# CS262 Lecture 17: Human Population Genomics

Scribe: Arushi Raghuvanshi

March 2016

## 1 Lecture Overview

In this lecture, we continue to discuss multiple sequence alignment and human population genomics. MUSCLE and PROBCONS are two methodologies for multiple sequence alignment that we will analyze in detail. MUSCLE is much faster, but PROBCONS is slightly more accurate. Both of these algorithms are often used for protein sequence alignments. The biggest application of this is phylogenetics.

For human population genomics, we discuss evolution and migration theories as well as association studies. Association studies are useful, because they allow us to draw correlations between genomes and phenotypes. For example, WTCCC found correlations between common diseases and a few genes. However, we still do not know how to predict phenotypes from a genome. This is an open problem, that will hopefully be solved soon.

## 2 Multiple Sequence Alignment (contd)

In the previous lecture, we talked about multiple sequence alignments, and we started talking about two highly successful multiple alignment systems that put together a lot of topics that we've discussed in this course. The two systems are MUSCLE and PROBCONS, described in more detail below. Both of these systems are typically used for protein sequences. MUSCLE is much faster, and can bu used to align 1000s of protein sequences. PROBCONS is much slower, but slightly more accurate. Because of the time constraint, it can only be used to align a few hundred sequences.

### 2.1 MUSCLE

MUSCLE stands for Multiple Sequence Alignment by Log Expectation. You can use it here: `http://www.ebi.ac.uk/Tools/msa/muscle/`

This method is concentrated on being extremely fast. We want to avoid an $O(N^2L^2)$ and higher computational steps. Since N and L are roughly the same order of magnitude (about 1000), we want to avoid anything of the 4th power of N and L combined.

In order to do a multiple sequence alignment, we can take the following steps:

1. Build a good phylogeny tree, so you can do a good progressive alignment.
2. Given a good tree, we can create an alignment in $O(L^2N)$ time, where each step corresponding to an internal node of the tree.

Note that we need distances to create a tree. Computing the distances by going through every pair of sequences and doing an alignment would take $O(N^2L^2)$ time. In order to avoid the $O(N^2L^2)$ running time, the MUSCLE algorithm uses the following heuristics to build the tree.

**MUSCLE algorithm:**

1. Create a draft of the distances by counting the number of k-mer matches. This is a very fast measurement of all pairwise distances, because counting k-mer matches is simple (essentially linear time). Therefore computing the metric $D_{DRAFT}(x, y)$ takes $O(N^2L * log(L))$ time. We usually use k as about 3.
2. Build a draft tree $T_{DRAFT}$ based on the draft distances distances using UPGMA. We use UPGMA, because it was empirically shown to work better.
3. Create a progressive alignment over $T_{DRAFT}$, resulting in a multiple alignment $M_{DRAFT}$. This takes $O(NL^2)$ time.
4. Measure new Kimura-based distances $D(x, y)$ based on $M_{DRAFT}$. These new distances will be more accurate than the original $D_{DRAFT}$ distances.
5. Now we can build the tree T based on D
6. Do a progressive alignment over T, to build M
7. Perform an iterative refinement as follows.
    (a) Tree partitioning. Take a random branch, and split M on it. Then, realign the two resulting profiles.
    (b) If the new alignment $M'$ has a better sum-of-pairs score, then accept the refinement.
    (c) Keep repeating these steps for many rounds or until convergence.

Note there are many scoring functions we could use. The sum-of-pairs score was chosen by testing 32 scoring functions, and picking the one that worked best empirically.

## 2.2 PROBCONS

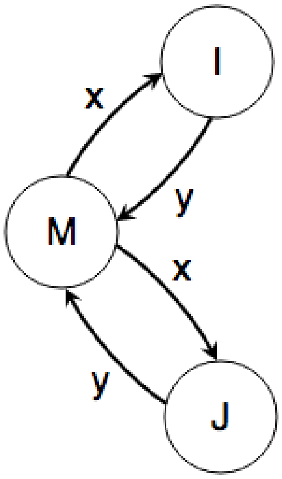
PROBCONS is short for Probabilistic Consistency-based Multiple Alignment of Amino Acid Sequences. You can learn more about it here: `http://probcons.stanford.edu/`.

The idea of this algorithm is to use the consistency heuristic.

We can illustrate the idea of **consistency** as follows:

- We are given a number of sequences to align
- We start by aligning two sequences x and y across a progressive alignment setting.
- As we align x and y, it may be about the same score to align $x_i$ to $y_j$ and $y'_j$.
- However, when we align both to some sequence z later, it might lead to favoring one over the other. So we should consider z when deciding to align $x_i$ to $y_j$ or $y'_j$

PROBCONS tries to use this idea to create an optimal alignment based on a probabilistic model. It is much slower than MUSCLE, but was shown to be slightly more accurate. We can think of it as a pair-HMM for alignment with the following structure.



**PROBCONS algorithm:**

1. Compute the posterior probability that $x_i$ aligns to $y_j$ for each x,y,i,j. Fill the result into posterior matrices $M_{xy}$. $M_{xy}(i,j) = P(x_i \ y_i)$. Recall that the posterior probability is the probability of alignment given the observed sequences, and it can be calculated using an HMM. This takes $O(N^2L^2)$ time, which is already greater than the MUSCLE running time.

2. Do a re-estimation of posterior matrices $M'_{xy}$ with probabilistic consistency as follows. (Note that comparing all pairings takes $O(N^3L^3)$ time in the worst case which is infeasible on real data. We do this re-estimation as a faster heuristic. It is still $O(N^3L^3)$ in the worst case, but in practice is linear with a very large constant.)

   (a) $M'_{xy}(i,j) = \frac{1}{N}\sum_z \sum_k M_{xz}(i,k)xM_{yz}(j,k)$

   (b) $M'_{xy} = Avg_z(M_{xz}M_{zy}$

3. Now, for every pair x,y compute the maximum expected accuracy alignment as follows.

   (a) $A_{xy}$: alignment that maximizes $\sum_{aligned(i,j)inA} M'_{xy}(i,j)$

   (b) Define $E(x,y) = \sum_{aligned(i,j)inA_{xy}} M'_{xy}(i,j)$

4. Now build the tree T with hierarchical clustering using similarity measure $E(x,y)$. Note that this tree is not a phylogeny tree. It is the tree that looks like it will give us the fewest possible errors.

5. Do a progressive alignment on T to maximize $E(.,.)$. Note that this alignment is optimizing a different objective than the standard Needleman-Wulch alignment. This alignment maximizes the expected number of correct positions ($\sum_{aligned(i,j)} M'_{xy}(i,j)$)

6. Do iterative refinement by randomized partitioning as follows.

   (a) Split sequences in M in two subsets by flipping a coin for each sequence

   (b) Realign the two resulting profiles

   (c) If the projected profiles are better, accept

   (d) Repeat these steps for many iterations or till convergence.

**Algorithm Analysis**

The re-estimation heuristic assumes independence of the the probabilities of the 3 alignments. In the worst case, step 2 would still take $O(N^3L^3)$ time. However, in practice, these matrices are extremely sparse. We drop any probability less than .001 to 0. Then, the sparse matrix multiplication is essentially only calculating about 10 positions per entry in x, y, z. Therefore, the re-estimation step is essentially linear in time with some significantly large constant in the length of the sequences. Therefore, the second step is about the same or faster than the first step. However, PROBCONS still has a bottleneck at the second step if there many sequences, so it shouldn't be used to align more than a few hundred sequences.

## 2.3 Further Discussion

Trees are a very intuitive way to approach multiple sequence alignment, but not necessarily the most effective.

In this class we focus on distance matrices for tree generation and sequence alignment. Distances correspond to substitutions to a site, which is an intuitive metric for how related different sequences are.

There are researches that focus on many different techniques for phylogeny trees. One such technique is Parsimony based. In parsimony based trees, the goal is to minimize the number of mutations to explain a set of observed sequences. These methods simultaneously find the tree and the associated mutations.

For more information on parsimony techniques, refer to this paper: `http://bioinformatics.oxfordjournals.org/content/23/2/e123.short`

## 2.4 Genome Evolutionary Rate Profiling (GERP)

### 2.4.1 Motivation

Once we have a multiple alignment, and a tree that corresponds to substitutions per site, how how do we use this for useful analysis? One simple question:

> Which regions of the genome are conserved across evolution?

This is a simple question, but it doesn't have a simple solution. It has been answered in a multitude of ways that aren't principled. The idea is to quantify the conserved regions as the regions where the number of substitutions per site is much less than some average location in the genome. But at what scale do we do that?

Understanding how much of the human genome has evolutionary pressure to be preserved is interesting, because it highlights the part of the genome that is essential for a functional phenotype (purifying selection). The rest of the genome is presumably junk.

Many techniques have been applied to this question. Each gives a different estimate for how much of the genome is conserved.

### 2.4.2 ENCODE project

ENCODE stands for ENCyclopedia Of DNA Elements. In 2003, the project was launched to identify all functional elements in the human genome. They used both computational and experimental methods.

One experimental method is sequence probing the human genome in multiple cell lines, and doing a sequencing of the RNA products. You can take DNA of a culture of cells, use DNase, and parts that were cut were those that were presumably active at the time.

The ENCODE project found that 80 percent of the human genome is functional, but received a lot of backlash for making this claim. In reality, many of these functions are not very important, because they don't actually have a reflection in the phenotype.

You can read more about ENCODE here: `http://www.genome.gov/ENCODE/`
and here: `http://www.nature.com/nature/journal/v489/n7414/full/nature11247.html`

### 2.4.3 GERP

GERP is one technique for identifying conserved region.

**GREP approach**

- Given a tree and the multiple alignment, you can ask if every position is faster or slower than the average substitutions per site for the tree.
- Multiply the tree by a scaling factor k (¡1 means fewer subs per site, ¿1 more subs per site).
- For each position, find the maximum likelihood k scaling factor that makes this pattern of substitutions most likely.
- Highly conserved positions will have very small k, highly diverse positions will have very large k.
- Each position has a number of substitutions kS - S. This is a measure of how many substitutions we have additionally or fewer than expected
- For a window of any length, consider if this is likely to happen by chance or not.
- Negative values are associated with rejected substitutions. So conserved areas will have many negative values that are very unlikely to happen by chance.

The GREP methodology finds that 8 percent of the human genome is preserved. Some other methods have lower estimations of about 5 percent. Note the assumption in many of these methods (including GERP) is that mutations are independent. This isn't a very good assumption in principle, because once a gene sequence is broken, there's no more pressure to preserve a broken gene.

# 3 Human Population Genomics

## 3.1 Human population migrations

### 3.1.1 Out of Africa, Replacement Evolution

- 30 million years ago was the age of the great apes

- Homo sapiens began to evolve about 200,000 years ago

- Humans moved out of Africa about 50,000 years ago, replacing other contemporaries (e.g. Neandertals).

- This suggests that people today share a relatively modern African ancestry

- We can trace ancestry of the paternal line through the y chromosome. The y chromosome does not recombine, so you get it entirely from your father. Therefore, it makes sense to talk about the ancestor of the y-chromosome (Adam). This genetic Adam is estimated to have lived about 120,000 to 340,000 years ago.

- We can trace the ancestry of the maternal lineage through the mitochondria (since it is only passed through mother and only present in the egg). The genetic Eve is estimated to have lived 99,000 to 150,000 years ago.

### 3.1.2 Multiregional Evolution

- This is an alternative hypothesis for how humans evolved

- It suggests that great apes settled in many areas and humans evolved separately with regional varieties

- This hypothesis was largely debunked from both fossil records and genetics

- This hypothesis was historically used with racist connotations to describe certain groups of people as superior (i.e. Nazis)

- We now know that Neanderthals were not a predecessor of modern humans, but their contemporary

- There was enough breeding between Neanderthals and humans that people today with European and Asian ancestry are about 4 % neanderthal.

- People today with African ancestry are more purely homo sapien than people from Europe or Asia

- It seems like it was a quite common phenomenon for there to be breading between Neanderthal and Homo Sapiens, this leads to the idea of **coalescence**, that population genetics relates to historical genetic diversity

### 3.2 Alleles and Mutations

Most positions in the human genome, are diallelic, meaning they have two alleles. For diallelic positions, the majority of the population usually has one of the alleles, and a small proportion of the population has the other. There are also very few triallelic and a handful of quadallelic positions.

Between the paternal germ line and the sperm or egg genome, the accumulated number of mutations is about 100. The father is the biggest contributor to mutations, and the age of father is correlated to the number of mutations.

We have 700 billion new mutations in the currently existing new generation. We have about 230 mutations per position, dividing by three we see that about 80 individuals have every alternative base.

The probability of heterozygosity is the probability of picking 2 alleles at random with replacement that are different. This is given by the following equation:
$H = \frac{4\nu}{1+4\nu}$

### 3.3 The Neanderthal Whole Genome

#### 3.3.1 Identical by descent (IBD)

Two pieces of the genome are identical by descent if they are derived by the same very recent ancestor (i.e. you and your sister will have 50% of the genome IBD).

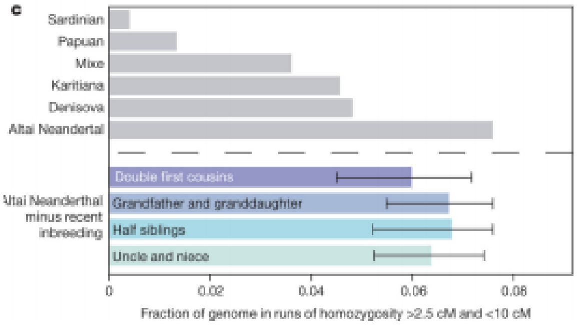
Looking across an entire genome, we can note how much is identical to the other part. There are positions where the two alleles often differ given the distribution in the population. If we see a region with a lack of these mutations, that region will be IBD.

In small communities, individuals often have a high degree of IBD. This sometimes happens in larger populations as well that share some similarities.

#### 3.3.2 Sequenced Neanderthal

The sequenced Neanderthal had a very high degree of IBD. 8 percent was identical. This amount of IBD is much more than very small populations and is equivalent to an extremely high degree of inbreeding (i.e. double first cousins). Consider the graph below that further illustrates how unusual this is by modern standards.

One possible explanation is that the individual Neanderthal that was sequences was an anomaly. The other more likely explanation is that this is how Neanderthals lived. They existed in very small tribes and kept breading with each other. This is an incredibly interesting result, because it illustrates how genetics can give us insights into cultural phenomenons of ancient times.

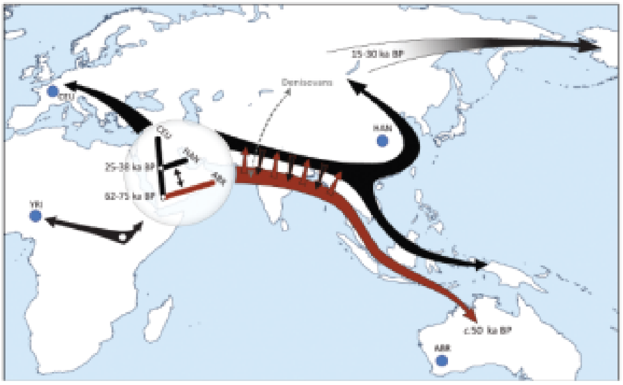


## 3.4 Aboriginal Australian

An analysis of Aboriginal Australians showed an updated version and further support of the out of Africa evolution theory.

There was a waive of migration earlier than the main wave of migration. People from this first wave went to Australia and South East Asia. The second migration wave caught up and interbred with the first wave, but some went to different places as illustrated in the diagram below.

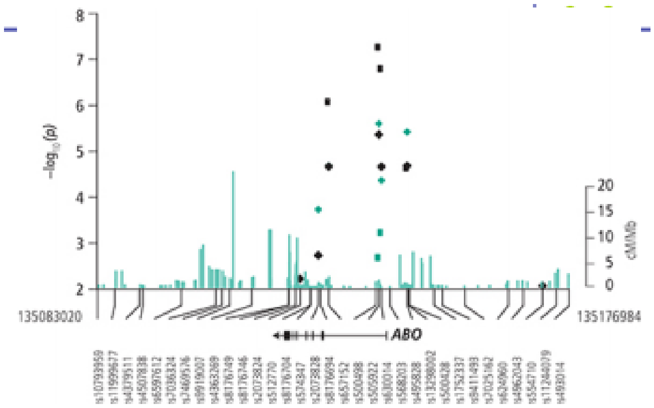
Due to this first wave, Aboriginal Australians have a very high degree of descent from the pure ancient homo sapien species.



## 3.5 Association Studies

Association studies is a very important topic in population genetics. We want to link alleles or versions of our genome with specific phenotype traits. In particular, disease pheotypes are important to study.

We can study association in a binary setting with a set of genomes that result in a disease and a cohort of genomes that are non-diseased. We then find alleles that have a much higher frequency in the disease cohort to the non-diseased ones. We can get the p values and create a Manhatten plot as illustrated below. On the x axis of the plot, write the genome. On the y axis, write $-log_{10}$ of the p value. If the p value is small, the negative log of p will be a very high value. Many high values in nearby positions in plot is due to linkage to equilibrium. This means that alleles associated with a trait will have nearby alleles also associated with that trait.



Note that it is difficult to find causal allele, but we can find many important ones.

## 3.6 Wellcome Trust Case Control Consortium (WTCCC)

WTCCC was the first important large scale association study. It was conducted for 7 frequent diseases. For each disease, there was 2k subjects that had the disease and 3k without.

They weren't sequenced (because at the time it was too expensive). Instead, micro array based genotyping was done. 500k positions were genotyped for each individual.

They formed Manhatten plots and for each disease found a few genes that slightly increase your probability of having the disease. The results of this study slightly disappointing, because while there was some association, we realized that it is not easy to explain disease phenotypes with a handful of genes.

## 3.7 Heritability Environment

When studying twins separated at birth, we can compare correlation of a given trait with genetics. Study after study has shown that many characteristics of a person are very highly genetic traits (including intelligence, extroversion, disorders, etc.). While we know that they are correlated, however, we can't yet predict it from the genome. This is a open problem that will hopefully be solved in the upcoming years.