

INVESTIGATING CONSISTENCY BETWEEN CURATED BINDING PROFILES AND PWMs DERIVED FROM PROTEIN-DNA STRUCTURE MODELS

Tsung-Ting Hsieh, Mei-Ju May Chen, and Chien-Yu Chen*

*Dept. of Bio-Industrial Mechatronics Engineering, National Taiwan University,
Taipei, 106, Taiwan*

**Email: cychen@mars.csie.ntu.edu.tw*

Prediction of transcription factor binding sites (TFBSs) is an important issue in study of regulatory networks. There are about 500 experimentally determined or predicted binding profiles curated in the database TRANSFAC for human transcription factors (TFs). This study employs two scoring functions to investigate consistency between curated profiles and position weight matrices (PWMs) derived from structure models. The first scoring function is named 'contact model', which was previously demonstrated to be able to generate binding profiles similar to those determined by experiments based on existing protein-DNA complexes. In this study, we collected 178 protein-DNA complexes for 29 human TFs, each of which has at least one annotated matrix in TRANSFAC. The PWMs constructed by the contact model are compared with the corresponding annotated matrices. It is surprising that more than 30% of the 45 profiles derived by the structure models are not consistent with the curated PWMs (similarity smaller than 0.7). Similar situation was observed on results of another scoring function, an all-atom knowledge-based energy function developed for predicting protein-DNA interactions. After examining both the consistent and the inconsistent positions on the dissimilar cases, it is observed that some profile pairs are similar except for few positions that should be inserted or deleted in one of the two PWMs to improve the alignment. We conclude that it deserves more advanced studies to examine the annotated matrices by incorporating sophisticated energy functions with structure models of protein-DNA complexes to further improve the prediction accuracy of TFBS discovery.

* Corresponding author.