Example 4: Music. See Figure 2. This example taken from Tom Collins, Robin Laney, Alistair Willis, and Paul H. Garthwaite. Chopin, mazurkas and Markov. Significance, 8(4):154-159, 2011. doi:10.1111/j.1740-9713.2011.00519.x.

Example 5: A DNA sequence. State space is $\{A, C, G, T\}$. Need to specify transition probabilities P_{AA}, P_{AC}, P_{AG} etc. Then we obtain a random sequence by specifying a starting state and recording each state visited. An example of a random sequence looks as follows: AAGCTGCGTGTGGGGGAAGGAACTTTTGCGTGTTAGTA

The $m-$*step transition probability* is the probability of going from state i to state j in exactly m steps, $P_{ij}(m) = \Pr(X_{n+m} = j | X_n = m)$. Hence the $m-$step transition matrix is $P_m = [P_{ij}(m)]$.

A result known as the *Chapman-Kolmogorov equations* tells us $P_{m+n} = P_m P_n$ (where the right-hand side is just standard matrix multiplication). In particular, this result tells us that $P_n = P^n$, that is, the n-step transition matrix is just the nth power of the (one-step) transition matrix.

15 Introduction to genetics and genetic terminology

The history of life can be viewed, in a rather mundane way, as a long running and very complex stochastic (or random) process.

At a very basic level, and after many simplifying assumptions, we can think of the historical process explaining the relationships between species as a tree. The points where the tree splits are speciations and the leaves of the tree are different species. The past is back at the base or root of the tree and time increases from the root to the tips. Information is passed along the tree (away from the root) from one generation to the next via genetic material.

Genetic material is thought to be the only means by which biological information is passed from parent to offspring. The process of copying genetic material is imperfect, so that children will differ slightly from the parent. These imperfections consist of errors in the copying, known as mutations, and can be thought of as a stochastic process.

The fundamental objects we will be studying are sequences of characters representing biological macromolecules: DNA (Deoxyribonucleic), RNA (Ribonucleic acid) and proteins. DNA are RNA are the primary forms of genetic material. The characters in DNA and RNA sequences are drawn from 4 letter alphabets: DNA has $\Omega = \{A, C, G, T\}$ while RNA has $\Omega = \{A, C, G, U\}$. The A stands for adenine, C for cytosine, G for guanine, T for thymine and U for uracil. These are known as nucleobases or simply bases, with C, T, U being *pyramidines* and A, G being *purines*. Protein sequences consist of the twenty amino acids that are represented by the alphabet $\{A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$ (that is, all the letters except $\{B, J, O, U, X, Z\}$). We will refer to the bases in an DNA/RNA sequence or the amino acids in a protein sequence as *residues*.

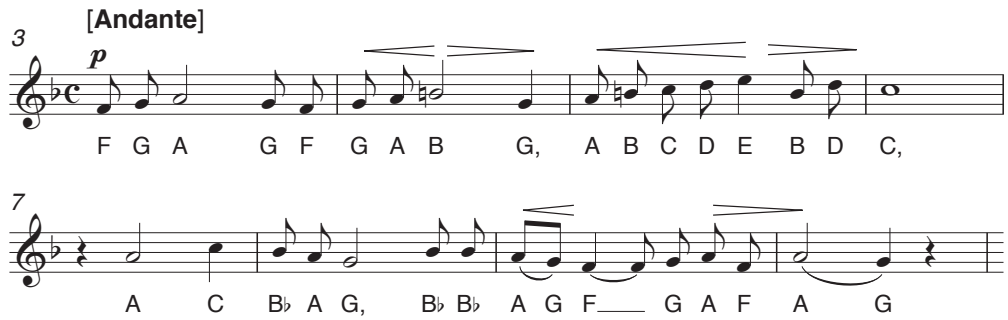

Figure 1. Bars 3–10 of the melody from "Lydia", Op. 4 No. 2, by Gabriel Fauré (1845–1924).

Table 1. Transition matrix for the material shown in Figure 1. The *i*th row and *j*th column records the number of transitions from the *i*th to the *j*th state in the melody, divided by the total number of transitions from the *i*th state.

Pitch class	*F*	*G*	*A*	*B♭*	*B*	*C*	*D*	*E*
F	0	3/4	1/4	0	0	0	0	0
G	2/7	0	4/7	1/7	0	0	0	0
A	1/8	1/2	0	0	1/4	1/8	0	0
B♭	0	0	2/3	1/3	0	0	0	0
B	0	1/3	0	0	0	1/3	1/3	0
C	0	0	1/3	1/3	0	0	1/3	0
D	0	0	0	0	0	1/2	0	1/2
E	0	0	0	0	1	0	0	0

1. A, G, F, G, F, G, A, B, G,
   F, G, F, G, A, B, D, E,
   B, C, A, F, G,
   B♭, A, F, G, A, G, A, B, G, A.

2. A, G, A, B, D, C, B♭, A, F,
   G, F, A, B, D, C, A, G,
   A, G, F, A, F,
   A, F, G, F, G, A, G, F, A, G.

3. F, A, B, G, F, G, F, G, A,
   B, C, A, G, F, G, F, G,
   B♭, A, G, A, G,
   A, F, G, B♭, A, B, G, F, G, A.

Figure 2: An example showing how a piece of music can be modelled as a Markov chain. The original piece, a fragment of Lydia by Fauré, is shown at the top. Just the pitches are considered in this simple Markov model. The transition matrix between pitches (centre) is constructed from empirical counts of the observed transitions. Three random realisations of the process are given at the bottom.

In eukaryotes (organisms with cells that have a nucleus), the three types of sequences related to each other by the Central Dogma of Molecular Biology that states, DNA makes RNA makes Protein. Or, more prosaically, DNA is *transcribed* into a type of RNA called mRNA that is then *translated* into protein.

There are some good animations showing how translation and transcription work at `www.hhmi.org/biointeractive/animations/index.html`, in particular see the DNA transcription and translation animations. A Japanese anime style film of the central dogma is also worth a look: `http://www.youtube.com/watch?v=-ygpqVr7_xs`.

Parts of the the DNA sequence encode information for proteins. These regions are known as genes and must be transcribed to RNA before being built into proteins. When the DNA is transcribed to RNA, all bases are copied exactly except that T (thymine) is transcribed as U (uracil). Once copied, the RNA is edited at splice sites so that only exons remain (the introns are edited out). This leaves the *messenger RNA*, mRNA, which is then translated to a protein sequence (poly-peptide chain). This translation occurs via the genetic code which translates consecutive triples of RNA bases (known as a codon) into one of the 20 amino acids. There are $4^3 = 64$ possible codes since there is an alphabet of 4 bases. 60 of these code for proteins, 1 (AUG) is a start codon and 3 (UAA, UGA and UAG) are stop codons signalling the start or finish of a protein. A particular amino acid may be encoded by just one codon (e.g. AUG→Methionine(M)) or up to 6 (e.g. any of UUA, UUG, CUU, CUC, CUA, CUG→Leucine (L)). Once the poly-peptide chain is formed it folds into three dimensional molecule, taking on a particular structure.

Example: The sequence `atgaggttgacgctactttgttgcacctggagggaa` can be split into codons `atg agg ttg acg cta ctt tgt tgc acc tgg agg gaa` which translate into the protein sequence `MRLTLLCCTWRE`. □

In this course, we are only interested in the primary structure of sequences, that is, the order in which residues occur along the sequence. We will ignore the secondary, tertiary and quaternary structure of proteins — secondary structure is the name for the regular substructures such as alpha helices and beta sheets, the tertiary structure are the three dimensional structures of single molecules while quaternary structure are the complex forms taken by collections of single protein molecules. The study of these more complex structures is known as structural bioinformatics.

When DNA is passed from one generation to the next, the copy made is not exact. There are a number of processes that cause differences to arise between the parent and child. Recombination is one such process and involves the mixing of the maternal and paternal copies of DNA when the gametes (eggs or sperm) are produced. Other processes are generally thought of as mutations. The simplest are *point mutations* where the offspring sequence differs from the parent sequence by a single base (residue). This type of mutation is called a *single nucleotide polymorphism*, abbreviated as SNP and pronounced 'snip'. *Insertions* (or deletions) refer to the child sequence gaining (losing) one or more base than the parent. Larger scale mutations include: *gene duplication* which is a large scale insertion where the child inherits extra copy of a region containing

a whole gene. Other large scale mutations include inversions (part of the sequence is reversed end to end) and translocations (a piece of the sequence is copied out of order).

Examples of mutations: Consider the short sequence `cgctcaccatgaagcgtttcactaat`. We demonstrate types of mutations showing the original sequence and a mutated version of it below with X marking the mutation.

- Single nucleotide polymorphism (SNP)
  ```
  cgctcaccatgaagcgtttcactaat
  cgctcgccatgaagcgtttcactaat
  .....X.....................
  ```
- Insertion
  ```
  cgctcacc----atgaagcgtttcactaat
  cgctcacctgatatgaagcgtttcactaat
  ........XXXX..................
  ```
- Deletion
  ```
  cgctcaccatgaagcgtttcactaat
  cgct----atgaagcgtttcactaat
  ....XXXX......................
  ```
- Duplication (the copied region is marked with parentheses). Note that duplication usually refers to gene duplication where whole genes are copied.
  ```
  cgctcaccatgaagcgtttcacta-----------at
  cgctcaccatgaagcgtttcactacaccatgaagcat
  ....(.........).........XXXXXXXXXXX..
  ```
- Inversion (again, this typically happens at a larger scale than shown here)
  ```
  cgctcaccatgaagcgtttcactaat
  cgctctaccagaagcgtttcactaat
  .....XXXXX.................
  ```

□

All these processes can be modelled and studied, with varying degrees of difficulty. We'll focus primarily on the question of how to align the sequences, how to identify regions of interest in sequences (for example, genes), and given aligned sequences, how can we reconstruct the evolutionary history (the tree) of those sequences. This last problem will require us to model the the mutation process where we restrict ourselves to looking at how point mutations arise.

The models we use will use are relatively simple, sometimes to the point of being downright crude. It is good to keep in mind the quote from the famous statistician George Box who said, "All models are wrong but some are useful".

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.