journal of autoimmunity

Journal of Autoimmunity 21 (2003) 65–75

www.elsevier.com/locate/issn/08968411

Cluster analysis of human autoantibody reactivities in health and in type 1 diabetes mellitus: a bio-informatic approach to immune complexity

Francisco J. Quintana[a], Gad Getz[b], Guy Hed[b], Eytan Domany[b], Irun R. Cohen[a*]

[a] *Department of Immunology, The Weizmann Institute of Science, Rehovot 76100, Israel*
[b] *Department of Physics of Complex Systems, The Weizmann Institute of Science, Rehovot 76100, Israel*

Received 30 October 2002; revised 11 March 2003; accepted 26 March 2003

Abstract

Informatic methodologies are being applied successfully to analyze the complexity of the genome. But beyond the genome, the immune system reflects the state of the body in health and disease. Traditionally, immunologists have reduced the immune system, where possible, to one-to-one relationships between particular antigens and particular antibodies or T-cell clones. Autoimmune diseases, caused by an immune attack against a body component, are usually investigated by following the response to single self-antigens. In this study, we apply informatics to analyze *patterns* of autoantibodies rather than single species of autoantibodies. This study was designed not to replace traditional approaches to immune diagnosis, but to test whether meaningful patterns of autoantibodies might exist. Using an unbiased solid-phase ELISA antibody test, we detected serum IgG and IgM antibodies in the sera of 20 healthy persons and 20 persons with type 1 diabetes mellitus binding to an array of 87 different antigens, mostly self-antigens. The healthy subjects manifested autoantibodies to a variety of self-antigens, many known to be associated with autoimmune diseases. We investigated the patterns of these autoantibodies using a coupled two-way clustering algorithm developed for analyzing data from gene arrays. We now report that the reactivity patterns of autoantibodies to particular subsets of self-antigens exhibited non-trivial structure, which significantly discriminated between healthy persons and persons with type 1 diabetes. The results show that despite the wide prevalence of autoantibodies, the patterns of reactivity to defined subsets of self-antigens can provide information about the state of the body.

Keywords: Bio-informatics; Autoantibodies; Immunological homunculus

1. Introduction

The great advances in information about the molecules and cells comprising the immune system have frustrated immunologists; the immune system is patently more complex than originally thought. Autoimmunity is a notable example of the problem. Traditionally, investigators and clinicians have focused on selected autoantibodies to study or diagnose specific autoimmune diseases [1–4]. They sought to establish a one-to-one relationship between a particular autoantibody and a particular disease. In practice, however, the presence of autoantibodies in healthy persons [5] complicates the serological diagnosis of autoimmune disease and confounds our understanding as to how the immune system actually discriminates the self from the non-self [6]. Immunology is in need of informatics.

For their part, complexity science people have not given much thought to the immune system, and have focused mostly on genomics or the nervous system. The immune system, however, is a suitable subject for informatics: like the central nervous system, the immune system is self-organizing [6–8]; unlike the central nervous system, the immune system is functionally accessible as a system at the cellular level both in vivo and in vitro [9].

The present study applies informatics to autoimmunity: we characterize a set of molecules recognized by

* Corresponding author. Tel.: +972-8-934-2911; fax: +972-8-934-4103.
E-mail address: irun.cohen@weizmann.ac.il (I.R. Cohen).

0896-8411/03/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0896-8411(03)00064-7

autoantibodies in healthy persons and test whether the global patterns of autoantibodies might discriminate between a state of health and an autoimmune disease, such as type 1 diabetes mellitus. This concept has been already tested in the past for other human autoimmune conditions using complex mixtures of undefined self-antigens [10–13]. However, our aim was to use an ELISA system able to detect even small amounts of low-affinity autoantibodies binding to an array of 87 different antigens of known identity without preconceived bias. We did not use reactivity thresholds, as is usually done to define a ‘negligible background’, and we did not restrict the detection to the specific high-affinity antibodies usually associated with autoimmune disease. For the purpose of this analysis, we define autoantibodies as antibodies that bind to self-molecules in the ELISA conditions used in this study. We imply nothing about function. To analyze the patterns of the autoantibodies, we applied a clustering algorithm and tested the statistical significance of the results. Particular sets of self-antigens, most of which are not known to be associated with type 1 diabetes, were found to discriminate between the patterns of autoantibodies of the healthy subjects and those of the type 1 diabetes patients. These results demonstrate that even the low-affinity autoantibody repertoire is structured and can yield information about the state of the body [14] when analyzed with suitable informatic tools.

2. Materials and methods

2.1. Antigens

The 87 antigens used in these studies are enumerated in Table 1. These antigens include proteins, peptides, nucleotides and phospholipids reported to interact with antibodies. The antigens are classified according to their cellular localization, tissue distribution or function.

2.2. Antibodies

The second antibodies used in the ELISA assay were F(ab′)2 goat anti-human IgG+IgM linked to alkaline phosphatase and goat anti-human IgM linked to horseradish peroxidase. These antibodies were purchased from Jackson ImmunoResearch Laboratories Inc. (West Grove, PA, USA), and were used at a final dilution of 1:1500 in bovine serum albumin 0.3%.

2.3. Test samples

Serum samples were collected at the Hadassah Medical Center (Jerusalem, Israel), under the supervision of Dr Rivka Abulafia-Lapid and Professor Itamar Raz, from 20 healthy young-adult blood donors, with no family history of diabetes, and from 20 unselected type 1 diabetes patients. Most of the type 1 diabetes patients, too, were young adults, 21–34 years old (95% confidence interval; median, 23 years). The diagnosis was made on the basis of accepted clinical criteria: hyperglycemia, ketonuria, low body weight, the absence of a family history of type 2 diabetes and a standard (anti-glutamic acid decarboxylase (GAD)) antibody assay [3]. The HLA genotypes were not tested. The sera were collected within 4–7 weeks of diagnosis (95% confidence interval; median, 6 weeks). Informed consent was obtained. The samples were stored at −20 °C without any additive. The T-cell proliferative responses of these type 1 diabetes patients and healthy blood donors had been studied previously. No significant difference was found between the groups in their T-cell responses to the foreign antigen tetanus toxoid, although the diabetic subjects manifested heightened responses to the 60 kDa heat shock protein (HSP) self-antigen [15].

2.4. Solid-phase antibody assay

A standard ELISA assay was used. Antigens (10 µg/ml in phosphate-buffered saline (PBS)) were coated in 96-well ELISA plates (Maxisorp; Nunc, Roskilde, Denmark) by overnight incubation at 4 °C. The plates were washed with PBS 0.05% Tween, and blocked for 2 h with bovine serum albumin 3% (Sigma, Rehovot, Israel). The serum samples were diluted 1:100 in bovine serum albumin 0.3%, and 50 µl was added to each well. After 3 h of incubation at 37 °C, the sera were removed and the plates were washed with PBS 0.05% Tween. Bound antibodies were detected with an appropriate alkaline phosphatase or horseradish peroxidase-conjugated second antibody (Jackson ImmunoResearch Laboratories Inc.), 50 µl incubated for 1.5 h at 37 °C. The plates were washed with PBS, and *p*-nitrophenol phosphate or 2,2′-azino-bis (3-ethylbenzthiazoline-6 sulfonic acid; both from Sigma) were added, and the optical density (OD) in each well was read at 405 nm using a spectrophotometer.

We optimized the conditions of the assay: a direct correlation between the OD readings and dilutions of the sera were found between 1:50 and 1:200 dilutions. Accordingly, we chose 1:100 as the standard dilution of the test sera. The relationship between the OD and the incubation time was linear during 45 min of incubation (mean $r^2 \pm$ standard deviation (SD)=0.98 ± 0.02). Therefore, we recorded the OD readings 30 min after the addition of the substrate. The assay was reproducible: the mean intra-assay coefficient of variation was 4.3%, and the mean inter-assay coefficient of variation was 9.5%. Correlation analysis of intra- and inter-assay variations yielded r^2 coefficient values 0.98 and 0.96, respectively, with P value <0.0001 for both.

2.5. Cluster analysis

The OD readings corresponding to the antibody reactivities of a group of N serum samples against a panel of M=176 different reactivities (87 antigens and one blank with two secondary antibodies) were placed in a matrix A, whose element A_{js} represents the extent to which the serum of subject s reacted with test antigen j (the secondary antibody was absorbed in the index j). The 'immune state' of subject s is represented by a vector $\mathbf{A}^{(s)}$ (of M components). Similarly, antigen j is represented by the (N component) vector $\mathbf{A}^{(j)}$.

The following normalization was used and done only once, for k=1,...,88 (IgM) and k=89,...,176 (IgM+IgG):

$$B_{js}=\log A_{js}-\frac{2}{M}\sum\nolimits_k \log A_{ks} \tag{1}$$

For j=1,2,...,88, the sum over k is from 1 to 88, and for j=89,90,...,176, the sum is from 89 to 176. For every subject s, this operation produces a mean-centered set of values; if all readings A_{js} are multiplied by a constant, it does not affect the B variables. We take the log, since the noise on the readings is multiplicative, and we mean center the variables to eliminate dependence on concentration fluctuations from subject to subject.

Next, we renormalized the data by subtracting from each element the average value of the elements in the same matrix row (corresponding to a particular antigen) and dividing by the SD of the row. The elements of the resulting renormalized submatrix are denoted by G_{js}—for each antigen, the mean of G_{js} vanishes and the sum of squares is 1. We analyzed the data using the method of Getz et al. [16] for analysis of gene expression. Thus, we identify subsets of K serum samples and cluster them on the basis of their reactivities to a selected subset of antigens. In this way, the analysis uses various submatrices of the total data matrix G, described in detail in Ref. [16]. Briefly, we first cluster all antigens (using data from all sera) and identify stable antigen clusters. Next, we cluster the sera using, one at a time, the groups of antigens that emerged as stable clusters in the first step.

To assign related antigens to the same cluster, we singled out 'close' pairs of highly correlated antigens, as well as pairs that were highly anti-correlated. This 'closeness' is measured by the 'distance' $d_{j,l}$ between antigens j and l, given by

$$(d_{j,l})^2=1-c_{j,l}^2 \tag{2}$$

where $c_{j,l}$ is the correlation coefficient of antigens j and l, as measured over the N samples, given by

$$c_{j,l}=\sum_{s=1}^{N} G_{js}G_{ls} \tag{3}$$

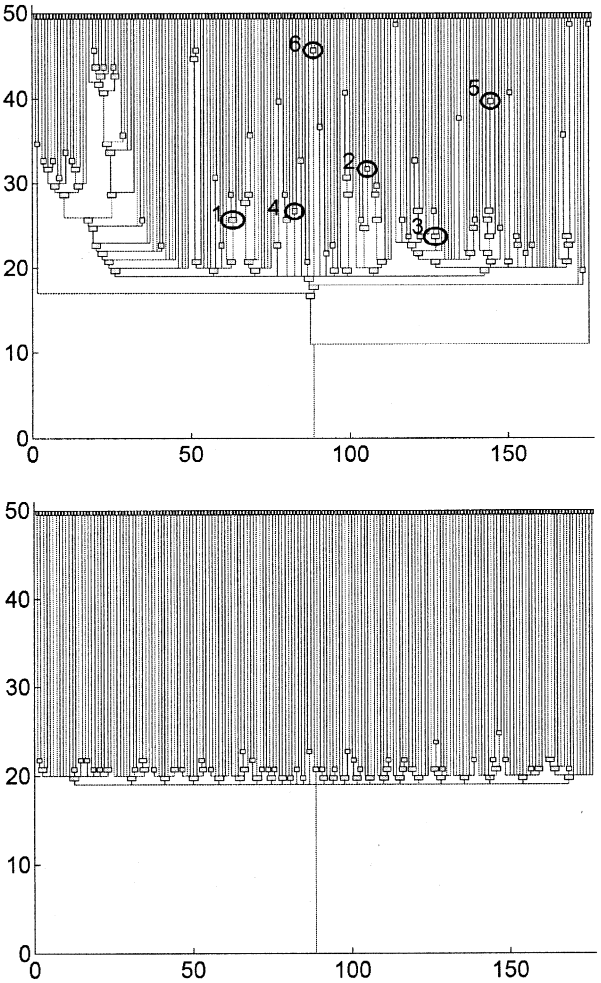

Fig. 1. Dendrograms of antigens obtained by clustering. (A) Dendrogram obtained from the original data matrix, using sera from healthy and type 1 diabetes subjects; the antigen clusters that are reported in Table 3 are circled and numbered. (B) Dendrogram of the antigens obtained by clustering a randomized matrix.

In contrast, the distance D_{sp} between subjects s and p is the Euclidian distance

$$D_{sp}^2=\sum\nolimits_j (G_{js}-G_{jp})^2 \tag{4}$$

These distance measures reflect similarity between pairs of subjects and pairs of antigens. Since the clustering method in this study relies heavily on proximity of pairs of points rather than wide separations, the results are only slightly sensitive to the precise measure used in Eq. (4).

We used an unsupervised clustering technique, the SPC clustering algorithm [17], which organizes the data in the form of a dendrogram, such as that shown in Fig. 1A, B. As a control parameter T increases to a value $T_1(C)$, a cluster C may be 'born' (when its 'parent' cluster breaks up into two or more subclusters, one of which is C). As T increases further, to $T_2(C)$>$T_1(C)$, C

itself breaks up and 'dies'. A main advantage of SPC is that it provides a quantitative stability index, $R(C)=T_2(C)/T_1(C)$ for any cluster C. The larger the value of $R(C)$, the more statistically significant and stable (against noise in the data and fluctuations) is C. We used SPC within a coupled two-way clustering approach [16] to identify subsets of serum samples and of antigens, allowing meaningful partitions of the samples to emerge. The clinical labels were then used to *evaluate* the results (not to produce them). If a cluster of serum samples contained predominantly subjects with the same diagnosis, the cluster's predictive capacity was estimated. The effectiveness of the resulting classification was measured in terms of the *number of classification errors*, N_e that were made by assigning the subjects of the cluster to one diagnosis and the rest of the subjects to the other diagnosis, healthy or type 1 diabetes. We evaluated *specificity* and *sensitivity*: *specificity* is the fraction of correctly diagnosed subjects present in a cluster (=1 for no false positives); *sensitivity* is the fraction of correctly diagnosed subjects that were included in the cluster, of the total number of subjects with the same diagnosis (=1 for no false negatives).

2.6. Combining classifiers

The sensitivity and specificity of 'final' classifier in this study were improved by combining several different sets of antigens. We classified a test sample as type 1 diabetes, for example, if it was so discriminated by a majority of the classifiers. Then we identified the samples, and evaluated the specificity and sensitivity of the combined classifiers.

2.7. Assessing statistical significance

Statistical analysis was carried out to test four questions: *first*, whether the pattern of antibody reactivity exhibited a non-trivial structure; *second*, how sensitive were the subject clusters to variations in the data—such as leaving out one subject; *third*, whether the clinical state of the subjects was reflected by their reactivity patterns; and *fourth*, how robust was the method, whether it is able to predict the clinical status of subjects with unknown clinical labels.

For the *first* question, we used the original data matrix and randomized all its entries, placing each in a random location. The randomized matrix was normalized and clustered; for every cluster C, the size (number of elements) $n(C)$ and its stability index $R(C)$ were recorded. For each n, we identified $R^*(n)$, the maximal value of R. This was repeated for *1000* random matrices and; for each size n, a histogram $P_n(R^*(n))$ was prepared, and the SD σ_n and maximal value $R^*_{\max}(n)$ were determined. The maximal values of R, which were found for the stable clusters of size n obtained from the real data, were compared with the extremal values found for the randomized matrices.

For the *second* question, we repeated the clustering of the subjects 40 times, leaving out one subject in each trial, and checked the effect on the resulting clusters.

For the *third* question, the P-values of the diagnostic labels were estimated by calculating the probability, P, that a previously selected 'discriminating' cluster of C subjects would produce e errors for randomly assigned clinical labels, with $e \leq N_e$, where N_e is the number of 'errors' for the real labels (of S diagnosed subjects). A high value of P indicates that the discrimination produced by this cluster is not related to the diagnosis in a statistically significant way. This probability $P(e \leq N|C,S,N)$ was determined by the Fisher's exact two-sided test [18].

For the *fourth* question, we simulated a situation in which the clinical labels of 30 subjects (15 diseased, 15 healthy) were known and the diagnosis of the remaining 10 subjects was hidden. All 40 subjects were clustered as described previously, using, one at a time, the reactivities of each one of the six antigen clusters shown in Table 3. In the resulting dendrogram of subjects (obtained for each one of the six antigen clusters), we identified the stable clusters of subjects. These clusters were candidates to serve as classifiers; among them, we selected the one with the lowest number of errors based on the 30 '*known*' labels. This subject cluster was then chosen as a classifier. Next, the 10 'unknown' samples were diagnosed on the basis of their affiliation with the classifier cluster. This process was repeated for each one of the six antigen clusters enumerated in Table 3, using 100 random choices of 10 subjects with hidden labels. In this way, we could determine the ability of classifiers, that were constructed using 30 labeled subjects, to correctly identify 10 'unknowns'.

2.8. Principal component analysis

For comparison, the data were also analyzed by standard principal component analysis (PCA) [19].

3. Results

3.1. Serum autoantibodies in healthy blood donors

We tested the repertoire of autoantibodies in the sera of 20 healthy individuals using an array of 87 different antigens (listed in Table 1) and detection antibodies directed to IgM or IgG+IgM. Table 2 summarizes the self-antigens most frequently recognized by the autoantibodies of the 20 healthy blood donors. For comparison, Table 2 also contains the bacterial antigens lipopolysaccharide (LPS) and purified protein derivative (PPD) of *Mycobacterium tuberculosis*. Thus, healthy persons may express a wide range of autoantibodies detectable by ELISA. To learn whether the patterns of such autoantibodies might be informative, we applied our clustering analysis to the healthy subjects and to a population of persons with the autoimmune disease type 1 diabetes mellitus.

3.2. Structure in the repertoire: self-antigen clusters

First, we clustered the antigens and then used the various antigen clusters as probes to cluster the subjects. To cluster the 87 antigens, we used the serum OD readings for all subjects (healthy or not). The serum reactivity data (for 2 isotypes × 87 antigens) exhibited a non-trivial structure, as is evident from comparing the dendrogram of Fig. 1A (obtained by clustering the original antigen matrix) with that of Fig. 1B (obtained by clustering a randomized matrix). The dendrograms obtained from the randomized matrices exhibited a sudden ‘melt down’ from a single cluster that contained all the points to small clusters with very low stability. The dendrogram of the actual data, in contrast, contained a cluster of five antigens, with stability index R=1.33. We tested 1000 different realizations of randomized matrices, and determined for each cluster size n the corresponding extremal value $R^*_{\max}(n)$ and SD σ_n. The real data yielded stable clusters of size n, whose stability indices R exceeded the corresponding random extremal value $R^*_{\max}(n)$ by at least $3\sigma_n$ (data not shown). Thus, the P-value for the presence of non-trivial structure in the antigen clusters was less than 0.001. Hence, groups of self-antigens do cluster together as collectives [7]; sera that react with one member of an antigen cluster will tend to react with other members of the antigen cluster.

3.3. Clusters as classifiers of type 1 diabetes

The identified subsets of antigens were then used to probe the sera of 20 healthy donors and 20 diabetes patients. Fig. 2 shows the dendrogram obtained using the IgM reactivities to insulin and to collagen I and the IgG+IgM reactivities to collagen I. This set of antigens generated a sensitivity of 85% and a specificity of 81% for diabetes. Other sets of antigens generated different dendrograms. Table 3 summarizes the findings and includes the P-values calculated for each cluster size C and error number N_e, using Fisher’s exact two-tailed test. This result shows the advantages of our methodology versus linear discrimination. If one distributes randomly two labeled groups, of 20 points in each, in 176 dimensional space, a separating linear manifold will be found with probability very close to 1 [20,21]. However, as we have shown, our cluster analysis definitely does not separate two randomly placed groups of points. Finally, we repeated the clustering process 40 times using the reactivities of antigen cluster 1 of Table 3, each time leaving out another subject. The resulting clusters were stable despite the omissions; 17 subjects appeared in all 40 trials in the cluster of the diabetes patients; 16 were indeed diseased and one was healthy.

The combined results produced an overall sensitivity of 95% and a specificity of 90%. Among the self-antigens that discriminated between type 1 diabetes and healthy sera were cardiolipin, collagen I, collagen X, cytochrome *c* P450, cartilage extract (a commercial preparation rich in collagen I), aldolase, acetylcholine receptor (AchR), heparin and insulin. Among this list of molecules, only insulin has been noted previously to be a self-antigen in type 1 diabetes [22]. Neither GAD nor the 60 kDa HSP appeared among the discriminatory antigen clusters, although both self-antigens have been implicated in type 1 diabetes in other assays [3,15].

Using the antigen clusters shown in Table 3, we chose classifier subject clusters (using 30 known labels) and tested the 10 ‘unknown’ subjects, obtaining the following results: for antigen cluster 1, all 100 trials selected the same ‘original’ subject cluster as in Table 3. For antigen cluster 2, the original subject cluster was selected in 78 trials, but two other clusters were also picked up, both 11 times. For antigen cluster 3, we found the original cluster 73 times, but two other clusters were also picked up, 22 and five times. For antigen cluster 4, the scores were 95 and 5. For antigen clusters number 5 and 6, the score was 100 selections of the respective original subject clusters.

In each one of the 100 simulations, we diagnosed the 10 ‘unknown’ samples on the basis of their affiliation with the selected classifier clusters, and for each simulation, used the majority rule described previously to combine the results obtained for the six antigen clusters. We found the following error distribution: 29 occurrences with 0 errors (for 10 predictions), 45 with 1 error, 21 with 2 errors and 5 with 3 errors. The average number of errors for the 10 ‘unknown’ subjects was 1.02; this is only slightly worse than the 3 errors obtained for the 40 subjects, using all their labels.

The PCA [19] has been regularly used to study patterns of autoantibodies to undefined antigens in tissue blots [23,24]. Applying PCA to our data, we identified the eigendirections associated with the leading eigenvalues; each eigendirection had significant projections onto more than 10 antigens. Next, we projected the data onto the plane spanned by the two leading directions. The best linear separator, found using non-normalized data, generated a comparable number of errors to that of antigen cluster 1 in Table 3 (data not shown). The PCA finds about 20 antigens that have sizeable contributions to the two leading principal directions. The present clustering method, in contrast, yields separation in several two- or three-dimensional spaces that are related to a few antigens.

Table 1
Antigens used (the catalogue number is given for those molecules purchased from Sigma)

Group	Function/structure	Number	Antigen	Sequence (when applicable)	Catalogue
Cellular structure	Cytoskeleton	1	Actin		A3653
		2	Tubulin		T4925
		3	Myosin		M6643
		4	Tropomyosin		T4770
		5	Vimentin		V4383
	Extracellular matrix	6	Fibronectin		F0895
		7	Collagen I		C7774
		8	Collagen II		C7806
		9	Collagen III		C4407
		10	Collagen IV		C7521
		11	Collagen V		C3657
		12	Heparin		H2149
		13	Laminin		L6274
		14	Collagenase		C9891
Cellular membranes	Phospholipids	15	Cardiolipin		C5646
		16	Glucocerebroside		G9884
		17	Phosphatidylethanolamine		P9137
		18	Cholesterol		C1145
Cellular metabolism	Glucose	19	Enolase		E0379
		20	Aldolase		A8811
		21	Acid phosphatase		P1774
	Apoptosis	22	Annexin 33 kDa		A9460
		23	Annexin 67 kDa		A2824
		24	Cytochrome *c* P450		C3131
	Monooxigenases	25	Catalase		C9322
		26	Peroxidase		P6782
		27	Tyrosinase		T7755
	Others	28	Ribonuclease		R4875
Nucleus	Protein	29	Histone II A		H9250
	DNA	30	Double-stranded DNA		D1501
		31	Single-stranded DNA		D1501
Plasma proteins	Carriers	32	Transferrin		T4132
		33	Fetuin		F2379
		34	Human serum albumin		A8763
		35	Bovine serum albumin		A9647
		36	Ovoalbumin		A5378
	Coagulation	37	Factor II		F5132
		38	Factor VII		F6509
		39	Fibrin		F5386
		40	Fibrinogen		F4883
	Complement	41	C 1		C2660
		42	C 1 q		C0660
Immune System	Cytokines	43	Interleukin 2		I2644
		44	Interleukin 10		I9276
		45	Interleukin 4		I4269
	Immunoglobulins	46	IgG		I8640
		47	IgM		I8260
		48	1E10 Fab[a]		
	TCR peptides	49	N4	ASSLWTNQDTQY	NA
		50	C9	ASSLGGNQDTQY	NA
Tissue antigens	Heat shock protein	51	HSP60[b]		
		52	p277	VLGGGVALLRVIPALDSLTPANED	NA
	Islet antigens	53	GAD		G2126
		54	Insulin		I0259
	CNS	55	Human MOG[c]		
		56	Murine MOG[c]		
		57	Human MOG p94–116[c]	GGFTCFFRDHSYQEEAAMELKVE	
		58	Rat MOG p35–55[c]	MEVGWYRSPFSRVVHLYRNGK	
		59	MBP[d]		
		60	Brain extract		B1877
	Muscle and skeleton	61	AchR[e]		
		62	Myoglobulin		M6036

Table 1 (*continued*)

Group	Function/structure	Number	Antigen	Sequence (when applicable)	Catalogue
Tissue antigens	Joints	63	Cartilage extract		C5210
	Thyroid	64	Thyroglobulin		T1001
	Blood cells and platelets	65	Hemoglobin A		H0267
		66	Spectrin		S3644
Foreign antigens	Proteins and peptides	67	TB PPD[f]		
		68	HSP65[g]		
		69	ecp27	KKARVEDALHATRAAVEEGV	NA
		70	mtp278	EGDEATGANIVKVALEA	NA
		71	GST[b]		
		72	KLH[h]		
		73	Pepstatin		P5318
		74	R13	EEEDDDMGFGLFD	NA
	Others	75	LPS		L3755
Synthetic polymers	Poly amino acids	76	Poly arginine		P3892
		77	Poly lysine		P4408
		78	Poly aspartic		P6762
		79	Poly glutamate		P4636
	Oligonucloetides	80	PolyA	A_{20}	NA
		81	PolyT	T_{20}	NA
		82	PolyC	C_{20}	NA
		83	PolyG	G_{20}	NA
		84	PolyATA	$AT_{18}A$	NA
		85	PolyTAT	$TA_{18}T$	NA
		86	CpG	TCCATGACGTTCCTGACGTT	NA
		87	GpC	TCCAGGACTTCTCTCAGGTT	NA

[a] Fab fraction generated from a monoclonal antibody directed to peptide p277.

[b] Recombinant protein expressed in bacteria and purified using standard procedures.

[c] Kindly provided by Professor Avraham Ben Nun (The Weizmann Institute of Science).

[d] Kindly provided by Dr Felix Mor (The Weizmann Institute of Science).

[e] Kindly provided by Professor Sara Fuchs (The Weizmann Institute of Science).

[f] Produced at the Statens Seruminstitut, Copenhagen, Denmark.

[g] Kindly provided by Professor R. van deer Zee (Utrecht University, The Netherlands).

[h] Purchased from Pierce (Oud Beijerland, The Netherlands), catalogue number 77153.

4. Discussion

The results reported in this article relate to general issues in both biology and informatics. In general, this study confirms that unselected patterns of reactivity can be informative [11,12,23–32] not only in the nervous system, but also in immunology, where the emphasis has been traditionally on specific, high-affinity antibody molecules and single clones of cells. Complex systems use arrays of signals to generate information, and biologists too have to exploit arrays of information. The highly non-trivial task of mining a fairly large array of antigen reactivities for different subjects was accomplished by an unsupervised clustering methodology that identifies (relatively small) correlated groups of antigens, whose reactivities may be used to separate the subjects according to known clinical labels (which are used only *a posteriori*). This method starts from an unsupervised 'holistic' approach that looks at a large number of antigens, and proceeds on a reductionist path, identifying small subsets of antigens that may be relevant to some particular differentiation.

Immunologically, we found that the autoantibodies of the healthy subjects detectible by ELISA bound to many self-antigens implicated in autoimmune diseases (Table 2): histone II A, and single- and double-stranded DNA—targeted in systemic lupus erythematosus [2]; HSP60, insulin and GAD—associated with type 1 diabetes mellitus [22]; myelin basic protein (MBP) and myelin oligodendrocyte glycoprotein (MOG)—associated with multiple sclerosis [33]; the AchR—targeted in myasthenia gravis [1]; tyrosinase—associated with vitiligo [34]; myosin—associated with polymyositis [35]; and cytochrome *c* P450—associated with autoimmune liver disease [36]. Other frequent self-antigens included the serum proteins fibrinogen and clotting factor VII, heparin, enzymes and globin molecules. The present study did not deal with the precise specificity and affinity of the individual antibodies, as is usually done to investigate autoimmunity.

The documentation of autoantibodies binding to self-antigens in healthy people is compatible with the concept of the immunological homunculus [7,37]. The term immunological homunculus refers to the observation

Table 2
Frequencies of autoantibodies in healthy humans (to limit the number of self-antigens shown, the only those antigens are included to which at least 35% of the healthy subjects responded with an OD of greater than 0.3 nm)

Antigen	Incidence of autoantibodies (%)	
	IgM+IgG	IgM
Tubulin	50	–
Myosin	40	–
Heparin	75	–
Acid phosphatase	35	–
Annexin 33 kDa	55	–
Cytochrome *c* P450	50	80
Catalase	65	–
Tyrosinase	–	45
Histone II A	65	45
Double stranded DNA	75	75
Single stranded DNA	100	95
Factor VII	70	100
Fibrinogen	90	–
HSP60	40	–
GAD	100	70
Insulin	35	35
MOG	–	95
MBP	35	–
AchR	90	75
Myoglobulin	65	35
Hemoglobin A	50	45
LPS	85	45
TB PPD	90	50

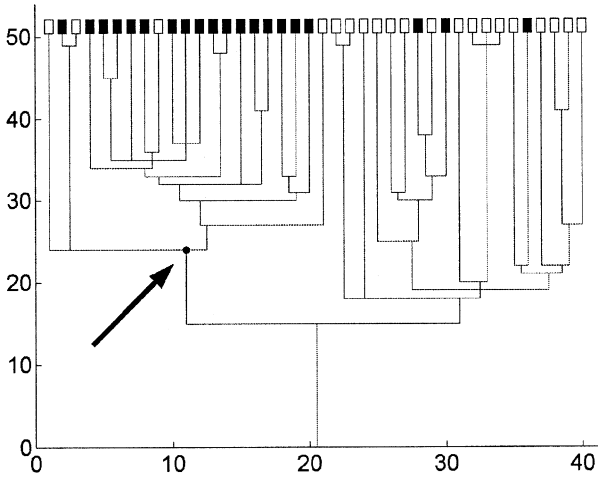

Fig. 2. Dendrogram of healthy subjects and type 1 diabetes subjects. Clustering was done using IgM reactivities to insulin, and IgM and IgM+IgG for collagen I. Healthy subjects are represented by white squares, and type 1 diabetes patients are represented by black squares. We used the cluster marked by the arrow to classify the subjects.

that T- and B-cell autoimmunity in healthy individuals is usually organized around particular sets of self-antigens [38,39]. Homunculus theory proposes that this natural autoimmunity is regulated by various mechanisms that prevent the transition of healthy autoimmunity to autoimmune disease [7,37,40–43]. Indeed, the development of autoimmune disease could be explained most simply by the failure of these control mechanisms [6,7]. The high prevalence of certain autoimmune reactivities shown in Table 2 might explain why the major autoimmune diseases are associated with the abnormal activation of just these autoimmune reactivities; they are already built into the healthy system [7,37]. The autoimmune disease process would seem to expand the quantity and select for high affinity of the particular autoantibodies involved in the disease. The natural autoimmune repertoire, which is quiescent in the healthy state, might serve as the ground for this pernicious autoimmunity. Clearly, we would like to know how natural autoimmunity to particular sets of self-antigens develops, what functions healthy autoimmunity might serve in body maintenance [44,45], how natural autoimmunity is controlled and how it deteriorates into autoimmune disease in certain persons [41,46]. The present study did not explore these biological questions, but rather focused on the informatic questions: (a) whether there are non-trivial structures and correlations in the reactivity patterns of autoantibodies and (b) whether the pattern of autoantibodies present in healthy persons might be distinguished from the pattern of autoantibodies present in an autoimmune disease, taking type 1 diabetes mellitus as our example.

Clinically, diagnostic tests for type 1 diabetes are constructed and standardized in a way that takes advantage of the greater amounts and higher affinities of the autoantibodies that are produced when an autoimmune disease process becomes activated; in clinical testing, the natural autoantibodies we detected, presented in Table 2, are buried in the background [47,48]. The association of type 1 diabetes with antibodies to GAD, an accepted assay, is observed, for example, only when the GAD antibody assay is done according to a standard protocol [3]. Our aim in this study, in contrast to standard procedure, was not to analyze different types of patients and the quantities or affinities of their particular autoantibodies, but to test for the presence of informative patterns in the global array of autoantibodies. This approach has been used in the past for the study of several autoimmune conditions in humans [10–13]. However, these previous studies did not use large panels of defined self-antigens and were not applied to human type 1 diabetes mellitus.

Coutinho and colleagues pioneered the study of patterns of autoantibodies binding to undefined antigens in blots of tissue extracts [23,24]; they used PCA, alone or in combination with hierarchical clustering [23], to study their results. The present work extends the study of patterns to defined self-antigens and is based on a novel cluster analysis [16] that identifies small groups of antigens whose reactivity patterns reflect a subject's clinical state. In this study, we show that healthy subjects and type 1 diabetes subjects can be distinguished, despite the

Table 3
Clustering of type 1 diabetes and healthy human serum samples (the antigen clusters shown in Fig. 1A were used to classify the subjects)

Cluster number	Antigens	Sensitivity (%)	Specificity (%)	Cluster Size	Number of errors	*P*-value
1	IgM to collagen I IgG+M to collagen I IgG+M to insulin	85	81	21	7	8.8×10^{-5}
2	IgM to hAchR IgM aldolase	85	74	23	9	0.0011
3	IgM to cartilage extract IgG+M to cardiolipin IgM to cardiolipin	55	100	11	9	0.00015
4	IgM to poly arginine IgM to heparin	80	70	23	11	0.0095
5	IgM to collagen X IgG+M collagen X	95	63	30	12	0.0084
6	IgM to cytochrome *c* P450 IgG+M cytochrome *c* P450	70	67	21	13	0.056
	Combination (more than three classifiers)	95	90			

presence of various autoantibodies in both groups, by the patterns of their autoantibodies. Furthermore, the patterns of reactivity appear to be disease-specific; type 1 diabetes subjects could be efficiently separated from persons with type II diabetes or Behçet's disease (manuscript in preparation). In other words, patterns of autoantibody reactivity may provide information beyond that seen in a simple one-to-one relationship between an antibody and an antigen [7,11,12,23–32,49]. The present work extends previous findings made in humans (reviewed in Refs. [39,50]) and introduces a novel two-step approach of first clustering the antigens and then using antigen clusters to cluster the subjects.

It is to be noted that we estimated the statistical significance of the clusters by empirically testing the frequency with which similar results could arise by chance using scrambled data. One thousand such computer experiments proved the significance of the antigen clusters derived from the real data. Furthermore, the statistical significance of the separation of diseased and healthy subjects, based on our clusters, was calculated analytically [18]. Finally, the clustering proved robust: the clinical labels of 10 'unknown' subjects could be clustered on the basis of the remaining 30, with a success rate of 90%.

We do not yet know the biological relevance of the particular autoantibodies or of their patterns to the pathophysiology of disease. It is to be noted that the discrimination between the healthy blood donors and the type 1 diabetes patients by clustering does not mean that the informative antibody reactivities are directly involved in the disease process. The differences in autoantibody patterns could have resulted from the disease itself, or from genetic or environmental factors associated directly or indirectly with susceptibility to the disease. Even factors such as age, gender and immunization history were not controlled. Nevertheless, the present findings fit the renewed appreciation of the importance of collective patterns in living systems. Collective interactions that form distinct reactivity patterns bear meaning in signal transduction, gene activation, neoplastic transformation, cell movement, organogenesis, brain function and almost any other subject presently of interest to biologists [51–57]. Biology has succeeded in reducing complex systems to component cells and molecules, but the emergent properties of living systems cannot easily be reduced to the one-to-one relationships of single components; informatic analysis of arrays of data is required. Like other complex systems, the immune system can be mined for information by studying arrays of data. Indeed, the repertoire of autoantibodies present in the individual is much closer to the individual's life experience than are the individual's genes. The immune system, like the brain, is an adaptive bio-informatic system in its own right [58].

Acknowledgements

I.R.C. is the incumbent of the Mauerberger Chair in Immunology, and Director of the Center for the Study of Emerging Diseases. E.D. is the incumbent of the H. J. Leir Professorial Chair. His research was partially supported by GIF—the Germany–Israel Science Foundation. We thank Yeda Ltd, Dr Isaac Shariv CEO, for funding this study.

References

[1] Al-Lozi M, Pestronk A. Organ-specific autoantibodies with muscle weakness. Curr Opin Rheumatol 1999;11:483–8.

[2] Pisetsky DS. Anti-DNA and autoantibodies. Curr Opin Rheumatol 2000;12:364–8.

[3] Verge CF, Stenger D, Bonifacio E, Colman PG, Pilcher C, Bingley PJ et al. Combined use of autoantibodies (IA-2 autoantibody, GAD autoantibody, insulin autoantibody, cytoplasmic islet cell antibodies) in type 1 diabetes: combinatorial islet autoantibody workshop. Diabetes 1998;47:1857–66.

[4] Zauli D, Cassani F, Bianchi FB. Auto-antibodies in hepatitis C. Biomed Pharmacother 1999;53:234–41.

[5] Avrameas S, Guilbert B, Dighiero G. Natural antibodies against tubulin, actin, myoglobin, thyroglobulin, fetuin, albumin and transferrin are present in normal human sera, and monoclonal immunoglobulins from multiple myeloma and Waldenstrom's macroglobulinemia may express similar antibody specificities. Ann Immunol 1981;132:231–6.

[6] Cohen IR. Discrimination and dialogue in the immune system. Semin Immunol 2000;12:215–9.

[7] Cohen IR. Tending Adam's garden: evolving the cognitive immune self. London: Academic Press, 2000.

[8] Coutinho A, Forni L, Holmberg D, Ivars F, Vaz N. From an antigen-centered, clonal perspective of immune responses to an organism-centered, network perspective of autonomous activity in a self-referential immune system. Immunol Rev 1984; 79:151–68.

[9] Abbas AK, Lichtman AH, Pober JS. Cellular and molecular immunology. 2nd ed. Philadelphia: W.B. Saunders, 1994.

[10] Sharshar T, Lacroix-Desmazes S, Mouthon L, Kaveri S, Gajdos P, Kazatchkine MD. Selective impairment of serum antibody repertoires toward muscle and thymus antigens in patients with seronegative and seropositive myasthenia gravis. Eur J Immunol 1998;28:2344–54.

[11] Ronda N, Haury M, Nobrega A, Kaveri SV, Coutinho A, Kazatchkine MD. Analysis of natural and disease-associated autoantibody repertoires: anti-endothelial cell IgG autoantibody activity in the serum of healthy individuals and patients with systemic lupus erythematosus. Int Immunol 1994;6:1651–60.

[12] Sundblad A, Ferreira C, Nobrega A, Haury M, Ferreira E, Padua F et al. Characteristic generated alterations of autoantibody patterns in idiopathic thrombocytopenic purpura. J Autoimmun 1997;10:193–201.

[13] Stahl D, Lacroix-Desmazes S, Heudes D, Mouthon L, Kaveri SV, Kazatchkine MD. Altered control of self-reactive IgG by autologous IgM in patients with warm autoimmune hemolytic anemia. Blood 2000;95:328–35.

[14] Coutinho A, Avrameas S. Speculations on immunosomatics: potential diagnostic and therapeutic value of immune homeostasis concepts. Scand J Immunol 1992;36:527–32.

[15] Abulafia-Lapid R, Elias D, Raz I, Keren-Zur Y, Atlan H, Cohen IR. T cell proliferative responses of type 1 diabetes patients and healthy individuals to human hsp60 and its peptides. J Autoimmun 1999;12:121–9.

[16] Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. Proc Natl Acad Sci U S A 2000; 97:12079–84.

[17] Blatt M, Wiseman S, Domany E. Super-paramagnetic clustering of data. Phys Rev Lett 1996;76:3251–5.

[18] Fisher RA. The logic of inductive inference. J R Stat Soc 1935; 98:39–82.

[19] Duda OR, Hart PE, Stork DG. Pattern classification. 2nd ed. New York: Wiley, 2001.

[20] Cover TM. Geometrical and statistical properties with application in pattern recognition. IEEE Trans Electr Comput 1965; 14:326–34.

[21] Gardner E. Maximum storage capacity in neural networks. Europhys Lett 1987;4:471–85.

[22] Tisch R, McDevitt H. Insulin-dependent diabetes mellitus. Cell 1996;85:291–7.

[23] Nobrega A, Haury M, Grandien A, Malanchere E, Sundblad A, Coutinho A. Global analysis of antibody repertoires. II. Evidence for specificity, self-selection and the immunological 'homunculus' of antibodies in normal serum. Eur J Immunol 1993;23:2851–9.

[24] Haury M, Grandien A, Sundblad A, Coutinho A, Nobrega A. Global analysis of antibody repertoires. 1. An immunoblot method for the quantitative screening of a large number of reactivities. Scand J Immunol 1994;39:79–87.

[25] Haury M, Sundblad A, Grandien A, Barreau C, Coutinho A, Nobrega A. The repertoire of serum IgM in normal mice is largely independent of external antigenic contact. Eur J Immunol 1997; 27:1557–63.

[26] Malanchere E, Marcos MA, Nobrega A, Coutinho A. Studies on the T cell dependence of natural IgM and IgG antibody repertoires in adult mice. Eur J Immunol 1995;25:1358–65.

[27] Mouthon L, Nobrega A, Nicolas N, Kaveri SV, Barreau C, Coutinho A et al. Invariance and restriction toward a limited set of self-antigens characterize neonatal IgM antibody repertoires and prevail in autoreactive repertoires of healthy adults. Proc Natl Acad Sci U S A 1995;92:3839–43.

[28] Mouthon L, Lacroix-Desmazes S, Nobrega A, Barreau C, Coutinho A, Kazatchkine MD. The self-reactive antibody repertoire of normal human serum IgM is acquired in early childhood and remains conserved throughout life. Scand J Immunol 1996; 44:243–51.

[29] Nobrega A, Grandien A, Haury M, Hecker L, Malanchere E, Coutinho A. Functional diversity and clonal frequencies of reactivity in the available antibody repertoire. Eur J Immunol 1998; 28:1204–15.

[30] Vasconcellos R, Nobrega A, Haury M, Viale AC, Coutinho A. Genetic control of natural antibody repertoires. I. IgH, MHC and TCR beta loci. Eur J Immunol 1998;28:1104–15.

[31] Nobrega A, Haury M, Gueret R, Coutinho A, Weksler ME. The age-associated increase in autoreactive immunoglobulins reflects a quantitative increase in specificities detectable at lower concentrations in young mice. Scand J Immunol 1996;44:437–43.

[32] Nobrega A, Stransky B, Nicolas N, Coutinho A. Regeneration of natural antibody repertoire after massive ablation of lymphoid system: robust selection mechanisms preserve antigen binding specificities. J Immunol 2002;169:2971–8.

[33] Link H, Baig S, Jiang YP, Olsson O, Hojeberg B, Kostulas V et al. B cells and antibodies in MS. Res Immunol 1989;140:219–26 [discussion p. 45–8].

[34] Jimbow K. Biological role of tyrosinase-related protein and its relevance to pigmentary disorders (vitiligo vulgaris). J Dermatol 1999;26:734–7.

[35] Erlacher P, Lercher A, Falkensammer J, Nassonov EL, Samsonov MI, Shtutman VZ et al. Cardiac troponin and beta-type myosin heavy chain concentrations in patients with polymyositis or dermatomyositis. Clin Chim Acta 2001;306:27–33.

[36] Boitier E, Beaune P. Xenobiotic-metabolizing enzymes as autoantigens in human autoimmune disorders. An update. Clin Rev Allergy Immunol 2000;18:215–39.

[37] Cohen IR. The cognitive paradigm and the immunological homunculus. Immunol Today 1992;13:490–4.

[38] Goldrath AW, Bevan MJ. Selecting and maintaining a diverse T-cell repertoire. Nature 1999;402:255–62.

[39] Lacroix-Desmazes S, Kaveri SV, Mouthon L, Ayouba A, Malanchere E, Coutinho A et al. Self-reactive antibodies (natural autoantibodies) in healthy individuals. J Immunol Methods 1998; 216:117–37.

[40] Cohen IR. Regulation of autoimmune disease physiological and therapeutic. Immunol Rev 1986;94:5–21.

[41] Hurez V, Kaveri SV, Kazatchkine MD. Expression and control of the natural autoreactive IgG repertoire in normal human serum. Eur J Immunol 1993;23:783–9.

[42] Kumar V, Sercarz E. Distinct levels of regulation in organ-specific autoimmune diseases. Life Sci 1999;65:1523–30.

[43] Bach JF, Chatenoud L. Tolerance to islet autoantigens in type 1 diabetes. Annu Rev Immunol 2001;19:131–61.

[44] Moalem G, Leibowitz-Amit R, Yoles E, Mor F, Cohen IR, Schwartz M. Autoimmune T cells protect neurons from secondary degeneration after central nervous system axotomy. Nat Med 1999;5:49–55.

[45] Schwartz M, Cohen IR. Autoimmunity can benefit self-maintenance. Immunol Today 2000;21:265–8.

[46] Moudgil KD, Sercarz EE. The self-directed T cell repertoire: its creation and activation. Rev Immunogenet 2000;2:26–37.

[47] Kyriatsoulis A, Manns M, Gerken G, Lohse AW, Ballhausen W, Reske K et al. Distinction between natural and pathological autoantibodies by immunoblotting and densitometric subtraction: liver–kidney microsomal antibody (LKM) positive sera identify multiple antigens in human liver tissue. Clin Exp Immunol 1987;70:53–60.

[48] Bouanani M, Dietrich G, Hurez V, Kaveri SV, Del Rio M, Pau B et al. Age-related changes in specificity of human natural autoantibodies to thyroglobulin. J Autoimmun 1993;6:639–48.

[49] Ferreira C, Mouthon L, Nobrega A, Haury M, Kazatchkine MD, Ferreira E et al. Instability of natural antibody repertoires in systemic lupus erythematosus patients, revealed by multiparametric analysis of serum antibody reactivities. Scand J Immunol 1997;45:331–41.

[50] Stahl D, Lacroix-Desmazes S, Mouthon L, Kaveri SV, Kazatchkine MD. Analysis of human self-reactive antibody repertoires by quantitative immunoblotting. J Immunol Methods 2000;240:1–14.

[51] Augenlicht LH, Bordonaro M, Heerdt BG, Mariadason J, Velcich A. Cellular mechanisms of risk and transformation. Ann NY Acad Sci 1999;889:20–31.

[52] Berman DE, Dudai Y. Memory extinction, learning anew, and learning the new: dissociations in the molecular machinery of learning in cortex. Science 2001;291:2417–9.

[53] Downward J. The ins and outs of signalling. Nature 2001; 411:759–62.

[54] Kalir S, McClure J, Pabbaraju K, Southward C, Ronen M, Leibler S et al. Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. Science 2001;292:2080–3.

[55] Moser B, Loetscher P. Lymphocyte traffic control by chemokines. Nat Immunol 2001;2:123–8.

[56] Rao CV, Arkin AP. Control motifs for intracellular regulatory networks. Annu Rev Biomed Eng 2001;3:391–419.

[57] Wilkie AO, Morriss-Kay GM. Genetics of craniofacial development and malformation. Nat Rev Genet 2001;2:458–68.

[58] Atlan H, Cohen IR. Immune information, self-organization and meaning. Int Immunol 1998;10:711–7.