Large-scale gene regulatory motif discovery and categorisation with NestedMICA

Matias Piipari [1], Thomas Down [2], Tim Hubbard [1]

Background: Our understanding of DNA specificities of transcription factors is largely recorded in databases containing DNA sequence motifs (position frequence or weight matrices). Previously it has been shown that there are familial tendencies in these DNA sequence motifs that are predictive of the family of factors that binds them and indeed motifs belonging to different transcription factor families have been studied using unsupervised and supervised machine learning methods in an attempt to predict the binding domain for sequence motifs and to sensitively find motifs from novel sequence sets that fit these tendencies. However, a natural probabilistic model for recurring patterns in a set of sequence motifs is still an open research problem. Sequence motif discovery algorithms also often present the problem of reporting duplicate or closely related motifs multiple times. There is a clear need for a mathematical framework and tools for computational biologists used for profiling relatedness of gene regulatory sequence motifs.

Results: We propose a generative model for nucleotide sequence weight matrices termed the 'metamotif'. This model can be used to summarise recurring patterns in a set of weight matrices. A nested sampling based algorithm for parameter estimation of metamotifs from a set of motifs, as well as two practical uses for the model will be discussed: a motif classification task, as well as use as a weight matrix prior in a Bayesian model discovery algorithm.

Conclusions: The metamotif model is successfully applied to a weight matrix classification problem where sequence motif features are used to predict the type of a sequence motif on the level of its TRANSFAC family and superfamily. We also show that metamotifs can be applied as informative priors in a motif discovery algorithm to dramatically increase the sensitivity to discover motifs. Both the transcription factor type prediction tool and the informative prior are also made available for the use use of computational biologists through a web server and a new release of the NestedMICA motif discovery tool.

Project URL: `http://www.sanger.ac.uk/Software/analysis/nmica/`
SVN repository: `http://www.derkholm.net/svn/repos/nmica/`
License: LGPL

Affiliations:

1) Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire (UK)
2) Wellcome Trust/Cancer Research UK Gurdon Institute, Cambridge, Cambridgeshire (UK)