Clustering of incomplete, high dimensional and dependent biological data with SpaCEM^3

Matthieu VIGNES[1], Juliette BLANCHET[2], Damien LEROUX[1], and Florence FORBES[3]

[1] BIA Unit, INRA Toulouse, chemin de Borderouge, BP 52627, 31326 Castanet-Tolosan Cedex, France
[2] SLF, Fluelastrasse, 11, 7260 Davos Dorf, Switzerland
[3] Mistis Project, INRIA Rhône-Alpes, 655, av. de l'Europe, Montbonnot, 38334 Saint Ismier Cedex, France
SpaCEM3-help@lists.gforge.inria.fr

Abstract: *The SpaCEM^3 software makes the most of variational estimates in Markov Random Field (MRF) models to cluster (i) high-dimensional (ii) dependent and (iii) incomplete data. This methodology has been applied to draw meaningful modules of genes from high-throughput data. To download the latest version, visit http://spacem3.gforge.inria.fr/.*

Keywords: Markov Random Fields, variational methods, biological data integration.

1 Introduction

To our knowledge, few of the clustering algorithms available to biologists for molecular biology data analysis both model observations as measures carried out on individuals and integrate interaction data. Either individuals are assumed independent or observations are transformed into a pairwise metrics, the choice of which is often critical. We developed an efficient statistical toolkit based on integrated Markovian models and EM algorithms for parameter inference to deal with features of high-troughput biological data. Main advantages of our approach are 1) the integrated handling of missing observations, 2) modelling of network information if available and 3) appropriate treatment of high-dimensional noisy data. The SpaCEM^3 software (Spatial Clustering with EM and Markov Models), which implements our approach, is now in a mature version; the GUI makes it possible for biologists to use it on their own, still relying on powerful recent and still active developments (*e.g.* to deal with additional features of biological data like spurious interactions in databases) of algorithms devoted to the inference of probabilistic graphical models. We present it here along with key modelling aspects and a biological application.

2 Approach

For clarity purpose, we restrict our presentation to the case of transcript levels and interaction data. The latter is retrieved from databases and allows us to build a graph where nodes represent genes and edges stem from direct interaction: confirmed by experts so that edge weights can be fixed or putative where weights need to be estimated. The analysis is then recast into a biological object clustering framework. A *Hidden MRF* (HMRF) is used to model individual measures (with probabilistic distributions) and graph interactions (the neighbourhood graph structure is Gibbsian, see [3]). The main originality of SpaCEM^3 is that model estimation is based on an EM algorithm variational approximation ([1]) in a mean field-like setting. In this context, two specific features of the models available in the software are: (i) a modelling (*e.g.* Gaussian) of class-dependent distributions built for

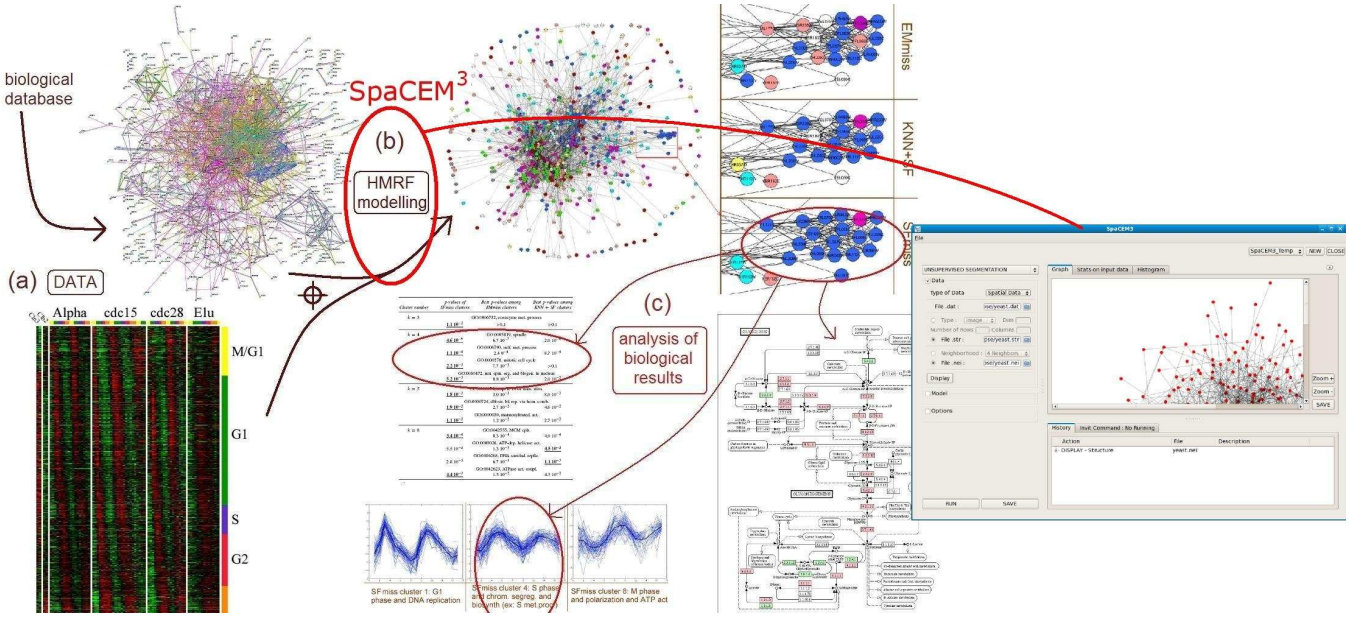

Figure 1. Graphical summary of an analysis workflow: (a) data extraction from relevant databases, (b) Specification of the HMRF settings in SpaCEM^3, model inference and results visualization. (c) Downstream biological analysis for biological cluster relevance: modularity, over-represented GO terms, expression levels profiles and link to pathways.

high-dimensional data and (ii) an integrated treatment of data with missing observations in a HMRF context [4]. This tackles the *missing value* issue in microarrays in a probabilistic framework and still enables *a posteriori* inference of incomplete observations without imposing any pre-processing of the data.

As it can be useful for comparison, other standard algorithms are also available in SpaCEM^3 along with classical imputation techniques for missing data. Model selection can be performed using criteria such as BIC or ICL approximated in the Markovian case ([2]). Lastly, SpaCEM^3 allows the user to simulate the different models presented above.

3 Recovering biological knowledge from biological data with SpaCEM^3

Figure 1 shows a typical biological data analysis sequence with SpaCEM^3: data retrieval which are specifically modelled in a HMRF context by the user (model choice, parameter estimation, data and results visualization) and investigation of different biological features of the obtained clustering.

References

[1] G. Celeux, F. Forbes and N. Peyrard, EM procedures using mean field-like approximations for Markov model-based image segmentation, *Pat. Rec.*, 36:131–144, 2003.

[2] F. Forbes and N. Peyrard, Hidden Markov random field model selection criteria based on mean field-like approximations, *IEEE PAMI*, 25:1089–1101, 2003.

[3] C.M. Bishop, *Pattern Recognition And Machine Learning*, Springer, 2008.

[4] J. Blanchet and M. Vignes, A model-based approach to gene clustering with missing observations reconstruction in a Markov Random Field framework, *J. Comput. Biol.*, 16:475–86, 2009.