D-Light - A Sophisticated Tool for Exploration of cis-regulatory Elements

Zuzan C.J.[2], Laimer J.[1,2], Sophie A. Blank [1], Paul Neulinger [1], Alexander P. Seitinger [1], Alexandra M. Simader [1] und Peter Lackner [2]

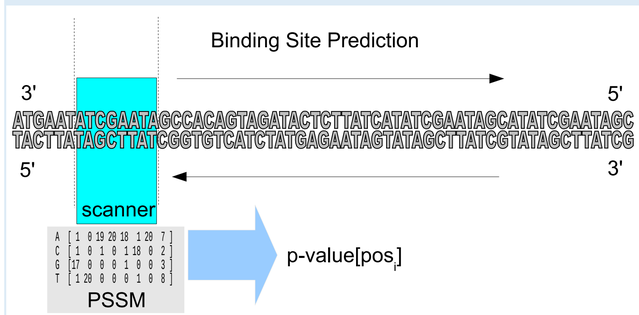
[1] Upper Austria University of Applied Sciences, Department of Biomedical Informatics [2] University of Salzburg, Faculty of Natural Sciences, Department of Molecular Biology

Introduction

Transcription factor (TF) proteins determine the time point and the amount of a gene being transcribed to RNA. From laboratory experiments short DNA fragments binding to TFs can be obtained, allowing for generation of positional frequency matrices (PFMs) and further prediction of transcription factor binding sites (TFBS). Analysis of regulatory effects is challenging as combinations of diverse TFBS and DNA accessibility have to be considered.

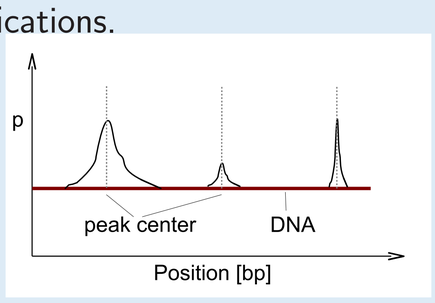
We developed *D-Light*, a new software offering storage and management of DNA-promoter sequences, annotation data and PFMs for multiple genomes. All data are stored in a relational database to provide complex queries and short response times. Users can add data such as promoter sequences, PFMs or custom annotation to the prefilled *D-Light* database. Users interact with *D-Light* via an intuitive, platform independent GUI.

Annotation Types

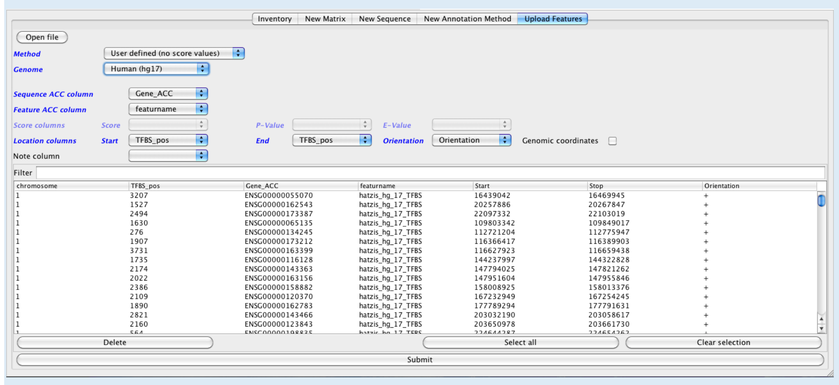


A PFM is evaluated against each position in both reading directions on all stored promoters. A p-value is computed to estimate the likelihood of a TFBS occurrence at a given position. Results above a minimum cutoff are stored in *D-Light* .

Other annotations comprise experimentally determined binding sites or DNA modifications.



Data Upload

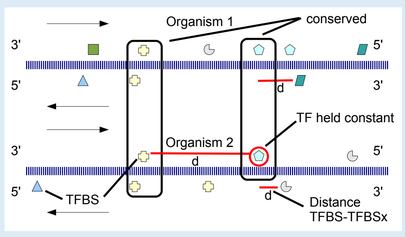


The database is prefilled with promoters and predicted TFBSs during installation of the *D-Light* server.

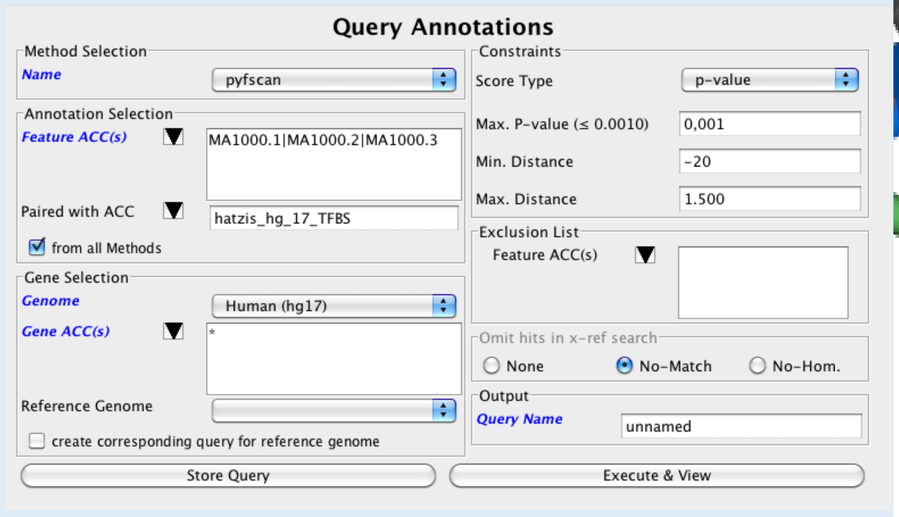
Promoters of different sizes have been prepared for 25 major genomes available at UCSC[1]. PFMs are taken from JASPAR[2]. Subsequently, users can add their own PFMs, promoters or annotations via the GUI.

Queries

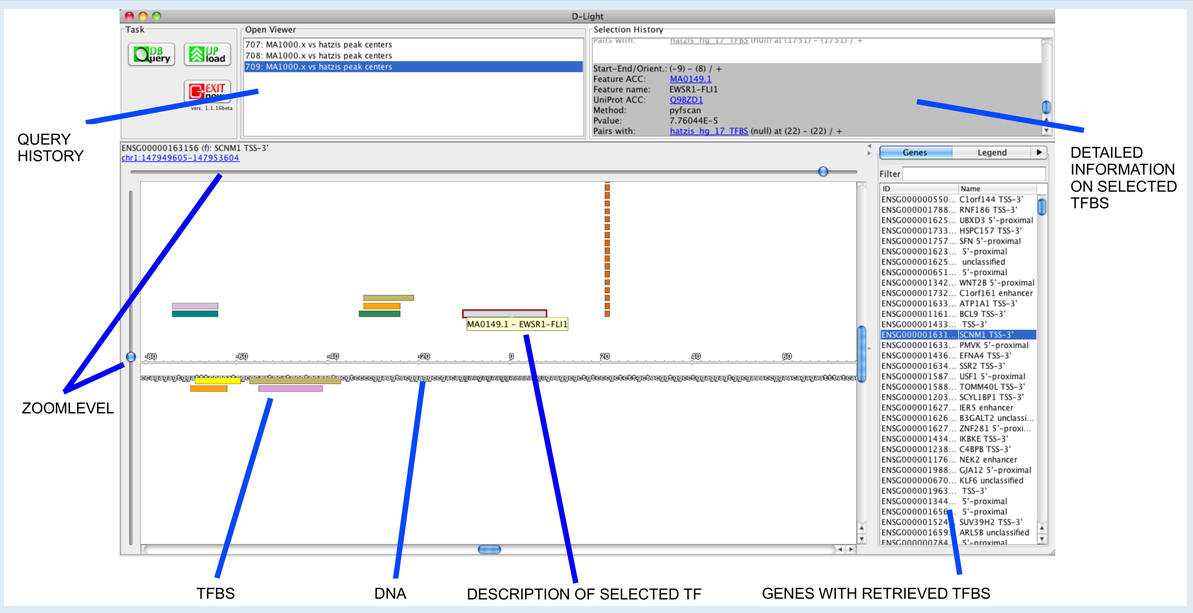
The most simple search granted by *D-Light* is to retrieve TFBSs occurrences of one TF. More complex queries allow searching of combinations of TFBSs within a certain sequential distance.



Homology restriction additionally requires two TFBSs to occur in the same place in two different organisms. This reduces false positive TFBSs and favors retrieval of housekeeping and development related TFBSs.



Result Visualization



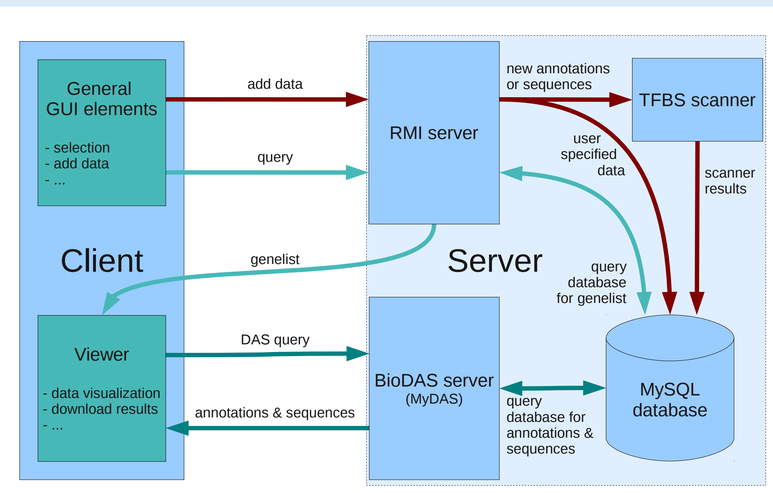
Case Study

In a case study we examined potential co-factors of transcription factor TCF4, which is involved in colon cancer. Centers of experimentally found[6] probability-peaks of TCF4 binding were uploaded and neighboring TFBSs queried in *D-Light* 's database. The searched area was spanned from the peak to 250 base pairs upstream and downstream. Background TF-counts were collected from the adjacent 250 base pairs on either side. Results were exported from *D-Light* and TFBS occurrences were counted and summed up for each TF. Foreground sums divided by background counts indicate the over-representation of a TF near the TCF4 peak-centers. The following over-representation sorted table thus lists possible co-factors for TCF4.

matrix	over-repr.	description
MA0164.1	2.05556	Nr2e3, retina development (rod differentiation, S-cone pathway inhibition; molecular "switch" function)
MA0133.1	1.95652	BRCA1, breast cancer
MA0043.1	1.75	HLF, hepatic leukemia factor
MA0117.1	1.69697	Mafb, rat lens development, "v-maf musculoaponeurotic fibrosarcoma"
MA0062.2	1.68605	GABPA , (=NRF2,NTF2); hypoxia/reoxygenation; implication to tumor biology (via prx1 gene)
MA0092.1	1.61111	Hand1::Tcfe2a, unknown function; transcript expressed in heart and neural crest
MA0073.1	1.57957	RREB1, signal transduction in thyroid, cancer and other cells

Technical Background

D-Light Is a client-server based system. The server runs on Linux, the client is implemented in Java and thus OS independent. One key feature is the use of BioDAS protocol used by most important bioinformatics centers. Another is the replaceable scanner part.



Queries from the graphical userinterface are directed to Java-RMI on the server side. From there calls to scanner and database are effectuated. In response the server sends back BioDAS formatted annotation and sequence data the client visualizes. *D-Light* is available freely under GPL licence.

References

[1] **The UCSC Genome Browser database: update 2010.** Rhead B *et al.* Nucleic Acids Res. 2010. http://genome.ucsc.edu/

[2] **JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles.** Portales-Casamar E *et al.* Nucleic Acids Res. 2010. http://jaspar.cgb.ki.se/

[3] **The distributed annotation system.** Dowell RD *et al.* BMC Bioinformatics 2001. http://www.biodas.org/wiki/Main_Page

[4] **MyDas** http://code.google.com/p/mydas/

[5] **Genoviz Software Development Kit: Java tool kit for building genomics visualization applications.** Helt GA *et al.* BMC Bioinformatics 2009.

[6] **Genome-Wide Pattern of TCF7L2/TCF4 Chromatin Occupancy in Colorectal Cancer Cells.** Hatzis *et al.* Mol Cell Biol. 2008.

http://biwww.che.sbg.ac.at/dlight/

clemensjohannes.zuzan@sbg.ac.at