**A MODULAR METAPROFILE-BASED SYSTEM FOR PREDICTION AND ANALYSIS OF EUKARYOTIC PROMOTERS.**

JUNIER THOMAS, KROGH ANDERS, BUCHER PHILIPP

Swiss Institute of Bioinformatics & Swiss Institute for Experimental Cancer Research

Ch. des boveresses 155 CH-1066 Epalinges, Switzerland

Phone-#:(+41 21) 692 5892 FAX-#:(+41 21) 652 6933

e-mail: pbucher@cmpteam4.unil.ch

Keywords: eukaryotic promoters, transcription control elements, CpG islands, hidden Markov models, metaprofiles

We have developed a modular system based on a new concept called metaprofile for developing promoter prediction algorithms, and for evaluating the diagnostic power of promoter-associated sequence features such as transcription factor binding sites and CpG islands. The system consists of training data, a hierarchically organised set of component methods, and procedures for integrating component methods. The current implementation uses TATA-box, GC-box, CCAAT-box, and CpG island predictors as component methods.

The systems can be operated in three different modes: training, prediction, and evaluation mode. The evaluation mode is a combination of training and prediction mode connected by a cross-validation scheme.

The training data consist of positive and negative datasets. The positive set contains sequences with labeled transcription start sites. In the application presented this set is derived from the EPD database. The negative data set simply consists of unlabeled sequences, human genome sequences in our example. This set is not necessarily devoid of promoters. The performance of a prediction algorithms is quantitatively expressed by receiver-operating characteristics (ROC).

Component methods exist in three states: initial state, trained state, and null state. The initial state is supposed to reflect common knowledge which cannot and need not to be subjected to cross-validation. The trained state is a refined version of the initial state obtained after training of certain parameters with the data. The training method needs to be defined in advance and is considered an integral part of a prediction module. The null state is a similarly structured neutral module that can be plugged in into the final prediction algorithm instead of the trained version. Such a neutral module is not supposed to have any effect on prediction performance.

The integration of component methods into a synthetic promoter prediction machine follows a hierarchical tree defining the unidirectional dependences between the modules. Each module typically exploits one particular sequence feature associated with a molecular function, in our case promoter function. A sequence function-prediction system based on this principle is called a "metaprofile". The integration of component methods may be achieved by many different mechanisms. Here, the promoter element predictors are simple hidden Markov models (HMMs) which can be incorporated into a more complex model performing signal-based promoter prediction. By contrast, the integration of signal-based promoter prediction with CpG islands analysis works by combining the output of independently operating stand-alone programs. The metaprofile itself is implemented as a collection of perl scripts and Makefiles gluing together the individual software components.

The experimental metaprofile for eukaryotic promoters has been developed with two objectives in mind: (i) maintenance of an evolvable promoter prediction machine which can be improved incrementally by adding new training data and component methods. (ii) objective and quantitative evaluation of the diagnostic value of sequence properties reported to be associated with promoter function. The later application is conceptually more original and has the potential of leading to new insights on transcriptional cotrol mechanisms. Computer experiments consisting of knocking out individual component methods of a synthetic promoter prediction algorithm and measuring the resulting effects on performance constitute a new approach to investigate synergisms and antagonisms between transcriptional control elements.