Integration of protein-protein and protein-DNA interaction. Approach for protein complexes-on-DNA database

Sasha Belostotsky

Institute for Information Transmission Problems (Kharkevich Institute) of Russian Academy of Sciences (Mathematic methods and models in bioinformatics lab), Bolshoy Karetny per. 19, Moscow, 127994, Russia,

alexbel.system@gmail.com, belostotsky@iitp.ru

There are many approaches targeting examination of gene expression regulation. One of them is concentrated on single sites search, others are on tandem repeats, palindromes and clusters search. But as far as author knows there have not been any approaches for reconstruction of protein complexes from genomic data. Still protein complexes form everywhere in cell and play central role in processes regulation, among transcription is not the exception. Here such integrative approach is presented starting from genome and targeting protein complexes. Some problems arising from reverse task are also discussed.

Protein complexes binding DNA are reconstructed from ChIP-seq peak data for different genes and regulons. Efficacy of such approach is demonstrated on promoter-proximal ChIP-seq peaks that give relevant results on protein complexes if compared with experimentally known data about protein complexes. Histone modification enzymes such as histoneacetyltransferases, histonedeacethylases and chromatin remodeling machines are involved also, so multilevel system take place in this case. Approach can be algorithmically realized in a very simple way dealing with simple databases queries (in this case it was MySQL driven database) and standard protein-protein interaction database tools (in this case data from different protein-protein interaction database was integrated in local database). Such approach cannot be used now on binding sites of transcription factors because experimentally ones are few and predicted ones are everywhere in genome, even in DNAse1-hypersensitive regions that are thought to be regulatory ones and were used in this approach.

Protein complexes surely can be reconstructed from ChIP-seq data so it means that ChIP-seq data and protein-protein interaction data is abundant enough. Such protein complexes include different protein activities so it can be used for pretty wide analysis of genome regulation. Of cause such approach can be used on any proteins set. Perspectives of such approach are different but it seems that most challenging one is to obtain protein complexes for following protein-protein docking procedure to have well-resolved protein complexes on which many hypotheses can be tested.