NEXT GENERATION GENE FINDING: USING CONDITIONAL RANDOM FIELDS TO SEARCH FOR ANTIBIOTIC RESISTANCE GENES

Mariana Buongermino Pereira[1], **Erik Kristiansson**[1] **and**
Marina Axelson-Fisk[1]
[1] Department of Mathematical Sciences, Chalmers University of Technology
and University of Gothenburg, Sweden

Traditionally, bioinformatics has successfully used hidden Markov models (HMMs) to search for genes in biological sequences. The advantages of HMMs over other analysis methods is that they have high accuracy, provide a probabilistic interpretation of the results, and assure an appropriate framework in which the model can be tested and validated. However, HMMs are limited to deal with the local dependencies in biological sequences, while in nature long-range and complex dependencies are common.

Conditional random fields (CRFs) can be seen as an extension of HMMs. They preserve the advantages of HMMs, while having the ability to include complex, long-range dependencies into the model. This is achieved because CRFs are discriminative models while HMMs are generative, which means that HMMs treat the observed sequence in a temporal order in order to be able to "generate" it again. In contrast, CRFs consider the entire sequence without this temporal restriction making the model more flexible. This flexibility also allows dependencies to be tackled implicitly, meaning that unknown dependencies may also be included in the model. Finally, CRFs allow the components in the model to be both probabilistic and non-probabilistic, facilitating heterogenous information to be added to the model.

In this project, we apply CRFs to identify antibiotic resistance genes that have been horizontally transferred by integrons. Integrons are genetic elements whose main function is to aid the incorporation of new genes into the genome. Resistance genes incorporated this way have a complex structure. For instance, its dependence on the integrase located upstream of the genes is a long-range dependency suitable to be studied using CRFs. Another important dependency is the presence of a characteristic motif located after the incorporated genes. These motifs have a secondary structure similar to non-coding RNA, where they fold to form a regulatory element using Watson-Crick complementary base-pairing. Thus, these motifs have also a local complex dependency structure. In the present work, a model is created to identify integrons and the inserted genes. The model is developed within the CRF framework where we describe both local and long-range dependencies. The ultimate aim is to use the model to study the evolution of antibiotic resistance and the sharing of resistance genes between integrons.

Keywords: hidden Markov models, conditional random fields, bioinformatics, gene finding, integrons, antibiotic resistance.

References:

Axelson-Fisk, M. (2010). *Comparative Gene Finding: Models, Algorithms and Implementation.* London: Springer.

Lafferty, J., A. McCallum, F. Pereira (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML.* 282–289.