DE NOVO TRANSCRIPTION FACTOR BINDING SITE DISCOVERY:

DE NOVO TRANSCRIPTION FACTOR BINDING SITE DISCOVERY:

Sherschel, James

Name: scherschel.pdf
Size: 1.195Mb
Format: PDF

View/Open

Permanent Link: http://hdl.handle.net/1805/1967
Date: 2009-10-29

Abstract:

Abstract Computational methods have been widely applied to the problem of predicting regulatory elements. Many tools have been proposed. Each has taken a different approach and has been based on different underlying sets of assumptions, frequently similar to those of other tools. To date, the accuracy of each individual tool has been relatively poor. Noting that different tools often report different results, common practice is to analyze a given set of regulatory regions using more than one tool and to manually compare the results. Recently, ensemble approaches have been proposed that automate the execution of a set of tools and aggregate the results. This has been seen to provide some improvement but is still handled in an ad hoc manner since tool outputs are often in dissimilar formats. Another approach to improve accuracy has been to investigate the objective functions currently in use and identify additional informational statistics to incorporate into them. As a result of this investigation, one statistical measure of positional specificity has been demonstrated to be informative. In this context, this thesis explores the application of three simple models for the positional distribution of transcription factor binding sites (TFBS) to the problem of TFBS discovery. As alternate measures of positional specificity, log-likelihood ratios for the three models are calculated and treated as features to classify TFBSs as biologically relevant or irrelevant. As a verification step, randomly generated positional distributions are analyzed to demonstrate the robustness and accuracy of the log-likelihood ratios at classifying data from known distributions using a simple classifier. To improve classification accuracy, a support vector machine (SVM) approach is used. Subsequently, randomly generated sequences seeded with TFBSs at positions chosen to conform to one of the three models are analyzed as an additional verification step. Finally, two types of sets of real regulatory region sequences are analyzed. First, results consistent with the literature are obtained in three cases for genes experimentally determined to be co-expressed during mouse thymocyte maturation,

and a novel role is predicted for three families of TFBSs in single positive (SP) T-cells. Second, the mouse and human ¯real|| sets from Tompa et al' s ¯Assessment of Computational Motif Discovery Tools|| are analyzed, and the results are reported.

This item appears in the following Collection(s)

- Informatics Graduate Theses and PhD Dissertations (/handle/1805/303)
- Informatics School Theses and Dissertations (/handle/1805/954)

 Show Statistical Information (#)

My Account

- Login
- Register

Statistics

- Most Popular Items
- Statistics by Country
- Most Popular Authors