Multivariate time-series segmentation for immunological data analysis

Alberto Castellini and Giuditta Franco

Verona University, Department of Computer Science,
Strada Le Grazie 15, 37134 Verona, Italy,
{alberto.castellini,giuditta.franco}@univr.it

DataMod 2017 – 6th International Symposium *From Data to Models and Back*

Introduction. Developing new methodologies for (biomedical, clinical, immunological, molecular) data analysis is a main research interest in both bioinformatics and computational biology. In particular, efficient and fast computational methods are proposed in the literature of machine learning (for data clustering, and feature extraction) to infer new knowledge from given data. A vivid interest is particularly devoted to immunological research, as most of the human deseases are induced by some fall or misplay in our body defence system. More specifically, a recent broad interest is focused on the lifetime aging of immune system, in terms of changes of immune mechanisms of an individual during his/her infancy, growing age, mature age, and senescence. Indeed, an age-related decline, referred to as immunosenescence, seems characterized by a decrease in cell-mediated immune functions, where defects in T- and B-cell functions coexist, and has social/commercial impact related to vaccination in elderly persons [14].

Here we present a couple of different models, published in [2] and in [4], respectively based on MP systems and on piecewise linear segmentation. Moreover, we discuss our current work on an immunological dataset analysis, we are approaching by a combination of time-series clustering and segmentation methodologies

Models for an immunological dataset. The immune network theory, formulated by Jerne [8] and subsequently developed by Parelson [12], attempted to use mathematical formalisms (such as differential equations) to describe the dynamics of lymphocyte interactions from a quantitative and systemic point of view. Here we follow this direction, and propose possible data-driven network models to describe relationships among cell quantities of specific eight peripheral B lymphocyte subpopulations. In collaboration with Antonio Vella from the Polyclinic Hospital of the University of Verona, we were given a dataset with the measured amounts of cells exhibiting the combinations of receptor clusters *CD27*, *CD23* and *CD5* in almost six thousands patients. Such a matrix may be seen either as a set of cross-sectional data or as eight time-series, if the age of single patients is taken into account.

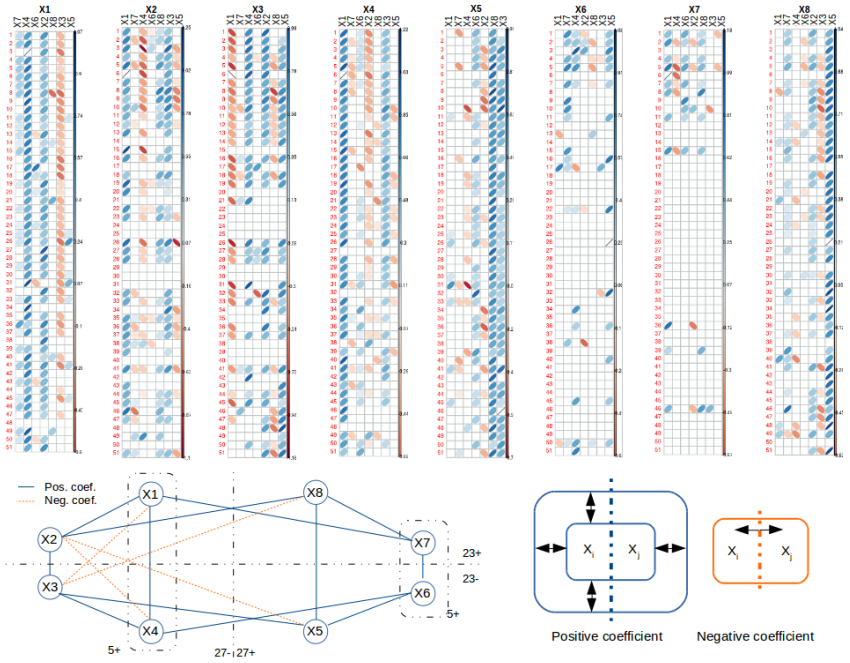

Fig. 1. Interactions among B cell subpopulations emerging from clinical data. Matrices represent the eight statistically valid multivariate models, with coefficient p-values less than 0.05. Rows represent age intervals and columns cell types. Blue (right-up diagonals) positive values and red (left-up diagonals) negative values of the model coefficients.

In [4] a constant (to all age intervals) network was provided by setting general assumptions, while an age-dependent one was found by restricting statistical thresholds to validate our multivariate linear models.

An evident property of the network in Figure 1 is to be tripartite, having blocks with $\{x_2, x_3\}$ and $\{x_6, x_7\}$ which are not directly connected each other, as both interact with the central block $\{x_1, x_4, x_5, x_8\}$. Such a network indicates that any couple of cell phenotypes which are different for the activation of only one receptor have direct proportional quantities, that is, if one of them increases/decreases, the other does the same. In terms of binary (or bit) strings, if we identify the subpopulation with a combination of binary states of three CD receptors (see Table I), then we may see the blue lines in the network (see Figure 1) as all the possible one-bit changes in a graph with eight nodes of degree 3. This phenomenon indicates that one receptor at the time may be lost or activated in each of these B cell phenotypes during our life. The four red lines in our network denote that namely CD23 may be expressed together with a second receptor, only when the third one is not expressed, and these interactions are negative, meaning that if the quantity of one of the involved phenotypes increases/decreases the other ones do the opposite (decrease/increase, respec-

tively). In other terms, cell phenotypes having expressed only one between CD27 and CD5, may loose it, together with the expression state change of CD23, and when this transformation happens it is not reversible, because if the quantity of these cells decreases then the other increases.

In [2], a previous network model was proposed for the above dataset, describing a possible sequence of (ex-vivo observed) B cell maturation steps in human body. It was based on Metabolic P systems, with linear regulation maps, generated by regression techniques based on genetic algorithms. MP systems are discrete dynamical systems [10,11], introduced in the context of membrane computing [7,1], by giving a deterministic perspective where multiset rewriting rules are equipped with state functions that determine the quantities of transformed elements. An algebraic formulation of their dynamics, combined with methods of statistical regression, provided systemic solutions (MP systems) to generate observed time series of given phenomena [3].

Ongoing work. Segmentation of multiple time series is a complex problem, since different data sequences may show different aspects of the underlying processes, and these aspects could also have non-synchronous evidence. Main methodologies in this field take inspiration and extend methods of motif discovery in time series [9,5] or are based on motif clustering [6,13]. Our cross-sectional data may be reduced to multivariate time series if we sort them according to the age of patients.

It turns out that the choice of the information measure has a strong influence on the identification of segments (i.e., time intervals/clusters of time points) and change points. A possible measure is represented by the parameters of the (multivariate) mathematical models fitting the data in each segment, since they represent some aspects of the information in the segment itself. Comparing these parameters between couples of adjacent segments and maximizing their differences is a way to identify good segmentations. If linear regression models are used, as in [4], then predictor coefficients may be mutually compared, while if probabilistic models are employed, then means and covariance matrices can be compared. In our case the information contained in covariance matrices is of particular interest, because it is a proxy for relationships between two cell types.

More complex models, considering for instance the dynamics of the multivariate time series in each segment, could be used to detect more subtle/advanced properties of the information contained in those segments. Since complex models usually have a large number of coefficients, a trade-off between model complexity and sensitivity to specific informational changes must be considered. The number of segments or change points must be also chosen in order to maximize the goodness of fit of the models in each segment. Since the choice of the best partition, according to the requirements here defined, is a time consuming task, several heuristics were proposed in the literature.

We are currently testing some of them and extending their capabilities to improve their performance in our specific application. Namely we are employing a very recent approachwhere subsequence clustering of multivariate time series

is profitably used for discovering repeated patterns in temporal data. Once these patterns have been discovered, the initial dataset can be interpreted as a temporal sequence of only a small number of states (namely clusters or segments). Patterns are defined by Markov Random Field (MRF) characterizing the interactions between different variables in typical subsequences of specific clusters. Based on this graphical representation, a simultaneous segmentation of time series data is efficiently realized.

References

1. L. Bianco, F. Fontana, G. Franco, and V. Manca. P systems for biological dynamics. In G. Ciobanu, M.J. Perez-Jimenez, and G. Păun, editors, *Applications of Membrane Computing*, pages 81–126. Springer, 2006.
2. A. Castellini, G. Franco, V. Manca, R. Ortolani, and A. Vella. Towards an MP model for B lymphocytes maturation. In *Unconventional Computation and Natural Computation*, volume 8553 of *LNCS*, pages 80–92, Berlin, Germany, 2014. Springer.
3. A. Castellini, G. Franco, and R. Pagliarini. Data analysis pipeline from laboratory to MP models. *Natural Computing*, 10(1):55–76, 2011.
4. A. Castellini, G. Franco, and A. Vella. Age-related relationships among peripheral B lymphocyte subpopulations. In *2017 IEEE Congress of evolutionary computation*, LNCS, Berlin, Germany, To appear. Springer.
5. B. Chiu, E. Keogh, and S. Lonardi. Probabilistic discovery of time series motifs. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 493–498. ACM, 2003.
6. F. Duchêne, C. Garbay, and V. Rialle. Learning recurrent behaviors from heterogeneous multivariate time-series. *Artif. Intell. Med.*, 39(1):25–47, 2007.
7. G. Franco and V. Manca. A membrane system for the leukocyte selective recruitment. In C. Martin-Vide, G. Mauri, G. Păun, Gh.and Rozenberg, and Salomaa A., editors, *Revised Papers of Membrane Computing – WMC 2003*, volume 2933 of *LNCS*, pages 181–190. Heidelberg Germany, Springer-Verlag, 2004.
8. N. K. Jerne. Towards a network theory of the immune system. *Annales d'immunologie*, 125C(1-2):373–389, January 1974.
9. J. Lin, E. Keogh, S. Lonardi, and P. Patel. Finding motifs in time series. In *Proceedings of the Second Workshop on Temporal Data Mining*, 2002.
10. V. Manca. *Infobiotics: Information in biotic systems*. Springer, 2013.
11. V. Manca, A. Castellini, G. Franco, L. Marchetti, and R. Pagliarini. Metabolic p systems: A discrete model for biological dynamics. *Chinese Journal of Electronics*, 22:717–723, 2013.
12. A.S. Perelson. Immune network theory. *Immunological Reviews*, 110:5–33, 1989.
13. A. Vahdatpour, N. Amini, and M. Sarrafzadeh. Toward unsupervised activity discovery using multi-dimensional motif detection in time series. In *Proc. of the 21st Int. Joint Conference on Artifical Intelligence*, IJCAI'09, pages 1261–1266, 2009.
14. D. Weiskopf, B. Weinberger, and B. Grubeck-Loebenstein. The aging of the immune system. *Transplant International*, 22:10411050, 2009.