InMoDe: Tools for Learning and Visualizing
Intra-Motif Dependencies of DNA Binding Sites

Ralf Eggeling
Ivo Grosse
Jan Grau

Department of Computer Science, University of Helsinki, Finland
Institute of Computer Science, MLU Halle–Wittenberg & iDiv, Germany
Institute of Computer Science, MLU Halle–Wittenberg, Germany

Introduction

The position weight matrix (PWM) model and its visualization by sequence logos used to be the state of the art for motif representation in DNA sequence analysis.

However, recent studies suggest that taking into account dependencies between binding site positions often yields a better motif representation. Here, we present the Java application InMoDe [1], a collection of tools for learning, leveraging and visualizing **in**tra-**mo**tif **de**pendencies of putative higher order. Availability:

`www.jstacs.de/index.php/InMoDe`

Methodology

Binding sites fully observable?	Probabilistic Model	Learning Approach
Yes	• allows dependencies of maximal order d among adjacent motif positions • sparse parameterization via parsimonious context tree (PCT) at each position [2]	• structure learning: selecting each PCT so that the BIC score of the model is maximized • huge PCT search space, but finding globally optimal PCTs tractable up to depth six [3]
No	• PCT-based model from above wrapped in OOPS model • latent variabels for (i) binding site position, (ii) strand orientation, (iii) motif type	• finding globally optimal model is intractable • stochastic EM algorithm [2]; M step performs structure learning as above

Conditional sequence logo
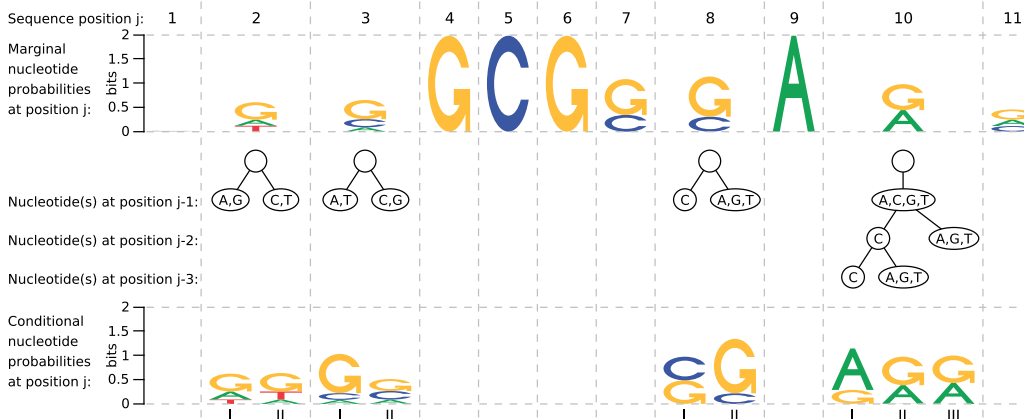
- direct visualization of learned PCTs and conditional probabilities
- contains traditional sequence logo (top)
- learned with from E2F1 binding sites (JASPAR) with SimpleMoDe (d=3)
- plot further refined with VisualizationApp (color scheme, labels)

InMoDe tools in brief

Learning

- **SimpleMoDe**
  learning a model from aligned sequences

- **DeNovoMoDe**
  motif discovery from, e.g., ChIP-seq data

- **MixtureMoDe**
  learning multiple models, clustering and structure learning simultaneously

- **FlexibleMoDe**
  flexible choice of latent variables, allows weighted input

Application

- **ScanApp**
  finding significant motif occurrences in given sequences

- **ClassificationApp**
  classifying sequences according to different motif models and/or background models

- **VisualizationApp**
  plotting of customized conditional sequence logos

User interfaces
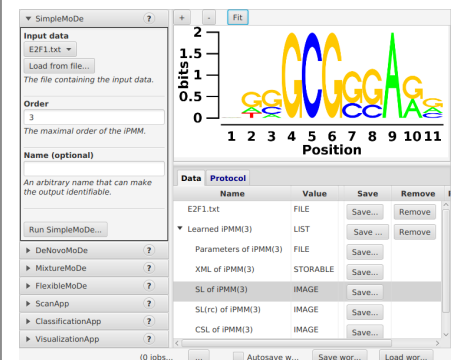
- GUI
- Command line interface
- Galaxy integration

[1] R. Eggeling, I. Grosse, J. Grau. *InMoDe: tools for learning and visualizing intra-motif dependencies of DNA binding sites.* Bioinformatics (2017)

[2] R. Eggeling, T. Roos, P. Myllymäki, I. Grosse. *Inferring intra-motif dependencies of DNA binding sites from ChIP-seq data.* BMC Bioinformatics (2015)

[3] R. Eggeling, M. Koivisto. *Pruning rules for learning parsimonious context trees.* Proceedings of UAI (2016)