Cell Stress and Chaperones (2015) 20:391–395
DOI 10.1007/s12192-014-0555-y

SHORT COMMUNICATION

Data mining-based statistical analysis of biological data uncovers hidden significance: clustering Hashimoto's thyroiditis patients based on the response of their PBMC with IL-2 and IFN-γ secretion to stimulation with Hsp60

Lucio Tonello · Everly Conway de Macario · Antonella Marino Gammazza · Massimo Cocchi · Fabio Gabrielli · Giovanni Zummo · Francesco Cappello · Alberto J. L. Macario

Received: 2 October 2014 / Revised: 31 October 2014 / Accepted: 5 November 2014 / Published online: 19 November 2014

Abstract The pathogenesis of Hashimoto's thyroiditis includes autoimmunity involving thyroid antigens, autoantibodies, and possibly cytokines. It is unclear what role plays Hsp60, but our recent data indicate that it may contribute to pathogenesis as an autoantigen. Its role in the induction of cytokine production, pro- or anti-inflammatory, was not elucidated, except that we found that peripheral blood mononucleated cells (PBMC) from patients or from healthy controls did not respond with cytokine production upon stimulation by Hsp60 in vitro with patterns that would differentiate patients from controls with statistical significance. This "negative" outcome appeared when the data were pooled and analyzed with conventional statistical methods. We re-analyzed our data with non-conventional statistical methods based on data mining using the classification and regression tree learning algorithm and clustering methodology. The results indicate that by focusing on IFN-γ and IL-2 levels before and after Hsp60 stimulation of PBMC in each patient, it is possible to differentiate patients from controls. A major general conclusion is that when trying to identify disease markers such as levels of cytokines and Hsp60, reference to standards obtained from pooled data from many patients may be misleading. The chosen biomarker, e.g., production of IFN-γ and IL-2 by PBMC upon stimulation with Hsp60, must be assessed before and after stimulation and the results compared within each patient and analyzed with conventional and data mining statistical methods.

Keywords Hashimoto's thyroiditis · Hsp60 · IL-2 · IFN-γ · Data mining · Clustering · Delta values

Lucio Tonello, Everly Conway de Macario, Francesco Cappello, and Alberto JL Macario contributed equally to the present work.

L. Tonello · M. Cocchi · F. Gabrielli · F. Cappello
University of Human Sciences and Technology (LUdeS University), Lugano, Switzerland

E. Conway de Macario · A. J. L. Macario
Department of Microbiology and Immunology, School of Medicine, University of Maryland at Baltimore; and Institute of Marine and Environmental Technology (IMET), Columbus Center, Baltimore, MD 21202, USA

A. Marino Gammazza · F. Cappello · A. J. L. Macario
Istituto Euro-Mediterraneo di Scienza e Tecnologia (IEMEST), Palermo, Italy

A. Marino Gammazza · G. Zummo · F. Cappello (✉)
Department of Experimental Biomedicine and Clinical Neurosciences, University of Palermo, Via del Vespro 129, 90127 Palermo, Italy
e-mail: francapp@hotmail.com

M. Cocchi
Department of Veterinary Sciences, University of Bologna, Ozzano dell'Emilia, BO, Italy

Introduction

In a recent paper, we showed that Hsp60 levels are increased in thyroid tissue and blood of Hashitomo's thyroiditis (HT) patients compared to healthy individuals (controls) (Marino Gammazza et al. 2014). The levels of Hsp60 were measured by ELISA in blood and by immunohistochemistry and immunofluorescence in thyroid tissue (surgical specimens) and cells (fine needle aspiration samples from patients). We examined Hsp60 localization on subcellular compartments and found the chaperonin also on the surface of oncocytes.

These findings supported the notion that Hsp60 plays a role in the autoimmune manifestations of HT. To further assess the possible role of Hsp60 as a player in the autoimmune pathogenesis of HT, we evaluated (amino acid sequence and 3D structural comparisons by bioinformatics) the similarity between the chaperonin and two thyroid-specific molecules, thyroglobulin (TG), and thyroid peroxidase (TPO). We found that the Hsp60 molecule shares regions of similarity with TG and TPO. Subsequently, we confirmed by ELISA that there is cross-reactivity between Hsp60 and both thyroid proteins.

We also assessed (by flow cytometer) levels of cytokines in conditioned medium of peripheral blood mononuclear cells (PBMC) isolated from HT patients and controls, stimulated by recombinant Hsp60. This test was done because: (a) it has been suggested that cytokines are also involved in the pathogenesis of thyroid disease via stimulation of the immune system and direct targeting thyroid follicular cells (Mikoś et al. 2014); and (b) it has been postulated that Hsp60 can stimulate cytokine release in some pathologic conditions (Tsan and Gao 2004). We analyzed our PBMC stimulation results with a standard statistical method (hypothesis testing), and no differences in the levels of the cytokines assayed (IL-2, IL-4, IL-6, IL-10, TNF, and IFN-γ) were detected between PBMC controls (no in vitro stimulation with Hsp60) and stimulated with Hsp60, either from healthy controls or from HT patients. Hence, our data showed that Hsp60 did not stimulate cytokine production in PBMC from HT patients or normal subjects in ways that would allow a distinction between the two groups.

In view of the fact that there are contradictory reports on whether or not Hsp60 does stimulate cytokine production by different types of immune cells, we decided to re-examine our data utilizing non-conventional statistical methods, considering the possibility that conventional methods may miss key differences between groups. Furthermore, it is accepted that conventional hypothesis testing statistical methods do not provide a full description of the populations being compared and can fail to detect differences between them. Because of this, the application of non-conventional statistical methods to biological and medical samples is gaining acceptance in Biology and Medicine. Along these lines, we analyzed our data on cytokine secretion, using a nonparametric hypothesis testing, followed by another different approach named classification and regression tree learning algorithm. This is a data mining methodology, used in various scientific disciplines, including, lately, in the biological and medical fields (Richette et al. 2014; Wong et al. 2014). By applying this non-conventional statistical strategy, we could find in our data information of interest, which had not been evidenced by the conventional approach used earlier [1].

Methods and results

The results from HT and controls reported previously (Marino Gammazza et al. 2014) were assessed with the Kolmogorov-Smirnov and Lilliefors tests for normality. The outcome showed that the investigated parameters had to be considered "not normal". So, in order to find out possible differences between the two populations, the nonparametric Mann-Whitney U test was applied. The comparisons made were for all cytokines measured (see list above) between cytokines before, after, and between before and after the Hsp60 treatment. No parameter showed a statistically significant difference ($p<.05$). Hence, the cytokines evaluated started from similar values (no statistical difference), and, after the treatment, they reached similar values as well (Table 1).

In view of these results and considering the need to clarify whether or not cytokines participate as pathogenic factors in HT, the main aim of the present study was to elucidate the effect of the Hsp60 on cytokine production in HT patients. For this purpose, a new parameter was defined and used in the calculations. This new parameter is the Delta value, i.e., the difference of each cytokine's level, before and after the treatment of PBMC with Hsp60 (e.g., for IL2, it is: Delta IL2=[IL2 level after the treatment]−[IL2 level before the treatment]). The comparison of HT and controls was carried out using the U test as before (Marino Gammazza et al. 2014) because of the outcome of the Kolmogorov-Smirnov and Lilliefors test for normality. Again, no parameter showed a statistically significant difference (Table 1). These results fully agree with those previously reported (Marino Gammazza et al. 2014).

To summarize our observations, thus far, standard statistic methodologies indicated that HT and controls started with similar cytokine levels, and when treated with Hsp60, it showed similar final values (no statistical differences). Likewise, focusing on the Delta values (i.e., after-minus-before treatment values, as explained earlier), no statistically significant differences were unveiled. The conclusion would then be, according to the results from standard statistical methods, Hsp60 acts in the same way on the two populations investigated, i.e., HT and controls, and no differences become apparent in cytokine levels between the two populations.

We then applied a nonparametric classifier called the standard classification and regression tree (SCRT), which from our experience in other situations, seemed to be appropriate for our sets of data. SCRT, together with similar methodologies known as ID3 and C4.5 algorithms, is a member of the family known as decision tree learning algorithm. They are part of the classical artificial intelligence and machine learning

Table 1 Cytokine secretion by peripheral blood mononucleated cells (PBMC) before and after Hsp60 treatment in vitro: comparison of Hashimoto's thyroiditis patients (HT) with healthy individuals (controls (CT))

Cytokine	Before HSP60 treatment			After HSP60 treatment			Delta values		
	HT	CT	*p*	HT	CT	*p*	HT	CT	*p*
IFN-y	0.018[a]	0.172	0.096	0.096	0.31	0.364	0.078	0.138	0.791
	(0.031552)	(0.195323)		(0.106583)	(0.351821)		(0.092592)	(0.356832)	
TNF	7.036	5.348	0.496	14.082	8.398	0.910	7.046	3.05	0.623
	(4.522989)	(2.301366)		(15.70327)	(5.67462)		(14.63528)	(4.02791)	
IL10	4.364	3.361482	0.734	4.958	6.007	0.650	0.594	0.891	0.473
	(5.116)	(4.38554)		(3.216616)	(4.834067)		(1.645007)	(1.108347)	
IL6	86.749	26.962	0.076	107.483	104.918	0.571	20.734	77.956	0.241
	(76.61398)	(13.00919)		(45.99601)	(88.3114)		(60.06125)	(85.53102)	
IL4	0.219	0.422255	0.791	0.199	0.232	0.273	−0.02	0.094	0.791
	(0.138)	(0.212122)		(0.430515)	(0.266074)		(0.353113)	(0.228823)	
IL2	0.085	0.098686	0.364	0.045	0.499	0.227	−0.04	0.319	0.273
	(0.18)	(0.194822)		(0.057783)	(0.691334)		(0.123558)	(0.571828)	

[a] Each cytokine value (pg/ml) is expressed as mean and (SD). The *p* value was calculated according to the Mann-Whitney *U* test

algorithms (Russell and Norvig 2009). These algorithms are widely employed and are available in many analytic software products. The one we used here comes from the data analysis software STATISTICA (StatSoft, Inc., 2011; STATISTICA, data analysis software system, version 10. www.statsoft.com) and it is a comprehensive implementation of the methods described as CART® (Breiman et al. 1984).

The goal of the algorithm is to create a mathematical model called decision tree that classifies the value of a target grouping variable and based on other input variables. It is a supervised training algorithm, so the construction of the tree is made on a database of class-labeled training tuples. The resulting tree is like a flowchart scheme starting from a topmost node, called root, Fig. 1. The root denotes a test on an attribute, namely an input variable. According to the outcome of the test, a particular branch is chosen. The selected branch leads to a new node. This node, in turn, denotes a new test on an attribute and each of its daughter branches represents a possible outcome of the new test, and so on until a terminal node (a node with no branch) called leaf is met. A leaf is labeled with a target class; actually, each node is labeled with a target class but only a leaf-node class is fundamental for classification. So, starting from the root and answering to each node's questions, it is arrived at a leaf, which represents a class membership. Note that the attribute, the input variables, of each node of the tree are chosen by the learning algorithm itself, in an autonomous way, and only the ones considered necessary by the algorithm are chosen.

In our case, in order to study the possible effect of Hsp60, the different Delta values for cytokines are the input variables and the pathologic (HT) or non-pathologic (controls) populations are the two target classes. The decision tree learning algorithm constructs the classification tree based on our database. The classification tree is the classifier. Then, assuming a new subject, whose cytokines are known but whose pathologic status is not, starting from the root, answering to the tests regarding some cytokines values and following the correct branch accordingly, it arrives at a leaf that classifies the subject as pathologic or non-pathologic.

Since the cardinality of the database used in this study is low (n=20), we did not expect to build a definitive classifier but the idea was to use the resulting tree as a simple data mining technique in order to better understand the data under investigation and, eventually, disclose hidden differences or associations.

The outcome of the decision tree learning algorithm on the database is the classification tree, represented in Fig. 1. The tree found is a classifiers' tool. It classifies pathologic and non-pathologic subjects according to their value of Delta IL2 and Delta IFN-y. The outcome tree suggests for the populations investigated that: (i) By using the distinctive cytokine-response to Hsp60 treatment in HT and controls, it is possible to sort out the two populations, suggesting they differ in one or more biological-pathological properties; and (ii) Just knowing Delta IL2 and Delta IFN-y, it was possible to arrive at the classification indicated in (i) above.

This decision tree is build up starting from a low cardinality database (n=20), and it was not validated or verified with other sets of data. Therefore, it cannot be taken into account for a definitive application as a classifier. On the other hand, it can be seen as a useful clue to unveil the cytokine secretion response by PBMC to Hsp60 in the HT patients and healthy controls. In any case, it can be considered as a data mining tool, hence a possible source of information.

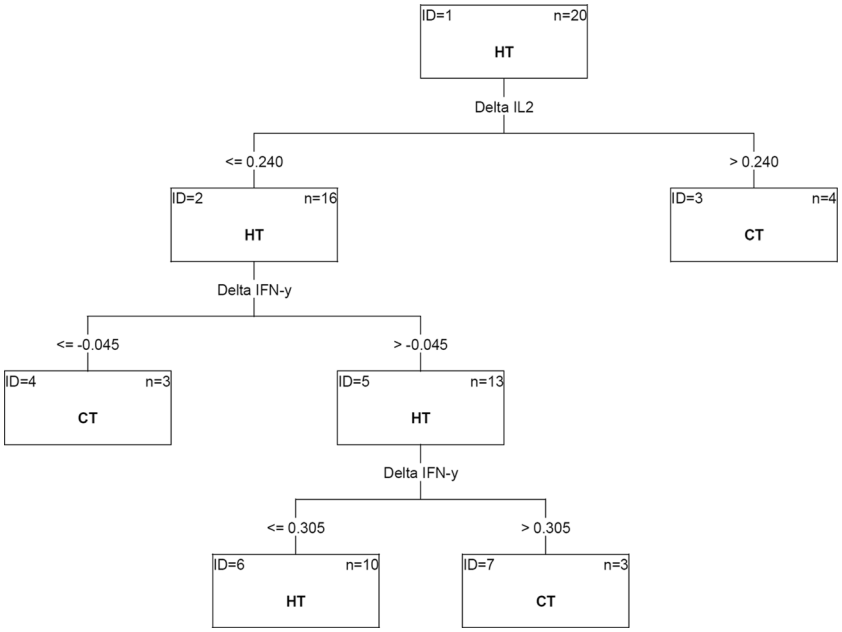

Fig. 1 The classification tree. Each *rectangle* is a node. In the upper left corner of each node indicated the identification number (*ID*) of the node. In the upper right corner, *n* indicates the number of elements (subjects) from the available database that were classified (or tested) at that node. The main label of a node, in the center of the rectangular box, identifies its class (*HT*, Hashimoto's thyroiditis; and *CT*, controls), while the class label of the leaves (node without daughters denotes the subject classification. The procedure starts from the root (ID=1), tests the cytokine labeled just under the root itself (Delta IL2), and follows the branch according to the answer specified on it. These steps are repeated until a leaf appears. The label of the leaf is the class of the subject being tested

The decision tree suggested us to consider Delta IL2 and IFN-y only: according to the tree, they are the main and sufficient biomarkers for classifying the pathologic population. So, we considered the two Delta values by looking at them in a scatterplot with the Delta-IFN-y in the X-axis and the Delta-IL2 in the Y-axis (Fig. 2). The distribution pattern clearly indicates that the pathological subjects are clustered together in a defined area with albescence of overlapping with non-pathological subjects. In conclusion, the scatterplot shows that HT subjects have smaller Delta absolute values of IFN-y and IL2 than healthy subjects, with cutoff values to be defined.

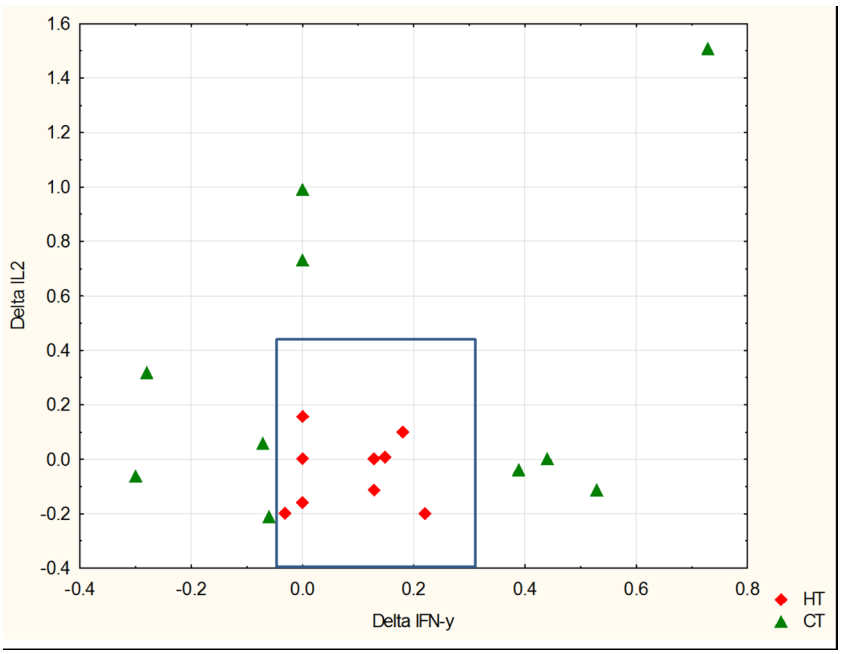

Fig. 2 The scatterplot. Scatterplot of Delta IFN-y (i.e., the difference between the IFN-y level after HSP60 treatment minus the level before the treatment) versus Delta IL2 (the difference between the IL2 level after Hsp60 treatment minus the level before the treatment). *Red rhombuses* represent HT subjects and *green triangles* represent controls. A *blue rectangle* frames the pathologic area in the scatterplot, namely, a cluster of HT subjects

To display the cluster formally, e.g., a rectangle-shaped cluster, we first chose the middle points between the closer data belonging to the opposite classes (i.e., HT and controls), and second, we determined the lines touching them and that are parallel to the X and Y axes (Fig. 2). In this way, the cluster depicted in Fig. 2 could be defined as follows:

$$-0.450 < \text{Delta IFN-y} < 0.305 \text{ and } \text{Delta IL2} < 0.445$$

In summary, the classification tree places the delta values of IFN-y and IL2 in a peculiar position. Then, drawing a scatterplot using the data suggested by the tree leads to an even clearer result: a cluster displayed in an easy-to-visualize graph showing that the two populations, HT and controls, differ from one another and are amenable to a distinctive classification.

Discussion and conclusions

The central aim of this pilot study was to determine the potential of non-conventional statistical methods for unveiling distinctive features that would identify Hashimoto's thyroiditis cases from individual not carrying the disease, according to the response of their PBMC to stimulation with Hsp60 in vitro. One added objective was to establish if the levels of cytokines in response to stimulation with Hsp60 can be considered good candidates as biomarkers for HT. We built a small database and performed a conventional statistical analysis but no significant results came out, confirming our previous observations (Marino Gammazza et al. 2014). Then, we used another approach, applying a data mining method involving classical artificial intelligence and machine learning algorithms named classification tree. The final outcome, a classification tree, was not considered a definitive classification tool but an investigation method (a data mining technique, in its deepest meaning). The initial result indicated that the search should focus on just two cytokines, IFN-y and IL2.

A scatterplot of the results pertaining to these two cytokines showed clusterization of the pathologic population (HT patients). It also showed that the populations investigated could be classified considering their IFN-y and IL2 reaction to Hsp60. The pathologic population showed a reduced reaction to Hsp60 treatment, considered as absolute value. One can hypothesize that this effect could be due to the fact that leucocytes obtained from HT patients are too much stressed to react to Hsp60 stimulation in the same manner as the leukocytes isolated from control subjects. However, further investigations on larger populations are necessary to establish the multiple possible uses of the type of analysis presented in this report. Nonetheless, the results reported here demonstrate that a data mining method coming from classical artificial intelligence and machine learning algorithms named classification tree can be used with benefits when the routine, standard statistical analyses fall short of providing definitive answers.

Acknowledgments This work was done under the umbrella of the agreement between the Euro-Mediterranean Institute of Science and Technology (IEMEST; Italy) and the Institute of Marine and Environmental Technology (IMET; USA) signed in March 2012. This paper is IMET contribution no. 14-137. AJLM, AMG, and FC were partially supported by IEMEST. LT, MC, and FG were partially supported by LUdeS.

References

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA

Marino Gammazza A, Rizzo M, Citarrella R, Rappa F, Campanella C, Bucchieri F, Patti A, Nikolic D, Cabibi D, Amico G, Conaldi PG, San Biagio PL, Montalto G, Farina F, Zummo G, Conway de Macario E, Macario AJL, Cappello F (2014) Elevated blood Hsp60, its structural similarities and cross-reactivity with thyroid molecules, and its presence on the plasma membrane of oncocytes point to the chaperonin as an immunopathogenic factor in Hashimoto's thyroiditis. Cell Stress Chaperones 19:343–353

Mikoś H, Mikoś M, Obara-Moszyńska M, Niedziela M (2014) The role of the immune system and cytokines involved in the pathogenesis of autoimmune thyroid disease (AITD). Endokrynol Pol 65:150–155

Richette P, Clerson P, Bouée S, Chalès G, Doherty M, Flipo RM, Lambert C, Lioté F, Poiraud T, Schaeverbeke T, Bardin T (2014) Identification of patients with gout: elaboration of a questionnaire for epidemiological studies. Ann Rheum Dis. 2014 May 5

Russell S, Norvig P (2009) Artificial intelligence: a modern approach. Pearson College Div., 3rd edition

Tsan MF, Gao B (2004) Cytokine function of heat shock proteins. Am J Physiol Cell Physiol 286:C739–C744

Wong HR, Lindsell CJ, Pettilä V, Meyer NJ, Thair SA, Karlsson S, Russell JA, Fjell CD, Boyd JH, Ruokonen E, Shashaty MG, Christie JD, Hart KW, Lahni P, Walley KR (2014) A multibiomarker-based outcome risk stratification model for adult septic shock. Crit Care Med 42:781–789