Last modified August 19, 2015

# Elementary Sequence Analysis

**Brian Golding, Dick Morton and Wilfried Haerty**

Department of Biology
McMaster University
Hamilton, Ontario
L8S 4K1

These notes are in Adobe Acrobat format (they are available upon request in other formats) and they can be obtained from the website http://helix.biology.mcmaster.ca/courses.html. Some of the programs that you will be using in this course and which will be run locally can be found at http://evol.mcmaster.ca/p3S03.html.

The "blue text" should designate links within this document while the "red text" designate links outside of this document. Clicking on the latter should activate your web browser and load the appropriate page into your browser. If these do not work please check your Acrobat reader setup. The web links are accurate to the best of our knowledge but the web changes quickly and we cannot guarantee that they are still accurate. The links designated next to the JAVA logo,  , require that JAVA be installed on your computer.

These notes are used in Biology 3S03. The purpose of this course is to introduce students to the basics of bioinformatics and to give them the opportunity to learn to manipulate and analyze DNA/protein sequences. Of necessity only some of the more simple algorithms will be examined.

The course will hopefully cover . . .

- databases of relevance to molecular biology.
- some common network servers/sites that provide access to these databases.
- methods to obtain sequence analysis software and data.
- methods of sequence alignment.
- methods of calculating genetic distance.
- methods of phylogenetic reconstruction.
- methods for detecting patterns and codon usage.
- methods for detecting gene coding regions.

The formal part of the course will consist of two approximately one hour lectures each week. Weekly assignments will be be provided to practice and explore the lecture material. In addition there will be an optional tutorial to help students with these assignments or other problems. These assignments will be 40% of your grade and three, in class quizzes will make up the remainder.

We would appreciate any comments, corrections or updates regarding these notes.

Golding@McMaster.CA Morton@McMaster.CA HaertyW@McMaster.CA

This document was entirely constructed with pdfLaTeX. Enormous kudos to all those that make this great software free to everyone.

# Table of Contents in Brief

In order to speed download, I place here links to the individual chapters in pdf format. The contents of these are shown on the following 'Contents' pages but note that the links will function only for the individual chapter included here.

Preliminaries

Basic Unix

Genomics

Databases

Sequence File Formats

Sequence Alignment

Distance Measures

Database Searching

Reconstructing Phylogenies

Pattern analysis

Exon analysis

# Contents


**1 Preliminaries** **1**

1.1 Resources . . . 1

1.1.1 Electronic Resources . . . 1

1.1.2 Textbooks . . . 2

1.1.3 Journal sources . . . 7

1.2 Biological preliminaries . . . 10

1.2.1 Some notes on terminology . . . 10

1.2.2 Letter Codes for Sequences . . . 11

**2 Computer skills preliminaries** **13**

2.1 UNIX Operating Systems . . . 13

2.1.1 Logging on/off . . . 14

2.1.2 UNIX File System . . . 14

2.1.3 Commands . . . 17

2.1.4 Help . . . 19

2.1.5 Redirection . . . 20

2.1.6 Shells . . . 20

2.1.7 Special 'hidden' files . . . 21

2.1.8 Background Processes . . . 21

2.1.9 Utilities . . . 22

2.1.10 Editors . . . 22

2.2 Exchange among computers . . . 24

2.2.1 ssh . . . 24

2.2.2 Mail . . . 24

2.3 Scripts-Languages . . . 25

2.4 Obtaining LINUX . . . 25

**3 Genomics** **27**

3.1 Where the data comes from . . . 27

3.2 How DNA is sequenced . . . 27

3.3 First Generation Methods . . . 28
3.4 The reality of sequencing includes errors . . . 32
3.5 From sequence to genome . . . 33
3.6 Second (Next) Generation Sequencing . . . 37
3.7 Paired sequences . . . 43
3.8 Third Generation Sequencing . . . 44
3.9 Upcoming Sequencing Technologies . . . 45
3.10 Types of sequencing . . . 46
3.10.1 Exome sequencing . . . 46
3.10.2 RAD-tag seq . . . 47
3.10.3 BAsE-seq . . . 47
3.10.4 RNA-seq . . . 47
3.10.5 BS-seq . . . 48
3.10.5.1 TAB-seq . . . 48
3.10.5.2 NOMe-seq . . . 49
3.10.6 Regulatory sequencing: DNase-seq/FAIRE-seq . . . 49
3.10.7 ChIP-seq . . . 49
3.10.7.1 CLIP-seq . . . 49
3.10.8 Hi-C . . . 50
3.11 Other kinds of biological data . . . 50
3.11.1 Microarrays . . . 51
3.11.2 Mass spectrometry methods . . . 56
3.11.3 Textual information . . . 56

**4 Databases 59**
4.1 Introduction . . . 59
4.2 N.C.B.I. . . . 62
4.3 E.M.B.L. . . . 67
4.4 D.D.B.J. . . . 68
4.5 SwissProt . . . 69
4.6 Organization of the entries . . . 71
4.7 Other Major Databases . . . 73
4.8 Remote Database Entry retrieval . . . 76
4.8.1 Entrez . . . 76
4.8.2 NCBI retrieve . . . 77
4.8.3 EMBL get . . . 79
4.8.4 Others . . . 80
4.9 Reliability . . . 80

**5 Sequence File Formats** **83**
5.1 Genbank/EMBL . . . 83
5.2 FASTA . . . 85
5.3 FASTQ . . . 86
5.4 SAM/BAM format . . . 87
5.5 Stockholm format . . . 88
5.6 GDE . . . 90
5.7 NEXUS . . . 92
5.8 PHYLIP . . . 93
5.9 ASN . . . 94
5.10 BSML format . . . 97
5.11 PDB file format . . . 97

**6 Sequence Alignment** **103**
6.1 Dot Plots . . . 103
6.1.1 The Exact Way . . . 103
6.1.2 Identity Blocks . . . 105
6.2 Alignments . . . 113
6.2.1 The Needleman and Wunsch Algorithm . . . 113
6.2.2 The Smith-Waterman Algorithm . . . 116
6.3 Testing Significance . . . 117
6.4 Gaps and Indels . . . 120
6.4.1 "Natural" Gap Weights - Thorne, Kishino & Felsenstein . . . 120
6.5 Multiple Sequence Alignments . . . 121

**7 Distance Measures** **125**
7.1 Nucleotide Distance Measures . . . 125
7.1.1 Simple counts as a distance measure . . . 125
7.1.2 Jukes - Cantor Correction . . . 126
7.1.3 Kimura 2-parameter Correction . . . 128
7.1.4 Tamura - Nei Correction . . . 128
7.1.5 Uneven spatial distribution of substitutions . . . 129
7.1.6 Synonymous - nonsynonymous substitutions . . . 130
7.2 Amino acid distance measures . . . 130
7.2.1 PAM Matrices . . . 131
7.2.2 BLOSUM Matrices . . . 133
7.2.3 GONNET Matrix . . . 134
7.3 Gap Weighting . . . 135

**8 Database Searching** **137**
8.1 Are there homologues in the database? . . . 137
8.1.1 FASTA . . . 137
8.1.1.1 Instructions . . . 137
8.1.1.2 FASTA output . . . 139
8.1.1.3 FASTA format . . . 142
8.1.1.4 Statistical Significance . . . 144
8.1.2 BLAST . . . 145
8.1.2.1 BLAST output . . . 146
8.1.2.2 BLAST format . . . 150
8.1.3 MPsrch . . . 152
8.1.3.1 MPsrch output . . . 153
8.1.3.2 MPsrch format . . . 155
8.2 BLOCKS . . . 156
8.2.1 BLOCKS output . . . 157
8.2.2 Getting the Block . . . 158
8.3 SSearch . . . 164
8.4 Why you should routinely check your sequence . . . 164

**9 Reconstructing Phylogenies** **165**
9.1 Introduction . . . 165
9.1.1 Purpose . . . 165
9.1.2 Trees of what . . . 165
9.1.3 Terminology . . . 167
9.1.4 Controversy . . . 169
9.2 Distance Methods . . . 169
9.3 Parsimony Methods . . . 171
9.4 Other Methods . . . 174
9.4.1 Compatibility methods . . . 174
9.4.2 Maximum Likelihood methods . . . 174
9.4.3 Method of Invariants . . . 175
9.4.4 Quartet Methods . . . 176
9.5 Consensus Trees . . . 178
9.6 Bootstrap trees . . . 178
9.7 Warnings . . . 181
9.8 Available Packages . . . 182
9.9 PHYLIP . . . 186
9.9.1 PHYLIP Contents . . . 186

**10 Pattern Analysis** **199**
10.1 Base Composition: first order patchiness . . . 199
10.1.1 Genome Patchiness . . . 199
10.2 Dinucleotide Composition: second order patchiness . . . 200
10.3 Strand Asymmetry . . . 201
10.3.1 Chargaff's Rules . . . 201
10.3.2 Replication Asymmetry . . . 202
10.3.3 Transcriptional Asymmetry . . . 203
10.3.4 Codon Selection . . . 204
10.4 Simple Sequence Repeats . . . 204
10.5 Sequence Complexity . . . 204
10.5.1 Information Theory . . . 204
10.5.2 Sequence Window Complexity . . . 206
10.6 Finding Pattern in DNA Sequences . . . 207
10.6.1 Consensus Sequences . . . 207
10.6.2 Matrix Analysis of Sequence Motifs . . . 208
10.6.3 Sequence Conservation and Sequence Logos . . . 209

**11 Exon Analysis** **213**
11.1 Open Reading Frames . . . 213
11.2 Gene Recognition . . . 213
11.2.1 Splice Sites . . . 214
11.2.2 Codon Usage . . . 215
11.2.3 Gene Prediction Software . . . 218
11.2.4 Hidden Markov Models (HMM) . . . 219
11.2.5 Comparison of Programs . . . 219

# Chapter 7

# Distance Measures

## 7.1 Nucleotide Distance Measures

### 7.1.1 Simple counts as a distance measure

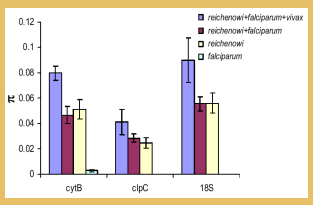
One of the most common measures used in computer algorithms for sequence analysis is some measure of the distance between two sequences. For many methods it is absolutely critical to get an accurate measure of distance. Past studies have shown that most algorithms that make use of a distance are not robust to small deviations in the distance matrices. This problem is also related to weighting differences between sequences.

Why bother about corrections for distances? Consider the analogy of the difference between a Toyota Corolla and a Honda Civic. This is not the same difference as that between a Civic and a Mercedes. They are each different cars but there is a greater qualitative difference between them (minimally, a big difference exists in the price between a Civic and a Mercedes but less so between a Civic and a Corolla).

The same thing applies for sequences. Two sequences that differ by an A and G do not have the same quality of difference as do two sequences that differ by an A and a T. The former substitution is a transition and can happen readily while the latter is a transversion and occurs far less frequently. Hence it would be desirable to weight or to treat these substitutions in a different fashion. There is no reason why we should have used 1 for a residue match and 0 for a mismatch in the section on alignments. You can use any value for these that you wish (and indeed 1 and 0 are particularly poor choices).

An example of just one interesting study (from among thousands) that use genetic distance measures is shown in the Box. Not only is this a typical (and, in my opinion, fascinating) application, but distance measures are a basic groundwork for much that follows.

The malaria parasite jumped hosts!




This figure is from Rich et al. 2009 and shows $\pi$, a measure of genetic distances due to polymorphisms within a species. The bars show $\pi$ for three genes from (right to left) (i) *Plasmodium faliciparum* alone (ii) *P. reichenowi* alone, (iii) *P. reichenowi* + *P. falciparum*, (iv) *P. reichenowi, P. falciparum, + P. vivax*.

Note that the genetic distances within *P. reichenowi* a parasite that infects chimpanzees is not increased by the addition of sequences from *P. falciparum*, the species that infects humans and cause malaria.

The authors reasoned that if the two malarial parasites cospeciated with humans and chimps they should be 5-7 million years old (this estimate is yet another application of distance measures). They went on to measure genetic distances for each gene and show that *P. falciparum* and *P. reichenowi* are too similar and suggest they originated from 10,000 to 1 million years ago.

For each gene, *cytB, clpC*, and *18S rRNA*, they estimated the best distance model as F81 + Γ, GTR + Γ, and HKY + I + Γ, respectively. It is these types of models we discuss in this section.

To construct a good distance measure requires more than solving a problem of simple weighting, there are also subtler problems. Lets assume for the moment that all mutations occur with equal frequency. Then you might think that the difference between two sequences could be calculated simply

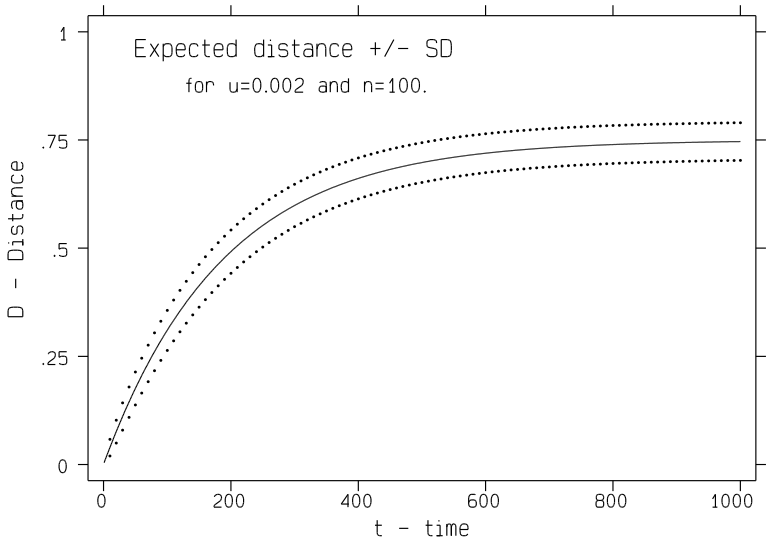



Figure 7.1: An asymptotic divergence with time.

by counting the number of nucleotide differences between the species. Lets consider how this difference, this measure of distance changes over time. Figure 7.1 shows the difference expected between two sequences that have diverged at increasing times into the past. The proportion of differences are calculated simply by counting the number of nucleotide differences divided by the total length of the sequence. Hence,

$$D = k/n,$$

$$Var(D) = D(1-D)/n,$$

where $n$ is the length of the sequence and $k$ is the number of nucleotides that differ. In Figure 7.1, $\mu$ is rate of substitutions for the sequences and $t$ is the length of time since the last common ancestor of these sequences. The rate of change is initially going up with a slope equal to twice what one might expect from the product of the mutation rate and time because both sequences are diverging from a common ancestor. Figure 7.1 shows that as the time of divergence increases the percent difference or the distance increases. Initially this occurs linearly however as time proceeds the measure of distance begins to slow its increase and finally reaches an asymptote of 0.75 and ceases to increase at all.

This is quite reasonable when you think about it. There are only four types of nucleotides. A random collection based on these four possibilities will have one quarter of them identical by chance alone. But this has lots of implications for the distances that are calculated between species. A pair at time $t = 20$ are expected to have $D = 7.6$ and a pair at $t = 40$ are expected to have $D = 14.4$. These can be easily distinguished. But a pair at $t = 500$ and $t = 1000$ have $D$'s of 69.8 and 74.6. These will be hard to tell apart. And yet, in both cases there is a doubling of the divergence between species pairs.

### 7.1.2 Jukes - Cantor Correction

As the time of divergence between two sequences increases the probability of a second substitution at any one nucleotide site increases and the increase in the count of differences is slowed. This makes these counts an undesirable measure of distance. In some way, this slow down must be accounted for. The solution to this problem was first noted by Jukes and Cantor (1969; Evolution of Protein Molecules, Academic Press). Instead of calculating distance as a simple count take the distance as

$$D_{JC} = -(\frac{3}{4}) \ln \left(1 - (\frac{4}{3})D\right),$$

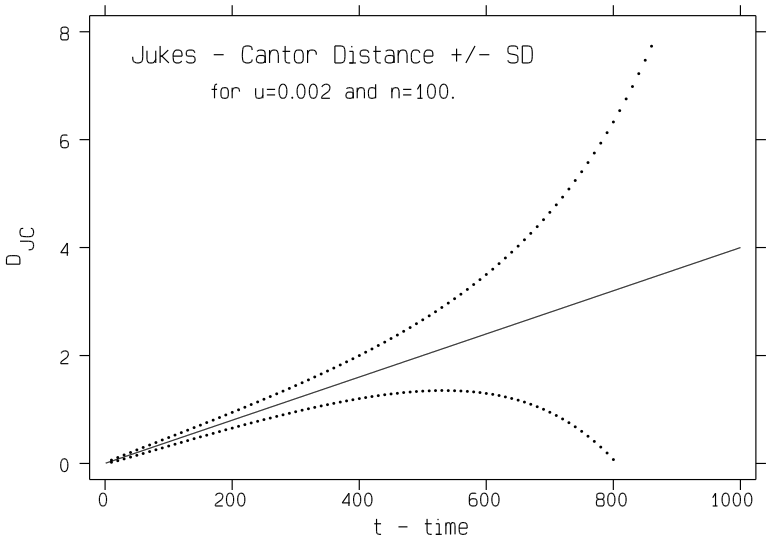



Figure 7.2: A correction leads to linear divergence with time.

$$Var(D_{JC}) = \frac{D(1-D)}{\left[n(1-(\frac{4}{3})D)^2\right]}$$

(Kimura and Ohta 1972; J. Mol. Evol. 2:87-90).

A plot of this function for the same range of parameters as in Figure 7.1 is given in Figure 7.2. This figure shows that this distance measure increases linearly with time (this is one property that is desirable for a distance measure). This is termed the Jukes & Cantor correction to distance and clearly indicates that divergence is a logarithmic function of time.

Observe the large increase in the variance as time increases. As $D$ gets closer and closer over time to 0.75 the variance increases. In the limit as $D$ approaches 0.75, the variance approaches infinity. This is an indication that the measure of distance becomes increasingly less reliable as time increases.

Note that in expectation $D$ is less than 0.75 but in reality a single instance of $D$ can be greater than 0.75. If this is the case then a Jukes-Cantor correction cannot be done and $D_{JC}$ is undefined because the argument of the logarithm will be negative. In this case you can apply a method developed by Tajima (1993, Mol. Biol. Evol. 10:677-688). He suggests using the modified estimator

$$D^*_{JC} = \sum_{i=1}^{k} \frac{k^{(i)}}{i(\frac{3}{4})^{i-1} n^{(i)}}$$

where

$$k^{(i)} = k!/(k-i)! \qquad \text{and} \qquad n^{(i)} = n!/(n-i)!$$

With variance

$$Var(D^*_{JC}) = D^*_{JC}(1-D^*_{JC})exp(8D^*_{JC}/3)/(n-1).$$

Here $k$ is the count of differences between the two sequences and $n$ is the length of these sequences. This is actually just a different formulation of the same quantity using a Taylor series expansion to avoid the logarithm. This estimator of distance is defined for all parameter values and actually has less bias than Jukes and Cantor's original correction for small levels of divergence. Tajima provides similar adjustments to all of the corrections noted below.

### 7.1.3 Kimura 2-parameter Correction

Note that this still does not correct for differences in the rates of transition and transversion. To do this you can use what is called the Kimura 2-parameter correction. This was a method established by Kimura (1980; J. Mol. Evol. 16:111-120) where the rates of transitions are assumed to be $\alpha$ and the rates of transversions are $\beta$. Then if the observed percentage of transitional differences are $P$ and the observed percentage of transversion differences are $Q$, the estimate of distance is

$$D_{K2p} = -(\frac{1}{2})\ln(1-2P-Q) - (\frac{1}{4})\ln(1-2Q)$$

and

$$Var(D_{K2p}) = [c_1^2P + c_3^2Q - (c_1P + c_3Q)^2]/n,$$

where $c_1 = 1/(1-2P-Q)$, $c_2 = 1/(1-2Q)$ and $c_3 = \frac{1}{2}(c_1+c_2)$. Again divergence follows a logarithmic function.

In this case you can also determine the rates of substitution via transitions and transversions separately. The rate of transition substitutions per site is

$$s = -(\frac{1}{2})\ln(1-2P-Q) + (\frac{1}{4})\ln(1-2Q)$$

$$Var(s) = [c_1^2P + c_4^2Q - (c_1P + c_4Q)^2]/n$$

where $c_4 = \frac{1}{2}(c_1 - c2)$. The rate of transversion substitutions per site is

$$v = -(\frac{1}{2})\ln(1-2Q)$$

$$Var(v) = c_2^2Q(1-Q)/n.$$

### 7.1.4 Tamura - Nei Correction

Hasegawa, Kishino and Yano (1985, J. Mol. Evol. 22:160-174) suggested a model that Tamura and Nei (1993, Mol. Biol. Evol. 10:512-526) have extended. They suggest a model with different rates of transversions $\beta$, and transitions as $\alpha_1$ and $\alpha_2$ between purines and between pyrimidines respectively. They also consider mutation rates that yield the observed frequency of A, T, C and G ($g_A, g_T, g_C, g_G$). In this case, it can be shown that the distance is

$$\begin{aligned} D_{TN} = & - (2g_Ag_G/g_R)\ln[1-(g_R/2g_Ag_G)P_1 - (1/2g_R)Q] \\ & - (2g_Tg_C/g_Y)\ln[1-(g_Y/2g_Tg_C)P_2 - (1/2g_Y)Q] \\ & - 2(g_Rg_Y - (g_Ag_Gg_Y/g_R) - (g_Tg_Cg_R/g_Y))\ln[1-(1/2g_Rg_Y)Q], \end{aligned}$$

where $P_1, P_2, Q$ are the proportions of transitions between A and G, between T and C, and the proportions of transversions. The variance has also be derived but is very complicated.

Other more complicated corrections are possible. For example, Felsenstein and Hasegawa have developed likelihood methods that find a maximum likelihood estimate of the distance between two sequences with mutation rates estimated from the actual sequences. It has also been demonstrated that such maximum likelihood estimates of distances are much more accurate than log-transform estimates Hoyle and Higgs (2003, Mol. Biol. Evol. 20:1-9)

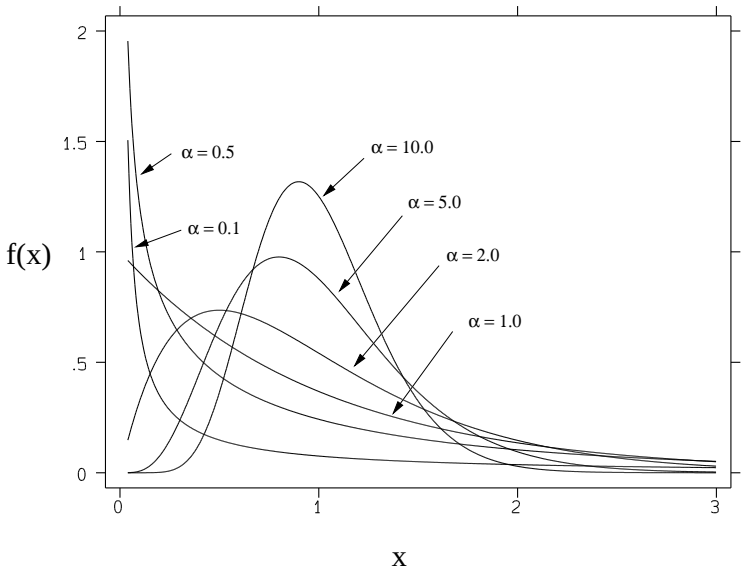



Figure 7.3: Shapes that gamma distributions can take.

### 7.1.5 Uneven spatial distribution of substitutions

A different sort of problem arises if substitutions are not equally spread throughout the sequence. In this case there are some spots that are "hot" - have had many substitutions and other spots that are "cold" - have had few substitutions. Hence some parts of the sequence may require strong correction for multiple substitutions and an excess of transitions/transversions while other parts of the sequence may require only minor correction.

It would be ideal if all spots along a sequence could have their own rate constants. But if this is permitted then there are so many parameters possible that any sort of data or observation could be simply explained by changing to the appropriate set of parameters.

The most common method to deal with this problem is to apply a gamma distribution to the distribution of substitutions along the sequence. The gamma distribution has been chosen because it is mathematically well characterized, it is a simple distribution, it is a continuous distribution, it is non-negative, and it can assume a variety of shapes. The gamma distribution has density

$$\begin{aligned} f(x) &= [\beta^{\alpha}/\Gamma(\alpha)]x^{(\alpha-1)}e^{-\beta x}, \quad x > 0 \\ &= 0, \quad elsewhere \end{aligned}$$

and where

$$\Gamma(\alpha) = \int_0^{\infty} x^{(\alpha-1)}e^{-x}dx.$$

A plot of gamma distributions is given in Figure 7.3. In distance measures, it is generally used with $\alpha$ set to $\mu^2/Var(\mu)$ and $\beta$ set equal to $\alpha$, where $\mu$ is the overall rate of substitution. This provides an interpretation such that $\alpha$ is the inverse of the coefficient of variation of substitutions among sites squared. Therefore the smaller the parameter $\alpha$ the higher the extent of variation in substitution rate. The distribution is completely determined by this mean rate of substitution and its coefficient of variation. Therefore only one extra parameter is being used to determine a variety of distributional shapes. Since the mean of the gamma distribution is $\alpha/\beta$, the mean will always be one in this case. Thus for each of these variety

of distributions the relative rate per site is constant. But unless $\alpha$ is very large, some sites in the sequence will have rates well above the mean and some well below the mean.

All of the distance measures discussed in previous sections can be corrected to include a gamma distribution for the distance measure. For example, the Jukes - Cantor correction becomes

$$D_{\widetilde{JC}} = (\frac{3}{4})\alpha \left[ (1 - (\frac{4}{3})D)^{-1/\alpha} - 1 \right],$$

$$Var(D_{\widetilde{JC}}) = D(1-D) \left[ (1 - (\frac{4}{3})D)^{-2(1/\alpha+1)} \right] /n.$$

In general it would be desirable to estimate the value of the gamma parameter $\alpha$ and this can be done easily (given the above interpretation) and it is done by some algorithms that you might run across. However, it is also very common for algorithms that include this correction to simply request a value for $\alpha$ from the user. Studies of many amino acids sequences have suggested that often $\alpha < 2$ and one program package uses a default value of $\alpha = 1$. A typical example of extreme variation would be the $\alpha = 0.47$ that has been noted for some immunoglobulin genes (here much of the variation is probably due to differential selection). Values typical for your own applications will have to be calculated if you are using a program that requests supplied values.

The codes used!

As you read papers in the scientific literature about genetic distances you will read code that states they used the "HKY + Γ + I" model. As explained in this chapter there are several models that can be used to estimate distances. These range from simple to very complex. Since in general it is not good to over-parameterize a model, a simpler model is preferred if it adequately fits the data. To test this fit a series of hierarchical tests can be applied and have been implemented by Posada & Crandall 1998.

The models that they test include

| | |
|---|---|
| JC | Jukes and Cantor (1969) |
| K80 | Kimura (1980) (=K2P) |
| HKY | Hasegawa, Kishino, Yano (1985) |
| TN | Tamura and Nei (1993) |
| TNef | Tamura-Nei equal frequencies |
| K81 | Two transversion-parameters model 1 (=K81=K3P) (Kimura, 1981) |
| K81uf | K81 with unequal frequencies |
| TIMef | Transitional model equal frequencies |
| TIM | Transitional model |
| TVMef | Transversional model equal frequencies |
| TVM | Transversional model |
| SYM | Symmetrical model (Zharkikh, 1994) |
| GTR | General time reversible (=REV) (Tavare, 1986) |

In addition to these, the rates can be gamma distributed, Γ, and the model might include some sites that are considered invariant.

Hence you arrive at codes such as "HKY + Γ + I; an HKY model with gamma distributed rates and invariant sites".

### 7.1.6 Synonymous - nonsynonymous substitutions

Substitutions that result in amino acid replacements are said to be nonsynonymous while substitutions that do not cause an amino acid replacement (such as a GGG codon to GGC codon change - both codons still encode glycine) are said to be synonymous substitutions. Because of the difference in their effects on the physiology of the organism, synonymous and nonsynonymous substitutions can have quite different dynamics. For example, synonymous substitutions usually occur at a much faster rate than do nonsynonymous substitutions. Hence, for coding sequence it is often desirable to separate these two.

The most common method to estimate these parameters separately is via an algorithm set out by Li, Wu & Luo (1985; Mol. Biol. Evol. 2:150-174). It is somewhat complicated and I refer you to their paper for a complete description. Basically it counts the number of sites that are potentially 4-way, 2-way or 0-way degenerate (the third position of a glycine codon being 4-way degenerate, any second codon position being 0-way degenerate). It then counts the number of differences at each site of each category keeping tract of transversions and transitions. It then calculates

$$K_S \quad \text{and} \quad K_A$$

the rate of synonymous and nonsynonymous substitutions. It has been found that $K_A$ can have large variation and great changes between/within specific organisms. On the other hand, $K_S$ is generally less variable (though still shows more variation than would otherwise be predicted) and shows less changes between/within organisms.

## 7.2 Amino acid distance measures

Distance is powerful in the sense that it can be used with anything that can be measured. For example, the distance could be based on the strength of an immunological reaction. Using this method any form or measure of distance can be used

and different types of measures can be combined into one. Hence, distances can be used with restriction site data, with allozyme data, with data on quantitative characters, with DNA fingerprints or even with real finger fingerprints. Methods to correct this type of data are not well developed because these are not as well defined characteristics.

Even with amino acids, the corrections can not be done easily and/or without some large bias. A Jukes-Cantor correction is possible. It is simply

$$D_{JC} = -(\frac{19}{20}) \ln(1 - (\frac{20}{19})D),$$

or more commonly just

$$D_{JC} = -\ln(1 - D).$$

But this assumes (as does the nucleotide Jukes-Cantor correction) that for all characters the rate of substitution from one amino acid and to some other amino acid are equal and independent of the residue. This is not true of DNA and is even less true for proteins. Amino acids like cysteine and proline are very important for the structure and function of proteins. Amino acids such as tryptophan have bulky side groups and can not be inserted easily into any site in a peptide. Because of this most amino acid distances use empirical weighting schemes. The most popular of these empirical measures is the PAM family of matrices.

### 7.2.1 PAM Matrices

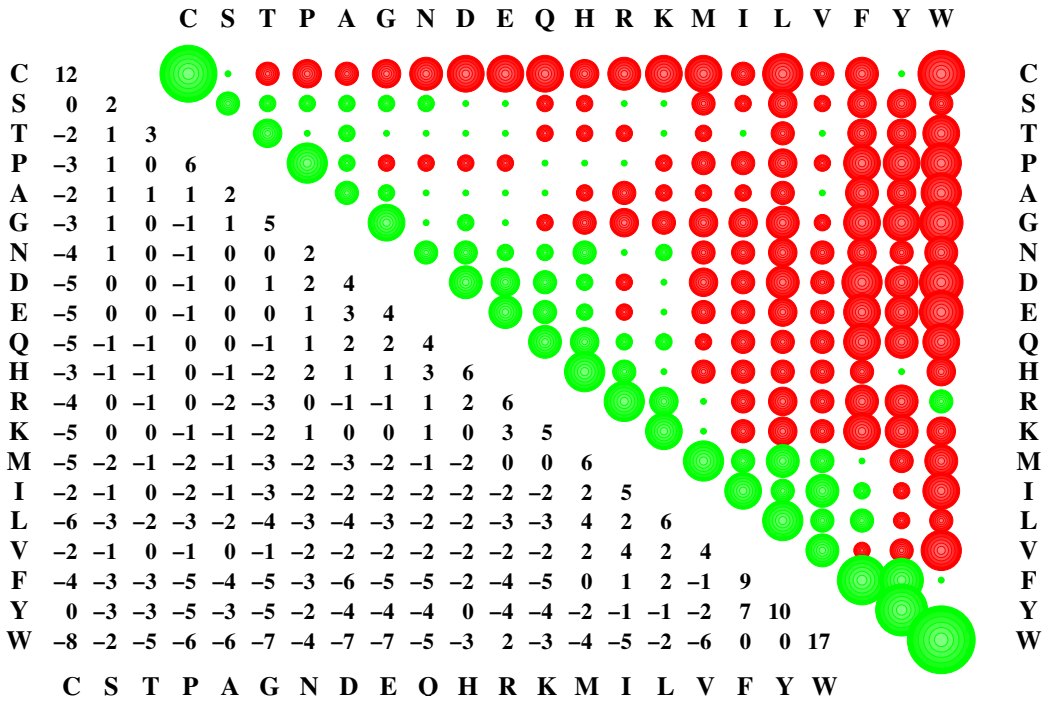
There are several common ways in which weights can be applied for amino acid differences. Karlin and Ghandour (1985, PNAS 82:8597-8601) proposed a method of weights based on chemical, functional, charge and structural properties of the amino acids. Similarly Doolittle proposed weights based on the structural similarities and the ease of genetic interchange (Feng, Johnson and Doolittle 1985 J. Mol. Evol. 21: 112-125). However, by far the most common and most famous way to assign weights is to use Dayhoff's PAM250 matrix. This is a matrix of weights that is derived from how often different amino acids replace other amino acids in evolution (see M.O. Dayhoff, ed., 1978, Atlas of Protein Sequence and Structure, Vol. 5). This was based on a data base of 1,572 changes in 71 groups of closely related proteins appearing in earlier volumes of this amazing predecessor to electronic databases. PAM stands for percent accepted mutations and these were inferred from the types of changes observed in these proteins. Every change was tabulated and entered in a matrix enumerating all possible amino acid changes.

In addition to these counts of accepted point mutations an idea of the relative mutability of different amino acids were calculated. The information about the individual kinds of mutations and about the relative mutability of the amino acids can be combined into one distance-dependent "mutation probability matrix". The elements of this matrix give the probability that the amino acid in one column will be replaced by the amino acid in some row after a given evolutionary interval. For example, a matrix with an evolutionary distance of 0 PAMs would have ones on the main diagonal and zeros elsewhere. A matrix with an evolutionary distance of 1 PAM would have numbers close to one on the main diagonal and small numbers off the main diagonal. One PAM would correspond to roughly 1% divergence in a protein (one amino acid replacement per hundred). The model of evolution that Dayhoff used assumed that proteins diverged as a result of accumulated, uncorrelated mutations. They treat the $PAM-1$ matrix as a first order Markov chain transition model. To derive a mutational probability matrix for a protein sequence that has undergone N percent accepted mutations, a $PAM-N$ matrix, the $PAM-1$ matrix is multiplied by itself $N$ times. This results in a family of scoring matrices.

By trial and error Dayhoff *et al.* found that for weighting purposes a PAM 250 matrix works well for distant relationships. At this evolutionary distance (250 substitutions per hundred residues) only one amino acid in five remains unchanged and the percent divergence has increased to roughly 80%. However, the amino acids vary greatly in their mutability. According to Dayhoff *et al.* roughly 55% of the tryptophans, 52% of the cysteines and 27% of the glycines would still be unchanged, but only 6% of the highly mutable asparagines would remain. Several other amino acids particularly alanine, aspartic acid, glutamic acid, glycine, lysine and serine are more likely to occur in place of an original asparagine than asparagine itself at this evolutionary distance.

From this matrix an odds matrix is constructed. This matrix takes the elements of the previous matrix ($M_{ij}$) and divides

Table 7.1: The log odds matrix for PAM250 (multiplied by 10). The numbers in the lower left give the log odds. For the diagram in the upper right, green/red circles are proportional to the odds of an interchange more/less likely than chance alone.



each term by the frequency of the replacement residue. Hence, each term now gives the probability of replacement, $j$ to $i$ per occurrence of residue $j$.

By tradition the $\log_{10}$ of this matrix is used as weights (this is because to calculate the odds for the whole matrix requires taking the product of changes for all sites of the protein. Before calculators it was easier to find the sum of the log's rather than the product sum). This log odds PAM 250 matrix is shown in Table 7.1 (also note that amino acids have been sorted according to their similarity in this matrix).

Residue pairs with scores above 0 replace each other more often as alternatives in related sequences than in random sequences. This can be an indication that both residues can carry out similar functions. A score exactly equal to zero indicates amino acid pairs that are found as alternatives at exactly the frequency predicted by chance. Residue pairs with scores less than 0 replace each other less often than in random sequences and might be an indication that these residues are not functionally equivalent.

Some of the properties that are visible from this matrix and go into its makeup are - size, shape, local concentrations of electric charge, conformation of van der Waals surface, ability to form salt bonds, hydrophobic bonds, and hydrogen bonds. Interestingly, these patterns are imposed principally by natural selection and only secondarily by the constraints of the genetic code. This tends to indicate that coming up with your own matrix of weights based on some logical features may not be very successful because your logical features may have been over-written by other more important biological considerations.

Some of the problems with this measure of distance are that it assumes that all sites are equally mutable. But this is clearly false. Another problem is that by examining proteins with few differences, the highly mutable amino acids have been stressed. Lastly, due to the collection of proteins known at that time, the matrix is biased because it is based mainly on small globular proteins.

### 7.2.2 BLOSUM Matrices

The BLOSUM matrices originate with a paper by Henikoff and Henikoff (1992; PNAS 89:10915-10919). Their idea was to get a better measure of differences between two proteins specifically for more distantly related proteins. While this bias limits the usefulness of BLOSUM matrices for some purposes, for other programs such as FASTA, BLAST, etc. it should do substantially better. This is because the need for an accurate measure of distance is not as great when peptides are more closely related.

They use the BLOCKS database to search for differences among sequences but only among the very conserved regions of a protein family. Hence the term BLOSUM is from BLOcks SUbstitution Matrix. They first collect all of the sequences in the BLOCKS database and then for each one they sum the number of amino acids in each site to get a frequency table ($q_{ij}, \quad i,j = 1..20$) of how often different pairs of amino acids are found together in these conserved regions. Hence the observed frequency of occurrence of one amino acid is

$$p_i = q_{ii} + \sum_{i \neq j} q_{ij}/2$$

Given pairs should occur with expected frequencies

$$e_{ij} = p_i^2, \quad \text{if } i = j$$

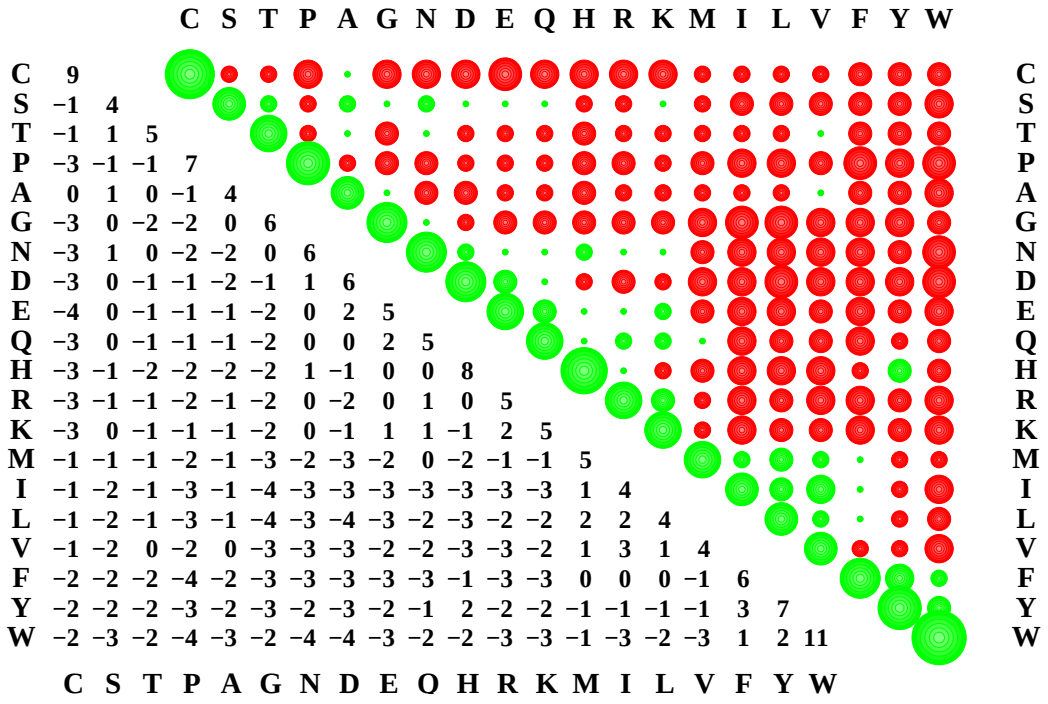
and

$$e_{ij} = 2p_i p_j, \quad \text{if } i \neq j.$$

The odds matrix is $q_{ij}/e_{ij}$. Generally log's are taken of this matrix to give a $\log(odds)$ or *lod* matrix such that

$$s_{ij} = 2\log_2(q_{ij}/e_{ij}).$$

Table 7.2: The log odds matrix for BLOSUM62. The numbers in the lower left give the logs odds, while in the diagram to the upper right, green/red circles are proportional to the odds of an interchange more/less likely than chance alone.




Hence if the observed number of differences between a pair of amino acids is equal to the expected number then $s_{ij} = 0$. If the observed is less than expected then $s_{ij} < 0$ and if the observed is greater than expected $s_{ij} > 0$.

All of this gives the BLOSUM matrix. Different levels of the BLOSUM matrix can be created by differentially weighting the degree of similarity between sequences. Sequences that belong to the same family (within a block) up to a critical level of similarity are clustered so that they are treated as a single entry. For example, a BLOSUM62 matrix is calculated from protein blocks such that if two sequences are more than 62% identical, then the contribution of these sequences is weighted to sum to one. In this way the contributions of multiple entries of closely related sequences is reduced.

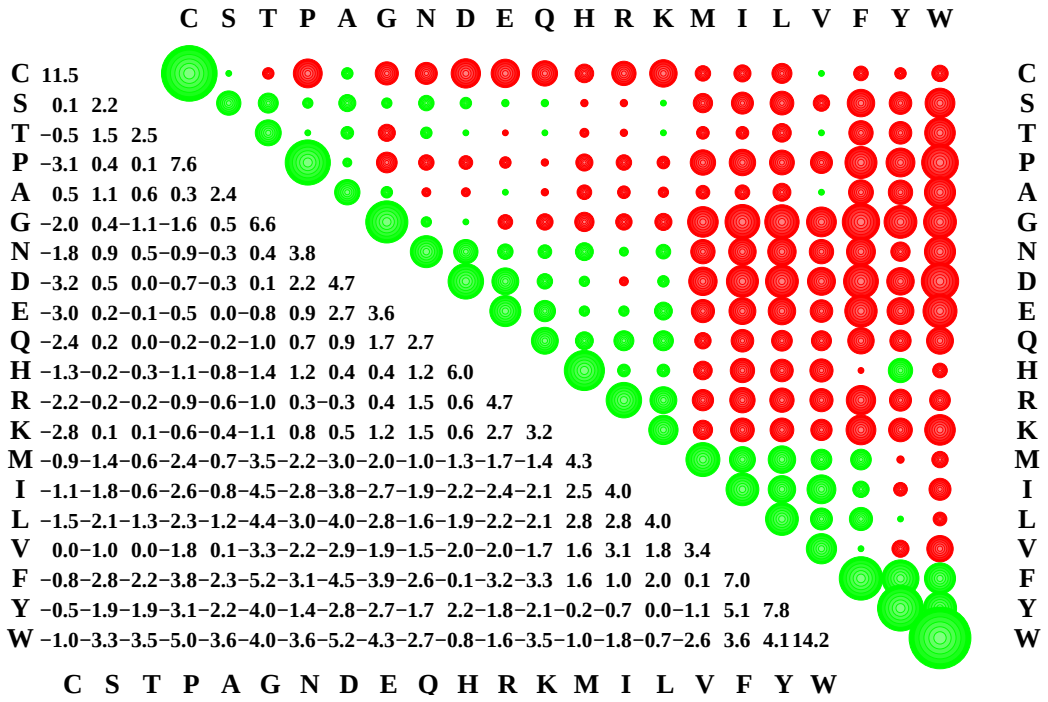
The BLOSUM62 matrix is given in Table 7.2. If the BLOSUM62 matrix is compared to PAM160 (it's closest equivalent) then it is found that the BLOSUM matrix is less tolerant of substitutions to or from hydrophilic amino acids, while more tolerant of hydrophobic changes and of cysteine and tryptophan mismatches.

One of the significant disadvantages of the BLOSUM matrices is that they are not Markov chain matrices. Therefore Veerassamy et al. 2003, J. Comput. Biol 10:997-1010 developed a probability transition matrix, based on the BLOSUM matrices, that can be used in a Markov chain model. This is implimented as the PBM model in the PHYLIP package of programs (see below).

### 7.2.3 GONNET Matrix

A different method to measure differences among amino acids was developed by Gonnet, Cohen and Benner (1992; Science 256:1443-1445) using exhaustive pairwise alignments of the protein databases as they existed at that time. They used classical distance measures to estimate an alignment of the proteins. They then used this data to estimate a new distance matrix. This was used to refine the alignment, estimate a new distance matrix and so on iteratively. They noted that the distance matrices (all first normalized to 250 PAMs) differed depending on whether they were derived from distantly or closely homologous proteins. They suggest that for initial comparisons their resulting matrix should be used in preference to a PAM250 matrix, and that subsequent refinements should be done using a PAM matrix appropriate to the distance between proteins.

Table 7.3: The log odds GONNET matrix. The numbers in the lower left give the log odds, while in the diagram to the upper right, green/red circles are proportional to the odds of an interchange more/less likely than chance alone.




Their matrix is given in Table 7.3 and has been normalized to a PAM distance of 250. The matrix elements are ten times the logarithm of the probability that the amino acids are aligned, divided by the probability that these amino acids would be aligned by chance.

In addition they used these alignments to make an estimate of appropriate gap penalties. From this empirical data they suggest that $P$, the probability of a gap of length $k$ should follow a relation such that

$$10\ln(P) = -36.31 + 7.44\ln(\text{PAM distance}) - 14.93\ln(k).$$

This relation would give the most accurate answer but if the PAM distance is not available, they suggest

$$10\ln(P) = -20.63 - 1.65(k-1).$$

## 7.3 Gap Weighting

Gap penalties are a field where a great deal more work is required. They are often applied without much justification. Dayhoff suggested using a gap penalty of 6 with PAM250 matrices. Henikoff & Henikoff suggest using a gap penalty of 8 with BLOSUM62 matrices. As noted in the previous section Gonnet, Cohen and Benner suggested yet another gap penalty. There is little reason, other than empirical support, for their choices. The MEGA program package and the PHYLIP program package go to the extreme of ignoring all gaps and any missing data. They do this because there is no accurate way to weight changes due to indels relative to substitutions. Never-the-less, the indels do contain some information but the current challenge is to correctly extract it in a precise manner. In this respect, the approach being taken by Thorne (mentioned in the section on alignments) holds great promise. Barring such a sophisticated approach, it is suggested that you use a variety of gap penalties (from some slight to some significant punishment) and from these determine the effects that this has on your results.

As an example of these problems, consider the following three sequences

```
--GCAAACT
--GCAAGCC
ATGCTAGCC
```

Which pair of sequences has the smallest distance? If gaps are ignored then the second and third sequences are closest with one difference. But if gaps are considered (and if each gapped position is counted as one) then the first and second sequences are closest. If gaps are weighted differently then the answer might depend on the particular weighting.