BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 13 2014, pages 1923–1925
doi:10.1093/bioinformatics/btu115

Sequence analysis Advance Access publication February 26, 2014

LASAGNA-Search 2.0: integrated transcription factor binding site search and visualization in a browser

Chih Lee* and Chun-Hsi Huang*
Department of Computer Science and Engineering, University of Connecticut, Storrs, CT 06269, USA

Associate Editor: John Hancock

ABSTRACT

Summary: LASAGNA-Search 2.0 is an integrated webtool for transcription factor (TF) binding site search and visualization. The tool is based on the LASAGNA (Length-Aware Site Alignment Guided by Nucleotide Association) algorithm. It eliminates manual TF model collection and promoter sequence retrieval. Search results can be visualized locally or in the University of California Santa Cruz Genome Browser. Gene regulatory network inference based on the search results offers another way of visualization. A list of TFs and target genes is all a user needs to start using the tool. LASAGNA-Search 2.0 currently offers 1792 TF models and supports 15 species for automatic promoter retrieval and visualization in the University of California Santa Cruz Genome Browser. It is a user-friendly tool designed for non-bioinformaticians and is suitable for research and teaching. We describe important changes made since the initial release.

Availability and implementation: LASAGNA-Search 2.0 is freely available without registration at http://biogrid.engr.uconn.edu/lasagna_search/.

Contact: chihlee@engr.uconn.edu or huang@engr.uconn.edu

Received on January 2, 2014; revised on February 10, 2014; accepted on February 20, 2014

1 INTRODUCTION

A transcription factor (TF) is a protein or protein complex that regulates gene expression by physically binding to the regulatory regions of its target genes. Because of the similarity shared among binding sites of a TF, a model can be built from the known TF binding sites (TFBSs) to search for novel binding sites in unannotated DNA sequences. Various assays can be used to identify binding sites of a TF. One of these is the high-throughput ChIP-seq (Johnson *et al*., 2007) method, which combines chromatin immunoprecipitation (Gilmour and Lis, 1984) with massively parallel DNA sequencing. Briefly, DNA bound by the TF in question is sequenced, producing short DNA sequences or reads. By aligning these reads to a reference genome, we obtain the number of reads covering each genomic location or the binding signals, which are used by a peak finding algorithm to locate signal peaks in the genome. Projects such as the ENCyclopedia of DNA Elements (ENCODE) project (ENCODE Project Consortium, 2012) generated huge amounts of ChIP-seq data for human and mouse TFs across different cell types. This helps scientists better understand TF binding for the studied species. ChIP-seq data, however, are not available for lesser-studied TFs and species. As a result, TFBS search tools are still routinely used by scientists (Del Campo *et al*., 2014; Martin, 2013) and databases curating TFBSs and TF models (Kiliç *et al*., 2014; Mathelier *et al*., 2013) are released or updated.

LASAGNA-Search 2.0 is an integrated webtool that allows users to perform TFBS search and visualization without leaving the website. It offers collections of 1792 TF models built from either TFBSs or position-specific weight matrices (PWMs), where the TFBSs and PWMs were collected from TRANSFAC (Matys *et al*., 2006), JASPAR (Mathelier *et al*., 2013), UniPROBE (Newburger and Bulyk, 2009), ORegAnno (Griffith *et al*., 2008) and PAZAR (Portales-Casamar *et al*., 2009). To use custom TF models, the tool accepts unaligned variable-length binding sites of a TF as well as TF models in the form of PWMs. In case TFBSs are provided, the LASAGNA algorithm (Lee and Huang, 2013a) is used to align variable-length TFBSs before building a TF model. While users can provide DNA sequences in the FASTA format to be scanned for putative binding sites, LASAGNA-Search supports automatic promoter sequence retrieval for 15 species. Users can use official gene symbols, RefSeq mRNA accession numbers or NCBI Gene IDs to search for genes and retrieve promoter sequences relative to the transcription start sites. Promoter sequences retrieved via LASAGNA-Search come with genomic locations, which allow the search results to be visualized with other annotations with ease. Users can either visualize hits locally or in the University of California Santa Cruz (UCSC) Genome Browser with a mouse click. Finally, significant binding sites of TF A found in the promoter of gene B suggest interaction between TF A and gene B. Enabled by our tool, search results can hence be visualized as a network of TFs, TF coding genes and target genes.

Since the initial release of LASAGNA-Search (Lee and Huang, 2013b), we received constructive and encouraging feedback from the users, which resulted in LASAGNA-Search 2.0. In this note, we describe the improvements, discuss the implications and outline future work.

2 NEW FEATURES

2.1 New TF models based on TFBSs in the PAZAR database

The PAZAR database (Portales-Casamar *et al*., 2009) offers a platform for users to start curation projects. A record stores one

*To whom correspondence should be addressed.

annotation for one sequence from either a TF–gene interaction or gene expression experiment. Hence, binding sites of a TF can be extracted from TF–gene interaction records in the PAZAR projects. As more than one project may curate binding sites of a particular TF, we aggregated records containing TF–gene interaction information from all the public projects. All the files in the general feature format dated 20120117 were downloaded. We group TFBSs by TF and species, that is, human TF A and mouse TF A are considered two TFs.

A binding site was filtered out if it is <4 or >1000 bases long. To verify a binding site, we searched for it in the vicinity of the curated genomic location in the reference genome. The binding site was discarded if it could not be located within five bases of the curated location. As we collected binding sites for a TF across all the public projects in PAZAR, a TFBS may be curated by more than one project, resulting in multiple copies of the TFBS in our collection. Therefore, for each pair of overlapping binding sites, we kept only the shorter one if the overlap is >80% of the shorter one in length. A model was built for a TF if it has at least 10 binding sites.

The LASAGNA-ChIP algorithm (Lee and Huang, 2013a) was used to align the binding sites of a TF because some of the projects contain TFBSs identified by ChIP-seq and ChIP-chip experiments. As reported by Johnson *et al.* (2007), about 94% of the actual binding sites can be located within 50 bases of signal peaks. However, no clipping was done for sequences produced by ChIP-seq experiments because information about the signal peak is not available in PAZAR. The new collection contains 66 TF models, 39, 20 and 7 of which are human, mouse and rat, respectively.

2.2 Support for more species

The initial release of LASAGNA-Search 2.0 supports seven species for automatic promoter sequence retrieval and visualization in the UCSC genome browser. This is an important feature because the genomic locations of sequences retrieved by LASAGNA-Search are available. The genomic locations of user-supplied sequences in the FASTA format, however, are unknown. Consequently, putative binding sites found in user-supplied sequences can only be visualized locally at LASAGNA-Search not in the UCSC genome browser. To make LASAGNA-Search more useful, we have added support for eight additional species since the initial release. The newly added species are *Bos taurus*, *Sus scrofa*, *Ovis aries*, *Gallus gallus*, *Canis lupus familiaris*, *Felis catus*, *Xenopus (Silurana) tropicalis* and *Danio rerio*. More importantly, we have automated adding support for a new species. As long as information about transcription start sites is available in the UCSC genome browser database, a new species can be added with ease. Hence, we encourage users to request for new species to meet their research or teaching need.

2.3 Hardware and user interface

LASAGNA-Search was initially deployed on a shared web server. To better serve the users, LASAGNA-Search 2.0 is now powered by a cluster with a load balancer to direct incoming traffic, making it more scalable. This enabled us to add more options to retrieve longer promoter sequences, which were capped at 1000 bps because of limited computation capacity. To make TF model selection easier, the model information page with a sequence logo does not need to be opened in a new tab. It can now be displayed when mouse pointer hovers over. To visualize hits locally, LASAGNA-Search displays only the region of promoter from the first hit to the last hit. An option has been added to display the entire promoter, and hence, the saved images are aligned by transcription start sites.

3 CONCLUSION AND FUTURE DIRECTION

LASAGNA-Search 2.0 was designed to be a user-friendly and an easy-to-use webtool for TFBS search and visualization. Users do not need to be bioinformaticians and can start using the tool immediately. Besides research, LASAGNA-Search 2.0 is well-suited for teaching and class projects at the undergraduate level, where students have limited experience in programming. To keep the tool useful, we plan to include support for species on users' request and keep the TF model collections up-to-date by incorporating new database releases such as JASPAR 2014 (Mathelier *et al.*, 2013). To easily use ChIP-seq data, we plan to interface with a genomic cloud computing platform such as the Illumina BaseSpace. This way users can easily use motifs discovered in ChIP-seq experiments of one species to scan sequences of another species. A genomic cloud computing platform allows users to perform computation-intensive analyses in a browser, which is in line with LASAGNA-Search.

ACKNOWLEDGEMENTS

The authors are indebted to users of LASAGNA-Search for their feedback, which greatly improved LASAGNA-Search.

Funding: National Science Foundation (grant numbers CCF-0755373, OCI-1156837) (in part).

Conflict of Interest: none declared.

REFERENCES

Del Campo,E.P. *et al.* (2014) CTCF and CTCFL mRNA expression in 17β-estradiol-treated MCF7 cells. *Biomed. Rep.*, **2**, 101–104.

ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Gilmour,D.S. and Lis,J.T. (1984) Detecting protein-DNA interactions *in vivo*: distribution of RNA polymerase on specific bacterial genes. *Proc. Natl Acad. Sci. USA*, **81**, 4275–4279.

Griffith,O.L. *et al.* (2008) ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.*, **36**, D107–D113.

Johnson,D.S. *et al.* (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

Kiliç,S. *et al.* (2014) CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. *Nucleic Acids Res.*, **42**, D156–D160.

Lee,C. and Huang,C.-H. (2013a) LASAGNA: a novel algorithm for transcription factor binding site alignment. *BMC Bioinformatics*, **14**, 108.

Lee,C. and Huang,C.-H. (2013b) LASAGNA-Search: an integrated webtool for transcription factor binding site search and visualization. *BioTechniques*, **54**, 141–153.

Martin,L.J. (2013) Implications of adiponectin in linking metabolism to testicular function. *Endocrine*, 1–13.

Mathelier,A. *et al.* (2013) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D147.

Matys,V. *et al.* (2006) TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.

Newburger,D.E. and Bulyk,M.L. (2009) UniPROBE: an online database of protein binding microarray data on proteinDNA interactions. *Nucleic Acids Res.*, **37**, D77–D82.

Portales-Casamar,E. *et al.* (2009) The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Res.*, **37**, D54–D60.