Part IV – Comparative Genomics

**Slide 1:** Comparative Genomics

**Slide 2:** In this slide deck, we’ll examine the importance of genome comparisons. We will introduce different scoring matrices that are used to quantify the similarities in nucleotide and protein alignments. Finally, we will end by doing an exercise to learn about some different alignment tools.

**Slide 3:** Before the advent of sequencing technologies, organisms were classified based on morphological characteristics. Genome sequences now make it possible to create more accurate classifications based on molecular information.

**Slide 4:** The first question to ask then is how to compare different genes or genomes or proteins?

The top of the slide shows two protein sequences aligned with each other. Identical amino acids at the same position are highlighted by an asterisk (\*) symbol. In this representation, amino acids that are not identical but share similar properties are represented by a colon (:) In places where there are non-identical amino acids which are not in the same chemical/physical properties family, there is no symbol or a gap (-). Not all representation schemes show the same symbols, but it’s usually easy to figure out by inspection.

**Slide 5:** We can talk about two kinds of alignments: In this slide, we’re defining **Global Alignment** which is spread across the length of the sequence to include as many matches as possible. This method is used in cases where we know the two sequences come from organisms that are closely related and the sequences are of about equal length. This is based on the Needleman-Wunsch algorithm.

**Slide 6:** The second kind of alignment is **Local Alignment.** These are confined to small regions of strong similarity between two sequences. These are useful to look at sequences which are distantly related but may still have a small region of conserved DNA or protein domains. This algorithm is based on the Smith-Waterman algorithm and forms the basis of the BLAST search tool.

**Slide** 7**:** How do we know when an alignment is good? In this example, we can easily see by eye which alignment has more matches (Hint: It’s alignment I). But if we want a quantitative measure we need some kind of scoring system. The scoring system is a set of rules used to evaluate sequence alignments.

**Slide 8:** Definition of a scoring system – developed for understanding the likelihood by which one character state (meaning a nucleotide or an amino acid) will be changed to another character state over time. We’ll first talk about scoring systems for nucleotides and then for proteins.

**Slide 9:** One of the simplest models of substitution is the Jukes-Cantor model, which assumes that all nucleotide substitutions occur with equal probabilities. These are represented by similar sized arrows and similar sized circles.

**Slide 10:** A more realistic model is Kimura’s 2 parameter model which takes into account the probability that a transition is more likely than a transversion.

**Slide 11:** A third model is the HKY85 model, which considers unequal base frequencies and unequal probabilities of substitution. The thicker arrows indicate higher probability of substitution. Also, the size of the circle represents the base frequencies. The HKY85 is the most complicated model of the three we’re presenting and tries to mimic real-life substitution changes. There are many others as well (you’ll see this when you do the Phylogeny Exercise after Slide Deck V - Phylogenetics).

**Slide 12:** For proteins, the same general idea holds true – we assign scores to each place along a string of letters. Retention of the letter gets a high score, substitution gets a lower score.

Here, a reminder about the chemical and physical properties of amino acids. Amino acids with similar properties are grouped together in this diagram. In cases where there is a mutation resulting in a change of amino acid in a sequence, we can make an inference about whether the observed change might affect the structure and hence the function of the protein. For example, substitution of a threonine for a serine in a sequence might not be too radical because the two are close in size and both have hydroxyl groups. On the other hand, substitution of glutamate for phenylalanine might have big effects because you’ve exchanged a negatively charged amino acid for one that is aromatic.

Scoring matrices for proteins then take into account the “cost” of substitution of one amino acid for another based on the chemical and physical properties of the amino acid side chains.

**Slide 13:** There are two main scoring systems for amino acid substitutions. The first one is the PAM matrix. PAM stands for Point Accepted Mutation. This image of the PAM matrix is showing the probability of an original amino acid represented in the columns being replaced by another amino acid represented in the rows over a defined interval of time.

What’s shown here is the PAM1 matrix, developed by Margaret Dayhoff and colleagues at Georgetown University in the late 1970’s. They were interested in comparing proteins that were closely related phylogenetically, ones that shared 85% similarity. You’ll notice the highest numbers are on the diagonal, which indicates that when comparing closely related proteins from different species, most of the time, the amino acid at a given spot stays the same. As an example, 9867 times of 10,000 comparisons, an alanine remained as an alanine, for example. Or tryptophan remained as a tryptophan 9976 times of 10,000 comparisons. Similarly, there are a lot of zeros in the matrix where a substitution of a particular amino acid with another never occurred. But looking at our alanine example again, we see that 21 times a glycine is found in alanine’s place in a companion protein and 18 times a valine is found in alanine’s place. This makes sense because alanine and glycine and valine are all small amino acids. If we look at tryptophan again, we see that most of the possible substitutions are 0 (meaning they were never observed). Dayhoff and colleagues define the interval of one PAM as a unit of evolutionary divergence in which 1 amino acid per 100 residues has changed between the sequences.

**Slide 14:** The PAM 1 matrix was created for alignments of proteins that share 85% identity. But obviously sometimes we’d like to compare proteins that are less related. Dayhoff and colleagues developed a series of matrices that can be used depending on how closely or distantly related the proteins are. In this slide is shown the PAM 250 matrix which is derived by multiplying the PAM1 by itself 250 times. This new matrix can now be used to compare proteins that share 20% amino acid identity. Note also that in this representation there are actual **scores:** keeping an amino acid the same gives you a high positive score, but substitutions give you a range of scores from low positive scores to negative scores. Example – keeping Cysteine in place results in a score of 12, but if substituted with Tryptophan, this results in a score of -8. This make sense because cysteine and tryptophan have very different chemical and physical properties. Note also that only the “bottom” half of the matrix is shown in this representation. PAM matrices by definition are symmetrical – substituting a cysteine for a tryptophan gives the same score as substituting a tryptophan for a cysteine.

**Slide 15:** Another scoring matrix is BLOSUM, Blocks Amino Acid Substitution Matrices. It is derived from the BLOCKS database of conserved regions of protein families. BLOSUM62 is the most commonly used matrix and is used for sequences that share 62% identity. Importantly, BLOSUM62 is the default scoring matrix in the BLAST search tool at NCBI.

**Slide 16:** Here is a comparison between the PAM and BLOSUM scoring matrices. PAM is used to score alignments between closely related protein sequences an it is based on global alignments. Whereas BLOSUM is used to score alignments between evolutionary divergent protein sequences and is based on local alignments.

**Slide 17:** When we align sequences together to compare them, we can either do it in a pair-wise manner or use a multiple sequence alignment approach. We’ll discuss pair-wise alignment first. The alignments help us to calculate how similar two sequences are. This sequence similarity could result in functional and structural similarity. Also, sequences that are similar to each other may be evolutionarily related.

**Slide 18:** There are different methods for pair-wise alignments.

One of the approaches for pair-wise alignment is creation of dot-matrix which provides a visual way to examine a comparison. An example is on the next slide.

**Slide 19:** This is an example of a dot matrix plot comparing two sequences of words: SEQUENCE ANALYSIS IS FUN and FUN SEQUENCES ARE FUN. Matches between the two sequences are represented by dots. If the two sequences are highly identical we would see the dots clustering on the main diagonal. In our example, we see areas where there are gaps because the two sequences do not match up. Large scatter in the dots indicates lower similarity between sequences.

**Slide 20:** A more practical way to do pairwise alignments is to use dynamic programming, where an algorithm matches all possible characters and uses a scoring system for matches (i.e. PAM or BLOSUM for proteins; Jukes-Cantor, etc. for nucleotides), mismatches and gaps. Again, this is method used by BLAST to find orthologs.

**Slide 21:** Multiple-sequence alignments forms the basis of comparative genomics. We can get information about assigning sequences to gene families or to functional domains of proteins. These patterns of similarities can be used to build statistical models and eventually form the basis of phylogenies.

**Slide 22:** There are several programs that can be used for multiple sequence alignment – For global alignment programs include T-Coffee, CLUSTAL-W and MUSCLE. For local alignments, we can use the MAFT program. Each of these programs use a different algorithm for making the alignment and we must be aware of our data in terms of sequence similarity and our requirement when picking a software to use. In addition, some programs require more computational time, although they provide robust alignments. For most purposes, we recommend MUSCLE. This program will be used in the Comparative Genomics exercise that follows this slide deck.

**Slide 23:** In summary, this slide deck covers the differences between local and global alignments, and pairwise and multiple sequence alignments. We also covered different scoring matrices that are used for nucleotide and protein sequence comparisons. Now we will do an exercise using a multiple alignment tool (MUSCLE) to make comparisons between different protein sequences.

Please do the Comparative Genomics Exercise before continuing.

The next slide deck in the series is Part V – Phylogenetics