Bioinformatics Advance Access published February 3, 2011

SpaCEM^3, a software for biological module detection when data is incomplete, high dimensional and dependent.

Matthieu Vignes [1], Juliette Blanchet [2] Damien Leroux[1] and Florence Forbes [3]

[1]INRA Toulouse - BIA Unit, Castanet Tolosan, France
[2]EPFL (SB/IMA/STAT), Lausanne, Switzerland
[3]INRIA Grenoble Rhône-Alpes - Team Mistis, Montbonnot, France

Associate Editor: Dr. Trey Ideker

ABSTRACT

Summary: Among classical methods for module detection, SpaCEM^3 provides ad hoc algorithms that were shown to be particularly well adapted to specific features of biological data: high-dimensionality, interaction between components (genes) and integrated treatment of missingness in observations. The software, currently in its version 2.0, is developed in C++ and can be used either *via* command line or with the GUI under Linux and Windows environments.

Availability: The SpaCEM^3 software, a documentation and datasets are available from `http://spacem3.gforge.inria.fr/`.

Contact: SpaCEM3-help@lists.gforge.inria.fr

1 INTRODUCTION

Few of the module detection algorithms made available to scientists for molecular biology data analysis (see Kim et al., 2005 and references therein for a representative list) directly model observations as measures carried out on individuals and integrate interaction data. Generally, either the latter is omitted *i.e.* individuals are considered as independent or individual observations are transformed into a pairwise metrics, the choice of which is intricate and can have a terrible impact on downstream analysis.

The SpaCEM^3 software (Spatial Clustering with EM and Markov Models) provides efficient statistical tools to deal with high-troughput biological data such as gene expression data. Its main advantages are 1) the possibility to handle missing observations and 2) the possibility to integrate available interaction network information. In a gene expression context, such interactions either come from prior knowledge or from measures like two-hybrid experiments. The integrated Markovian approach, which is at the heart of the software, is presented in Section 2. An example of application follows in Section 3.

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. Such interactions range from confirmed by expert to simply putative (Fig. 1 (a)); fixed weights can hence be assigned to edges or need to be estimated. The analysis is then recast into a biological object clustering framework. A *Hidden Markov Random Field* (HMRF) is used to model individual measures and graph interactions. The main originality of SpaCEM^3 is that model estimation is based on a variational approximation described in Celeux et al. (2003) for an Expectation-Maximization (EM) algorithm in a mean-field like setting. In this context, two new features of the models available in the software are:

- a modelling (*e.g.* Gaussian) of class-dependent distributions specifically built for high-dimensional data (Bouveyron et al., 2007). It has been adapted to the Markovian setting and used on biological data in Vignes and Forbes, 2009,
- an integrated treatment of data with missing observations in a HMRF context (Blanchet and Vignes, 2009). 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. We chose to present this feature in Section 3 for an illustrative use of SpaCEM^3 in the context of Molecular Biology.

Furthermore, the software provides extensions of the standard HMRF model such as *Triplet Markov models* (Blanchet and Forbes, 2008) that allow objects to be assigned to subclasses (possibly common to different clusters). It introduces an additional blanket that could, for example, encode genetic dependencies. Applications are at present limited to supervised classification (so a training set is needed). As it can be useful for comparison, other standard algorithms are also available in SpaCEM^3: k-means, Iterated Conditional Modes (ICM), standard EM and variants. The software also includes classical imputation techniques for missing data: by zeros, (un)conditional line/column mean/median, last observation carried forward, KNN-imputation. In addition, model selection can be performed using criteria such as the *Bayes Information Criterion* (BIC, Schwarz, 1978) or *Integrated Complete Likelihood* (ICL, Biernacki et al., 2000) approximated in the Markovian case (Forbes and Peyrard, 2003). Lastly, SpaCEM^3 allows the user to simulate the different models presented above.

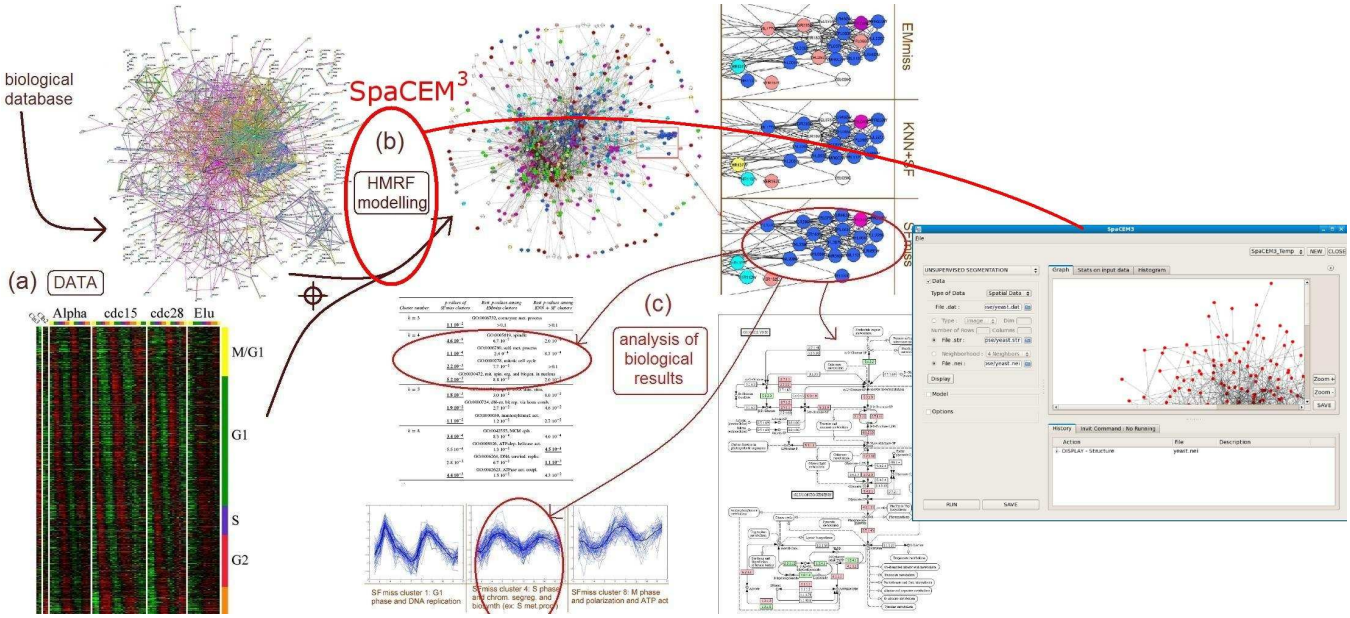

Fig. 1. Graphical summary of the data analysis workflow of Blanchet and Vignes (2009): (a) data from relevant databases are extracted. (b) The SpaCEM^3 software allows the user to specify the HMRF settings, to solve the model and scan the results in the GUI. (c) Downstream biological analysis for biological module relevance: modularity of the network, over-represented GO terms, expression levels profiles and link to pathways.

3 RECOVERING BIOLOGICAL KNOWLEDGE FROM BIOLOGICAL DATA WITH SPACEM^3

Figure 1 shows a typical biological data analysis sequence with SpaCEM^3. First, data are retrieved from relevant databases to set a network topology or neighbourhood between biological components (here from `http://string.embl.de/`) and individual measures carried out on these components (yeast cell cycle DNA microarray images; white spots represent missing observations in Fig. 1 (a)). A HMRF model integrates individual measures and graph interactions in the SpaCEM^3 software. It allows the user to specify the model, estimate parameters and visualize the results in the GUI. In the absence of gold standard to assess the accuracy of our results, we investigated different biological features of the obtained modules: network modularity, associated Gene Ontology (GO) terms, gene expression profiles and connection to metabolic pathways (from KEGG `http://www.genome.jp/kegg/`).

4 CONCLUSION

SpaCEM^3 provides a stand-alone analysis tool to retrieve meaningful moduless of biological objects. It relies on powerful recent developments on algorithms devoted to the inference of probabilistic graphical models so that complex individual and interaction biological data can be modelled together as shown in Vignes and Forbes, 2009; Blanchet and Vignes, 2009. The GUI makes it easy for biologists to use. Further developments of the software will follow theoretical work under progress to deal with additional features of biological data: possibly spurious interactions in databases and application to genetical genomics to reconstruct biological networks.

ACKNOWLEDGEMENT

The authors would like to thank Sophie Chopart for her work on the software.

REFERENCES

Biernacki,C. et al. (2000) Assessing a mixture model for clustering with the integrated complete likelihood, *IEEE Trans PAMI*, 22, 719-725.

Blanchet,J. and Forbes,F. (2008) Triplet Markov fields for the supervised classification of complex structure data, *IEEE Trans PAMI*, 30, 1055-67.

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

Bouveyron,C. et al. (2007) High dimensional data clustering, *Comput. Statist. Data Analysis*, 52, 502–519.

Celeux,G. et al. (2003) EM procedures using mean field-like approximations for Markov model-based image segmentation, *Pat. Rec.*, 36, 131-144.

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

Kim,D.-W. et al. (2005) Detecting clusters of different geometrical shapes in microarray gene expression data, *Bioinformatics*, 21, 1927–1934.

Schwarz,G. (1978) Estimating the dimension of a model, *Ann. Stat.*, 6, 461-464.

Vignes,M. and Forbes,F. (2009) Gene clustering via integrated Markov models combining individual and pairwise features, *IEEE/ACM TCBB*, 6, 260-270.