Making sense of too many too short biological sequences

Stefan Canzar
Gene Center Ludwig-Maximilians-Universität München, Münich

10 May 2018

Abstract

Advances in computer science and the explosion of large-scale, quantitative experiments have created a data-driven revolution in biology. In particular, next-generation sequencing (NGS) technology allows us to rapidly sequence many millions of DNA molecules and has been used to address a wide range of fundamental biological questions. Every short DNA sequence ('read') generated by NGS instruments carries little information by itselft, and thus the reconstruction of the desired biological measurement involves solving a complex computational puzzle. My talk focuses on engineered algorithms that piece reads together to reconstruct biologically meaningful measurements. In the first part, I will introduce a novel method to reconstruct the cellular transcriptome from NGS data. We determine transcripts that improve prediction by solving an (NP-hard) optimization problem, digging out even low-expressed transcripts. In the second part I will introduce a novel method to compare phylogenetic trees.