

Decoding Novel Genomes: From Microbiomes to the Eukaryota

Mark Borodovsky
Department of Biomedical Engineering
and Division of Computational Science
Georgia Institute of Technology
Atlanta, Georgia, USA

Providing effective means for fast and accurate biological interpretation of newly sequenced genomic DNA is an important and long standing problem of computational genomics. The complexity of the task varies among genomes but is never simple. Currently, for each new genome a custom built annotation pipeline is constructed by integration of existing ab initio and comparative genomic methods. Still, a satisfactory and consistent solution of the jigsaw puzzle of genome annotation frequently requires additional experimental efforts (such as EST/cDNA sequencing, etc.)

A genome annotation pipeline centers around one or several gene prediction algorithms. Gene finding by similarity search has an advantage of generating insights into functional role of predicted gene and protein. Still, a pipeline must include an ab initio gene finding algorithm able to identify genes missed by similarity detection methods, the genes whose sequences do not carry conservation detectable by alignment to related genomes or whose protein products do not show significant similarity to previously known proteins.

Current ab initio gene finding algorithms use the concepts of statistical analysis and optimization. One of the mainstream approaches is the reformulation of the problem as finding the optimal parse of the genomic sequence into fragments with distinct statistical characteristics. This problem can be reduced to a classic task for dynamic programming: finding an optimal path through a network with weights/scores assigned to nodes and vertices.

An adequate assignment of weights/scores presents a significant challenge. This task is related to identification of parameters of statistical models (such as hidden Markov models) representing various types of functional sequences and sites in a given genome. A conventional approach to parameter estimation is the use of manually curated sets of training sequences. However, such sets may become available only at well advanced stages of genome sequencing projects.

In this lecture we will consider the general schemes of ab initio gene prediction. We also will describe the approaches to “supervised” and “unsupervised” estimation of model parameters. Interestingly, the unsupervised approach has proved itself to be very important for two rapidly developing branches of genomics: i/ for prokaryotic metagenomes obtained directly from environment and becoming a rich source of information about non-cultivated microbial species and ii/ for “compact” eukaryotic genomes, such as fungi, which rather short genome size allow to obtain full genome sequence in a relatively short time.