Age-related relationships among peripheral B lymphocyte subpopulations

Alberto Castellini, Giuditta Franco
Department of Computer Science
University of Verona, Italy
Email: {alberto.castellini, giuditta.franco}@univr.it

Antonio Vella
Polyclinic Hospital of Verona
Medical University, Verona, Italy
Email: antonio.vella@univr.it

***Abstract*—An immunological data-driven model is here proposed, for age related changes in the network of relationships among cell quantities of eight peripherical B lymphocyte subpopulations, that is, cells exhibiting all combinations of three specific receptor clusters (CD27, CD23, CD5). We model phenomena based on real immunological data (quantities of cells exhibiting CD19, characterizing B lymphocytes) of about six thousands patients having ages between one day and ninetyfive years, by suitably combining traditional data analysis methods, such as piecewise linear regression models. With relaxed values for statistically significant models (coefficient p-values bounded by 0.05), we found a network holding for all ages, that likely represents the general assessment of adaptive immune system for healthy human beings. When statistical validation comes to be more restrictive, we found that some of these interactions are lost with aging, as widely observed in medical literature. Namely, interesting (inverse or directed) proportions are highlighted by pure data analysis among quantities of a partition of peripherical B lymphocytes.**

I. Introduction

According to the immune network theory formulated by Jerne [1] (called also Jernes hypothesis), and further developed by Perelson [2], the immune response is based not only on the interaction of B-cells and antigens but also on the interactions of B-cells with other B-cells. Despite outstanding achievements on this front, we currently have more questions than answers. The idea of an immune network, in terms of idiotype anti-idiotype relations, as a regulatory mechanism of the immune system is quite attractive, since it enables us to explain the selectivity (specificity) and the clonal selection. Nevertheless, it remains unclear how the network of immune cells is organized, how it operates, and how it exerts control over autoimmunity and/or infection and there are no concrete applications in modern and fundamental immunology [3].

An unconventional way to look at the immune system is to consider it as a whole, as a discrete unit made up of relations between components of the system, mainly cells and antibodies (situated within a defined space as seen by an external human observer). According to this view, immunological activity may be seen as an inward movement inside the immune system, occurring through various known and not known mechanisms, in order to primarily preserve its organization (relations among components) as adaptation to the immediate surrounding environment. Among the effects of the mechanisms occurring inside the immune system to maintain its state of activation, there is the changes in amounts of cells present in different anatomical location of the immune system. An external observer can see relationships among amounts of different cells, that change in course of the life, as an epiphenomenon of mechanisms that occur among components (cells, antibodies and molecules) inside the system and the surrounding environment.

We have developed our computational model as a contribution along the above perspective on immunological networks, by providing a careful examination of age related changes of cell amounts in networks of specific types of lymphocytes, at a different grade of maturation level. The determination of B lymphocytes with a defined phenotype in peripheral blood is an increasingly requested exam by physicians for the purpose of therapy and diagnosis of immunological diseases, mainly immunodeficiency and autoimmunity. In addition, during the last years many biological therapies that target B cells were introduced in therapy. Therefore we think that finding relations and/or rules among peripheral blood B subsets, being representative of all circulating B lymphocytes [4], along the age space as representative of the course of life, could be interesting for main basic knowledge as well as for discoveries with diagnostic and therapeutic purpose.

A. Our model

We aim at representing the evolution of relationships among quantities of peripherical subpopulations of B lymphocytes, by means of a set of networks, where each network corresponds to a model for an age interval. Starting from our database of patients, we split the entire range of ages in segments in which relationships among cell types are statistically significant, and represent by a graph their interactions, where cell types are nodes and variation relationships (i.e., linear model coefficients) are represented by directed edges. An illustration of our method to generate networks is depicted in Fig. 1.

A previous quantitative model has been developed on this dataset in [5], with completely different modeling methods, aimed at simulating the maturation dynamics among cell subpopulations, starting from their quantities and computing network fluxes by computing flux regulation functions by multiple linear regression and genetic algorithms [6], [7]. A key point in that modeling work was the use of MP sytems [8]–[10], equivalent to ODE and hybrid Petri nets [11],

978-1-5090-4601-0/17/$31.00

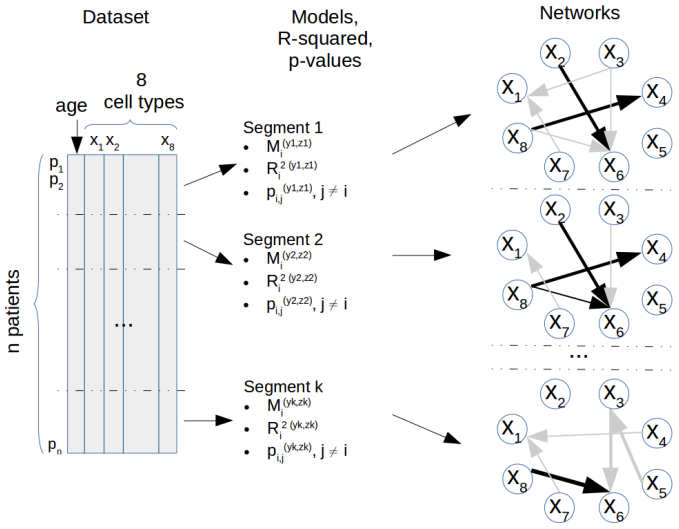


Fig. 1: **Pictorial exemplification of network generation.** Dataset is segmented in order to maximize model performance. Networks are generated from linear models with good statistical performance in terms of R^2 and coefficient p-values; the evolution of interactions among amounts of different cell types, across patient age, is shown on the right, where gray arrows represent negative coefficients, black arrows positive coefficients, and edge orientation the role of the two involved nodes in the model.

and able to reproduce with a simple grammar structure even complex non-linear dynamics. More qualitative (theoretical) models of immunological phenomena have been proposed in the context of maximally parallel multiset grammars [12], so called P-systems. Indeed, immunological dynamics have clear distinct checkpoints which may be well modelled by discrete systems [13]. More fundamentally, here we start from cell quantity series and employ methods from multiple time series segmentation to find relationships (possible reciprocal dependencies) among cell quantities in specific range of ages.

Our methodology represents an approach trasversal to those typical of artificial immune systems (AIS), as we model real immunological data by combining traditional data analysis methods, rather than developing new immune inspired methods to solve optimization problems. Our work is described in the following, along with four main sections. First the necessary background, goal and motivation are given, in both immunological and computational terms. Our network model is presented in details in Section II, along with the algorithm generating it, and is followed by a discussion of the results in Section III. Our aim to continue, deepen, and expand this research work is outlined and discussed as a conclusion section of the paper.

B. Immunology background

The main actors of the adaptive immune system are B and T lymphocytes, that origin from a progenitor derived from a stem cell in the bone marrow, and once mature circulate in the peripheral districts of the immune system and of tissues, through the blood vessels and capillaries that drain the cells in the lymphatic vessels, and once within the blood stream through the vessel "thoracic duct" they return to peripheral blood. In [14] a qualitative model was proposed for leukocyte recruitment, that plays a critical role in the immune response. B lymphocytes maturate in the bone marrow, and are related to humoral immunity, as they secrete antibodies, whereas T cells migrate and maturate in the thymus, and are related to cell immunity.

The activation of the immune response is generated in different phases. The antigen is processed into peptides, assembled with molecules of the major histocompatibility complex class I and II (MHC I, II) and expressed on the cell membrane of antigen presenting cells (APC), mainly dendritic cells. B and T lymphocytes recognize antigen peptide associated with molecules of MHC I, II through their T and B cell receptor that occur respectively in T cell cortex area (T cells), and B cells follicles of lymph nodes and spleen white pulp. B lymphocytes and T lymphocytes initiate the process (by recognizing antigenic peptides) driving the immune response that could develop in activating, inhibiting and/or regulatory [15]. Naive B cells are generally activated by antigen thought the help of T cells, even if some B cells do not need the T cell contribution, and are defined as T cell independent B cells, as they do not need the help of T cell. They are located external at the marginal zone of the spleen lymphoid tissue. In peripheral lymph nodes or spleen B cells maturate and proliferate into memory and antibody-secreting plasma cells that migrate to the bone marrow [16].

In this work we have focused on B lymphocytes, distinguished into naive, that express CD5 and CD23 at different combinations, and memory B lymphocytes, that are distinguished into marginal zone and follicular B cells, both expressing CD27 [17].

C. Dataset

Peripheral blood corresponds to blood recovered by puncture at the arm brachial vein, a blood compartment reflecting the immunological response that occurs in the peripheral districts of the organism rather than in the peripheral blood. Flow cytometry technique combined with the use of specific conjugated antibodies allows to quantify the number of cells present in a examined biological liquid. The lymphocyte count is performed with criteria established by an immunological nomenclature: cluster of differentation (CD), which takes into account the anatomical location and/or immunological function, as well as the biology of molecules expressed by cells during their maturation and/or differentation. The result of the mentioned lab exam is called immunophenotype, that contemplates the enumeration of B lymphocytes, T, Natural Killer (NK) and their sub-populations [18].

In years 2001-12, at University Hospital of Verona, we collected data by assessing the enumeration of peripheral blood lymphocytes surface expressing CD19, CD5, CD23 and CD27 from 5,954 patients of all ages undergoing peripheral

Cell phenotype	Population size variables	Binary triples
CD5+ CD23+ CD27-	X1	110
CD5- CD23+ CD27-	X2	010
CD5- CD23- CD27-	X3	000
CD5+ CD23- CD27-	X4	100
CD5- CD23- CD27+	X5	001
CD5+ CD23- CD27+	X6	101
CD5+ CD23+ CD27+	X7	111
CD5- CD23+ CD27+	X8	011

TABLE I: **Variables of the models.** B cell phenotype of 8 subpopulations are identified by presence/absence of CDs, abstractly described by random variables (assuming quantities of corresponding cell, in each patient) or by binary triples.

blood immunophenotyping exam. The median age of the patients was 37 years (range: 0-95 years). There were 2,910 males and 3,045 females (male/female ratio: 0.95).

The majority of lymphocyte biological variability is age dependent [19] and could make difficulties to interpret results obtained from immunological phenotypes. The distributions of the values for these parameters were compared with the medians of reference values published in the literature, and it was found that most of the values from the subjects included in the database were close to the medians in the literature, probably due to adaptation capacity of the immune system that only when it becomes cronically activated and exhausted show real signs of anomalous conditions, as happens in HIV infection or leukemia [20], [21].

In this work we focused on the analysis of eight B cell subsets (see Table I) as in [5], where more details may be found, both on the biological process involving these immune cells and on the process of data collection. The starting point of our computational work is then a cross-sectional dataset which counts 5,954 observations and 8 variables. We will call X the data matrix, n the number of patients, eight is the number of cell types and $x_{i,j}$ the number of cells x_j in patient p_i.

Expression of receptor CD27 usually indicates the transition from naive B cells to memory cells, even if some memory B cells do not express CD27 molecules. In addition, some not conventional memory B cells that reassemble marginal zone B cells phenoptype express the molecules CD27 [22]. B cells corresponding to variables X1-X4 in Table I have a phenotype as described for naive B cells as they express also IgM and IgD on membranes. X5 has a phenotype compatible with memory cell, as they do not express IgM and IgD, instead cells express IgG or IgA or IgE. It was reported that B cells expressing molecules as in X5 when further activated could maturate in plasmacells that migrate back to the bone marrow [16]. Moreover, marginal zone memory B cells expressing on membrane CD27, CD38, IgM and IgD maturate into plasmacells. Plasmacells as detected in peripheral blood are recognized by a bright CD27 and CD38 molecules expression. Finally, X6-X8 have a phenotype compatible with those described as memory cells from which Cronic Lymphatic Leukemia (CLL) tumoral cells could originate [23].

D. Time series segmentation

Problems related to the partitioning of data sequences into a suitable number k of contiguous segments have been widely tackled in literature. Each segment is expected to contain "homogeneous" data, according to some criterion, while different segments are expected to show non-homogeneous data behaviours. Piecewise Linear Regression (PLR) is one of the most frequently used representations for time series, reduced to k straight lines [24]. An optimal solution to the segmentation problem of a single time series can be computed by dynamic programming methods [25], having a time complexity of $O(n^2k)$. Since a quadratic algorithm is often not adequate in practice, some heuristics were proposed that approximate the optimal solution and have better time performance [26]. Other heuristics guarantee specific error bounds or impose constraints on the structure of segment representatives to improve the performance.

There exist three major approaches to time series segmentations [24], [26], namely *i) sliding window methods*, in which a segment (of data) is elongated until the fitting error of the model with the data in the segment exceeds some bounds: the process is then iterated on data points not belonging to the current segmentation, *ii) top-down methods*, where the complete time series is recursively partitioned with a strategy which minimizes the fitting error, until all segments have an error lower than some threshold, *iii) bottom-up methods*, in which data points are merged to generate larger segments until these segments exceed some error threshold. Given the data points in a time series segment, *linear regression* is often used to compute the segment representative, namely the approximation line which better fits data points in the least squares sense.

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 [27], [28] or are based on motif clustering [29], [30].

Our cross-sectional data may be reduced to multivariate time series if we sort them according to the age of patients. The approach described in the next section is based on PLR and generates a set of networks describing relationships among cell quantities, over different age intervals, from our cross-sectional dataset.

II. METHODS

We analyzed our multiple time series $\{x_1(p), x_2(p), \ldots, x_8(p)\}$, where $x_i(p)$ denotes the quantity of cell type X_i for patient p, with the methodoloty described in the following. Starting from an initial set of infants (our youngest patients), a network inference method based on linear regression [31] is employed to generate a multivariate linear model with eight linear equations $x_i = \sum_{j \neq i} x_j \cdot \beta_j$ (one for each cell/random variable). The performance of the entire network is computed after elimination of outliers, as the sum of the performance of each model: $P^{(y,z)} = \sum_{i=1}^{8} R^2_{x_i}$, where

$R^2_{x_i}$ is the coefficient of determination of the multivariate linear model generated using x_i as a dependent variable, and the others as independent variables, y is the index of the first (i.e., youngest) patient of the interval and z is the index of the last (oldest) one. The above two-steps process is iterated on other intervals, obtained by elongating the current interval (y, z) until $P^{(y,z)}$ reaches its first local maximum. More precisely, to define the next segment, an initial interval is elongated with new older patients, until the longest one exhibits a multivariate model with a better performance.

Table II formalizes the methodology described above, in terms of the algorithm NetGen(q), which generates network models on intervals of ages that exhibit very similar behaviour in terms of cell amount relationships. It initially starts with a contiguous set (initial segment) $\mathcal{I}$ of q (a few) infants, and defines next intervals (y is the variable denoting the youngest patient of the current interval, and z is the variable denoting the oldest one), as in Table II. In particular, the *repeat-until* cycle in instruction (4) iterates on interval starting points (index y), while the *repeat-until* cycle in instruction (5) iterates on interval final points (index z) and the *for* cycle in instruction (7) iterates on cell types (index i). Given a specific interval of ages and corresponding patients (y, z), eight multiple linear models $M_i^{(y,z)}, i = 1, \ldots, 8$ are computed, which all together form a single network $N^{(y,z)}$ for that interval. Model performance $R_i^{2(y,z)}$ and coefficient p-values $p_{i,j}^{(y,z)}, j \neq i$ are also considered in network $N^{(y,z)}$ since an edge between cell types X_i and X_j is drawn *iff* the related model coefficient is not null, model performance is greater than a threshold τ_{R^2} and coefficient p-value is less than a threshold τ_p. Network performance $P^{(y,z)}$ are then computed from single model performances by instruction 8. The *repeat-until* loop (5) is iterated until network performance $P^{(y,z)}$ do not decrease, then a new interval is generated by the *repeat-until* loop (4) by considering older patients. The algorithm returns the set of models $M_i^{(y,z)}$ for $i = 1, \ldots, 8$, and related performance, for each interval (y, z).

In this way, we dynamically generate a reliable partition of the entire range of ages in 51 subintervals, with 51×8 multivariate models (statistically) well fitting the data from corresponding age intervals. Each equation $x_i = \sum_{j \neq i} x_j \cdot \beta_j$ may be visualized by a directed graph as in Figure 1, where an oriented edge from cell type c_j to cell type c_i with $i \neq j$ appears in the graph if the coefficient β_j is different from zero. We set the coefficient β_j equal to zero *iff* its p-value p_j is (strictly) greater than 0.05. Moreover, such an edge changes color according to the sign of β_j (green/standard line for positive and orange/dashed line for negative) and may be weighted by the corresponding coefficient p-value. Positive coefficients in the model represent variations inducing the same effect (i.e., an increase/decrease of x_j induces an increase/decrease of x_i) while negative coefficients represent variations inducing opposite effect (i.e., an increase/decrease of x_j determines a dicrease/increase of x_i, respectively) – see Figure 4.

NetGen(q)

Input: q # minimum number of patients in a segment

1. $y := 1$; # Index of the first patient of segment
2. $z := q - 1$; # Index of the last patient of segment
3. $P^{(1,q-1)} := 0$; # Initial model performance
4. **repeat**

5. **repeat**
6. $z := z + 1$;

7. **for** $i = 1$ **to** 8 **do** # i: variable index
   a. Compute model $M_i^{(y,z)} : x_i = \sum_{j \neq i} x_j \cdot \beta_j$ (from data of patients in the current interval);
   b. Remove outliers (in the current segment);
   c. Compute model performance $R_i^{2(y,z)}$ and coefficient p-values $p_{i,j}^{(y,z)}, j \neq i$;
   end for

8. $P^{(y,z)} := \sum_{i=1}^{8} R_i^{2(y,z)}$;

until $P^{(y,z)} > P^{(y,z-1)}$;

9. **Output:** multivariate model $M_i^{(y,z)}$ for $i = 1, \ldots, 8$, related performance $P^{(y,z)}$ and coeff p-values;

10. $y := z$; $z := y + p - 1$; # Indices update

until $y \geq 5,954$.

TABLE II: Algorithm for the generation of age-dependent models, which are networks of relationships among B lymphocyte subpopulations.

The idea behind this algorithm is the way to choose subintervals which maximize the average performance among the eight models generated (one for each variable) by using patients in this range of ages. More complex strategies considering simultaneously model performance, adjusted coefficient p-values and possibly other measures are under development. In this initial phase, we are interested to keep all possible coefficients different than zero, in order to find general connections among cell amount variations. This is why the threshold for coefficient p-values is 0.05, and why we did not employ regularization-based algorithms, which however could be used in a future investigation. Besides, outlier detection is here performed by a procedure based on median absolute deviation (MAD), resulting in data reduction as in Figure 2, while multivariate outlier detection methods [32] are being considered for future developments.

In conclusion, networks $N^{(\bar{y},\bar{z})}$ for 51 intervals $(\bar{y}, \bar{z})$ of patients as in the bottom picture of Fig. 3 have been generated, from their multivariate models $M^{(\bar{y},\bar{z})}$, with their performance in terms of average R^2 and p-values. We may notice the intervals have an average of patients greater than 100, even if a major number of intervals has less than 50 patients (widely distributed in age). Indeed a couple of intervals have more than 200 and even 250 patients, and age distribution along the 51 intervals may be seen in Figure 3 as well.

We defined a threshold $\tau_{R^2} = 0.5$ for model performance

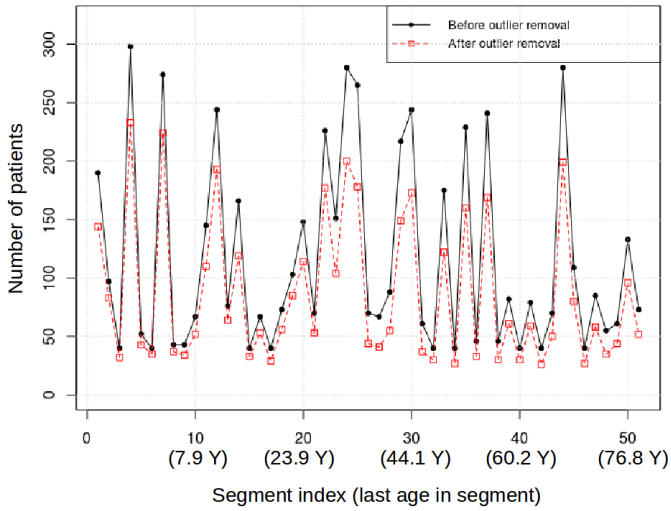
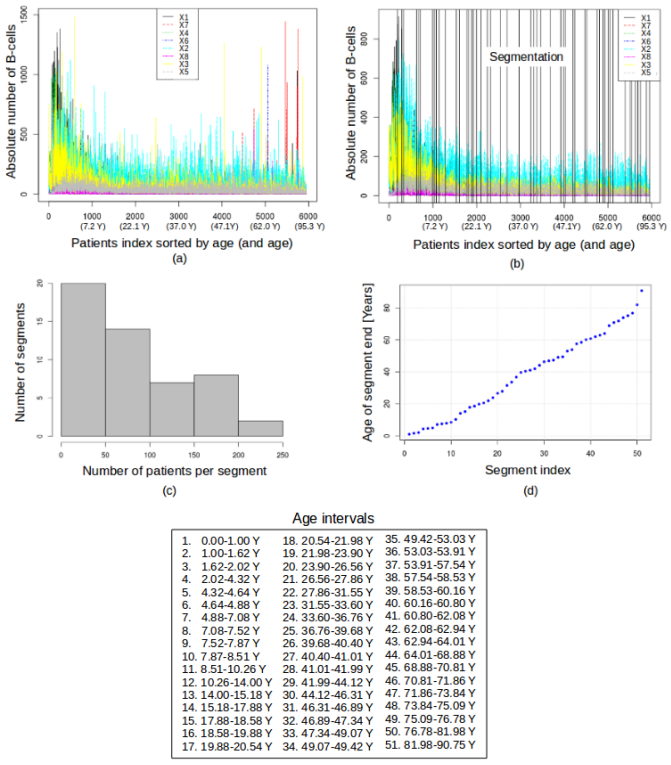


Fig. 2: Data distribution is not altered by outliers elimination – the number of outliers varies on intervals.



Age intervals

1. 0.00-1.00 Y	18. 20.54-21.98 Y	35. 49.42-53.03 Y
2. 1.00-1.62 Y	19. 21.98-23.90 Y	36. 53.03-53.91 Y
3. 1.62-2.02 Y	20. 23.90-26.56 Y	37. 53.91-57.54 Y
4. 2.02-4.32 Y	21. 26.56-27.86 Y	38. 57.54-58.53 Y
5. 4.32-4.64 Y	22. 27.86-31.55 Y	39. 58.53-60.16 Y
6. 4.64-4.88 Y	23. 31.55-33.60 Y	40. 60.16-60.80 Y
7. 4.88-7.08 Y	24. 33.60-36.76 Y	41. 60.80-62.08 Y
8. 7.08-7.52 Y	25. 36.76-39.68 Y	42. 62.08-62.94 Y
9. 7.52-7.87 Y	26. 39.68-40.40 Y	43. 62.94-64.01 Y
10. 7.87-8.51 Y	27. 40.40-41.01 Y	44. 64.01-68.88 Y
11. 8.51-10.26 Y	28. 41.01-41.99 Y	45. 68.88-70.81 Y
12. 10.26-14.00 Y	29. 41.99-44.12 Y	46. 70.81-71.86 Y
13. 14.00-15.18 Y	30. 44.12-46.31 Y	47. 71.86-73.84 Y
14. 15.18-17.88 Y	31. 46.31-46.89 Y	48. 73.84-75.09 Y
15. 17.88-18.58 Y	32. 46.89-47.34 Y	49. 75.09-76.78 Y
16. 18.58-19.88 Y	33. 47.34-49.07 Y	50. 76.78-81.98 Y
17. 19.88-20.54 Y	34. 49.07-49.42 Y	51. 81.98-90.75 Y

Fig. 3: **Data segmentation.** Data profiles are reported, before and after segmentation of intervals of patients. At the bottom, ages are reported related to the intervals computed by the algorithm NetGen(q), while the histogram illustrates number of patients per interval.

and a threshold $\tau_p = 0.05$ for coefficient p-values. Coefficients of models having performance less than τ_{R^2} were not considered in the networks. Moreover, coefficients of models having performance greater than τ_{R^2} but also p-value greater than τ_p were set to zero. All the other cases have corresponding edges in the network. For each interval of ages computed in this way, a combination of eight networks is produced as a final network (of cell types for patients having age in that range), which is analyzed in its evolution in the next section.

III. Results

In every segment of patients $(\bar{y}, \bar{z})$, we have combined together the eight linear models, one for each variable corresponding to one subpopulation (or cell type, see Table I), to form one network $N^{(\bar{y},\bar{z})}$. Out of models with $R^2 \geq 0.5$ those with coefficient p-values smaller than either 0.05 or 0.01 were considered, respectively reported in Figure 4 and Figure 5.

Let us consider, for instance, the first (top-left) heatmap of Figure 4, representing models for cell X_1. Each row represents an age interval and each column a cell type. Heatmap cells represent model coefficients, where blue (right-up diagonals) represent positive values and red (left-up diagonals) negative values. For example, the first cell of the first heatmap represents the coefficient of subpopulation X_7 (see column name) in the linear model for subpopulation X_1 (see heatmap name) in the first age interval (see row index). This cell is blue (right-up diagonal), meaning that X_7 has a positive effect on X_1 in that age-interval, and the color (or gray) intensity represents the "weight" of the coefficient (i.e., its absolute value).

The network in Figure 4 has 16 non-oriented edges, as representing relationships with the same coefficient sign, in the opposite directions, for the whole life long. However, if we restrict ourself to more significant edges (that is, coefficient p-values smaller than 0.01 rather than 0.05) we obtain the subnetwork in Figure 5, which has only 8 (of the 16 shown in Fig. 4) edges surviving for the life, while a couple of edges are present only up to about 35 years old, all three edges involving x_7 are present only up to 18 years, and the edge $x_2 - x_4$ is present only for children (up to 10 years old). According to the meta-dynamics of the self-organized immune system [33] a constant network among B cell subpopulations could be the underlying structure of internal interactions helpful for adaption and survival of the system itself.

We notice that both our models resulted in undirected graphs, meaning that for any couple of cells in the network the respective coefficients in the model have the same sign (and comparable stastistical significance). In other terms, whenever $x_i = \sum_{j \neq i} x_j \cdot \beta_j$ and $x_k = \sum_{j \neq k} x_j \cdot \alpha_j$, if X_i and X_k are two different nodes connected in the network by an edge, then β_k and α_i are different than zero and have the same sign.

By looking at matrices in Figures 4 and 5 long columns of coefficients (corresponding to one specific variable) keep the same sign along the intervals (when different than zero). Only in sporadic cases, when this does not happen, we carried out a partial residuals analysis to check that they were not statistically significant cases. A similar validation has been done for the rare cases of intervals having all coefficients of one model equal to zero (corresponding to one zero-row in the matrices). In order to obtain a first "stable" network (involving a widely major part of patients), as in Figure 4, we assumed that only coefficients having p-values less than 0.05 *for at least*

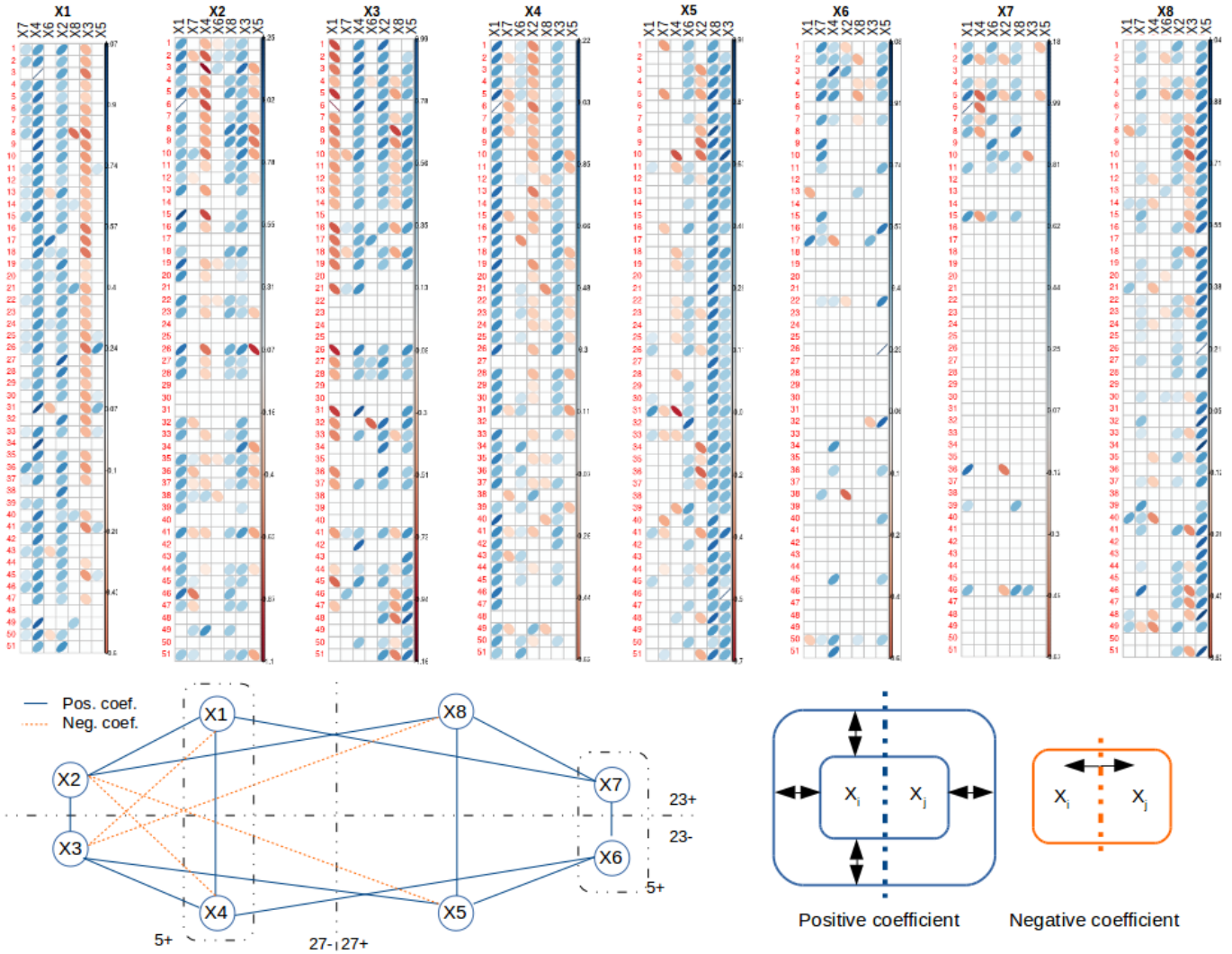


Fig. 4: **Interactions among B cell subpopulations emerging from clinical data.** Matrices represent the eight statistically valid multivariate models, with coefficient p-values less than 0.05. Each row represents an age interval and each column a cell type. Blue (right-up diagonals) represent positive values and red (left-up diagonals) negative values of the model coefficients. For each pair of variables, relationships were found (with no order) having the same coefficient sign for all the 51 intervals (along the entire columns). Corresponding network turns out constant for the whole life, where blue/plain lines denote direct proportionality between quantity of corresponding cell/nodes and orange/dashed line denote inverse proportionality. As sketched on the right hand, positive coefficients in the model correspond to same variations (an increase/decrease of x_j induces an increase/decrease of x_i), negative coefficients correspond to opposite variations (an increase/decrease of x_j determines a dicrease/increase of x_i).

34 intervals are different than zero (i.e. they may appear in the graph).

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 4) 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, rispectively). 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 tranformation happens it is not reversible, because if the quantity of these cells decreases then the other increases.

An evident property of the network in Figure 4 is to be

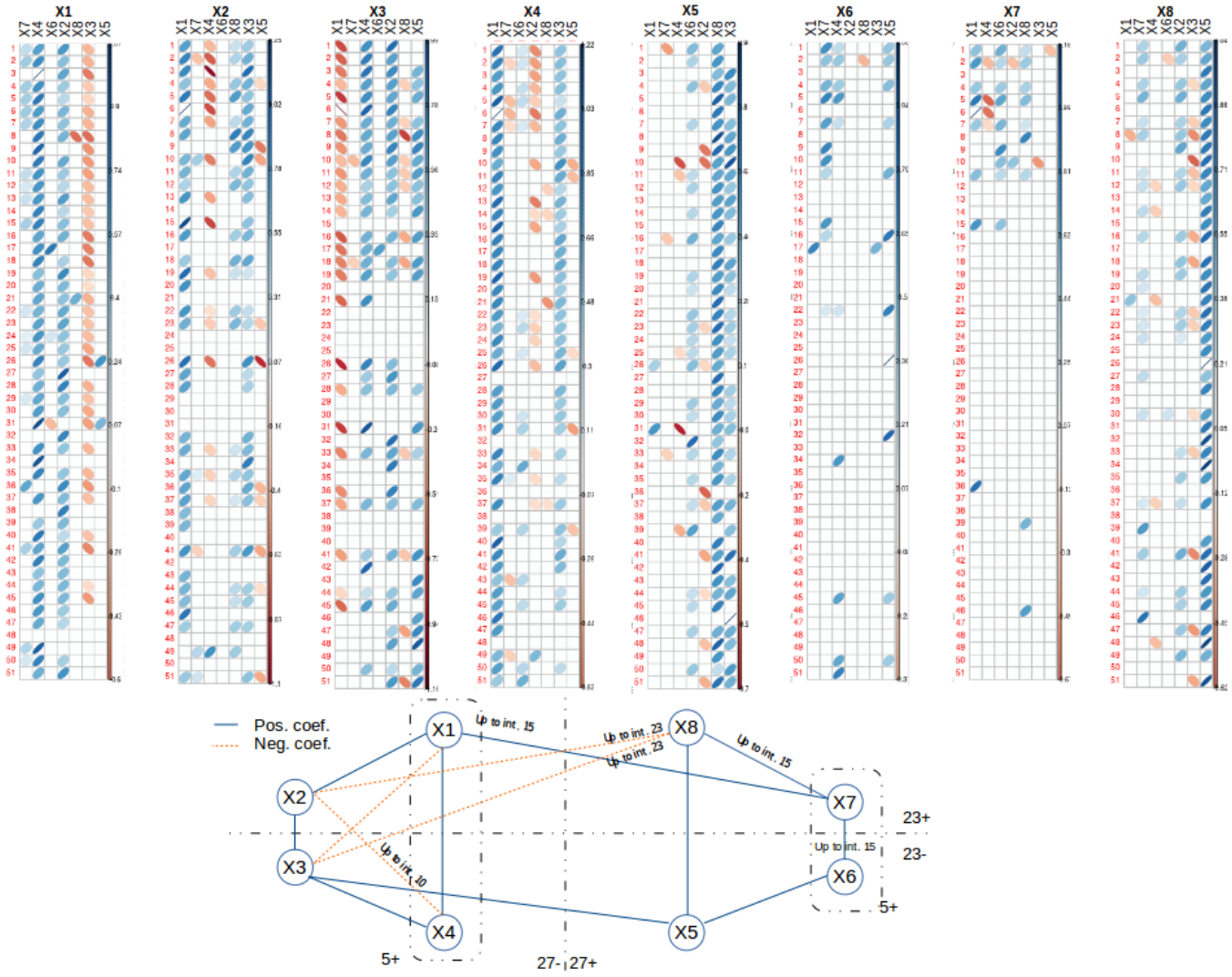


Fig. 5: **Interactions among B cell subpopulations emerging from clinical data.** Matrices represent statistically valid multivariate models, with coefficient p-values less than 0.01. Each row represents an age interval and each column a cell type. Blue (right-up diagonals) represent positive values and red (left-up diagonals) negative values of the model coefficients. Corresponding network has seven of the blue connections X1-X2, X2-X3, X3-X4, X4-X1, X3-X5, X5-X8, X5-X6, and one red connection X1-X3, all constant for the life of our patiens, two (direct and inverse, respectively) connections X2-X8 and X3-X8 holding up to interval 23, three X1-X7, X6-X7, X8-X7 holding up to interval 15, and one direct connection X2-X4 up to interval 10 (see Figure 3).

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 $\{x_1, x_4, x_5, x_8\}$. In the network in Figure 5 (which is a specific subnetwork) such a structure is lost, however, the cycle $\{x_1, x_2, x_3, x_4\}$ keeps its identity constant, and x_5 keeps all its positive connections with the rest of the network. Only one negative interaction is mainteined for the whole life $\{x_1, x_3\}$, while two of the other three last only for few first years (10 years for $\{x_2, x_4\}$ and 35 for $\{x_3, x_8\}$). In both networks, node degree denotes a greater number of interactions involving naive cells $\{x_2, x_3\}$ and scarse interactions among memory cells $\{x_6, x_7\}$, as it is known in medical literature. In particular, we notice that cell phenotype X_7, having all CD expressed, becomes soon (after 18 years of life) an isolated node in out network.

In this context it is known that as people age, the immune system becomes less able to distinguish self from nonself (that is, to identify foreign antigens). As a result, autoimmune disorders become more common, and this slowdown may be one reason that cancer is more common among adults and older people. Older people have smaller amounts of complement proteins and do not produce as many of these proteins as younger people do in response to bacterial infections. These changes may partly explain why pneumonia, influenza, infectious endocarditis, and tetanus are more common among older people and result in death more often.

The network in Figure 5 is in a sense more statistically reliable (because of the p-value threshold set to 0.01), and seems to confirm some ongoing experiments of B-cell metilation, since in the cycle $\{x_1, x_2, x_3, x_4\}$, constant for the whole life and in both our models a constant level of metilation is mainteined, while passing to memory cell X5 there is a rough change, meaning that a clear differentiation or maturation happens between the two blocks of cells $\{x_1, x_2, x_3, x_4\}$ and $\{x_5, x_6, x_7, x_8\}$, as described in subsection I-B.

IV. Conclusion

Our general target here is to extract knowledge from a significantly large dataset of clinical data. Linear modeling is the first and correct choice when the interval of patients is small enough to show a linear behaviour, then we focused on the choice of segmentation of the patients range, by looking at the values in the dataset as temporal series of eight random variables (where the time is given by the exact age of patient - we have enough observations to assume that the time incrementation is constant). A constant (to all segments) network was provided by setting general assumptions, while an age-dependent one was found by restricting statistical thresholds to validate our multivariate linear models. These networks seem to find confirmation in medical literature, since two (maturation) cycles are present in both networks, one among naive B cells and the other among memory ones. This is however an ongoing research, which we aim at developing in many directions. Namely, we indend to keep investigating the changes of our network during aging, by variously refining the modelling process. Highlighting of hide relationships between lymphocyte subsets could enable system approaches to a better use of the phenotyping lymphocytes test for the medical purposes. For example, the cellular origin of Chronic Lymphocytic leukemia (CLL) is still debated, although some information about the adaption of cellular immunological network is critical to understanding its pathogenesis [23].

References

[1] N. Jerne, "Towards a network theory of the immune system," *Ann. Immunol (Inst. Pasteur)*, vol. 125C, pp. 373–389, 1974.

[2] A. Perelson, "Immune network theory," *Immunological Reviews*, vol. 110, pp. 5–33, 1989.

[3] I. Menshikov, L. Beduleva, M. Frolov, N. Abisheva, T. Khramova, E. Stolyarova, and K. Fomina, "The idiotypic network in the regulation of autoimmunity: Theoretical and experimental studies," *J. Theor Biol.*, vol. 21, no. 375, pp. 32–9, 2015.

[4] V. Ganusov and J. Auerbach, "Mathematical modeling reveals kinetics of lymphocyte recirculation in the whole organism," *PLOS – Computational Biology*, vol. 10, no. 5, p. e1003586, 2014.

[5] A. Castellini, G. Franco, V. Manca, R. Ortolani, and A. Vella, "Towards an MP model for B lymphocytes maturation," in *Proceed. of UCNC*, ser. LNCS, vol. 8553. Springer, 2014, pp. 80–92.

[6] A. Castellini, M. Zucchelli, M. Busato, and V. Manca, "From time series to biological network regulations: an evolutionary approach," *Molecular BioSystems*, vol. 9, pp. 225–233, 2013.

[7] A. Castellini, D. Paltrinieri, and V. Manca, "MP-GeneticSynth: inferring biological network regulations from time series," *Bioinformatics*, vol. 31, no. 5, pp. 785–787, 2015.

[8] A. Castellini, G. Franco, and R. Pagliarini, "Data analysis pipeline from laboratory to mp models," *Natural Computing*, vol. 10, no. 1, pp. 55–76, 2011.

[9] V. Manca, A. Castellini, G. Franco, L. Marchetti, and R. Pagliarini, "Metabolic P systems: A discrete model for biological dynamics," *Chinese Journal of Electronics*, vol. 22, no. 4, pp. 717–723, 2013.

[10] V. Manca, *Infobiotics – Information in Biotic Systems*. Springer, 2013.

[11] A. Castellini, G. Franco, and V. Manca, "Hybrid functional petri nets as MP systems," *Natural Computing*, vol. 9, no. 1, pp. 61–81, 2010.

[12] G. Franco, N. Jonoska, B. Osborn, and A. Plaas, "Knee joint injury and repair modeled by membrane systems," *BioSystems*, vol. 91, no. 3, pp. 473–88, 2008.

[13] V. Manca, G. Franco, and G. Scollo, "State transition dynamics: basic concepts and molecular computing perspectives," in *Molecular Computational Models: Unconventional Approaches*, M. Gheorghe, Ed. Elsevier, 2005, ch. 2, pp. 32–55.

[14] G. Franco and V. Manca, "A membrane system for the leukocyte selective recruitment," in *Membrane Computing*, ser. LNCS, vol. 2933. Springer, 2004, pp. 181–190.

[15] D. D. Chaplin, "Overview of the immune response," *J Allergy Clin Immunol*, vol. 125, no. 2, pp. S3–23, 2010.

[16] S. Nutt, P. Hodgkin, D. Tarlinton, and L. Corcoran, "The generation of antibody-secreting plasma cells," *Nat Rev Immunol.*, vol. 15, no. 3, pp. 160–71, 2015.

[17] Y. Wu, D. Kipling, and D. K. Dunn-Walters, "The relationship between CD27 negative and positive B cell populations in human peripheral blood," *Frontiers in Immunology*, vol. 2, no. 81, 2011.

[18] F. Craig and K. Foon, "Flow cytometric immunophenotyping for hematologic neoplasms," *Blood*, vol. 111, no. 8, pp. 3941–67, 2008.

[19] D. Veneri, R. Ortolani, M. Franchini, G. Tridente, G. Pizzolo, and A. Vella, "Expression of CD27 and CD23 on peripheral blood B lymphocytes in humans of different ages," *Blood Transfus*, vol. 7, pp. 29–34, 2009.

[20] L. Al-Harthi, S. MaWhinney, E. Connick, R. Schooley, J. E. Forster, C. Benson, M. Thompson, F. Judson, and A. Landay, "Immunophenotypic alterations in acute and early HIV infection," *Clin. Immunol.*, vol. 125, no. 3, pp. 299–308, 2007.

[21] D. Goldman, "Chronic lymphocytic leukemia and its impact on the immune system," *Clin. J. Oncol. Nurs.*, vol. 4, no. 5, pp. 233–36, 2000.

[22] J. Fecteau, G. Côté, and S. Néron, "A new memory CD27- IgG+ B cell population in peripheral blood expressing VH genes with low frequency of somatic mutation," *J Immunol.*, vol. 177, no. 6, pp. 3728–36, 2006.

[23] M. Seifert, L. Sellmann, J. Bloehdorn, F. Wein, S. Stilgenbauer, J. Dürig, and R. Küppers, "Cellular origin and pathophysiology of chronic lymphocytic leukemia," *JEM*, vol. 209, no. 12, pp. 2183–2198, 2012.

[24] E. Keogh, S. Chu, D. Hart, and M. Pazzani, "Segmenting time series: A survey and novel approach," in *In an Edited Volume, Data mining in Time Series Databases. Published by World Scientific*. Publishing Company, 1993, pp. 1–22.

[25] R. Bellman and R. Roth, "Curve fitting by segmented straight lines," *Journal of the American Statistical Association*, vol. 64, pp. 1079–1084, 1969.

[26] E. Terzi and P. Tsaparas, "Efficient algorithms for sequence segmentation," in *Proceedings of the 2006 SIAM International Conference on Data Mining*. SIAM, 2006, pp. 316–327.

[27] 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.

[28] 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*, ser. KDD '03. ACM, 2003, pp. 493–498.

[29] F. Duchêne, C. Garbay, and V. Rialle, "Learning recurrent behaviors from heterogeneous multivariate time-series," *Artif. Intell. Med.*, vol. 39, no. 1, pp. 25–47, Jan. 2007.

[30] 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*, ser. IJCAI'09, 2009, pp. 1261–1266.

[31] D. Marbach and et al., "Wisdom of crowds for robust gene network inference," *Nature Methods*, vol. 9, no. 8, pp. 796–804, Jul. 2012.

[32] P. Filzmoser, R. Maronna, and M. Werner, "Outlier identification in high dimensions," *Computational Statistics and Data Analysis*, vol. 52, no. 3, pp. 1694–1711, Jan. 2008.

[33] F. Varela and A. Coutinho, "Second generation immune networks," *Immunology Today*, vol. 12, pp. 159–166, 1991.