Bulletin of Mathematical Biology (2001) **63**, 1063–1078
doi:10.1006/bulm.2001.0259
Available online at http://www.idealibrary.com on IDEAL®

Estimating the Size of the Olfactory Repertoire

LIRAN CARMEL AND DAVID HAREL
Department of Computer Science and Applied Mathematics,
The Weizmann Institute of Science,
Rehovot 76100,
Israel

DORON LANCET
Department of Molecular Genetics,
The Weizmann Institute of Science,
Rehovot 76100,
Israel

The concept of shape space, which has been successfully implemented in immunology, is used here to construct a model for the discrimination power of the olfactory system. Using reasonable assumptions on the behaviour of the biological system, we are able to estimate the number of distinct olfactory receptor types. Our estimated value of around 1000 receptor types is in good agreement with experimental data.

1. INTRODUCTION

Biological systems that are to encounter diverse stimuli do not usually develop a specific receptor for each stimulus type. Rather, they normally adopt a strategy of building a set of nonspecific receptors. The term 'nonspecificity' means that each receptor type responds, with different affinities, to a variety of stimuli. A stimulus introduced into the system elicits a typical, characteristic, response pattern, known as its *spectrum* or its *fingerprint*. The number of different receptor types in such a nonspecific system is referred to as the *repertoire size*. Two typical examples of such systems are the immune system and the olfactory system. The repertoire size of the former is apparently on the order of 10^7 [see, for example, Janeway and Travers (1997), part 2, Chapter 3], while the repertoire size of the latter is assessed to be in the range of 300–1000 [see, for example, Mori and Yoshihara (1998)][†].

In this paper we concentrate on estimating the size r of the olfactory repertoire. The reasons for r being at its current size are not well-understood, and revealing them is an intriguing puzzle. Lancet *et al.* (1993) used a mathematical description of the receptor–ligand interaction and utilized some other measured quantities

[†]More accurately, the repertoire size of most mammals is probably around 1000, among which only about 300 are active receptors in humans.

0092-8240/01/061063 + 16 $35.00/0

to obtain theoretical estimations of r in the range 300–1000. We use the same mathematical description for modelling the interactions, but take a totally different approach to the estimation of r. We use several simple and intuitive assumptions to derive an order of magnitude estimation of r, around 1000, well in agreement with current assessments.

Section 2 presents our underlying model. In Section 3 we derive a formula for the repertoire size of the olfactory system. Section 4 is devoted to a discussion.

2. The Model

Let us assume that odour molecules, hereinafter the ligands, can be described as vectors in some metric space E, and that their fingerprints can be described as vectors in a different metric space, F. By metric space we mean a space on which a distance function is defined, so that there is a meaning to terms such as 'close' ligands, 'distant' ligands, 'close' fingerprints and 'distant' fingerprints. The metric of E is assumed to measure how similar ligands are with respect to their binding affinities to the olfactory receptors. In that sense 'close' ligands elicit similar responses of the sensory system, and 'distant' ligands elicit different response of the sensory system. Whenever we use the terms 'close' and 'distant' with respect to ligands, it should be understood with respect to their binding affinities, and one should keep in mind that 'identical' ligands in E do not have to be structurally identical. The metric of F is assumed to measure how similar ligands are with respect to their perception. We used these abstract spaces to estimate the olfactory repertoire size, by identifying that the interaction between olfactory receptors and odour molecules is just a mapping between these two spaces $f : E \to F$. Simple biological requirements imply basic properties of the mapping f:

(1) *Ligand discrimination*: We require that different ligands will not have the same fingerprint. Using the notation H_i to denote elements in E, we demand that $f(H_1) \neq f(H_2)$ whenever $H_1 \neq H_2$.
(2) *Proximity preservation*: Since the perception is determined from the fingerprints, we require that close ligands are mapped into close fingerprints, or, more formally, that $\nabla_H f$ (the gradient of f with respect to H) is small. In a sense, this requirement determines the robustness of the system. Biological stimuli of the same nature are never expected to produce an identical response of the sensory system, but we still expect their fingerprints to be close.
(3) *Remoteness preservation*: The most tricky observation is that distant ligands (eliciting different response in the sensory system) must have distant fingerprints. For if different ligands yield similar fingerprints, the classification process in the brain may wrongly identify them as being of the same type.

The last two observations are analogous to saying that there must exist some sort of correlation between distances in E and in F. Close objects in E must

be mapped into close objects in F, and distant objects in E must be mapped into distant objects in F.

A realization of such a 'ligand space' E was first introduced by Perelson and Oster (1979), in the context of immunology theory. They argued that both receptors and ligands can be described as N-dimensional vectors in a space E which they dubbed the *shape space*. The underlying logic of such a description is that the strength of the noncovalent binding between two proteins is determined by the properties of the *binding sites* of these proteins, and that these properties can be characterized by a (relatively short) list of N numbers. Each such number, or coordinate, is customarily called a *subsite*, and is identified with properties such as charge, hydrophobicity, and geometric shape. We identify E with the shape space throughout the rest of the paper.

Several implementations of the shape space concept have been successfully used in immunology, especially in the fields of immune networks and cross-reactivity [(Segel and Perelson, 1988; Weisbuch, 1990; De Boer and Perelson, 1991; Lancet *et al.*, 1993; Weisbuch and Opera, 1994; Smith *et al.*, 1997; Detours *et al.*, 1997), see also reviews in Perelson and Weisbuch (1997) or Perelson and Wiegel (1999)].

In this paper we adopt a popular realization of shape space used in certain variations by most of these authors. This realization, that we dub the *discrete shape space*, takes both ligands and receptors to be N-dimensional vectors over a finite alphabet of size S, $E = \{1, 2, \ldots, S\}^N$. The *match*, L, between a receptor $R = (R_1, R_2, \ldots, R_N)$ and a ligand $H = (H_1, H_2, \ldots, H_N)$ is just

$$L = \sum_{i=1}^{N} \delta_{S+1, R_i + H_i}, \tag{1}$$

where $\delta_{x,y}$ is 1 for $x = y$ and 0 otherwise. This law is an implementation of a complementarity principle, which states that each pair of corresponding coordinates of the shape space match if their sum equals $S + 1$. The number $L \in [0, N]$ is a measure of the interaction strength between the receptor and the ligand. Using this shape space, and under the assumptions of the model, Lancet *et al.* (1993) demonstrated that L is proportional to $\ln K$, where K is the *association constant*, or *affinity*, of the receptor–ligand interaction. If r receptors interact with a certain ligand, the fingerprint of this ligand can be represented by the r-tuple $(L_1, L_2, \ldots, L_r)$, and accordingly we define the ligand fingerprint space F as $F = \{0, 1, \ldots, N\}^r$.

The concept of shape space has been used almost exclusively in immunology theory. Lancet *et al.* (1993) claimed that these notions apply also to other nonspecific sets of receptors, including the olfactory system. They assumed a discrete shape space and used immunological experimental data to determine that the best fit values of N and S are 10 and 8, respectively. In the absence of measured values for olfaction, we adopt the assumption of Lancet *et al.* (1993) and take these values to also hold for olfaction.

Equipped with realizations of E and F, we can now turn to define distance functions on both spaces. Such distance functions can be chosen in various ways. Since we are only after an order of magnitude estimation for r, the calculations are insensitive to the details of the distance functions.

The natural choice of distance in the shape space E is

$$d_{H_1H_2} = \sum_{i=1}^{N} (1 - \delta_{(H_1)_i(H_2)_i}), \tag{2}$$

which is the number of different subsites in H_1 and H_2. For example, the distance between the two ligands $H_1 = (1, 2, 2, 3, 1, 1)$ and $H_2 = (1, 2, 3, 3, 1, 2)$ is $d_{H_1H_2} = 2$.

The natural choice of distance in F-space is the *Manhattan distance* [see definition in e.g., Kohonen (1997)],

$$D^r_{H_1H_2} = \sum_{k=1}^{r} |L_{2k} - L_{1k}|, \tag{3}$$

where L_{ik} is the match between receptor R_k and ligand H_i, and r is the total number of receptors.

3. Estimation of the Repertoire Size

In the previous section we employed simple biological observations to put three constraints on the function f. We now use these constraints to get lower bounds on the repertoire size, r. These bounds will be functions of N and S, and for their numerical evaluation we use the proposed values of $N = 10$ and $S = 8$, taken from Lancet *et al.* (1993).

3.1. ***First constraint: ligand discrimination.*** For this requirement to hold, a necessary condition is that the number of possible fingerprints should exceed the number of possible ligands. The overall number of ligands in the E-space is S^N, while the overall number of fingerprints in the F-space is $(1 + N)^r$. Therefore, we must have

$$(1 + N)^r \geq S^N,$$

or

$$r \geq \frac{N \ln S}{\ln(1 + N)}. \tag{4}$$

For $N = 10$ and $S = 8$ this yields

$$r \geq 9. \tag{5}$$

This is indeed a lower bound, but we can do better. Assuming a uniform distribution of the subsites, the probability of two subsites being complementary is just $1/S$. Therefore, the probability function (PF) for getting a match L between an arbitrary receptor and an arbitrary ligand is simply the binomial, namely

$$P(L) = \binom{N}{L}\left(\frac{1}{S}\right)^{L}\left(1-\frac{1}{S}\right)^{N-L}. \tag{6}$$

The most probable match is $L_0 = \lfloor (N+1)/S \rfloor$ (the notion $\lfloor . \rfloor$ stands for the closest integer from below), and the probability of this being the case is denoted by $P_0 = P(L_0)$. For r receptors, the most probable fingerprint is $\bar{L}_0 = (L_0, L_0, \ldots, L_0)$ and its probability is P_0^r. Out of a total of S^N ligands, $P_0^r \cdot S^N$ of them will have the most probable fingerprint, $\bar{L}_0$. We may thus pose a constraint requiring that on the average there will be no more than one ligand with that fingerprint,

$$P_0^r \cdot S^N \leq 1,$$

or

$$r \geq \frac{N \ln S}{\ln(1/P_0)}. \tag{7}$$

This inequality is more restrictive than the previous one. In the case of $N = 10$ and $S = 8$ it yields

$$r \geq 22. \tag{8}$$

This lower bound ensures, in a probabilistic sense, that the fingerprint mechanism is discriminatory, i.e., that no two ligands have the same fingerprint.

3.2. ***Second constraint: proximity preservation.*** Let H_1 and H_2 be two ligands, and let their distance in E be $d_{H_1H_2}$. Then, for a single receptor their distance $D^1_{H_1H_2}$ in F is bounded from above by $d_{H_1H_2}$, since the response of the receptor to identical subsites is the same. Similarly, for r receptors $D^r_{H_1H_2}$ is bounded by

$$D^r_{H_1H_2} \leq r \cdot d_{H_1H_2}.$$

For this reason, the distances in F are bounded proportionately to the distances in E, and the second constraint is satisfied.

3.3. ***Third constraint: remoteness preservation.*** The first constraint indeed yielded a lower bound on r, but this bound is not very informative, being much smaller than the accepted repertoire size. The second constraint was shown to be an inherent property of the mapping f. It is really the third constraint that is the interesting one. For the clarity of the presentation, we first focus on a system with only a single receptor type, $r = 1$, and then show what happens when $r > 1$.

3.3.1. *Single receptor.* Let H_1 and H_2 be two ligands, whose distance in E is $d_{H_1H_2}$. Let $g(D^1_{H_1H_2}|d_{H_1H_2})$ be the probability that the distance between the two ligands in F will be $D^1_{H_1H_2}$. How can we calculate $g(D^1_{H_1H_2}|d_{H_1H_2})$? Assume an arbitrary receptor R, whose match with H_1 is L_1 and whose match with H_2 is L_2. For every subsite i we define a function

$$\Delta h_i = \delta_{S+1,R_i+(H_1)_i} - \delta_{S+1,R_i+(H_2)_i} \tag{9}$$

such that $D^1_{H_1H_2} = |L_2 - L_1| = \left|\sum_{i=1}^{N} \Delta h_i\right|$.

Assume for the moment that H_1 and H_2 differ by a single subsite only, say i_0. Then $D^1_{H_1H_2} = |\Delta h_{i_0}|$, and Δh_{i_0} can take on the values 0 and ± 1 with the following probabilities:

- $(H_1)_{i_0}$ matches R_{i_0} and $(H_2)_{i_0}$ matches R_{i_0}: impossible, $\Delta h_{i_0} = 0$ with probability zero.
- $(H_1)_{i_0}$ matches R_{i_0} and $(H_2)_{i_0}$ does not match R_{i_0}: $\Delta h_{i_0} = +1$ with probability $1/S$.
- $(H_1)_{i_0}$ does not match R_{i_0} and $(H_2)_{i_0}$ does not match R_{i_0}: $\Delta h_{i_0} = 0$ with probability $(S-2)/S$.
- $(H_1)_{i_0}$ does not match R_{i_0} and $(H_2)_{i_0}$ matches R_{i_0}: $\Delta h_{i_0} = -1$ with probability $1/S$.

Now assume that H_1 and H_2 differ by exactly $d_{H_1H_2} = b$ subsites, and let us ignore the identical subsites, treating H_1 and H_2 as pointwise distinct vectors of length b. Denote by n_0 the number of subsites for which $\Delta h_i = 0$, by n_+ the number of subsites for which $\Delta h_i = +1$, and by n_- the number of subsites for which $\Delta h_i = -1$. Obviously $n_0 + n_+ + n_- = b$. The number of possible arrangements of the triple (n_0, n_+, n_-) is the multinomial

$$\binom{b}{n_0, n_+, n_-} \equiv \frac{b!}{n_0! n_+! n_-!},$$

and the probability of obtaining each of them is

$$\frac{(S-2)^{n_0}}{S^b}.$$

Therefore, the total probability of obtaining a specific triple (n_0, n_+, n_-) is

$$\frac{(S-2)^{n_0}}{S^b}\binom{b}{n_0, n_+, n_-}. \tag{10}$$

A value of $D^1_{H_1H_2} = 0$ can be obtained in various ways. It is possible that $n_0 = b$ and $n_+ = n_- = 0$ (i.e., all the b subsites yield $\Delta h_i = 0$); the probability for this is

$$\frac{(S-2)^b}{S^b}\binom{b}{b, 0, 0}.$$

Alternatively, it is possible that $n_0 = b - 2$ and $n_+ = n_- = 1$. The probability for this is

$$\frac{(S-2)^{b-2}}{S^b}\binom{b}{b-2,1,1}.$$

We can continue in such a manner, and the final result is

$$g(D^1_{H_1H_2} = 0 | d_{H_1H_2} = b) = \frac{(S-2)^b}{S^b}\binom{b}{b,0,0} + \frac{(S-2)^{b-2}}{S^b}\binom{b}{b-2,1,1} + \frac{(S-2)^{b-4}}{S^b}\binom{b}{b-4,2,2} + \cdots \quad (11)$$

Similarly

$$g(D^1_{H_1H_2} = 1 | d_{H_1H_2} = b) = 2 \cdot \left[\frac{(S-2)^{b-1}}{S^b}\binom{b}{b-1,1,0} + \frac{(S-2)^{b-3}}{S^b}\binom{b}{b-3,2,1} + \cdots\right],$$

$$g(D^1_{H_1H_2} = 2 | d_{H_1H_2} = b) = 2 \cdot \left[\frac{(S-2)^{b-2}}{S^b}\binom{b}{b-2,2,0} + \frac{(S-2)^{b-4}}{S^b}\binom{b}{b-4,3,1} + \cdots\right], \quad (12)$$

(the multiplication by 2 is due to the symmetry of the exchange $n_+ \leftrightarrow n_-$). Of course, $g(D^1_{H_1H_2} > b | d_{H_1H_2} = b) = 0$. We computed these series using Matlab ®, and the results for $N = 10$ and $S = 8$ are presented in Fig. 1. The crucial observation in Fig. 1 is that there is a large overlap between the different PFs $g(D^1_{H_1H_2} | d_{H_1H_2})$. For example, for any value of $d_{H_1H_2}$ there is a nonnegligible probability of getting $D^1_{H_1H_2} = 0$. This is, of course, a violation of our third constraint—distant ligands ($d_{H_1H_2} = 10$) can have close, or even identical, fingerprints ($D^1_{H_1H_2} = 0$). This result is, of course, expected considering our sole receptor. It is implausible to expect a single receptor to adequately discriminate between any two ligands.

For the third constraint to hold we should obtain much smaller overlaps between the different PFs. In Section 3.3.2 we show that this can be achieved by increasing the number of receptor types, r. But before we do that, we need a way to measure the amount of overlap between the different PFs. Let $f_1(x)$ and $f_2(x)$ be two PFs with means E_1 and E_2, and variances V_1 and V_2. We choose to express their overlap by the parameter ν defined as

$$\nu \equiv \frac{\sigma_1 + \sigma_2}{|E_1 - E_2|}, \quad (13)$$

with $\sigma_1 = \sqrt{V_1}$, $\sigma_2 = \sqrt{V_2}$ the standard deviations of $f_1(x)$ and $f_2(x)$. ν measures the inverse of the difference between the means in units of the average standard

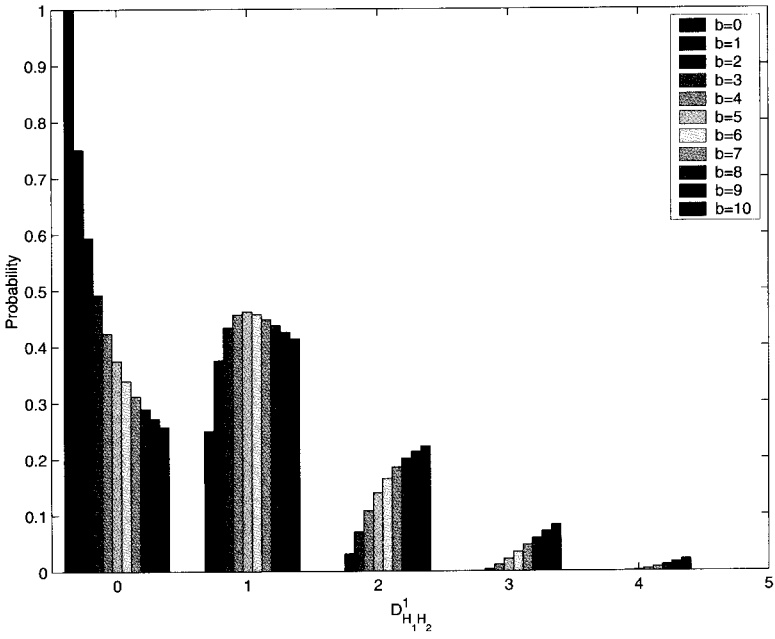

Figure 1. The functions $g(D^1_{H_1H_2}|d_{H_1H_2}=b)$ for the case $N=10$ and $S=8$. The large overlap between the different functions is clearly observed (a colour version of this figure can be found at `http://www.wisdom.weizmann.ac.il/~harel/rep.figs`).

deviation (up to a factor of 2). It is straightforward that the bigger the ν, the larger the overlap. To use (13) in our case, we must find expressions for the means and variances of our functions $g(D^1_{H_1H_2}|d_{H_1H_2})$. The full derivation of these magnitudes is shown in Appendix A, and the final result is as follows [see (A2), (A5) and (A7)]:

$$E^1(b) \equiv E\left(D^1_{H_1H_2}|d_{H_1H_2}=b\right) = \frac{1}{S^b}\sum_{n_0=0}^{b}\binom{b}{n_0}(S-2)^{n_0}R(b-n_0) \tag{14}$$

$$V^1(b) \equiv V\left(D^1_{H_1H_2}|d_{H_1H_2}=b\right) = \frac{2b}{S} - [E^1(b)]^2, \tag{15}$$

where in (14) $R(l)$ is a function that obeys the recursive formula

$$R(l+2) = 4R(l) + \frac{2}{l+1}R(l+1); \qquad R(0)=0; \quad R(1)=2; \quad l \geq 0.$$

Substituting (14) and (15) in (13), the overlap between any pair of functions $g(D^1_{H_1H_2}|d_{H_1H_2}=b)$ and $g(D^1_{H_1H_2}|d_{H_1H_2}=b')$ can be found by,

$$\nu(b,b') = \frac{\sigma^1(b)+\sigma^1(b')}{|E^1(b)-E^1(b')|},$$

with $\sigma^1 = \sqrt{V^1}$. It is natural to identify the ability of the receptors system to distinguish between the cases $d_{H_1H_2} = b$ and $d_{H_1H_2} = b'$ with $\nu^{-1}(b, b')$. We therefore define the *discrimination power* of the system, i.e., its ability to determine whether case $d_{H_1H_2} = b$ has occurred, by

$$\mu^1(b) \equiv \min_{b'} \nu^{-1}(b, b') = \frac{1}{\max_{b'} \nu(b, b')}.$$

The superscript 1 denotes the number of receptors. In our case $\max_{b'} \nu(b, b') = \nu(b, b+1)$, and therefore

$$\mu^1(b) = \frac{E^1(b+1) - E^1(b)}{\sigma^1(b+1) + \sigma^1(b)}. \tag{16}$$

To better understand the power of this definition, let us introduce the notion of *noise*. Let the noise in the olfactory system be represented by an integer q, $0 \leq q \leq N$, such that the two ligands H_1 and H_2 are considered identical if $d_{H_1H_2} \leq q$. In terms of the noise, the system's discrimination power is considered satisfactory if $\mu(q)$ is large enough. But how much is enough? We take the normal distribution as a reference. Let $f_1(x)$ and $f_2(x)$ be two Gaussians with the same σ and with $|\mu_1 - \mu_2| = 2k\sigma$. Then we know that k of 3–4 yields practical separation of the two distributions. Therefore, we choose $k = 3.5$ as a representative value (overlap between the Gaussians of 0.023267%) so that the threshold value of μ is

$$\mu = \frac{2k\sigma}{2\sigma} = k = 3.5.$$

The third constraint is fulfilled only if the μ of the system exceeds the value $k = 3.5$. Just to see how poor the performance of a single receptor is, we calculated $\mu^1(1) = 0.1897$, $\mu^1(2) = 0.1244$, and $\mu^1(3) = 0.0916$.

3.3.2. *Multiple receptors.* If we increase the number of receptors, r, the random variable $D^r_{H_1H_2}$ is given by equation (3) to be a sum of r identical random variables:

$$D^r_{H_1H_2} = D^1_{H_1H_2} + D^1_{H_1H_2} + \cdots + D^1_{H_1H_2} \qquad (r \text{ times}). \tag{17}$$

Therefore, the PF $g(D^r_{H_1H_2}|d_{H_1H_2} = b)$ is the r-times convolution of $g(D^1_{H_1H_2}|d_{H_1H_2} = b)$:

$$g(D^r_{H_1H_2}|d_{H_1H_2} = b) = g(D^1_{H_1H_2}|d_{H_1H_2} = b) \otimes \cdots \otimes g(D^1_{H_1H_2}|d_{H_1H_2} = b) \qquad (r \text{ times})$$

and

$$E^r(b) \equiv E(D^r_{H_1H_2}|d_{H_1H_2} = b) = rE^1(b),$$

$$V^r(b) \equiv V(D^r_{H_1H_2}|d_{H_1H_2} = b) = rV^1(b).$$

Now the discrimination power of the system becomes

$$\mu^r(q) = \sqrt{r} \cdot \mu^1(q). \tag{18}$$

Clearly, for a large enough repertoire r we can make the discrimination power as high as we like. It should be pointed out that since $\mu^1(q)$ is a monotonically decreasing function of q, a system that discriminates for a noise level q will surely do so for any noise level $q' < q$. The smallest r that makes $\mu^r(q) \geq 3.5$ for $q = 1, 2$ and 3 is

$$r = 341 \quad \text{when} \quad q = 1,$$

$$r = 792 \quad \text{when} \quad q = 2,$$

$$r = 1459 \quad \text{when} \quad q = 3.$$

In Figs 2 and 3 we show the PFs of a system with 341 and 792 receptors, respectively. As expected, Fig. 2 shows that the PF for $d_{H_1H_2} = 1$ is well discriminated for a system of 341 receptors, and Fig. 3 shows that the PF for $d_{H_1H_2} = 2$ is well discriminated for a system of 792 receptors.

What is the 'appropriate' value of q? Given a noise level q, there will be

$$\sum_{k=1}^{q} \binom{N}{k} (S-1)^k$$

ligands which are considered identical to some given ligand. Therefore, our formal ligand space is constructed of

$$\frac{S^N}{\sum_{k=1}^{q} \binom{N}{k}(S-1)^k} \tag{19}$$

different ligands. For $N = 10$ and $S = 8$, we get a total of 15 000 000 different ligands when $q = 1$, 470 000 different ligands when $q = 2$, and 25 000 different ligands when $q = 3$. Comparing to common estimations on the real world, the truth is probably somewhere between $q = 2$ and $q = 3$, giving a repertoire size on the order of 1000. When q increases, the ligand space contains fewer distinct ligands, but their separation becomes harder (r should be increased), since there is more noise in the system.

This estimation of q is based on the number of distinct ligands, see equation (19). We should keep in mind however that 'ligands' here are not equivalent to odour chemicals. Rather, they are related to binding sites. Two different chemicals with the same binding site are considered, here, identical ligands.

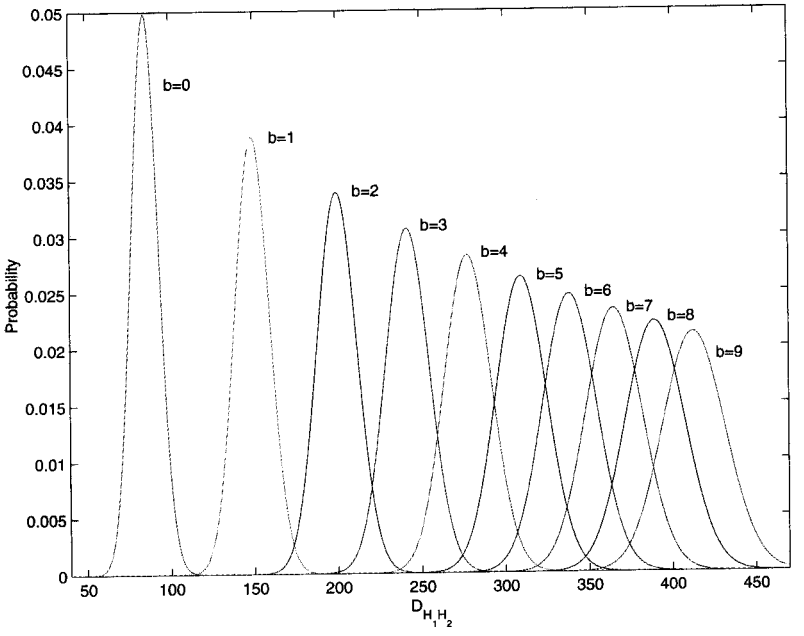

Figure 2. The PFs of a system of 341 receptors for $N = 10$ and $S = 8$ (a colour version of this figure can be found at `http://www.wisdom.weizmann.ac.il/~harel/rep.figs`).

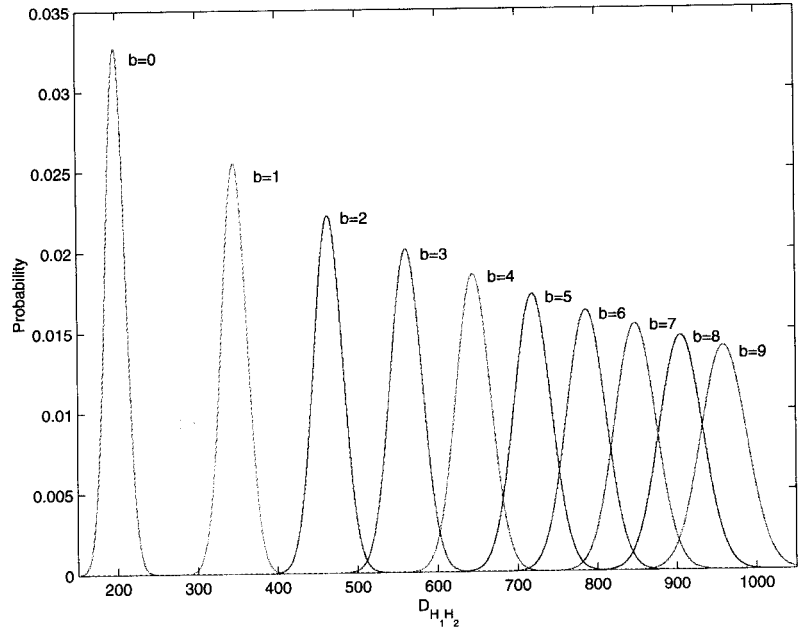

Figure 3. The PFs of a system of 792 receptors for $N = 10$ and $S = 8$ (a colour version of this figure can be found at `http://www.wisdom.weizmann.ac.il/~harel/rep.figs`).

Table 1. Calculation of the relevant noise level and repertoire size for different model parameters.

N	S	Number of ligands in ligand space	Estimated repertoire size
10	9	54 100 ($q = 2$)	834–1498
		1 800 000 ($q = 3$)	
10	7	168 000 ($q = 2$)	758
11	8	145 000 ($q = 3$)	1460
9	8	73 500 ($q = 2$)	792

4. Discussion

The main result of this paper is a derived estimation that the olfactory system should contain about 1000 receptor types. This result is consistent with other estimates of the size of the olfactory repertoire and with experimental observations. We want to emphasize, however, that the exact values we have computed, and from which we derive our estimate (341, 792, and 1459), should not be taken as accurate predictions, but rather as order of magnitude indicators. The exact values depend on assumptions regarding the noise level, the threshold discrimination power k, the parameters N and S, and the discrete shape space model. We do not really expect the final results to lead to the exact size of the repertoire, but we do think that the estimation of 1000 is indeed valid.

Indeed, we claimed that the results are insensitive to the details of the model. For example, picking a different shape space model will undoubtedly shift the numerical results, but we believe that the order of magnitude will remain unchanged. No model will yield a good discrimination power with a single receptor, and since the expression $\mu^r(q) = \sqrt{r} \cdot \mu^1(q)$ is valid for any model, there will always be some r for which $\mu^r(q)$ will cross the threshold value. If the model used is designed to fit experimental data [like the one we used, from Lancet *et al.* (1993)] we expect its global properties to be similar to ours. To test this hypothesis we repeated the calculations with different model parameters. The estimations of the repertoire size indeed remain practically unchanged, as can be seen in Table 1.

We close by discussing an intriguing question: the values that we used for N and S were actually measured in Lancet *et al.* (1993) for general immunological data, so why are our results not valid for estimating the size of the immunological repertoire? Why is the immunological repertoire, which is estimated to be of the order of magnitude 10^7, much higher than would have been concluded from our work? We do not know the answer to this, but we can speculate in several ways. First, the immune system is required for a more complex mission. It must discriminate between self-molecules and external molecules, which may give rise to even stronger restrictions on the mapping f. Second, the mechanism of stimulus recognition in the immune system is much more complicated than in the olfactory

system. It involves several different biological systems that change in time during the response. Our simplistic description of the mapping is probably no longer valid for the immune system. Third, as pointed out by one of the referees of the paper, the immune system takes decisions at the clonal level, and thus the clones must be much more specific than odour receptors [see, for example, Borghans *et al.* (1999)]. Consequently, the immune repertoire has to be much more diverse than that of the olfactory system. Finally, there does not seem to be a wall-to-wall consensus about the value of 10^7 [see, for example, discussions in Cohen (2000), Arstila *et al.* (1999), Keşmir *et al.* (2000)]. Our work may suggest a significantly lower repertoire size for the immune system.

ACKNOWLEDGEMENTS

We would like to thank Lee Segel for many fruitful discussions and stimulating ideas, and to Sol Efroni for suggesting some of the references.

APPENDIX A

Let $E^1(b) = E(D^1_{H_1H_2}|d_{H_1H_2} = b)$ and $V^1(b) = V(D^1_{H_1H_2}|d_{H_1H_2} = b)$ be, respectively, the conditional mean and conditional variance of $D^1_{H_1H_2}$ under the PF $g(D^1_{H_1H_2}|d_{H_1H_2})$. Let us consider only the b different subsites of H_1 and H_2, and ignore the identical subsites. Then, in accordance with definition (9), $E^1(b)$ is

$$E^1(b) = \sum_{\text{all possible } h} \left[\left| \sum_{i=0}^{b} h_i \right| \cdot P(h) \right],$$

with $P(h)$ the probability of obtaining $h = (h_1, h_2, \ldots, h_b)$. As in Section 3.3.1, let n_0 and $n_\pm$ be the number of subsites for which Δh_i equals 0 and ± 1, respectively. By this notation

$$\left| \sum_{i=0}^{b} h_i \right| = |n_+ - n_-| = |b - n_0 - 2n_-|.$$

$P(h)$ is given by expression (10), so that we obtain

$$E^1(b) = \sum_{\substack{n_0, n_-, n_+ \\ n_0 + n_+ + n_- = b}} \binom{b}{n_0, n_+, n_-} \frac{(S-2)^{n_0}}{S^b} |b - n_0 - 2n_-|,$$

or

$$E^1(b) = \frac{1}{S^b} \sum_{n_0=0}^{b} \binom{b}{n_0} (S-2)^{n_0} \sum_{n_-=0}^{b-n_0} \binom{b-n_0}{n_-} |b - n_0 - 2n_-|. \tag{A1}$$

To further simplify this expression, let us define a new function

$$R(l) = \sum_{k=0}^{l} \binom{l}{k} |l - 2k|,$$

so that (A1) becomes

$$E^1(b) = \frac{1}{S^b} \sum_{n_0=0}^{b} \binom{b}{n_0} (S-2)^{n_0} R(b - n_0). \tag{A2}$$

$R(l)$ can be found by examining, separately, the cases where l is even or odd.

- For even l the middle term ($k = l/2$) always vanishes ($l - 2k = 0$), and it suffices to sum only the terms for which $l - 2k$ is positive, and to multiply by 2,

$$\begin{aligned} R(l) &= 2 \sum_{k=0}^{l/2} \binom{l}{k} (l - 2k) = 2l \sum_{k=0}^{l/2} \binom{l}{k} - 4 \sum_{k=0}^{l/2} k \binom{l}{k} \\ &= 2l \left[2^{l-1} + \frac{1}{2} \binom{l}{l/2} \right] - 4l \cdot 2^{l-2} = l \binom{l}{l/2}. \end{aligned} \tag{A3}$$

- For odd l the series is symmetric, and it suffices to sum only up to $k = (l-1)/2$ (keep $l - 2k$ positive), and to multiply by 2,

$$\begin{aligned} R(l) &= 2 \sum_{k=0}^{(l-1)/2} \binom{l}{k} (l - 2k) = 2l \sum_{k=0}^{(l-1)/2} \binom{l}{k} - 4 \sum_{k=0}^{(l-1)/2} k \binom{l}{k} \\ &= 2l \cdot 2^{l-1} - 4 \left[l \cdot 2^{l-2} - l \binom{l-2}{(l-1)/2} \right] = (l+1) \binom{l}{(l-1)/2}. \end{aligned} \tag{A4}$$

These two expression are unified if we define $R(l)$ in terms of a recursion relation. Actually, it can be shown that $R(l)$ satisfies

$$R(l+2) = 4R(l) + \frac{2}{l+1} R(l+1); \quad R(0) = 0; \quad R(1) = 2; \qquad l \geq 0. \tag{A5}$$

The variance is obtained by using the relation

$$V^1(b) = E[(D^1_{H_1H_2})^2 | d_{H_1H_2} = b] - [E^1(b)]^2.$$

The first term on the right-hand side can be found in the same fashion as $E^1(b)$. In analogy with (A1)

$$E[(D^1_{H_1H_2})^2 | d_{H_1H_2} = b] = \frac{1}{S^b} \sum_{n_0=0}^{b} \binom{b}{n_0} (S-2)^{n_0} \sum_{n_-=0}^{b-n_0} \binom{b-n_0}{n_-} (b - n_0 - 2n_-)^2.$$

Now

$$\sum_{k=0}^{l} \binom{l}{k} (l-2k)^2 = l^2 \sum_{k=0}^{l} \binom{l}{k} - 4l \sum_{k=0}^{l} k \binom{l}{k} + 4 \sum_{k=0}^{l} k^2 \binom{l}{k}$$
$$= l^2 \cdot 2^l - 4l \cdot l \cdot 2^{l-1} + 4 \cdot (l^2 + l) \cdot 2^{l-2} = l \cdot 2^l,$$

and therefore

$$E[(D^1_{H_1H_2})^2 | d_{H_1H_2} = b] = \frac{1}{S^b} \sum_{n_0=0}^{b} \binom{b}{n_0} (S-2)^{n_0} (b - n_0) 2^{b-n_0}$$

$$= \frac{b}{S^b} \sum_{n_0=0}^{b} \binom{b}{n_0} (S-2)^{n_0} 2^{b-n_0} - \frac{1}{S^b} \sum_{n_0=0}^{b} n_0 \binom{b}{n_0} (S-2)^{n_0} 2^{b-n_0}. \quad \text{(A6)}$$

The summation in the first term of (A6) is just the binomial expansion of S^b. To find the second term we use the fact that

$$f(x) = (x+a)^b = \sum_{k=0}^{b} \binom{b}{k} x^k a^{b-k}$$

implies

$$\frac{d}{dx} f(x) = b(x+a)^{b-1} = \frac{1}{x} \sum_{k=0}^{b} k \binom{b}{k} x^k a^{b-k}.$$

Therefore, the summation in the second term of (A6) is $b(S-2)S^{b-1}$. All in all, we have

$$E[(D^1_{H_1H_2})^2 | d_{H_1H_2} = b] = b\left(1 - \frac{S-2}{S}\right) = \frac{2b}{S},$$

and the variance is thus

$$V^1(b) = \frac{2b}{S} - [E^1(b)]^2. \quad \text{(A7)}$$

REFERENCES

Arstila, T. P., A. Casrouge, V. Baron, J. Even, J. Kanellopoulos and P. Kourilsky (1999). A direct estimate of the human $\alpha\beta$ T cell receptor diversity. *Science* **286**, 958–961.

Borghans, J. A. M., A. J. Noest and R. J. De Boer (1999). How specific should immunological memory be? *J. Immunol.* **163**, 569–575.

Cohen, I. R. (2000). *Tending Adam's Garden*, Academic Press.

De Boer, R. J. and A. S. Perelson (1991). Size and connectivity as emergent properties of developing immune network. *J. Theor. Biol.* **149**, 381–424.

Detours, V., R. Mehr and A. S. Perelson (1999). A quantitative theory of affinity-driven T cell repertoire selection. *J. Theor. Biol.* **200**, 389–403.

Janeway, C. and P. Travers (1997). *Immunobiology: the Immune System in Health and Disease*, 3rd edn, Current Biology Ltd., Garland.

Keşmir, C., J. A. M. Borghans and R. J. De Boer (2000). Diversity of human $\alpha\beta$ T cell receptors. *Science* **288**, 1135a.

Kohonen, T. (1997). *Self-organizing Maps*, 2nd edn, Springer.

Lancet, D., E. Sadovsky and E. Seidemann (1993). Probability model for molecular recognition in biological receptor repertoires: Significance to the olfactory system. *Proc. Natl. Acad. Sci.* **90**, 3715–3719.

Mori, K. and Y. Yoshihara (1998). Molecular recognition and olfactory processing in the mammlian olfactory system. *Prog. Neurobiol.* **13**, 479–493.

Perelson, A. S. and G. F. Oster (1979). Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self-non-self discrimination. *J. Theor. Biol.* **81**, 645–670.

Perelson, A. S. and G. Weisbuch (1997). Immunology for physicist. *Rev. Mod. Phys.* **69**, 1219–1267.

Perelson, A. S. and F. W. Wiegel (1999). Some design principles for immune system recognition. *Complexity* **4**, 29–37.

Segel, L. A. and A. S. Perelson (1988). Computations in shape space: a new approach to immune network theory, in *Theoretical Immunology Part Two, SFI Studies in the Sciences of Complexity*, A. Perelson (Ed.), Addison-Wesley Publishing Company, pp. 321–343.

Smith, D. J., S. Forrest, R. R. Hightower and A. S. Perelson (1997). Deriving shape space parameters from immunological data. *J. Theor. Biol.* **189**, 141–150.

Weisbuch, G. (1990). A shape space approach to the dynamics of the immune system. *J. Theor. Biol.* **143**, 507–522.

Weisbuch, G. and M. Oprea (1994). Capacity of a model immune network. *Bull. Math. Biol.* **56**, 899–921.

Received 28 February 2001 and accepted 25 July 2001