Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 1691–1695, March 1993
Immunology

Predicting the size of the T-cell receptor and antibody combining region from consideration of efficient self–nonself discrimination

(epitopes/self-nonself discrimination)

JEROME K. PERCUS*, ORA E. PERCUS*, AND ALAN S. PERELSON†‡

*Courant Institute of Mathematical Sciences, New York University, New York, NY 10012; and †Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545

Communicated by Stirling A. Colgate, November 18, 1992

ABSTRACT **The binding of antibody to antigen or T-cell receptor to major histocompatibility complex–peptide complex requires that portions of the two structures have complementary shapes that can closely approach each other. The question that we address here is how large should the complementary regions on the two structures be. The interacting regions are by necessity roughly the same size. To estimate the size (number of contact residues) of an optimal receptor combining region, we assume that the immune system over evolutionary time has been presented with a large random set of foreign molecules that occur on common pathogens, which it must recognize, and a smaller random set of self-antigens to which it must fail to respond. Evolutionarily, the receptors and the molecular groups that the immune system recognizes as epitopes are imagined to have coevolved to maximize the probability that this task is performed. The probability of a receptor matching a random antigen is estimated from this condition. Using a simple model for receptor–ligand interaction, we estimate that the optimal size binding region on immunoglobulin or T-cell receptors will contain about 15 contact residues, in agreement with experimental observation.**

The mammalian immune system, to provide protection against pathogenic organisms, must be able to recognize an extremely large number of foreign molecular antigens. The actual number of antigens that the immune system can recognize is unknown but for antibodies it has been estimated to be greater than 10^{16} (1). At the same time, the immune system must fail to be responsive to a presumably smaller number of self-antigens. The binding of receptor to antigen occurs by a generalized lock and key fit of the two structures, involving geometrical and charge complementarity and other factors. The region of an antigen that interacts with a receptor is called an antigenic determinant or epitope. Recognition in the immune system is carried out by receptors on the surface of B and T lymphocytes. The immunoglobulin and T-cell receptor (TCR) repertoires in a mouse are estimated to contain on the order of 10^7 different receptors generated from a much larger potential repertoire of germ-line-encoded receptors (2, 3). With this limited number of receptors, the immune system seems capable of recognizing essentially any antigen. Immunologists thus speak of the repertoire as being complete (4). Perelson and Oster (5) introduced the idea of shape space to show in a quantitative way that a repertoire of 10^7 randomly made receptors is essentially complete. To accomplish this degree of recognition each receptor must be capable of recognizing many different epitopes.

Evolution has shaped the repertoires of immunoglobulin and TCR genes found in vertebrates. Here we suppose that frequently encountered pathogens provide an evolutionary driving force for recognition of epitopes that are common to large classes of microorganisms, for example, polysaccharides found in bacterial cell walls. Evidence that at least some of these antigens are treated specifically by the immune system is found in the fact that polysaccharides stimulate B cells in the absence of T-cell help. At the same time, some epitopes may be expected to be found on self-molecules and on foreign molecules. Although deletion or inactivation of self-reactive clones is known to occur during B-cell and T-cell development, it still seems reasonable that over evolutionary time genes coding for the recognition of common self-antigens would tend to be eliminated and mechanisms would be discovered to minimize interactions with common self-antigens. Thus, we suggest that repertoires are shaped to recognize common features of pathogens that are not simultaneously present on self-molecules. Work by Chalufour *et al.* (6) and Claverie *et al.* (7), in which the sequences of known epitopes were compared with sequences of self-proteins, indicates that epitopes are sequences that are unusually "rare" in known proteins. Similar results have been obtained by Ohno (8). Since only a small fraction of proteins have been sequenced so far, this result is still preliminary but is supportive of our general hypothesis.

If this very general picture is indeed the main story, at least one vestige of the receptor–antigen coevolution should be apparent in present-day organisms. In particular, the size of the region that is complementary in shape between receptors and antigens should be optimal for the recognition and nonrecognition tasks that are to be performed. Size is commonly measured by the number of contact residues and we shall use that convention here. Recent x-ray crystallographic studies indicate that approximately 15 amino acids of an antigenic protein contact the antibody combining site. For example, Amit *et al.* (9) found that in an antigen (lysozyme)–antibody complex, the interface is tightly packed with 16 lysozyme and 17 antibody residues making close contact. Sheriff *et al.* (10) looking at a second epitope on lysozyme found that it is composed of three sequentially separated subsites containing a total of 14 residues in direct contact with the antibody. Surprisingly, Cygler *et al.* (11) found that even with an oligosaccharide antigen there are 15 residues in contact with the antigen. Ajitkumar *et al.* (12) found that T cells recognize a region on peptide–major histocompatibility complex (MHC) complexes of approximately 600 $Å^2$, which is roughly the same contact area as in the antigen–antibody complexes of native proteins such as lysozyme and influenza virus neuraminidase (9, 10, 13). Thus, a common property of receptors is that they seem to interact with regions of about 15 amino acids.

Below we construct a simple model of receptor–antigen matching, which when optimized for maximal recognition of foreign epitopes and minimal recognition of self-epitopes,

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: MHC, major histocompatibility complex; TCR, T-cell receptor.

‡To whom reprint requests should be addressed.

predicts that the optimal sizes of receptor combining regions and epitopes are approximately 15 amino acids. It will turn out that this estimate of the optimal epitope size is insensitive to the precise matching criterion between receptor and antigen that we use and appears to be very robust.

RESULTS

The size of an epitope bears no relationship to the overall size of the antigen, but it should roughly correspond to the size of the receptor combining site. We imagine that epitopes occur at random locations in antigens. We assume that the probability of binding between antigen and receptor depends in part upon epitope size. Data consistent with this assumption have been obtained for the binding of oligosaccharides to antidextran antibodies (14). In our theory, r will denote the number of complementary units (amino acids) on receptors and antigens that need to interact to generate the minimal affinity needed for B-cell or T-cell activation.

Below we calculate $Pr(N,N';n)$, the probability that a receptor repertoire of size n has the property that all of N foreign antigens are recognized by at least one receptor in the repertoire but that none of N' self-antigens are recognized. We will then determine the value of r that maximizes this probability and associate this value of r with the most likely size of epitopes and receptor combining regions.

Let P_S be the probability that a random receptor recognizes a random antigen, so that the corresponding complementary probability, $P_F = 1 - P_S$, is the probability that a random receptor fails to recognize a random antigen. Then

$$\begin{aligned} Pr(N,N';n) &= Pr[\text{each of } N \text{ antigens is recognized by at least one of the } n \text{ receptors}] \\ &\quad \times Pr[\text{none of the } N' \text{ self antigens are recognized by any of the } n \text{ receptors}] \\ &= (1 - P_F^n)^N P_F^{nN'}. \end{aligned} \quad [1]$$

The value of P_F that maximizes $Pr(N,N';n)$ is

$$P_F = \left(1 + \frac{N}{N'}\right)^{-1/n} \approx 1 - \frac{1}{n}\ln\left(1 + \frac{N}{N'}\right). \quad [2]$$

(The approximation can be shown to be extremely accurate for the parameter ranges of interest.)

If the current immune system has optimized $Pr(N,N';n)$, then we would expect the probability of recognizing a random antigen, $P_S = 1 - P_F$, to be computable from Eq. 2. Hence

$$P_S \approx \frac{1}{n}\ln\left(1 + \frac{N}{N'}\right). \quad [3]$$

A mouse contains about 2×10^8 B lymphocytes. If each lymphocyte type grows into a clone of size 20 during its development in the bone marrow, then the animal will contain approximately 10^7 different clones at any one time; i.e., $n = 10^7$ (15). Although the number of foreign and self-epitopes that the immune system deals with is unknown, let us assume $N = 10^{16}$, which is the lower limit proposed by Inman (1). Now if the number of self-epitopes $N' = 10^6$ (i.e., 10 epitopes in each of the 10^5 or so self-proteins coded for in the human genome), then by Eq. 3, $P_S = 2.3 \times 10^{-6}$. This is somewhat smaller than empirical estimates. For example, Cancro *et al.* (16) estimate that in BALB/c mice there are 13.0 ± 2.3 anti-influenza hemagglutinin-specific B cells per 10^6 splenic B cells. Since hemagglutinin is a protein found on the surface of a naturally occurring virus, it is a good choice for comparison with our theory. However, as a protein it may have multiple epitopes, which may explain the 5-fold higher measured frequency than our theory predicts. One might question the assumption of 10 epitopes per self protein or 10^{16} foreign epitopes. However, due to the logarithmic nature of Eq. 3, our estimate of P_S is not very sensitive to changes in N and N'.

As in the classical lock and key picture, a portion of the antigen (or peptide–MHC complex) needs to be roughly complementary in shape to the receptor combining region. Since it is not yet possible to solve folding problems to determine the shapes of receptors and antigens, we resort to a simple model in which antigen and receptor are each modeled as strings or sequences of letters, an approach initiated by Farmer *et al.* (17) and recently used by De Boer and Perelson (18) and Celada and Seiden (19). To determine recognition, we posit that the strings are composed of m letters, each letter having the property that it is complementary to one other letter in the alphabet. For example, if the strings are binary strings, then 1 and 0 are defined as being complementary. To be more realistic, one can think of protein antigens composed of m classes of amino acids, say for $m = 3$, positive, negative, and neutral. To determine complementarity, we assume that due to geometric constraints only l sites on the receptor and antigen sequences can contact each other and demand that a continuous sequence of at least r complementing pairs are required to generate the affinity of interaction needed to trigger B-cell or T-cell activation.

Denote a matching or a complementary pair by the symbol x and noncomplementation by y. If receptor and antigen strings are each constructed with the m units chosen at random, then at each position complementation occurs with probability $1/m$ and noncomplementation occurs with probability $(m - 1)/m$. The probability of recognition P_S then translates into the probability of at least one sequence of at least r contiguous xs out of a total of l entries. This matching problem is not novel and has a history going back at least to de Moivre (see ref. 20).

A general mathematical treatment of this problem, presented in a form that permits generalization to various relaxed matching criteria, will be presented elsewhere. For present purposes, a simple argument suffices. A rigorous analysis shows that the probability of a long matching region is very small, and hence when $m^{-r} << 1$, to a good approximation the various contributing possibilities can be regarded as independent. Starting at the leftmost site of the l-site sequence, r contiguous xs occur with probability m^{-r}. Thereafter, runs of r xs can start at $l - r$ possible sites. Each such run is preceded by a mismatch y, for a net probability of $m^{-r}(m - 1)/m$. We conclude, on adding up these probabilities, that

$$P_S = m^{-r}[(l - r)(m - 1)/m + 1]. \quad [4]$$

By assuming $l >> r > 1$ {i.e., dropping the negligible $[(m - 1)/m]r$ and 1 in Eq. 4}, we obtain

$$r = -\ln_m P_S + \ln_m[l(m - 1)/m]. \quad [5a]$$

Experimental estimates for P_S can be used, or we can substitute the optimal value of P_S given by Eq. 3 to obtain

$$r = \ln_m(nl) - \ln_m\left[\frac{m}{m-1}\ln\left(1 + \frac{N}{N'}\right)\right]. \quad [5b]$$

Because of their logarithmic nature both formulas make similar predictions. In particular, Eq. 5b is very insensitive to the population sizes N and N' of foreign and self-antigens.

Having derived Eqs. 5a and 5b, we now estimate r. The antibody and TCR repertoire sizes are both estimated to be $n \approx 10^7$. The entire variable region of a receptor is not accessible to antigen since some residues are buried in the

interior of the molecule. If we estimate that roughly half of the amino acids are accessible, then $l \approx 100$. Lastly, we need to determine $1/m$, the probability that two amino acids are complementary under the assumption that the different types of amino acids are present in equal quantities. To study the properties of amino acid sequences various "chemical alphabets" have been defined (21, 22). For our purposes a three-letter charge alphabet might be appropriate, where the letters represent positive, negative, and neutral amino acids. Empirical estimates of P_S are on the order of 10^{-5} (16). With $m = 3, l \approx 100, n = 10^7, P_S = 10^{-5}, N \approx 10^{16}$, and $N' \approx 10^6$, Eqs. **5a** and **5b** predict $r \approx 14.3$ and 15.6, respectively. If $l = 70$ rather than 100, then Eq. **5b** predicts $r \approx 15.3$; if $N/N' = 10^9$ rather than 10^{10}, then Eq. **5b** predicts $r \approx 15.7$, illustrating the insensitivity of r to our precise choice of l and the ratio of foreign to self-antigens. These predicted values of r are consistent with the various experimental determinations on the number of contact residues between antibody combining sites and protein antigens and the size of the region on the MHC–peptide complex that interacts with the TCR (9, 10, 12).

If a four-letter functional alphabet were used, with the letters representing acidic, basic, polar uncharged, and hydrophobic nonpolar residues, then Eqs. **5a** and **5b** give $r \approx 11.3$ and 12.5, respectively. Thus the prediction of the theory is sensitive to the choice of m. One way to estimate m is to note that each amino acid in a receptor can on average bind to $20/m$ amino acids on an antigen. Thus for $m = 3$ this implies that on average there would be seven possible complementary amino acids. The recent data on peptides bound to the MHC class I molecule HLA-B27 (23) indicate that, at the nine positions of the peptide, the number of allowed substitutions at each of the positions is 6, 1, 8, 11, 8, 7, 7, 11, and 6 or on average 7.2 amino acids can be at each position of the peptide. Such data, while still preliminary, are consistent with $m = 3$ for peptide–MHC interactions. It is not unlikely that similar values of m would be found for TCR–MHC–peptide or immunoglobulin–antigen binding, since similar physical forces govern the interactions between these molecules. However, the binding of peptide to MHC seems to be associated with the folding and assembly of MHC class I molecules (24) and thus may be somewhat unique. A second way to estimate m is to note that if $r = 15$, then there are 20^{15} possible epitopes. If the repertoire is complete, 10^7 receptors must be able to recognize the 20^{15} epitopes, or each receptor must recognize 3.2×10^{12} epitopes. This number should equal $(20/m)^{15}$, the number of complementary epitopes of length 15. Solving for m one finds $m = 2.9$.

One feature of our results that is surprising is that the dominant effect in determining the optimal epitope size turns out not to be self–nonself discrimination. From Eq. **5b**, keeping only the largest term on the righthand side, $r \approx \ln_m n$, or $n = m^r$. For $m = 3$ and $r = 15$, $m^r \approx 1.4 \times 10^7$, which is approximately the estimated repertoire size. Thus, we conclude that the optimal value of r generates the maximal number of epitopes that an immune system with 10^7 receptors can detect.

If the arguments given above are relevant to the processes involved in the evolution of immunoglobulin and TCR genes, then one might argue that in evaluating r one should use the potential repertoire rather than the expressed repertoire. For immunoglobulins the potential repertoire is estimated to be of order 10^{10} (2). With this value of n, Eq. **5b** predicts that the optimal value of r is 21.9 for $m = 3$ and 17.5 for $m = 4$. Again, quite plausible values. The potential TCR repertoire has been estimated to be 10^{16} (3). One might thus imagine that evolution would have a much harder task in evolving a set of TCR gene segments that would be optimal at recognizing foreign molecules and not recognizing self-molecules. Using $n = 10^{16}$, we find from Eq. **5b** that the optimal values of r are 34.5 and 27.4 for $m = 3$ and 4, respectively. These values are not consistent with observation and thus suggest that the TCR gene families are not optimized for self–nonself recognition and hence other mechanisms, such as thymic deletion, are needed.

Even if the optimal value of r is obtained, as may be the case for immunoglobulins, this does not imply that the immune system has succeeded in doing self–nonself discrimination purely at the level of the selected gene families. At the optimal value of r, the probability of achieving self–nonself discrimination, as given by Eqs. **1** and **2**, is extraordinarily small, $\simeq (N'/eN)^{N'} \simeq 10^{-10^7}$ and, as we well know, self–nonself discrimination has not been obtained at the level of receptors. However, we could still imagine an evolutionary tendency toward self–nonself discrimination by receptors that would lead to an optimal value of r. For example, if discrimination is achieved for some small values of N and N', then the probability of achieving discrimination for an incrementally larger system, say of size $N + \delta N, N' + \delta N'$, is no longer vanishing small and in principle is achievable (25).

DISCUSSION

Immunoglobulins and TCRs have similar-sized combining sites (12). One might then ask if this is coincidence or if there is some underlying principle that has led to the evolution of antigen receptors with a characteristic size binding site. There is a very simple argument that points to the existence of an optimal size for combining regions. The size of the combining region determines the number of amino acid contact residues. If the number of contact residues is too small, then the receptor will lack specificity. Conversely, if the number of contact residues is too large then the receptor will be so specific that the available repertoire will not be sufficient to recognize pathogens efficiently. Here we have developed a theoretical framework with which one can make quantitative predictions of the optimal number of contact residues.

We have formulated the receptor–antigen recognition problem abstractly. We represent receptors and antigen in terms of strings of letters chosen from an alphabet of size m with the property that each letter is complementary to one other letter. In binary sequences, 1 would be complementary to 0. In a three-letter charge alphabet, positive (+) and negative (−) would be complementary, as would be uncharged (0) and uncharged (0). Binary strings, $m = 2$, have found wide use in the last 5 years to represent antibodies and antigens (17–19, 26) in problems involving the determination of complementarity for a number of reasons. (*i*) It is convenient and elegant. String-matching algorithms are fast and can be used to compute a degree of match between molecules that can then be used to represent the affinity of the interaction (27). (*ii*) String matching seems to capture the essence of the process, details of which should be of secondary importance for the matters of issue here.

By using the example of $m = 3$, it is easy to see why an evolutionary strategy in which receptors evolved only to recognize common foreign antigens would not be useful. Given any set of foreign antigens represented by random ternary strings, e.g.,

$$\begin{array}{c} 00+-0++0 \\ +-00-+00 \\ 00+--+00 \\ ++--0+-- \\ --+00000 \\ 00-+00+0 \end{array}$$

one can see that patterns such as 00 or 0+ occur in almost all of the strings. Thus, receptors that recognized any of these

"epitopes" would recognize most antigens in the above list. However, such receptors would not be useful because strings, such as 0+, would also be expected to appear in any population of self-molecules represented in the form of ternary strings. A similar phenomenon is observed when proteins are analyzed for their amino acid content. Ohno (8) observed that two totally unrelated proteins, on the average, share 30 identical tripeptides, two tetrapeptides, and one pentapeptide per 500 residues. Hence receptors that recognized short patterns would be expected to bind molecules throughout the body. Thus, to be useful, receptors must recognize long strings that uniquely characterize the antigen population. Recognizing each antigen by its entire string would require a receptor for each possible antigen and would not be consistent with reasonable repertoire sizes. Thus, antigens need to be recognized over regions that are long but not too long. The regions that are recognized are epitopes and an optimal epitope size should result once the repertoire size and recognition task are specified.

Given repertoire completeness, the most important recognition task that the immune system must perform is self–nonself discrimination. Here we have formulated this task probabilistically and asked what is the optimal epitope size that simultaneously maximizes the probability of recognizing a large set of foreign molecules while minimizing the probability of recognizing a smaller set of self-molecules. Recent work by Claverie *et al.* (7) suggests that this optimization problem may in fact underlie some of self–nonself discrimination. Claverie *et al.* (7) examined the amino acid sequences of known epitopes and compared them with peptides from self-molecules contained in the available protein sequence databanks. They found that epitopes could be characterized in a statistically significant way as sequences that were rare in the currently known population of self-molecules. They then used this as a basis for predicting epitopes of the human immunodeficiency virus. Further, it is known that antibodies can act as antigens and it is this property of antibodies that is the basis of idiotypic network theory. Chalufour *et al.* (6) thus examined the variable regions of a set of antibodies and again found that their sequences were rare in the set of all peptides, consistent with the hypothesis that sequences that are not present in self-molecules are selected as epitopes.

In the string-matching model of receptor–antigen recognition, we assumed that the recognition would be of sufficient strength to signal the B cell or T cell if binding occurred over a sequence of at least r contiguous positions. We then predicted r by using two approaches: (*i*) we used empirical estimates of the probability of an antibody recognizing a random antigen, and (*ii*) we maximized the probability of having a repertoire that recognized all of N foreign antigens and none of N' self-antigens. Interestingly, both approaches predicted that r should be about 15, corresponding to measured sizes of antibody combining sites (9–11, 13).

The closeness of fit of the predicted value of r with that observed, although surprising and well worth taking note of, should not be given excessive credence. Given the approximations in the theory, what is significant is that the predicted value of r is roughly correct, not that it is 15. Clearly, the particular values of N,N' and m that are used are only rough estimates, and by assuming that each letter in a sequence has probability $1/m$ of matching, a letter in the complementary sequence is only an approximation. Also, we have tacitly assumed that the accessible regions of both antigen and receptor can be modeled by sequences that are both of length l, in register. If instead the receptor region were of length l' $< l$, there are clearly $l - l' + 1 + 2(l' - r)$ relative positions, without imposing register, which have an overlap of at least r. Note that if $l' = r$, then Eq. **4** is again recovered. If gaps in complementation are permitted, e.g., due either to discontinuous epitopes or to a relaxed matching criterion where each element need not match, the complementary configurations are again increased in number. For example, one can show (25) that if the combining site instead of being composed of one uninterrupted sequence of at least r matches is an area composed of at least t linear pieces each contributing at least r' matches, then the equivalent of Eq. **5b** is

$$r't = \ln_m(nl^t) - \ln_m\left[\left(\frac{m}{m-1}\right)^t \ln\left(1 + \frac{N}{N'}\right)\right]. \qquad [6]$$

Thus, allowing the $r = r't$ matching sites to be distributed in pieces increases the optimal size of the combining site, but not significantly until l^t approaches the order of n. For example, with two pieces ($t = 2$) and all other parameters as in Eq. **5b**, $r = r't = 19.46$ when $m = 3$, and $r = 15.59$ when $m = 4$. While this effect increases the size of r, any geometric fitting requirements for molecules in three dimensions would decrease this number. One would imagine that these effects are small on the logarithmic scale of Eq. **5**, and hence predicted sizes of combining sites with $r \approx 15$ should be robust.

We thank Rob De Boer, Philip Seiden, Doron Lancet, Lee Segel, and George Bell for helpful discussions. This work was performed under the auspices of the Department of Energy and supported in part by the Applied Mathematical Sciences Program under Contract DE-FG02-88ER85052 (O.E.P.), by National Institutes of Health Grant AI28433 (A.S.P.), and by the Santa Fe Institute through their Theoretical Immunology program.

1. Inman, J. K. (1978) in *Theoretical Immunology*, eds. Bell, G. I., Perelson, A. S. & Pimbley, G. H., Jr. (Dekker, New York), pp. 243–278.
2. Berek, C. & Milstein, C. (1988) *Immunol. Rev.* **105**, 5–26.
3. Davis, M. M. & Bjorkman, P. J. (1988) *Nature (London)* **334**, 395–402.
4. Coutinho, A. (1980) *Ann. Inst. Pasteur Immunol. (Paris)* **131D**, 235–253.
5. Perelson, A. S. & Oster, G. F. (1979) *J. Theoret. Biol.* **81**, 645–670.
6. Chalufour, A., Bougueleret, L., Claverie, J.-M. & Kourilsky, P. (1987) *Ann. Inst. Pasteur/Immunol.* **138**, 671–685.
7. Claverie, J.-M., Kourilsky, P., Langlade-Demoyen, P., Chalufour-Prochnicka, A., Dadaglio, G., Tekaia, F. & Bougueleret, L. (1988) *Eur. J. Immunol.* **18**, 1547–1553.
8. Ohno, S. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 3065–3068.
9. Amit, A. G., Mariuzza, R. A., Phillips, S. E. V. & Poljak, R. J. (1986) *Science* **233**, 747–753.
10. Sheriff, S., Silverton, E. W., Padlan, E. A., Cohen, G. H., Smith-Gill, S. J., Finzel, B. C. & Davies, D. R. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 8075–8079.
11. Cygler, M., Rose, D. R. & Bundle, D. R. (1991) *Science* **253**, 442–445.
12. Ajitkumar, P., Geier, S. S., Kesari, K. V., Borriello, F., Nakagawa, M., Bluestone, J. A., Saper, M. A., Wiley, D. C. & Nathenson, S. G. (1988) *Cell* **54**, 47–56.
13. Colman, P. M., Laver, W. G., Varghese, J. N., Baker, A. T., Tulloch, P. A., Air, G. M. & Webster, R. G. (1987) *Nature (London)* **326**, 358–363.
14. Kabat, E. A. (1976) *Structural Concepts in Immunology and Immunochemistry* (Holt, Rinehart & Winston, New York), 2nd Ed.
15. Klinman, N. R., Press, J. L., Sigal, N. H. & Gearhart, P. J. (1976) in *The Generation of Diversity: A New Look*, ed. Cunningham, A. J. (Academic, London), pp. 127–149.
16. Cancro, M. P., Gerhard, W. & Klinman, N. R. (1978) *J. Exp. Med.* **147**, 776–786.
17. Farmer, J. D., Packard, N. H. & Perelson, A. S. (1986) *Physica D (Amsterdam)* **22**, 187–204.
18. De Boer, R. J. & Perelson, A. S. (1991) *J. Theor. Biol.* **149**, 381–424.
19. Celada, F. & Seiden, P. E. (1992) *Immunol. Today* **13**, 56–62.
20. Uspensky, J. V. (1937) *Introduction to Mathematical Probability* (McGraw-Hill, New York), pp. 77–84.

21. Karlin, S., Ost, F. & Blaisdell, B. E. (1989) in *Mathematical Methods for DNA Sequences*, ed. Waterman, M. S. (CRC, Boca Raton, FL).
22. Karlin, S., Bucher, P., Brendel, V. & Altschul, S. F. (1991) *Annu. Rev. Biophys. Chem.* **20**, 175–203.
23. Jardetzky, T. S., Lane, W. S., Robinson, R. A., Madden, D. R. & Wiley, D. C. (1991) *Nature (London)* **353**, 326–329.
24. Silver, M. L., Parker, K. C. & Wiley, D. C. (1991) *Nature (London)* **350**, 619–622.
25. Percus, J. K., Percus, O. E. & Perelson, A. S. (1992) in *Theoretical and Experimental Insights into Immunology*, eds. Perelson, A. S. & Weisbuch, G. (Springer, New York).
26. Perelson, A. S. (1988) in *Theoretical Immunology, Part Two*, ed. Perelson, A. S. (Addison–Wesley, Reading, MA), pp. 377–401.
27. Perelson, A. S. (1989) *Immunol. Rev.* **110**, 5–36.