Deriving Shape Space Parameters from Immunological Data[1]

Derek J. Smith
Department of Computer Science
University of New Mexico
Albuquerque, NM 87131 USA
dsmith@cs.unm.edu

Stephanie Forrest
Department of Computer Science
University of New Mexico
Albuquerque, NM 87131 USA
forrest@cs.unm.edu

Ron R. Hightower
Department of Computer Science
University of New Mexico
Albuquerque, NM 87131 USA
high@cs.unm.edu

Alan S. Perelson
Theoretical Division
Los Alamos National Laboratory
Los Alamos, NM 87545 USA
asp@t10.lanl.gov

Abstract

We present a method for deriving shape space parameters that are consistent with immunological data, and illustrate the method by deriving shape space parameters for a model of cross-reactive memory. Cross-reactive memory responses occur when the immune system is primed by one strain of a pathogen and challenged with a related, but different, strain. Much of the nature of a cross-reactive response is determined by the quantity and distribution of the memory cells, raised to the primary antigen, that cross-react with the secondary antigen. B cells with above threshold affinity for an antigen lie in a region of shape space that we call a *ball of stimulation*. In a cross-reactive response, the intersection of the balls of stimulation of the primary and secondary antigens contains the cross-reactive B cells and thus determines the degree of cross-reactivity between the antigens. We derive formulas for the volume of intersection of balls of stimulation in different shape spaces and show that the parameters of shape space, such as its dimensionality, have a large impact on the number of B cells in the intersection. The application of our method for deriving shape space parameters indicates that, for Hamming shape spaces, twenty to twenty-five dimensions, a three or four letter alphabet, and balls of stimulation of radius five or six, are choices that match the experimental data. For Euclidean shape spaces, five to eight dimensions and balls of stimulation with radius about twenty percent of the radius of the whole space, match the experimental data.

Keywords: immune system, cross-reactive memory, original antigenic sin, shape space

[1]To appear in The Journal of Theoretical Biology.

1 Introduction

Cross-reactive memory is observed when an individual develops memory to one strain of a pathogen and is challenged with a related strain. Vaccination with cowpox to protect against smallpox is an example of an early use of cross-reactive memory (Jenner, 1798; Ada, 1993). Cross-reactive memory also occurs in the natural immune response to pathogens that mutate. Francis (1953) observed that the immune response to influenza was often a recall of the response to a prior influenza infection. Fazekas de St. Groth & Webster (1966) studied this phenomenon and called it “original antigenic sin”. Their work, and that of others (Deutsch & Bussard, 1972; Gerhard, 1978; Yarchoan & Nelson, 1984), revealed that some memory cells specific for the primary antigen were also cross-reactive with the secondary antigen. Cross-reactive memory is often useful in that memory to a one strain of a pathogen can protect against other strains. It has also been suggested that memory to the primary antigen may be maintained by challenge with cross-reactive antigen (Angelova & Shvartsman, 1982; Matzinger, 1994). However, cross-reactive memory can also be a problem because memory cells highly specific for the secondary antigen are not formed if the antigen is cleared too quickly by memory cells of the primary antigen. The same effect can potentially cause vaccine failure; a vaccine might be cleared by memory cells of a prior infection without inducing memory to the vaccine components.

Most of the experimental work on cross-reactive memory has been performed using two strains of a single organism, or two related haptens, as primary and secondary antigens. Experiments on more than two antigens would be useful in order to better understand, for example, sequential infections with influenza (Angelova & Shvartsman, 1982), pathogenesis of HIV and multivalent vaccine design. We have developed a model and computer simulation of cross-reactive memory in order to perform *in machina* multi-antigen experiments with the goal of helping to understand the immune response to mutating pathogens and sequential multivalent vaccines. An advantage of computer simulations is that all of the data in the simulation are easily measured; a disadvantage is that the simulation may omit or distort important aspects of the system being modeled. We describe some of our efforts to calibrate a model with immunological data to make it closer to the *in vivo* reality.

Much of the character of a cross-reactive response is determined by the quantity and distribution of the population of memory cells, raised to the primary antigen, that also react with the secondary antigen. Figure 1 shows that if we consider the cells that respond to the primary and secondary antigens as sets, then the cells that react with both antigens, the cross-reactive cells, lie in the intersection of the sets. If the antigens are closely related, then there are a large number of cells in the intersection and there will probably be a strong cross-reactive response. If the antigens are less closely related, then the number of cells in the intersection is small, and there will probably be only a weak cross-reactive response. Because the number and distribution of cells in the intersection plays a significant role in the cross-reactive response, it warrants careful study in a model being used to study cross-reactive responses. We would like to know how the intersection varies as a function of the antigenic differences between the primary and secondary antigen. The number of cells in the intersection will depend on how we choose to model antibody-antigen interactions, and what parameters values we choose for this aspect of the model. In this paper we derive parameters, from experimental data, for a model of cross-reactive memory.

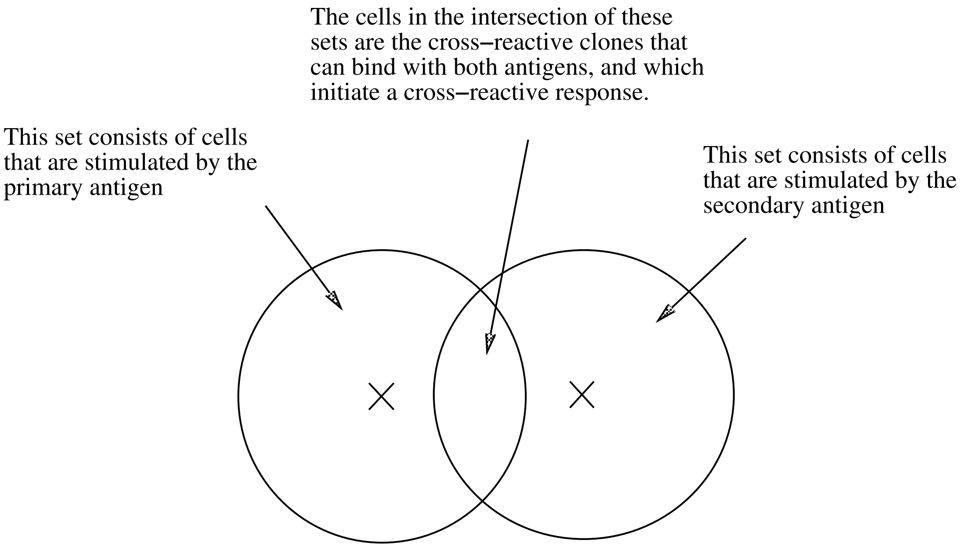

Figure 1: The cells that cause a cross-reactive response are those in the intersection of the set of cells stimulated by the primary antigen and the set of cells stimulated by the secondary antigen.

2 Shape Space Model of Antibody-Antigen Interactions

Antibody-antigen binding affinity[2] is based on complementarity between regions of the antigen and antibody. An abstract model of this was introduced by Perelson & Oster (1979). In this model antibodies and antigens are considered as points in a "shape space" and the distance between an antibody and an antigen is a measure of their affinity for each other. Thus, antibodies within an affinity cutoff for clonal selection by an antigen form a ball in shape space called a ball of stimulation. In a cross-reactive response, each antigen forms such a ball and the intersection of the balls contains the cross-reactive antibodies, thus determining the degree of cross-reactivity between the antigens. Consequently, the Venn diagram representation in Figure 1 can also be interpreted as a shape space diagram.

In order to make shape space more quantitative Perelson & Oster (1979) represented the "generalized shape" of antibodies and antigens with a set of real valued coordinates $<a_1, a_2 \ldots a_n>$. Thus, mathematically, each antibody and antigen could be regarded as a point in an n-dimensional real-valued space. The affinity between an antigen and antibody was related to the distance between them, which was measured as the square root of the sum of the squares of the distances between the values in each dimension. For example if the coordinates of an antibody are $<a_1, a_2 \ldots a_n>$ and the coordinates of an antigen are $<b_1, b_2 \ldots b_n>$, then the distance between them is $\surd\{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2\}$. Shape spaces that use real-valued coordinates, and that measure distance this way, are called *Euclidean* shape spaces (Segel & Perelson, 1988; DeBoer

[2]We refer to the binding of antibodies and antigen, however, this analysis could also be applied to the binding of the T cell receptor with antigen presented on MHC.

et al., 1992).

An alternative to Euclidean shape space is *Hamming* shape space, in which antigens and antibodies are represented as sequences of symbols (Farmer *et al.*, 1986; DeBoer & Perelson, 1991; Seiden & Celada, 1992; Weisbuch & Oprea, 1994; Hightower *et al.*, 1995; Perelson *et al.*, 1996; Detours *et al.*, 1996). Such sequences can be can loosely interpreted as peptides and the different symbols as properties of either the amino acids or of equivalence classes of amino acids of similar charge or hydrophobicity. The mapping between sequence and shape is not fully understood, so for the purposes of this paper we assume that sequence and shape are equivalent. This assumption is reasonable in some situations, for example Champion *et al.* (1975) showed that for azurins, lysozymes, and alpha subunits of tryptophan synthetase, that sequence difference was correlated with the degree of antigenic difference. However, for some antigenic determinants, a single amino acid change can cause a large change in antigenic difference. For such cases a different type of analysis would be needed.

In order to measure the affinity between sequences, we need to define what symbols are complementary so the Hamming distance can be calculated. Any choice is equivalent mathematically, hence for simplicity we choose symbols to be complementary to themselves. For example, let CADBCADB be an antigen and CBDBCDDB an antibody, these are "complementary" in six out of eight places, and thus have a reasonably high affinity for each other. Shape spaces which measure contiguous complementary symbols (Percus *et al.*, 1993), or use other rules for complementarity between symbol sequences (Weisbuch & Oprea, 1994; Detours *et al.*, 1996), have also been used.

A shape space will have different properties depending on the number of dimensions, n, the radius of a ball of stimulation, r, and, in the case of Hamming shape space, on the number of symbols in each dimension, k. As an example of how some properties are sensitive to n, k and r, Figure 2 plots the volume of the intersection of two balls of stimulation, as a function of the sequence difference between the antigens. The formula for the intersection volume is derived in appendix A. Thus, in a model where the volume of the intersection is important, as in a model of cross-reactive memory, shape space parameters must be chosen carefully.

3 The Method and its Application

The method of deriving shape space parameters from immunological data consists of the following steps: (i) determine properties that are important to represent correctly in the model, (ii) estimate data values, from immunological experiments, that characterize these properties, (iii) derive equations for these data values as a function of the parameters of the model, (iv) equate the immunological data values with the model equations, and (v) solve the equations for the model parameters. For a model of cross-reactive memory, an important property to represent correctly is cross-reactivity, and the ideal data values would be the number of B cells in the intersection of the balls of stimulation of antigens of varying sequence differences.

When the sequence difference is zero, the intersection volume is the volume of a ball of stimulation. What has been measured experimentally is the proportion of B cells that respond to an antigen, and from this we can estimate the absolute number of B cells in a ball of stimulation. Estimates for the proportion, P_{exp}, range from 10^{-5} to 10^{-4} (Edelman, 1974; Nossal & Ada, 1971; Jerne,

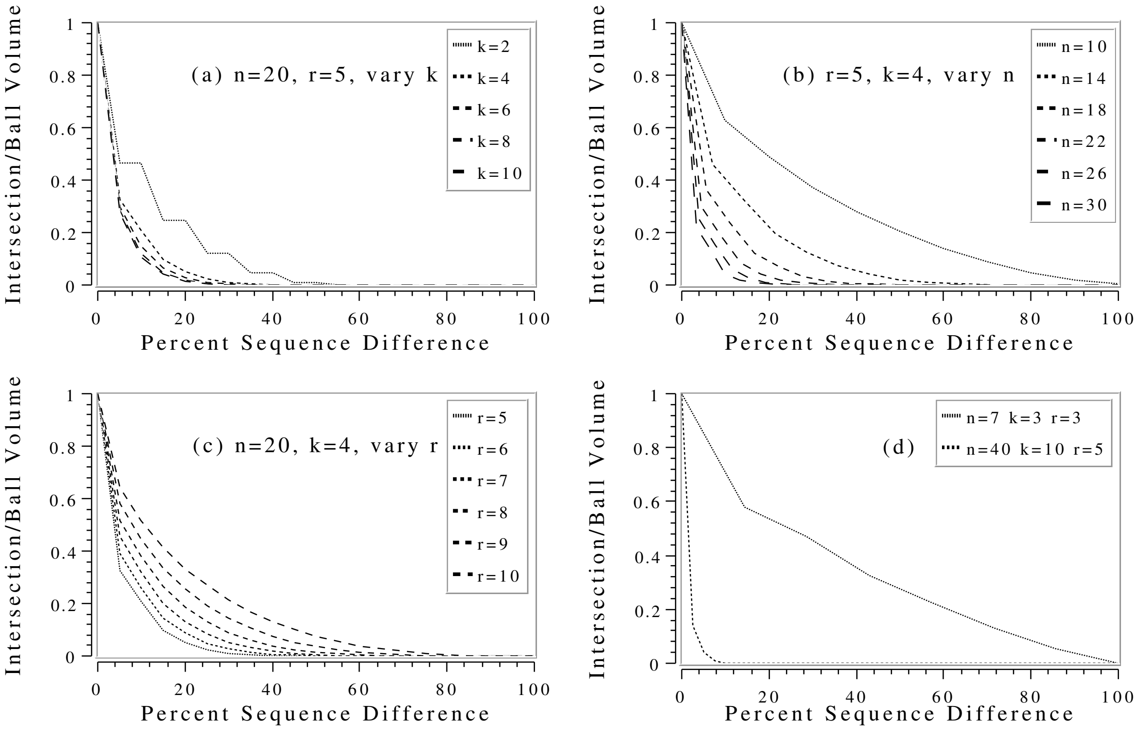

Figure 2: Panels (a), (b) and (c) show that as n or k increase, or as r decreases, the intersection volume, as a function of the sequence difference, falls off more quickly. Panel (a) shows that the binary alphabet, $k = 2$, has an unusual property; every other increase in sequence difference does not cause a decrease in the intersection volume (Kanerva, 1988). Panel (d) shows $n, k,$ and r set at extreme values to illustrate how much the curves can differ.

1974). The equation from the model for the proportion, P, of B cells responding, is volume of a ball of stimulation divided by the volume of the space. Equating the experimental data values and the formula from the Hamming space model we have

$$\frac{\sum_{0 \le i \le r} \binom{n}{i} (k-1)^i}{k^n} = 10^{-5} \text{ to } 10^{-4}. \tag{1}$$

The size, S, of the shape space needs to be sufficiently large to be able to represent all possible antibodies. Based on the number of gene segments used to encode antibodies, the number of possible antibodies, S_{exp} is thought to be at least 10^{10} (Berek & Milstein, 1988; Lodish *et al.*, 1995). Including the effects of somatic hypermutation the number of possible antibodies is many orders of magnitude higher (Lodish *et al.*, 1995); for example it might be as high as 10^{16}. Again equating the experimental data values and the formula from the model we have

$$k^n = 10^{10} \text{ to } 10^{16}. \tag{2}$$

Another data value that can be extracted from experiment is the sequence difference at which the intersection volume of the balls of stimulation goes to zero, i.e. the sequence difference at which two antigens no longer cross-react. We call this distance the "cross-reaction cutoff". It is more intricate than the above equations and is derived in the following subsection.

3.1 Cross-Reaction Cutoff

The experimental data for the cross-reaction cutoff comes from two sources: East *et al.* (1980) and Champion *et al.* (1975). East *et al.* (1980) primed rabbits[3] with beef myoglobin and split them into five groups. Each group received a second injection of myoglobin from one of beef, sheep, pig, whale or chick. The antibody titer to beef myoglobin was plotted against the percent sequence difference between the myoglobins given in the primary and secondary injections. These data are almost ideal, but not quite. The antibody titer was measured at the peak of the secondary response, however we need the number of cells at the beginning of the secondary response. These values are related, but the dynamics of the immune response makes the relation complex. When there are no cross-reactive antibodies, the relation is simple; we can assume there were no cross-reacting cells at the beginning of the response, and thus determine the cross-reaction cutoff. East *et al.* (1980) estimate this point, C_{exp}, to occur in the range of 33 to 42% sequence difference between the primary and secondary antigens.

A second source of the data value for the cross-reaction cutoff comes from Champion *et al.* (1975). In these experiments, seven groups of rabbits were primed with one of seven bacterial azurins. At ten to twelve weeks the rabbits were boosted with the same strain with which they were primed. At twenty to twenty-five weeks the rabbits were boosted again on three successive days, with the same strain, and then bled one week later and the antisera purified. Micro-complement fixation assays were used to determine how well each antisera fixed complement to each of the heterologous azurins. As with East *et al.* (1980), these data are almost ideal, but not quite. The problem is that the distribution of antibodies is not uniform, as it has been biased by affinity maturation during the hyperimmunization. In order to use these data we would need to know the bias due to affinity maturation, and that is not available. We can however again determine the cross-reaction cutoff, which this time is at 40% sequence difference between the antigens.

In order to properly match the model to the experimental cross-reaction cutoff, we need to take into account that memory B cells are more easily stimulated than naive B cells. Fish *et al.* (1989) showed that clonal expansion of memory B cells required a lower affinity antibody-antigen interaction than clonal expansion of naive B cells. In their experiments, A/J mice primed with *p*-azophenylarsonate (Ars) responded predominantly with clones derived from a single V_H gene segment, $V_H Id^{CR}$, and when primed with *p*-azophenylsulfonate (Sulf), no such clones were elicited. However, in mice primed with Ars and challenged with Sulf, clones originally encoded by the $V_H Id^{CR}$ were present. We take this greater sensitivity of memory B cells into account by increasing the radius of the ball of stimulation of the secondary antigen as shown in Figure 3. Thus, when we calculate the volume of the intersection in the model, we must take the radius of the first (naive) ball of stimulation as r, and the radius of second (memory) ball of stimulation as $r \times M_{exp}$.

Two further factors need to be taken into account before relating the cross-reaction cutoff in

[3] Values for our other parameters were taken from experiments in mice, however this parameter is taken from experiments in rabbits.

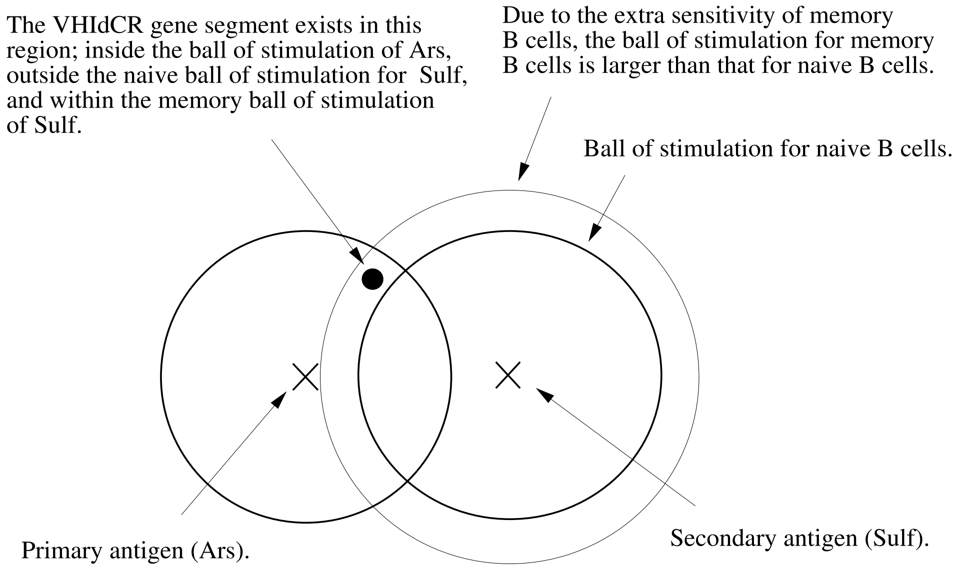

Figure 3: The extra sensitivity of memory B cells results in a ball of stimulation for the secondary antigen that is larger than that of the primary antigen. The amount that the memory ball of stimulation is greater than the naive ball of stimulation is the value M_{exp} as explained in the main text.

experiments with that in the model. First, vertebrate immune systems only express a portion of their total number of possible B cell specificities at any one time. For the mouse, this number, the expressed repertoire, E_{exp}, is in the range 10^7 to $5{\times}10^7$ (Köhler, 1976; Klinman *et al.*, 1976; Klinman *et al.*, 1977). In contrast the formulas in the model give answers in terms of the number of all possible B cell specificities. Second, an experiment might not be able to detect less than a critical number of B cells in the intersection of balls of stimulation, whereas the model can detect a single B cell. We take these factors into account by equating the ratio of the intersection volume and shape space volume in the model, with the ratio of the number of B cells an experiment can detect in the intersection (which we assume to be a single B cell specificity) and the size of the expressed repertoire. Thus we have

$$\frac{I(n, k, r, rM_{exp}, c)}{k^n} = \frac{1}{10^7 \text{ to } 10^8}, \tag{3}$$

where c ranges from 33 to 42% sequence difference, and where the intersection volume, I, is defined in appendix A.

3.2 Solving for Shape Space Parameters

Given values from experiments for M_{exp}, P_{exp}, E_{exp}, S_{exp} and C_{exp}, Equations 1 through 3 can be solved[4] for the Hamming shape space parameters n, k and r. For example, reasonable values to choose from current immunological data are $C_{exp} = 0.33 - 0.42, P_{exp} = 10^{-5}, S_{exp} = 10^{12}, M_{exp} = 1.2$ and $E_{exp} = 10^{7}$, which give shape space parameters $\{n = 20, k = 4, r = 5\}$ or $\{n = 25, k = 3, r = 6\}$. Table 1 shows solutions for different values of the immunological data.

			$S = 10^{10}$	$S = 10^{12}$	$S = 10^{14}$	$S = 10^{16}$
			$n - k - r$	$n - k - r$	$n - k - r$	$n - k - r$
$M = 1.0$	$P = 10^{-5}$	$E = 10^{7}$	6-13-3 7-12-3 8-11-3	5-17-5 6-15-5 8-13-4 9-13-4 10-12-4 11-12-4	4-23-7 5-20-7 6-18-6 7-17-6 8-16-6	4-27-9 5-23-8 6-21-8 8-18-7 9-17-7
		$E = 10^{8}$	3-21-4 4-17-3	3-25-6		3-34-10
	$P = 10^{-4}$	$E = 10^{7}$	2-33-6		2-47-10	2-53-13
		$E = 10^{8}$	2-33-6	2-40-8	2-47-10	
$M = 1.2$	$P = 10^{-5}$	$E = 10^{7}$	3-21-4 4-17-3	3-25-6 4-20-5		
		$E = 10^{8}$			2-47-9	2-53-11
	$P = 10^{-4}$	$E = 10^{7}$	2-33-6	2-40-8	2-47-10	
		$E = 10^{8}$				

Table 1: The values for k, n and r, that satisfy the immunological data, in a Hamming shape space, for different values of M_{exp}, P_{exp}, E_{exp} and S_{exp}. Multiple entries indicated multiple solutions and blank entries indicate no solutions.

4 Using the Same Method on Euclidean Shape Space

The method for deriving shape space parameters can be applied to other shape spaces, other experimental data, and other properties we choose to satisfy. As an example, we now use the method to derive parameters for Euclidean shape space.

After Perelson & Oster (1979), we place a limit on the magnitude of each shape space parameter and normalize distances with respect to this distance, so that distances are in the range zero to one. We use $\hat{r}$ for the normalized radius and $\hat{c}$ for the normalized cross-reaction cutoff.

[4] We obtain solutions by considering k's in the range 2 to 20, and for each k, derive n and r from Equations 2 and 1 respectively, and accept solutions that give C between 0.33 and 0.42. Solutions are necessarily approximate because only integer values make sense in a Hamming shape space.

For Euclidean shape space, equation 1 becomes

$$\hat{r}^n = 10^{-5} \text{ to } 10^{-4}, \tag{4}$$

where the left hand side of this equation is derived at the end of appendix B, and equation 3 becomes

$$\frac{I(n, \hat{r}, \hat{r}M, \hat{c})}{\text{Ball}(n, 1)} = \frac{1}{10^7 \text{ to } 10^8}, \tag{5}$$

where $\hat{c}$ ranges from 0.33 to 0.42, I is now the Euclidean intersection volume and Ball$(n, 1)$ is the volume of the n-dimensional Euclidean shape space normalized to radius 1; both quantities are derived in appendix B.

			$C = 0.33$	$C = 0.36$	$C = 0.42$
			$n/\hat{r}$	$n/\hat{r}$	$n/\hat{r}$
$M = 1.0$	$P = 10^{-5}$	$E = 10^7$	7/0.19	8/0.24	9/0.28
		$E = 10^8$	6/0.15	7/0.19	8/0.24
	$P = 10^{-4}$	$E = 10^7$	5/0.16	5/0.16	6/0.22
		$E = 10^8$	5/0.16	5/0.16	6/0.22
$M = 1.1$	$P = 10^{-5}$	$E = 10^7$	7/0.19	7/0.19	8/0.24
		$E = 10^8$	6/0.15	7/0.19	7/0.19
	$P = 10^{-4}$	$E = 10^7$	5/0.16	5/0.16	6/0.22
		$E = 10^8$	5/0.16	5/0.16	5/0.16
$M = 1.2$	$P = 10^{-5}$	$E = 10^7$	6/0.15	7/0.19	8/0.24
		$E = 10^8$	6/0.15	6/0.15	7/0.19
	$P = 10^{-4}$	$E = 10^7$	5/0.16	5/0.16	5/0.16
		$E = 10^8$	4/0.10	5/0.16	5/0.16

Table 2: Solutions of Equations 4 and 5, for the model parameters n and $\hat{r}$, that satisfy the immunological data in a Euclidean shape space, for different values of $M_{exp}, P_{exp}, E_{exp}$ and C_{exp}.

Solutions of Equations 4 and 5 (Table 2) indicate that n, the number of dimensions, is not very sensitive to the biological values; five to eight dimensions is about right for a Euclidean shape space that satisfies the immunological data for a cross-reactive memory model. In general, the number of dimensions increases as: C_{exp} increases, E_{exp} decreases, P_{exp} decreases, or M_{exp} decreases.

5 Discussion

The intersection volume between balls of stimulation for primary and secondary antigen encounters plays an important role in cross-reactive memory responses. Choices of shape space parameters have a significant effect on the intersection volume predicted by our model. Thus, care must be taken when choosing shape space parameters. We have selected immunological data that are important for a model of cross-reactive memory, and have shown how we can derive shape space parameters from these data. A comparison of our findings for Euclidean and Hamming shape spaces (Figure 4) shows agreement in the intersection volume at zero sequence difference and zero intersection as would be expected as these were the data points for which the equations were solved. However, the intersection volumes differ between these points.

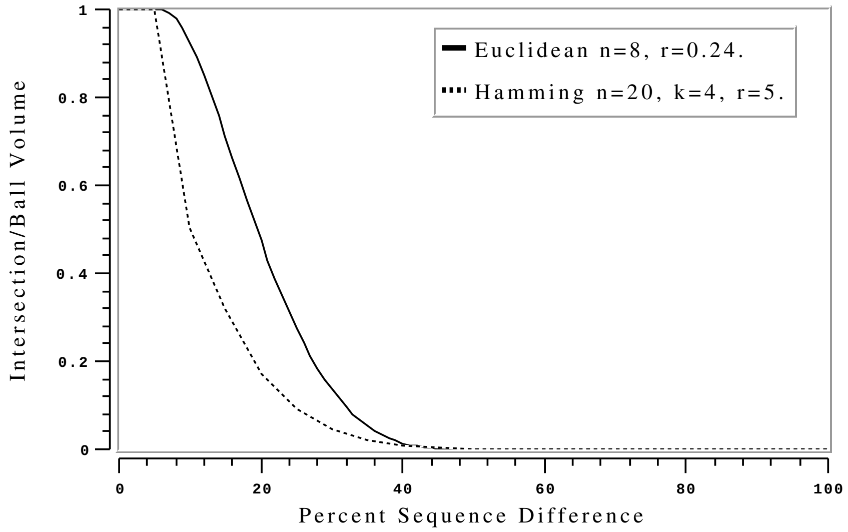

Figure 4: A comparison of the intersection volume, as a function of sequence difference, for a Euclidean and Hamming shape space. For both spaces, the shape space parameters were derived from the immunological data $M_{exp} = 1.2, P_{exp} = 10^{-5}, E_{exp} = 10^{7}, C_{exp} = 0.42$, and, for Hamming space, $S_{exp} = 10^{12}$.

Experiments could be done to test the qualitative relationships in this paper, if we assume that antigenic difference is proportional to sequence difference. Ideal data would give the intersection volume, at various sequence differences, for multiple antigens. As an example, Gerhard (1978) measured the degree of cross-stimulation between various strains of influenza with known sequence differences. Such data could be used to further determine the appropriate choice of shape space parameters.

In prior work, Perelson & Oster (1979) estimated the number of dimensions for a Euclidean shape space to be between five and ten. This agrees well with our calculations which suggest five to eight dimensions (Table 2). Percus *et al.* (1993) used a variation on Hamming shape space in which the complementary symbols had to be contiguous. Using self-nonself discrimination arguments, they predicted about a fifteen symbol binding region for strings made from a three symbol alphabet, and a nineteen symbol binding region when the complementary symbols could be in two contiguous regions. Here we find that a twenty symbol binding region, with a four symbol alphabet and balls of stimulation of radius five, which gave a minimum binding region of fifteen symbols, or a twenty-five symbol binding region, with a three symbol alphabet and balls of stimulation of radius six, which gave a minimum binding region of nineteen symbols, to be consistent with immunological data on cross-reactivity. X-ray crystallographic analysis has shown that a typical antibody-antigen binding site is about seventeen amino acids (Amit *et al.*, 1986), and that there might some gaps in the binding. This matches well with our derivation of a twenty symbol binding region, and a ball of stimulation of radius five.

Segel & Perelson (1988) and DeBoer *et al.* (1992) simulated immunological processes in one and two dimensional Euclidean spaces because it facilitated analysis and comprehension of the dynamics. However, our calculations suggest that a Euclidean shape space between five and eight dimensions is more consistent with the immunological data on cross-reactivity. Binary alphabets are common when Hamming shape spaces are used. Seiden & Celada (1992) used a Hamming space with a binary alphabet and eight to fourteen dimensions. This allowed them to express the complete repertoire which was important for their experiments. Farmer *et al.* (1986) used a binary alphabet and thirty-two dimensions, and Hightower *et al.* (1995) used a binary alphabet and sixty-four dimensions. Binary alphabets with a multiple of thirty-two dimensions are an obvious choice for the efficiency of computer simulations. However, binary alphabets have a stair-step intersection volume, as shown in Figure 2a, and thus might not be a good choice for a model of cross-reactive memory.

It may be tempting to suggest that real antibodies and antigens can be characterized by five to eight Euclidean parameters, or by twenty or so four-symbol Hamming parameters. Either of these statements may or may not be true, but they should not be inferred by the work presented here. What this work shows is how to choose Hamming and Euclidean shape space parameters of a model so that they will match a chosen set of immunological observations.

We have shown, for Hamming shape space, that alphabet sizes of three and four, with the number of dimensions in the mid to low twenties, and balls of stimulation of radius five to six, are good parameters for use in a model of cross-reactive memory. For Euclidean shape space we have shown that, for a wide range of immunological data, five to eight dimensions and balls of stimulation of normalized radius 0.15 to 0.22 are good parameters. We have also shown that, for Hamming shape space, binary alphabets have a stair-step intersection volume, and that large alphabets only satisfy the constraints when we choose extreme values of the immunological data.

Acknowledgments

The authors gratefully acknowledge the ongoing support of the Santa Fe Institute and its Joseph P. and Jeanne M. Sullivan program in theoretical immunology. DJS also acknowledges the University of New Mexico Computer Science Department AI fellowship and Digital Equipment Fellowship, and the help of Terry Jones, Wendell Miller, Ronald Moore, Mihaela Oprea, Francesca Shrady, Paul Stanford and Bernhard Sulzer. Portions of this work were performed under the auspices of the U.S. Department of Energy. This work was also supported by ONR (N00014-95-1-0364), NSF (IRI-9157644), and the Los Alamos National Laboratory LDRD program.

REFERENCES

Ada, G. L. (1993). Vaccines. In: *Fundamental Immunolgy, Third Edition* (Paul, W. E., ed.) chapter 37, pp. 1309–1352. Raven Press, New York.

Amit, A. G., Mariuzza, R. A., Phillips, S. E. V. & Poljak, R. J. (1986). Three-dimensional structure of an antigen-antibody complex at 2.8 angstrom resolution. *Science* **233**, 747–753.

Angelova, L. A. & Shvartsman. (1982). Original antigenic sin to influenza in rats. *Immunology* **46**, 183–188.

Berek, C. & Milstein, C. (1988). The dynamic nature of the antibody repertoire. *Immunol. Rev.* **105**, 5–26.

Beyer, W. H. (1981). *CRC Standard Mathematical Tables, 26th Edition*. CRC Press, Florida.

Champion, A. B., Soderberg, K. L., Wilson, A. C. & Ambler, R. P. (1975). Immunological comparison of azurins of known amino acid sequence: Dependence of cross-reactivity upon sequence resemblance. *J. Mol. Evol.* **5**, 291–305.

DeBoer, R. J. & Perelson, A. S. (1991). Size and connectivity as emergent properties of a developing immune network. *J. Theoret. Biol.* **149**, 381–424.

DeBoer, R. J., Segel, L. A. & Perelson, A. S. (1992). Pattern formation in one and two dimensional shape space models of the immune system. *J. Theoret. Biol.* **155**, 295–333.

Detours, V., Sulzer, B. & Perelson, A. S. (1996). Size and connectivity of the idiotypic network are independent of the discreteness of the affinity distribution. *J. Theoret. Biol.* **183**, 409–416.

Deutsch, S. & Bussard, A. E. (1972). Original antigenic sin at the cellular level. I. Antibodies produced by individual cells against cross-reacting haptens. *Eur. J. Immunol.* **2**, 374–378.

East, I. J., Todd, P. E. & Leach, S. J. (1980). Original antigenic sin: Experiments with a defined antigen. *Mol. Immunol.* **17**, 1539–1544.

Edelman, G. M. (1974). Origins and mechanisms of specificity in clonal selection. In: *Cellular Selection and Regulation in the Immune System* (Edelman, G. M., ed.) pp. 1–38. Raven Press, New York.

Farmer, J. D., Packard, N. H. & Perelson, A. S. (1986). The immune system, adaptation, and machine learning. *Physica D* **22**, 187–204.

Fazekas de St. Groth, S. & Webster, R. G. (1966). Disquisitions of original antigenic sin. I. Evidence in man. *J. Exp. Med.* **124**, 331–345.

Fish, S., Zenowich, E., Fleming, M. & Manser, T. (1989). Molecular analysis of original antigenic sin. I. Clonal selection, somatic mutation, and isotype switching during a memory B cell response. *J. Exp. Med.* **170**, 1191–1209.

Francis, T. (1953). Influenza, the newe acquayantance. *Ann. Intern. Med.* **39**, 203–221.

Gerhard, W. (1978). The analysis of the monoclonal immune response to influenza virus. III. The relationship between stimulation of virus-primed precursor B cells by heterologous viruses and reactivity of secreted antibodies. *J. Immunol.* **120**, 1164–1168.

Hightower, R. R., Forrest, S. & Perelson, A. S. (1995). The evolution of emergent organization in immune system gene libraries. In: *Proceedings of the Sixth International Conference on Genetic Algorithms* (Eshelman, L. J., ed.) pp. 344–350, San Francisco, CA. Morgan Kaufman.

Jenner, E. (1798). *An Inquiry into the Causes and Effects of the Variolae Vaccinae*. Low, London.

Jerne, N. K. (1974). Clonal selection in a lymphocyte network. In: *Cellular Selection and Regulation in the Immune System* (Edelman, G. M., ed.) pp. 39–48. Raven Press, New York.

Kanerva, P. (1988). *Sparse Distributed Memory*. MIT Press, Cambridge, MA.

Klinman, N. R., Press, J. L., Sigal, N. H. & Gerhart, P. J. (1976). The acquisition of the B cell specificity repertoire: the germ-line theory of predetermined permutation of genetic information. In: *The Generation of Antibody Diversity* (Cunningham, A. J., ed.) pp. 127–150. Academic Press, New York.

Klinman, N. R., Sigal, N. H., Metcalf, E. S., Gerhart, P. J. & Pierce, S. K. (1977). *Cold Spring Harbor Symp. Quant. Biol.* **41**, 165.

Köhler, G. (1976). Frequency of precursor cells against the enzyme beta-galactosidase: an estimate of the balb/c strain antibody repertoire. *Eur. J. Immunol.* **6**, 340–347.

Lodish, H., Baltimore, D., Berk, A., Zipursky, S. L., Matsudaira, P. & Darnell, J. (1995). *Molecular Cell Biology*. Scientific American Books, New York.

Matzinger, P. (1994). Immunological memories are made of this? *Nature* **369**, 605–6.
Nossal, C. J. V. & Ada, G. L. (1971). *Antigens, Lymphoid Cells and The Immune Response*. Academic Press, New York.
Percus, J. K., Percus, O. E. & Perelson, A. S. (1993). Predicting the size of the T cell receptor and antibody combining region from consideration of efficient self-nonself discrimination. *Proc. Natl. Acad. Sci. USA* **90**, 1691–1695.
Perelson, A. S. & Oster, G. F. (1979). Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self- non-self discrimination. *J. Theoret. Biol.* **81**, 645–670.
Perelson, A. S., Hightower, R. & Forrest, S. (1996). Evolution and somatic learning in V-region genes. *Res. Immunol.* **147**, 202–208.
Segel, L. A. & Perelson, A. S. (1988). Computations in shape space: A new approach to immune network theory. In: *Theoretical Immunology, Part Two, SFI Studies in the Sciences of Complexity* (Perelson, A. S., ed.) pp. 321–343. Addison-Wesley, Reading, MA.
Seiden, P. E. & Celada, F. (1992). A model for simulating cognate recognition and response in the immune system. *J. Theoret. Biol.* **158**, 329–357.
Weisbuch, G. & Oprea, M. (1994). Capacity of a model immune network. *Bull. Math. Biol.* **56**, 899–921.
Yarchoan, R. & Nelson, D. L. (1984). Specificity of in vitro anti-influenza virus antibody production by human lymphocytes: Analysis of original antigenic sin by limiting dilution cultures. *J. Immunol.* **132**, 928–935.

Appendix

A Intersection Volume in Hamming Shape Space

Consider an n-dimensional Hamming space with alphabet size k. Let I and J be points (strings of n symbols) in the space at Hamming distance s from each other. Let K be a point at distance i from I and j from J, and let $N_{i,j}$ be the number of all such points. Figure 5 shows the three strings I, J, and K, structured in a way to illustrate that the symbols of K can be partitioned into five groups. These strings can, without loss of generality, be manipulated to fit this template, because the space has an automorphism which maps any three points to these templates; the order of presentation of the dimensions, and the choice of symbols for each dimension, do not alter any of the aspects of the space that interest us.

The partitions a, b, c, d, and e of Figure 5 can be described in words as follows:

a. those that are the same as both I and J.

b. those that are different from I and J in a place where I and J are the same.

c. those that are the same as I in a place where J differs from I.

d. those that are the same as J in a place where I differs from J.

e. those that are different from both I and J in a place where I and J differ.

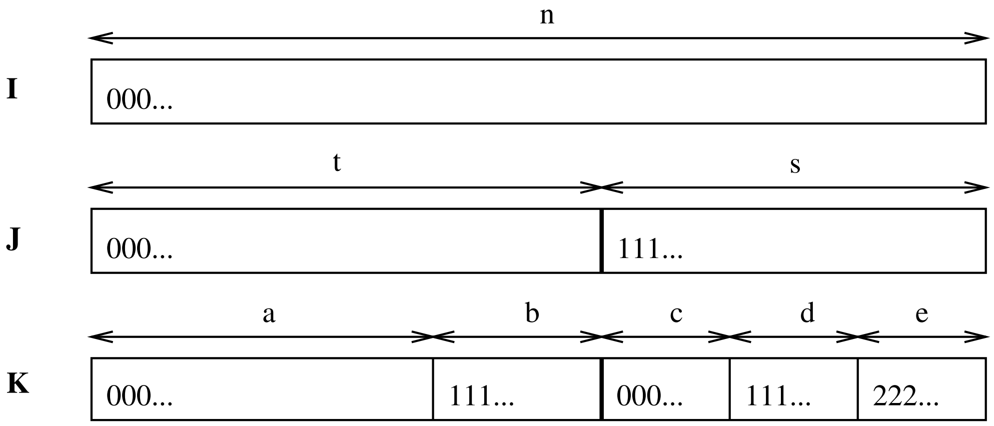

Figure 5: This figure shows the five groupings of the symbols of K, and how these groups relate to the groupings in I and J.

The Hamming distance between two strings is the number of symbols that are different between the strings. The Hamming distance between I and J is s, the Hamming distance between K and I is the sum of b, d and e, and the Hamming distance between K and J is the sum of b, c and e (Figure 5).

Thus for K to be Hamming distance i from I and j from J we must have (Figure 5)

$$i = b + d + e, \tag{6}$$

and

$$j = b + c + e. \tag{7}$$

Similarly the distance, s, between I and J is the sum of the sizes of the partitions c, d and e, thus we have

$$s = c + d + e, \tag{8}$$

and

$$t = a + b. \tag{9}$$

Because of equations (8) and (9) and that the length of the strings is n, we get

$$n = t + s. \tag{10}$$

Let C be the set of 5-tuples, $\{a, b, c, d, e\}$, that satisfy equations (6), (7), (8), (9) and (10). Then,

$$N_{i,j} = \sum_{\{a,b,c,d,e\}\epsilon C} \binom{t}{a, b} 1^a (k-1)^b \binom{s}{c, d, e} 1^c 1^d (k-2)^e,$$

where k is the number of symbols in the alphabet.

If we write the multinomials as binomials, and substitute in a rearrangement of equation (10) for t we get

$$N_{i,j} = \sum_{\{b,d,e\}\epsilon C'} \binom{n-s}{b}(k-1)^b \binom{s}{d}\binom{s-d}{e}(k-2)^e.$$

The five equations, (6) through (10), constrain the values of the five free variables of a, b, c, d and e, to one degree of freedom, i.e. the choice of either a, b, c, d or e, will determine the remaining four values. If we choose d as the free variable, then

by (7) minus (8) we get

$$b = d + j - s, \tag{11}$$

and by (11) into (6) we get

$$e = i + s - 2d - j, \tag{12}$$

which gives

$$N_{i,j} = \sum_{0 \le d \le s} \binom{n-s}{d+j-s}(k-1)^{d+j-s}\binom{s}{d}\binom{s-d}{i+s-2d-j}(k-2)^{i+s-2d-j}.$$

Let r_0 and r_1 be the radii of the Hamming balls[5] about I and J respectively. Then the intersection volume is the sum of $N_{i,j}$ for all $0 \le i \le r_0$ and $0 \le j \le r_1$, thus

$$I(n, k, s, r_0, r_1) = \sum_{\substack{0 \le i \le r_0 \\ 0 \le j \le r_1}} N_{i,j}. \tag{13}$$

B Intersection Volume in Euclidean Shape Space

The volume of the intersection of two Euclidean balls of stimulation can be calculated as the sum of the two shaded segments in Figure 6. Beyer (1981) gives formulas for 2- and 3-dimensional segments; here we derive the volume of n-dimensional segments.

The volume of a segment is the integral, along the line S_1S_0, from the circumference of the ball to the line AB. The integrand is an $n-1$ dimensional ball with radius $\sqrt{r^2 - (r-x)^2}$, where x is the distance from the circumference along S_1S_0. Thus, for a segment of width x',

$$\text{Segment } (n, r, x') = \int_0^{x'} \text{Ball } (n-1, \sqrt{r^2 - (r-x)^2}) \; dx,$$

where

$$\text{Ball}(n, r) = \text{C}(n) r^n$$

[5] A Hamming ball is the set of points that are within a distance r of a particular point. We call r the radius of the Hamming ball. The *ball* and *radius* terminology is by analogy with the usual, Euclidean, notion of balls and radii.

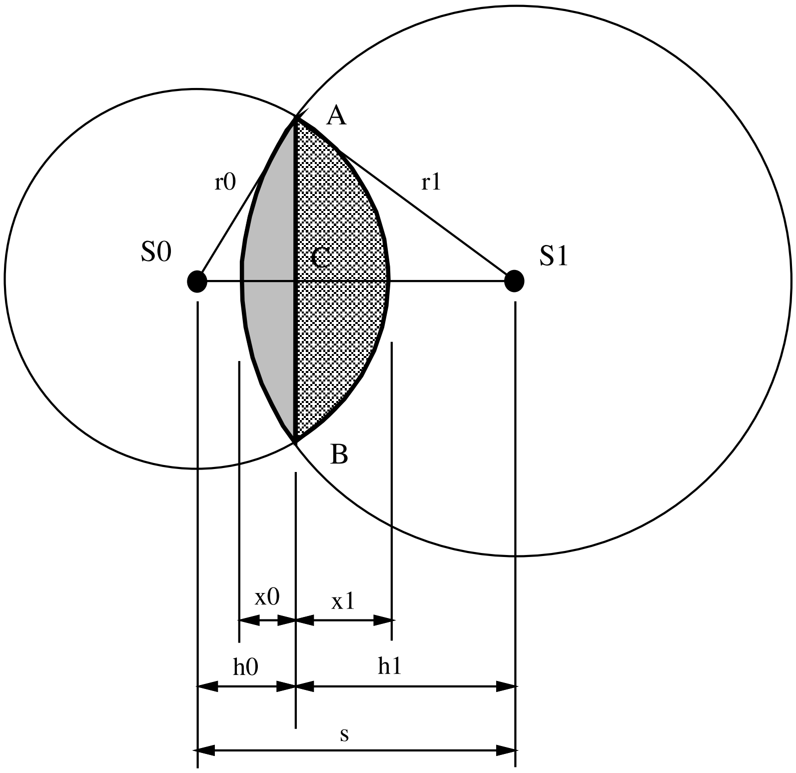

Figure 6: The two shaded segments in this figure, when added together, form the intersection of the balls. The volume of a segment is calculated by integrating along the line S_0S_1. The key to the calculation is that the integrand is a ball in the next lower dimension whose radius is the distance from S_0S_1 to the circumference of the segment's circle (which varies as x moves along S_0S_1 during the integration).

and

$$\mathrm{C}(n) = \frac{2\pi^{(n/2)}}{n\Gamma(n/2)} = \frac{2^n\pi^{((n-1)/2)}((n-1)/2)!}{n!}.$$

Thus

$$\text{Segment}(n, r, x') = \mathrm{C}(n-1)\int_0^{x'} \left(r^2 - (r-x)^2\right)^{(n-1)/2} \quad dx$$

The intersection is the sum of the S_0 segment and the S_1 segment, i. e. ,

$$I(n, r_0, r_1, s) = \text{Segment}(n, r_0, x_0) + \text{Segment}(n, r_1, x_1)$$

where

$$x_0 = \min(2r_0, 2r_1, \max(0, x_0')),$$
$$x_1 = \min(2r_0, 2r_1, \max(0, x_1')),$$
$$x_0' = r_0 - h_0,$$

$$x_1' = r_1 - h_1,$$
$$h_0 = \frac{s^2 - (r_1^2 - r_0^2)}{2s},$$

and

$$h_1 = s - h_0.$$

Figure 7 plots the Euclidean intersection, as a function of sequence difference for 1, 2, 3, 5, 10 and 20 dimensions, with $\hat{r} = 0.5$.

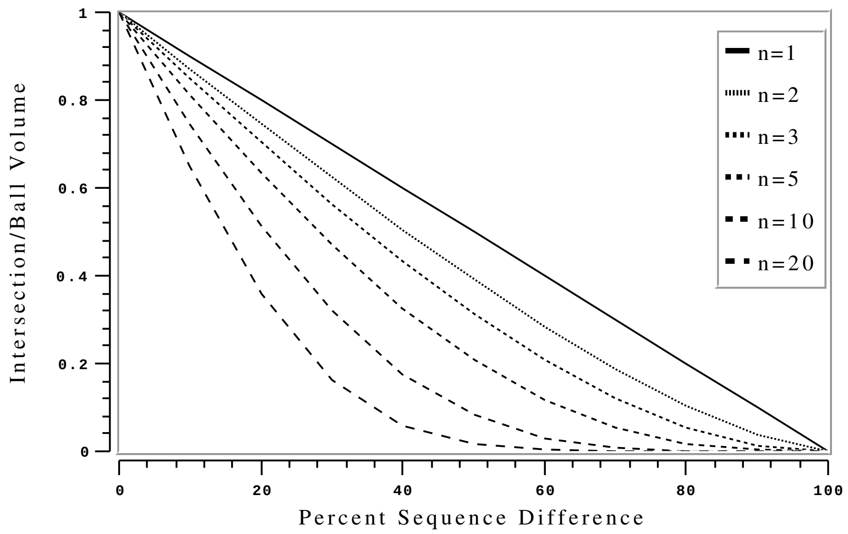

Figure 7: The intersection volume as a function of the sequence difference for Euclidean balls in 1, 2, 3, 5, 10 and 20 dimensions, with $\hat{r} = 0.5$.

The left hand side of equation 4, in the main text, is derived as follows

$$\frac{\text{Ball}(n, \hat{r})}{\text{Ball}(n, 1)} = \frac{\text{C}(n)\hat{r}^n}{\text{C}(n)1^n} = \hat{r}^n.$$