Ab initio characterization of promoter regions based on Conditional Random Fields

Ígor Bonadio, Mauro de Medeiros, Alan Mitchell Durham

Programa de Pós Graduação em Ciência da Computação da USP, Programa de Pós Graduação em Bioinformática da USP, Instituto de Matemática e Estatística da USP

Gene prediction aims to find the location of genes in a genome. However, current gene prediction programs just identify the coding regions and not the promoter regions. Identifying the promoter region involves correctly locating the transcription start site (TSS), which is a difficult task due the lack of a strong signal around this site. Many techniques were developed but their number of false positives is too high for practical use. In this project we propose a new method based on Conditional Random Fields (CRF) that presents a much better prediction rates that previous algorithms. With our approach we are able to predict not only the coding region but also to approximately locate the TSS, TATA-box and CCAAT-box. The use of CRFs enables us to effectively combine the annotation generated from a traditional GHMM-based gene predictor with information about the nucleotide composition of the intergenic region, the distance distribution between start codons and TSSs and between TSSs and TATA-boxes. We validated our methodology using the PlantProm database, which have annotation of the promoter region of 579 plant genes (monocots and dicots) including experimentally verified TSSs, putative TATA-boxes and putative CCAAT-boxes. Our approach was able to approximate the TSS location with a much higher precision than other approaches. In particular, 74.95% of the TSSs were identified with maximum error of up to 30 nucleotides, 58.03% with an error of up to 20 nucleotides, and 35.92% with a maximum error of only 10 nucleotides. This first modeling of the promoter region can help reduce false positives in the process of *ab initio* TFBS discovery. We plan in the near future to investigate more sophisticated models of promoter regions with other signals such as Y-Patch, DPE, MTE, INR, DCE and MDE.