Discriminating coding and non-coding RNAs using comparative sequence analysis

In my talk, I will first briefly review challenges and the current state-of-the-art for genome-wide annotation of non-coding RNAs. To accurately locate non-coding RNAs in a genome it turned out to be critical to know what parts are actually coding. Although there are many sophisticated protein gene finders and very good annotations exist for most model organisms, there are also ambiguous and non-standard situations in which these programs fail.

We have therefore developed a new algorithm called "RNAcode", a program to detect coding regions in multiple sequence alignments that is optimized for emerging applications not covered by current protein gene finding software. Our algorithm combines evolutionary information from nucleotide substitution and gap patterns in a unified framework and also deals with real-life issues such as alignment and sequencing errors. It uses an explicit statistical model with no machine learning component and can therefore be applied "out of the box", without any training, to data from all domains of life.

I will demonstrate how RNAcode was used in combination with mass spectrometry experiments to predict and confirm seven novel short peptides in E. coli that have evaded annotation so far. As another example of a typical application, I will show how RNAcode can be used together with the structural RNA gene finder RNAz to study ambiguous cases of dual function genes that function on both the RNA and protein level.