### 15.1 Summary of above

- We model genetic sequences: think of them as strings of letters.
- There are 3 types of sequence, DNA, RNA or Protein.
- DNA sequences are composed of the 4 letters, or bases, $\{A, C, G, T\}$, RNA is made of the bases $\{A, C, G, U\}$ while protein sequences are made up of the 20 amino acids.
- The three types of sequence are related by the central dogma of molecular biology: DNA is transcribed to RNA and then translated to protein.
- Protein sequences fold up into more complex structures. We will ignore this structure in this introductory course.
- DNA is copied from parent to child.
- At copying, mutations are introduced.
- Mutations may be single nucleotide polymorphisms (SNPs), insertions, deletions or of other types.
- We use a tree to model the history of relationships between individuals (which are represented by their sequences).

To model the complex random process of genetic mutation and inheritance, we will need tools from applied probability and statistics. The next few sections are concerned with introducing the main tools and concepts that we will use for our study. All of you will have previously encountered at least some of the ideas we discuss here but, as with the linear algebra sections, it helps to review the main points before plunging in to new material.

## 16 Alignment

### 16.1 Homology

Homology (from the Greek, to agree) is a crucial concept in biology referring to traits or, in the case of sequences, sequence regions that share a common ancestry. We expect homologous regions to be similar to each other where the level of similarity will depend on how recently they shared a common ancestor.

Thus to say two regions are homologous is an evolutionary hypothesis. Mrs Darwin (in Carol Ann Duffy's poem from the collection The World's Wife) was making an evolutionary hypothesis of homology:

7 April 1852
Went to the Zoo.

I said to Him —
Something about that Chimpanzee over there reminds me of you.

The claim does not imply that the regions share a similar function now or, depending on the time since divergence, that they even particularly similar, just that they share a common ancestor. Therefore, sequences are either homologous or not, there are no degrees of homology. We often infer homology between two sequences when they are similar but we must be careful as we can get similarity without homology. Homologous sequences are sometimes referred to as *homologs*.

There are two main ways that we get similarity without homology: either by chance or by convergent evolution. Similarity by chance will occur even in completely random sequences on a finite alphabet. In two random sequences of four letters, we would expect similarity by chance of 25%.

Convergent evolution occurs when similar functions evolve independently of each other. An example of this are the wings of birds and insects. We don't believe these two very different creatures had a common ancestor that evolved wings but that wings evolved indecently of each other in the insect and bird lineages. Thus, while wings in a fly and a sparrow may be superficially similar, they are not considered homologous. The same applies to sequences that code for similar proteins (i.e., have similar function) but have evolved independently. Such traits/regions are called analogous.

We distinguish between two types of homology: orthology and paralogy:

**Orthology** occurs when two genes are separated by a speciation event and evolve independently from there on.

**Paralogy** occurs when a region of the genome is duplicated in the same genome (a duplication event) and they evolve in parallel in the same genome. The two copies are said to be paralogs.

## 16.2 Pairwise alignment

Given two sequences, if they are homologues, how do they align with each other? That is, exactly which sites in the sequence are homologous with each other?

We consider pairs of sequences, $x$ and $y$ of length $m$ and $n$, respectively. $x_i$ is the $i$th symbol of $x$. These symbols are usually the 4 DNA or RNA bases or the 20 amino acids. We refer to the symbols as *residues*.

We will allow *gaps* to be introduced in either sequence to allow them to align better. Biologically, gaps correspond to insertions or deletions in the sequence.

Clearly, there are many ways of aligning a pair of sequences (how many?), but what is the best alignment?

Example: x = GAATTC and y = GATTA

```
GAATTC or GAATTC-
GA-TTA    -GATT-A
```

are two possible alignments. □

## 16.3 Scoring alignments

The best alignment will depend on how we score alignments. It is easy to come up with different scoring regimes (e.g., score 1 for a match, -1 for a mismatch) but we really want to compare two models — that the similarity we see is just chance vs. that the sequences are homologs.

We initially consider alignments without gaps.

### 16.3.1 Model of non-homologous sequences

The most basic model is that each letter appears with some probability, letter $a$ appears with probability $q_a$ (note that the probabilities summed over the alphabet are 1), that each site is independent and that the each sequence is independent. Then the probability of seeing sequence $x$ is

$$P(x) = \prod_{i=1}^{n} q_{x_i}$$

and the joint likelihood of an alignment is just the joint probability of the sequences $x$ and $y$,

$$P(x,y) = P(x)P(y) = \prod_{i=1}^{n} q_{x_i} \prod_{i=1}^{m} q_{y_i} = \prod_{i=1}^{n} q_{x_i} q_{y_i}.$$

### 16.3.2 Model of homologous sequences

An alternative model is that the two sequences are related and the probability of seeing the pair of residues $a$ (from $x$) and $b$ (from $y$) aligned at a locus is $p_{ab}$. The probability of the alignment is then the product of the loci,

$$P(x,y) = \prod_{i=1}^{n} p_{x_i y_i}.$$

To compare these two models, we take ratio of these likelihoods:

$$\frac{\prod_{i=1}^{n} p_{x_i y_i}}{\prod_{i=1}^{n} q_{x_i} q_{y_i}} = \prod_{i=1}^{n} \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}.$$

It is easier to work with log-likelihoods (and addition) than likelihoods and multiplication, so we take the log of this quantity to get

$$S = \sum_i s(x,y_i)$$

where

$$s(a,b) = \log\left(\frac{p_{ab}}{q_a q_b}\right).$$

The score of the alignment then is the sum over the local score $s(a,b)$. $s$ can be thought of as a matrix and is often referred to as a *score matrix* or *substitution matrix*.

There are various methods for deciding reasonable values for the entries of the matrix, discussed below. Note that even when ad hoc values are chosen for the matrix, the underlying probabilities, $p_{ab}$ and $q_a$ can be derived by reversing the above argument. That is, when a score matrix is selected we are making implicit assumptions about the $q_a$s and $p_{ab}$s.

Example: If x = GAATTC and y = GGATTA are aligned as

```
GAATTC
GGATTA
```

where $s(a,b) = 2$ if $a = b$ and $s(a,b) = -1$ if $a \neq b$, the alignments scores $2 - 1 + 2 + 2 + 2 - 1 = 6$. □

## 16.4 Choosing the substitution matrix

For protein sequences, the quantities $p_{ab}$ and $q_a$ have been empirically estimated to produce score matrices. In particular, the BLOSUM (BLocks SUbstitution Matrix) matrix of which there are various types, e.g. Blosum 45 and Blosum 80.

These matrices were calculated by studying a large number of confirmed alignments where there was considerable agreement between the sequences. The relative frequency of residues was calculated (to estimates for the $q_a$s.) and the relative frequencies of pairs of residues was calculated (to give estimates for the $p_{ab}$s). The relative frequencies were then scaled to give integer entries in the matrix. The number after the matrix represents the similarity of the sequences used to estimate the matrix, so matrices with higher numbers are used for less divergent sequences.

The Blosum matrices are generally the most effective and widely used but see also PAM (Point Accepted Mutation) matrices.

### 16.4.1 Scoring gaps

To make sequences align fully, we add gaps to one sequence or the other. A gap in $x$ corresponds to an insertion in $y$ with respect to $x$ or a deletion in $x$ with respect to $y$. Adding gaps comes with a penalty, so reduces the score for the match.

For a gap of length $k$, write $\gamma(k)$ for the penalty. We consider two forms for $\gamma$.

A *linear penalty* is defined by $\gamma(k) = -dk$ for some $d > 0$. That is, each deleted base adds a penalty of $d$.

An *affine penalty* is defined by $\gamma(k) = -d - (k-1)e$ where $d > e > 0$. $d$ is the gap open penalty and $e$ is the gap extension penalty. The affine penalty is more biologically

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **G** | 7 | | | | | | | | | | | | | | | | | | |
| **P** | -2 | 9 | | | | | | | | | | | | | | | | | |
| **D** | -1 | -1 | 7 | | | | | | | | | | | | | | | | |
| **E** | -2 | 0 | 2 | 6 | | | | | | | | | | | | | | | |
| **N** | 0 | -2 | 2 | 0 | 6 | | | | | | | | | | | | | | |
| **H** | -2 | -2 | 0 | 0 | 1 | 10 | | | | | | | | | | | | | |
| **Q** | -2 | -1 | 0 | 2 | 0 | 1 | 6 | | | | | | | | | | | | |
| **K** | -2 | -1 | 0 | 1 | 0 | -1 | 1 | 5 | | | | | | | | | | | |
| **R** | -2 | -2 | -1 | 0 | 0 | 0 | 1 | 3 | 7 | | | | | | | | | | |
| **S** | 0 | -1 | 0 | 0 | 1 | -1 | 0 | -1 | -1 | 4 | | | | | | | | | |
| **T** | -2 | -1 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | 2 | 5 | | | | | | | | |
| **A** | 0 | -1 | -2 | -1 | -1 | -2 | -1 | -1 | -2 | 1 | 0 | 5 | | | | | | | |
| **M** | -2 | -2 | -3 | -2 | -2 | 0 | 0 | -1 | -1 | -2 | -1 | -1 | 6 | | | | | | |
| **V** | -3 | -3 | -3 | -3 | -3 | -3 | -3 | -2 | -2 | -1 | 0 | 0 | 1 | 5 | | | | | |
| **I** | -4 | -2 | -4 | -3 | -2 | -3 | -2 | -3 | -3 | -2 | -1 | -1 | 2 | 3 | 5 | | | | |
| **L** | -3 | -3 | -3 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | -1 | 2 | 1 | 2 | 5 | | | |
| **F** | -3 | -3 | -4 | -3 | -2 | -2 | -4 | -3 | -2 | -2 | -1 | -2 | 0 | 0 | 0 | 1 | 8 | | |
| **Y** | -3 | -3 | -2 | -2 | -2 | 2 | -1 | -1 | -1 | -2 | -1 | -2 | 0 | -1 | 0 | 0 | 3 | 8 | |
| **W** | -2 | -3 | -4 | -3 | -4 | -3 | -2 | -2 | -2 | -4 | -3 | -2 | -2 | -3 | -2 | -2 | 1 | 3 | 15 |
| **C** | -3 | -4 | -3 | -3 | -2 | -3 | -3 | -3 | -3 | -1 | -1 | -1 | -2 | -1 | -3 | -2 | -2 | -3 | -5 | 12 |
| | **G** | **P** | **D** | **E** | **N** | **H** | **Q** | **K** | **R** | **S** | **T** | **A** | **M** | **V** | **I** | **L** | **F** | **Y** | **W** | **C** |

Figure 3: The Blosum 45 score matrix. The matrix is symmetric as $s(a, b) = s(b, a)$.

appropriate as insertions or deletions are typically created in a single event rather than building up one residue at a time.

More complex gap penalties can be used, for example, we may wish to have different penalties for gaps matched with different residues, or non-linear functions of gap length. Such penalties come at the cost of more difficult implementation.

In the algorithms below, we'll first consider the simple case of the linear penalty.