The danger model in deciphering autoimmunity

Anders A. Tveita

Abstract

Autoimmunity has been a topic of intensive research for several decades, yet amazingly, no uniform hypothesis exists to explain the basis for the spectrum of autoantibody specificities seen in autoimmune diseases. It therefore seems appropriate to consider whether our current framework for understanding tolerance, and thus the mechanisms controlling the initiation and perpetuation of autoimmunity, may be faulty. Adapting the paradigm of Matzinger—the ‘danger model’, a case can be made for a perspective that appreciates the fundamental role of the tissues in controlling immune response, favouring a shift of focus in studies on the initiation of autoimmunity. Applying the elements of this model, I set forth a number of scenarios for how autoreactivity could emerge, with emphasis on the likely sources of the involved autoantigens and the functional basis of their appearance. The emerging picture is one in which disruption of tissue homeostasis takes centre stage, with the antigen-presenting cells as the key players.

Key words: Danger signalling, Tolerance, Autoimmunity, Systemic lupus erythematosus, Autoreactivity, Apoptosis.

Introduction

The autoimmune diseases form an enigmatic class of conditions in which the immune system is engaged in a destructive attack against varying repertoires of self structures. Despite decades of intensive studies of immune function and regulation of immunity in both human and animal model systems, key elements in the pathogenesis of autoimmunity still appear obscure. It is tempting to speculate that this might relate to problems with the conceptual framework of understanding immunological tolerance and its breakdown in autoimmunity. A fresh perspective is offered by approaching the immune system by the premises laid down by Polly Matzinger in her ‘danger model’ [1–6]. Even today, more than a decade after its emergence, this model offers a fundamentally different interpretation of classic and newly emerging concepts in immunology.

The danger model suggests that most autoimmune states do not stem from deficiencies in the immune response, but rather from defects in the normal physiology of tissues [4]. This review, therefore, will focus on the fundamental principles by which the immune system interacts with the tissues making up the living human, rather than on minute details of immunology.

A dissection of autoimmunity as a concept has to start with definitions. Autoimmunity is a situation in which lymphocytes reactive against self structures are activated and allowed to operate to an extent that causes sustained self-reactivity and tissue damage. This rather pragmatic definition excludes situations with transient activation of autoreactive lymphocytes, such as when fighting off an infection [7]. Indeed, such autoreactivity might be expected to be present in any situation where there is tissue damage, as discussed below. In this lies one of the fundamental postulates of the danger model—that autoreactivity per se is a normal aspect of the physiological process of tissue inflammation, whereas the defining characteristic of pathological autoimmunity (autoimmune disease) is the persistence of autoreactivity caused by a continued supply of danger signals (chronic autoreactivity).

The danger model ties the emergence of autoreactivity to the release of danger signals. Therefore, it seems possible that the dominating targets of autoreactivity might provide clues to the origin of the autoimmune response and point to potential underlying causes of danger signalling, as discussed below. For simplicity, in the following, the term ‘site of damage’ refers not only to the injured tissue structure, but also includes draining lymph nodes, as this forms the point of initial contact for lymphocyte recruitment.
Danger revisited

Matzinger defines danger signals as structures released or produced by cells or tissues undergoing stress or abnormal cell death (or even normal cell death, if the cells are not properly scavenged and go into secondary necrosis) that induce activation of resting antigen-presenting cells (APCs). A growing number of endogenous substances are being identified as potential danger signals [e.g. adenosine-5'-triphosphate (ATP), hyaluronan breakdown products, transcription factors such as high-mobility group box 1 (HMGB-1), certain complement components, S100 protein family, etc.] [8], suggesting that the term might encompass a large number of hitherto unappreciated biochemical alterations occurring in the micro-environment resulting from cellular damage [9].

The principle substrate of danger need not be anything more than unresolved cellular damage, as on the biochemical level there is no fundamental difference in the way various kinds of tissue damage presents themselves—there is no a priori means of telling what caused the damage or what kind of response would be more favourable. The logical response is therefore a broad one: flood the area with APCs to see what is there, and then bring in the effector cells and scavengers to start removing it. When the job is effectively done, danger subsides, inflammation is dampened and the process dies out.

Approaching autoimmunity

We are now seeing the arena in which autoimmunity unfolds. There is no reason why autoimmune diseases should require added complexities; autoimmunity should stem from breaches of the homeostatic apparatus presented above. Essentially this is a matter of statistical likelihood, with the effectiveness of tolerogenic (signal 1) vs immunogenic (signal 1 + signal 2) antigen presentation being the crucial determinant. A shift in balance towards the latter occurs whenever danger signals are released and trigger the expression of signal 2, so what we should be looking for is a source of prevailing tissue injury.

The importance of the autoantigens vs the danger signals

Most autoimmune diseases are characteristically associated with autoreactivity against particular subsets of antigens. A key issue is therefore to what extent an autoimmune disease is dependent on reactivity against such families of antigens. In other words, are the autoreactive targets determined by the underlying pathogenetic processes, or are they merely a reflection of the location in which the release of danger signals occur? Moreover, will autoreactivity persist upon the removal of the main target structures? How important is a particular subset of autoantigens to the maintenance of an autoimmune disease, and what classes of antigens can serve as sustained inducers and targets of autoreactivity? Certain organ-specific antigens with limited expression are presented by a small number of cells, increasing the ease with which an immunogenic presentation could be attained if danger signals are persistently released in the containing areas. The small number of cells provides scarce tolerogenic display in the normal state, whereas release of danger signals will generate an inflammatory focus, causing the emergence of autoreactivity centred at antigens confined to the inflamed area. Conceptually, such antigens should therefore provide good candidates for maintenance of an autoimmune response as long as they are presented in a dangerous context. Other, more widely distributed antigens would be presented on non-activated cells in other parts of the body, effectively terminating autoreactivity against such structures. By this line of reasoning, looking at the autoreactive target structures could provide clues to the origin of the danger signals, but might say nothing about the nature of the inciting tissue damage.

The possibility that the efficacy of tolerogenic antigen display differs for various cellular constituents provides a model to explain the distinct panels of target antigens seen in autoimmune diseases. It is obvious that, under normal circumstances, inefficient tolerization of these antigens does not pose a problem, and probabilistic reasoning does not in itself provide an answer to why autoimmunity can persist. It does however offer an explanation as to why MHC loci are consistently identified as susceptibility loci for autoimmune diseases [10, 11]. Subtle differences in MHC Class I or MHC Class II epitope binding preferences might alter the ease with which tolerance towards particular autoantigens is breached, and thus increase the likelihood of autoimmunity emerging against these structures. Indeed, data exist in support of an association between MHC haplotype and autoantigen profile in diseases such as SLE [12, 13]. With these thoughts on predisposing factors in mind, we now turn to the core of the matter—the pathogenetic basis of sustained autoimmunity.

The pathogenesis of autoimmune diseases

Antigen display is a fundamental responsibility of most cells, and so the emergence of autoreactivity is a signal that this process is not properly maintained. Consequently, autoreactivity is based on increased immunogenic presentation, diminished tolerogenic display or a combination of these. Considering the autoantibody repertoires and the clinical manifestations of autoimmune diseases, it is obvious that a fundamental difference between systemic and organ-restricted disease is the distribution of the involved autoantigens. Maintaining autoreactivity towards ubiquitous antigens should intuitively originate in widespread deficiencies of some aspect of APC function, whereas reactivity towards tissue-restricted antigens could be envisioned to occur on the basis of more ‘local’ types of aberrancy in antigen display or tissue homeostasis.

There appears to be a tendency to try to pinpoint a particular subset of antigens that constitute the pathogenetic basis of the disease being studied. The search for direct causality between autoreactivity and clinical
manifestations might be a considerable source of confusion in designing and interpreting studies on autoimmunity, especially the systemic diseases, in which the source of the triggering autoantigens is not obvious, as discussed below. An additional and important cause for caution lies in the heterogeneity of laboratory and clinical features of the individual autoimmune syndromes, with the possibility that some current disease entities represent artificial clustering of aetiologically unrelated illnesses. Such precautions might be particularly relevant to the systemic diseases, with conditions overlapping clinically and serologically, further adding to the uncertainty.

Systemic autoimmunity: SLE

SLE is considered a prototypic systemic autoimmune disease, and has been extensively studied for decades. Particular interest has been paid to the nephritis accompanying the disease. To some extent, the dominating autoantibody specificities in SLE appear to correlate with specific end-organ manifestations. Autoantibodies against chromatin particles [double-stranded DNA (dsDNA) and nucleosomes] are by many regarded as pivotal in the development of lupus nephritis [14, 15]. Nucleosomes have been found to be deposited within the glomerular matrix and capillary membranes in patients with lupus nephritis, and these deposits co-localize with glomerular autoantibody deposits [16]. A key problem in this respect is the origin of such extracellular chromatin deposits. Are they the result of an intrinsic renal pathology, or is deposition caused by passive accumulation of pre-formed circulating immune complexes?

The example of lupus nephritis effectively demonstrates major problems associated with studies of systemic autoimmunity. The ubiquitous nature of the major autoantigens makes it hard to infer what the origin of autoimmunity is, and the resulting inflammatory response against exposed antigenic structures makes it difficult to separate primary and secondary events.

Targets of autoreactivity in systemic autoimmunity

Autoantibody production in systemic autoimmune diseases is remarkable in being largely directed at antigens with more or less universal tissue distribution, including DNA/chromatin and RNA-containing complexes. There are few situations that should allow for persistent autoreactivity against such very abundant cellular constituents, and appreciating this considerably limits the possibilities of how systemic autoimmunity could conceivably be initiated.

Autoantibody production requires antigen presentation on activated APCs. The most obvious source of autoantigens for display on APCs is phagocytosed dead cell remnants in the form of apoptotic bodies or necrotic debris, with the former being the predominant one in the healthy individual. A large number of in vitro studies have indicated that phagocytic ingestion of normal apoptotic cells is immunologically silent or even triggers anti-inflammatory cytokine production by the phagocytosing cell [17, 18]. The topographic origin of an autoimmune response should be a site where APCs are effectively activated by danger signals and where autoantigens are displayed in a manner that is not sufficiently countered by tolerogenic display on non-activated APCs. Within a danger-signalling framework, the mechanistic explanation for an autoimmune disease, therefore, must encompass two basic elements: (i) a sustained source of danger signals and (ii) a means to explain the selection of target autoantigens.

Listed below are four hypothetical scenarios that could fulfill these criteria, and thus form the pathogenic basis of an autoimmune response. In various forms, elements of each of these scenarios have all been suggested as aetiological factors in SLE.

Scenario 1: waste management and apoptotic cell antigens. MHC Class II antigen presentation is largely restricted to APCs, and accordingly, T-helper and B-cell activation is controlled by these cell types. As a consequence, tolerance is closely reliant on presentation by non-activated APCs. Through the process of cross-presentation, phagocytosed material is also made accessible for presentation on MHC Class I, which gives APCs a role in the regulation of cytotoxic T cells as well. From these considerations, it follows that normal APC function is of key importance to the maintenance of tolerance.

It is possible that some antigens, by virtue of their relative inaccessibility to MHC loading compartments, would only rarely be presented on the cell surface. When cells die, however, such antigens would be sequestered in apoptotic blebs and taken up by phagocytosing cells. Upon phagocytosis, these antigens are made accessible for MHC Class I and MHC Class II presentation, forming a population of dead cell ‘ghost antigens’ effectively only presented on APCs. The appearance of such antigens on APCs would not constitute a problem, provided phagocytosis of apoptotic cells is maintained as an immunologically silent process, which is the case for normal apoptotic cell clearance.

However, if the phagocytic process triggers danger signalling, and thus activates the phagocytosing APC, a fertile ground is created for autoreactivity against the ghost antigens. Such a situation could emerge if phagocytosis was inefficient, allowing apoptotic cells to undergo secondary necrotic transformation. Indeed, there have been reports of impaired phagocytic function in SLE [19]. Deficiencies of several soluble factors, notably C1q and serum amyloid P, as well as a growing number of cell surface receptors have been implicated as possible sources of aberrant apoptotic cell clearance [20, 21]. C1q deficiency is a rare hereditary disease known to elicit a lupus-like syndrome, and it would seem plausible that deficiencies of similar factors could serve as aetiological factors in SLE [22].

Another important aspect of apoptotic cell death is its immunomodulatory effects on surrounding cells, including APCs. It is tempting to speculate that alterations in these interactions could serve as a permissive factor for the development of autoimmune diseases. Recent reports
indicate the existence of several subtypes of apoptotic cell death, characterized by the presence or absence of certain secreted or membrane-bound antigens that bring about different immunological responses [23]. By means of one or more ill-defined ligand(s), apoptotic cells under some circumstances appear to elicit an anti-inflammatory response in contacting cells [24]. In contrast, surface expression of caerulein on apoptotic cells has been shown to cause rapid activation of APCs, and preventing its binding on the surface blocks this effect altogether [25]. Subsequent release of potential danger signals including heat-shock proteins and HMGB1, such as by secondary necrotic transformation, may further influence the APC response and mode of antigen display. In these ways, factors in the local micro-environment of the dying cell might serve as sources of immunological priming against phagocytosed self antigens, either through prolonged persistence caused by inefficient uptake or by their inappropriate extracellular exposure by secretion or surface display. Indeed, there are data supportive of this type of mechanism influencing the effects of apoptotic cells on macrophages from mouse models of both SLE and diabetes mellitus [26].

Scenario 2: digestion and fate of antigens ingested by APCs. Looking beyond the uptake of apoptotic debris, it is obvious that the further fate of ingested material is important in avoiding danger signalling. Several important danger-sensing structures, including Toll-like receptors (TLR) 7 and 9, are active within endocytic compartments and could be triggered by contact with internalized structures like chromatin and certain RNA structures. It is known that genomic DNA contains CpG dinucleotide motifs, and under certain conditions has the ability to activate TLR9 signalling [27]. Therefore, if phagocytosed chromatin is not promptly and completely degraded, one could risk the activation of the APC by retained DNA. At the same time, accumulation of undegraded chromatin would provide an abundant source of antigens accessible for MHC presentation. Effectively, incomplete digestion of DNA could therefore provide both a danger signal and a source of autoreactivity focused against nuclear antigens. Similarly ill-defined cellular defects causing accumulation of endogenous DNA elements within the cytosol appear to trigger the so-called INF-stimulated DNA response. This mechanism serves as a source of pro-inflammatory signalling, causing autoinflammatory conditions with lupus-like manifestations (e.g. chilblain lupus caused by mutations in the *trex1* gene) [28].

Kawane et al. [29] reported that mice deficient in the endonuclease DNase II, thought to play a key role in the degradation of nucleosomal DNA in phagocytosed material in macrophages, are incapable of degrading ingested DNA, possibly triggering ‘activation of the innate immune system’. Several other DNase deficiencies have been proposed as potential sources of anti-DNA autoimmunity, and data to support this include reports of DNase I deficiency in two SLE patients [30] and findings of lupus-like disease manifestations in murine knockouts of DNase I [31], as well as reports of DNase-γ (DNase1L3) mutations in spontaneous models of murine SLE [32]. Of note, other ANAs are also reported to be present in these mice, although at lower levels.

Scenario 3: cryptic antigens—‘altered self’. Antigens processed in an unusual manner can lead to the formation and presentation of ‘altered self’ structures, which are not part of the natural processed and presented self antigen repertoire. Several lines of experiments have demonstrated that such modified self structures provide efficient immunogens, and they have been suggested to play a central role in autoimmune disease pathogenesis provided that the cryptic antigens are accessible as targets in vivo. For instance, it has been shown that immunization of healthy mice with modified antigens, such as de-methylated syngeneic chromatin, triggers autoimmune disease, not seen by immunization with the native antigen [33, 34]. Acetylated DNA similarly accelerates disease manifestation in lupus-prone mice [35]. Representing a class of antigens towards which the immune system is not normally tolerized, the functional immunogenicity of ‘altered self’ antigens when introduced with adjuvants is no surprise. By themselves, however, most altered self structures are not known to be immunogenic (i.e. serve as danger signals). The crucial factor would therefore be the process by which they are brought about. An obvious source of such alterations would be an inflammatory process, with oxidative damage and release and activation of potent proteases, for example, a burn injury or a chronic infectious process. The accompanying release of danger signals would then activate APCs, and offer both native and altered self structures for T-cell activation. In some instances, the altered antigen could in itself serve as a danger signal, such as might be the case for hypomethylated DNA, which likely obtains the ability to function as a TLR9 ligand. Likewise, loss of a mannosidase in mice causes lupus-like disease likely precipitated by the unmasking of TLR ligands [36]. In this case, danger signals and target antigens differ, demonstrating that the underlying defect does not necessarily dictate the targets of the ensuing autoimmune response. For most antigens, the potential for ‘de-masking’ of danger signals by biochemical modification remains unknown. Consequently, rather than offering a complete scenario for the origin of autoimmunity, the altered self model circles out a subset of epitopes towards which tolerance is easily broken in the face of danger-driven APC activation. Although these antigens might not always be a pathogenetic factor (i.e. the danger signal), there is still a possibility, therefore, that they serve as a permissive factor in the maintenance of autoreactivity analogous to antigens with limited tissue distribution.

Scenario 4: autoantigens in complexes. The hapten-carrier model was proposed to explain how poor immunogens, such as DNA, can become targets of an immune response when in complex with a more potent immunogenic structure. According to this model, by
virtue of its physical association with the carrier protein, the hapten follows along upon antibody-mediated internalization of the carrier in APCs or carrier-specific B cells. The hapten is then presented alongside carrier epitopes, and serves as a T-cell target in a process initiated by an anti-carrier response. There is no contradiction between the danger model and the hapten-carrier principle. The fundamental property of a hapten carrier is that it introduces a protein antigen that is not effectively displayed on host cells, analogous to ‘altered self’ antigens, presented in a dangerous context, for example, through the use of adjuvants. Also, there is the possibility that certain hapten structures may represent ligands to yet undescribed receptors, thus serving as danger signals in their own right [37]. For instance, HMGB1, a well-recognized endogenous danger signal, has been shown to serve as an effective adjuvant, boosting the immune response against nucleosomes, which by themselves are poor immunogens. HMGB1-containing nucleosome complexes activate dendritic cells in vitro, in a fashion that is competitively inhibited by addition of a recombinant form of a DNA-binding domain of HMGB1 [38]. One possible explanation for these findings would be that HMGB1 serves as a hapten structure by (i) its ability to bind nucleosomal DNA and (ii) to serve as a danger signal. This effectively co-localizes the danger signal with a number of nuclear antigens. Furthermore, HMGB1 has been shown to induce cross-presentation of ingested antigens by APCs, thus also facilitating a cytotoxic T-cell response [39]. Similar mechanisms might explain the effect of using heat-shock protein complexes as a means of boosting vaccination against various tumour antigens [40].

Immunization with a polyomavirus transcription factor in complex with mammalian DNA induces transient anti-dsDNA production in healthy mice. Induction of transgenic expression of this transcription factor in kidneys following such immunization causes sustained production of anti-dsDNA in the absence of a viral infection [41]. The hapten-carrier model demonstrates that autoreactivity towards DNA can emerge from the initial breach of tolerance to a smaller set of antigens within a chromatin complex. Although artificial, this system demonstrates an expansion of autoreactivity that might be relevant to the initiation of autoimmune disease.

If one accepts that transient autoreactivity is a natural aspect of the response to tissue injury, the appearance of autoantibodies would be expected during the course of substantial inflammatory processes. Such antibodies would react with their antigens to form immune complexes, and facilitate Fc-mediated clearance by APCs. If formation of immune complexes is increased, or if their clearance is insufficient, a ‘hapten-carrier’-like situation could emerge, with the antibody functioning as a ‘carrier’, facilitating Fc-mediated uptake and high-level presentation of complex-contained antigens [42]. Uptake of immune complexes can induce different responses in various APCs, depending on contrasting effects by uptake via particular subtypes of Fc receptors, referred to as ‘inhibitory’ or ‘stimulatory’ depending on the response of the APC upon cross-linking of receptors. Of note, the inhibitory Fc receptor, FcRIIb, seems to seques ter immune complexes on APCs in a way that prevents their presentation in MHC. It has been demonstrated that retroviral delivery of FcRIIb to lupus-prone mice can eliminate autoantibody production and prevent the disease [43]. If the balance between stimulatory and inhibitory Fc receptors present on APCs is somehow shifted in favour of the former, this could provide a driving force for persisting autoimmunity by promoting the presentation of immune complex-captured autoantigen(s) in a dangerous context. According to this scenario, the lupus-defining defect is ascribed to a hyper-reactive response to ‘physiological autoreactivity’, with immunogenic uptake of immune complexes serving as a sustainment mechanism, providing a constant source of danger signals. If this is the case, looking at the phenotype of APCs upon challenge with immune complexes would seem a logical first step in revealing the underlying mechanism(s).

Systemic autoimmunity: is there a unifying concept?

The above scenarios describe conditions under which a persisting immune response against tissue antigens could be initiated and maintained. They all relate to defects or aberrancies in the clearance of cellular debris, which leads to the retention of cellular remnants. The continued presence of such cellular constituents could, by itself, provide the danger signals needed to initiate and maintain a cellular and humoral immune response. There would thus not be any absolute requirement for an external event to trigger autoimmunity, although exogenously inflicted tissue injury (ultraviolet light, hypoxia, infections, trauma, etc.) could conceivably accelerate or expand the process, hastening the appearance of clinical manifestations (e.g. lupus flares). A cautious reservation is the likely possibility that some autoimmune diseases might actually evolve from unappreciated persistent infections. In such a situation, the continued infliction of tissue injury by the infectious agent would provide the driving force of the immune response. In contrast to ‘true’ autoimmune diseases, the autoimmunity of persisting infections is an aspect of a biologically meaningful but inadequate immune response. However, studies on autoimmunity-prone animals bred in strictly germ-free environments confirm that autoimmunity exists also in the absence of infectious agents, that is, ‘true’ sustained autoimmunity is a reality [44, 45]. Importantly, neither of the above-mentioned scenarios is exclusive. Rather, they complement each other and could be aspects of deficiencies affecting both the processing of the antigen in the dying cell and its uptake, digestion and presentation by APCs. A number of specific molecular defects that relate to these processes have been proposed as possible contributors to SLE and other systemic autoimmune diseases, with the list still growing. The idea that deficiencies related to apoptosis or clearance of dead cells form the basis of SLE pathogenesis is now widely held within the research community. Figure 1 is an attempt to summarize factors that
have been postulated to be important contributors to lupus pathogenesis, and show how they all fit together, with danger signalling being the central element. The emerging picture is one in which this network of dying cells, their phagocytosis and presentation by APCs and the accompanying danger signals forms the basis of the lupus condition.

Concluding remarks

Is there anything to suggest that systemic and organ-specific autoimmunity have a common basis? Following the reasoning of the danger model, the maintenance of tolerance is a common responsibility of cells of all tissues, and the breach of tolerance tells us that some of the cells containing the autoantigen(s) in question are not doing their job correctly or are inappropriately processed and cleared upon dying. By considering that transient autoimmunity could be a biologically meaningful or at least acceptable part of an immune response, the focus in understanding autoimmune diseases is shifted from the process of lymphocyte activation to the recruitment and activation of APCs by danger signals released by cells in the involved tissue(s).

This text does not provide a definite answer to the fundamental question—what triggers autoimmune diseases? I have presented several scenarios that I imagine could touch upon the answer. More such elements probably exist, and new ones might emerge as our knowledge of the interplay between the innate immune system and the tissues increases. The key to answering such questions may lie very close, it is just a matter of getting the context right.
Rheumatology key messages

- Aberrancies of apoptosis can release danger signals, serving as a potential driving force for autoimmunity.
- Autoimmune diseases may relate to tissue-restricted pathology rather than dysfunctions in the immune system itself.
- The continued presence of danger signals separates autoimmune diseases from ‘physiological’ autoreactivity.

Acknowledgements

I would like to thank Dr Ole Petter Rekvig, Polly Matzinger, Tony Marion and Svetlana N. Zykova for helpful discussions and critical review of the manuscript.

Funding: This work was supported by internal funding from the University of Tromsø.

Disclosure statement: The author has declared no conflicts of interest.

References

et al


www.rheumatology.oxfordjournals.org