

# 1 Virus Simulator

Our group is currently working on various projects involving the study of virus evolution [McCauley & Hein, 2006, de Groot *et al.*, 2006, McCauley *et al.*, 2006]. To test whether devised methods work, and to investigate the true nature of the evolutionary process, requires a program which can simulate the evolution of viruses over time. Developing such a virus simulator is an exciting, challenging and highly useful project with the potential of many applications in the bioinformatic field.

There are various models available for the evolution of sequences, some more realistic than others. Over time nucleotides can mutate or even be deleted and inserted, as shown in figure 1. The rates at which this happens, are obviously highly dependent on the functionality of the genomic region. Additionally, not just single nucleotides undergo this process, but entire chunks of DNA can get inserted and deleted. RNA secondary structure also lays down constraints on the evolution of the sequence, and a model for this would need to be incorporated into a more advanced sequence simulator.



Figure 1: A reference sequence above with a descendant subsequence below. We observe one substitution (C → A), one deletion (A → -) and one insertion (- → C) having occurred over time.

In the case of viruses however, things are complicated even further, due to them generally have a very complex structure. Viruses are generally constrained in length and for maximal efficiency, this often results in the presence of overlapping coding regions, where one nucleotide is involved in simultaneously coding for more than one gene (see figure 2).

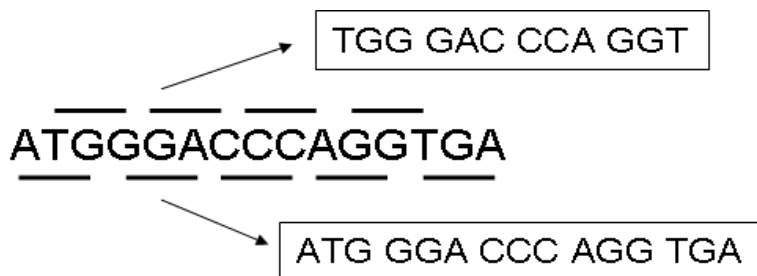


Figure 2: A multiple coding region, where the above bars indicate the triplets encoding one gene, and the below ones indicate the triplets encoded by the other gene. The boxes show the different coding sequences for each gene.

Within a gene, one amino acid is made up of a nucleotide triplet, where the same amino acid can be encoded for by several different triplets — for example CCT, CCG, CCA and CCC all code for the amino acid proline. When one of these nucleotides mutates, this may or may not result in a change of

amino acid. If indeed it does, we call this a non-synonymous substitution. Some viral regions are under heavy evolutionary constraints, meaning that very few non-synonymous mutations are observed. Other regions are encouraged to mutate faster than average, for example to avoid immune response. The process underlying this phenomenon is called *selection*, where positive and negative selection correspond to fast and slow evolution. In a multiple coding region however, as shown in figure 3, a mutation may result in a non-synonymous substitution in one gene and in a synonymous one in the other, thus complicating the concept of selection within our evolutionary model.

Looking back at the top sequence in figure 1, it is 33 nucleotides long and if we only allowed substitution events, there would be 3 alternative nucleotides at each position. Thus there are 99 sequences that are reachable from that sequence in 1 substitution. Since we can specify the basic rates of substitution and selection strength against an amino acid change, we can write down exactly the probability that it will jump to any of those 99 neighbour sequences. The same is the case for allowing insertions and deletions.

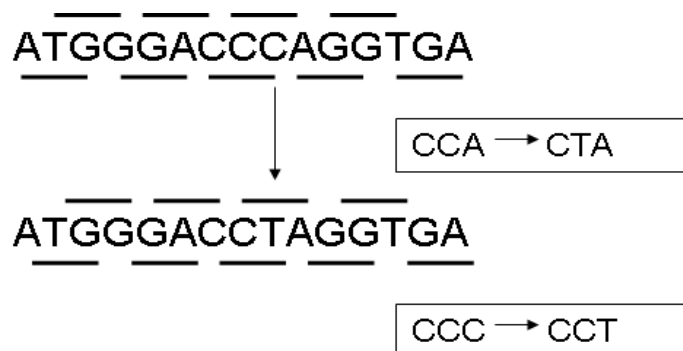


Figure 3: A C → T mutation causing a non-synonymous change from proline to leucine above, and a synonymous change below.

When testing various methods to analyze the viral genome, it is important to know how statistically significant one's observations are. For this one needs a simulator, which mocks the evolution of viral sequences. One wishes to make statements about what effect recombination, or secondary structure conservation could have on one's results, as well as how far apart sequences must be to even be able to pick up on selection levels. Additionally, developing a realistic virus simulator could be a step towards understanding the evolutionary process better, since one could simulate from different models and see which one fits the real data best.

The simulator development can be phased into a basic, regular and advanced simulator, dependent on how many levels of realism are desired. The basic simulator is a guaranteed reachable goal, while the creation of an advanced simulator is both a major challenge and would be viewed as a major success.

The things that could be drawn into account, when attempting to model sequence evolution in a realistic way, include:

- Nucleotide sequence evolution
- Overlapping reading frames

- Codon bias
- Genome structure evolution
  - Indels
  - Loss/gain of start/stop codons
- Heterogeneity & stationarity of selection
- RNA secondary structure
- Insertion/deletion of new genes

## 1.1 Nucleotide Sequence Evolution

There are several models which describe the substitution process underlying the evolution of nucleotide sequences, the most general being the the General Time Reversible (GTR) model (see [Tavaré, 1986]). On top of this we must draw the possibility into account of nucleotides being added or deleted over time.

## 1.2 Overlapping Reading Frames

Since three nucleotides code for one amino acid, each locus may be coding for up to three genes simultaneously. Overlapping regions are likely to evolve in a different way to single coding regions, due to the multiple coding constraints they are under. One paper postulating a model for the evolution in overlapping regions is by [Hein & Støvlbæk, 1995], however an improvement on this would be desirable.

## 1.3 Codon bias

It is known that the usage of some codons is much more common than that of others [Sharp & Li, 1987] – this can be related to G-C content, gene expression and translational efficiency. We could include a  $64 \times 1$  codon usage matrix and weigh the mutation probabilities of each nucleotide by the appropriate matrix entry to account for codon bias. In overlapping regions one would presumably multiply up the weights.

## 1.4 Genome Structure Evolution

Genome structure may evolve over time, meaning that start and stop codons of a gene can change. This obviously has a large impact on nucleotide sequence evolution, since something that priorly has been non-coding could suddenly become functional and thus potentially more constrained.

## 1.5 Heterogeneity & stationarity of selection

When modelling selection across the genome we must draw possible intra- and intergenic evolutionary pressures into account. A simulator would, for each nucleotide position in the sequence, have a certain selection level attached to it, and let the locus evolve accordingly. It is also possible to let this selection strength change over time, especially in the case of a region newly becoming coding.

## 1.6 RNA secondary structure

We may fix an RNA secondary structure on top of the sequence and either insist on it being conserved or evolve it in conjunction with the sequence [Holmes, 2004]. We could have two different mutational models for stem and loop regions, as well as draw into account co-evolution of two sites due to stem base-pairing.

## 1.7 Insertion/deletion of new genes

At a very small rate we could allow for insertion and deletion of new genes. Whether this would occur as a gain/loss of functionality (e.g. an elongation/shortening of an open reading frame) or a large indel event is optional.

## References

- [de Groot *et al.*, 2006] de Groot,S., Mailund,T., Hein,J. (2006) Comparative Annotation of Viral Genomes with Non-Conserved Gene Structure, *Bioinformatics*, In Review
- [Durbin *et al.*, 1998] Durbin,R., Eddy,S., Krogh,A., Mitchison,G. (1998) Biological Sequence Analysis, *Cambridge University Press*
- [Hein & Støvlbæk, 1995] Hein,J., Støvlbæk,J. (1995) A maximum-likelihood approach to analyzing nonoverlapping and overlapping reading frame, *Journal of Molecular Evolution*, **40(2)**, 181-189.
- [Holmes, 2004] Holmes,I. (2004) A probabilistic model for the evolution of RNA structure, *BMC Bioinformatics*, **5:166**
- [McCauley & Hein, 2006] McCauley,S., Hein,J. (2006) Using HMMs and observed evolution to annotate viral genomes, *Bioinformatics*, Advance Access published online on April 13, 2006
- [McCauley *et al.*, 2006] McCauley,S., de Groot,S., Mailund,T., Hein,J. (2006) Annotation of Selection Strengths in Viral Genomes, *About to be submitted*
- [Rogozin *et al.*, 2002] Rogozin,I., Spiridinov,A.N., Sorokin,A.V., Wolf,Y.L., Jordan,I.K., Tatusov,R.L., Koonin,E.V. (2002) Purifying and directional selection in overlapping prokaryotic genes, *Trends in Genetics*, **18(5)**, 228-232.
- [Sharp & Li, 1987] Sharp,P.M., Li,W.H. (1987) The codon adaptation index — a measure of directional synonymous codon usage bias, and its potential applications, *Nucleic Acids Research*, **15(3)**, 1281-1295.
- [Tavaré, 1986] Tavaré,S. (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. In: Miura RM (ed) Some mathematical questions in biology DNA sequence analysis, *American Mathematic Society*, Providence, 57-86
- [Yang & Swanson, 2002] Yang,Z.H., Swanson,W.J. (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes, *Molecular Biology and Evolution*, **19(1)**, 49-57.