Representation in the (Artificial) Immune System

Chris McEwan and Emma Hart

Napier University, Edinburgh, Scotland
{c.mcewan, e.hart}@napier.ac.uk

Abstract. Much of contemporary research in Artificial Immune Systems (AIS) has partitioned into either algorithmic machine learning and optimisation, *or*, modelling biologically plausible dynamical systems, with little overlap between. We propose that this dichotomy is somewhat to blame for the lack of significant advancement of the field in either direction and demonstrate how a simplistic interpretation of Perelson's shape-space formalism may have largely contributed to this dichotomy. In this paper, we motivate and derive an alternative representational abstraction. To do so we consider the validity of shape-space from both the biological and machine learning perspectives. We then take steps towards formally integrating these perspectives into a coherent computational model of notions such as life-long learning, degeneracy, constructive representations and contextual recognition – rhetoric that has long inspired work in AIS, while remaining largely devoid of operational definition.

1 Introduction

Perelson's shape-space formalism has become the *de facto* representational abstraction in Artificial Immune Systems (AIS). Briefly: receptors and their ligands are represented as points in an n-dimensional "binding parameter" space, with a contiguous *recognition region* surrounding each point to account for imperfect matching. Ligands and receptors that have intersecting regions are said to have *affinity*. Although biologically simplistic, for theoretical immunologists the shape-space has a certain heuristic value in quantifying gross properties of the immune repertoire, away from the complex bio-chemical process of protein binding (see Sect. 2 and [13] for more details).

This abstraction has been adopted wholesale by the AIS community as isomorphic with the vectorial representation of a data-set: each data-point being an artificial *antigen*, perhaps falling under the recognition region of some artificial *antibody*. These problem representations dominate all classical immune-inspired algorithms – clonal selection, negative/positive selection and idiotypic networks[1]. Indeed, such practice is prescribed as a methodology for AIS [17].

[1] There are some modern exceptions, e.g. Greensmith [30] and Nanas et al. [48], though neither of these works could be considered representative of the field of AIS.

In short, the "pattern-matching" aspect of ligand binding in the immune system is abstracted as some variation on vector correlation or distance metrics. Whilst pragmatic from a computational perspective, this abstraction both distances algorithms from the underlying biology and necessarily reduces to *augmented* instance-based methods of machine learning; methods that have largely fallen out of favour due to theoretically and practically undesirable properties (discussed in Sect. 3). Proposed immune-inspired augmentations do not address these undesirable properties; rather, they obfuscate them behind complex biological mechanisms and metaphors. As such, a potentially powerful method for building autonomous learning systems is ultimately undermined by foundational issues with its representational abstraction, rather than issues with the underlying immunological inspiration.

Our goal in this paper is to motivate and derive an alternative representational abstraction. To do so we will consider both the biological and machine learning perspectives, with a view towards integrating these ideas into a coherent computational model of *immunological learning* – rhetoric that has long inspired work in AIS, while remaining largely devoid of operational definition. As a precursor to more empirical work, we trace a formal path from the current status quo to our proposed abstraction and illustrate the theoretical properties of one explicit instantiation of our ideas.

The paper is structured as follows: in Sect. 2 we briefly review the biological foundations of ligand binding and the shape-space abstraction. In Sect. 3 we consider the theoretical, computational and practical issues with the shape-space as a computational abstraction. In Sect. 4 we motivate an alternative abstraction, based on recent advances in statistical learning. Finally, in Sect. 5 we bring these ideas together and analyse the properties of a toy algorithm based on our alternative abstraction. We conclude, in Sect 5.4 with some general comments and future work.

2 A Biological Perspective

The shape-space is simply a mathematical abstraction; its biological validity is a controversial issue. It would be easy to dismiss lack of biological fidelity as a straw-man argument from a computer scientist. However, we intend to demonstrate that a more biologically valid abstraction makes for a more valid computational abstraction. Before we do so, we briefly review the necessary biological ideas that motivate our work.

2.1 Motivating shape-space

Perelson and Oster introduced the shape-space as a simple quantitative model of the immune repertoire [50]. Its purpose was to answer questions such as "*given n receptors, what is the probability that a random antigen is recognised?*". The following discussion is largely taken from [51], where the reader is directed for further details.

Given a recognition region of volume v_i and the total volume of shape-space V, the probability p that an antigen is recognised is $p = \frac{v_i}{V}$. The probability P that an antigen is *not* recognised by one of n receptors is thus $(1 - p)^n$, which, assuming a Poisson distribution for antibodies, is approximated by e^{-np}. Experimental results suggest that $p \approx 10^{-5}$ of the immune repertoire respond to any given epitope, so this suggests that a value of $n = 10^6$ would be sufficient to ensure negligible chance of escaping detection. Such a repertoire would be "complete". This figure is in agreement with experimental estimation of the size of the smallest known immune system which, Perelson suggests, is because a smaller immune system would offer little protective advantage, e.g. if $n = 10^5$, then $P = e^{-1} \approx 0.37$.

A similar argument can be made for variations on the question "*what is the probability that a repertoire of n recognises all N foreign epitopes and none of N' self epitopes?*". The reader is directed to the literature for the details; here there are two key points to appreciate: (i) this model is a heuristic that does not attempt to define the parameters of the space – it only assumes that they could be defined in principle; (ii) the volume of the recognition region p is based on an experimentally derived number, not a geometric argument, which again, would require defining the degrees of freedom in the shape-space.

In the original shape-space, an affinity measure $\alpha(x, y)$ only implies some form of x and y. This is in stark contrast to the computational abstraction, where x, y and α all have explicit form. Later work in immunology simulated repertoires of bit-string shapes and string matching affinity functions (see e.g. [22]). It has been argued, convincingly, that these "physical" representations are unsuitable for computational learning [23, 24]. This really highlights where computer scientists' and immunologists' interpretations become incompatible. Immunologists assume that "shapes" are relatively small and simple; computer scientists assume that by abstracting the notion of shape, they can produce the same logical behaviour in a different domain, where the shapes are actually large and complex. In much of what follows, our contention is with the interpretation of shape-space by computer scientists, however it is worthwhile to pursue the biological argument some more, to justify our alternative.

2.2 Rethinking shape-space

The issues with shape-space as a theoretical abstraction were most notably asserted by Carneiro and Stewart [13]. Their argument is straight-forward: for a theoretical immunologist, deriving an affinity function and its dimensions from the limited experimental knowledge of known binding relationships is clearly ill-posed and data-dependent. Alternately, experimentally validating the parameters of the real shape-space is a "remote goal", which would likely result in a "highly complex, irregular and discontinuous" affinity function. Carneiro and Stewart criticise theoreticians' tendency to not distinguish clearly between these two, quite different, interpretations of shape-space, and thus, overlook the obvious difficulties with either. Furthermore, Carneiro and Stewart's experimental work suggests that shape complementarity is a necessary, but not sufficient, condition for recognition – there is a relational aspect, not accounted for by the classical lock-and-key metaphor.

Carneiro suggests that immunological models should be robust to the exact nature of the affinity relationship. In his own work, this took the form of binding occurring probabilistically without regard for position in shape-space. Receptors bind to multiple points that have no geometric relationship to each other. As such, the resulting model's immune-like behaviour is not bound to, or a side-effect of, any topological properties of the space it operates in [11, 12, 40].

2.3 The degenerate immune response

An issue gaining recent interest in both immunology and it's computational abstractions is *degeneracy* in ligand binding [21, 47]. It has been increasingly recognised that antibody can bind to many distinct epitopes (*poly-recognition*) and, similarly, an epitope can select many "specificities" of clone (*poly-clonality*). These ideas are quite contrary to the original clonal selection principle, which relies on approximate specificity to ensure self-reactive clones could be identified and deleted. For similar reasons, these ideas are also not comfortably expressed in the shape-space, where a clone's identity is its co-ordinate and binding strength is a function of metric distance in shape.

There are several authors, in both AIS and immunology, who embrace this degeneracy as an important feature of the immune system. However it remains to determine how immunological specificity emerges from these degenerate interactions [33]. Timmis and Andrews [47] have shown interest in how this degeneracy can be applied as a novel engineering paradigm. This work is still nascent, however it is clearly difficult to derive compelling engineering results based on interpreting immunologists' descriptions of phenomena. To complement this work, we seek something more formally tangible.

2.4 The cognitive immune system(s)

A large part of justification of AIS as a research paradigm is based on the rich cognitive metaphors attributed to the immune system – learning, memory, recognition, regulation and so on. By studying the components and mechanisms of the immune response, it is hoped that we can derive principles for producing autonomous computational systems with qualitatively similar behaviours[2].

There are two key figures in theoretical immunology that have promoted these cognitive metaphors: Francisco Varela and Irun Cohen. Both have quite different perspectives, but their influence on the AIS research program cannot be underestimated. Still, it remains for computer scientists to instantiate these influential ideas, and progress in this respect has been much less impressive. It

[2] Rather than immune systems for computers!

is our opinion, that the reason for this limited success is because these ideas do not translate into the shape-space abstraction. We briefly outline the key ideas of both authors here, which we will later reconsider under our alternative abstraction.

Varela was an influential cyberneticist, cognitive scientist and theoretical immunologist. His ideas were largely driven by his phenomenological/constructivist philosophical leanings and his early work with Maturana on the so-called *autopoietic* theory of the biology of cognition and behaviour [46]. In immunology, Varela (with Coutinho, Stewart and others) elaborated on Jerne's controversial Idiotypic Network theory, where lymphocyte co-recognition acts as an additional feedback mechanism to antigen stimulation [57, 18]. They developed simple mathematical models of how tolerance might emerge from the "self-knowledge" embodied in the immune repertoire – in contrast to the teleological "self-blindness" imposed by Burnet's clonal selection theory. Varela referred to the immune system as a "cognitive network" [56], much like the neural system, though an order of magnitude larger and inherently mobile. His work with Bersini is largely responsible for bringing these ideas into the computational domain, where they were applied with some success to control and reinforcement learning problems [7, 6, 5]. Like much of the cybernetics movement, this work was, in many respects, before its time and has somewhat languished. Varela et al's ideas significantly predate the current interest in complex interaction networks and emergent behaviours.

Irun Cohen is an experimental and theoretical immunologist who has expressed several radical ideas that have generated interest in the AIS community [16, 15]. The relation is quite natural, Cohen commonly refers to the immune system as a "computational system" (and also as a cognitive system, though this interpretation of cognition seems weaker than Varela's). Essentially, he sees the immune system as performing a non-classical distributed computation of the state of the body, with feedback mechanisms that govern the computation's evolution [35, 14]. The purpose of this computation is *maintenance* (inflammation, healing, garbage collecting and so on), with the immune response reduced to an extremal form of this maintenance. One of the key ideas that have caught computer scientists' attention is *co-respondence* [15] – how coherent system-wide responses emerge from the local interactions of diverse, contradictory components with limited sensing and effecting capabilities. Cohen is also vocal regarding the importance of degeneracy, pleitropy and other forms of beneficial redundancy.

Many of these influential ideas are more descriptive than formally (or experimentally) quantified. Later, we will return to these ideas of constructing internal representations and integrating diverse responses of simple components; but armed with a theoretical foundation for thinking about such ideas.

2.5 Ligand binding *in vivo*

It is important to realise that an epitope (binding region) is not a predefined object. It is an arbitrary discontinuous region on the three-dimensional surface of a molecule. It comes into being as an epitope by virtue of binding to a receptor, that is, in the context of a particular interaction [20]. The whole surface may have, so to speak, “epitope potential”. To appreciate what makes up the binding surface, it is useful to elaborate on the basics of protein structure.

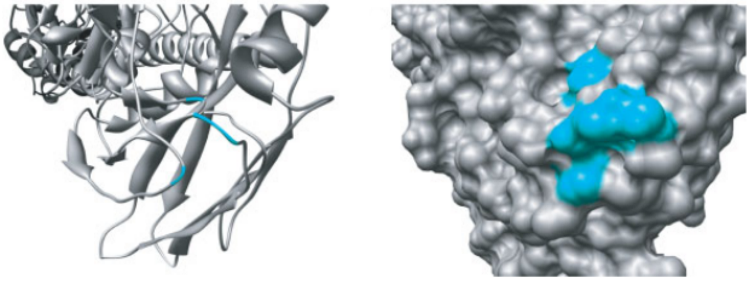

Fig. 1. A discontinuous epitope on a protein consists of residues that are distant in the primary sequence, but close when the protein is folded into its native three-dimensional structure. All of the residues are required for recognition by the antibody and thus are not epitopes on their own. Approximately 90% of ligands are discontinuous. Reproduced, in part, with permission from [20].

A protein is a long chain of shorter structures, called peptides, which are themselves, chains of amino acids. Laid out as a long chain, this is referred to as the protein's *primary structure*. During synthesis, the protein undergoes a complex folding process which, ultimately, results in a three-dimensional *tertiary structure* where some peptides are buried inside the structure and others are brought together on the surface. The significance of this is that different immune components sense different aspects of the protein: antigen presented to T-Cells are broken back down into peptide fragments; antibody, however, bind to the surface of the tertiary structure. As such, they recognise “features” that are distributed across the primary structure, but local (though not necessarily contiguous) on the Tertiary structure. To quote Janeway [36, Sect. 3.11]:

> Antigen recognition by T-cell receptors clearly differs from recognition by B-cell receptors and antibodies. Antigen recognition by B cells involves direct binding of immunoglobulin to the intact antigen and [...] antibodies typically bind to the surface of protein antigens, contacting amino acids that are discontinuous in the primary structure but are brought together in the folded protein. T cells, on the other hand, were found to respond to short contiguous amino acid sequences in proteins. These sequences were often buried within the native structure of the protein and thus could not be recognised directly by T-cell receptors unless some unfolding of the protein antigen and its ‘processing’ into peptide fragments had occurred.

So here we have two fundamental aspects to ligand binding that are not accommodated in the shape-space abstraction, and by extension, not addressed in current AIS models of immune behaviour. On the one hand, there is the different modes of sensing between B and T cells. On the other, the fact that antibody (and thus, B-Cells) do not bind to discrete units, but to appropriate structures brought together on the complex surface of a folded protein. It is this latter aspect that will be particularly relevant in what follows.

3 A Computational Perspective

The shape-space is not really immune inspired, it is just an intuitive mathematical abstraction. However, a good mathematical abstraction is not necessarily a good computational abstraction. A large sub-field of computer science is approximation of intractable mathematical operations that are, nevertheless, intuitive in principle. The so-called "Curse of Dimensionality" [3] is a recurrent motivation for such approximation – where a crucial quantity (e.g. the number of function evaluations in an optimisation problem) grows intractable with the dimensionality of the space. The curse has many different faces. Here we highlight the problem from a computational learning perspective, where low dimensional intuitions about "distance" and "density" turn out to be highly inadequate, thus undermining the very concepts that traditional AIS abstractions build upon.

3.1 The *non*-immune-inspired foundations of AIS

By treating ligands and receptors as points in shape-space, the "pattern matching" behaviour of an artificial immune system necessarily reduces to an augmented instance-based methods from machine learning. Instance-based methods (such as nearest-neighbour classifiers and density estimators – see e.g. [32]) are flexible, non-parametric and, as such, have a fairly natural mapping to biological metaphors: *the population is the model* and classification is performed across the population, at runtime. By exploiting a sense of *locality* between data-points they tend to be able to fit arbitrary complex decision boundaries or represent dense, nonlinear regions in unlabelled data.

This locality is a key parameter that can be used to trade-off representational power (high locality to model fine grained patterns) and stability (low locality to limit the effect of perturbations and noise in the data). Ideally, we would like high representational power and stability, but more often these factors are inversely proportional to each other. In the context of statistics, this is the classical bias-variance trade-off, though we avoid using these terms as they have quite specific technical meaning. The key point, is that how well instance-based methods perform depends crucially on just how local this sense of locality is.

Unfortunately, locality is where the curse of dimensionality hits hardest: as the dimensionality of a space increases, its volume increases exponentially faster. This simple fact compounds several issues that undermine classical instance-based methods, and thus, any AIS built upon a shape-space abstraction. We demonstrate these issues now.

3.2 Theoretical issues

To simplify our treatment, we will assume our shape-space and recognition region are cuboid rather than spherical. The spherical case has already been discussed by Stibor *et al.* in the context of negative selection algorithms [55]. The reason for our simplification is two-fold: firstly, in some respects the overall message is diluted by the mathematics of approximating the volume of a hyper-sphere and the particularly bizarre result that follows – as dimensionality increases, the volume of the hyper-sphere approaches zero. Secondly, the hyper-sphere tends to be implicit: at any given point we are usually only interested in straight-line distances as defined by an appropriate metric. Our intention is to illustrate that these simpler scenarios are also cursed.

In a rectangular space, the volume is obviously the product of the lengths L_i of each dimension – L^d if the space is cuboid. Similarly, the coverage of any $l < L$ cuboid region in this space is $\frac{l^d}{L^d}$. Clearly, the coverage of any fixed-size region grows slower than the volume as dimensionality increases – much slower if $l << L$. In turn, the stability of estimations made in that region also decrease: partly because the region covers less of the space; and partly because as data points are redistributed in the new dimension, they can easily move outside regions they where once covered by. In the worst case, there are no neighbours in any fixed locale, and thus, no generalisation power – our learning machine reduces to a rote lookup table.

Conversely, in order to capture the same volume of space as dimensionality increases, the range of coverage of each dimension rapidly approaches 100% [32]. For example, to capture 25% of a two dimensional "volume" requires covering 50% of each dimension. More generally, to capture v volume of a d dimensional cuboid space requires covering $v^{\frac{1}{d}}$ of each dimension (e.g. $0.5 = 0.25^{\frac{1}{2}}$). Even for modest dimensionality and small volume, this value rapidly approaches 1.0, as illustrated in Fig. 2. This dramatically reduces the representational power of our learner as desireable non-linearities from localisation are lost. The learning machine reduces to either a poor approximation of a linear classifier, or worse, the trivial *maximum a priori* classifier[3].

These effects of dimensionality are not limited to our representational power and stability. Any metric defined across an increasing volume becomes increasingly meaningless as data points tend to become equidistant [1, 8]. Figure 2 illustrates the difference in pairwise distances between 10 uniformly distributed points as the dimensionality of the unit-space is increased. It is clear that all distances are converging. Between the pressures on both stability and representational power of the recognition region, this results in a very fine line between

[3] Always predicting the most common class. This classifer is the epitome of stable – its decision does not depend on the data – but is clearly a poor learning algorithm. These reductions are further illustrated in Sect. 4.1.

a region capturing either none, or all, of the datapoints. That is, any discriminatory power in pairwise distances is lost[4].

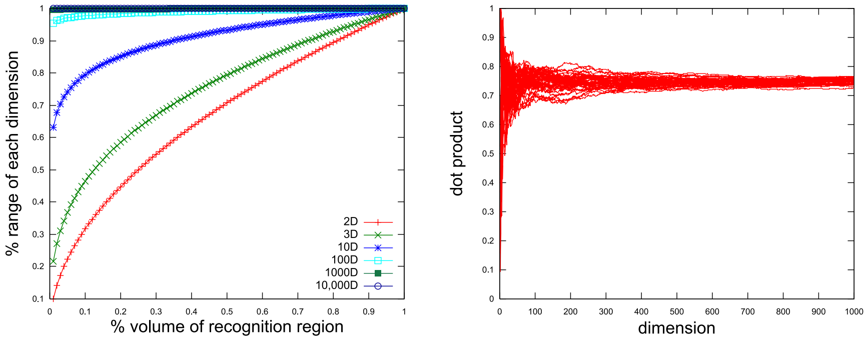

Fig. 2. The curse of dimensionality in shape-space. **Left:** $v^{\frac{1}{d}}$ for different dimensional spaces, adapted from [32]. Capturing even a very small percent of a modest dimensional space requires significant coverage of the range of each dimension. **Right:** convergence of pairwise measures (in this case, the dot product) between 10 uniformly distributed points as the dimensionality increases. See text for a description of the consequences of these effects.

This convergence to equidistant is in some respects an artifact of the unit-space: outside the unit hypercube, the pairwise distances clearly grow as component dimensions are added; however, the relative significance of those distances is reducing at a much faster rate. Consider a low dimensional analogue: to be 10 feet apart may be a significant measurement for two people in a room, but may be indistinguishable from noise for a GPS satellite. In this limited example, adding dimensions is somewhat akin to moving exponentially further away.

Similarly, an inconsequentially small amount of noise, added to many dimensions, can still add up to a significant displacement from the original position. In high dimensions, a scalar metric cannot differentiate between two vectors that differ slightly across all dimensions (and may be the same after accounting for noise) or differ significantly on just a few dimensions (and are clearly different in some respect).

The two key assumptions in instance-based methods are based on low-dimensional intuitions: that there are dense regions in the space that can be generalised, compressed or sparsely represented; and that the distance between points is a meaningful proxy for comparison, discrimination and localisation. The validity of both assumptions is, unfortunately, a rapidly decreasing function of dimensionality.

[4] Of course, one can still select e.g. the k “nearest” neighbours. The point, is that they may not be significantly nearer than the other $(N-k)$ neighbours – an implicit assumption in these methods.

3.3 Computational issues

In an n-dimensional shape-space, the search space for the immune repertoire is of the order $O(c^n)$ where c is a constant (e.g. $c = 2$ in a binary space). This exponential scaling is computationally abhorrent for the typically large n involved in machine learning. Even if we assume, as is normal, that the active repertoire at any time is a sparse sampling of the shape-space – in high-dimensional space antigen are also distributed sparsely. An n-dimensional hypercube has 2^n "corners" where most of the volume is concentrated – more corners than antigen.

It is often not possible to scale the available data at the same rate as the increase in volume. Even if the data is available, instance-based methods must keep all (or a representative majority) of the data in memory and, due to lack of any *a priori* model, must perform "lazy learning" at runtime. So both space and time scalability of these algorithms is quite poor. Even an $O(m)$ algorithm that scales linearly with the size of the dataset, hides the fact that m should, ideally, scale in proportion to the increase in dimensionality. In a typical AIS algorithm, generation of the affinity matrix ($O(m^2)$ space), pairwise interactions ($O(m^2)$ time) and the inner-loops of hyper-mutation and affinity maturation (say e.g. $O(km)$ in space and time, where k is the average mutations per clone and each mutation must be evaluated for fitness) place these algorithms far from $O(m)$.

Indeed, any algorithm that scales, in space, with the number of data-points is a questionable choice for representing a system that exhibits life-long learning with no terminating condition: naively, $m \rightarrow \infty$. One can impose conditions and mechanisms to "forget" old and irrelevant data-points, but in addition to being vast, the shape-space is relatively uninteresting. Points appear, persist and decay without elucidating on the underlying process that generates these points.

Given that m should scale in proportion to n, one might consider working in the space of feature-feature relations. Here, there is a substantial initial computational cost (e.g. $O(n)$ or $O(n^2)$ rather than in $m << n$) but this cost is likely constant, or a slowly growing function of time. This space is inherently more dynamic: each new feature vector reinforces the relations between its features, and these relations illicit underlying structure in the data generation mechanism. To our knowledge, only Nanas et al's *Nootropia* [48] performs in such a space.

3.4 Practical Issues

In some respects, the preceding arguments are purely academic. Instance-based methods are still popular in data-mining and applied machine learning. There are perhaps no other methods that can compete, on an intuitive level, for a practitioners attention – even if that intuition turns out to be wrong. Some very general practical issues of instance-based methods also translate comfortably into the AIS setting.

Firstly, it is common wisdom that these methods can be very sensitive to noise – particularly dimensions that are not correlated with the learning task, but nevertheless the algorithm attempts to accommodate. Typically, data-mining and applied machine learning begins with a pre-processing stage, where data-derived statistics are used to reduce the dimensionality prior to learning – either by selecting only relevant features (*feature selection*) or by projecting the data onto a low dimensional space of abstract feature combinations (*feature extraction*). A reasonable question might be why a "learning algorithm" should not be expected to derive these relevancies during its execution. This question is particularly pertinent to any algorithm that claims to be based on the immune system, where noise is a hallmark of its operational environment (e.g. degeneracy, redundancy and pleitropy). It seems reasonable that an immune-inspired system should be robust to such noise and should not depend excessively on ad-hoc external preprocessing, particularly if this can be integrated into its internal dynamics.

Secondly, and more crucially, it is well appreciated that the success of applying these methods is very sensitive to the choice of representation and metric. It must be so, because everything else follows from these definitions. Such warnings are echoed in the AIS literature: Timmis and Freitas recently stressed how important it is for practitioners to *not* accept the standard representations and affinity measures when applying AIS to data-mining problems [24]. They suggest that these must be carefully designed by the practitioner prior to plugging them into the chosen AIS algorithm. But therein lies the rub: for most learning tasks, *representation is the difficult problem.*

Almost by definition, a good representation makes learning simpler; at the very least, it significantly reduces the impact of a particular algorithm choice. It hardly seems satisfactory to ask the practitioner to *a priori* solve the learning problem, so that the algorithm need only churn through the numbers. Simplified learning tasks can be solved without recourse to complex immunological mechanisms. However, Varela and Cohen have promoted the view that the immune system can derive its own internal representations as part of its autonomous operation. Understanding how this can be achieved *in silico* may well benefit from insight into immunological mechanisms.

In both cases, the practical realities of the shape-space abstraction directly contradict the immunological inspiration.

4 A Statistical Learning Perspective

Having now discussed the theoretical, computational and practical issues of the shape-space abstraction, we begin to look towards an alternative. The statistical learning community have approached these problems from several angles. It is unlikely that the dynamics of the immune system will reduce to standard statistical procedures, but it will be useful to appreciate these ideas on their own terms, before moving on to our own abstraction.

4.1 Simple classifiers

In many respects the parametric linear classifier and the non-parametric nearest-neighbour classifier are the Alpha and Omega of statistical learning. Their relation is easily understood algebraically if we dismiss the ad-hoc modifications and abstractions that define specific algorithms. We begin with the ubiquitous fitting of a line by the method of least-squares

$$y = X'w$$

where X is an $n \times m$ column matrix of data vectors, w an n-dimensional weight vector (to be found), and y an m-dimensional vector of class labels or function values for each column of X. We will assume $y = [-1, +1]$ to blur any distinction between classification and regression. This equation is solved using the Moore-Penrose Pseudo-Inverse of X[5]

$$\begin{aligned} w_* &= X^{-1}y' \\ &= (XX')^{-1}Xy' \\ &= X(X'X)^{-1}y' \end{aligned}$$

Plugging the optimal solution w_* back into the original equation we can calculate the deviations of our estimated $\hat{y}$ from the given y. It is well known that this method minimises the squared error of such deviations: $\sum_i (y_i - \hat{y_i})^2$. Alternatively, using a new matrix $\hat{X}$ of test data, we can estimate the unknown $\hat{y}$ values:

$$\begin{aligned} \hat{y} &= \hat{X}'w_* \\ &= [\hat{X}'(XX')^{-1}X]y' \\ &= \hat{X}'X[(X'X)^{-1}y'] \\ &= \hat{X}'X\alpha \end{aligned}$$

The first bracketed term is the so-called "hat matrix" because it takes y onto $\hat{y}$. Note that this matrix is completely data dependent: it is the same for all possible y (dichotomies over X). The second bracketed term highlights the derivation of a vector α which we will use below. We are now in a position to demonstrate the relaxation from a parametric linear classifier to a non-parametric "non-linear" nearest-neighbour classifier

$$\hat{y} = \langle w_*, \hat{x} \rangle = \langle (\sum_n \alpha_i x_i), \hat{x} \rangle = \sum_n \alpha_i \langle x_i, \hat{x} \rangle \approx \sum_n y_i \langle x_i, \hat{x} \rangle \approx \sum_{\langle x_i, \hat{x} \rangle \geq \epsilon} y_i$$

[5] If the reader is not familiar with this technique, the pseudo-inverse reduces to the regular inverse for a fully determined system and behaves sensibly in over and under-determined systems: solving $\text{argmin}_w \, ||y - X'w||_2$ when there are no solutions, and $\text{argmin}_w \, ||w||_2$ s.t. $y = X'w$ when there are infinite solutions.

where the first step relies on the fact that $w_* = X\alpha$ can be represented as a linear combination of the training data. The second step is simply a regrouping of terms. The approximation then occurs: first by replacing the matrix inversion in the derivation of α with the unprojected y values; and then by ignoring training data where the dot product is below – or equivalently, where the "distance" is above – a threshold ϵ.

In this case ϵ is the radius of our "recognition region". Moving backwards through this derivation, we can now see how increasing this radius in response to increased dimensionality, first approaches the *maximum a priori classifier* and then, by weighting y_i values by the dot product $\langle \hat{x}, x_i \rangle$, approaches a poor approximation to the linear classifier. By avoiding the necessary (but costly) matrix inversions, we are essentially "solving" $y = X'w$ as $w = Xy$.

Now this is hardly the status quo of modern statistical learning. However, it is a small step to appreciate one crucial aspect of the status quo: producing powerful classifiers need not require complex and advanced learning mechanisms, but rather, requires tackling the real, difficult problem of representation.

4.2 A simple classifier with powerful representational capabilities

The highlight of our previous derivation was a duality between parametric linear and non-parametric, "non-linear" nearest neighbour classifiers. It is a small notational and conceptual step[6] to go from

$$\hat{y} = \sum_i^n w_i \hat{x}_i = \sum_j^m \alpha_j \langle x_j, \hat{x} \rangle \tag{1}$$

to a more abstract and richer

$$\hat{y} = \sum_i^n w_i \beta_i(\hat{x}) = \sum_j^m \alpha_j K(x_j, \hat{x}) \tag{2}$$

Both of these abstractions allow us to introduce non-linear transformations in the representation (where we need them) but maintain linearity in the parameters (where it is analytically convenient to do so). A simple linear classification in the transformed space is equivalent to a complex non-linear classification in the original space.

For example, on the right-hand-side of Eq. (2) Kernel methods [53] expect the kernel function K to return a dot-product in *some* high-dimensional non-linearly transformed space, while employing the "kernel trick" to avoid the computational

[6] Unfortunately, it is not such a small step formally. The full details are outside the scope of this paper, suffice to say, it is mostly additional layers of abstraction that add little, in the way of understanding, over the linear algebra (but add a lot in terms of generality and power). The generalisation of linear models with functions is common practice in statistics, see e.g. [32, Chapter 9]. A concise description of Hilbert space (where functions form a vector space) can be found in [37, Chapter 5].

burden of explicitly working in that space. The “trick”, is to derive a high dimensional dot product in terms of the low dimensional vectors – e.g. the polynomial kernel $K(x, y) = (1+\langle x, y\rangle)^p$. However, this ingenious technique for transforming low(er) dimensional, non-linearly separable data into high(er) dimensional linearly separable data only attacks one aspect of the curse of dimensionality. The computational cost of *increasing* the dimensionality is circumvented; however, almost all popular kernel tricks work by transforming either the dot-product or Euclidean distance of the original vectors. If the arguments to the kernel function are already high dimensional, then these measures are already cursed, in the sense discussed in Sect. 3.2, prior to any transformation.

Conversely, on the left-hand-side of Eq. (2), any nonlinear transformation implied by a kernel function can be explicitly represented by $\beta_i(\hat{x})$ – with the computational burden of working in the higher dimensional space. In fact, β_i can represent any transformation, with $\beta_i(\hat{x}) = \hat{x}_i$ reducing to the standard linear model. An alternative strategy then, is rather than enriching the feature space of our data (and parameters), to consider each β_i as a classifier and transform directly onto a classification decision [29]. Our search space is enriched to the space of classifiers; our regression function becomes a weighted vote amongst an *ensemble* of classifiers. Crucially, observe that this reintroduces the notion of a *population*, but not a population based on the training instances.

4.3 Many simple classifiers with weak representational capabilities

The general goal of ensemble methods is to improve on the classification performance of a single learning machine, by integrating the results of many diverse learners. Diversity can be specified in different ways: different learning algorithms; the same algorithm with different parameters or trained on different subsets of the training data, and so on (see e.g. [54]).

Boosting [26, 38] has emerged as one of the most radical and successful instantiations of ensemble learning. The radical aspect is the formal demonstration of the equivalency of *weak* and *strong* learnability: a weak learner, performing only slightly better than random guessing, can be aggregated into an arbitrarily strong learning algorithm[7].

$$\hat{y} = strong(x) = \sum_i w_i weak_i(x)$$

Intuitively, Boosting can be seen as the integration of many cheap heuristics that will often fail, but have some edge over random guessing (rather than the integration of a few, strong classifiers as employed by ensemble methods in general). There are still gaps in the theoretical understanding of Boosting, but it is clear that a key aspect of it’s success is that, during training, learner diversity

[7] This formal demonstration only holds in the PAC learning framework (see [38, 52] for background and proofs), though the same intuition has been applied very successfully outside of this framework.

is augmented by dynamically reweighting the training examples (see Alg. 1). During training, successfully classified data have their weight decreased, forcing learners in later iterations to compensate any predecessor's bias and concentrate on data that is causing continued classification error. During classification, integration across the ensemble increases the confidence in any particular classification decision, by averaging out the variance of the weak components.

```
1. Initialise with uniform distribution over data
while $error > desired$ and $rounds < max$ do
    2.1 Generate a new weak learner with current distribution
    if $error \geq 0.5$ then
        continue
    end
    2.2 Weight learner based on performance
    2.3 Reweight data distribution based on performance
    2.4 Add learner to ensemble
end
```

Algorithm 1: Pseudo-code for the Adaboost algorithm

Contrary to the trade-off inherent in instance-based methods, we see both an increase in representational power (through diversity) *and* an increase in stability (through integration). Contrary to the theoretical limit of standard weighted majority voting [43], the diversity in data seen by each learner allows the boosted system to perform *better* than its single best component.

With a background in computational learning theory, Boosting shares its foundations with the seminal online learning work of Littlestone, Warmuth et al. [43, 42]. Leading from this, the training procedure has also been shown to have a game theoretic interpretation as learning an optimal mixed-strategy in iterated games [25]. The statistical perspective of gradient descent in function (rather than parameter) space, most notably pursued by Friedman [28] and Breiman [9], is easier to state given our preceding exposition and leads to a natural connection between Boosting and dictionary-based basis decomposition methods, popular in signal processing and compression [27, 58]. Exploring this connection allows us to emphasise the representation learning behaviour within the realms of algebra, rather than introduce the nomenclature of the Boosting community. This will be the last piece of the puzzle in developing our representational abstraction.

4.4 Learning Representations

Consider a vector of digital samples (e.g. an audio signal). Our goal is to find a compact representation of this signal from a "dictionary" of atomic basis functions. Given a signal s, dictionary β and a representation r of basis co-efficients, our two operations are *analysis* ($r = \beta s$) and *synthesis* ($s = \beta^{-1} r$). A good representation is sparse (many of the basis coefficients are zero) and it may be acceptable to trade-off some reconstruction error in the synthesis stage for additional sparsity (e.g. lossy compression). A good basis generalises over possible signals and is fast to compute and invert.

Classical methods include the Spectral Decomposition, Fourier analysis and Wavelet analysis [44]. However, any particular form of basis function may render some representations difficult (e.g. Fourier basis have difficulty representing sharp spikes as sine waves) so we would like to have many diverse basis functions; more than we actually need to represent a given signal. Such an "overcomplete" dictionary renders any solution non-unique – a given signal will have many possible decompositions [41]. Using methods for solving under-determined systems of equations, one can add additional constraints to regain uniqueness (such as minimum L_1 norm of r) and attempt to solve by global optimsation. An alternative strategy is to greedily construct a sparse representation. A canonical algorithm in this respect is Matching Pursuit [45] (Algorithm 2).

```
1. $r = []$
while $||s||_2 > \epsilon$ do
    2.1 $\beta_t = \text{argmax}_{\beta_i} \langle \beta_i, s \rangle$
    2.2 $w_t = \langle \beta_t, s \rangle$
    2.3 $s = s - w_t \beta_t$
    2.4 $r.add(w_t, \beta_t)$
end
```

Algorithm 2: Pseudo-code for the Matching Pursuit Algorithm

The relation to Boosting is quite straight-forward: find the basis with maximum correlation with the signal; weight it by that correlation; subtract the weighted basis from the existing signal; repeat till the residual is small enough[8]. The end result is a sparse representation of the original signal in terms of the chosen basis functions. Convergence is guaranteed if the basis span the space, though the rate of convergence depends on the "coherence" between the signal and the dictionary elements [45].

The overarching issue with matching pursuit is finding the *best* basis at each cycle of the algorithm's main loop. Typically, only a subset of the dictionary is considered for searching, or some (hierarchical) structure is imposed on the dictionary to limit the search space [39]. The dictionary can be fixed *a priori* or, more appropriate from our perspective, adaptively learned in response to the incoming signals (see e.g. [2]).

Matching Pursuit is necessarily sub-optimal. Of course, optimality is always with respect to a measure and renders solutions fragile to changes in that measure or it's underlying assumptions. In general, sparse representations from overcomplete dictionaries can be very robust to noise and perturbations of the data.

[8] The primary difference with boosting is that the "residual" in matching pursuit is the reconstruction error $||s_{t-1} - w_t \beta_t||_2$. As a supervised learning strategy, the "residual" in boosting is a function of the classification error, e.g. $||y_{t-1} - f_t(X_t)||_2$.

Having a large set of specialised bases carries gains in flexibility over a small set of general bases. In contrast to “purer” algebraic orthonormal basis, overcomplete dictionaries can more efficiently represent non-orthogonal density distributions. For example, an orthonormal basis, such as the eigenvectors, can provide perfect reconstruction with a minimal set of bases; but that representation will necessarily be distributed across *all* bases. A coherent non-orthogonal dictionary can isolate the representation in a select few [41].

5 Representation in the (Artificial) Immune System

The preceding sections have discussed problems with learning models built upon the shape-space abstraction and traced a reasonably formal path from these simple instance-based methods to modern learning methods. The key idea was powerful representations rather than powerful algorithms. We then took this idea to an extreme of measurably weak algorithms, that nevertheless maintain powerful representational capabilities across the ensemble. We also demonstrated how, within the same framework, these representations can be learned in addition to any decision function built on top of that representation.

With the preceding statistical insights, and recalling our earlier discussion of ligand binding, our proposed change of abstraction is now quite easy to state:

Proposition 1. *Epitopes are not points in a high-dimensional feature space, they are a subset of surface correlated peptides. The immune repertoire is not a population of centroids, prototypes, or support vectors, but an overcomplete dictionary of basis functions; an ensemble of weak learners.*

Given the preceding exposition, we hope that this proposition now seems inevitable; even obvious. It is sufficiently general to accommodate many instantiations, though in Section 5.3 we will elaborate with a more concrete example. First we wish to highlight the general benefits of this change of perspective, away from the details of a specific algorithm.

5.1 Potential contribution *to* Artificial Immune Systems

Quite simply, in order to construct an internal representation one needs building blocks: this is precisely what basis functions are. There is nothing “constructive” in mirroring training data and generalising dense regions. A repertoire of basis functions can partition the space; representing compound structures as the sum of diverse, overlapping subspaces. This leads neatly to our previous discussion in degeneracy, redundancy and pleitropy. Basis functions incur many-to-one “binding” (*poly-recognition*) and many-to-one “activation” (*poly-clonality*) by definition – many data points overlap an individual bases’ subspace and an individual data point overlaps many base subspaces. When that basis is also over-complete, beneficial redundancy and robustness are also formally demonstrable, as previously discussed.

We invite the reader to consider Cohen's *co-respondence*, where coherent system-wide responses emerge from the interactions of simple components with contradictory responses [15]. What better way to base a quantification of such an ideas than with the strength of weak learnability? All the necessary ideas are in place: (*i*) randomly generated weak components; (*ii*) the feedback cycle between antigen presentation, clone activation and antibody production that modifies the the data distribution in the environment; and (*iii*) integration across the population to increase confidence in the final system's output.

As we have repeatedly stressed throughout, the shape-space antagonises realising these ideas as computational abstractions: it ties ligand and receptor identities to co-ordinates; assumes affinity is isometric centred on these co-ordinates; and portrays the repertoire as a collection of points attempting to cover the space. We tentatively suggest that the proposed abstraction also renders some of the open problems and future directions (e.g. [24, 31]) of AIS tractable. We will illustrate this point further in Sect. 5.3.

5.2 Potential contribution *from* Artificial Immune Systems

As was noted in a recent workshop report [19] under the heading "What is real-time autonomous learning?":

> ...in most ML [machine learning] applications, the intelligent aspects of learning are managed by the human supervisor (and not by the learning algorithm). Typically this human supervisor must: select the training examples; choose the representation of the training examples; choose the learning algorithm;choose the learning rate; decide when to stop learning and choose the way in which the performance of the learning algorithm is evaluated. This absolute requirement for a human expert supervisor precludes ubiquitous use of ML. It also means that ML has not yet captured the essence of learning.

Clearly, the immune system is real-time autonomous. It is, perhaps, this autonomy that should be the inspiration for algorithms based on immunological metaphors and models, rather than, or in addition to, superficial two-class classification or anomaly detection behaviour.

However, in demonstrating how the shape-space inhibits realising these ideas, we justified our change of abstraction with reference to statistical methods that are also *not* autonomous. Both Boosting and Matching Pursuit are greedy algorithms. Matching Pursuit does not learn how to structure its dictionary; is only concerned with efficiently representing individual signals, rather than the signal "environment"; and because of its myopic nature, errors in earlier iterations can persist and propagate. Boosting is usually framed in a batch supervised learning framework, resulting in a finite training period prior to static deployment (although, see [49]). Boosting can also be very sensitive to outliers and noise in the class labels. None of these aspects are conducive to autonomous learning.

By reframing repertoire generation and management in the context of learning an adaptive dictionary of base functions or classifiers (rather than prototypical points in a shape-space), we think the immunological perspective has much to contribute to this autonomy – somewhere between the analytical extremes of greedy and global optimisation. Indeed, this is an area where AIS already contribute, albeit in a different search space.

Crucially, this change of abstraction does *not* render previous AIS contributions invalid[9]. By working at a lower level, we simply reinterpret what structures these immune-inspired processes are performed over.

5.3 Sketch of an alternate system

To better illustrate these ideas we will sketch a caricature of a system based on our previous algebraic analysis – with the explicit understanding that we are not suggesting that AIS should reduce to algebra, or that the immune system is in the business of least-squares regression. That said, any qualitative similarity with formal techniques, at least, helps intuit that the system might be doing something useful, which can be difficult from an arbitrary algorithm based on a textbook description of biological processes.

We return to the linear classifier, but leave nearest-neighbour behind

$$\begin{aligned}\hat{y} &= \hat{X}'w_* \\ &= \hat{X}'[(XX')^{-1}][Xy] \\ &= \hat{X}'G^{-1}\hat{w}\end{aligned}$$

where we have highlighted two components, G and $\hat{w}$, because they are both amenable to online updating. Firstly, the pseudo-weight vector $\hat{w} = Xy = \sum_t y_t x_t$ represents the correlation between each feature (i.e. "peptide") and the class label. Secondly, the matrix $G = XX' = \sum_t \mid x_t\rangle\langle x_t \mid$ represents the correlation between features[10]. Translating to an online setting is straight-forward:

$$\hat{w}_t = \hat{w}_{t-1} + y_t x_t \tag{3}$$

$$G_t = G_{t-1} + \mid x_t\rangle\langle x_t \mid \tag{4}$$

To understand the qualitative effect of inverting G, we must first consider our representation learning process. The famous Spectral Theorem [4] states that G can be decomposed as a superposition of basis vectors

[9] e.g. Structural and parametric plasticity, exploration and exploitation in hypermutation and affinity maturation, meta-stability, lifelong learning, rapid adaptation and homeostatic behaviour. See [17] for a general overview of much of these ideas.

[10] We have used braket notation to emphasise the outer-product $\mid x\rangle\langle y \mid = xy'$ resulting in a matrix – not to be confused with the scalar inner-product $\langle x, y\rangle = x'y$.

$$G = \sum_i \lambda_i \mid \varphi_i\rangle\langle\varphi_i \mid .$$

The canonical basis are, of course, the eigenvectors, and the eigen-decomposition is a standard method of inverting a matrix by inverting the eigenvalues:

$$G^{-1} = \sum_i \frac{1}{\lambda_i} \mid \varphi_i\rangle\langle\varphi_i \mid .$$

The eigen-decomposition is also the heart of many unsupervised learning techniques and common pre-processing steps prior to supervised learning (e.g. Principle Component Analysis [32]).

However, we will be concerned with an alternative decomposition – the generation and regulation of basis functions is now analogous with the development of the immune repertoire. Ignoring the complexities of repertoire meta-dynamics[11], this can be considered as a Matching Pursuit like process over G, as illustrated in Algorithm (3). We have deliberately left the details of this algorithm undefined – the point we wish to make is that (*i*) it's form is similar to many AIS algorithms; (*ii*) the parallel, evolutionary search and repertoire meta-dynamics of AIS is contrary to greedy optimisation methods such as Matching Pursuit (Alg. 2) and Boosting (Alg. 1); and (*iii*) the search space and representation are completely different from traditional AIS.

Assuming a repertoire of basis functions, we turn to the qualitative effects of inverting G. The inversion of the eigenvalues $\frac{1}{\lambda_i}$ has a simple effect on the summation – dominant eigenvectors, with large eigenvalues, are de-emphasised (large denominator), and weak eigenvectors, with low eigenvalues, are emphasised (small denominator). Typically, the dominant eigenvectors are interpreted as capturing coarse grained (high variance, low frequency) structure in G; the weaker eigenvectors interpreted as noise. Given each B-Clone is now a basis (where λ_i represents the clone's contribution to approximating G) we can achieve a similar effect to the algebraic inversion by exploiting the well known "bell-shaped" dose response curve of the immune system [40]. Under this curve, the extremes of overt and weak recognition invoke little response, but clones with intermediate levels of stimulation flourish. In effect, we are tolerating *both* coarse grained structure *and* fine grained noise. This is an intuitively reasonable strategy for extracting meaningful structure, without reference to any immunology.

Although we do not want to digress into a discussion on possible supervisory or feedback signals, the vector $\hat{w}$ also has an intuitive interpretation. In the lymph nodes, T-Cells co-ordinate between the adaptive repertoire and other innate cells that sample peptides and chemical signals from the tissues. The net effect of this sampling is an association between peptide fragments and levels of

[11] The "meta-dynamics" cover random generation, evolutionary hyper-mutation, competitive exclusion and regulation of the repertoire. It is one of the most well modelled aspects of the adaptive immune system and the inspirational basis of many AIS.

```
$repertoire = []$
while $true$ do
    for $Clone\ c_i\ in\ repertoire$ do
        $s_i = stimulation(c_i, G_t)$
        $proliferate(c_i, s_i)$
        $mutate(c_i, s_i)$
        $\Psi = \Psi + secrete(c_i, s_i)$
    end
    $G_{t+1} = G_t - \gamma\Psi$
    $cull(repertoire)$
    $populate(repertoire)$
    $metadynamics()$
end
```

Algorithm 3: Illustration of basis discovery and decomposition in the immune repertoire. The form is similar to standard AIS algorithms, however the search space and representation are quite different. The evolutionary-based search and repertoire metadynamics is different from greedy optimisation.

endogenously defined "danger". This is not unlike the correlation between class labels and features in $\hat{w}$, though the mechanism is quite different from $(Eq.3)$.

Finally, in least squares, the projection of $\hat{w}$ across G^{-1} then mixes these two forms of correlation, resulting in final, optimal weighting w_* of features. Let us consider an algebraically equivalent, though *semantically* quite different, interpretation of this step, for a system attempting to classify a point $\hat{x}$:

$$\begin{aligned}
\hat{y} &= \hat{x}' w_* \\
&= \hat{x}' G^{-1} \hat{w} \\
&= \hat{x}' \left[\sum_i \frac{1}{\lambda_i} \mid \varphi_i \rangle \langle \varphi_i \mid \right] \hat{w} \\
&= \sum_i \frac{1}{\lambda_i} \hat{x}' \mid \varphi_i \rangle \langle \varphi_i \mid \hat{w} \\
&= \sum_i \frac{1}{\lambda_i} \langle \varphi_i, \hat{w} \rangle \langle \varphi_i, \hat{x} \rangle
\end{aligned}$$

Note that the classification decision depends on the contribution of each basis in approximating G and the correlation of each basis with both the query point $\langle \varphi_i, \hat{x} \rangle$ and the vector of feature-label correlations $\langle \varphi_i, \hat{w} \rangle$. The interpretation is straight-forward: each B-Clone's response is a function of its fitness λ_i, its recognition of $\hat{x}$ and the contextual state $\hat{w}$ of the peptides that make up its receptor. Note also, that without leaving the algebra we have a coherent, albeit simplified, formal basis for many key ideas in AIS: online learning, degeneracy, constructive representations and contextual recognition.

Simplifying the final step, which is outside the scope of this paper, we abstract our basis as a weak learning algorithm, where weakness is imposed *a priori* by sensing only subspaces

$$\hat{y} = \sum_i \varphi_i(\hat{x}; \hat{w}, \lambda_i) \tag{5}$$

The result is an expression reminiscent of co-respondence and the strength of weak learnability. Although we have necessarily glossed over many details, let us analyse the gross properties of this algorithm.

Recall that each feature is analogous with a peptide. A given immune repertoire will have roughly $O(n)$ T-Cells capable of recognising distinct peptide fragments. A B-Clone receptor's capacity to bind to epitopes is entirely dependent on appropriate peptides being brought together on the surface of environmental proteins. This information is encoded in G[12]. In a maximal simplification, if we assume that a B-Clone can recognise aspects of k nearby surface peptides, then we can also expect a search space of the order $O\binom{n}{k}$ for the immune repertoire. This scaling is much slower than the polynomial $O(n^k)$ in the worst case[13]. Furthermore, this worst case that is only realistic if each peptide is uniformly likely to appear close to another, which almost certainly never holds in both the biological and computational perspective. Redundancy is typically rife: it is this redundancy that makes learning feasible.

Note that in this form, recognition is no longer a function of an affinity metric in high dimensional space. Each B-Cell receptor defines a different k-dimensional *subspace*. This is the individual cell's weakness as a component, but the strength of the repertoire as a whole: it can construct and adapt its representation based on many low-dimensional perspectives.

Of course, nothing is for free. The implicit cost in all of the above is that G scales $O(n^2)$ where $n >> m$. This may not be so bad. In certain domains, such as statistical natural language processing, the dimension n is very high, but sparsely represented in any given vector. This is because different vectors populate different subspaces – a notion lost in the traditional shape-space. This means that G may also be very sparse, and often we can explicitly control this sparsity by choosing how long-range we are still willing to accept features as "correlated". The lower the correlation distance, the sparser G becomes. When G is sparse, it can be stored and updated independently of both n and m – roughly $O(k^2)$ where $k << n$ is the average number of features.

In conclusion, we highlight several aspects of this toy algorithm that are not so toy-like:

[12] As a sum of vector outer-products this is essentially just feature correlation. However, the algorithm is indifferent to how G is constructed. A much richer matrix-based (or graph-based) representation of component peptides and their surface dependencies is possible, given that X is now implicit.

[13] Given $n! < n^k(n-k)!$ then $\binom{n}{k} = \frac{n!}{k!(n-k)!} < \frac{n^k(n-k)!}{k!(n-k)!} < n^k$.

- This system can perform *online* – it is not a batch learning strategy (unlike regular parametric methods) and does not scale with the size of the dataset (unlike non-parametric instance-based methods).
- This basis generation method is evolutionary adaptive and performed in parallel (unlike Matching Pursuit and Boosting). The repertoire meta-dynamics ensure complexity is regulated and non-beneficial redundancy is reduced.
- There are separate processes generating G (Eq. 4), $\hat{w}$ (Eq. 3) and the repertoire (Alg. 3). The result is an unsupervised *representation learning* process, that is directed by an (unspecified) reinforcement or supervised *task learning* process. The system can also perform "semi-supervised" because G is enriched independently of $\hat{w}$.
- This system reduces to something statistically sound when the immunological dynamics are removed. Furthermore, the immunological mechanisms are justified based on the equivalent algebraic and algorithmic mechanisms, not by appealing to metaphors.

Taken together, we propose this is as a possible sketch of an online autonomous learner: by playing to the strengths of the immune metaphor it can learn its representation and adapt to changes in the underlying data generating mechanism; by reconsidering the representational foundations, it becomes more computationally and analytically powerful.

5.4 Conclusion and future work

We have illustrated how the shape-space is practically and theoretically inferior as a representational abstraction for machine learning in AIS. No amount of immune-inspired mechanisms can compensate for this low-level inferiority and, as such, the potential value in these mechanisms is undermined. This is an unnecessary waste, as the potential value of immune-inspired processes is not tied to the shape-space. We have attempted to derive an alternative representational abstraction that, both, empowers existing AIS algorithms and provides operational definitions for some of the rhetoric that inspires contemporary AIS.

The ideas discussed in this paper are somewhat reminiscent of AIS' roots in Holland's Learning Classifier Systems (LCS) [34, 10]. However, our derivation is based on modern advances in computational and statistical learning that have only become widespread in the 15 years since Perelson, Farmer and Packard's seminal paper [22]. AIS and LCS have proven notoriously difficult to analyse. Even if these methods can only serve as a first-order approximation to immune dynamics, they provide a formal foundation based on online learning, signal processing, iterated games and adaptive control systems. It is perhaps a matter of taste, but we propose that this is a more convincing foundation than nearest-neighbour classification in a vectorial shape-space.

Basis decomposition is a powerful methodology with a rich history across mathematics, theoretical and applied sciences. There is little need for us to justify this as a reasonable approach, only to highlight it as a compelling alternative to the approaches currently dominating AIS. Similarly, the burgeoning literature on Boosting, random subspace and ensemble methods is, to some extent, self-validating. While AIS continue to be based on instance-based methods they will necessarily be excluded from the theoretical advances in these fields.

Whether the immunological inspiration contributes back sufficiently to these fields still remains to be seen. It is, at the very least, plausible given that AIS already contribute between the extremes of greedy and global optimisation. This is a future aspect of our work, which relies on additional material not relevant to this paper. At this point, we simply hope the reader finds the presented argument compelling enough to consider thinking outside the shape-space.

References

1. Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensionalspace. *Lecture Notes in Computer Science*, 1973:420–434, 2001.
2. Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54:4311–4322, 2006.
3. R. Bellman. *Adaptive Control Processes: A Guided Tour.* Princeton University Press, 1961.
4. Richard Bellman. *Introduction to Matrix Analysis.* SIAM Classics, 1997.
5. H. Bersini. Immune network and adaptive control. In Paul Bourgine and Francisco Varela, editors, *Proceedings of the first European Conference on Artificial Life*, 1991.
6. H. Bersini. Reinforcement and recruitment learning for adaptive process control. In *Proceedings of the International Fuzzy Association Conference onARtificial Intelligence in Real Time Control*, 1992.
7. H. Bersini. *Artificial Immune Systems and Their Applications*, chapter The Endogenous Double Plasticity of the Immune Network and the Inspirationto be drawn for Engineering Artifacts. Springer-Verlag, 1999.
8. Kevin Beyer, Jonathan Goldstein, and Uri Ramakrishnan, Raghu andShaft. When is "nearest neighbor" meaningful? *Lecture Notes in Computer Science*, 1540:217–235, 1999.
9. Leo Breiman. Prediction games and arcing algorithms. *Neural Comp.*, 11(7):1493–1517, October 1999.
10. Martin V. Butz. Learning classifier systems. In *GECCO '07: Proceedings of the 2007 GECCO conference companion on Genetic and evolutionary computation*, pages 3035–3056, New York, NY, USA, 2007. ACM.
11. Jorge Carneiro, Antonio Coutinho, Jose Faro, and John Stewart. A model of the immune network with b-t cell co-operation. i - prototypicalstructures and dynamics. *Journal of Theoretical Biology*, 182:513–529, 1996.
12. Jorge Carneiro, Antonio Coutinho, and John Stewart. A model of the immune network with b-t cell co-operation. ii - thesimulation of ontogenisis. *Journal of Theoretical Biology*, 182:531–547, 1996.
13. Jorge Carneiro and John Stewart. Rethinking shape space: Evidence from simulated docking suggeststhat steric shape complementarity is not limiting for antibody-antigenrecognition and idiotypic interactions. *J.Theor.Biol*, 169:391–402, 1994.

14. I. R. Cohen. Immune system computation and the immunological homunculus. In O. Niestrasz et al., editor, *MoDELS 2006*, pages 499–512, 2006.
15. Irun R. Cohen. *Tending Adam's Garden: Evolving the Cognitive Immune Self.* Academic Press, 2004.
16. Irun R. Cohen and Lee A. Segel. *Design Principles for the Immune System and Other Distributed AutonomousSystems.* Oxford University Press Inc, 2001.
17. Leandro N. De Castro and Jonathan Timmis. *Artificial Immune Systems: A New Computational Intelligence Approach.* Springer Verlag London, 2002.
18. Vincent Detours, Hugues Bersini, John Stewart, and Francisco Varela. Development of an idiotypic network in shape space. *Journal of Theoretical Biology*, 170(4):401–414, October 1994.
19. R. Douglas and T. Sejnowski. Final workshop report: Future challenges for the science and engineering of learning. Technical report, National Science Foundation, 2007.
20. Jason A. Greenbaum et al. Towards a consensus on datasets and evaluation metrics for developing b-cell epitope prediction tools. *Journal of Molecular Recognition*, 2007.
21. Kai W. Wucherpfennig et al. Polyspecificity of t cell and b cell receptor recognition. *Seminars in Immunology*, 19:216–224, 2007.
22. J. D. Farmer, N. H. Packard, and A. S. Perelson. The immune system, adaptation and machine learning. *Physica*, 22:187–204, 1986.
23. A. Freitas and J. Timmis. Revisiting the foundations of artificial immune systems: A problem-oriented perspective. In *ICARIS 2003 : international conference on artificial immune systems*, 2003.
24. A. A. Freitas and J. Timmis. Revisiting the foundations of artificial immune systems for datamining. *IEEE Transactions on Evolutionary Computation*, 11-4:521–540, 2007.
25. Y. Freund and R. Schapire. Game theory, on-line prediction and boosting. In *9th Annual Conference on Computational Learning Theory*, 1996.
26. Yoav Freund and Robert E. Schapire. A decision theoretic generalisation of on-line learning and an applicationto boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.
27. J. Friedman. Greedy function approximation: A gradient boosting machine. IMS 1999 Reitz Lecture, 1999.
28. J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting, 1998.
29. Jerome H. Friedman. Recent advances in predictive (machine) learning. In *PHYSTAT2003*, 2003.
30. Julie Greensmith and Uwe Aickelin. The deterministic dendritic cell algorithm. In *Proceedings of the seventh internation conference on Artificial ImmuneSystems (ICARIS 2008)*, 2008.
31. Emma Hart and Jonathan Timmis. Application areas of ais: The past, the present and the future. In *ICARIS 2005, LNCS 3627*, 2005.
32. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning.* Springer, 2001.
33. U. Hershberg, S. Solomon, and IR Cohen. What is the basis of the immune system's specificity? In V.Capasso, editor, *Mathematical Modelling & Computing in Biology and Medicine*, pages 377–384, 2003.
34. John Holland. *Adaptation in Natural and Artificial Systems.* MIT Press, 1992.
35. Edith I. R. Cohen. Real and artificial immune systems: computing the state of the body. *Nature Reviews Immunology*, 7:569–74, 2007.

36. Charles A Janeway, Paul Travers, Mark Walport, and Mark Schlomchik. *Immunobiology*. Garland, 2001.
37. Frederick W. Byron Jr. and Robert W. Fuller. *Mathematics of Classical and Quantum Physics*. Dover, 1992.
38. Michael J. Kearns and Umesh V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994.
39. S. Krstulovic and R. Gribonval. Mptk: Matching pursuit made tractable. In *Acoustics, Speech and Signal Processing (ICASSP 2006)*, 2006.
40. K. Leon, J. Carneiro, R. Perez, E. Montero, and A. Lage. Natural and induced tolerance in an immune network model. *Journal of Theoretical Biology*, 193:519–534, 1998.
41. Michael S. Lewicki and Terrence J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, 2000.
42. Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.
43. Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.
44. Sridhar Mahadevan. *Representation Discovery using Harmonic Analysis*. Morgan and Claypool, 2008.
45. Stephane G. Mallat. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing, 1993*, 41:3397–3415, 1993.
46. Humbert R. Maturana and Francisco J. Varela. *Autopoiesis and Cognition: The Realization of the Living*. Kluwer Academic Publishers, 1979.
47. M. Mendao, J. Timmis, P. S. Andrews, and M. Davies. The immune system in pieces: Computational lessons from degeneracyin the immune system. In *Foundations of Computational Intelligence (FOCI 2007)*, 2007.
48. Nikoloas Nanas, Victoria S. Uren, and Anne de Roeck. Nootropia: A user profiling model based on a self-organising termnetwork. In *ICARIS 2004, LNCS 3239*, 2004.
49. N. Oza and S. Russell. Online bagging and boosting. In *Artificial Intelligence and Statistics 2001*, pages 105–112. Morgan Kaufmann, 2001.
50. A. S. Perelson and G. Oster. Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self non-self discrimination. *Journal of Theoretical Biology*, 81:645–670, 1979.
51. Alan S. Perelson and Gerard Weisbuch. Immunology for physicists. *Review of Modern Physics*, 69, 1997.
52. Robert E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.
53. John Shaw-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2006.
54. Marina Skurichina and Robert P. W. Duin. Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis and Applications*, 5:121–135, 2002.
55. Thomas Stibor, Jonathan Timmis, and Claudia Eckert. On the use of hyperspheres in artificial immune systems as antibodyrecognition regions. In *ICARIS 2006*, 2006.
56. F. Varela, A. Coutinho, B. Dupire, and N. M. Vaz. *Theoretical Immunology, vol. II*, chapter Cognitive networks: Immune, neural and otherwise. Addison-Wesley, 1988.
57. Francisco J. Varela and Antonio Coutinho. Second generation immune networks. *Immunology Today*, 12(5):159–166, 1991.
58. Pascal Vincent and Yoshua Bengio. Kernel matching pursuit. *Machine Learning*, 48:169–191, 2001.