J. theor. Biol. (2002) **216,** 327–336
doi:10.1006/jtbi.2002.2538, available online at http://www.idealibrary.com on IDEAL®

Test of a Statistical Model for Molecular Recognition in Biological Repertoires

SHAI ROSENWALD†, RAN KAFRI† AND DORON LANCET*†

†*Department of Molecular Genetic and the Crown Human Genome Center, The Weizmann Institute of Science, Rehovot* 76100, *Israel*

(*Received on* 21 *February* 2001, *Accepted in revised form on* 14 *January* 2002)

A chance encounter between members of a random repertoire and a molecular target is characteristic of different biological systems, including the immune and olfactory pathways as well as combinatorial libraries. In such systems, the affinity between the target and members of the repertoire is distributed with a probability function describing the propensity of obtaining a particular affinity value. We have previously proposed a phenomenological receptor affinity distribution (RAD) formalism, which describes this probability function based on simple statistical considerations. In the present analysis, we use published data from diverse experimental systems, including phage display libraries, immunoglobulins and enzymes, to test the RAD model and to compare it to other affinity distribution formalisms. The RAD model is found to provide the best description for binding data for over eight orders of magnitude on the affinity scale, and to account for a relationship between repertoire size and the maximal obtainable affinity within different repertoires. This approach points to a potential universality of the rules that govern affinity distributions in biology.

Introduction

The notion of probabilistic recognition between randomly encountered molecules is central to many biological phenomena. This is particularly evident in biological repertoires, which have evolved to contain enough molecular diversity so as to bind any randomly encountered ligand with a functionally sufficient affinity. The immune repertoires, immunoglobulins and T-cell receptors, provide the most well-known examples for systems displaying probability-based interactions. Other examples include the multi drug resistance (MDR) proteins that underlie the cellular efflux of a large array of compounds (Bolhuis *et al.*, 1997); the olfactory receptor repertoire, which recognizes multitudes of odorants (Buck & Axel, 1991; Lancet, 1986; Lancet & Ben-Arie, 1993); and biotransformation enzymes such as Cytochromes P450, which provide examples of "probabilistic catalysis"—a phenomenon often based on chance interaction between enzyme and substrate (Cupp & Tracy, 1998; Nebert & Gonzalez, 1990). Probabilistic interactions are also at work in catalytic antibodies, whereby members of the immune repertoire are shown to selectively interact with arbitrary transition-state analogs, and consequently to catalyse diverse chemical reactions (Janda *et al.*, 1997; Schultz & Lerner, 1995).

*Author to whom correspondence should be addressed.
E-mail: doron.lancet@weizmann.ac.il

0022-5193/02/$35.00/0

Probability-based recognition is also at the core of the field of combinatorial chemistry. Here, ligand repertoires are used to find new binders for specific molecular targets (Collins, 1997; Hoogenboom, 1997; Lohse & Szostak, 1994, 1996; Plunkett & Ellman, 1997; Scott & Smith, 1990). In an antithesis of rational design, a large pre-prepared random repertoire is screened and often found to contain high-affinity ligands for a pre-selected macromolecular targets.

A corollary of probabilistic recognition is the concept of "affinity distribution", a formal depiction of the statistics that govern the interactions within ligand and receptor repertoires. Such a distribution constitutes a frequency histogram for the affinities obtained when a single target is tested against numerous members within a repertoire. While the existence of such an affinity distribution may hardly be disputed (Burnet, 1963; Inman, 1978; Kauvar *et al.*, 1995; Lancet *et al.*, 1994a; Levitan, 1997, 1998; Macken & Perelson, 1991; Mandecki *et al.*, 1995; Richards, 1975; Vant-Hull *et al.*, 1998), its particular functional shape has been explored only to a limited extent.

An intriguing suggestion is that biological recognition between receptors and ligands obeys a simple, perhaps universal, statistical law (Inman, 1978; Lancet *et al.*, 1993, 1994a). In other words, it is possible that a simple mathematical model could describe the affinity distribution for many different repertoire types, including receptor multi-gene families and combinatorial ligands such as in phage display libraries (Scott & Smith, 1990) or SELEX (Tuerk, 1997).

A central question related to probabilistic recognition is how large and complex should a random repertoire be, so as to ensure, with high probability, that at least one member will manifest a desired affinity value towards a target. This is equivalent to asking how many ligands, on average, should one try randomly before getting a ligand with an affinity value higher than a pre-set threshold. We have previously suggested that this can be answered based on concrete knowledge of the mathematical shape of the entire affinity distribution $\Psi(K)$ for the repertoire in question. It was argued that for any specified value of the desired affinity K^*, the necessary repertoire size would be roughly equal to $1/\Psi(K^*)$ (Lancet *et al.*, 1993).

Several attempts have been aimed at predicting the functional shape of affinity distributions. Among the first predicted affinity distributions was the Sips distribution (Sips, 1948). Due to the existence of an easy test for its appropriateness, as it was often applied to describing the binding of a hapten to heterogeneous polyclonal antibodies, it remained highly attractive for many years. Additional attempts were aimed at deriving numerical procedures for the computation of the desired distribution from experimental binding curves (Bowman, 1963; Erwin, 1976; Bruni *et al.*, 1984) or attempting qualitative predictions, based on intuitive combinatorics (Burnet, 1963).

More recently, an affinity distribution was proposed, based on the assumption of five types of non-covalent interactions (Inman, 1978). An attempt at predicting the probability functions and probability density functions for the high affinity range ($\log K > 5$) was described, based on minimum cross-entropy procedure, resulting in a bimodal distribution (Yee, 1991). In modeling the immune system, Farmer *et al.* assumed only three levels of affinity values with arbitrary probabilities (Farmer *et al.*, 1986) and later an exponential decrease of probabilities, to obtain an affinity value above a certain threshold, was imposed (Detours *et al.*, 1996). Several other models assumed a continuous log-normal distribution (Goldstein, 1975; Macken & Perelson, 1991). Only two of these models have been practically tested against experimental data using a narrow range of affinities (Inman & Barnett 1988; Yee, 1991).

This paper is concerned with a receptor affinity distribution (RAD) model, based on simplified assumptions on the statistics of non-covalent molecular complementarity (Lancet *et al.*, 1994a, 1993). This model, based on a binomial distribution, was tested using ligand titrations for a heterogeneous immunoglobulin mixture. The RAD formalism was subsequently used in modeling the affinity values involved in phage display procedure (Levitan, 1998) as well as in the application of a theoretical distribution for analysing the procedure of

affinity fingerprinting (Kauvar *et al.*, 1995). A similar statistical approach has been independently pursued in the analysis of idiotypic networks in the immune system (Detours *et al.*, 1996; Farmer *et al.*, 1986). An alternative approach, based on extreme value theory, was used to fit the high-affinity tail of a distribution stemming from a low molecular weight ligand screen (Young *et al.*, 1997).

In the standard experimental approach of molecular searches within random ligand repertoires, attention is usually directed to relatively few binders that show the highest affinity (Aujame *et al.*, 1997; Burton, 1995; Griffiths & Duncan, 1998; Hoogenboom, 1997; Scott & Smith, 1990). Data on such high-affinity ligands are, however, not sufficient to gain comprehensive insight on questions related to a complete affinity distribution, which requires knowledge of the entire affinity value range.

We describe here, an analysis of data from the few past studies that actually provided data on a wide range of affinities for large groups of randomly selected ligands towards targets such as immunoglobulins (Inman & Barnett 1988; Varga *et al.*, 1991) and enzymes (Kauvar *et al.*, 1995). In addition, we include in the analysis maximal affinity data published for combinatorial libraries. The analyses provide strong support for the statistical approach embodied in the RAD model, and lend credence to the notion that this model may serve as a universal tool for analysing molecular recognition in biological repertoires.

Statistical Model

The RAD model (Lancet *et al.*, 1993, 1994a, b) is a general statistical description for ligand and receptor repertoires. The RAD model assumes that each molecule, when binding to a target, forms L "formal interactions", each generating an equal free energy contribution (α kcal mol^{-1}) to the overall binding free energy. The variable L is assumed to obey a simple statistical law, which underlies the affinity distribution. In one of the embodiments of the RAD model, ligands and receptors are represented by random strings (with length B) over an alphabet of size S, and L is computed based on a string complementarity rule. The assessment of binding potential through string representations of molecular surfaces has been employed in various previous reports (Farmer *et al.*, 1986; Lancet *et al.*, 1993, 1994a). One particularly prominent series of studies used string complementarity to model antigen–antibody binding in the immune system (Detours *et al.*, 1999, 2000; Detours & Perelson, 1999, 2000; Perelson & Weisbuch, 1997; Smith *et al.*, 1999). In string representation models, complementarity between two surfaces is defined by any number of rules and the degree of complementarity can be quantified and used as a measure of affinity between the two molecules. Employing string complementarity on RAD formalism (Lancet *et al.*, 1993) results in a binomial distribution for L:

$$\Psi_{(L)} = \frac{B!}{L!(B-L)!}\, p^L (1-p)^{(B-L)}, \qquad (1)$$

where B is the maximal possible number of formal interactions (string length), and p is equal to $1/S$.

The total binding free energy is given by $\Delta G = -\alpha L$, and therefore the affinity, i.e. the thermodynamic equilibrium association constant K, is given by

$$\mathrm{Log}\, K = \alpha L / 2.3RT, \qquad (2)$$

where R is the gas constant and T is the absolute temperature. This formalism assumes additivity in free energy of binding, an assumption which is well supported by experimental results (Horovitz, 1996; Horovitz & Rigbi, 1985). The RAD formalism results in the prediction of a specific functional shape for the entire affinity distribution (Fig. 1).

Such string complementarity formalism combined with the principle of additivity in free energy of binding may be shown to be readily applicable to simple molecular recognition systems such as the complementarity between DNA strands. In this case, where a string complementarity rule is rather natural, B is the length of an oligonucleotide and $p = \frac{1}{4}$ (derived from $S = 4$) is the probability of Watson–Crick base pairing. The binomial RAD model has, however, been suggested to

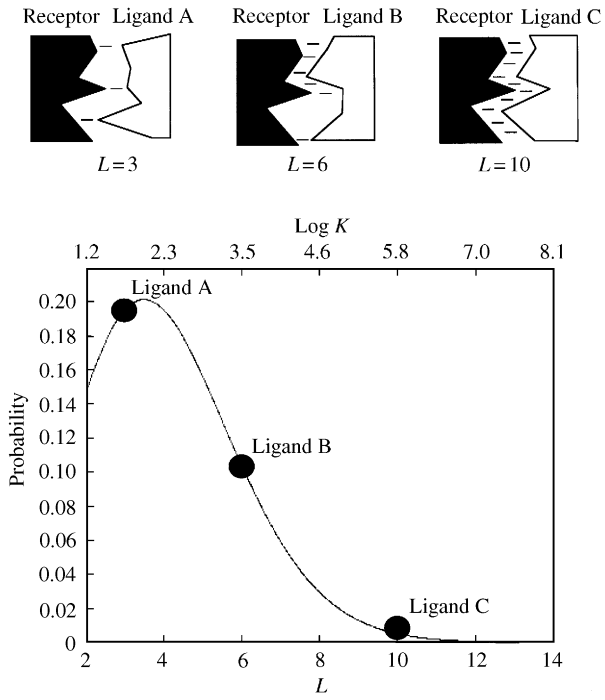

FIG. 1. A schematic representation of the notion of affinity distribution. The probability function $\Psi(L)$ for a given number of formal interactions L is plotted vs. L. The value of L increases with the degree of ligand–receptor complementarity, as shown on top. The curve shown is a binomial distribution with $p = 0.1$ and $B = 40$, equivalent to the Poisson distribution with $\lambda = 4$ (see text). The energy of interaction for each of the ligands is the number of such formal interactions L times an energy parameter α, yielding the affinity scale shown on top (with $\alpha = 0.8$). It is seen that above a certain threshold value of L ($L = 3.5$), the probability function $\Psi(L)$ decreases with increasing L.

be also applicable to more general cases of ligand–receptor interaction where string complementarity serves as a useful phenomenological framework (Kauvar *et al.*, 1995; Lancet *et al.*, 1993; Levitan, 1998).

In the present analysis, we explore the Poisson approximation of the binomial distribution. This is because it may be demonstrated that the binomial RAD model is overdetermined, whereby over a certain range, different B and p pairs fit the data equally well, as long as their product is unchanged (Lancet *et al.*, 1993). A single parameter Poisson-based model is also advantageous, because it is "open ended", as it does not set a maximum for the number of formal interactions, thus allowing to conveniently model high affinity values. For the Poisson distribution, the probability for L formal interactions is computed as

$$\Psi(L) = \frac{\lambda^L}{L!} e^{-\lambda}, \tag{3}$$

where λ is equal to the mean of L. Equations (2) and (3) were used to derive the probability function $\Psi(K)$ as

$$\Psi(K) = \frac{\lambda^{RT/\alpha \ln K}}{((RT/\alpha) \ln K)} e^{-\lambda}. \tag{4}$$

Since K is a continuous variable while L is a discrete variable, values of $RT \ln K/\alpha$ are sorted into bins according to discrete values of L. A numerical fit to experimental affinity data using the probability function $\Psi(K)$ allows a description in terms of the model's two free parameters, λ and α.

Methods

DATA ANALYSES

Three previously published data sets were analysed and studied. The first data set (Varga *et al.*, 1991) was obtained by measuring the degree of inhibition exhibited by 1949 different water-soluble ligands at concentrations of 10, 1 mM and 1 mg ml^{-1} on the binding of water-soluble mouse monoclonal anti-dinitrophenyl, IgE (aDNP), to a stationary phase polysterene-bound hapten, DNP-Gly. The percent inhibition values were calculated from the uptake of radiolabelled IgE with and without the inhibitory ligands. For compounds with high inhibition, the molar concentration necessary for 50% inhibition was determined as described (Varga *et al.*, 1991).

In the second data set (Inman & Barnett, 1988), 85 different small ligands were tested against an IgG1 monoclonal antibody specific for 2,4-dinitrophenyl hapten. For the evaluation of binding constants, the author employed a quantitative affinity chromatography procedure (Inman, 1983), whereby the retention values of the mobile phase antibody was measured with and without addition of mobile phase ligands, and used to calculate the 85 association constants.

The third data set (Kauvar *et al.*, 1995) consisted of a matrix of binding potencies of eight enzymes to 122 random diverse ligands. The

ligand-binding potencies were quantified through the concentrations required to inhibit 50% of the enzyme activity (IC_{50}), as measured by the reduction of the formation rate of the product.

The percent inhibition values I from the radioimmunoassay data (Varga *et al.*, 1991) were translated by us into affinity values K, using the equation

$$\frac{(100 - I)KH}{100(1 + KH)} = \frac{KH}{1 + KH + K_A A}, \qquad (5)$$

where H is the free molar concentration of the ligand, approximated by the total ligand concentration, since the ligands (~ 0.9 μM, J.M. Varga, Pers. Comm.) are at 10 000-fold excess over the antibody [1.8×10^{-10} M, computed from the published data (Varga *et al.*, 1991)]. A is the free molar concentration of the inhibited ligand, and $K_A = 6.5 \times 10^7$ M^{-1} is its affinity towards the antibody target (Eshhar *et al.*, 1980).

The ligand molar concentrations were estimated from the published weight concentration (1 g/l) using the average $M_w = 293 \pm 170$ g mol^{-1} for the 87 compounds for which a molecular mass was available. This was justified by our analysis (not shown) indicating no detectable correlation between M_w and percent inhibition (correlation coefficient 0.023) and between M_w and the affinity constant K (correlation coefficient 0.087).

NUMERICAL PROCEDURES

The ligand affinity data were transformed numerically to cumulative distribution functions, to eliminate a dependence on affinity bin size. The residual minimization for curve fitting was done using Gauss–Newton method implemented in Matlab 4.2c (Matlab, 1994). An interpolation procedure was employed to transform the true independent variable L to the derived variable Log K, based on eqn (2).

Results and Discussion

We have analysed three published data sets containing screening results of large ligand repertoires against single protein targets, in which the affinity data for all ligands have been reported. One additional point represents a maximal affinity value from a limited screen of a random collection of small ligands (Inman & Barnett, 1988).

The data were plotted as distribution functions vs. affinity (Fig. 2. See also Table 1). A monotonically decreasing function typical of cumulative distribution functions was obtained, covering a range of relatively low affinity values (Log K between 2 and 7). Two of the data sets were individually examined for their correspondence to three different models, the Poisson-based RAD model, Inman's multispecificity model and a log-normal model. This resulted in adequate numerical fits for both the log-normal model and the RAD model, as well as the multispecificity model with somewhat lower RMS (Fig. 2). The fitted RAD model predicted a value of λ (the mean number of formal interactions) between 9 and 12 (Fig. 2). The best-fit energy parameter α obtained for the Poisson RAD was found to be within a rather narrow range of 0.2 to 0.4 kcal mol^{-1}. The data for the third system, which involved a much larger error range, were not subjected to a parameter fit procedure, but appeared to be in good agreement with one of the other two datasets.

The data and the computed curves were then replotted double logarithmically, and extended to a higher affinity range [Fig. 2(b), (d), and (f)]. In order to examine the validity of the model in this realm, one would ideally need to examine the entire affinity distributions within large combinatorial repertoires with 10^7–10^9 ligands, that might manifest affinities up to the nanomolar range (Collins, 1997; Hoogenboom, 1997; Lorsch & Szostak, 1994; Plunkett & Ellman, 1997). However, most of the screening results of such libraries published so far, report only one or very few maximal affinity ligands. The use of such data is still possible, based on an extreme value formalism, with two approximations: (a) assuming that the probability of the highest affinity ligand in a repertoire is equal to the inverse repertoire size (Lancet *et al.*, 1994a); and (b) assuming that in the high affinity range, where the distribution curve descends relatively steeply, cumulative probabilities may be adequately approximated by the standard probability values provided in the library data.

TABLE 1
*The values obtained by the numerical fits of the models to the different datasets as shown in Fig. 2**

	a	v	RMS
(*Multispecificity*)			
Varga	8.80E − 02	1.30E + 02	3.50E − 02
Inman	2.00E − 01	4.40E + 01	4.70E − 03
Varga + high	6.50E − 02	1.60E + 02	9.30E − 01
Inman + high	3.80E − 02	1.60E + 02	1.60E + 00
	μ	σ	RMS
(*Lognormal*)			
Varga	3.40E + 00	1.20E + 00	2.40E − 03
Inman	2.60E + 00	7.50E − 01	1.30E − 03
Varga + high	3.40E + 00	5.80E − 01	6.30E − 01
Inman + high	1.80E + 00	1.10E + 00	4.20E − 01
	λ	α	RMS
(*Poisson*)			
Varga	1.10E + 01	4.20E − 01	1.50E − 03
Inman	1.00E + 01	3.60E − 01	
Varga + high	3.70E + 01	1.40E − 01	1.30E + 00
Inman + high	9.20E + 00	3.40E − 01	1.20E − 01

*For each model the table lists the values obtained for its two free parameters and the corresponding value of RMS (root mean square of residuals). The free parameters used for the multispecificity model are the number of van der Waals interactions, v, and their corresponding free energy contribution, a. The free parameters for log-normal model are the mean, μ, and S.D., σ. The free parameters underlying the Poisson RAD are λ and α accounting for the number of formal interactions and mean interaction energy, respectively.

A set of such published maximal affinity values from combinatorial ligand screens (Barrett *et al.*, 1992; Clackson *et al.*, 1991; Cwirla *et al.*, 1990; Griffiths *et al.*, 1993; Hoogenboom, 1997; Inman & Barnett, 1988; Kramer *et al.*, 1995; Martin *et al.*, 1996; Osbourn *et al.*, 1996; Parsons *et al.*, 1996; Schier *et al.*, 1995; Valadon *et al.*, 1996) has been plotted (Fig. 2). These data constitute phage display measurements, mainly from peptide and antibody libraries, but also a minibodies library (Martin *et al.*, 1994).*

The high affinity data points show a roughly linear double logarithmic relationship, with a slope in the range of −1.2 to −2 [Fig. 2(b), (d), (f)] thus providing a rationalization for the previously observed linear double logarithmic behavior (Bradbury, 1997; Hoogenboom, 1997). It is seen that through the Poisson-based RAD or similar distributions one can adequately represent all experimental data covering the range of both low affinity and high affinity data.

Non-covalent interactions between ligands and receptors have been extensively analysed in terms of the relevant molecular forces. The contributions of various elementary interactions to the binding may be computed by force field calculations, and docking programs may be used to predict the configuration within specific ligand–receptor pairwise complexes (Kuntz *et al.*, 1982). Such analyses, which depend on detailed knowledge of specific molecular structures, are suitable for studies of specific binding partners. Yet, when studying receptor and ligand repertoires, such a detailed approach may be less relevant, and it may become highly significant to pursue the statistical behavior of the entire ensemble of binders. This may provide an insight to the general laws that govern non-covalent binding in biological systems.

*The term minibody was given by the authors to describe a small polypeptide fragment with a pre-determined structure or a novel function.

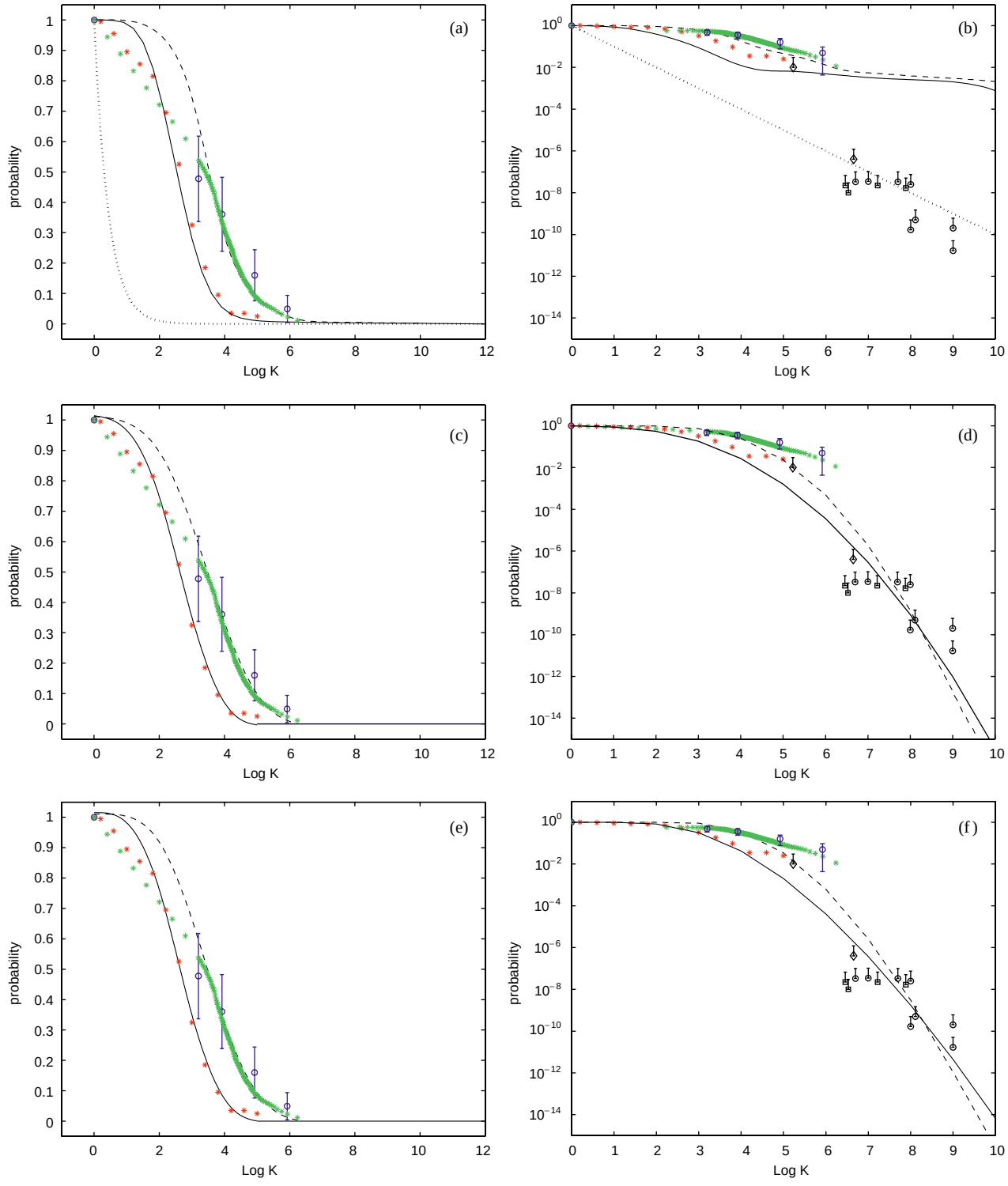

FIG. 2. Best fits of the cumulative distribution functions derived from Inmans model of multispecificity (a, b), the log-normal model (c, d), and the Poisson RAD (e, f). The cumulative distribution of the replotted data represents the probability of an affinity equal or higher than the specified $\log K$. Three reanalysed datasets are shown, as published by Inman and Barnett (1988) (★), Varga *et al.* (1991) (☆) and Kauvar (Kauvar *et al.*, 1995) (○). For each model, the parameter fitting was performed separately with the Varga data (— —) set and the Inman data set (——). In the fitting of the Varga data set, only data points with $\log(K) > 3.7$ were used, and for the Inman data set, a cutoff of $\mathrm{Log}(K) > 2$ was utilized (a,c,e). The Kauvar data set was not independently analysed by parameter fit because of the coarse affinity scale and large variance in the original data. The resulting parameters obtained through the least-squares analysis are given in Table 1 along with their corresponding RMS (root mean square) values. Also shown is a computed curve based on a log-linear model (2a, b) (· · · · · ·), with a slope of -1 (Mandecki *et al.*, 1995). To verify the validity of the suggested models in the high-affinity realm, the models were fitted to the two low affinity data sets, each combined with the high affinity data (b,d,f). These results were plotted as double logarithmic plots. The maximal affinity data is for peptides (□) and antibody (○) phage display libraries as well as minibody libraries (◇). These are plotted against inverse theoretical library size as an approximation for the probability of the maximal affinity. The unidirectional error bars indicate an average suggested correction of half-order of magnitude for the true size of the ligand repertoire due to a decrease in phage viability caused by the displayed fragments (Hoogenboom, 1997). The parameter values for the three models are presented in Table 1.

doi: 10:1006/jtbi2002.2538

The most crucial aspect of such a statistical view of biological recognition rests in the notion of affinity distributions—the relative pre-ponderance of target affinity values within a ligand repertoire (Kauvar *et al.*, 1995; Lancet *et al.*, 1994a; Levitan, 1997; Richards *et al.*, 1975). An important notion pursued in the present paper is that certain biological affinity distributions may manifest continuity. Accordingly, we conjectured that individual measurements for several hundred randomly chosen ligands (low affinity domain), and data on "winning" ligands in combinatorial libraries (high affinity domain), may be described by the same mathematical function. We therefore analysed such seemingly disparate measurements and attempted to fit them concomitantly to specific forms of affinity distributions.

The Poisson-based RAD model, whose simplicity is deemed advantageous, is depicted here as a limiting case of the previously proposed binomial distribution RAD model. The values of the λ parameter obtained in all the analyses were in the range of 8–16, suggesting that in a distribution of affinities for randomly selected ligands the typical value of the number of formal interactions is about a dozen. This agrees well with an analysis based on protein data bank structures for ligand–receptor pairs (M. Levitt, Pers. Comm.). When 26 different pairs of structures for ligand–protein complexes with ligands up to $\sim$300 Da were analysed, the average number of contacts (>2 Å^2 in area) was 12.

In the computer fits for the Poisson model, each of the formal interactions was found to be associated with a free energy change of around $-1/3$ kcal mol^{-1}, as indicated by the best-fit value for the parameter α. Such a value is well within the range of energies ($\sim 0.5\ RT$) typical of individual non-covalent contacts, e.g. hydrogen bonds or van der Waals interactions (Chaires, 1997; Klebe & Bohm, 1997; Levitt, 1974).

In the past, only a few attempts have been made to analyse low affinity data for ligand collections. In one case, a multi-parameter binome-related model of binding multispecificity (Inman, 1978) was used to examine a subset of the hapten binding data (Inman & Barnett, 1988; Varga *et al.*, 1991) which are fully reanalysed here. Inman's formalism, similar to our own, modeled the binding process as a set of local weak Van der Waals interactions (with interaction energy a) occurring between the binding surfaces. The number of successful interactions, v, was sampled from a binome distribution and was linearly related to the total binding strength. But in addition, the formalism also described the possibility that a single higher energy interaction, such as hydrogen bonding or salt bridges, may occur. While this multispecificity model fitted well the low affinity data [Fig. 2(a)], it is found here to strongly deviate from the data in the high-affinity part of the experimental distribution [Fig. 2(b)]. Other models which have been proposed include a log-normal (Goldstein, 1975; Macken & Perelson, 1991; Vant-Hull *et al.*, 1998) and a log-linear (Mandecki *et al.*, 1995) distribution. Parameters for both have been originally selected based on simple assumptions regarding the prevalence of maximal affinity ligands. The log-normal distribution with the originally published fitted parameters (Vant-Hull *et al.*, 1998) does not adequately explain the experimental data (not shown). However, we demonstrate here that a free parameter fit for the log-normal distribution results is a match to the experimental data that is as effective as that for the Poisson distribution. While the latter distribution has an advantage of providing insight into the multiplicity of elementary interactions and their energetics, using a log-normal distribution may be equally useful in some cases (Segre' *et al.*, 2001).

An important result of the present affinity distribution analysis is an extrapolation of the fitted curves from the low affinity to the high-affinity domain. Thus, based on the parameter fitted to data derived from relatively small ligand libraries (100–2000 members), which statistically attain affinities in the range of 10 mM to 1 μM, it is possible to derive information pertinent to the much higher affinities encountered within larger libraries. The probability and affinity range within which the combinatorial library data reside is found to fit adequately the extrapolated Poisson-based RAD curves as well as the log-normal curves fitted to the low affinity data. The curves corresponding to some of the other models also pass through this area, but

predict a slope considerably different from the experimental value of about −2 in the high affinity range (Bradbury, 1997; Hoogenboom, 1997).

The affinity ranges spanned by the combinatorial libraries of particular sizes is rather broad. This is probably the result of the fact that the panning selection procedures have a considerable stochastic component in it (Levitan, 1998). Furthermore, it has been argued that the panning procedure is limited in its ability to get the highest affinity in certain situations (Balass *et al*., 1996) and that the quoted size for many libraries constitutes an overestimate, relative to the effective number of expressed ligands (Hoogenboom, 1997). The latter effects would result in an under-estimate of the probability values. If corrected [as indicated by the unidirectional error bars, Fig. 2(b), (d), and (f)] it could bring more points into the range of the extrapolated low affinity curves.

Based on its ability to describe affinity patterns over a wide affinity range, more than eight orders of magnitude, the RAD model can serve as a conceptual framework for analysing molecular recognition in biological repertoires. In the future, it could be used to analyse immune phenomena such as self vs. non-self recognition (Mouthon *et al*., 1996; Nobrega *et al*., 1993), human olfactory threshold variability (Lancet *et al*., 1993, 1994a), and sensitivity to drugs based on changes in the Cytochromes P450 repertoire. A detailed knowledge of the statistics of affinity distributions can, in parallel, lead to a better understanding of molecular selection and *in vitro* evolution process such as combinatorial library panning and SELEX (Levitan, 1997, 1998; Mandecki *et al*., 1995; Vant-Hull *et al*., 1998) and high throughput screens for pharmaceutical drugs (Kauvar *et al*., 1995; Young *et al*., 1997).

Doron Lancet holds the Ralph and Lois Silver Chair in Human Genomics. Supported by the Crown Human Genome Center, a Ministry of Science grant to the National Laboratory for Genome Infrastructure, by the National Institutes of Health (DC00305), the Krupp foundation, and the Weizmann Institute Glasberg, Levy, Nathan Brunschwig and Levine funds. We thank Prof. M. Levitt, Stanford University, for the enlightening discussions and assistance with the PDB analysis.

REFERENCES

Aujame, L., Geoggroy, F. & Sodoyer, R. (1997). High affinity human antibodies by phage display. *Hum Antibodies* **8,** 155–168.

Balass, M., Morag, E., Bayer, E. A., Fuchs, S., Wilchek, M. & Katchalski-Katzir, E. (1996). Recovery of high-affinity phage from a nitrostreptavidin matrix in phage-display technology. *Anal. Biochem.* **243,** 264–269.

Barrett, R. W., Cwirla, S. E., Ackerman, M. S., Olson, A. M., Peters, E. A. & Dower, W. J. (1992). Selective enrichment and characterization of high affinity ligands from collections of random peptides on filamentous phage. *Anal. Biochem.* **204,** 357–364.

Bolhuis, H., Van, V. H., Poolman, B., Driessen, A. J. & Konings, W. N. (1997). Mechanisms of multidrug transporters. *FEMS Microbiol. Rev.* **21,** 55–84.

Bowman Jr., A. F. (1963). A method for the determination of heterogeneity of antibodies. *J. theor. Biol.* **4,** 242–253.

Bradbury, A. (1997). Recent advances in phage display: the report of the Phage Club first meeting. *Immunotechnology* **3,** 227–231.

Bruni, C., Gandolfi, A. & Germani, A. (1984). Analysis of the parameter constraints for a proposed antibody affinity distribution. *J. theor. Biol.* **109,** 71–76.

Buck, L. & Axel, R. (1991). A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65,** 175–187.

Burnet, F. M. (1963). Cold spring harbor symp. *Quant. Biol.* **32,** 1.

Burton, D. R. (1995). Phage display. *Immunotechnology* **1,** 87–94.

Chaires, J. B. (1997). Energetics of drug–DNA interactions. *Biopolymers* **44,** 201–215.

Clackson, T., Hoogenboom, H. R., Griffiths, A. D. & Winter, G. (1991). Making antibody fragments using phage display libraries. *Nature* **352,** 624–628.

Collins, J. (1997). Phage display. In: *Annual Reports in Combinatorial Chemistry and Molecular Diversity* (Ellington, A. D., Moos, W. H., Pavia, M. R., Kay, B. K., eds), pp. 210–262. The Netherlands: ESCOM.

Cupp, M. J. & Tracy, T. S. (1998). Cytochrome P450: new nomenclature and clinical implications. *Am Fam. Physician* **57,** 107–116.

Cwirla, S. E., Peters, E. A., Barrett, R. W. & Dower, W. J. (1990). Peptides on phage: a vast library of peptides for identifying ligands. *Proc. Natl Acad. Sci. U.S.A.* **87,** 6378–6382.

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

Detours, V., Mehr, R. & Perelson, S. A. (2000). Deriving quantitative constraints on T cell selection from data on the mature T cell repertoire *J. Immunol.* **164,** 121–128.

Detours, V. & Perelson, S. A. (1999). Explaining high alloreactivity as a quantitative cosequence of affinity-driven thymocyte selection. *Proc. Natl Acad. Sci. U.S.A.* **96,** 5153–5158.

Detours, V. & Perelson, S. A. (2000). The paradox of alloreactivity and self MHC restriction: quantitative analysis and statistics. *Proc. Natl Acad. Sci. U.S.A.* **97,** 8479–8483.

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. theor. Biol.* **183,** 409–416.

ERWIN, P. M. & ALADJEM, F. (1976). The heterogeneity of antibodies with respect to equilibrium constants. Calculation by a new method using delta functions and analysis of the results. *Immunochemistry* **13,** 873–883.

ESHHAR, Z., OFARIM, M. & WAKS, T. (1980). Generation of hybridomas secreting murine reaginic antibodies of anti-DNP specificity. *J. Immunol.* **124,** 775–780.

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

GOLDSTEIN, B. (1975). Theory of hapten binding to IgM: the question of repulsive interactions between binding sites. *Biophys. Chem.* **3,** 363–367.

GRIFFITHS, A. D. & DUNCAN, A. R. (1998). Strategies for selection of antibodies by phage display. *Curr. Opin. Biotechnol.* 9, 102–108.

GRIFFITHS, A. D., MALMQVIST, M., MARKS, J. D., BYE, J. M., EMBLETON, M. J., MCCAFFERTY, J., BAIER, M., HOLLIGER, K. P., GORICK, B. D. & HUGHES-JONES, N. C. (1993). Human anti-self antibodies with high specificity from phage display libraries. *EMBO J.* **12,** 725–734.

BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. & BOURNE. (2000). The protein data bank. *Nucleic Acids Research*, **28,** 235–242.

HOOGENBOOM, H. R. (1997). Designing and optimizing library selection strategies for generating high-affinity antibodies. *Trends Biotechnol.* **15,** 62–70.

HOROVITZ, A. (1996). Double-mutant cycles: a powerful tool for analyzing protein structure and function. *Fold Des.* **1,** R121–126.

HOROVITZ, A. & RIGBI, M. (1985). Protein–protein interactions: additivity of the free energies of association of amino acid residues. *J. theor. Biol.* **116,** 149–159.

INMAN, J. K. (1978). The antibody combining region: speculations on the hypothesis of general multispecificity. In: *Theoretical Immunology* (Bell, G. I. P. A. S., Pimbly Jr. G. H., eds), New York: Marcel Dekker.

INMAN, J. K. (1983). A study of multispecific interactions by quantative affinity chromatography. In: *Affinity Chromatography and Biological Recognition* (Chaiken, I. M., Wilcheck, M. Parkih, I., eds), Orlando, Academic Press.

INMAN, J. K. & Barnett, A. L. (1988). Affinities of antibodies for diverse ligands—theoretical and practical aspects. In: *Protein Recognition of Immobilized Ligand*; *UCLA Symposia on Molecular and Cellular Biology* (Hutchens, T. W., ed.), pp. 35–44. New York: Liss A. R.

JANDA, K. D., LO, L. C., LO, C., SIM, M. M., WANG, R., WONG, C. H. & LERNER, R. A. (1997). Chemical selection for catalysis in combinatorial antibody libraries. *Science* **275,** 945–948.

KAUVAR, L. M., HIGGINS, D. L., VILLAR, H. O., SPORTSMAN, J. R., ENGQVIST-GOLDSTEIN, A., BUKAR, R., BAUER, K. E., DILLEY, H. & ROCKE, D. M. (1995). Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* **2,** 107–118.

KLEBE, G. & BOHM, H. J. (1997). Energetic and entropic factors determining binding affinity in protein-ligand complexes. *J. Recept. Signal. Transduct. Res.* **17,** 459–473.

KRAMER, A., VAKALOPOULOU, E., SCHLEUNING, W. D. & SCHNEIDER-MERGENER, J. (1995). A general route to fingerprint analyses of peptide–antibody interactions using a clustered amino acid peptide library: comparison with a phage display library. *Mol. Immunol.* **32,** 459–465.

KUNTZ, I. D., BLANEY, J. M., OATLEY, S. J., LANGRIDGE, R. & FERRIN, T. E. (1982). A geometric approach to macromolecule–ligand interactions. *J. Mol. Biol.* **161,** 269–288.

LANCET, D. (1986). Vertebrate olfactory reception. *Annu. Rev. Neurosci.* **9,** 329–355.

LANCET, D. & BEN-ARIE, N. (1993). Olfactory receptors. *Curr. Biol.* **3,** 668–674.

LANCET, D., HOROVITZ, A. & KATCHALSKI-KATZIR, E. (1994a). Molecular recognition in biology: models for analysis of protein–ligand interactions. In: *The Lock-and-Key Principle* (Behr, J. P., ed.), pp. 25–71. New York: John Wiley & Sons Ltd.

LANCET, D., KEDEM, O. & PILPEL, Y. (1994b). Emerge of order in small autocatalytic sets maintained far from equilibrium: application of a probabilistic receptor affinity distribution (RAD) model. *Ber Bunseg. Phys. Chem.* **98,** 1166–1169.

LANCET, D., SADOVSKY, E. & SEIDEMANN, E. (1993). Probability model for molecular recognition in biological receptor repertoires: significance to the olfactory system. *Proc. Natl Acad. Sci. U.S.A.* **90,** 3715–3719.

LEVITAN, B. (1997). Models and search strategies for applied molecular evolution. In: *Annual Reports in Combinatorial Chemistry and Molecular Diversity* (Ellington, A. D., Moos, W. H., Pavia, M. R. & Kay, B. K., eds), pp. 95–152. The Netherlands: ESCOM.

LEVITAN, B. (1998). Stochastic modeling and optimization of phage display. *J. Mol. Biol.* **277,** 893–916.

LEVITT, M. (1974). Energy refinement of hen egg-white lysozyme. *J. Mol. Biol.* **82,** 393–420.

LORSCH, J. R. & SZOSTAK, J. W. (1994). *In vitro* selection of RNA aptamers specific for cyanocobalamin. *Biochemistry* **33,** 973–982.

LOHSE, P. A. & SZOSTAK, J. W. (1996). Ribozyme-catalysed amino-acid transfer reactions. *Nature* **381,** 442–444.

MACKEN, C. A. & PERELSON, A. S. (1991). Affinity maturation on rugged landscapes. In: *Molecular Evolution on Rugged Landscapes*: *Santa Fe Institute Studies in the Sciences of Complexity* (Perelson, A. S., Kauffman, S. A., eds), Reading, MA: Addison-Wesley.

MANDECKI, W., CHEN, Y. C. & GRIHALDE, N. (1995). A mathematical model for biopanning (affinity selection) using peptide libraries on filamentous phage. *J. theor. Biol.* **176,** 523–530.

MARTIN, F., TONIATTI, C., SALVATI, A. L., CILIBERTO, G., CORTESE, R. & SOLLAZZO, M. (1996). Coupling protein design and *in vitro* selection strategies: improving specificity and affinity of a designed beta-protein IL-6 antagonist. *J. Mol. Biol.* **255,** 86–97.

MARTIN, F., TONIATTI, C., SALVATI, A. L., VENTURINI, S., CILIBERTO, G., CORTESE, R. & SOLLAZZO, M. (1994). The affinity-selection of a minibody polypeptide inhibitor of human interleukin-6. *EMBO J.* **13,** 5303–5309.

MATLAB, (1994). *Statistics TOOLBOX For Use with MATLAB, User's Guide*. Natick, MA., The MATH WORKS Inc.

MOUTHON, L., LACROIX-DESMAZES, S., NOBREGA, A., BARREAU, C., COUTINHO, A. & KAZATCHKINE, M. D. (1996). The self-reactive antibody repertoire of normal human serum IgM is acquired in early childhood and remains conserved throughout life. *Scand. J. Immunol.* **44,** 243–251.

NEBERT, D. W. & GONZALEZ, F. J. (1987). P450 genes: Structure, evolution, and regulation. *Annu. Rev. Biochem*. **56,** 945–993.

NOBREGA, A., HAURY, M., GRANDIEN, A., MALANCHERE, E., SUNDBLAD, A. & COUTINHO, A. (1993). Global analysis of antibody repertoires. II. Evidence for specificity, self-selection and the immunological ''homunculus'' of antibodies in normal serum. *Eur. J. Immunol.* **23,** 2851–2859.

OSBOURN, J. K., FIELD, A., WILTON, J., DERBYSHIRE, E., EARNSHAW, J. C., JONES, P. T., ALLEN, D. & MCCAFFERTY, J. (1996). Generation of a panel of related human scFv antibodies with high affinities for human CEA. *Immunotechnology* **2,** 181–196.

PARSONS, H. L., EARNSHAW, J. C., WILTON, J., JOHNSON, K. S., SCHUELER, P. A., MAHONEY, W. & MCCAFFERTY, J. (1996). Directing phage selections towards specific epitopes. *Protein Eng.* **9,** 1043–1049.

PERELSON, S. A. & WEISBUCH, G. (1997). Immunology for physicists. *Rev. Mod. Phys.* **69,** 1219–1267.

PLUNKETT, M. J. & ELLMAN, J. A. (1997). Combinatorial chemistry and new drugs. *Sci. Am.* **276,** 68–73.

RICHARDS, F. F., KONIGSBERG, W. H., ROSENSTEIN, R. W. & VARGA, J. M. (1975). On the specificity of antibodies. *Science* **187,** 130–137.

SCHIER, R., MARKS, J. D., WOLF, E. J., APELL, G., WONG, C., MCCARTNEY, J. E., BOOKMAN, M. A., HUSTON, J. S., HOUSTON, L. L. & WEINER, L. M. (1995). *In vitro* and *in vivo* characterization of a human anti-c-erbB-2 single-chain Fv isolated from a filamentous phage antibody library. *Immunotechnology* **1,** 73–81.

SCHULTZ, P. G. & LERNER, R. A. (1995). From molecular diversity to catalysis: lessons from the immune system. *Science* **269,** 1835–1842.

SCOTT, J. K. & SMITH, G. P. (1990). Searching for peptide ligands with an epitope library. *Science* **249,** 386–390.

SEGRE', D., SHENHAV, B., KAFRI, R. & LANCET, D. (2001). The molecular roots of compositional inheritance. *J. theor. Biol.* **213,** 481–491.

SIPS, R. J. (1948). On the structure of a catalyst surface. *J. Chem. Phys.* **16,** 490–495.

SMITH, J. D., FORREST, S., ACKLEY, H. D. & PERELSON, S. A. (1999). Variable efficacy of repeated annual influenza vaccination. *Proc. Natl Acad. Sci. U.S.A.* **24,** 14 001–14 006.

TUERK, C. (1997). Using the SELEX combinatorial chemistry process to find high affinity nucleic acid ligands to target molecules. *Meth. Mol. Biol.* **67,** 219–230.

VALADON, P., NUSSBAUM, G., BOYD, L. F., MARGULIES, D. H. & SCHARFF, M. D. (1996). Peptide libraries define the fine specificity of anti-polysaccharide antibodies to Cryptococcus neoformans. *J. Mol. Biol.* **261,** 11–22.

VANT-HULL, B., PAYANO-BAEZ, A., DAVIS, R. H. & GOLD, L. (1998). The mathematics of SELEX against complex targets. *J. Mol. Biol.* **278,** 579–597.

VARGA, J. M., KALCHSHMID, G., KLEIN, G. F. & FRITSCH, P. (1991). Mechanism of allergic cross-reaction—I. Multispecific binding of ligands to a mouse monoclonal anti-DNP IgE antibody. *Mol. Immunol.* **28,** 641–654.

YEE, E. (1991). Reconstruction of the antibody affinity distribution from experimental binding data by a minimum cross-entropy procedure. *J. theor. Biol.* **153,** 205–227.

YOUNG, S. S., SHEFFIELD, C. F. & FARMEN, M. (1997). Optimum utilization of a compound collection or chemical library for drug discovery. *J. Chem. Inf. Comput. Sci.* **5,** 892–899.