

Biological Sciences, Medical Sciences and Physical Sciences, Applied Mathematics.

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

Francisco J Quintana¹, Gad Getz², Guy Hed², Eytan Domany² and Irun R Cohen¹

¹Department of Immunology and ²Department of Physics of Complex Systems, The Weizmann Institute of Science, Rehovot, Israel

Corresponding author: Prof. Irun R Cohen, Department of Immunology, The Weizmann Institute of Science, Rehovot 76100, Israel.

Fax: + 972 8 934 4103, Tel: + 972 8 934 2911.

E-mail: Irun.Cohen@weizmann.ac.il.

Text pages: 19

Figures: 4

Tables: 3

Words in the abstract: 162

Total number of characters: 45958

Abstract

Informatic methodologies are being applied successfully to analyze the complexity of the genome. But beyond the genome, coping with the environment depends greatly on the immune system. Here we apply informatics to analyze the patterns of autoantibodies in the sera of 20 healthy persons and 20 persons with type 1 diabetes mellitus. Using an unbiased solid-phase antibody test, we detected serum IgG and IgM antibodies 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 analysing 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. Thus, autoantibodies are common, but patterns of reactivity to defined subsets of self-antigens can provide information about the state of the body.

Introduction

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 of how the immune system actually discriminates the self from the non-self (6).

The present study had two objectives: to characterize the set of molecules recognized by autoantibodies in healthy persons and to learn whether the global patterns of autoantibodies might discriminate between a state of health and an autoimmune disease such as type 1 diabetes mellitus. We did not use reactivity thresholds, as is usually done to define a negligible background—but rather developed a system to detect even small amounts of autoantibodies binding to an array of 87 different antigens without preconceived bias. 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 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 the autoantibody repertoire is structured and can yield information about the state of the body when analyzed with suitable informatic tools.

Materials and Methods

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.

Antibodies

The secondary 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 Labs. Inc. (West Grove, Pennsylvania, USA), and were used at a final dilution of 1:1500 in bovine serum albumin 0.3 %.

Serum Samples

Serum samples were collected at the Hadassah Medical Center (Jerusalem, Israel) under the supervision of Dr. Rivka Abulafia-Lapid and Prof. Itamar Raz from 20 healthy adult blood donors with no family history of diabetes, and from 20 non-selected type 1 diabetes patients. Informed consent to test the sera was obtained. The samples were stored at -20°C without any additive.

Solid-phase antibody assay

We used a standard ELISA assay. Antigens (10 µg/ml in phosphate-buffered saline) were coated in wells of 96-well ELISA plates (Maxisorp:Nunc, Roskilde, Denmark) by overnight incubation at 4°C. The plates were washed with phosphate-buffered saline 0.05 % Tween, and blocked for 2 hours 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 hours of incubation at 37°C, the sera were removed and the plates washed with phosphate-buffered saline 0.05 % Tween. Bound antibodies were detected with an appropriate alkaline phosphatase or horseradish peroxidase conjugated second antibody (Jackson ImmunoResearch Labs. Inc.), 50 µl incubated for 1.5 hours at 37°C. The plates were washed with phosphate-buffered saline, 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 by testing the serum dilutions and the incubation times for each antigen. For most of the antigens, we found a direct correlation between the OD readings and dilutions of the sera between 1:50 and 1:200 (Figure 1A). Accordingly, we chose 1:100 as the standard dilution of the test sera. The kinetics of incubation after the addition of the substrate for the alkaline phosphatase are shown in Figure 1B. The relationship between the OD and the incubation time was linear during 45 minutes of incubation (mean $R^2 - SD = 0.98 - 0.02$). Therefore, we recorded the OD readings 30 minutes 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-assay (not shown) and inter-assay variation (Figure 1C) yielded values for the r^2 coefficient of 0.98 and 0.96, respectively, $p < 0.0001$ for both.

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 2 secondary antibodies) were placed in a matrix A , whose element A_j^i represents the extent to which the serum of subject i reacted with test antigen j , (the secondary antibody was absorbed in the index j). The "immune state" of subject i is represented by a vector \mathbf{A}^i (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 $k=1, \dots, 88$ (IgM) and $k=89, \dots, 176$ (IgM + IgG):

$$T_k^i = \frac{M}{2 \sum_k \log A_k^i} \cdot \log A_k^i$$

We analyzed the data using a method introduced recently by Getz et al (7) for analysis of gene expression. In this method, 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, our analysis uses various submatrices of the total data matrix T . Before each clustering process, 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 standard deviation of the

row. The elements of the resulting renormalized submatrix are denoted by G_j^i . The correlation coefficient of antigens j and l , as measured over the K samples, is given by

$$c_{j,l} = \sum_i^K G_j^i G_l^i$$

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

$$d_{j,l} = 1 - c_{j,l}^2$$

In contrast, the distance $D_{i,k}$ between subjects i and k is the Euclidian distance:

$$D_{i,k} = \sqrt{\sum_j (G_j^i - G_j^k)^2}$$

Unsupervised clustering techniques were used to explore the structure of the data in order to reveal, in an unbiased way, the natural classes therein. We used the SPC clustering algorithm (8), which organizes the data in the form of a dendrogram, such as those shown in Figs. 3 and 4. 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. One of the main advantages 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 $R(C)$, the more statistically significant and stable (against noise in the data and fluctuations) is C . We used SPC in conjunction with a coupled two-way clustering approach (7). This method identifies subsets of serum samples and of antigens. When a set of serum samples is represented by the reactivities of a set

of selected antigens and clustered accordingly, meaningful partitions of the samples can 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 was tested for the quality of its predictive capacity. 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 report the *specificity* and *sensitivity* of each classification: *specificity* is the fraction of correctly diagnosed subjects present in a cluster; *sensitivity* is the fraction of correctly diagnosed subjects that were included in the cluster, out of the total number of subjects with the same diagnosis.

Combining classifiers

The sensitivity and specificity of our C_{final} classifier was improved by combining the predictions that arose from using several different sets of antigens. This was done by classifying an unknown serum sample as type 1 diabetes, for example, if it was so discriminated by a majority of the classifiers. Subsequently, the identity of the serum samples was disclosed, and the groups obtained by combining our classifiers (healthy or type 1 diabetes) were evaluated in terms of specificity and sensitivity, as described above.

Assessing statistical significance

Statistical analysis was done to test two questions: first, whether the pattern of antibody reactivity exhibited a non-trivial structure, and second, whether the clinical state of the subjects was reflected by their reactivity patterns to sets of antigens. The statistical significance of the results was measured by their respective P-values: the probability of obtaining equal or better results using randomized data. We used the original data matrix that produced the original clusters, but randomized the clinical labels of the subjects. A randomized matrix was normalized as was the original matrix. We measured the P-values empirically by repeatedly applying our analysis to different randomizations of the data and calculating the fraction of instances that yielded a structure as stable as that produced by the original data. We examined the dendrograms generated by the randomized data to test whether they contained any cluster C whose stability index $R(C)$ exceeded one of the R values that were obtained for the stable clusters derived from the original reactivity matrix.

The P-values of the diagnostic labels were estimated by testing the probability P that a previously selected discriminating cluster of C subjects would produce the same number N_e of errors (or less) for the randomized labels. 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 is given by

$$P(e \leq N_e | C, S, N) = \sum_{e=0}^{N_e} \sum_{x=\max(0, C-(N-S))}^{\min(C, S)} \delta\{e, \min[S + C - 2x, N - S - (C - 2x)]\} \binom{C}{x} \binom{N-C}{S-x} / \binom{N}{S}$$

where S is the number of diagnosed subjects and e the number of classification errors.

The outer sum collects the contribution from different number of errors, e , and the inner sum goes over all possible sizes of the intersection between the cluster and the diagnosed subjects. The δ function ensures that the probability of only the cases with exactly e errors are added in the inner sum.

RESULTS

Serum autoantibodies in healthy blood donors

We tested the repertoire of autoantibodies in the sera of 20 healthy individuals using detection antibodies directed to IgM or IgG + IgM. Figure 2 shows the results obtained for one representative serum. It can be seen that the serum contained autoantibodies of both isotypes to many self-antigens. 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 gram-negative lipopolysaccharide (LPS) and purified protein derivative of *Mycobacterium tuberculosis* (PPD). Thus healthy persons may express a wide range of autoantibodies. To learn whether the patterns of 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.

Structure in the repertoire: self-antigen clusters

Our strategy was first to cluster the antigens and then to use the various antigen clusters as probes to cluster the subjects. To cluster the antigens, we used the serum OD readings for all subjects (healthy or not) to cluster the 87 antigens. The serum reactivity data (for 2 isotypes \times 87 antigens) exhibited a non-trivial structure, as is evident from comparing the dendrogram of Fig. 3a (obtained by clustering the original antigen matrix) with that of Fig. 3b (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 5 antigens, with stability index $R=1.33$. We tested 1000 different realizations of

randomized matrices, and did not find a single instance that produced a cluster with greater stability. 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 (9); sera that react with one member of an antigen cluster will tend to react to other members of the antigen cluster.

Clusters as classifiers of type 1 diabetes

The antibody reactivities directed to subsets of antigens identified in this first round of analysis were used as probes to discriminate between the different serum samples obtained from the 20 healthy donors and the 20 diabetes patients. Figure 4 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 that separated the healthy and diabetic sera into different clusters; Table 3 summarizes the findings and includes the P-values calculated for each cluster size C and error number N_e . 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, AchR, heparin, and insulin. Among this list of molecules, only insulin has been noted previously to be a self-antigen in type 1 diabetes (10).

Discussion

In this work we studied the repertoire of human autoantibodies in health and in type 1 diabetes using an experimental approach that did not bias our observations or down-grade self-reactivity in the healthy controls; we did not use absorptions or threshold titrations designed to establish a negative "background". Instead, we built an ELISA assay that would uncover reliable data, and we analyzed the actual OD values for each one of the 87 antigens listed in Table 1. The OD readings and the prevalence of antibodies to self-molecules were of the same magnitude as the values obtained for the control bacterial antigens LPS or PPD. Moreover, the patterns of autoantibody reactivity did discriminate between health and disease. Therefore, the reactivities that we detected are likely to be due to specific antibodies and not to artefacts.

Note that the autoantibodies of the healthy subjects bound to many self-antigens implicated in autoimmune diseases (Table 2): histone IIA, and single- and double-stranded DNA — targeted in systemic lupus erythematosus(2); heat shock protein 60 (HSP60), insulin and glutamic acid decarboxylase (GAD) — associated with type 1 diabetes mellitus (10); myelin basic protein (MBP) and myelin oligodendrocyte glycoprotein (MOG) — associated with multiple sclerosis(11); the acetylcholine receptor (AchR) — targeted in myasthenia gravis (1); tyrosinase— associated with vitiligo(12); myosin — associated with polymyositis (13); and cytochrome C P450 — associated with autoimmune liver disease (14). Other frequent self-antigens included the serum proteins fibrinogen and clotting factor VII, heparin, enzymes and globin molecules.

The documentation of autoantibodies to a set of defined antigens in healthy people is compatible with the concept of the immunological homunculus (9, 15). The term immunological homunculus refers to the observation that T-cell and B-cell autoimmunity in healthy individuals is usually organized around particular sets of self-antigens (16, 17). Homunculus theory proposes that this natural autoimmunity is

regulated by various mechanisms that prevent the transition of healthy autoimmunity into autoimmune disease (9, 15, 18-21). Indeed, the development of autoimmune disease could be explained most simply by the failure of these control mechanisms (6, 9). 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 (9, 15). Clearly, we would like to know how natural autoimmunity develops to particular sets of self-antigens, what functions healthy autoimmunity might serve in body maintenance (22, 23), how natural autoimmunity is controlled, and how it deteriorates into autoimmune disease (19, 24). The present study did not explore these biological questions, but rather focused on the informatic question of 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.

Type 1 diabetes is caused by autoimmune T cells that destroy the insulin-producing beta cells of the pancreas (21). Many self-antigens have been found to mark the diabetogenic autoimmune process (10), and autoantibodies to GAD and to insulin, among others, are used as markers to identify persons who have started to destroy their beta cells (4). But how can autoantibodies to GAD and insulin be associated with the type 1 diabetes if, as shown in Table 2, many healthy persons produce these autoantibodies? Quite simply, the diagnostic tests are constructed and standardized in a way that takes advantage of the greater amounts and higher affinities of these antibodies that are produced when the autoimmune disease process becomes activated; the natural autoantibodies are buried in the background (25, 26). Our aim here, in contrast to traditional procedure, was not to analyze the quantities or affinities of particular autoantibodies, but to test for the presence of informative structure in the global array of autoantibodies. Coutinho and his colleagues pioneered the study of patterns of autoantibodies binding to undefined antigens in blots of tissue extracts; they have used

principal component analysis to study their results (27). The present work extends the study of patterns to defined self-antigens and uses a cluster analysis.

Here we show that healthy subjects and type 1 diabetes subjects can be distinguished, despite the presence of various autoantibodies in both groups, by the patterns of their autoantibodies. In other words, patterns of autoantibody reactivity can be more informative than any simple one-to-one relationship between an antibody and an antigen (9). We do not yet know the biological relevance of the particular autoantibodies or of their patterns to the pathophysiology of disease, but 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 (28-34). 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 are required.

Note that we estimated the statistical significance of the clusters by empirically testing the frequency with which similar results could arise by chance using scrambled classes of test subjects. One thousand such computer experiments proved the significance of the real data. The immune system can be viewed as a system that continuously reflects and responds to the state of the body and its evolving needs for maintenance and repair, as well as for defense (6, 9). We are extending our antibody analyzes, and we find that patterns of autoantibodies to particular subsets of self-antigens may discriminate type 1 diabetes from type 2 diabetes, both types of diabetes from healthy persons, and the dynamic evolution of type 1 diabetes over time (in preparation). Hence, analysis of the autoantibody repertoire can provide an informative view of the state of the body. 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 (35).

Acknowledgements: Prof. Irun R. Cohen is the incumbent of the Mauerberger Chair in Immunology, Director of the Robert Koch Minerva Center for Research in Autoimmune Disease, and Director of the Center for the Study of Emerging Diseases. Eytan Domany 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. (1999) *Curr Opin Rheumatol* **11**, 483-8.
2. Pisetsky, D. S. (2000) *Curr Opin Rheumatol* **12**, 364-8.
3. Zauli, D., Cassani, F. & Bianchi, F. B. (1999) *Biomed Pharmacother* **53**, 234-41.
4. Verge, C. F., Stenger, D., Bonifacio, E., Colman, P. G., Pilcher, C., Bingley, P. J. & Eisenbarth, G. S. (1998) *Diabetes* **47**, 1857-1866.
5. Avrameas, S., Guilbert, B. & Dighiero, G. (1981) *Ann Immunol* **132**, 231-6.
6. Cohen, I. R. (2000) *Semin Immunol* **12**, 215-9; discussion 257-344.
7. Getz, G., Levine, E. & Domany, E. (2000) *Proc Natl Acad Sci U S A* **97**, 12079-84.
8. Blatt, M., Wiseman, S. & Domany, E. (1996) *Physical Review Letters* **76**, 3251-3255.

9. Cohen, I. R. (2000) *Tending Adam's Garden: Evolving the Cognitive Immune Self* (Academic Press, London).
10. Tisch, R. & McDevitt, H. (1996) *Cell* **85**, 291-7.
11. Link, H., Baig, S., Jiang, Y. P., Olsson, O., Hojeberg, B., Kostulas, V. & Olsson, T. (1989) *Res Immunol* **140**, 219-26; discussion 245-8.
12. Jimbow, K. (1999) *J Dermatol* **26**, 734-7.
13. Erlacher, P., Lercher, A., Falkensammer, J., Nassonov, E. L., Samsonov, M. I., Shtutman, V. Z., Puschendorf, B. & Mair, J. (2001) *Clin Chim Acta* **306**, 27-33.
14. Boitier, E. & Beaune, P. (2000) *Clin Rev Allergy Immunol* **18**, 215-39.
15. Cohen, I. R. (1992) *Immunol Today* **13**, 490-4.
16. Goldrath, A. W. & Bevan, M. J. (1999) *Nature* **402**, 255-62.
17. Lacroix-Desmazes, S., Kaveri, S. V., Mouthon, L., Ayoub, A., Malanchere, E., Coutinho, A. & Kazatchkine, M. D. (1998) *J Immunol Methods* **216**, 117-37.
18. Cohen, I. R. (1986) *Immunol Rev* **94**, 5-21.
19. Hurez, V., Kaveri, S. V. & Kazatchkine, M. D. (1993) *Eur J Immunol* **23**, 783-9.
20. Kumar, V. & Sercarz, E. (1999) *Life Sci* **65**, 1523-30.
21. Bach, J. F. & Chatenoud, L. (2001) *Annu Rev Immunol* **19**, 131-61.
22. Moalem, G., Leibowitz-Amit, R., Yoles, E., Mor, F., Cohen, I. R. & Schwartz, M. (1999) *Nat Med* **5**, 49-55.
23. Schwartz, M. & Cohen, I. R. (2000) *Immunol Today* **21**, 265-8.
24. Moudgil, K. D. & Sercarz, E. E. (2000) *Rev Immunogenet* **2**, 26-37.
25. Kyriatsoulis, A., Manns, M., Gerken, G., Lohse, A. W., Ballhausen, W., Reske, K. & Meyer zum Buschenfelde, K. H. (1987) *Clin Exp Immunol* **70**, 53-60.

26. Bouanani, M., Dietrich, G., Hurez, V., Kaveri, S. V., Del Rio, M., Pau, B. & Kazatchkine, M. D. (1993) *J Autoimmun* **6**, 639-48.
27. Haury, M., Grandien, A., Sundblad, A., Coutinho, A. & Nobrega, A. (1994) *Scand J Immunol* **39**, 79-87.
28. Augenlicht, L. H., Bordonaro, M., Heerdt, B. G., Mariadason, J. & Velcich, A. (1999) *Ann N Y Acad Sci* **889**, 20-31.
29. Berman, D. E. & Dudai, Y. (2001) *Science* **291**, 2417-9.
30. Downward, J. (2001) *Nature* **411**, 759-62.
31. Kalir, S., McClure, J., Pabbaraju, K., Southward, C., Ronen, M., Leibler, S., Surette, M. G. & Alon, U. (2001) *Science* **292**, 2080-3.
32. Moser, B. & Loetscher, P. (2001) *Nat Immunol* **2**, 123-8.
33. Rao, C. V. & Arkin, A. P. (2001) **3**, 391-419.
34. Wilkie, A. O. & Morriss-Kay, G. M. (2001) *Nat Rev Genet* **2**, 458-68.
35. Atlan, H. & Cohen, I. R. (1998) *Int Immunol* **10**, 711-7.

Figures and Tables

Figure 1. Optimization of the ELISA assay. A. Titration of the IgG + IgM autoantibody reactivities. Various dilutions of a healthy serum were tested for binding to each of the antigens listed in Table 1. Only representative reactivities are shown. B. Kinetics of the alkaline phosphatase substrate reaction. Representative reactivities are depicted. C. Reproducibility of the assay. The serum used for the experiments described in A and B was analyzed on two different days for its reactivity against the panel of

antigens listed in Table 1. The OD readings obtained for each antigen the two days are plotted one against the other. The parameters corresponding to the adjustment of the diagonal to a linear trendline are shown in the inset.

Figure 2. Autoantibodies of a healthy blood donor. OD readings obtained with a representative serum are depicted. Each antigen is indicated by a number, and grouped, according to the list in Table 1. IgG + IgM (black columns) and IgG (white columns) reactivities are shown.

Figure 3. 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 randomised matrix.

Figure 4. 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.

Table 1: Antigens used.

The catalogue number is given for those molecules purchased from Sigma.

Group	Function / Structure	#	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
	Glucose	19	Enolase		E0379
		20	Aldolase		A8811

Cellular Metabolism		21	Acid Phosphatase		P1774	
	Apoptosis	22	Annexin 33 kDa.		A9460	
		23	Annexin 67 kDa.		A2824	
		24	Cytochrome C P450		C3131	
	Monoxygenases	25	Catalase		C9322	
		26	Peroxidase		P6782	
		27	Tyrosinase		T 7755	
	Others	28	Ribonuclease		R4875	
	Nucleus	Protein	29	Histone II A		H 9250
		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	
	Cytokines	43	Interleukin 2		I2644	
		44	Interleukin 10		I9276	

Immune System		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	GGFTCFFRDHSYQEEAAMELKVE	(c)	
		58	rat MOG p35-55	MEVGWYRSPFSRVVHLYRNGK	(c)	
		59	MBP		(d)	
		60	Brain Extract		B1877	
	Muscle & Skeleton	61	AchR		(e)	
		62	Myoglobin		M6036	
	Joints	63	Cartilage Extract		C5210	
	Thyroid	64	Thyroglobulin		T1001	
	Blood Cells & Platelets	65	Hemoglobin A		H0267	
		66	Spectrin		S3644	
			67	TB PPD		(f)
			68	HSP65		(g)

Foreign Antigenes	Proteins & Peptides	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 aminoacids	76	poly Arginine		P3892	
		77	poly Lysine		P4408	
		78	poly Aspartic		P6762	
		79	poly Glutamate		P4636	
	Oligonucleotides	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	TCCATGACGTTCTGACGTT		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 Prof. Avraham Ben Nun (The Weizmann Institute of Science).
- (d) Kindly provided by Dr. Felix Mor (The Weizmann Institute of Science).

- (e) Kindly provided by Prof. Sara Fuchs (The Weizmann Institute of Science).
- (f) Produced at the Statens Seruminstitut, Copenhagen, Denmark.
- (g) Kindly provided by Prof. R. van deer Zee (Utrecht University, The Netherlands).
- (h) Purchased from Pierce (Oud Beijerland, The Netherlands), catalogue number 77153.

Table 2: Frequencies of autoantibodies in healthy humans. To limit the number of self-antigens shown, the table includes only those antigens to which at least 35 % of the healthy subjects responded with an OD of greater than 0.3.

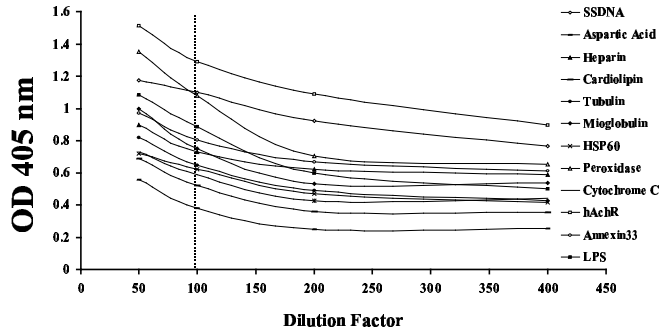
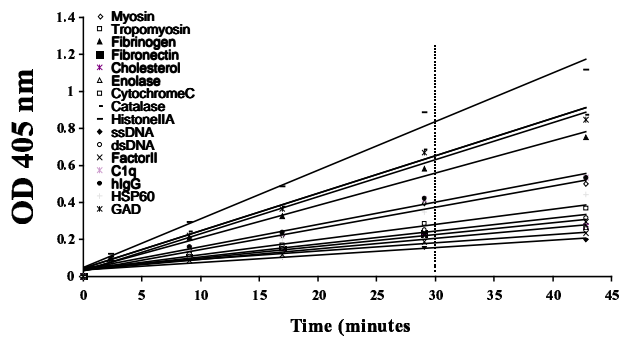
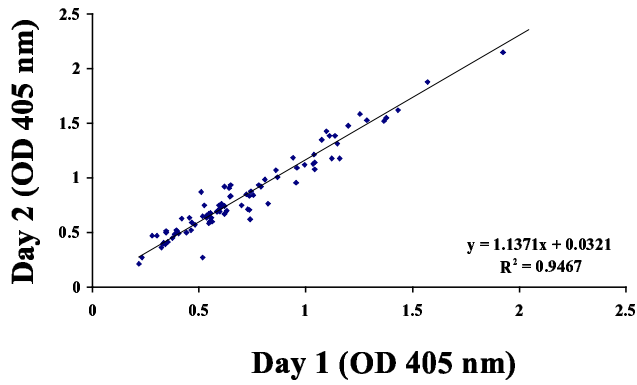
Incidence of autoantibodies (%)

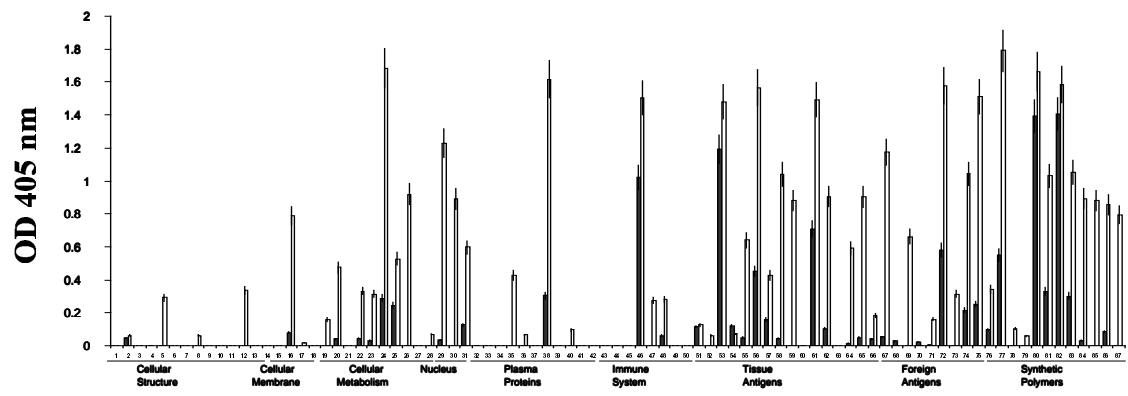
Antigen	IgM + IgG	IgM
Tubulin	50	-
Myosin	40	-
Heparin	75	-
Acid Phosphatase	35	-
Annexin 33 KDa.	55	-
Cytochrome C P450	50	80
Catalase	65	-
Tyrosinase	30	45
Histone II A	65	45
DS DNA	75	75
SS DNA	100	95
Factor VII	70	100
Fibrinogen	90	-
HSP60	40	-
GAD	100	70
Insulin	35	35
MOG	-	95
MBP	35	-
AchR	90	75
Myoglobulin	65	35
Hemoglobin A	50	45

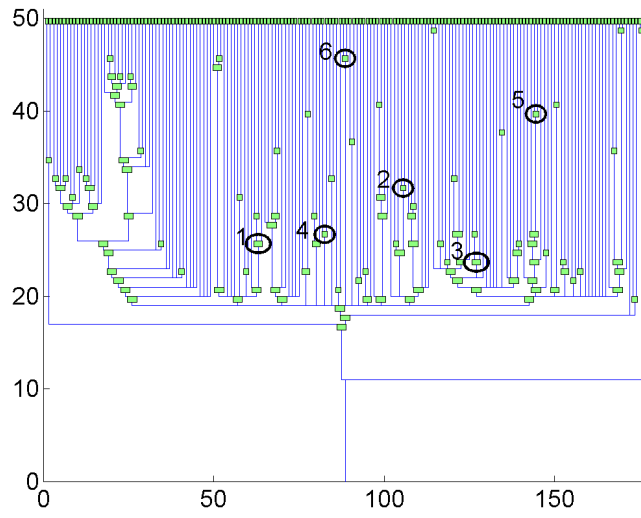
LPS	85	45
TB PPD	90	50

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

Cluster number	Antigens	Sensitivity	Specificity	Cluster Size	No. of Errors	P-Value
1	IgM to Collagen I IgG+M to Collagen I IgG+M to Insulin	85%	81%	21	7	$8.8 \cdot 10^{-5}$
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 3 classifiers)	95%	90%			

A**B****C**



A**B**