Alignment of large RNA viral genomes

16.08.11 Zsuzsanna Sükosd, Jørgen Kjems & Jotun Hein

The genomes of RNA viruses (eg. HIV or influenza) contain several layers of information. For example, they code for functional proteins (sometimes in overlapping reading frames), RNA secondary structure, splice sites, and noncoding RNAs, all of which act together to hijack the host cell's resources to replicate the virus. Figure 1 illustrates the organization of the HIV-1 genome.

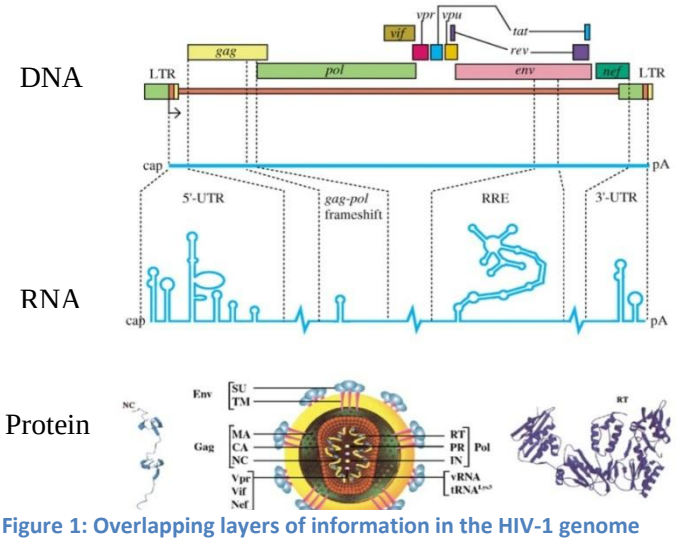

Figure 1: Overlapping layers of information in the HIV-1 genome

Ordinary sequence alignment methods assume a uniform evolutionary model over the entire sequence. However, viral genomes are clearly under non-uniform evolutionary pressures: in a simple model, some parts evolve to conserve protein coding, some parts evolve to conserve RNA secondary structure, and some parts conserve both.

Hidden Markov models [1] have long been used both to distinguish coding and non-coding regions in sequences, and to produce sequence alignments. In this project, the goal is to couple these approaches, and produce a multiple alignment of large viral genomes on the basis of two coupled Hidden Markov Models: one is responsible for determining the local context, and the other is responsible for the evolutionary model.

The advantage of this approach is that it will allow the solving of the alignment and annotation problems simultaneously. On the other hand, combining HMMs might be theoretically challenging [2], which may make it necessary to make compromises in the model. For example, instead of actually coupling the two HMMs, each genome can first be annotated using one of them to find eg. "Protein-coding region", "Structured region", etc. The ends of these regions can then be used as (fixed or probabilistic) "pivots" for the second HMM to find an optimal global alignment.

References:

[1] Durbin et al.: Biological Sequence Analysis, Cambridge University Press

[2] http://www.stats.ox.ac.uk/__data/assets/file/0016/3328/combinedHMMartifact.pdf

[3] Eddy, SR: Multiple alignment using Hidden Markov Models, Proc Int Conf Intell Syst Mol Biol. 1995;3:114-20.

[4] Sinha S, and He X, 2007 MORPH: Probabilistic Alignment Combined with Hidden Markov Models of cis-Regulatory Modules. PLoS Comput Biol 3(11): e216

[5] McCauley, S and Hein, J: Using hidden Markov models and observed evolution to annotate viral genomes. Bioinformatics (2006) 22(11): 1308-1316

[6] http://www.stats.ox.ac.uk/__data/assets/file/0016/3328/combinedHMMartifact.pdf