancestral nodes. This is called marginalization. In this particular example

$$P(x_1, x_2, x_3, x_4 | T) = \sum_{x_5} \sum_{x_6} \sum_{x_7} P(x_1, x_2, \ldots, x_7 | T)$$

There are $4^{n-1}$ terms in here, but we can use the following factorization trick:

$$= \sum_{x_7} \left[ P(x_7) \left( \sum_{x_5} P(x_5|x_7, t_5)P(x_1|x_5, t_1)P(x_2|x_5, t_2) \right) \left( \sum_{x_6} P(x_6|x_7, t_6)P(x_3|x_6, t_3)P(x_4|x_6, t_4) \right) \right]$$

Let $T_i$ be the subtree below $i$. In this case, our $2n - 1 \times 4$ dynamic programming array computes $L[i, b]$, the probability $P(T_i | x_i = b)$ of observing $T_i$, if node $i$ contains base $b$. Then we want to compute nothing but the probability of observing $T = T_{2n-1}$, which is

$$\sum_b P(x_{2n-1} = b)L[2n-1, b]$$

Note that for each ancestral node $i$ and its children $j, k$, we have

$$L[i, b] = \left( \sum_c P(c|b, t_j)L[j, c] \right) \left( \sum_c P(c|b, t_k)L[k, c] \right)$$

Subject to the initial conditions for the leaf nodes, i.e. for $i \leq n$:

$$L[i, b] = 1 \text{ if } x_i = b \text{ and } 0 \text{ otherwise}$$
Note that we still do not have the values $P(x_{2n-1} = b)$. It is usually assigned equally or from some prior distribution, but it does not affect the results greatly. The final step is of course to multiply all the probabilities for individual sites to obtain the probability of observing the set of entire sequences. The overall complexity of this algorithm is $O(nmk^2)$ where $n$ is the number of leaves (taxa), $m$ is the sequence length, and $k$ is the number of characters.

There are advantages and disadvantages of this algorithm. Such as

**Advantages:**

1. Inherently statistical and evolutionary model-based.
2. Usually the most consistent of the methods available.
3. Used for both character and rate analyses
4. Can be used to infer the sequences of the extinct ancestors.
5. Account for branch-length effects in unbalanced trees.
6. Nucleotide or amino acid sequences, other types of data.

**Disadvantages:**

1. Not as simple and intuitive as many other methods.
2. Computationally intense Limited by, number of taxa and sequence length).
3. Like parsimony, can be fooled by high levels of homoplasy.
4. Violations of model assumptions can lead to incorrect trees.

### 22.4.2 Search

In general, searching over the space of all trees is an extremely costly task. The number of full rooted trees with $n+1$ leaves is the $n$-th catalan number

$$C_n = \frac{1}{n+1} \binom{2n}{n} \approx \frac{4^n}{n^{3/2}\sqrt{\pi}}$$

Moreover, there exist consideration for the weights in Maximum Likelihood methods. So It is an NP-Hard probem to maximize the score absolutely for all trees. So Heuristic algorithms are used to search for good solutions in the tree space. The general framework for such search algorithms may be described as follows:

**Initialization:** First some tree is taken as the base of iteration (randomly or according to some other prior, or from the distance based direct algorithms).

**Proposal:** Propose a new tree by randomly modifying the current tree slightly.

**Score:** Score the new proposal according to the methods described above.

**Select:** Randomly select the new tree or the old tree (corresponding probabilities according to the score(likelihood) ratio.

**Iterate:** Repeat to proposal step unless some termination criteria is met (some thresold score or number of steps reached.

the basic idea here is the heuristic assumption that the scores of closely related trees are somewhat similar. So good solutions may be obtained by successive local optimization, which is expected to converge towards a overall good solution.
Tree Proposal

There are some methods for proposing a tree slightly different from one given. One such method is the Nearest Neighbour Exchange (NNI), illustrated below.

There is also another method called Tree Bisection and Join (TBJ) etc. The important criteria for such proposal rules is that:

(a) The tree space should be connected, i.e. any pair of trees should be obtainable from each other by successive proposals.

(b) An individual new proposal should be sufficiently close to the original. So that it is more likely to be a good solution by virtue of the proximity to an already discovered good solution. If individual steps are too big, the algorithm may move away from an already discovered solution (also depends on the selection step).

Selection

Analogous to Proposal, there is no fixed way to select a new proposal or retain the previous. The general rules of thumb is

1. If the new one has a better score, always accept it.

2. If it has a worse score, there should be some probability of selecting it, otherwise the algorithm will soon fixate in a local minima, ignoring better alternatives a little far away.

3. There should not be too much probability of selecting an worse new proposal, otherwise, it risks rejecting a known good solution.

It is the trade-off between the steps 2 and 3 that determines a good selection rule. Metropolis Hastings is a Markov Chain Monte Carlo Method (MCMC) that defines specific rules for exploring the state space in a way that makes it a sample from the posterior distribution. These algorithms work somewhat well in practice, but there is no guarantee for finding the appropriate tree. So a method known as bootstrapping is used, which is basically running the algorithm over and over using parts of the sequences is used. If the resulting trees tend to fit inside the original tree, it denotes a reliable tree.
22.5 Possible Theoretical and Practical Issues with Discussed Approach

A special point must be made about distances. Since distances are typically calculated between aligned gene sequences, most current tree reconstruction methods rely on heavily conserved genes, as non-conserved genes would not give information on species without those genes. This causes the ignoring of otherwise useful data. Therefore, there are some algorithms that try to take into account less conserved genes in reconstructing trees but these algorithms tend to take a long time due to the NP-Hard nature of reconstructing trees.

Additionally, aligned sequences are still not explicit in regards to the events that created them. That is, combinations of speciation, duplication, loss, and horizontal gene transfer (hgt) events are easy to mix up because only current DNA sequences are available. (see [9] for a commentary on such theoretical issues) A duplication followed by a loss would be very hard to detect. Additionally, a duplication followed by a speciation could look like an HGT event. Even the probabilities of events happening is still contested, especially horizontal gene transfer events.

Another issue is that often multiple marker sequences are concatenated and the concatenated sequence is used to calculate distance and create trees. However, this approach assumes that all the concatenated genes had the same history and there is debate over if this is a valid approach given that events such as hgt and duplications as described above could have occurred differently for different genes. [8] is an article showing how different phylogenetic relationships were found depending on if the tree was created using multiple genes concatenated together or if it was created using each of the individual genes. Conversely, additional [4] claims that while hgt is prevalent, orthologs used for phylogenetic reconstruction are consistent with a single tree of life. These two issues indicate that there is clearly debate in the field on a non arbitrary way to define species and to infer phylogenetic relationships to recreate the tree of life.

22.6 Towards final project

22.6.1 Project Ideas

1. Creating better distance models such as taking into account duplicate genes or loss of genes. It may also be possible to analyze sequences for peptide coding regions and calculate distances based on peptide chains too.

2. Creating a faster/more accurate search algorithm for turning distances into trees.

3. Analyze sequences to calculate probabilities of speciation, duplication, loss, and horizontal gene transfer events.

4. Extending an algorithm that looks for HGTs to look for extinct species. A possible use for HGTs is that if a program were to infer HGTs between different times, it could mean that there was a speciation where one branch is now extinct (or not yet discovered) and that branch had caused an HGT to the other extant branch.

22.6.2 Project Datasets

1. 1000 Genomes Project http://www.1000genomes.org/

2. Microbes Online http://microbesonline.org/

22.7 What Have We Learned?

In this chapter, we have learnt different methods and approaches for reconstructing Phylogenetic trees from sequence data. In the next chapter, its application in gene trees and species trees and the relationship between those two will be discussed, as well as modelling phylogenies among populations within a species and between closely related species.
Bibliography

[1] 1000 genomes project.


CHAPTER TWENTYTHREE

PHYLOGENOMICS II

Guest Lecture by
Matt Rasmussen
Scribed by Jerry Wang and Dhruv Garg

Figures

23.1 Heuristic tree search in character-based reconstruction .......................... 258
23.2 Scoring metric for heuristic tree search .............................................. 259
23.3 Site evolution uses the Jukes-Cantor model ......................................... 259
23.4 Sequence evolution is the product of site evolution ............................... 259
23.5 Sequence evolution over an entire tree. .............................................. 259
23.6 Peeling Algorithm ............................................................................. 260
23.7 Gene Family Evolution: Gene Trees and Species Trees .......................... 260
23.8 Maximum Parsimony Reconciliation (MPR) ......................................... 260
23.9 Maximum Parsimony Reconciliation Recursive Algorithm .................... 261
23.10 Using species trees to improve gene tree reconstruction ....................... 261
23.11 We can develop a model for what kind of branch lengths we can expect. We can use conserved gene order to tell orthologs and build trees. ............................. 261
23.12 Branch length can be modeled as two different rate components: gene specific and species specific ................................................................. 262
23.13 The Wright-Fisher model .................................................................. 262
23.14 The Wright-Fisher model continued over many generations and ignoring the ordering of chromosomes ................................................................. 262
23.15 The coalescent model ........................................................................ 263
23.16 Geometric probability distribution for coalescent events in k lineages .... 263
23.17 Multispecies Coalescent Model. Leaf branches track one lineage. There is a lag time from when population separated and when two actual gene lineages find a common ancestor. The rate of coalescent slows down as N gets bigger and for short branches. Deep coalescent is depicted in light blue for three lineages. The species and gene tree are incongruent since C and D are sisters in gene tree but not the species tree. There is a $\frac{4}{3}$ chance that incongruence will occur because once we get to the light blue section, the Wright-fisher is memory less and there is only $\frac{1}{3}$ chance that it will be congruent. Effect of incongruence is called incomplete lineage sorting ......................................................... 264
23.18 MPR reconciliation of genes and species tree. ..................................... 264
23.19 Inaccuracies in gene tree. .................................................................. 265
23.1 Introduction

Guest lecturer Matt Rasmussen, a former student of Manoliss presented our second phylogenomics lecture. The lecture finished explaining max likelihood methods for phylogenetics, and then progressed to more advanced uses of phylogenetics such as inferring orthologs, paralogs, gene duplication and gene loss. This led to learning across gene trees and modeling populations and allele frequencies.

In previous lectures, we studied various algorithms to obtain phylogenetic species trees. Similar studies can be performed to study phylogeny of gene families, or sets of orthologous and paralogous genes. Given multiply aligned sequences, several techniques discussed in previous lectures could be employed for constructing a gene tree, including nearest neighbor joining, and hierarchical clustering. If in addition to the aligned genes, we also have a species tree (which can often be taken as a given for sufficiently diverged species), then we should be able to formulate a consistent view of the evolutionary process; namely, we hope to map the gene tree onto the species tree. These mappings between the two trees are called reconciliations. The standard phylogenomic pipeline can be summarized as follows:

1. Blast protein sequences against each other to score similarities.
2. Use this metric to cluster genes into families of relatedness.
3. Build multiple alignments.
4. From the alignments, build gene trees.
5. Reconcile the gene tree to the species tree.

23.1.1 Phylogenetics

The two main pipelines for building trees are distance-based and character-based. Last lecture focused on distance-based pipeline. In distance-based, you form a pair-wise distance matrix using Jukes-Cantor or Kimura. Then use Neighbor Joining or UPGMA to reconstruct a tree from the matrix. Distance based pipelines use a fixed number of steps so and UPGMA and NJ have a running time is $O(n^3)$. However, there are flaws to this approach. Distance-based metrics are overly simplified and under measure the rate of mutation because a nucleotide that gets mutated back to its original form is counted as not mutated.

Today’s lecture focuses on the character-based pipeline, which is NP Hard so we have to resort to heuristics. The basic idea is we want to search through different trees and test each one. We start with an initial tree, then compute the probability/likelihood, then explore the tree space using methods such as nearest neighbor interchange (NNI), compute the score again, loop and then return the tree with the highest score as the answer. Using NNI, we can go to all trees in the tree space. The problem is that the tree space is very big (why it is NP hard).

![Figure 23.1: Heuristic tree search in character-based reconstruction](image)

For the scoring metric, we want to maximize

Using the Felsenstein peeling algorithm, we can efficiently compute $P(X|T, B)$ by building up a dynamic programming problem. We can first look at site evolution along a single branch, then build on that and
look at sequence evolution and then look at site evolution along an entire tree. Site evolution uses the Jukes-Cantor model and has the definition

\[
\begin{align*}
\mathbb{P}(x_1, x_2, x_3, x_4 | T, t) &= \sum_{x_5} \sum_{x_6} \sum_{x_7} P(x_1 | x_5, t_1) P(x_2 | x_5, t_1) P(x_3 | x_6, t_3) P(x_4 | x_6, t_4) \\
&= \sum_{x_7} P(x_7) [\sum_{x_5} P(x_5 | x_7, t_5) P(x_1 | x_5, t_1) P(x_2 | x_5, t_1)] [\sum_{x_6} P(x_6 | x_7, t_6) P(x_3 | x_6, t_3) P(x_4 | x_6, t_4)]
\end{align*}
\]

The Peeling algorithm builds a DP table. Each entry contains the probability of seeing the leaf data below node \(I\), given that node \(I\) has base \(a\) at site \(j\). The leaves of the table are initialized based on the observed sequence. Entries are populated in post-order traversal. The runtime of the Peeling algorithm is \(O(nm k^2)\). The Peeling algorithm scores one tree and we need to use the search algorithm to search for more trees. The runtime is for one tree while the entire runtime depends on how many trees you want to look at.
23.2 Inferring Orthologs/Paralogs, Gene Duplication and Loss

There are two commonly used trees. The species tree uses morphological characters, fossil evidence, etc to create a tree of how species are related (leaves are species). Gene trees look at specific genes in different species (leaves are genes).

Reconciliation is an algorithm to figure out how the gene tree fits inside the species tree. It maps the vertices in the gene tree to vertices in the species tree.

We want to minimize the duplication/loss so we want to map events as low in the tree as possible to when they happened to minimize loss.

Duplication events map to the same as both of its children. Loss event maps to gap in the mapping. Gene tree accuracy is important; even one branch misplaced can dramatically increases error.

23.3 Learning Across Gene Trees

If we knew the species tree we could know beforehand that we expect the branch to be longer. We can develop a model for what kind of branch lengths we can expect. We can use conserved gene order to tell orthologs and build trees.
When gene is fast evolving in one species, it is fast evolving in all species. We can model a branch length as two different rate components. One is gene specific (present across all species) and a species specific which is customized to a specific species. This method greatly improves reconstruction accuracy.

23.4 Modeling Population and Allele Frequencies

People keep sequencing genomes so looking at how populations evolve is becoming more and more important and feasible. The Wright-fisher model is used to study drifts, bottlenecks, etc. The coalescent model combines the Wright-fisher with trees. Wright-fisher was designed to study the effect of finite population sizes. We
Figure 23.12: Branch length can be modeled as two different rate components: gene specific and species specific.

need to assume population size is fixed at $N$, random mating, non-overlapping generations.

Figure 23.13: The Wright-Fisher model

Continue for many generations and ignore ordering of chromosomes.

Figure 23.14: The Wright-Fisher model continued over many generations and ignoring the ordering of chromosomes.
The coalescent model only focuses on the genealogy. It only is concerned about the lineages we have sequences for; do not have to worry about others. It is a probabilistic model that works backwards in time to find when they have common ancestors.

Say we have $2N$ individuals, what is the probability that $k$ lineages do not have any coalescent events in parental generation? What is the probability that the first coalescent of $k$ lineages is at $t$ generations? This process can be seen as a geometric distribution.

Can repeat to find when all individuals coalesce. Each branch of species tree can be seen as having its own Wright-Fisher inside of it.

### 23.5 SPIDIR: Background

As presented in the supplementary information for SPIDIR, a gene family is the set of genes that are descendents of a single gene in the most recent common ancestor (MRCA) of all species under consideration. Furthermore, genetic sequences undergo evolution at multiple scales, namely at the level of base pairs, and at the level of genes. In the context of this lecture, two genes are orthologs if their MRCA is a speciation event; two genes are paralogs if their MRCA is a duplication event.

In the genomic era, the species of a modern genes is often known; ancestral genes can be inferred by reconciling gene- and species-trees. A reconciliation maps every gene-tree node to a species-tree node. A common technique is to perform Maximum Parsimony Reconciliation (MPR), which finds the reconciliation
Figure 23.17: Multispecies Coalescent Model. Leaf branches track one lineage. There is a lag time from when population separated and when two actual gene lineages find a common ancestor. The rate of coalescent slows down as $N$ gets bigger and for short branches. Deep coalescent is depicted in light blue for three lineages. The species and gene tree are incongruent since C and D are sisters in gene tree but not the species tree. There is a $\frac{2}{3}$ chance that incongruence will occur because once we get to the light blue section, the Wright-fisher is memory less and there is only $\frac{1}{3}$ chance that it will be congruent. Effect of incongruence is called incomplete lineage sorting.

$R$ implying the fewest number of duplications or losses using the recursion over inner nodes $v$ of a gene tree $G$. MPR fist maps each leaf of the gene tree to the corresponding species leaf of the species tree. Then the internal nodes of $G$ are mapped recursively:

$$R(v) = MRCA(R(right(v)), R(left(v)))$$

If a speciation event and its ancestral node are mapped to the same node on the species tree. Then the ancestral node must be an duplication event.

Using MPR, the accuracy of the gene tree is crucial. Suboptimal gene trees may lead to an excess of loss and duplication events. For example, if just one branch is misplaced (as in Figure 23.2) then reconciliation infers 3 losses and 1 duplication event. In [6], the authors show that the contemporaneous current gene tree methods perform poorly (60% accuracy) on single genes. But if we have longer concatenated genes, then accuracy may go up towards 100%. Furthermore, very quickly or slowly evolving genes carry less information as compared with moderately diverging sequences (40-50% sequence identity), and perform correspondingly worse. As corroborated by simulations, single genes lack sufficient information to reproduce the correct species tree. Average genes are too short and contains too few phylogenetically informative characters. While many early gene tree construction algorithms ignored species information, algorithms like SPIDIR capitalize on the insight that the species tree can provide additional information which can be leveraged for gene tree construction. Synteny can be used to independently test the relative accuracy of different gene tree reconstructions. This is because syntenic blocks are regions of the genome where recently diverged organisms have the same gene order, and contain much more information than single genes.

Figure 23.18: MPR reconciliation of genes and species tree.

There have been a number of recent phylogenomic algorithms including: RIO [2], which uses neighbor
joining (NJ) and bootstrapping to deal with incogruencies, Orthostrapper [7], which uses NJ and reconciles to a vague species tree, TreeFAM [3], which uses human curation of gene trees as well as many others. A number of algorithms take a more similar track to SPIDIR [6], including [4], a probabilistic reconciliation algorithm [8], a Bayesian method with a clock,[9],and parsimony method using species tree , as well as more recent developments: [1] a Bayesian method with relaxed clock and [5], a Bayesian method with gene and species specific relaxed rates (an extension to SPIDIR).

23.6 SPIDIR: Method and Model

SPIDIR exemplifies an iterative algorithm for gene tree construction using the species tree. In SPIDIR, the authors define a generative model for gene-tree evolution. This consists of a prior for gene-tree topology and branch lengths. SPIDIR uses a birth and death process to model duplications and losses (which informs the prior on topology) and then then learns gene-specific and species-specific substitution rates (which inform the prior on branch lengths). SPIDIR is a Maximum a posteriori (MAP) method, and, as such, enjoys several nice optimality criteria.

In terms of the estimation problem, the full SPIDIR model appears as follows:

$$
$$

The parameters in the above equation are: $D =$ alignment data , $L =$ branch length $T =$ gene tree topology , $R =$ reconciliation , $S =$ species tree (expressed in times) , $\Theta =$ (gene and species specific parameters [estimated using EM training], $\lambda$, $\mu$ dup/loss parameters)). This model can be understood through the three terms in the right hand expression, namely:

1. the sequence model$-- P(D|T,L)$. The authors used the common HKY model for sequence substitutions, which unifies Kimura’s two parameter model for transitions and transversions with Felsenstein’s model where substitution rate depends upon nucleotide equilibrium frequency.

2. the first prior term, for the rates model$-- P(L|T,R,S,\Theta)$, which the authors compute numerically after learning species and gene specific rates.

3. the second prior term, for the duplication/loss model$-- P(T,R|S,\Theta)$, which the authors describe using a birth and death process.

Having a rates model is very rates model very useful, since mutation rates are quite variable across genes. In the lecture, we saw how rates were well described by a decomposition into gene and species specific rates. In lecture we saw that an inverse gamma distribution appears to parametrize the gene specific substitution rates, and we were told that a gamma distribution apparently captures species specific substitution rates. Accounting for gene and species specific rates allows SPIDIR to build gene trees more accurately than previous methods. A training set for learning rate parameters can be chosen from gene trees which are congruent to the species tree. An important algorithmic concern for gene tree reconstructions is devising a fast tree search method. In lecture, we saw how the tree search could be sped up by only computing the full $\text{argmax}_{L,T,R,P}(L,T,R|D,S,\Theta)$ for trees with high prior probabilities. This is accomplished through a
computational pipeline where in each iteration 100s of trees are proposed by some heuristic. The topology prior \( P(T, R|D, S, \Theta) \) can be computed quickly. This is used as a filter where only the topologies with high prior probabilities are selected as candidates for the full likelihood computation.

The performance of SPIDIR was tested on a real dataset of 21 fungi. SPIDER recovered over 96% of the synteny orthologs while other algorithms found less than 65%. As a result, SPIDER invoked much fewer number of duplications and losses.

### 23.7 Conclusion

Incorporating species tree information into the gene tree building process via introducing separate gene and species substitution rates allows for accurate parsimonious gene tree reconstructions. Previous gene tree reconstructions probably vastly overestimated the number of duplication and loss events. Reconstructing gene trees for large families remains a challenging problem.

### 23.8 Current Research Directions

### 23.9 Further Reading

### 23.10 Tools and Techniques

### 23.11 What Have We Learned?

### Bibliography


24.1 Introduction

Humans share 99.9% of the same genetic information, and are 99% similar to chimpanzees. Learning about the 0.1% difference between humans can be used to understand population history, trace lineages, predict disease, and analyze natural selection trends. In this lecture, Dr. Reich explained how we can use this data to see evidence of gene flow between neanderthals and modern humans of Western Eurasian decent.

Last year, he examined India as a case of how genetic data can inform population history, which is included as an appendix.

24.2 Quick Survey of Human Genetic Variation

In the human genome, there is generally a polymorphism every 1000 bases, though there are regions of the genome where this rate can quadruple. These polymorphisms are markers of genetic variation. It is necessary to understand how genetic variation arises before attempting to analyze it. Single Nucleotide Polymorphisms (SNPs) are one manifestation of genetic variation. When SNPs occur, they are segregated according to recombination rates, advantages or disadvantages of the mutation, and the population structure that exists and continues during the lifespan of the SNP. Through the passing of generations, recombination splits the SNP haplotype into smaller blocks. The length of these blocks, then, is dependent on the rate of recombination and the stability of the recombination product. Therefore, the length of conserved haplotypes can be used to infer the age of mutation or its selection. An important consideration, though, is that the rate of recombination is not uniform across the genome; rather, there are recombination hot spots that can
skew the measure of haplotype age or selectivity. This makes the haplotype blocks longer than expected under a uniform model.

To Dr. Reich, every place in the genome is best thought of as a tree when compared across individuals. But, depending on where you look within the genome, this particular tree will be different than another particular tree you may get from a specific set of SNPs. The trick is to use the data that we have available on SNPs to infer the underlying trees, and then the overarching phylogenetic relationships. Take, for instance, the Y chromosome. It undergoes little to no recombination and thus can produce a high accuracy tree as it passed down from father to son. Likewise, we can take mitochondrial DNA, passed down from mother to child. While these trees can have high accuracy, other autosomal trees are confounded with recombination, and thus show lower accuracy to predict phylogenetic relationships. Gene trees are best made by looking at areas of low recombination, as recombination mixes trees. In general, there are about 1 to 2 recombinations per generation.

Humans show about 10,000 base-pairs of linkage, as we go back about 10,000 generations. Fruit fly linkage equilibrium blocks, on the other hand, are only a few hundred bases. Fixation will occur over time, proportional to the size of the population. For a population of about 10,000 it will take about 10,000 years to reach that point. When a population expands, genetic drift goes down. So, curiously enough, the variation in humans looks like what would have been formed in a population size of 10,000.

If long haplotypes are mapped to genetic trees, approximately half of the depth is on the first branch; most morphology changes are deep in the tree because there was more time to mutate. One simple model of mutation without natural selection is the Wright-Fisher neutral model which utilizes binomial sampling. In this model, a population will reach fixation (frequency 1), will die out (frequency 0), or continue to segregate. In the human genome, there are 10-20 million common SNPs. This is less diversity than chimpanzees, implying that humans are genetically closer to one another.

With this genetic similarity in mind, comparing human sub-populations can give information about common ancestors and suggest historical events. The similarity between two sub-populations can be measured by comparing allele frequencies in a scatter plot. If we plot the frequencies of SNPs across different populations on a scatterplot, we see more spread between more distant populations. The plot below, for example, shows the relative dissimilarity of European American and American Indian populations along with the greater similarity of European American and Chinese populations. The plots indicate that there was a divergence in the past between Chinese and Native Americans, evidence for the North American migration bottleneck that has been hypothesized by archaeologists. The spread among different populations within Africa is quite large. We can measure spread by Fst (which describes the variance).

![Figure 24.1: Similarity between two subpopulations can be measured by comparing allele frequencies in a scatterplot. The plots show the relative dissimilarity of European American and American Indian populations along with greater similarity of European American and Chinese populations.](image)

Several current studies have shown that unsupervised clustering of genetic data can recover self-selected labels of ethnic identity. In Rosenbergs experiment, a bayesian clustering algorithm was developed. They took a sample size of 1000 people (50 populations, 20 people per population), and clustered those people by their SNP genetic data but they didnt tag any of the people by populations, so they could see how the algorithm would cluster them without knowledge of ethnicity. They didnt know what the optimal number of clusters was, so they tried 2, then 3, then 4, then 5, et cetera. What they found was that with 2 clusters, East-
Asians and non-East-Asians were separated. With 3 clusters, Africans were separated from everyone else. With 4, East-Asians and Native Americans were separated. And then with 5, the smaller sub-populations began to emerge.

When waves of humans left Africa, genetic diversity decreased; the small numbers of people in the groups that left Africa allowed for serial founder events to occur. These serial founder events lead to the formation of sub-populations with less genetic diversity. This is evidenced by the fact that genetic diversity decreases moving out of Africa. West Africans have the highest diversity of any human sub-population.

### 24.3 Neanderthals and Modern Human Gene Flow

Recently, Dr. Reich worked with the Max Planck Institute as a population geneticist studying Neanderthal genetic data. He discussed with us the background of his research as part of the Neanderthal genome project, the draft sequence that they assembled, and then the evidence that's been compiled for gene flow between modern humans and Neanderthals.

#### 24.3.1 Background

Clear fossils from Neanderthals from 200,000 years ago exist in West Eurasia (Europe and Western Asia), which is far earlier than *Homo erectus*. The earliest fossils of us come from Ethiopia dating about 200,000 years ago. However, there is evidence that neanderthals and humans overlapped in time and space between 135,000 and 35,000 years ago.

The first place of contact could have occurred in The Levant, in Israel. There are human fossils from 120,000 years ago, then a gap, neanderthal fossils about 80,000 years ago, another gap, and then human fossils again 60,000 years ago. This is proof of an overlap in place, but not in time. In the upper paleolithic era, there was an explosion of populations out of Africa (the migration about 60,000 to 45,000 years ago). In Europe after 45,000 years ago, there are sites where neanderthals and humans exist in the fossil record side by side. (Eastern Eurasia is not well documented, so we have little evidence from there.) Since there is evidence that they co-existed, was there interbreeding? This is a question we wish to answer by examining population genomics.

Let's take a look at how you go about finding and sequencing DNA from ancient remains. First, you have to obtain a bone sample with DNA from a neanderthal. Human DNA and Neanderthal DNA is very similar (we are more similar to them than we are to chimps), so when sequencing short reads with very old DNA, it's impossible to tell if the DNA is neanderthal or human. So, the cave is first classified as human or non-human, which helps to predict the origin of the bones. In sites of findings, typically lots of trash is left behind which is used to help classify the site (stone tools, particular technologies, the way meat was cut off animals, other trash). These caves are made up of lots of trash, and only the occasional bone. Even if you have a bone, it is still very unlikely that you have any salvageable DNA. In fact, 99% of the sequence of Neanderthals comes from only three long bones found in one site: the Vindija cave in Croatia (5.3 Gb, 1.3x full coverage). The paleontologists chose to sacrifice the long bones because they were less morphologically helpful.

Next, the DNA is sent to an ancient-DNA lab. Since they are 40,000 year old bones, there is very little DNA left in them. So, they are first screened for DNA. If they have it, is it primate DNA? Usually it is DNA from microbes and fungi that live in soil and digest us when we die. If it is primate DNA, is it contamination from the human (archeologist or lab tech) handling it? The difference between human and neanderthal DNA is 1/600 bp. The size of reads from a 40,000 year old bone sample is 30-40 bp. The reads are almost always identical for a human and neanderthal, so it is difficult to distinguish them.

Only about 1-10% of the DNA on old bones is the primates DNA. 89 DNA extracts were screened for neanderthals, but only 6 bones were actually sequenced (requires lack of contamination and high enough amount of DNA). The process of retrieving the DNA requires drilling beneath the bone surface (to minimize contamination) and taking samples from within. For the three long bones, less than 1 gram of bone powder was able to be obtained. Then the DNA is sequenced and aligned to a reference chimp genome. It is mapped to a chimp instead of a particular human because mapping to a human might cause bias if you are looking to see how the sequence relates to specific human sub-populations.
Most successful finds have been in cool limestone caves, where it is dry and cold and perhaps a bit basic. The best chance of preservation occurs in permafrost areas. Very little DNA is recoverable from the tropics. The tropics have a great fossil record, but DNA is much harder to obtain. Since most bones don't yield enough or good DNA, scientists have the screen samples over and over again until they eventually find a good one.

24.3.2 Draft Sequence

The neanderthal DNA had short reads, about 37 bp on average. There are lots of holes due to mutations caused by time eroding the DNA. However, there some characteristic mutations that occur on DNA that has been sitting for very long periods of time. There is a tendency to see C to T mutations, and G to A mutations. Over time, a methyl group gets knocked off of a C, which causes it to resemble U. When PCR is used to amplify the DNA for sequencing, the polymerase sees a U and repairs it to a T. The G to A mutations are just the result of seeing that on the opposite strand, so really the important mutation to worry about is the C to T. This mutation is seen about 2% of the time! In order to combat this, scientists use a special enzyme now that recognizes the U, and instead of replacing it with a T, simply cuts the strand where it sees the mutation. This helps to identify those sites.

The average fragment size is quite small, and the error rate is still 0.1% - 0.3%. One way to combat the mutations is the note that on a double stranded fragment, the DNA is frayed towards the ends, where it becomes single stranded for about 10 bp. There tend to be high rates of mutations in the first and last 10 bases, but high quality elsewhere (so C to T mutations in the beginning and G to A in the end). In chimps, the most common mutations are transitions (purine to purine, pyrimidine to pyrimidine), and transversions are much rarer. The same goes for humans. Since the G to A and C to T mutations are transitions, it can be determined that there are about 4x more mutations in the Neanderthal DNA than if it were fresh by noting the number of transitions seen compared to the number of transversions seen (by comparing Neanderthal to human DNA). Transversions have a fairly stable rate of occurrence, so that ratio helps determine how much error has occurred through C to T mutations.

We are now able to get human contamination of artifact DNA down to around 1%. When the DNA is brought in, as soon as it is removed from the bone it is bar coded with a 4 bp tag (originally it was 4, now it is 7). That allows you to avoid contamination at any later point in the experiment, but not earlier. Extraction is also done in a clean room with UV light, after having washed the bone. Mitochondrial DNA is helpful for distinguishing what percent of the sample is contaminated with human DNA. Mitochondrial DNA is filled with characteristic event sites in different species - all Neanderthals were of one type, all humans of another (called reciprocally monophylogenetic). The contamination can be measured by counting the ratio of those sites. In the Neanderthal DNA, contamination was present, but it is only 0.5%.

In sequencing, the error rate is almost always higher than the polymorphism rate. Therefore, most sites in the sequence that are different from humans are caused by sequencing errors. So we can't exactly learn about Neanderthal biology through the sequence generated, but we can analyze particular SNPs as long as we know where to look. The probability of a particular SNP being changed due to an error in sequencing is only $\frac{1}{300}$ to 11000. So we can, in fact, still get lots of usable data from this after all.

After aligning the chimp, Neanderthal, and modern human sequences, we can measure the distance that Neanderthals are, on the scale from humans to chimps. This distance is only about 12.7% from the human reference sequence. a French sample measures about 8% distance from the reference sequence, and a Bushman about 10.3%. What this says is that the Neanderthal DNA is within our range of variation as a species.

24.3.3 Evidence for Gene Flow

1. First, lets look at a comparison test. Take two randomly chosen populations, sequence both, and for each different SNP, check to see which population the Neanderthal DNA matched. This was done for 8 sequences. When Eurasians were compared with Eurasians, there was little difference. When Africans were compared with Africans, there was also little difference. However, when Africans were compared with non-Africans, Neanderthal SNPs much more highly matched the non-African DNA.
This is evidence that there was mating and gene flow between Neanderthals and Eurasian modern humans.

2. Second, well take a look at a long range haplotype study done at Berkeley. These researchers picked long range sections of the genome and compared them among randomly chosen humans from various populations. When you look to see where the deepest branch of the tree constructed from that haplotype is, it almost always comes from an African population. However, occasionally non-Africans have the deepest branch. The study found that there were 12 regions where non-Africans have the deepest branch. When this data was used to analyze the Neanderthal genome, it was found that $\frac{10}{12}$ of these regions in non-Africans matched Neanderthals more than the matched the human reference sequence (a compilation of sequences from various populations). This is evidence of that haplotype actually being of Neanderthal origin.

3. Lastly, there is a bigger divergence than expect among humans. The average split between a Neanderthal and a human is about 800,000 years. The typical divergence between two humans is about 500,000 years. When looking at african and non-african sequences, regions of low divergence emerged in non-african sequences when compared with Neanderthal material. The regions found were highly enriched for Neanderthal material (94% Neanderthal), which would increase the average divergence between humans (as the standard Neanderthal - human divergence is about 800,000 years).

24.4 Discussion

There was an example of a 50,000 year old bone found in southern Siberia, where the mtDNA was sequenced, that appears to be an out-group to both Neanderthals and modern humans. It was a little finger bone of a child. It is twice as deep in the phylogenetic tree as either of them, and has 1.9x coverage. These pieces of the ancestral DNA puzzle help us piece together human history and before. They serve to help us understand where we came from.

The bottleneck caused by the migration from Africa is only one example of many that have occurred. Most scientists usually concentrate on the age and intensity of events and not necessarily the duration, but the duration is very important because long bottlenecks create a smaller range of diversity. One way to help tell the length of a bottleneck is to determine if any new variations arose during it, as that occurs during longer bottlenecks, and as they will help distinguish how long it lasted. That change in range of diversity is also what helped create the different human sub-populations that became geographically isolated. This is just another way that population genomics can be useful for helping to piece together information.

Today, Dr. Reich showed how genetic differences between species (specifically here within primates) can be used to help understand the phylogenetic tree from which we are all derived. We looked at the case study of comparisons with Neanderthal DNA, learned about how ancient DNA samples are obtained, how sequences are found and interpreted, and how that evidence shows high likelihood of interbreeding between modern humans (of eurasian descent) and Neanderthals. Those very small differences between one species and the next, and within species, are what allow us to deduce a great deal of this history through population genetics.

24.5 Current Research Directions

24.6 Further Reading

24.6.1 Fall 2009 Discussion Topic: Genomic Variation in 25 Groups in India

There is a general taxonomy for studying population relationships with genetic data. The first general type of study utilizes both phylogey and migration data. It fits the phylogenies to Fst values, values of sub-population heterozygosity (pioneered by Cavalli-Sforza and Edwards in 2267).TODO: cite the paper where this is discussed in more detail @scribe: This method also makes use of synthetic maps and Principal Components Analysis. [2] The primary downside to analyzing population data this way is uncertainty about results. There are mathematical and edge effects in the data processing that cannot be predicted.
Also, certain groups have shown that separate, bounded mixing populations can produce significant-seeming principal components by chance. Even if the results of the study are correct, then, they are also uncertain.

The second method of analyzing sub-population relationships is genetic clustering. Clusters can be formed using self-defined ancestry [1] or the STRUCTURE database. [3] This method is overused and can over-fit the data; the composition of the database can bias the clustering results.

Technological advances and increased data collection, though, have produced data sets that are 10,000 times larger than before, meaning that most specific claims can be disproved by some subset of data. So in effect, many models that are predicted either by phylogeography or migration or genetic clustering will be disproved at some point, leading to large-scale confusion of results. One solution to this problem is to use a simple model that makes a statement that is both useful and has less probability of being falsified.

Past surveys in India have studied such aspects as anthropometric variation, mtDNA, and the Y chromosome. The anthropometric study looked at significant differences in physical characteristics between groups separated by geography and ethnicity. The results showed variation much higher than that of Europe. The mtDNA study was a survey of maternal lineage and the results suggested that there was a single Indian tree such that age of lineage could be inferred by the number of mutations. The data also showed that Indian populations were separated from non-Indian populations at least 40,000 years ago. Finally, the Y chromosome study looked at paternal lineage and showed a more recent similarity to Middle Eastern men and dependencies on geography and caste. This data conflicts with the mtDNA results. One possible explanation is that there was a more recent male migration. Either way, the genetic studies done in India have served to show its genetic complexity. The high genetic variation, dissimilarity with other samples, and difficulty of obtaining more samples lead to India being left out of HapMap, the 1000 Genomes Project, and the HGDP.

For David Reich and collaborators study of India, 25 Indian groups were chosen to represent various geographies, language roots, and ethnicities. The raw data included five samples for each of the twenty-five groups. Even though this number seems small, the number of SNPs from each sample has a lot of information. Approximately five hundred thousand markers were genotyped per individual. Looking at the data to emerge from the study, if Principal Components Analysis is used on data from West Eurasians and Asians, and if the Indian populations are compared using the same components, the India Cline emerges. This shows a gradient of similarity that might indicate a staggered divergence of Indian populations and European populations.

### 24.6.2 Almost All Mainland Indian Groups are Mixed

Further analysis of the India Cline phenomenon produces interesting results. For instance, some Pakistani sub-populations have ancestry that also falls along the Indian Cline. Populations can be projected onto the principal components of other populations: South Asians projected onto Chinese and European principal components produces a linear effect (the India Cline), while Europeans projected onto South Asian and Chinese principal components does not. One interpretation is that Indian ancestry shows more variability than the other groups. A similar variability assessment appears when comparing African to non-African populations. Two tree hypotheses emerge from this analysis:

1. there were serial founder events in Indias history or
2. there was gene flow between ancestral populations.

The authors developed a formal four population test to test ancestry hypotheses in the presence of admixture or other confounding effects. The test takes a proposed tree topology and sums over all SNPs of \((Pp_1 \ Pp_2)(Pp_3 \ Pp_4)\), where P values are frequencies for the four populations. If the proposed tree is correct, the correlation will be 0 and the populations in question form a clade. This method is resistant to several problems that limit other models. A complete model can be built to fit history. The topology information from the admixture graphs can be augmented with Fst values through a fitting procedure. This method makes no assumptions about population split times, expansion and contractions, and duration of gene flow, resulting in a more robust estimation procedure.

Furthermore, estimating the mixture proportions using the 4 population statistic gives error estimates for each of the groups on the tree. Complicated history does not factor into this calculation, as long as the topology as determined by the 4-population test is valid.
Figure 24.2: Populations can be projected onto the principal components of other populations: South Asians projected onto Chinese and European principal components produces a linear effect (the India Cline), while Europeans projected onto South Asian and Chinese principal components does not.

Figure 24.3: An admixture graph that fits Indian history

These tests and the cline analysis allowed the authors to determine the relative strength of Ancestral North Indian and Ancestral South Indian ancestry in each representative population sample. They found that high Ancestral North Indian ancestry is correlated with traditionally higher caste and certain language groupings. Furthermore, Ancestral North Indian (ANI) and South Indian (ASI) ancestry is as different from Chinese as European.

24.6.3 Population structure in India is different from Europe

Population structure in India is much less correlated with geography than in Europe. Even correcting populations for language, geographic, and social status differences, the Fst value is 0.007, about 7 times that of the most divergent populations in Europe. An open question is whether this could be due to missing (largely India-specific) SNPs on the genotyping arrays. This is because the set of targeted SNPs were identified primarily from the HapMap project, which did not include Indian sources.
Most Indian genetic variation does not arise from events outside India. Additionally, consanguineous marriages cannot explain the signal. Many serial founder events, perhaps tied to the castes or precursor groups, could contribute. Analyzing a single group at a time, it becomes apparent that castes and subcastes have a lot of endogamy. The autocorrelation of allele sharing between pairs of samples within a group is used to determine whether a founder event occurred and its relative age. There are segments of DNA from a founder, many indicating events more than 1000 years old. In most groups there is evidence for a strong, ancient founder event and subsequent endogamy. This stands in contrast to the population structure in most of Europe or Africa, where more population mixing occurs (less endogamy).

These serial founder events and their resulting structure have important medical implications. The strong founder events followed by endogamy and some mixing have lead to groups that have strong propensities for various recessive diseases. This structure means that Indian groups have a collection of prevalent diseases, similar to those already known in other groups, such as Ashkenazi Jews or Finns. Unique variation within India means that linkages to disease alleles prevalent in India might not be discoverable using only non-Indian data sources. A small number of samples are needed from each group, and more groups, to better map these recessive diseases. These maps can then be used to better predict disease patterns in India.

24.6.4 Discussion

Overall, strong founder events followed by endogamy have given India more substructure than Europe. All surveyed tribal and caste groups show a strong mixing of ANI and ASI ancestry, varying between 35% and 75% ANI identity. Estimating the time and mechanism of the ANI-ASI mixture is currently a high priority. Additionally, future studies will determine whether and how new techniques like the 4-population test and admixture graphs can be applied to other populations.

24.7 Tools and Techniques

24.8 What Have We Learned?

Bibliography


25.1 Introduction

The ability to sequence find patterns in genomes is a fundamental tool in research into the evolution of organisms, especially Humans. There are various algorithms to find the similarities between the genome of our species and another, but how do we find the divergences and how long ago did they occur? In other words, how do we measure selection in the Human Genome? This lecture aims to outline the main criteria and features that are analyzed to answer these questions.

25.2 Polymorphisms

Polymorphisms are differences in appearance (Phenotype) amongst members of the same species. Since all polymorphisms have a genetic basis, they can be characterized into types based on the mutation in the genome.

**Single Nucleotide Polymorphisms (SNPs)**  
- The mutation of only a single nucleotide base within a sequence. In most cases, this type of change has no consequence. However, there are some cases where this causes a major change.
  - Prime example is when glutamic acid (GAG) valine (GTG) in hemoglobin and causes **Sickle Cell Anemia**

**Variable Number Tandem Repeats**  
- When the copying machinery in the body is transcribing repeats within the genome, but loses track of how many repeats its made and makes more repeats than there originally was.
  - Prime example is a triple CAG repeat which causes **Huntingtons disease**, where there is gradual muscle control loss and severe neurological degradation. Also an example of the phenomenon of anticipation where if it is inherited by the next generation, it becomes more severe in that generation.
Insertion/Deletion  – When certain nucleotide bases are just forgotten to be transcribed or extra nucleotides are transcribed into a sequence.
  - Worst when only 1 or 2 are deleted or added since this will shift the frame of reading codons (groups of 3 nucleotides) for instructions
  - Prime example is deletions in the CFTR gene, which codes for chloride channels in the lungs and may cause Cystic Fibrosis where the patient cannot clear mucous in the lungs and causes infection

25.2.1 Allele and Genotype Frequencies

Hardy Weinberg Principle: Allele and Genotype frequencies within a population will remain at constant equilibrium unless there is an outside influence and/or interaction.

Assumptions in Hardy-Weinberg:
  - All mating in the population is at random, there is no inbreeding or arranged pairings
  - There is no random mutations in the population
  - No migration of the species or introduction of another subpopulation into the general population; isolated
  - No natural selection, all individuals have equal probability of survival
  - The population observed is very large
  - Allele frequency drives future genotype frequency (Prevalent allele drives Prevalent genotype)

In a Hardy Weinberg Equilibrium, for two genotypes A and T, occurring with probability $p$ and $q = 1-p$, respectively, the probabilities of finding the homozygous AA or TT ($pp$ or $qq$, respectively) or heterozygous ($2pq$) genotypes can be described by the equation:

$$p^2 + 2pq + q^2 = 1$$

This equation gives a table of probabilities for each genotype, which can be compared with the observed genotype frequencies using statistical error tests such as the chi-squared test to determine if the Hardy-Weinberg model is applicable.

In a small population that violates one criteria of the Hardy-Weinberg, when genetic drift (a mutation) occurs, it will always either disappear (frequency $= 0$) from the population or become prevalent in a species - this is called fixation; in general, 99% of mutations disappear. Shown below is a simulation of a mutations prevalence in a finite-sized population over time: both perform random walks, with one mutation disappearing and the other becoming prevalent:

![Figure 25.1: Changes in allele frequency over time](image)

Once a mutation has disappeared, the only way to have it reappear is the introduction of a new mutation into the species population. For humans, it is believed that a given mutation, evolving neutrally, should fixate to 0 or 1 (within, e.g., 5%) within a few million years. However, under selection this will happen much faster.
25.2.2 Ancestral State of Polymorphisms

How do we determine for e.g., a SNP, which allele was the original and which is the mutation? Traces of the ancestral state can be done by comparing the genomes of a species and a closely related species, or outgroup (e.g. humans and chimpanzees) with a known phylogenetic tree. Mutations can occur anywhere along this tree sometimes mutations at the split fix differently in different populations (“fixed difference”), in which the entire populations differ in genotype. However, recently occurring mutations have not fixed, and a polymorphism will be present in one species but fully absent in the other (simultaneous mutations in both species are very rare). In this case, the “derived variant” is the version of the polymorphism appearing after the split, while the ancestral variant is the version shared by both species. There is one caveat: perhaps 1% of the time, something more complex (e.g. double mutations) may occur, and this simple model does not hold.

25.2.3 Measuring Derived Allele Frequencies

The derived allele frequency (i.e., the frequency of the derived variant in the population) can be very easily measured. In doing this there is an implicit assumption that the population is homogeneous. However, in fact there could be some unseen divide between two groups that cause them to evolve separately (shown below):

![Figure 25.2: Two isolated populations](image)

In this case, the prevalence of the variants among subpopulations is different, and Hardy-Weinberg is violated.

One way to quantify this is to use the Fixation Index (Fst) to compare subpopulations within a species. In reality, it is merely a proportion of the total heterozygosity found in a species in a given subpopulation. Fst estimates reduction in heterozygosity (2pq with alleles p and q) expected when 2 different populations are erroneously grouped given that there is in a population n alleles with frequencies pi where (1 ≤ i ≤ n) and homozygosity, G, of the population is given by:

\[ \Sigma_{i=1}^{n} p_i^2 \]

And the total heterozygosity in the population is given by 1-G.

\[ F_{st} = \frac{\text{Heterozygosity(total)} - \text{Heterozygosity(subpopulation)}}{\text{Heterozygosity(total)}} \]

In the case of the figure above, there is no heterozygosity between the populations, so \( F_{st} = 1 \). In reality within one species (e.g.: humans) the Fst will be small (0.0625). For actual tests, Fst is computed either by clustering sub-populations randomly, or using some simple/obvious characteristic, e.g., ethnicity.

25.3 Genetic Linkage

The second law in Mendelian inheritance is the law of independent assortment:

*Law of Independent Assortment: Alleles of different genes assort independently without influence from other alleles.*
When this “law” holds, there is no correlation between different polymorphisms. Then the probability of a haplotype (a given set of polymorphisms) is simply the product of the individual polymorphism probabilities.

Of course, at first sight this law seems impossible: genes lie on chromosomes, and chromosomes are passed on as units to offspring. Therefore, two polymorphisms occurring on the same chromosome should be perfectly correlated (perfectly linked). However, genetic recombination events, in which segments of DNA on homologous chromosomes are swapped, will over time reduce the correlation between polymorphisms. Over a suitably long time interval, recombination will remove the linkage between two polymorphisms; they are said then to be in equilibrium. When, on the other hand, the polymorphisms are correlated (either because they are both very recent, or as we will see below, by positive selection), we have Linkage Disequilibrium (LD). The amount of disequilibrium is the difference between the observed haplotype frequencies and those predicted in equilibrium.

We can use this LD measurement (= D) to find the difference between observed and expected assortments. If there are two Alleles (1 and 2) and two loci (A and B) we can calculate the haplotype probabilities and find expected allele frequency probabilities and compare to observed probabilities

- **Haplotype frequencies**
  - \( P(A_1) = x_{11} \)
  - \( P(B_1) = x_{12} \)
  - \( P(A_2) = x_{21} \)
  - \( P(B_2) = x_{22} \)

- **Allele Frequency**
  - \( P_{11} = x_{11} + x_{12} \)
  - \( P_{21} = x_{21} + x_{22} \)
  - \( P_{12} = x_{11} + x_{21} \)
  - \( P_{22} = x_{12} + x_{22} \)

- **D** = \( P_{11} \times P_{22} - P_{12} \times P_{21} \)

Using the value Dmax, the maximum value of D with given allele frequencies, it can compared to D in the relationship:

\[ D' = \frac{D}{D_{\text{max}}} \]

Where \( D' \) is equal the maximum linkage disequilibrium or complete skew for the given alleles and allele frequencies. Dmax can be found by taking the smaller of the expected haplotype frequencies \( P(A_1B_2) \) or \( P(A_2B_1) \). If there is full combination and complete independent sorting, then it is in complete equilibrium in which case \( D' = 0 \) for completely unlinked. Vice versa, a value of \( D = 1 \) will demonstrate total linkage.

The key point for this lecture is that mutations that have occurred relatively recently have not had time to be broken down by Linkage Disequilibrium. Normally, such a mutation will not be very common. However, if it is under positive selection, the mutation will be much more prevalent in the population than expected. Therefore, by carefully combining a measure of LD and derived allele frequency we can determine of a region is under positive selection.

Decay of Linkage Disequilibrium is driven by recombination rate and time (in generations) and has an exponential decay. For a higher recombination rate, linkage disequilibrium will decay faster in a shorter amount of time. The hard part in the test for natural selection is actually getting the background recombination rate, as it is difficult to estimate and depends on the location in the genome. Comparison of genomic data across multiple species helps a lot in determining these background rates.

### 25.3.1 Correlation Coefficient \( r^2 \)

Answers how predictive is allele at locus A of allele at locus B

\[ r^2 = \frac{D^2}{P(A_1)P(A_2)P(B_1)P(B_2)} \]
As the value of $r^2$ approaches 1, the more it can be said that two alleles at two loci can be correlated. There may also be Linkage Disequilibrium between two haplotypes, but the haplotypes are not correlated at all. The correlation coefficient is particularly interesting when studying disease mapping where information at locus A does not predict a disease whereas locus B does. Or there could be the case where neither locus A nor locus B predict disease but loci AB where A1B2 does predict a particular disease. With many linkages however, we can hope to generalize and tag many haplotypes.

25.4 Natural Selection

In the mid 2300s there were many papers already on the theory of evolution; however, Darwin and Wallace provide mechanisms of natural selection that lead to evolution. Not until 70 years later (1948) did we have an example of how this could occur in humans: J.B.S Haldanes Malaria Hypothesis showed a correlation between genetic mutations in red blood cells and the distribution of malaria prevalence. It turned out that individuals with these mutations (e.g., sickle cell mutations) had a resistance to malaria. Therefore, there is a direct environmental pressure for genetic mutation.

Another example is LCT, lactose tolerance (lasting into adulthood). However, these explicit examples were very hard to construct, since the investigators did not have genetic data. Now, in hindsight, we can search the genome for regions with the same patterns as these known examples to identify regions undergoing natural selection. This begs the question, what are these signals?

25.4.1 Genomics Signals of Natural Selection

- Exponential prevalence of a feature in sequential generations
- Mutation that helps species prosper

Tests

- Long range correlations (iHs, Xp, EHH): If we tag genetic sequences on an allele and try to reconstruct, we end up with a broken haplotype and the number of breaks or color changes is directly correlated to the number of recombinations and to how old or long ago it occurred.

![Figure 25.3: Genomic Signals of Natural Selection](image)

- SWEEP program developed by Pardis Sabeti, Ben Fry and Patrick Varilly. SWEEP detects evidence of natural selection by analyzing haplotype structures in the genome and using Long Range Haplotype (LRH) test. It looks for high frequency alleles with long range Linkage Disequilibrium This suggests that there was a large scale proliferation of a haplotype that occurred at a rate greater than recombination could break it from its markers.

- High Frequency derived Look for large spikes in the frequency of derived alleles in set positions.

- High Differentiation (Fst) Large spikes in differentiation at certain positions.
Using these tests, we can investigate selected regions. One problem is that, while a single SNP may be under positive selection and proliferate, nearby SNPs will hitchhike along. It is difficult to distinguish the SNP under selection from the hitchhikers with only one test. Under selection, these tests are strongly correlated; however, in the absence of selection they are generally independent. Therefore, by employing a composite statistic built from all of these tests, it is possible to isolate out the individual SNP under selection.

Examples where a single SNP has been implicated in a trait:

- Chr15 Skin pigmentation in Northern Europe
- Chr2 Hair traits in Asia
- Chr10 Unknown trait in Asia
- Chr12 Unknown Trait in Africa

The International HapMap Project aims to catalog the genomes of humans from various countries and regions and find similarities and differences to help researchers find genes that will benefit the advance in disease treatment and administration of health related technologies.

25.5 Current Research Directions

25.6 Further Reading

- Application to new data sets as they become available
- Additional genotyping and sequencing
- Functional validation
- Design of model organisms

25.7 Tools and Techniques

25.8 What Have We Learned?

Bibliography
26.1 Introduction

Mark J. Daly, Ph.D., is an Associate Professor at the Massachusetts General Hospital/Harvard Medical School and an Associate Member of the Broad Institute. This lecture explains how statistical and computational methods can aid researchers in understanding, diagnosing, and treating disease. The problem of identifying genetic variation which can explain phenotypic variation is known as association mapping. This problem is particularly important to solve for disease phenotypes (e.g., susceptibility). Historically, the method of choice for solving this problem was linkage analysis. However, advances in genotyping technology have allowed for a more powerful method called genome-wide association. More recent advances in genomic data have allowed for novel integrative analyses which can make more powerful predictions about diseases.

26.2 Goals of investigating the genetic basis of disease

Any discussion about the basis of disease must consider both genetic and environmental effects. However, it is known that many traits, for example those in Figure 26.1, have significant genetic components. Formally, the heritability of a phenotype is the proportion of variation in that phenotype which can be explained by genetic variation. The traits in Figure 26.1 are all at least 50% heritable.

Accurately estimating heritability involves statistical analyses on samples with highly varied levels of shared genetic variation (e.g., twins, siblings, relatives, and unrelated). Studies on the heritability of Type 2 diabetes, for example, have shown that given you have diabetes, the risk to the person sitting next to you (an unrelated person) increases by 5–10%; the risk to a sibling increases by 30%; and the risk to an identical twin increases by 85–90%.
Having established that there is a genetic component to disease traits of interest, what are the goals of understanding this component? There are three main goals:

- Directing downstream research in disease
- Potential for improved diagnostics
- Enabling rational drug development

Identifying genetic variants which explain variation in the disease trait obviously contributes to our ability to understand the mechanism (the biochemical pathways, etc.) by which the disease manifests. Moreover, those variants can be used in genetic screens to test for increased risk for the disease trait. But the last goal is of particular interest because strong evidence suggests we do not really know how to develop effective drugs to target particular diseases. For example, in the last 50 years, no truly novel compounds have been developed to treat various psychiatric disorders such as schizophrenia.

Figure 26.2 depicts the cycle of drug development. First, researchers hypothesize a possible target of interest that might be related to a disease. They evaluate the biochemistry of this target, test the target in model organisms, and then finally perform clinical trials in humans. However, the vast majority of drugs which make it through this process end up being ineffective in treating the disease for which they were originally designed. This result is mainly a consequence of poor hypotheses about the basis of the disease in question.

Statins are a prominent example of highly effective drugs developed after work on understanding the genetic basis of the disease trait they are targeted at. Dr. Michael Brown and Dr. Joseph Goldstein won the Nobel Prize in Physiology or Medicine in 1985 for their work on the regulation of LDL cholesterol metabolism [5]. They were able to isolate the cause of extreme familial hypercholesterolemia (FH), a Mendelian disorder, to mutations of a single gene encoding an LDL receptor. Moreover, they were able to identify the biochemical pathway which was affected by the mutation to create the disease condition. Statins target that pathway, making them useful not only to individuals suffering from FH, but also as an effective treatment for high LDL cholesterol in the general population.

### 26.3 Linkage Analysis

Historically, researchers have used linkage analysis to determine genetic variants which explain phenotypic variation. The goal is to determine which variants contribute to the observed pattern of phenotypic variation in a pedigree. Figure 26.3 shows an example pedigree in which squares are male individuals, circles are female individuals, couples and offspring are connected, and individuals in red have the trait of interest.

![Pedigree Example](image_url)

Figure 26.1: Examples of diseases and quantitative traits which have genetic components
6.047/6.878 Lecture 24: Medical Genetics – The Past to the Present

Figure 26.2: The drug development process

Figure 26.3: A pedigree which shows the inheritance of some trait

Linkage analysis relies on the biological insight that genetic variants are not independently inherited (as proposed by Mendel). Instead, meiotic recombination happens a limited number of times (roughly once per chromosome), so many variants cosegregate (are inherited together). This phenomenon is known as linkage disequilibrium (LD).

As the distance between two variants increases, the probability a recombination occurs between them increases. Thomas Hunt Morgan and Alfred Sturtevant developed this idea to produce linkage maps which could not only determine the order of genes on a chromosome, but also their relative distances to each other. The Morgan is the unit of genetic distance they proposed; loci separated by 1 centimorgan (cM) have 1 in 100 chance of being separated by a recombination. Unlinked loci have 50% chance of being separated by a recombination (they are separated if an odd number of recombinations happens between them). Since we usually do not know \textit{a priori} which variants are causal, we instead use genetic markers which capture other variants due to LD. In 1980, David Botstein proposed using single nucleotide polymorphisms (SNPs), or mutations of a single base, as genetic markers in humans [4]. If a particular marker is in LD with the actual causal variant, then we will observe its pattern of inheritance contributing to the phenotypic variation in the pedigree and can narrow down our search.

The statistical foundations of linkage analysis were developed in the first part of the 20th century. Ronald Fisher proposed a genetic model which could reconcile Mendelian inheritance with continuous phenotypes such as height [8]. Newton Morton developed a statistical test called the LOD score (logarithm of odds) to test the hypothesis that the observed data results from linkage [20]. The null hypothesis of the test is that the recombination fraction (the probability a recombination occurs between two adjacent markers) $\theta = 1/2$ (no linkage) while the alternative hypothesis is that it is some smaller quantity. The LOD score is essentially a log-likelihood ratio which captures this statistical test:

$$\text{LOD} = \frac{\log(\text{likelihood of disease given linkage})}{\log(\text{likelihood of disease given no linkage})}$$

The algorithms for linkage analysis were developed in the latter part of the 20th century. There are two main classes of linkage analysis: parametric and nonparametric [25]. Parametric linkage analysis relies on a model (parameters) of the inheritance, frequencies, and penetrance of a particular variant. Let $F$ be the set of founders (original ancestors) in the pedigree, let $g_i$ be the genotype of individual $i$, let $\Phi_i$ be the phenotype of individual $i$, and let $f(i)$ and $m(i)$ be the father and mother of individual $i$. Then, the
Figure 26.4: Representing a particular pattern of inheritance as an inheritance vector

likelihood of observing the genotypes and phenotypes in the pedigree is:

$$L = \sum_{g_1} \cdots \sum_{g_n} \prod_i \Pr(\Phi_i|g_i) \prod_{f \in F} \Pr(g_f) \prod_{i \in F} \Pr(g_i|g_{f(i)}; g_{m(i)})$$

The time required to compute this likelihood is exponential in both the number of markers being considered and the number of individuals in the pedigree. However, Elston and Stewart gave an algorithm for more efficiently computing it assuming no inbreeding in the pedigree [7]. Their insight was that conditioned on parental genotypes, offspring are conditionally independent. In other words, we can treat the pedigree as a Bayesian network to more efficiently compute the joint probability distribution. Their algorithm scales linearly in the size of the pedigree, but exponentially in the number of markers.

There are several issues with parametric linkage analysis. First, individual markers may not be informative (give unambiguous information about inheritance). For example, homozygous parents or genotyping error could lead to uninformative markers. To get around this, we could type more markers, but the algorithm does not scale well with the number of markers. Second, coming up with model parameters for a Mendelian disorder is straightforward. However, doing the same for non-Mendelian disorders is non-trivial. Finally, estimates of LD between markers are not inherently supported.

Nonparametric linkage analysis does not require a genetic model. Instead, we first infer the inheritance pattern given the genotypes and the pedigree. We then determine whether the inheritance pattern can explain the phenotypic variation in the pedigree.

Lander and Green formulated an HMM to perform the first part of this analysis [15]. The states of this HMM are inheritance vectors which specify the result every meiosis in the pedigree. Each individual is represented by 2 bits (one for each parent). The value of each bit is 0 or 1 depending on which of the grand-parental alleles is inherited. Figure 26.4 shows an example of the representation of two individuals in an inheritance vector.

Each step of the HMM corresponds to a marker; a transition in the HMM corresponds to some bits of the inheritance vector changing. This means the allele inherited from some meiosis changed, i.e. that a recombination occurred. The transition probabilities in the HMM are then a function of the recombination fraction between adjacent markers and the Hamming distance (the number of bits which differ, or the number of recombinations) between the two states. We can use the forward-backward algorithm to compute posterior probabilities on this HMM and infer the probability of every inheritance pattern for every marker.

This algorithm scales linearly in the number of markers, but exponentially in the size of the pedigree. The number of states in the HMM is exponential in the length of the inheritance vector, which is linear in the size of the pedigree. In general, the problem is known to be NP-hard (to the best of our knowledge, we cannot do better than an algorithm which scales exponentially in the input) [22]. However, the problem is important not only in this context, but also in the contexts of haplotype inference or phasing (assigning alleles to homologous chromosomes) and genotype imputation (inferring missing genotypes based on known genotypes). There have been many optimizations to make this analysis more tractable in practice [1, 9–14, 16, 18].

Linkage analysis identifies a broad genomic region which correlates with the trait of interest. To narrow down the region, we can use fine-resolution genetic maps of recombination breakpoints. We can then identify the affected gene and causal mutation by sequencing the region and testing for altered function.
26.4 Genome-wide Association Studies

Linkage analysis has proven to be highly effective in studying the genetic basis of Mendelian (single gene) diseases. In the past three decades, thousands of genes have been identified as contributing to Mendelian diseases. Figure 26.5 shows this explosion in published associations. We have identified the genetic basis of disease such as sickle cell anemia, cystic fibrosis, muscular dystrophy, and severe forms of common diseases such as diabetes and hypertension. For these diseases, mutations are severe and obvious; the environment, behavior, and chance have little effect.

However, most diseases (and many other traits of interest) are not Mendelian. These complex traits arise from the interactions of many genes and possibly the environment and behavior. A canonical complex trait is human height: it is highly heritable, but environmental factors can affect it. Recently, researchers have identified hundreds of variants which are associated with height [2, 19].

Linkage analysis is not a viable approach to find these variants. In the 1990s, researchers proposed a methodology called genome-wide association to systematically correlate markers with traits. These studies sample large pools of cases and controls, measure their genotypes at on the order of one million markers, and try to correlate variation in their genotypes with their variation in phenotype.

Genome-wide association studies (GWASs) are possible due to three advances. First, advances in our understanding of the genome and the creation of genomic resources. The key biological insight is the fact that humans are one of the least genetically diverse species. On the order of tens of millions of SNPs are shared between different human subpopulations. For any particular region of the genome, we observe only a limited number of haplotypes (allele combinations which are inherited together). Because of this high redundancy, we only need to measure a fraction of all the variants in the human genome in order to capture them all with LD. We can adapt the algorithms for inferring inheritance patterns in linkage analysis to impute genotypes for the markers which we did not genotype.

Genome resources allow us to carefully choose markers to measure and to make predictions based on markers which show statistically significant association. We now have the reference sequence of the human genome (allowing for alignments, genotype and SNP calling) and HapMap, a comprehensive catalog of SNPs in humans. We also have genome-wide annotations of genes and regulatory elements.

Second, advances in genotyping technology such as microarrays and high-throughput sequencing. Although there are many types of variation in the human genome (Figure 26.6 shows some examples), SNPs are the vast majority. They are also the easiest and cheapest to measure using these technologies. However, we still need to account for the other types of variants. Recently developed DNA microarrays can detect copy-number variation in addition to SNPs.

The third advance is a new expectation of collaboration between researchers. GWASs rely on large sample sizes to increase the power (probability of a true positive) of statistical tests. The explosion in the number of published GWASs has allowed for a new type of meta-analysis which combines the results of several GWASs for the same phenotype to make more powerful associations. Meta-analysis accounts for various technical and population-genetic biases in individual studies. Researchers who conduct GWASs are expected to collaborate with others who have conducted GWASs on the same trait in order to show replicability of results. By pooling
Figure 26.6: Different types of genetic variation

Figure 26.7: (A) Manhattan plot and (B) Q-Q plot for GWAS of Crohn’s disease

together the data, we also have more confidence in the reported associations.

The main problem in conducting GWASs is eliminating confounding factors. First, genotyping error, which is common enough to require special treatment regardless of which technology we use. To account for such errors, we use thresholds on metrics like minor allele frequency and deviation from Hardy–Weinberg equilibrium and throw out SNPs which do not meet the criteria. Second, systematic genetic differences between human subpopulations. There are several methods to account for this population substructure such as genomic control [6], structured association [24], and principal component analysis [21, 23]. Third, covariates such as environmental and behavioral effects. We can account for these by including them in our statistical model.

The statistical analysis involved in GWAS is fairly straightforward. We assume the effect of each SNP is independent and additive to make the analysis tractable. For each SNP, we perform a hypothesis test whose null hypothesis is that the observed variation in the genotype at that SNP across the subjects does not correlate with the observed variation in the phenotype across the subjects. Because we perform one test for each SNP, we need to deal with the multiple testing problem. Each test has some probability of giving a false positive result, and as we increase the number of tests, the probability of getting a false positive in any of them increases. There are several methods to account for multiple testing such as Bonferroni correction and measures such as the false discovery rate [3] and the irreproducible discovery rate [17].

In addition to reporting SNPs which show the strongest associations, we typically also use Manhattan plots to show where these SNPs are located in the genome and quantile-quantile (Q-Q) plots to detect biases which have not been properly accounted for. A Manhattan plot is a scatter plot of log-transformed p-values against genomic position (concatenating the chromosomes). In Figure 26.7A, the points in red are those
which meet the significance threshold. They are labeled with candidate genes which are close by.

A Q-Q plot is a scatter plot of log-transformed observed p-values against log-transformed expected p-values. We use uniform quantiles as the expected p-values: assuming there is no association, we expect p-values to be uniformly distributed. Deviation from the diagonal suggests p-values are more significant than would be expected. However, early and consistent deviation from the diagonal suggests too many p-values are too significant, i.e. there is some bias which is confounding the test. In Figure 26.7B, the plot shows observed test statistic against expected test statistic (which is equivalent). Considering all markers includes the Major Histocompatibility Complex (MHC), which is the region associated with immune response. This region has a unique LD structure which confounds the statistical analysis, as is clear from the deviation of the black points from the diagonal (the gray area). Throwing out the MHC removes much of this bias from the results (the blue points).

GWAS identifies markers which correlate with the trait of interest. However, each marker captures a neighborhood of SNPs which it is in LD with, making the problem of identifying the causal variant harder. Typically, the candidate gene for a marker is the one which is closest to it. From here, we have to do further study to identify the relevance of the variants which we identify. For example, we can identify biochemical pathways which are disrupted by the causal variant and develop novel therapeutics based on the identified disruption. We can use gene expression data to learn about the cellular context in which the variant operates. We can also use genome-wide maps of functional annotations to identify regulatory elements which are affected by variants.

It is important to note GWAS captures more variants than linkage analysis. Linkage analysis identifies rare variants which have negative effects. GWAS can also identify these variants, but in addition can identify rare variants which have protective effects. Linkage analysis cannot identify these variants because they are anti-correlated with disease status. More importantly, GWAS can identify common variants with smaller effect sizes. Linkage analysis relies on the assumption that a single variant explains the disease. But this assumption does not hold for complex traits such as disease. Instead, we need to consider many markers in order to explain the genetic basis of these traits.

We have learned several lessons from GWAS. First, fewer than one-third of reported associations are coding or obviously functional variants. Second, only some fraction of associated non-coding variants are significantly associated to expression level of a nearby gene. Third, many are associated to regions with no nearby coding gene. Finally, the majority of reported variants are associated to multiple autoimmune or inflammatory diseases. These revelations indicate that there are still many mysteries lurking in the genome waiting to be discovered.

26.5 Current Research Directions

26.6 Further Reading

26.7 Tools and Techniques

26.8 What Have We Learned?

In the past several decades, we have made huge advances in developing techniques to investigate the genetic basis of disease. Historically, we have used linkage analysis to find causal variants for Mendelian disease with great success. More recently, we have used genome-wide association studies to begin investigating more complex traits with some success. However, more work is needed in developing methods to interpret these GWAS and identifying causal variants and their role in disease mechanism. Improving our understanding of the genetic basis of disease will us develop more effective diagnoses and treatments.

Bibliography


CHAPTER	TWENTYSEVEN

MISSING HERETIBILITY

TODO: missing @scribe: add author

27.1 Introduction
27.2 Current Research Directions
27.3 Further Reading
27.4 Tools and Techniques
27.5 What Have We Learned?

Bibliography
28.1 Introduction

George Church discussed a variety of topics that have motivated his past and present research. He first discussed about reading and writing genomes, including his own involvement in the development of sequencing and the Human Genome Project. In that latter half, he discussed about his more recent endeavor, the Personal Genome Project, which he initiated in 2005.

28.2 Reading and Writing Genomes

As a motivation, consider the following question: Is there any technology that is not biologically motivated or inspired? Biology and our observations of it influence our lives pervasively. For example, within the energy sector, biomass and bioenergy has always existed and is increasingly becoming the focus of attention. Even in telecommunications, the potential of quantum-level molecular computing is promising, and is expected to be a major player in the future.

Church has been involved in molecular computing in his own research, and claims that once harnessed, it has great advantages over their current silicon counterparts. For example, molecular computing can provide at least 10% greater efficiency per Joule in computation. More profound perhaps is its potential effect on data storage. Current data storage media (magnetic disk, solid-state drives, etc.) is much less (billions times) dense than DNA. The limitation of DNA as data storage is that it has a high error rate. Church is currently involved in a project exploring reliable storage through the use of error correction and other techniques.

In a 2009 Nature Biotechnology review article [1], Church explores the potential for efficient methods to read and write to DNA. He observes that in the past decade there has been a $10\times$ exponential curve in both sequencing and oligo synthesis, with double-stranded synthesis lagging behind but steadily increasing. Compared to the $1.5\times$ exponential curve for VLSI (Moore’s Law), the increase on the biological side is more dramatic, and there is no theoretical argument yet for why the trend should taper off. In summary, there is great potential for genome synthesis and engineering.
Did You Know?

George Church was an early pioneer of genome sequencing. In 1978, Church was able to sequence plasmids at $10 per base. By 1984, together with Walter Gilbert, he developed the first direct genomic sequencing method [3]. With this breakthrough, he helped initiate the Human Genome Project in 1984. This proposal aimed to sequence an entire human haploid genome at $1 per base, requiring a total budget of $3 billion. This quickly played out into the well-known race between Celera and UCSC-Broad-Sanger. Although the latter barely won in the end, their sequence had many errors and gaps, whereas Celera’s version was much higher quality. Celera initially planned on releasing the genome in 50 kb fragments, which researchers could perform alignments on, much like BLAST. Church once approached Celera’s founder, Craig Venter, and received a promise to obtain the entire genome on DVD after release. However, questioning the promise, Church decided instead to download the genome directly from Celera by taking advantage of the short fragment releases. Using automated crawl and download scripts, Church managed to download the entire genome in 50 kb fragments within three days!

28.3 Personal Genomes

In 2005, George Church initiated the Personal Genome Project [2]. Now that sequencing costs have rapidly decreased to the point that we can currently get the entire diploid human genome for $4000 (compare to $3 billion for a haploid human genome in the Human Genome Project), personal genome and sequence information is becoming increasingly affordable.

One important application for this information is in personalized medicine. Although many diseases are still complicated to predict, diagnose, and study, we currently already have a small list of diseases that are highly predictable from genome data. Examples include phenylketonuria (PKU), BRCA-mutation-related breast cancer, and hypertrophic cardiomyopathy (HCM). Many of these and similar diseases are uncertain (sudden onset without warning symptoms) and not normally checked for (due to their relative rareness). As such, they are particularly suitable as targets for personalized medicine by personal genomes, because genomic data provide accurate information that otherwise cannot be obtained. Already, there are over 2500 diseases (due to ∼6000 genes) that are highly predictable and medically actionable, and companies such as 23andMe are exploring these opportunities.

As a final remark on the subject, Church remarked on some of his personal philosophy regarding personalized medicine. He finds many people reluctant to obtain their genomic information, and attributes this to a negative view among the general public toward GWAS and personalized medicine. He thinks that the media focuses too much on the failure of GWAS. The long-running argument against personalized medicine is that we should focus first on common diseases and variants before studying rare events. Church counterargues that in fact there is no such thing as a common disease. Phenomena such as high blood pressure or high cholesterol only count as symptoms; many ‘common diseases’ such as heart disease and cancer have many subtypes and finer categories. All along, lumping these diseases into one large category only has the benefit of teaching medical students and to sell pharmaceuticals (e.g., statins, which have fared well commercially but only benefit very few). Church argues that lumping implies a loss of statistical power, and is only useful if it is actually meaningful. Ultimately, everyone dies due to their own constellation of genes and diseases, so Church sees that splitting (personalized genomics) is the way to proceed.

Personal genomics provide information for planning and research. As a business model, it is analogous to an insurance policy, which provides risk management. As an additional benefit however, the information received allows for early detection, and consequences may even be avoidable. Access to genomic information allows one to make more informed decisions.

28.4 Current Research Directions

28.5 Further Reading

Personal Genome Project: [http://www.personalgenomes.org/](http://www.personalgenomes.org/)
28.6 Tools and Techniques

28.7 What Have We Learned?

Bibliography


