Lecture Notes: BIOL2007 Molecular Evolution

Kanchon Dasmahapatra (k.dasmahapatra@ucl.ac.uk)

Introduction

By now we all are familiar and understand, or think we understand, how evolution works on traits and characters – "survival of the fittest" and stuff like that. Heritable changes in traits and characters are a result of underlying changes at the level of DNA. What we are probably less acquainted with is how DNA itself changes and evolves. In this lecture we will explore the theme of molecular evolution, taking a look at how and why nucleotide and amino acid sequences evolve.

Background information

But first, some background information. DNA has many different roles. Most of our DNA does not code for any protein and is therefore different in character from sections of DNA that do code for proteins. In eukaryotes, genes (or stretches of DNA that code for a particular protein) frequently consist of exons and introns. After the DNA is transcribed into messenger RNA (mRNA), the intron regions are removed (spliced) and the exons are pasted back together. This modified mRNA is then translated into the protein. Therefore, while exons represent protein coding regions, introns do not.

Translation of the mRNA into the protein is carried out according to the genetic code that describes how each of the 20 amino acids is specified by a consecutive sequence of three of the four bases comprising DNA. These triplets of bases coding for each amino acid are called codons. As there are 64 such codons and only 20 amino acids, the same amino acid is often encoded by more than one codon. Codons which produce the same amino acid are called synonymous codons. Point mutations (replacement of one base by another) that do not change the amino acid coded by codons are known as synonymous or silent mutations. This is often the case for point mutations occurring in the third position of codons. While point mutations that change the coded amino acid are called non-synonymous mutations.

Modes of molecular evolution

There are a number of different ways or modes by which molecular evolution takes place. Most of you will be familiar with these mutational modes, so we won't go into a great deal of detail with them. Here is a list of some of these modes:

Single base pair changes, substitutions or point mutations
Insertions or deletions, also known as indels
Gene duplications - formation of multigene families and pseudogenes
Slippage – microsatellite length changes

Chromosomal mutations
Transposition

The Classical vs. the Balance Schools

Before the 1960s (in the days before there was any data about protein or DNA variation) there were two schools of population geneticists: the classical and balance schools. The classical school believed that polymorphisms, the existence of more than one allele in a population of genes, were rare. They argued that natural selection was a mainly purifying force that removed any deleterious alleles that may arise or would drive any advantageous alleles to fixation. Therefore, they believed that individuals were homozygous for most loci. In contrast, the balance school believed that polymorphisms were common. Polymorphisms at the various loci were thought to be maintained by different forms of balancing selection that favoured heterozygotes over homozygotes. Both schools of thought agreed that natural selection was the force driving molecular evolution.

In the mid-1960s the technique of protein electrophoresis was discovered allowing investigation into the levels of enzyme polymorphism. The results showed that large amounts of genetic variation was present in natural populations, appearing to vindicate the balance school's beliefs. The balance school held that these high levels of polymorphism were maintained by balancing selection. However, others argued that maintaining these high levels of polymorphism at thousands of loci by balancing selection would be very costly. Summed over multiple loci this high genetic load would be large enough to drive populations to extinction!

The neutral theory of molecular evolution

However, the high levels of polymorphism can explained without encountering excessive genetic load simply by dropping the assumption that natural selection is the driving force of molecular evolution and instead allowing the majority of mutations fixed to be neutral and therefore have no effect on fitness. Two papers, by Kimura in 1968 and by King and Jukes in 1969, first proposed this neutral theory of evolution. Since then it has become one of the most important and controversial theories in evolutionary biology.

In his paper, Kimura made some simple calculations.

If μ = mutation rate per gene per generation, and N = effective population size

Number alleles in a diploid population = $2N$

Number of new mutations per generation = $2N\mu$

Most of the time a new neutral allele will be quickly lost from the population by genetic drift. But sometimes it will drift into the population and get fixed, that is, it will replace (or substitute) the original allele in the population.

The probability that the new allele will drift to fixation = 1/2N (this is equivalent to the probability of reaching into a bag containing 2N black marbles and pulling out the only red marble in the bag).

Therefore, the rate of substitution of an allele by a new allele = $2N\mu \times 1/2N = \mu$

Basically meaning that the rate of neutral molecular evolution is independent of population size and is simply equal to the neutral mutation rate.

The average time for a neutral mutation to drift to fixation is 4N generations. Therefore, while the rate of origin and fixation of new mutations (μ) is independent of population size, the rate of progress of the mutation through the population is proportional to the population size. Therefore, under the neutral theory, polymorphisms in a large population are simply a result of lots of neutral mutations arising and passing through the population at a slow rate such that at any one time there are several different alleles at a particular locus drifting through the population.

According to the neutralists, most mutations are either deleterious and are selectively removed, or are "effectively neutral", in which case there is a small probability that they are fixed. Natural selection is incorporated, but as a purifying force, removing deleterious mutations and with only a small role in fixing new mutations. As we have seen above, the probability of fixation of a neutral allele by drift is 1/2N. If this probability is bigger than the selection pressure, the influence of drift is greater than that of selection and the mutation is effectively neutral. So, the neutral theory does not argue that most mutations are completely neutral, but that any selection pressures are outweighed by the effects of drift.

In contrast, according to the selectionists, mutations are fixed because they confer a selective advantage and that neutral mutations are rare.

Some predictions from the neutral theory

1. There is a constant rate, or molecular clock, of sequence evolution
2. There is an inverse relationship between the rate of substitution and the degree of functional constraint acting on a gene, such that functionally constrained genes or gene regions evolve at the lower rate and vice versa.

The molecular clock

A molecular clock is compatible with the neutral theory, as the rate of substitution of a neutral mutation is μ, and is not affected by population sizes or selective pressures. As long as μ is constant across species and most molecular evolution is neutral then the rate of evolution should be constant across lineages.

At first glance the evolution of sequences does indeed appear to be constant over time. However, on closer inspection, significant variation among lineages becomes evident. One way of testing the molecular clock is by using the relative rate test. A number of explanations have been put forward to explain deviations from molecular clock, such as differences between lineages in generation times, metabolic rate, DNA repair efficiency and even a bit of natural selection.

Functional constraints and the rate of substitution

According to the neutral theory most mutations are deleterious and the rest are neutral (advantageous mutations are very rare). However, genes will differ in the proportion of mutations that are deleterious. The higher the functional constraint on the gene, the greater is the strength of negative selection removing mutations. In a gene with high functional constraints the vast proportion of mutants will be deleterious and be removed by selection, leaving only a small fraction of neutral mutations which will result in a low rate of substitution. In a less constrained gene a larger fraction of the mutations will be neutral leading to a higher substitution rate.

Examples:
Variation in rates between and within genes
Substitution rates in non-coding regions – pseudogenes, introns.
Synonymous vs. non-synonymous mutations rates

Testing the neutrality of mutations using d_N/d_S:

1) Sequence copies of the gene of interest from a variety of species.
2) Construct a phylogeny of the species using the sequence or other data.
3) Identify synonymous and non-synonymous mutations.
4) Calculate the average synonymous rate of subsititution, d_S, the average non-synonymous rate of substitution, d_S, and the ratio, $\omega = d_N/d_S$.

We assume that synonymous mutations are neutral. As we have seen, due to functional constraints, in most genes $d_N < d_S$, and $\omega < 1$.
If $d_N > d_S$, $\omega > 1$, the coding changes are occurring more rapidly than silent changes. This is indicative of positive selection to change the amino acid sequence.

Positive selection – evidence against the neutral model?

Examples – mutation rates within the major histocompatibility complex and HIV envelop proteins.

However, the procedure described above for detecting positive selection is insensitive. The procedure calculates a single value of ω for a gene, averaging over the whole gene. It is possible that only a few parts of a protein are under strong positive selection. If this is the case, averaging

over the whole gene will mean that the $\omega > 1$ signal from the bits under positive selection will be swamped by the $\omega < 1$ signal coming from the majority of the gene.

Some improvements have been made to detecting positive selection, many of them coming from Zihang Yang here at UCL. e.g. lysine, see OHP.

The nearly neutral model

By the early 1970s it was becoming clear that the neutral theory was too simplistic. There was evidence for positive selection acting on mutations and the molecular clock did not tick at a perfectly constant rate. This gave rise to the nearly neutral model of molecular evolution. According to this theory, mutations in non-coding DNA and synonymous sites are still strictly neutral. However, non-synonymous mutations are no longer regarded as being neutral and are instead nearly neutral, being either slightly deleterious or slightly advantageous. Therefore, the nearly neutral model includes weak selection as well as genetic drift.

So who is correct, the neutralists or the selectionists?

It seems that both genetic drift and natural selection determine the evolution of mutations. Neutralists are probably correct in that most mutations are neutral, especially in non-coding DNA and synonymous sites. However, evidence of natural selection is sometimes evident at non-synonymous sites when molecular evolution over short evolutionary time periods are examined.

Reading

Chapter 7 – Models of molecular evolution
Page, R. D. M. and E. C. Holmes. 1998. Molecular Evolution: a phylogenetic approach. Blackwell Publishing.

Yang Z and J. P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution* 15:496-503.

Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature* 217:624-626.

King, L. C. and T. H. Jukes. 1969. Non-Darwinian evolution. *Science* 164:788-798.

Li, W. 1993. So, what about the molecular clock hypothesis? *Current Opinion in Genetics and Development* 3:896-901.