# LSM2241
# (Sequence Comparisons, Part I)
# Supplemental notes and FAQ

**Greg Tucker-Kellogg**
**dbsgtk@nus.edu.sg**

19 August 2015

## Overview

This first discussion of sequence analysis is often one of the most challenging topics of the term. Many students find the material unfamiliar, and I want to address some of the common questions that may still be with you after the lecture. I am trying to be as accurate as possible, without overburdening you with mathematics and theory that is not part of the course expectations. I have also included questions that came up in previous terms. I address some of these directly in the lecture this term, but they are discussed here with the hope that more explanation will be useful. Thank you very much for your questions during class, after class, and in email.

## Could you explain more clearly the distinctions between types of homologs?

As a reminder:

- *Speciation* creates **orthologs** (orthologous sequences)
- *Duplication* creates **paralogs**

Turning this around to conventional definitions, we can say

**Speciation** The origin of new species from a common ancestor.

**Ortholog** Orthologs are genes in different species that arise by speciation. Generally, orthologs retain the same function in different species.

**Paralog** Paralogs are genes that are related by duplication within a genome. The duplicate copy can evolve towards new function, so paralogs do not in general retain the same function.

**Homolog** gene related to another gene by decent from a common ancestral DNA sequence. Homologues include both orthologs and paralogs.

When we perform sequence comparisons, we examine sequences that exist today, and make inferences about their history — how they originated by different means, and in some cases what their common sequence ancestors must have looked like.

A few additional points to note:

- While for the next few weeks we focus on sequence *alignment*, it is important for you to be comfortable with the idea that relationships between sequences are often represented as branching trees. We'll learn how to construct and interpret such trees, called phylogenetic trees, in week 7, but for now just keep a few important features in mind:

1. the branching of a tree represents divergence in evolutionary history.
2. the lengths of branches often represent some quantitative measure of that divergence.
3. because of point 2, the total distance measured along the branches from one sequence to another is *often* a estimate (the meaning of which we'll discuss in week 7) of how far two sequences have diverged from a common ancestor.

- Since orthologs arise from speciation, trees of orthologous sequences reflect species history. In other words, if we use ortholog sequences to construct a tree between species, the resulting tree has the same or very similar structure to a tree constructed from other data showing relationships between species. On your handouts, you will note the close proximity of horse and zebra, and the mutual proximity of primates including humans.
- For years (decades!) the most common usage of "paralog" was to consider paralogs as homologous genes *within a species*. For example, all human globin are paralogs. But this is a somewhat limiting and imprecise definition. When looking at more than one species, we can make our definition of paralogs more precise to describe *when* duplication occurred relative to the species being compared. This more precise nomenclature was first proposed in [2].

### How is a substitution matrix used to score alignments?

An alignment is scored by *scoring each position separately and independently, and then adding them up*. Consider an alignment between two nucleic acid sequences

```
ATGGCGT
||| :||
ATG-AGT
```

Suppose we use a scoring matrix shown below to score this alignment. For any scoring matrix, you look up any match or mismatch in the matrix to identify the appropriate score. Please note that this scoring matrix (like the PAM and BLOSUM matrices) is symmetrical about the diagonal: a `G-C` change is equivalent to a `C-G` change.

| | A | C | G | T |
|---|---|---|---|---|
| A | 1 | -1 | -1 | -1 |
| C | -1 | 1 | -1 | -1 |
| G | -1 | -1 | 1 | -1 |
| T | -1 | -1 | -1 | 1 |

We haven't yet discussed the question of how to score gaps, much less how to find the best alignments. These questions will be discussed next week.

Since we haven't discussed how to score the gaps, let's ignore them. Using this scoring matrix would give an alignment score of

```
A T G G C G T
| | |   : | |
A T G - A G T
1+1+1  -1+1+1 = 4
```

Each number on the bottom row is looked up from the corresponding cell of the substitution matrix.

### Could you explain more clearly about where we get the PAM scoring matrices?

I'll give it a shot!

The first step is to count substitutions (or mutations) between sequences. Dayhoff [1] did this by aligning closely related sequences of protein families. Because they were closely related, she could be confident that the alignment was right, even with the tools she had in the 1970s. She then constructed evolutionary trees to describe the alignments, and identified the *minimum set of changes required to explain the data* from divergence from a common ancestor. That's the tree in the slide on "what is an accepted point mutation". Each change is called an "accepted point mutation", and a 1 gets added to a table of mutation counts in the appropriate position. This is called the mutation count matrix, and shown in Figure 1.

If you were looking naïvely at the sequences shown in that tree, you might believe that there must be a `D/C` substitution, since both `D` and `C` can occur at the second position of the sequence. But that is *not required* in an evolutionary model, and so it is not included in the count matrix. This approach is called *parsimony*, which tries to establish the minimum number of changes required to explain the observed sequences.

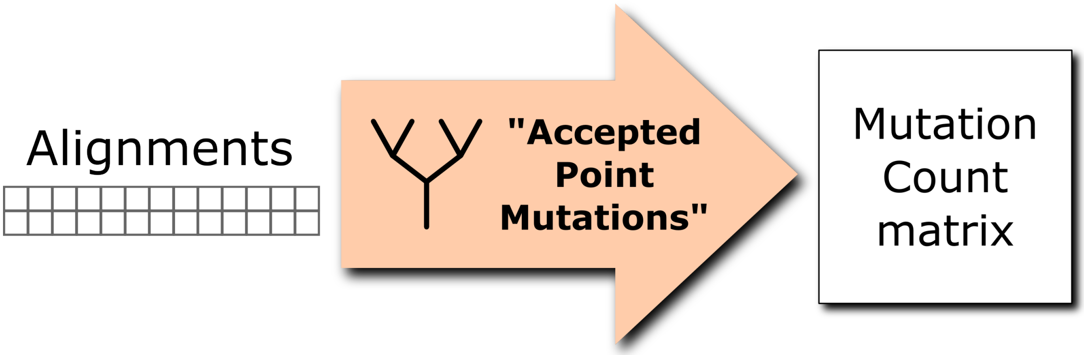

**Figure 1.** constructing a mutation count matrix from sequence alignments and accepted point mutations derived from evolutionary trees. This is the first step in constructing the PAM matrices

The PAM matrices go through successive stages of derivation. Only the last stage is used for scoring, and is the log odds matrix. The three stages are as follows:

| *Stage 1* | | *Stage 2* | | *Stage 3* |
|---|---|---|---|---|
| Mutation count matrix | → | Mutation probability matrix | → | scoring (log-odds) matrix |

Remember, *only the log odds matrix is used for scoring alignments*, and that is after all what we are trying to learn about. We went through the others so you would have an idea where the scoring matrix comes from, instead of having it dropped from the sky.

A graphical picture of the process is shown in Figure 2.

### More details (This part gets a little hairy.)

**Skip this section** unless you want to know more about the background, gain wisdom, and impress your friends taking LSM2251 and LSM2291.

The count matrix is just the counts of observed mutations in the setting studied by Dayhoff. It is symmetrical, but not normalized in any fashion whatsoever. The Gly-Asn cell, for example has the number "50" in it, because 5 observations of such mutations were made and the whole thing is multiplied by 10. (If you get a copy of the paper: occasionally, the ancestral amino acid was ambiguous, so some fractional entries would have appeared in the raw data. The whole thing is multiplied by 10 to make sure everything is an integer).

The mutation probability matrix *is* normalized, and has some special properties. Each column stands for one amino acid. The entries in that column are the probabilities that the amino acid will mutate given the

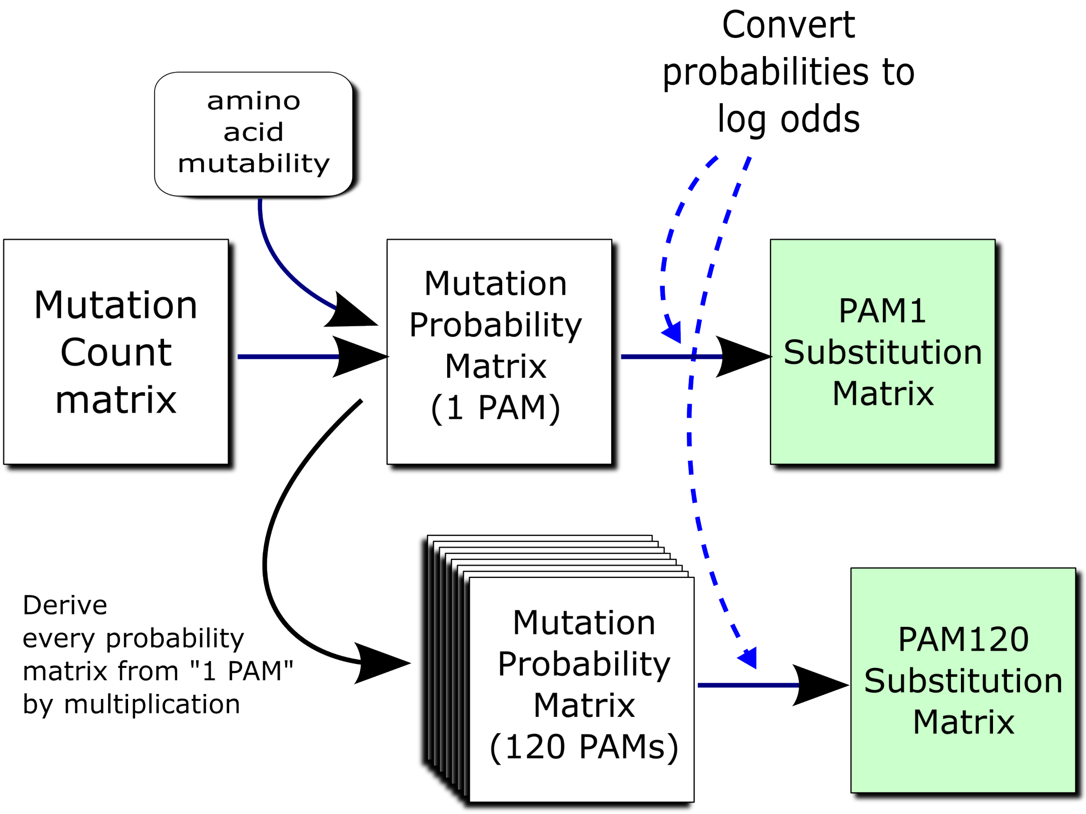
Convert probabilities to log odds
amino acid mutability
Mutation Count matrix
Mutation Probability Matrix (1 PAM)
PAM1 Substitution Matrix
Derive every probability matrix from "1 PAM" by multiplication
Mutation Probability Matrix (120 PAMs)
PAM120 Substitution Matrix

**Figure 2.** deriving a scoring matrix from counts of accepted point mutations. The mutation probability matrix for PAM1 is derived directly from the counts of APMs and the mutability of each amino acid. All the other mutation probability matrices are derived from the first. The resultant probability matrices are then converted to log odds form by the formula shown. The background values are the relative population of each amino acid. The example of PAM120 is shown, but every PAM matrix can be derived from this process.

opportunity to do so. The rows are each possible amino acid. The diagonals of the mutation probability matrix are the probabilities of *not* mutating. This matrix is not symmetrical because of the different amino acid frequencies. This matrix can be used to simulate the process of mutation: given an amino acid, choose at random the amino acid it might mutate into based on the probabilities in that column.

Why do we call this matrix a "Mutation Probability Matrix at 1 PAM"? Dayhoff cleverly introduced a scaling factor $\lambda$ which allowed her to decide how much mutation to allow. Remember that the diagonals $M_{ii}$ are the probabilities of each amino acid $i$ staying the same, right? So if we multiply each entry by the frequency of that amino acid to get $f_i M_{ii}$ and combine them up ($100 \times \sum f_i M_{ii}$) we get a number between 0-100 that is the percent of the time any amino acid (on average) will stay the same in a simulation. Dayhoff chose $\lambda$ so that

$$100 \times \sum f_i M_{ii} = 99$$

That is, the system is scaled so that there is a 99% chance of no mutation, but in an average sequence of 100 amino acids, one mutation will arise. That is what is meant by 1 PAM unit. The PAM1 mutation probability matrix can be used to simulate the process of introducing mutations, by taking each amino acid in the sequence, drawing a random number, deciding which if any amino acid it should be mutated to based on the corresponding probabilities, and moving on to the next one. 1% of the time, an amino acid will change.

In practice, the mutation probability matrix is a special type of mathematical object called a “stochastic matrix” for which there are numerical tools that do such simulations efficiently through matrix algebra. A more lengthy process can be simulated by repeating the process any number of times. Because of the special properties of stochastic matrices, a more lengthy random mutation process can also be simulated using a single matrix, which obtained by multiplying the probability matrix by itself the required number of times. That's how the higher numbered PAM matrices are generated.

*Did you skip that section? You really could have, you know.*

### Why do you call PAM1 a measure of evolutionary distance?

PAM1 is *defined* so that it can function as a unit of evolutionary distance. It is a small amount of expected evolutionary change (1% accepted point mutations) than can be modeled by a probability matrix. A longer process (e.g., 120 PAMs of evolutionary change) is modeled as a *Markov process*, a series of 120 steps of 1 PAM worth of change.

### When constructing sequence alignments, do we need to figure out which mutations must have occurred (like the Accepted Point Mutations of Dayhoff)?

*No.*

The PAM probability and scoring matrices were *derived* by inferring necessary accepted mutations from ancestral sequences. This is done once, and not by us. Once the matrices were created — and we should think of them as given to us — we use them for sequence alignment to compare sequences that exist today. We do not re-estimate the APMs.

### Why do the numbers in the BLOSUM matrices trend the opposite direction from the PAM matrices?

1. Because they are different matrices :-)

2. Because the numbers after PAM and BLOSUM have different meanings. The PAM1 scoring matrix is derived from the PAM1 mutation probability matrix, which is a relatively short period of sequence divergence. Larger PAM numbers (say, PAM250) mean successively greater opportunity for divergence. The BLOSUM numbers, on the other hand, refer to the maximum level of identity that is not condensed. Specifically, for BLOSUM80, sequences with 80% identity or greater are clustered together and replaced by a single sequence. (You can treat them as being effectively, but not exactly, ignored.) The consequence is that pairs of sequences with less than 80% identity are weighted more when deriving the BLOSUM80 scoring matrix. BLOSUM62 does the same with all sequences >62% identical. Therefore, smaller numbers in BLOSUM emphasize greater sequence divergence.

BLOSUM also compares sequences in observed alignments, rather than sequences with inferred ancestral sequences as in PAM. The PAM system has an evolutionary basis in it construction, but both systems used in scoring can be interpreted in an evolutionary framework when using them to score alignments.

### Why is this important?

If you never run a sequence analysis in your life, it may not be. But if you run an alignment or a sequence database search, you will have to make choices about which substitution matrix to use. Knowing what they mean will help you make appropriate choices.

Second, it is worth knowing that sequence comparison methods are based on models of sequence divergence through evolution. In the case of PAM matrices, this model is *explicit*. Understanding evolutionary models that use sequence data can not only help you understand the basis of sequence alignment but also stimulate in thinking deeply about evolution, the most important framework for understanding biology.

### I tried multiplying the numbers in PAM1 by 250, and did not get the numbers in PAM250. What's wrong?

Two reasons:

1. Only the mutation probability matrices work this way, and then only when rescaled to [0,1]. If you are using a scoring matrix, which is log-odds form, it won't work.
2. These are *matrix* multiplications, not integer multiplications. So just multiplying the individual numbers won't work.

### Where did you get the numbers for converting PAM matrices to log odds scores?

The top number, 0.55, is from the PAM250 mutation probability matrix. It's the entry at position $(W, W)$. Think of the meaning like this: on average, after 250 PAMs of divergence, a position that was a `W` will still be a `W` 55% of the time. Now, what about the 0.01? That's the chance of seeing a `W` at random. Tryptophan occurred only 1% of the time in the data set used to create the PAM matrices.

### The figure on the Zuckerkandl and Pauling stuff is in percentages, but the numbers add up to more than 100. What gives?

This is, admittedly, a confusing figure. Zuckerkandl and Pauling were looking – this should not surprise you by now – at the globin protein family. (This was 1965, after all, and protein sequences were hard to come by.) The numbers in the figure are not the frequency of substitution, but the percentage of sites that showed **any** such substitution. So, for example, in the top row you see the A (alanine) to D (aspartate) cell is labeled 28. That does *not* mean that alanine was substituted for aspartate 28% of the time. That means that 28% of *sites* containing alanine could be substituted for aspartate at least once. Because of this way of counting the substitutions, the percentages can add up to more than 100%.

Let's look at this number in more detail, and imagine how it could come about. What does that 28% number mean?

There were 18 globin sequences used by Zukerkandl and Pauling. Each globin sequence is roughly 142 amino acids. Suppose 7% of the amino acids in globins, on average, are alanine. That means that $142 \times 0.085 = 12$ positions could be described as "alanine positions". If they were all alanine, all of the time, there would be $12 \times 18 = 216$ alanines in the 18 globin sequences. How if a single aspartate appeared a single time in those 218 opportunities, that would mean 1/12 alanine positions could be substituted, so the cell would already be 8.3%. Three observations of D at three separate positions would result in a value of $3/12 = 25\%$. In this hypothetical example, 25% of the positions containing alanine could be considered able to be substituted to aspartate, even though only 3 substitutions out of 216 opportunities occurred.

### What did you mean by using different matrices for different amounts of divergence?

Let's perform a little experiment. The experiment is to compare the scores of pairwise alignments, one of closely related sequences, and one of distantly related sequences, as indicated in Table 1. In both cases the sequences being compared are homologues, but the percent identity is much higher in the case of the two closely related sequences.

The sequences are shown below

```
>gi|4504345|ref|NP_000508.1| hemoglobin subunit alpha [Homo sapiens]
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNA
VAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSK
YR

>gi|44955888|ref|NP_976312.1| myoglobin [Homo sapiens]
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATVL
```

| Comparison | Relationship | sequence 1 | sequence 2 |
|---|---|---|---|
| 1 | close | human hemoglobin subunit $\alpha$ NP_000508.1 | Macacca hemoglobin subunit $\theta$ NP_001038189.1 |
| 2 | distant | human hemoglobin subunit $\alpha$ NP_000508.1 | human myoglobin NP_976312.1 |

**Table 1.** Two sequence alignment problems

```
TALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFR
KDMASNYKELGFQG

>gi|113461965|ref|NP_001038189.1| hemoglobin subunit theta-1 [Macaca mulatta]
MVLSPADKSNVKAAWGKVGGHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTLA
VGHVDDMPHALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSK
YR
```

For the two comparisons listed above, I used a global alignment method discussed in Lecture 4 (Needleman Wunsch), and tabulated alignment scores for range of PAM matrices. A plot of the table is also shown. You will note that alignments using the lower numbered PAM matrices, which represent a smaller evolutionary distance and thus less opportunity for sequence divergence, fail to detect the homology between distantly related sequences. With a very high numbered PAM matrix the relationship would also be undetectable, but I stopped at PAM 500.

Like the practical exercise you went through, the reason this happens is because distant evolutionary relationships are best captured by some matrices (high numbered in the case of PAM). Low numbered PAM matrices simply fail to detect the relationship between distant homologues, and high numbered PAM matrices underestimate the significance of close homologues.

| PAM MATRIX | Comparison 1 (close) | Comparison 2 (distant) |
|---:|---:|---:|
| 10 | 1066.0 | 16.5 |
| 20 | 1018.0 | 25.0 |
| 30 | 1000.0 | 37.5 |
| 40 | 979.0 | 47.0 |
| 50 | 928.0 | 60.5 |
| 100 | 772.0 | 118.0 |
| 200 | 790.0 | 213.5 |
| 300 | 707.0 | 249.0 |
| 400 | 604.0 | 254.5 |
| 500 | 604.0 | 281.5 |

**Table 2.** Needleman-Wunsch alignments scores for two different scenarios using different PAM matrices

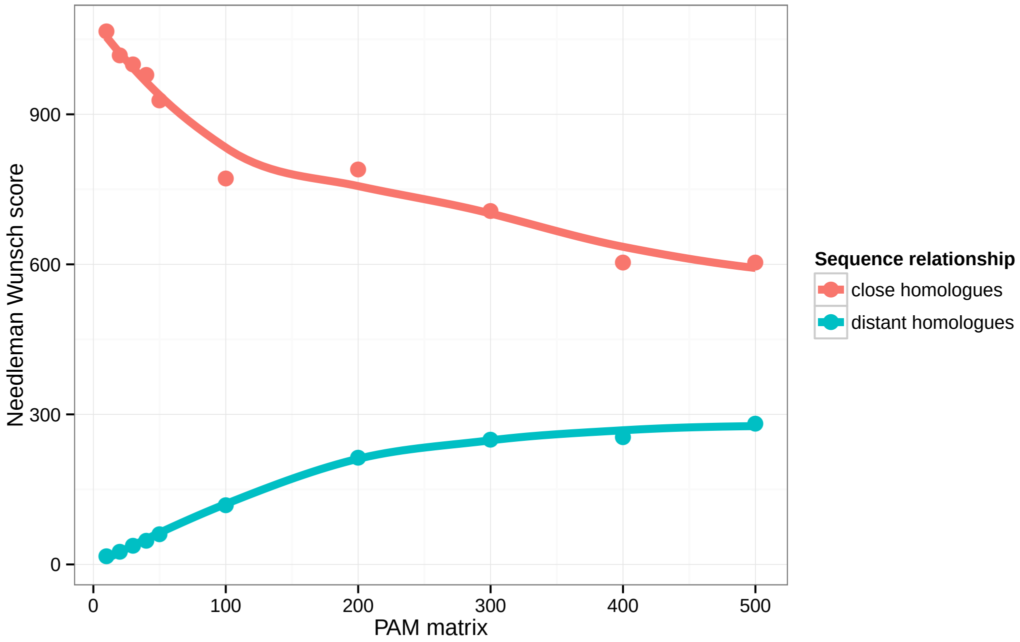

**Figure 3.** Global alignment scores of two sequence alignment problems listed in Table 2, using a range of PAM substitution matrices

# Bibliography

[1] MO O Dayhoff, RM M Schwartz, and BC Orcutt. “A Model of Evolutionary Change in Proteins”. In: *Atlas of Protein Sequence and Structure* 5 (1978), pp. 345–352 (cit. on p. 3).

[2] Erik L L Sonnhammer and Eugene V Koonin. “Orthology, paralogy and proposed classification for paralog subtypes”. In: *Trends in Genetics* 18.12 (Nov. 2002), pp. 619–620. DOI: `10.1038/nbt749` (cit. on p. 2).

*This document was typeset using LaTeX.*