

Computational approaches to the prediction of insertional RNA editing sites

Ralf Bundschuh, The Ohio State University; Department of Physics

The central dogma of molecular biology states that there are two steps in making a protein. First, a verbatim copy of the genomic sequence is created in the form of an mRNA during the process of transcription. Second, this sequence information is turned into a protein sequence using the genetic code during the process of translation. However, some organisms add an extra step called RNA editing in this process which changes, inserts, or deletes individual bases in the messenger RNA before it is translated into a protein. An especially striking example is the mitochondrion of the slime mold *Physarum polycephalum* which (among other editing events) inserts an extra C on average every 25 bases. It does so always at the same positions and extremely reliably. The consequence of such editing activity is that it is not obvious from the genomic sequence any more what protein it codes for and even if it codes for a protein at all. All gene finding methods - comparative or de novo - fail in such a situation. Indeed, in the mitochondrion of *Physarum* only 11 genes were known while in the mitochondrion of *Dictyostelium discoideum* – the genome of which has about the same length as the *Physarum* mitochondrial genome - 44 genes are known. Even for half of the genes that were known, the positions of the inserted Cs (the editing sites) were not known since experimentally finding these positions requires sequencing the edited mRNA which in turn relies on primers that have to be complementary to this unknown mRNA sequence.

In this talk I will give an overview over several computational approaches to the problem of gene finding and prediction of editing sites in the presence of insertional RNA editing. First, I will discuss how sequence alignment techniques can be used to compare protein sequences from other organisms to the genomic DNA sequence of the *Physarum* mitochondrion in order to identify new genes through comparative gene finding as well as in order to very reliably predict the position of editing sites. In fact, I will show four examples of previously unknown genes in which this method not only found the genes but was also used to choose primers to experimentally verify these genes. This first technique is only useful if one is interested in genes that one is explicitly looking for because one would expect to find them. Thus, I will discuss in the second part of my talk how we look for unexpected genes by modifying gene finding Hidden Markov models to take into account insertional editing as well as by modifying the Smith-Waterman sequence alignment algorithm to scan the mitochondrial genome of *Physarum* against the complete non-redundant protein database. Taking all methods together we now know over 30 putative genes some of which might even be interesting candidates for the (so far unknown) editing machinery itself.

This work has been supported by grant no. DMR-0404615 from the National Science Foundation.