Theoretical Principles of In Vitro Selection Using Combinatorial Nucleic Acid Libraries

Over the past decade, a new paradigm for drug discovery (Gold, 1995) and biological research (Gold et al., 1995) has been developed from technologies that integrate combinatorial chemistry with rounds of selection and amplification, a technique that is called in vitro selection. Systematic Evolution of Ligands by EXponential enrichment, or SELEX (Ellington and Szostak, 1990; Tuerk and Gold, 1990) is a flexible and extremely successful form of this technology that uses combinatorial libraries of oligonucleotides containing regions of randomized sequence as potential ligands. Oligonucleotide libraries (containing randomized regions) provide, after selection, compounds that bind tightly to the intended target. The process of in vitro selection was called SELEX by Tuerk and Gold (1990), while the selected compounds were called aptamers by Ellington and Szostak (1990). SELEX and in vitro selection (from oligonucleotide libraries) are identical. The selected and amplified bonding site (SAAB) technology (Blackwell and Weintraub, 1990) is a specialized form of SELEX directed toward finding naturally occurring sequences that bind proteins in vivo; however, the number of unique sequences used for SAAB analysis is usually much smaller than that used in most SELEX experiments, since the size of the binding area is usually well defined and thus the number of mutagenized nucleotides is small. SELEX and other adaptive molecular evolution techniques, such as phage display (Cwirla et al., 1990; Scott and Smith, 1990; Kay, 1994; Winter et al., 1994), gain much of their power from their ability to isolate individual molecules from vast molecular pools without resorting to cumbersome deconvolution or tagging methods commonly used in combinatorial chemistry schemes. Rather, these methods utilize iterative rounds consisting of ligand selection from combinatorial libraries followed by amplification of these selected ligands to form new libraries enriched for the particular function of interest, e.g., affinity binding or catalytic function. Such techniques enable quite rapid searches of enormous libraries (typically greater than 10^{15} potential ligands in the case of SELEX). SELEX has been used to discover high-affinity ligands to a wide variety of different molecular targets, including nucleic acid binding proteins, non–nucleic acid binding proteins, peptides, and small organic molecules (reviewed in Klug and Famulok, 1994; Gold, 1995; Gold et al., 1995). This unit presents a theoretical overview of in vitro affinity selection using SELEX technology.

A schematic representation of the SELEX process is shown in Figure 9.1.1 and may be used to describe SELEX performed with libraries of RNA, RNA derivatives, or DNA. For the purposes of developing a mathematical model, the SELEX process for affinity binding may be summarized in four steps: (1) generation of a library of potential ligands, (2) binding of the library to the target molecule, (3) partitioning of the bound ligands from the unbound ligands, and (4) amplification of the partitioned ligands to generate a new, enriched library, leading again to step 1. Repeated application of steps 2 to 4 results in an enriched pool composed of the sequences of interest. For selection of single-stranded DNA (*UNIT 9.2*), the two strands of the PCR-amplified pool of dsDNA must be denatured, and one of the strands isolated before binding with the target. For selection of RNA and RNA derivatives (*UNIT 9.3*), the PCR-amplified pool of dsDNA must be transcribed to form a pool of RNA before binding with the target. The partitioned RNA must then be reverse-transcribed into DNA before PCR amplification. For the present analysis, these enzymatic transformations—reverse transcription (RT), PCR, and transcription—are all assumed to be perfect, meaning that they do not affect the relative concentrations of the ligands. However, see Mathieu-Daudé et al., (1996) regarding imperfect amplification due to concentration differences, and Sun et al. (1996) for mathematical modeling of the amplification process taking stochastic effects into account.

SELEX is a very forgiving technology. High-affinity ligands to nearly any desired target may be found even when the selection conditions (protein and RNA concentrations, for example) are far from optimal. However, great savings in time and material, or perhaps even success with difficult targets, may be achieved by working at the optimal conditions. Determining what these conditions are demands a deeper understanding of the mechanisms of SELEX. We present such a theoretical model here. We first describe the characteristics of a ligand library, comprised of oligonu-

Contributed by Barry Vant-Hull, Larry Gold, and Dominic A. Zichi

Current Protocols in Nucleic Acid Chemistry (2000) 9.1.1-9.1.16

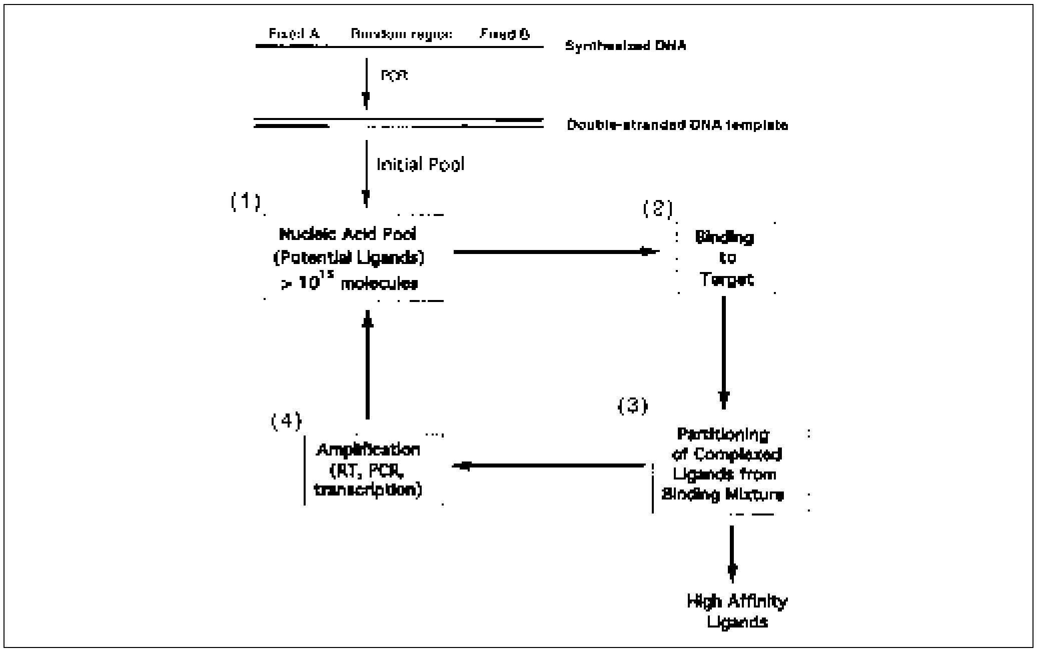


Figure 9.1.1 Schematic representation of the SELEX in vitro selection methodology. The initial random pool is derived from synthesized DNA oligonucleotides that are used directly for DNA SELEX or converted to double-stranded templates for transcription for RNA SELEX. Once the initial pool is created, the steps for a round of affinity binding SELEX are presented in the squares; (1) pool generation, (2) incubation with target, (3) partitioning, and (4) enzymatic amplification.

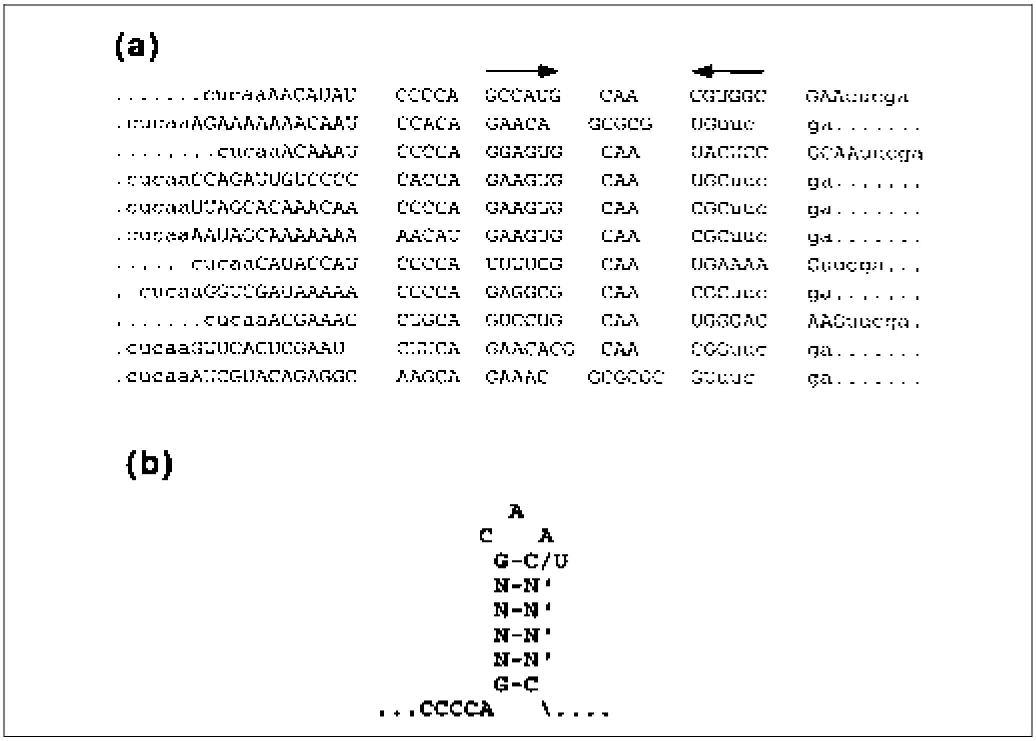
cleotide sequences, that are relevant to the model. We then describe the equilibrium selection model central to SELEX, incorporating these library characteristics. A demonstration of the model applied to experimental data is then presented. Analytical expressions for the optimal nucleic acid and protein concentrations are derived, these being two parameters easily varied during SELEX experiments. However, the formulas for optimal concentrations unfortunately depend on parameters that cannot easily be determined experimentally. We therefore introduce a new parameter, the signal-to-noise ratio, which allows the determination of near-optimal conditions based only on parameters that are easily determined experimentally.

NUCLEIC ACID LIBRARIES

In vitro selection is performed with nucleic acid libraries containing vast numbers of unique molecules, typically ~10^{15} sequences. Such large libraries are desirable in order to saturate the sequence space of longer randomized regions—a useful goal, as SELEX is often directed toward non–nucleic acid–binding proteins that are unlikely to have sites with high affinity and specificity to short sequence regions. Even with known nucleic acid–binding proteins, longer randomized regions may provide a larger contact surface, often making it possible to find sequences that bind with higher affinity than the wild-type binding sequences. The only practical limitation on library size is imposed by the volumes of material manipulated experimentally; 10^{15} random sequences are easily synthesized and readily processed. Each sequence in the library is composed of a random region of variable length sandwiched between two regions of fixed sequence used for primer binding sites during enzymatic processing. The length of the random region varies considerably among selection experiments. For affinity binding, most studies use between 20 and 60 nucleotides (Gold et al., 1995), while researchers performing catalytic selections typically use much larger random regions, the largest comprising >200 nucleotides (Hager et al., 1996; Breaker, 1997). The motivation for the difference in sequence length is that binding interactions with proteins and small molecules may require smaller molecular arrangements than those needed to carry out enzymatic activity. It is commonly believed that typical catalytic oligonucleotides have multiple secondary structural domains that may be required for activity, but this hypothesis still awaits rigorous proof.

A basic tenet of in vitro selection experiments is that the selected function of oligonucleotide molecules is conferred through their three-dimensional structures. These structures, usually supported by stacking interactions between adjacent base pairs, are a consequence of the individual sequences. The identification of conserved primary structural units (residues) and secondary structural units (e.g., helices and loops) from those sequences sharing a selected function allows one to define a motif required for the function. Once a motif is defined, it is easy to compute its frequency of occurrence in the initial pool. For example, SELEX-isolated sequences that bind with high affinity to the *E. coli* rho factor, displayed in Figure 9.1.2a, define the hairpin motif shown in Figure 9.1.2b. There are 4^4 combinations of base pairs forming the central stem (N-N′) and 2 bases (C/U) tolerated at position 15 out of 20 contiguous nucleotides defining the motif. Since the motif can start at 11 different positions within the 30-nucleotide random region used in this experiment, the sequence for such a motif occurs in the initial pool with a frequency of $(4^4 \times 2 \times 11)/4^{20} = 5 \times 10^{-9}$. As discussed below, such an estimate is always an upper limit on the frequency of actual high-affinity ligands. In addition, the actual frequency of high-affinity ligands is likely to be different than that calculated from the consensus motif, since this motif is usually defined from a sampling of relatively few sequences and is thus underdetermined. We show below that the frequency of selected motifs within the initial pool plays a central role in the progress of SELEX experiments.

For a continuous motif of length *m*, increasing the random region by *n* bases results in an *n*-fold increase for representation of that motif in the random sequence library. However, calculating the frequency of occurrence of a particular motif within the original sequence pool is certainly an upper limit on the number of active molecules with that motif, since this estimate does not take into account the likelihood that a particular sequence will fold into the motif of interest. As the length of the random region increases, the number of possible secondary and three-dimensional structures formed increases, decreasing the likelihood that the motif of interest is thermodynamically accessible in any particular sequence in which it occurs (Sabeti et al., 1997). It is difficult to estimate the loss of activity due to alternative folds, but, clearly, as the length of a sequence



```
.......cucaaAACAUAU   CCCCA  GCCAUG    CAA   CGUGGC  GAAcucga
.cucaaAGAAAAAAACAAU   CCACA  GAACA    GCGCG  UGuuc   ga.......
........cucaaACAAAU   CCCCA  GGAGUG    CAA   UACUCC  CCAAcucga
.cucaaCCAGAUUGUCCCC   CACCA  GAAGUG    CAA   UGCuuc  ga.......
.cucaaUUAGCACAAACAA   CCCCA  GAAGUG    CAA   CGCuuc  ga.......
.cucaaAAUAGCAAAAAAA   AACAU  GAAGUG    CAA   CGCuuc  ga.......
.....  cucaaCAUACCAU  CCCCA  UUUUCG    CAA   UGAAAA  Cuucga...
.  cucaaGGUUGAUAAAAA  CCCCA  GAGGCG    CAA   CCCuuc  ga.......
.......cucaaACGAAAC   CUGCA  GUCCUG    CAA   UGGCAC  AAGcucga.
.cucaaGUUCACUCGAAU    CUUCA  GAACACG   CAA   CGCuuc  ga.......
.cucaaAUCGUACAGAGGC   AAGCA  GAAAC    GCGCGC GUuuc   ga.......
```

Figure 9.1.2 Consensus sequences and motif obtained from *E. coli* rho SELEX (Schneider et al., 1993). The sequences (**a**) are aligned to reveal the consensus motif (**b**), a hairpin loop. Primer binding sites are denoted by lower case letters in (**a**) and N-N′ denote any Watson-Crick or G-U base pair in (**b**).

increases, these losses become more significant.

Complete coverage of all sequences for a given-size random region can only be achieved with relatively small random regions. Since the number of unique molecules in the starting pool is practically limited to 10^{15} molecules and there are 4^N sequences for a random region containing N positions, the set of all possible sequences (commonly called the sequence space) for libraries with $N > 25$ is not fully represented at the outset of the experiment (Ciesiolka et al., 1996). Further, most in vitro selection experiments do not rely on mutation to alter the makeup of sequences after selection; therefore the molecules possessing the desired functional activity must be present in the starting pool. As the size of the random region grows, coverage of the full sequence space diminishes exponentially, compounding the problem of culling rare sequences from a limited initial pool.

To help overcome this problem of limited sequence space coverage for libraries with large random regions, optimization of “lead” sequences isolated with in vitro selection is commonly performed with further rounds of selection, starting with a biased pool. Typically, the best sequence resulting from a selection is isolated from the enriched pool. A second pool of molecules is then constructed from this lead sequence by biasing the nucleotide content at each position to contain, for example, 70% original sequence and 10% the remaining three bases. This is a common strategy employed in catalytic RNA selection schemes; a rare sequence isolated from pools with >100 random bases is usually not optimal but is a good starting point for further selection experiments. Such strategies have resulted in increased activities more than several-fold over that of the starting sequence (Hager et al., 1996; Breaker, 1997). Of course, such a tactic never guarantees that the best sequence within the overall sequence space has been isolated.

The primary focus of the theoretical development to follow will be on in vitro affinity selection experiments—so-called SELEX. Mutation and recombination events are not deliberately included in most SELEX methodologies, although lack of perfect enzymatic fidelity during transcription, reverse transcription, and PCR amplification certainly introduces a small number of mutation events. The effects of these mutations will be ignored in the present analysis, and are expected to be small in any event. SELEX experiments are usually performed to isolate those sequences present in the initial pool that have the highest binding affinities to the target of interest. Restricting our discussion to these in vitro selection schemes allows for an enormous reduction in complexity of representation; the vast library of sequences can be formally mapped onto affinity distributions with no loss in generality. Coupling this reduction with an equilibrium model for binding, that is easily achieved experimentally, we can accurately model in vitro selection experiments and, therefore, identify those features of the process most critical to experimental success.

AFFINITY PROBABILITY DISTRIBUTIONS

It is now commonly accepted that nucleic acid sequences can fold into complex three-dimensional shapes bolstered by their secondary structures. It is the three-dimensional display of functional groups on nucleic acid oligomers that is responsible for their differential binding affinities to wide-ranging target molecules. This is true for small-molecule targets, where the oligonucleotides typically fold to engulf the target, as well as protein targets, where extensive surfaces of the macromolecules are in direct contact. Regardless of the particular target, the three-dimensional structures adopted by the sequences, in both their free and bound states, can in principle be mapped to free energies of binding for the target of interest. This is our basis for mapping linear sequences onto a probability distribution of binding affinities. For an equilibrium model of in vitro selection, all those sequences with the same binding affinities will partition in precisely the same way between target-bound and free in solution. In other words, averages computed over distinct sequences are mathematically identical to averages over the binding affinity distribution constructed from the vast (10^{15}) sequence library. This formally exact reduction in complexity is key to our theoretical development presented below. In the following discussion, we will denote an average over an affinity distribution $p(K_a)$ by $<...>$, i.e., $<f(K_a)> = \int dK_a\, p(K_a)\, f(K_a)$, where $f(K_a)$ is some function of affinity and $\int dK_a\, p(K_a)$ has been normalized to one.

In order to construct an equilibrium model for in vitro selection, it is necessary to explicitly define the affinity distribution of the library for the target of interest. Clearly, the details of such distributions will change depending on the target of interest. We show below that certain features of this distribution are key for assessing the progress of selection experiments. This probability distribution is most conveniently cast in terms of association constants, $p(K_a)$, but can also be cast in terms of binding free energies, since $\Delta G = -k_B T \ln(K_a)$, where k_B is the Boltzmann constant and T is the absolute temperature. An experimental determination of $p(K_a)$ is quite difficult and there exist very little data on which to base estimates. For a model of double-stranded DNA-protein interactions based upon independent base-pair contributions to affinity, the correlation between nucleic acid information content and protein binding affinity leads to a binding free energies that are normally distributed (Berg and von Hippel, 1987; Stormo and Yoshioka, 1991). Consequently, the probability distribution of K_a values for this model are log-normal. Such a profile clusters the majority of sequences around $<\Delta G>$ ($<\ln(K_a)>$), while those sequences with most favorable ΔG of binding [$\ln(K_a)$] occur relatively rarely. The frequency of the highest-affinity molecules, the so-called "winners," in the pool depends on the width of the distribution and the difference in affinity between the majority of the sequences and the rare winners. Whether the true affinity distributions are log-normal or skewed in some fashion, we expect the general profiles to be consistent with these features.

Although direct experimental determination of $p(K_a)$ is not practical, some useful data can be extracted from binding curves of the sequence pool. A pool binding curve obtained with a constant, small nucleic acid concentration and varying protein concentrations yields some limited data for the distribution of affinities. Specifically, the affinity measured using a standard binding curve analysis for bimolecular association yields a value for $K_{bulk} = e^{<\ln(Ka)>}$; that is, the measured affinity reflects the $\ln(K_a)$ averaged over the distribution. The initial asymptotic behavior of the pool binding curve, corresponding to low total protein concentration, is best described by $<K_a>$, the average of K_a over the distribution, and this may be quite different from the K_{bulk}. Similarly, the behavior of the binding curve where protein is in excess follows $<K_a^{-1}>$, whose inverse may differ considerably from $<K_a>$ and K_{bulk}. Figure 9.1.3 illustrates this for two distributions of $p(K_a)$. We show in the next section that progress of SELEX during initial rounds is relatively insensitive to the shape of the distribution $p(K_a)$, but depends on the following three key aspects of $p(K_a)$: (1) the $<K_a>$ of the pool, (2) the association constant K_w of the highest-affinity sequences in the pool (winners), and (3) the frequency f_w of these winners. We show below that for a round of SELEX $<K_a>$ and K_w are the critical features of $p(K_a)$ that determine the increase in the population of the winners, whereas the frequency of winners in the initial pool sets the scale for the number of rounds required for completion. In general, K_w is a measure of the amount of winning molecules bound to protein, while $<K_a>$ is a measure of the amount of total molecules bound. Therefore a ratio of these affinities reflects the possible enrichment during a round of SELEX. In general, we believe that the average affinity $<K_a>$ can be from two to twenty times higher than K_{bulk} for affinity distributions $p(K_a)$ of initial pools.

For the purpose of testing the mathematical model of SELEX (described below) against completed experiments, we must define the key features of $p(K_a)$. There is currently no good way of experimentally determining $<K_a>$ directly without knowing the affinity distribution, for which there is little data. We can, however, rely on assumptions to fill this gap and make reasonable choices for parameters describing $p(K_a)$. K_{bulk} may be easily measured experimentally and used to establish a lower limit for $<K_a>$. Since the binding curve of a pool of nucleic acids usually closely approximates that of a single defined ligand (see, e.g., Schneider et al., 1993), we assume that the vast majority of ligands cluster around a "bulk" affinity. The affinities of the winning ligands, K_w, however, are usually several orders of magnitude greater than this bulk affinity. K_w is easily obtained at the completion of SELEX by cloning and sequencing the enriched pool and measuring the affinity of individual clones for the target. As previously described, further sequence analysis of the enriched pool often leads to the identity of a winning motif, which then allows for an estimation of the winner's frequency in the initial pool. These frequencies are typically on the order of 10^{-9} to 10^{-13} (Gold et al., 1995) and provide some information about the affinity distribution. Assuming that binding affinities are distributed log-normally, and having determined K_{bulk}, K_w, and f_w, a value for the width of the distribution, and therefore $p(K_a)$, can be determined. Though there are many possible distributions, for the purposes of this discussion, we will consider the log-normal distribution.

The log-normal affinity distribution of ligands for the target protein has the probability density function:

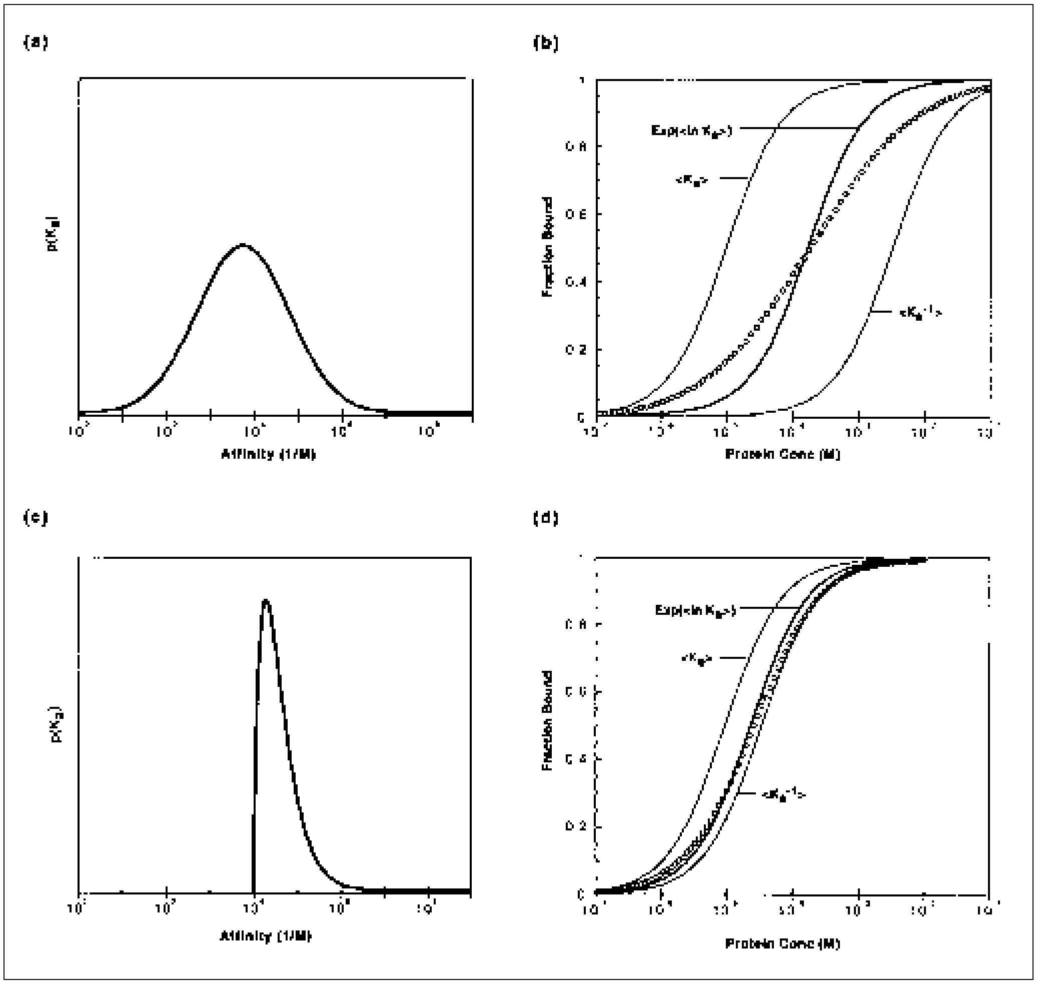


Figure 9.1.3 Affinity distributions and their resulting binding curves. A log-normal distribution of binding affinities is presented in (**a**) and a Poisson distribution in <lnK_a> is presented in (**c**). Calculated binding curves for these distributions are displayed in (**b**) and (**d**) and denoted by open circles. The one-component binding curves corresponding to affinities $e^{<\ln Ka>}$, <K_a>, and <K_a^{-1}> computed from the distributions are displayed as solid lines. Note that the binding curve derived from <lnK_a> best fits the overall binding curve generated from the distributions, while the low and high protein asymptotes are best fit by <K_a> and <K_a^{-1}>, respectively.

$$p(\ln K_a) = \frac{1}{\sqrt{2\pi\sigma}}\, e^{-(lnK_a - <\ln K_a>)^2/(2\sigma^2)}$$

Equation 9.1.1

where lnK_{bulk} = <lnK_a> is related to the bulk affinity of the initial pool, σ is the standard deviation of the log-normal distribution, and $p(\ln K_a)\ d\ln K_a$ is the fraction of ligands with affinities having logarithms equal to lnK_a ± (1/2)$d$lnK_a. In Equation 9.1.1), $K_{bulk} = e^{<\ln(Ka)>}$ is the affinity corresponding to the peak (and the midpoint) of the log-normal distribution. Since <lnK_a> can be determined experimentally, σ in Equation 9.1.1 is determined to match f_w at K_w.

A log-normal distribution for $p(K_a)$ allows an analytical calculation of the difference between <K_a> and K_{bulk}, given by the following expression:

$$\langle K_a\rangle = K_{bulk} e^{(\sigma^2/2)}$$

Equation 9.1.2

For instance, if a random pool of RNA has a measured bulk affinity of 10^6 M^{-1}, and has a winning motif with an affinity of 10^9 M^{-1} in the pool at a frequency of 10^{-10} (integrating the affinity distribution from ln K_w to infinity), then the average pool affinity <K_a> will be approximately three times greater than the measured bulk affinity.

AN EQUILIBRIUM MODEL FOR SELEX

For most SELEX applications, we have found the equilibrium model of Irvine et al. (1991) to be a good description of the in vitro selection process. This model assumes that there are no interactions among the different nucleic acid molecules in the pool and that no multiprotein aggregates form. Further, we replace the impractical summation over distinct sequences (on the order of 10^{15}) with an integral over the affinity probability distribution $p(K_a)$ for a particular protein, P. It is important to note that once $p(K_a)$ has been defined, there are no adjustable parameters in the model; the remainder of the variables are determined by the experimental conditions or are evaluated explicitly.

The total concentration of sequences with an affinity K_a is given by:

$$L_t(K_a) = p(K_a)L_t$$

Equation 9.1.3

where L_t is the total concentration of nucleic acid ligands. Concentrations of the ligand:protein complexes for a particular affinity K_a are given by:

$$[P : L(K_a)] = K_a[P][L(K_a)]$$

Equation 9.1.4

By applying mass conservation:

$$P_t = [P] + \int [P\colon L(K_a)]dK_a \qquad \text{(a)}$$

$$L_t(K_a) = [L(K_a)] + [P\colon L(K_a)] \qquad \text{(b)}$$

Equation 9.1.5

along with the equilibrium condition expressed in Equation 9.1.4, it is easily shown that the free protein concentration is given by:

$$[P] = \frac{P_t}{1 + \int_{(n)} \dfrac{K_a L_t(K_a)}{1 + K_a[P]} dK_a} = \frac{P_t}{1 + \langle \dfrac{K_a L_t}{1 + K_a[P]} \rangle}$$

Equation 9.1.6

where P_t is the total concentration of protein, both free and bound in complexes. Note that the integrated term in Equation 9.1.5a denotes the total amount of complex formed for all affinity species obtained by summing over the affinity distribution; a similar interpretation holds for the integrated term in Equation 9.1.6. Application of Equation 9.1.3 allows the integral in Equation 9.1.6 to be replaced with angle brackets (<...>). The only unknown in Equation 9.1.6 is [P] and may be conveniently solved for iteratively to self-consistency by choosing an initial free protein concentration of zero. Once [P] is determined, concentrations of all other species are easily determined with Equation 9.1.4 and Equation 9.1.5.

In the ideal case, all ligands that are complexed with protein would be carried into the next round of SELEX and all noncomplexed ligands would be lost. The optimal strategy for such an ideal partitioning would then be trivial: smaller concentrations of protein would always lead to better selection of the highest-affinity ligands over all other ligands.

Unfortunately, no partitioning method is ideal. Only a fraction of the protein:ligand complexes are recovered, and some portion of the noncomplexed ligands is partitioned along with the complexed ones. We call the fraction of correctly partitioning complexes the partitioning efficiency, *eff*. In the case of using nitrocellulose filters as a partitioning method, for example, the efficiency for different proteins may vary from closely approaching unity to as low as 0.1, where only 10% of the input protein (and complexes) is captured. For a given protein, the partitioning efficiency can be measured. We usually assume that this efficiency is the same for all protein:nucleic acid complexes for a particular protein. We define the background partitioning, *bg*, as that fraction of noncomplexed ligands which are recovered by the partitioning method. For nitrocellulose filters, the background is easily determined and is found to be typically on the order of 1% to 0.01%. We usually assume that this background value is the same for all ligands (however, see Concluding Remarks regarding ligands to nitrocellulose filters). We will show below that these deviations from ideality make the question of finding optimal conditions for SELEX an interesting one.

Application of the above considerations leads directly to an equation for determining the frequency, or probability $p_{(n+1)}(K_a)$, at which ligands $L(K_a)$ with an affinity K_a will exist in the sequence pool at round $n+1$ following a cycle of SELEX with a pool composition of $p_{(n)}(K_a)$ (Irvine et al., 1991):

$$p_{(n+1)}(K_a) = \frac{eff\, K_a[P][L(K_a)] + bg[L(K_a)]}{\int_{(n)} \{eff\, K_a[P][L(K_a)] + bg[L(K_a)]\} dK_a}$$

Equation 9.1.7

where $\int_{(n)} \ldots dK_a$ indicates a summation over the affinity distribution at round n, and thus represents the total amount of nucleic acid partitioned at round n. Note that Equation 9.1.7 is

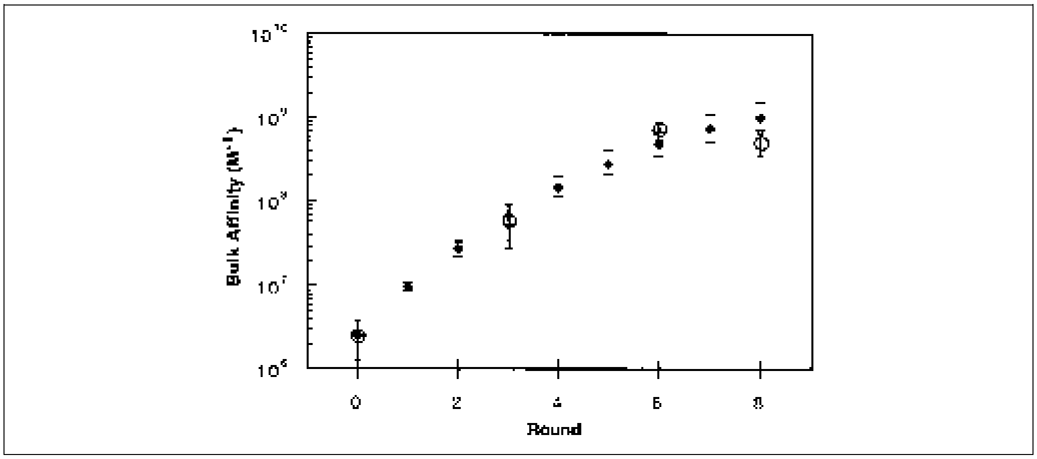


Figure 9.1.4 Comparison of enrichment obtained with experiment and theory. Enrichment, as measured by pool affinities determined during a SELEX experiment against *E. coli* rho factor (open circles) is compared to that calculated with the equilibrium model (filled diamonds). All parameters for this simulation, with no adjustable parameters, were obtained from Schneider et al. (1993). A log-normal initial distribution, parameterized as described in the text, was used in the simulation. The initial winner frequency is the parameter with the highest uncertainty in determination, and thus this parameter was varied as described below. Closed diamonds represent the simulation performed at an initial winner frequency of 10^{-10}. Bars above and below represent simulations using initial winner frequencies of 10^{-9} and 10^{-11}, respectively. The error bars on the experimental data are 95% confidence intervals found by fitting binding curves to the data of Schneider et al. (1993).

cast in terms of [*P*], determined from Equation 9.1.6, and [$L(K_a)$], determined from Equations 9.1.4 to 9.1.6, both evaluated with the pool at round *n,* and affinity distribution $p_{(n)}(K_a)$. The repeated application of this equation, with chosen concentrations of total protein and nucleic acid for each round, allows the initial affinity distribution to be propagated through multiple rounds of SELEX. The exponential enrichment is a consequence of equilibrium binding, i.e., $K_a = e^{-\Delta G/kBT}$, and is reflected in the changes in the affinity distributions from one round $p_{(n)}(K_a)$ to the next $p_{(n+1)}(K_a)$.

Equation 9.1.7 assumes that all partitioned sequences will be enzymatically processed with the same efficiency and with complete fidelity. In order to incorporate the effects of mutations during the RT, PCR, and transcription steps, a model for computing affinity as a function of sequence is required. As previously discussed, affinity selection using SELEX is designed to isolate those high-affinity sequences that exist in the initial pools, so treatment of mutations is not included here. To accurately model in vitro evolution, however, where mutations are a key characteristic of the technique, such a mapping is essential (Kauffman and Macready, 1995; Schuster, 1995). In our mathematical description here, however, we are concerned with selection and not evolution.

This completes our description of the mathematical model for SELEX. To assess the utility of this model, we present a comparison of simulated results to SELEX experimental results to find high-affinity oligonucleotides binding to *E. coli* Rho factor (Fig. 9.1.2), a transcription terminator (Schneider et al., 1993). For the simulated data, a log-normal distribution for $p(K_a)$ is assumed and parameterized with experimental data as described above. The remaining parameters, *bg*, *eff*, and the protein and RNA concentrations were taken from the SELEX experimental work. Progress of the enrichment during rounds of SELEX is monitored by pool affinity measured with binding curves. The simulated results compare quite well with the experiment (Fig. 9.1.4); the essential features of in vitro selection are quantitatively captured in the equilibrium model developed by (Irvine et al., 1991) and presented here in terms of affinity distribution functions.

We now address the issue of how a particular distribution affects the progress of SELEX experiments. We have constructed three distinct affinity distributions that have identical values for $<K_a>$, K_w, and f_w, and yet different overall distributions $p(K_a)$. We compare a two-point distribution consisting of the affinities $<K_a>$ and K_w, a log-normal distribution discussed in detail above (Fig. 9.1.3a), and a Poisson distribution $p(\ln K_a) = \alpha^2 x e^{-\alpha x}$, where $x = \ln K_a -$

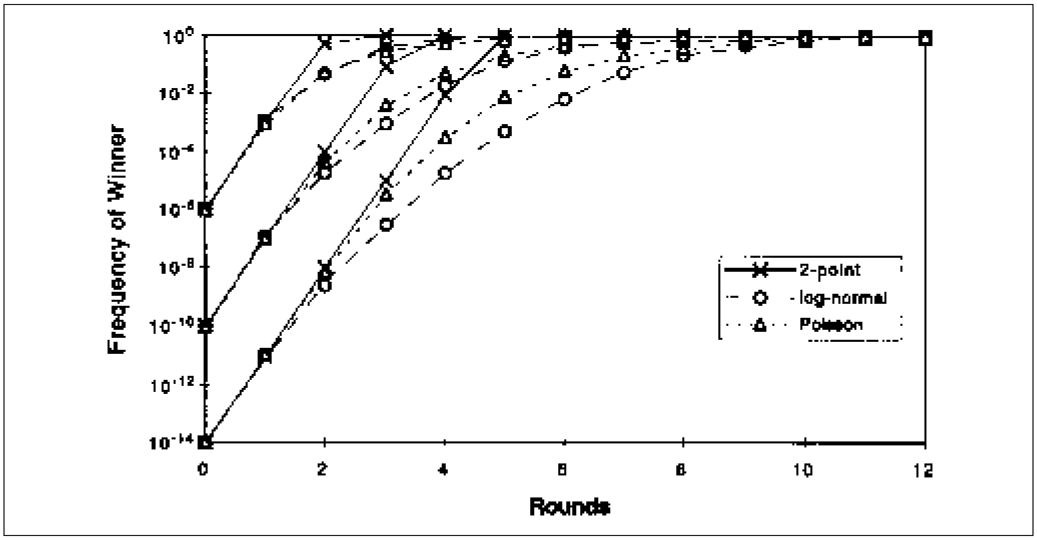


Figure 9.1.5 Comparison of three different initial affinity distributions versus progress of enrichment. The enrichment is measured here by the frequency of the winning ligands in the pool at each round for a two-point distribution (marked with x's) a log-normal distribution (marked with open circles) and a Poisson distribution (marked with triangles) of binding affinities.

<lnK_a> (Fig. 9.1.3c). The log-normal distribution spans the largest range of K_a values, followed by the Poisson distribution, and finally the two-point distribution. These initial distributions are used to initiate 12 rounds of SELEX performed in the absence of background and with identical conditions throughout. The results are displayed in Figure 9.1.5 for three different values of f_w. Sensibly, the in vitro selection is completed most rapidly for the case where the winning frequency is highest in the starting pool. In all cases of f_w, the enrichment for the first round, where the <K_a>, K_w, and f_w are well matched, is essentially the same for all three distributions, illustrating the importance that these features of the distribution have for enrichment. As the SELEX progresses, however, the distributions begin to differ in <K_a> and f_w, and some differences in enrichment progress begin to emerge. The two-point distribution moves towards full enrichment (100% K_w) fastest, followed by the Poisson, and finally the log-normal distribution. This is due to increased competition for binding reflected in the width of the distributions. Fewer species competing for limited binding helps drive the two-point distribution to completion, while the normal distribution takes many more rounds, on average, to completely saturate the high-affinity binders.

A simulation of SELEX requires that the affinity profile for the initial random pool of nucleic acids be defined. Although little is known about f_w and K_w at the outset of an experiment, there is, fortunately, a great deal of useful theory to help guide the design of in vitro selection experiments. In the next section, we present theoretical guides for SELEX that have been derived from the equilibrium model of SELEX presented here, even without a detailed knowledge of the affinity distribution. This set of analytical, theoretical results is, at least to first appearances, independent of the shape of the affinity profile.

OPTIMAL CONDITIONS FOR IN VITRO SELECTION

A variety of useful predictions may be derived from the SELEX theory as presented thus far without having to resort to computer simulations. The more useful of these predictions have to do with the conditions for optimal enrichment of the highest-affinity ligands. These conditions depend on the background, *bg*, the partitioning efficiency, *eff,* the affinity of the winning ligands, K_w, and the average pool affinity, <K_a>. It is important to remember that the pool affinity <K_a> used below is not necessarily the same as the affinity that is measured experimentally (see Fig. 9.1.3). This is especially true of the earliest rounds of SELEX.

There are two dimensionless terms that occur often in SELEX theory. We define these as:

$$\varepsilon = \frac{eff}{bg}$$

Equation 9.1.8

and

$$\boldsymbol{k} = \frac{K_{\mathrm{w}}}{K_{\mathrm{pool}}}$$

Equation 9.1.9

As defined, both the partitioning effectiveness ε and the affinity ratio ***k*** should be much greater than unity in most cases. The pool affinity is defined by the total complexed $[P{:}L(K_{\mathrm{a}})]$ and un-complexed $[L(K_{\mathrm{a}})]$ molecules summed over $p(K_{\mathrm{a}})$:

$$K_{\mathrm{pool}} = \int [P{:}L(K_{\mathrm{a}})]\, dK_{\mathrm{a}} / \{[P] \int [L(K_{\mathrm{a}})] dK_{\mathrm{a}}\}.$$

In other words, K_{pool} is the affinity for the pool as a whole that would result from a measurement of the total amount of ligand bound for the given protein and ligand concentration and distribution of affinities embodied in $p(K_{\mathrm{a}})$. It is important to note, however, that K_{pool} is not a true equilibrium constant since its value changes with total protein and ligand concentration. As the total amount of free protein goes to zero (excess nucleic acid), K_{pool} approaches $<K_{\mathrm{a}}>$. Approximating K_{pool} by $<K_{\mathrm{a}}>$ is actually quite good for most SELEX conditions since these conditions usually correspond to excess ligand concentrations.

The enrichment is defined as the factor by which the frequency of the "winner" in the affinity distribution changes between rounds of SELEX. Clearly, choosing experimental conditions to optimize enrichment leads to the greatest progress during a round of SELEX. Most of the discussion that follows focuses on guidelines to experimental conditions that maximize enrichment and therefore lead to most efficient in vitro selection schemes. It can be shown (Irvine et al., 1991) that the overall maximum enrichment is given by:

$$E^{\mathrm{opt}} = \left(\frac{1 = \sqrt{\varepsilon k}}{\sqrt{\varepsilon} + \sqrt{k}}\right)^2$$

Equation 9.1.10

This optimum may be achieved at any given total nucleic acid concentration by adjusting the total protein concentration, whereas the converse is not true. There are total protein concentrations whose corresponding optimal nucleic acid concentrations do not achieve the maximal enrichment possible; this behavior is discussed in detail below.

Optimal Concentrations

The effect of protein concentration on enrichment at several different background levels is shown in Figure 9.1.6a. At high protein concentrations (above $<K_a^{-1}>$) nearly all ligands are bound to an equal extent, and no selection takes place. At low concentrations, the amount of nucleic acid partitioned in the background overwhelms the complexed nucleic acid molecules, and again there is no selection. Clearly, in the absence of background, the optimal strategy is simply to use the lowest reasonable concentration of protein. However, even in the case of no background, it is important to note that the maximum enrichment is fixed at ***k***, or the ratio of K_{w} to K_{pool}. As selection proceeds, this ratio decreases since K_{pool} approaches K_{w} and enrichment slows.

The effect of protein concentration on enrichment at several different values of ***k*** is shown in Figure 9.1.6b. The bold curve in this plot (and in all plots in these two sets) is for conditions identical to the bold curve in Figure 9.1.6a. The above argument for the effect of protein concentration on enrichment holds here as well. What is startling is the similarity between the effects of ***k*** and the effects of ε on enrichment. It is clear from Equation 9.1.10 that the effects of ***k*** and ε are interchangeable for optimal enrichment. In fact, it can be shown that, in the general expression for enrichment as a function of total protein and nucleic acid concentration, the dimensionless quantities **k** and ε are mathematically interchangeable as well.

At a given total ligand concentration, the total protein concentration that leads to optimal enrichment is closely approximated by:

$$P^{\mathrm{opt}} = (L_t + K_{\mathrm{pool}}^{-1}) / \sqrt{\varepsilon k}$$

Equation 9.1.11

where P^{opt} is the optimal total protein concentration, and L_{t} is the total concentration of ligand (Irvine et al., 1991). As K_{w} increases compared to K_{pool} (increasing ***k***) the optimal protein concentration can be reduced, increasing the so-called stringency of selection. This reduction in protein concentration favors the high-affinity molecules in their competition for binding over lower-affinity species, thereby increasing enrichment. In general, the protein concentration that maximizes enrichment is calculated with respect to either L_{t} or K_{pool}^{-1}. For typical nucleic acid concentrations ~10^{-6} M and high-affinity pools ($>10^{6}$ M^{-1}), the total ligand concentration L_{t} dominates the choice of protein concentration, whereas for low-affinity pools ($<10^{6}$ M^{-1}) the affinity of the pool K_{pool}^{-1} sets the concentration range for total protein. For targets with little measurable binding to nucleic acid pools, high protein concentrations should be

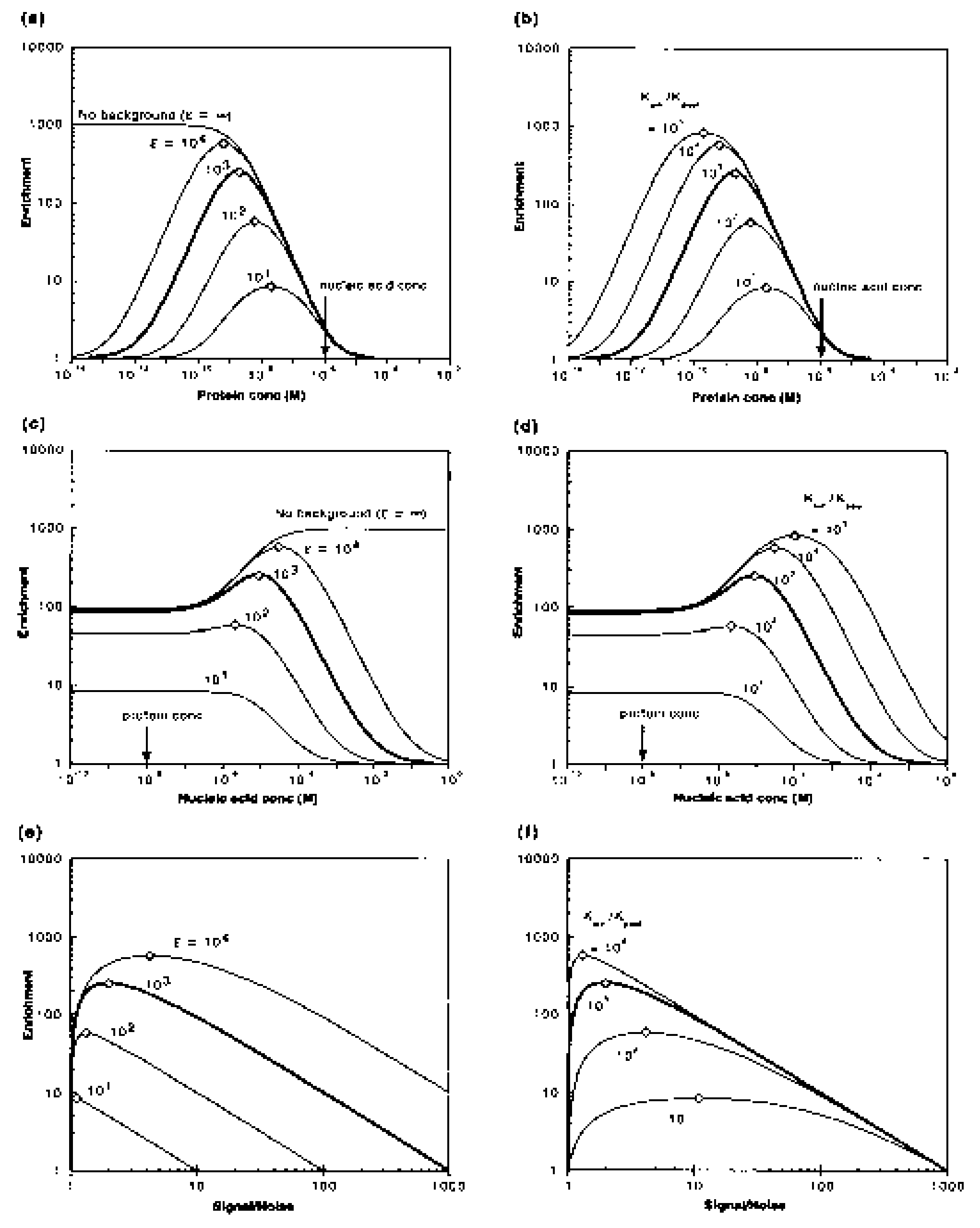


Figure 9.1.6 Enrichment as a function of key parameters for SELEX. The effects of protein concentration on enrichment are displayed for various values of the partitioning effectiveness ε (**a**) and the affinity ratio ***k*** (**b**) with fixed nucleic acid concentration. Similar plots for enrichment as a function of nucleic acid concentration are displayed in (**c**) and (**d**) and as a function of signal-to-noise ratio in (**e**) and (**f**). The optimal enrichments for each curve calculated with the approximations discussed in the text are denoted by the open diamonds. The bold solid line in each plot represents identical selection conditions ($P_t = 10^{-8}$ M, $L_t = 10^{-6}$ M, $K_{pool} = 10^6$ M^{-1}, $K_w = 10^9$ M^{-1}, *bg* = 0.1%, *eff* = 1), aside from the independently varied parameter for each plot.

employed; in extreme cases, protein can even be in excess with respect to nucleic acid!

The open symbols in Figures 9.1.6a and 9.1.6b represent the optimal enrichment and protein concentrations, calculated from Equation 9.1.10 and Equation 9.1.11, respectively. Even though Equation 9.1.11 is approximate, Figure 9.1.6 illustrates that it is clearly quite good. It is important to note that at high protein concentrations, the enrichment is reduced to one (i.e., no enrichment); all nucleic acid is bound by the protein whose concentration is in excess of even the highest dissociation constants (lowest affinity binders) and no enrichment occurs. At low protein concentration, the background overwhelms the correctly partitioned species, and again no enrichment occurs. In the absence of background, as noted above, the enrichment plateau at $E = \boldsymbol{k}$ is observed for low protein concentration.

The effect of nucleic acid concentration on enrichment at several different values of ε and $\boldsymbol{k}$ are shown in Figure 9.1.6c and d. As in Figures 9.1.6a and b, the effects of $\boldsymbol{k}$ and ε on enrichment are identical. For conditions in which nucleic acid is in excess, enrichment is dominated by the competition of high-affinity binders to the limited protein molecules; the same behavior is observed here for fixed protein concentration as for fixed nucleic acid previously discussed, although excess ligand appears to the right of center in Figures 9.1.6c and d, whereas it is to the left of center in Figures 9.1.6a and b. As seen in the fixed nucleic acid case, for zero background the limiting enrichment approaches $\boldsymbol{k}$, as nucleic acid concentration far exceeds that of protein. As total nucleic acid concentration decreases and protein is in excess, enrichment is qualitatively different than that observed above for fixed nucleic acid concentration. Here, enrichment is seen to plateau at protein excess, the plateau value depending on the fixed protein concentration. The enrichment plateau is governed by Equation 9.1.12:

$$E(L_t \rightarrow 0) = \frac{(1 + \varepsilon \boldsymbol{k} p)(1 + p)}{(1 + \varepsilon p)(1 + \boldsymbol{k} p)}$$

Equation 9.1.12

where $p = P_t K_{\text{pool}}$. At concentrations where Equation 9.1.12 is valid, protein is in excess over nucleic acid and there is no competition for binding among the ligands. Selection is driven purely by differences in affinity. It is important to note that, even under these conditions, enrichment occurs; these may be conditions necessary for early rounds of SELEX against targets with low overall pool affinity to nucleic acids and in situations of high background.

The optimal total nucleic acid concentration, L^{opt}, at a given total protein concentration is closely approximated by:

$$L^{\text{opt}} = P_t \sqrt{\varepsilon \boldsymbol{k}} - K_{\text{pool}}^{-1}$$

Equation 9.1.13

Equation 9.1.13 has an interesting symmetry with respect to Equation 9.1.11 for optimal protein concentration. For a given P_t, only one concentration L_t exists which maximizes enrichment as given by Equation 9.1.10. Similarly, that concentration L_t determines the same corresponding concentration P_t in order to maximize enrichment. It is important to note that the expression for optimal enrichment contains no dependence on either P_t or L_t, and yet there exist pairs of P^{opt} and L^{opt} that achieve this global enrichment maximum. In fact, for any given L_t, there always exists a P_t that allows for global enrichment. The converse, however, is not true. When $\boldsymbol{k}$ or ε are small, L^{opt} becomes negative for certain P_t values, indicating that there is no local optimal, as seen in Figures 9.1.6c and d, for $\boldsymbol{k}$ or ε = 10; maximal global enrichment can never be achieved for those values of P_t.

The Signal-to-Noise Ratio

A quantity that is routinely (and easily) measured at the bench is the signal-to-noise ratio, S, which is defined as the amount of oligonucleotide recovered during partitioning in the presence of protein, divided by the amount recovered in the absence of protein, which is due to background. The effect of signal-to-noise on enrichment at several different values of ε and $\boldsymbol{k}$ is shown in Figure 9.1.6, panels e and f. The signal-to-noise ratio that gives optimal enrichment is closely approximated by:

$$S^{\text{opt}} = 1 + \sqrt{\frac{\varepsilon}{k}}$$

Equation 9.1.14

which has the pleasing characteristic of being independent of either protein or ligand concentration. Indeed, S promises to be the reduced variable of choice for guiding SELEX experiments, as the enrichment E may be expressed as a function of S independent of protein or ligand concentration:

$$E = \frac{\varepsilon k(S - 1) + (\varepsilon - S)}{Sk(S - 1) + \boldsymbol{S}(\varepsilon - S)}$$

Equation 9.1.15

It is obvious from inspection of Equation 9.1.15 that the enrichment is sensibly reduced to 1, either when $S = 1$ (representing no partitioned complex), or when $S = \varepsilon$ (when all nucleic acid is partitioned). Equation 9.1.14 may be derived by setting the derivative of E with respect to S in Equation 9.1.15 equal to zero and solving for S, and assuming that ε, $\boldsymbol{k} >> 1$. As suggested by Equation 9.1.14 and

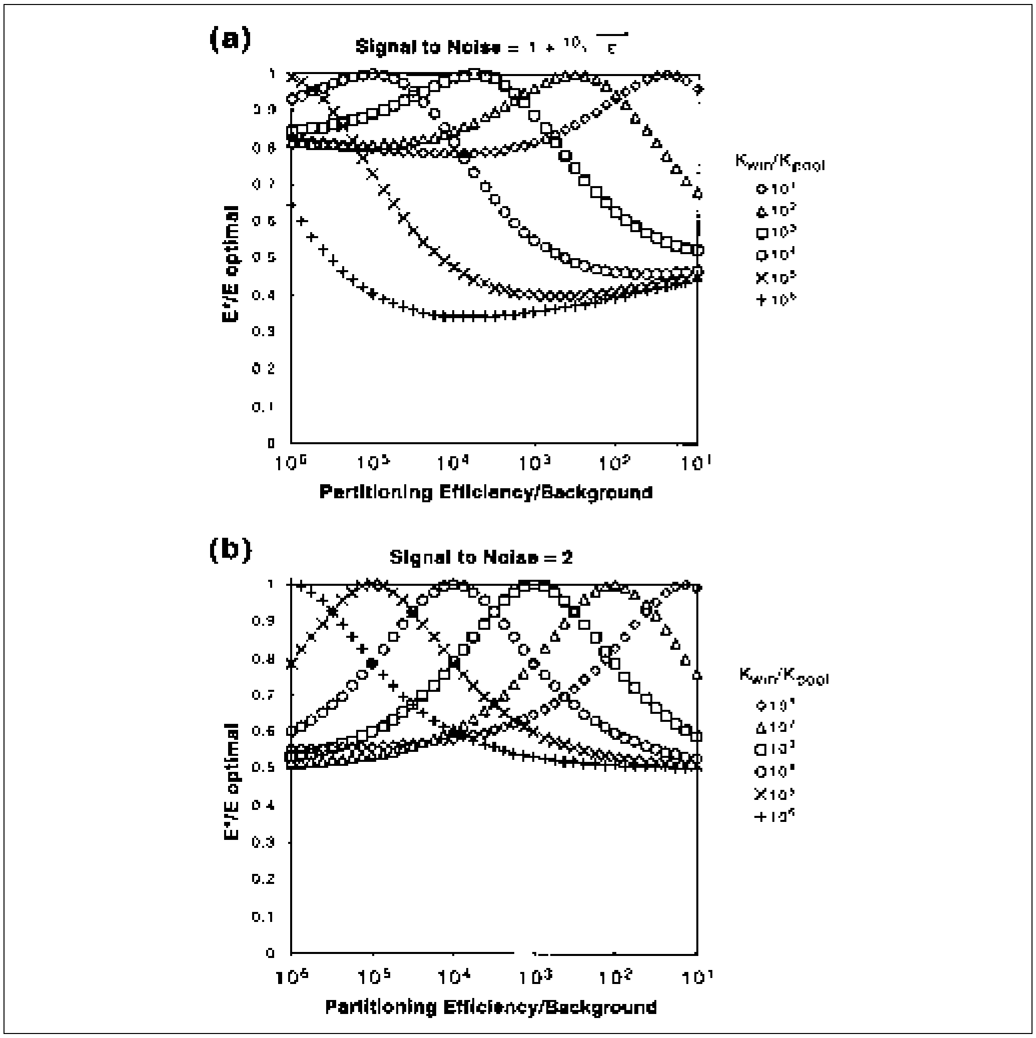


Figure 9.1.7 Comparison of optimal enrichment to signal-to-noise guidelines. The enrichments obtained using conditions that give signal-to-noise specified by Equation 9.1.16 are compared to the optimal enrichments as a function of the partitioning effectiveness ε for various values of the affinity ratio ***k*** in (**a**) while those obtained using conditions that fix ***S*** at two are displayed in (**b**). Note that enrichment is usually better than 50% of the optimal for these two approximations.

Equation 9.1.15, and by Figures 9.1.6, panels e and f, the effects of ***k*** and ε on enrichment as a function of signal-to-noise are not identical, as contrasted with protein and nucleic acid concentrations discussed above. Decreasing the background by a certain factor effectively shifts the enrichment curve up by this factor for large values of *S*. Decreasing ***k*** (decreasing the difference between K_{pool} and K_{w}) causes the enrichment curve to asymptotically approach a straight line (on a log/log plot) with intercepts on both the enrichment and signal-to-noise axes equal to ε. This suggests a simple geometric argument for choosing a near optimal signal-to-noise. A signal-to-noise ratio may be picked which is one-tenth of the way from the *y* axis (signal-to-noise = 1) to the *x* intercept (signal-to-noise = ε). To this, 1 is added to avoid the sudden drop-off in enrichment near the *y* axis. Formally, this is expressed as:

$$S^* = 1 + \sqrt[10]{\varepsilon}$$

Equation 9.1.16

which has a form somewhat similar to Equation 9.1.14. A key feature of Equation 9.1.16 is that it is independent of ***k***; no information about the winning affinity is required to find the optimal signal-to-noise. This information, however, is implicitly accounted for in the determination of *S*. A plot of the enrichment achieved using Equation 9.1.16, divided by the maximum enrichment as a function of ε and ***k***, is shown in Figure 9.1.7a. Even in the worst-case scenario, an enrichment that is at least one third of the maximum possible enrichment is achieved by picking conditions based only on easily measured quantities (signal-to-noise and background). In later rounds of SELEX, as ***k*** decreases (i.e., as the affinity of the pool approaches that of the winner), this strategy promises an enrichment at least 80% of the maximum.

Yet the signal-to-noise ratio allows an even simpler strategy. Setting the signal-to-noise ratio to 2 guarantees an enrichment that is at least 50% of the maximum value (Fig. 9.1.7b). Together, these plots suggest the following strategy. Initially, conditions are used that yield S close to 2 in the early rounds, and, as movement is seen in the bulk affinity of the pool, S is increased to values determined by Equation 9.1.16. This strategy can be realized by evaluating S for different ratios of P_t and L_t at each round of SELEX and carrying forward those conditions that closely match the desired value of S.

The relative insensitivity of enrichment to the signal-to-noise ratio at S >2 is this parameter's most exciting property. There is no effect of protein or nucleic acid concentration on these plots. Although one or both of these concentrations would have to be varied in order to obtain the desired signal-to-noise, it does not matter how this is done (a higher signal-to-noise could be obtained by lowering nucleic acid concentration or raising protein concentration, or a combination of both), and it is not necessary to know what these concentrations actually are. The range of signal-to-noise ratios that give near-optimal enrichment is also relatively insensitive to either background or the ratio of the winner affinity to the pool affinity. In nearly all cases, enrichment close to the optimal may be achieved by using signal-to-noise ratios between 2 and 4. Amazingly enough, the signal-to-noise ratio provides a way to optimize SELEX even while thrashing around in the dark, so to speak. That is, based on the signal-to-noise ratio, an experimenter may select conditions that lead to enrichment of the highest-affinity ligand that is within a factor of two of the best enrichment achievable, and this is possible without the benefit of any affinity data whatsoever!

To illustrate this point, we present a comparison of three simulated SELEX experiments. In each simulation, we begin with the same initial distribution chosen to be a log-normal affinity profile. For each profile, the SELEX experiments are performed as follows. The first SELEX simulation follows a set of conditions that optimize pool enrichment at each round, while the second two simulations utilize the signal-to-noise prescriptions outlined above. For the first signal-to-noise simulation, conditions are selected to yield S of 2, while the second signal-to-noise simulation employs Equation 9.1.16 to select appropriate conditions at each round. Results from these simulations are displayed in Figure 9.1.8. The signal-to-noise procedure performs remarkably well, tracking the optimal-condition SELEX

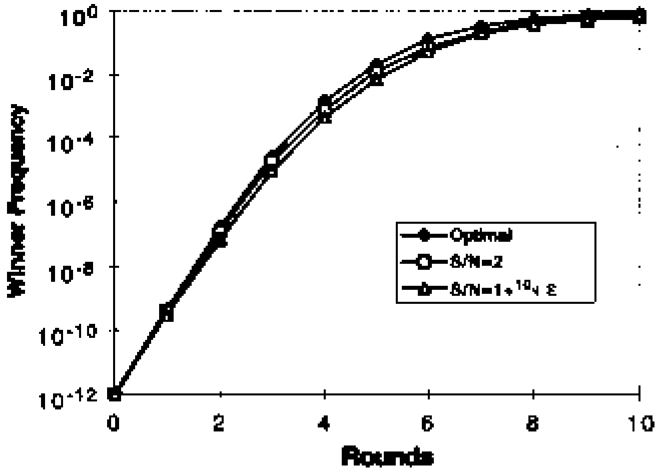


Figure 9.1.8 Progress of SELEX by different guidelines. The progress for rounds of SELEX performed under optimal conditions (filled diamonds) is compared to two signal-to-noise ***S*** guidelines, choosing conditions with fixed ***S*** of 2 (open squares) and choosing conditions that yield ***S*** given by Equation 9.1.16 (open diamonds). Both ***S*** guidelines yield progress remarkably close to optimal.

throughout the simulation. Even though we expect the signal-to-noise optimal to average only 75% maximal enrichment, the results in Figure 9.1.8 demonstrate significantly better performance. This is presumably due to the fact that enrichment that is less than optimal in one round leads to a potentially greater enrichment maximum in the next round compared to the corresponding round of fully optimized SELEX. Even though the near-optimal falls behind, it is able to keep up with the optimal conditions. We are currently working to further validate these simulation results with experimental verification for a number of protein targets.

As noted above, it is easy to experimentally evaluate the signal-to-noise for a particular set of conditions for a round of SELEX. Since we have shown that signal-to-noise can be used to adjust conditions to achieve near-maximal global enrichment, guiding automated SELEX is an obvious application of signal-to-noise theory; indeed, this was the motivation for exploring enrichment as a function of signal-to-noise. An automated system might easily be programmed to evaluate the signal-to-noise ratio and adjust protein or nucleic acid concentrations (or both) appropriately to achieve near-optimal enrichment. As all the other manipulations in SELEX follow standard protocols, the addition of a protocol for choosing selection conditions should allow for efficient automation of the entire SELEX process with performance close to the theoretical optimal.

CONCLUDING REMARKS

We have presented a powerful mathematical model for equilibrium in vitro selection experiments that allows for the identification of conditions leading to optimal enrichment for each round of the iterative SELEX procedure. We have cast the astronomical problem of summing over distinct sequences into a tractable integration over affinity distributions. Although a detailed knowledge of such distributions is currently unavailable, we have demonstrated the utility of our approach by both comparing simulations to experiment and by deriving an optimization guide that is independent of the details of the affinity distributions. We have demonstrated that optimization of SELEX conditions may be easily achieved by monitoring experimental signal-to-noise, a readily measured quantity. This approach promises to be quite powerful for choosing near-optimal conditions for equilibrium in vitro selection experiments.

Recent focus on the automation of SELEX has led to the development of microtiter plate formats for affinity partitioning. In addition to being a convenient format for automation, plate-SELEX offers the possibility of adding a kinetic-selection step to the process, less encumbered by rebinding events as compared to column formats. Ligands could therefore be subjected to selection pressures for kinetic characteristics, such as long off-rates, in addition to the usual high-affinity selection pressures. This is easily achieved through washing after high-affinity ligands have been captured by immobilized targets on the plates. The effects of such kinetic pressures can easily be included in the mathematical model for SELEX; dramatic enrichment for such kinetic characteristics is theoretically possible (Levitan, 1998). There is some experimental evidence to support this theory and more experiments are currently underway.

At this point, it is worth emphasizing that the above theory is predicated on the assumption that the background consists of nonspecific (not affinity-related) partitioning of the input oligonucleotides. Although this is a reasonable assumption, there are at least two other possible sources of background. Background could include outside contamination or artifactual, nonamplifiable signal (such as unbound radioactive label). In these cases, the actual background would be lower than what is measured, and the above strategies would have to be adjusted accordingly. Background could also, however, consist of sequences selected for their binding affinity to targets other than the target of interest. For example, sequences that bind to nitrocellulose, or sequences that bind to contaminants in the target preparation, may be partitioned along with specifically bound sequences. In this case, the behavior of the SELEX experiment is better described by a generalization of the SELEX theory to encompass multiple targets (Vant-Hull et al., 1998). Such a theory suggests that high-affinity ligands will evolve independently to each of the various targets in proportion to the concentrations of the respective targets. The SELEX experiment is still likely to be successful, however, as long as: (1) the concentration of the undesired target is not so great that ligands to the desired target cannot be found, (2) the partitioning of the undesired ligands does not preclude partitioning of the desired ligands (as may be the case when binders to the partitioning matrix are being evolved), or (3) the partitioning efficiency for the desired target is not low compared to the undesired targets. In this final case, ligands to the desired target may only be selected after a great many rounds (however, low partitioning efficiency has little effect in the case of a single target). In any case, the best overall strategy for in vitro selection is to make background as low as possible.

ACKNOWLEDGMENTS

We would like to thank our many colleagues at NeXstar Pharmaceuticals and the University of Colorado for the intellectual stimulus leading to this work. In particular, we wish to thank Doug Irvine for the initial development of the theory described here and for useful comments on this work, and Ed Brody for his careful reading of the manuscript. Finally, we acknowledge Matt Wecker-Wintersquash for timely supplies of Girl Scout cookies during the writing of these acknowledgments.

LITERATURE CITED

Berg, O.G. and von Hippel, P.H. 1987. Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* 193:723-750.

Blackwell, T.K. and Weintraub, H. 1990. Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. *Science* 250:1104-1110.

Breaker, R.R. 1997. In vitro selection of catalytic polynucleotides. *Chem. Rev.* 97:371-390.

Ciesiolka, J., Illangasekare, M., Majerfeld, I., Nickles, T., Welch, M., Yarus, M., and Zinnen, S. 1996. Affinity selection-amplification from randomized ribooligonucleotide pools. *Methods Enzymol.* 267:315-335.

Cwirla, S.E., Peters, E.A., Barrett, R.W., and 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.

Ellington, A.D. and Szostak, J.W. 1990. In vitro selection of RNA molecules that bind specific ligands. *Nature* 346:818-822.

Gold, L. 1995. Oligonucleotides as research, diagnostic, and therapeutic agents. *J. Biol. Chem.* 270:13581-13584.

Gold, L., Polisky, B., Uhlenbeck, O., and Yarus, M. 1995. Diversity of oligonucleotide functions. *Annu. Rev. Biochem.* 64:763-797.

Hager, A.J., Pollard, J.D., Jr., and Szostak, J.W. 1996. Ribozymes: Aiming at RNA replication and protein synthesis. *Chem. Biol.* 3:717-725.

Irvine, D., Tuerk, C., and Gold, L. 1991. SELEXION: Systematic evolution of ligands by exponential enrichment with integrated optimization by nonlinear analysis. *J. Mol. Biol.* 222:739-761.

Kauffman, S.A. and Macready, W.G. 1995. Search strategies for applied molecular evolution. *J. Theor. Biol.* 173:427-440.

Kay, B.K. 1994. Biologically displayed random peptides as reagents in mapping protein-protein interactions. *Persp. Drug Discovery Design* 2:251-268.

Klug, S.J. and Famulok, M. 1994. All you wanted to know about SELEX. *Mol. Biol. Rep.* 20:97-107.

Levitan, B. 1998. Stochastic modeling and optimization of phage display. *J. Mol. Biol*. 277:893-916.

Mathieu-Daudé, F., Welsh, J., Vogt, T., and McClelland, M. 1996. DNA rehybridization during PCR: the 'C_0t effect' and its consequences. *Nucl. Acids Res.* 24:2080-2086.

Sabeti, P.C., Unrau, P.J., and Bartel, D.P. 1997. Accessing rare activities from random RNA sequences: The importance of the length of molecules in the starting pool. *Chem. Biol.* 4:767-774.

Schneider, D., Gold, L., and Platt, T. 1993. Selective enrichment of RNA species for tight binding to *Escherichia coli* rho factor. *FASEB J.* 7:201-207.

Schuster, P. 1995. How to search for RNA structures. Theoretical concepts in evolutionary biotechnology. *J. Biotechnol.* 41:239-257.

Scott, J.K. and Smith, G.P. 1990. Searching for peptide ligands with an epitope library. *Science* 249:386-390.

Stormo, G.D. and Yoshioka, M. 1991. Specificity of the mnt protein determined by binding to randomized operators. *Proc. Natl. Acad. Sci. U.S.A.* 88:5699-5703.

Sun, F., Galas, D., and Waterman, M.S. 1996. A mathematical analysis of in vitro molecular selection-amplification. *J. Mol. Biol.* 258:650-660.

Tuerk, C. and Gold, L. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249:505-510.

Vant-Hull, B., Payano-Baez, A.R., Davis, R.H., and Gold, L. 1998. The mathematics of SELEX against complex targets. *J. Mol. Biol.*278:579-597.

Winter, G., Griffiths, A.D., Hawkins, R.E., and Hoogenboom, H.R. 1994. Making antibodies by phage display technology. *Annu. Rev. Immunol.* 12:433-455.

Contributed by Barry Vant-Hull, Larry Gold, and Dominic A. Zichi
NeXstar Pharmaceuticals
Boulder, Colorado