Cologne Colloquium on Theoretical Physics

Cologne Evolution Colloquium

Joint Seminar

Erik Aurell

KTH - Royal Institute of Technology, Stockholm

Minimal absent words in genome sequences

Absent words are sub-sequences of letters that cannot be found in a given text. Minimal absent words (MAWs) are absent words all of the sub-sequences of which can be found in the text. It has been observed for some time that the length distribution of MAWs in genome sequences have a curious two-mode structure with on the one hand many long fairly short MAWs ("the bulk") and on the other hand some very long MAWs ("the tail"). I will show that the first feature arises from statistical sampling of sub-sequences from a random genome while the second can be explained by a simple probabilistic model of genome evolution. Tail MAWs also seem to carry biological information as will be discussed in the talk. We were led to the study of MAWs from a biotechnological problem of optimal tag design for (tagged) RNA-sequencing. I will therefore also describe some of the results obtained by this method to distinguish primary from processed RNA in the human pathogen Enterococcus faecalis, including the discovery of many new non-coding genes in this organism.

Friday, January 15, 2016, 16:30
Institute for Theoretical Physics, New Building
Seminar Room 003, Ground Floor

Hosted by Michael Lässig