Abstract

In this thesis, we investigate various models of DNA regulatory sequence evolution as they apply to several different organisms. In Chapter 2 we focus on a model for humans, and in Chapter 3 we extend our study to include organisms with larger effective population sizes such as Drosophila (fruit flies) and yeast.

A regulatory sequence is a short sequence of DNA (in vertebrates, many are 6–9 nucleotides long) which is a binding site for transcription factors that promote or inhibit transcription of DNA to make proteins. Changes in regulatory sequences can cause changes in gene expression which in turn can lead to phenotypic evolution. Thus, one possible explanation for the substantial organismal differences between humans and chimpanzees is that there have been changes in gene regulation. Given what is known about transcription factor binding sites, this motivates the following probability question: given a 1000 nucleotide region in our genome, how long does it take for a specified 6–9 letter word to appear in that region in some individual? Stone and Wray (2001) studied this problem by simulation. In Chapter 2, we do a mathematical analysis of this problem focusing on words of length 6 and 8. We show that the average waiting time is 100,000 years for words of length 6, while for words of length 8, the waiting time has mean 375,000 years when there is a 7 out of 8 letter match in the population consensus sequence (an event of probability roughly 5/16) and has mean 650 million years when there is not. Fortunately, in biological reality, the match to the target word does not have to be perfect for binding to occur. If we model this by saying that a 7 out of 8 letter match is good enough, then the mean reduces to about 60,000 years. Hence, the inexactness of transcription factor binding is important to allow regulatory sequences to evolve at a reasonable rate.

In Chapter 3, we continue our study of how the tempo and mechanisms of regulatory sequence evolution depend on an organism's effective population size. The results above assume that the human effective population size is 10,000 which is relatively small compared to organisms such as Drosophila and yeast whose effective population sizes are in the millions. In addition, our previous results suggest that new regulatory sequences usually come from small modifications of existing sequence. Here we examine the waiting time for a pair of mutations, the first of which inactivates an existing transcription factor binding site and the second which creates a new one. Consistent with recent experimental observations for Drosophila and yeast, we find that a few million years is sufficient in these species, but for humans with a much smaller effective population size, this type of change would take more than 100 million years.