Discovering similar (epi)genomics feature patterns in multiple genome browser tracks

Montanari P(1), Ceol A(2), Bartolini I(1), Ciaccia P(1), Patella M(1), Ceri S(3), Masseroli M(3)

(1) School of Engineering, DISI - Università di Bologna, Mura Anteo Zamboni 7, 40126 Bologna, Italy
(2) Center for Genomic Science of IIT@SEMM, Fondazione Istituto Italiano di Tecnologia (IIT), Via Adamello 16, 20139 Milan, Italy
(3) Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy

Motivation

Next Generation Sequencing (NGS), with the high amount of heterogeneous data that it is generating, is opening many interesting practical and theoretical computational problems. Genome browsers, e.g. UCSC Genome Browser (Kuhn et al., 2013) or Integrated Genome Browser (IGB) (Nicol et al., 2009), allow visual inspection and identification of interesting patterns on multiple genome browser tracks, i.e. of sets of (epi)genomic regions/peaks at given distances from each other in different tracks. For example, such patterns can describe gene expression regulatory DNA areas including heterogeneous (epi)genomic features (e.g. histone modification and/or different transcription factor binding regions). Yet, once such patterns are visually identified in a genome section, the search of their occurrences along the whole genome is a complex computational task that is currently not supported, despite their discovery along the whole genome is very important for the biological interpretation of NGS experimental results and comprehension of biomolecular phenomena.
We defined an optimized pattern-search algorithm able to find efficiently, within a large set of (epi)genomic data, genomic region sets which are similar to a given pattern. We implemented it within an IGB plugin, which allows intuitive user interaction in both the visual selection of an interesting pattern on the loaded IGB tracks, and the visualization of occurrences of similar patterns identified along the entire genome.

Methods

Pattern matching is a recurrent problem in data science, typically solved by a cost based approach, where lower cost implies high similarity. Although the Best-Matching Problem (BMP) is suspected to be NP-hard, we propose an alternative Root-element approach (R-BMP) and a Dynamic Programming algorithm (DP-BMP) that lowers the complexity to order of O(M*N^2), with M and N the number of elements in a pattern and target track to be compared. Given the properties of the genomic data to which the algorithm will be applied (strictly increasing sequences, M << N), it is possible to obtain the best match for each element of the pattern in the target track through a binary search. With the resulting Windowed DP-BMP algorithm, the complexity can drop down to O(N*log(N)), making it applicable also to (very) large problem instances.
We extended this model to introduce the aspects missing in the base version, but critical for its application to NGS data: interval regions, multiple, partial, and negative tracks, region attribute matching, and top-K distinct matching. Interval regions can either be reduced to their centroid, or analyzed with an asymmetric approach, which takes into

account the region length. Negative matching tracks are considered when no region should be present for this track in the area of a result. Such regions are removed from the search space before search start. Partial matching are tracks that can be reasonably missing in the results. The cost for not matching an element in that track is consequently reduced. Region attributes can also be used to alter the cost of matching the elements in a track. Finally, in order to facilitate the discovery of pattern matching and increase the diversity of results, we implemented a top-K version of the algorithm, which compares the results produced and keep the best K disjoint results.
The algorithm has been implemented in Java 8, and integrated as a plugin for IGB.

Results
We extended IGB with a plugin, which provides biologists with a tool to visually inspect the genome browser's tracks to identify and select a pattern of possible interest, and search other instances of this pattern in the same or different tracks. It is also possible to load the pattern from a file or from a selection of tracks, related for instance to histone marks identified by chromatin immunoprecipitation sequencing (ChIP-seq), for which a peak should or should not be present in targeted regions.
A new track is created for each search query: all regions that match the pattern are highlighted and can be easily browsed. Because a similarity search can be repeated on different groups of samples (for instance treated/non treated, or control/disease), it is possible to compare the resulting tracks and to identify differences in (epi)genomic features, suggesting mechanisms for the response to treatments or for the investigated pathology.
Several "chromatin states" identified by different combination of histone marks were inferred during the Roadmap Epigenomics project (Ernst et al., 2012). Such patterns can be submitted to our plugin to infer the regulation state of genomic regions under different conditions. Because the tool is not limited to the analysis of ChIP-seq peaks, but can be applied to any (epi)genomic regions, analyses can be extended by integrating other features, such as differentially expressed genes (DEG), DNase I hypersensitive sites (DHS), transcription start sites (TSS), or single nucleotide polymorphisms (SNP).

References:

- Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods 2012; 9(3), 215–216.
- Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief. Bioinform. 2013; 14(2): 144–161.
- Nicol JW, Helt GA, Blanchard SG Jr, Raja A, Loraine AE. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics 2009; 25(20), 2730–2731.