J. Mol. Biol. (1991) 222, 739–761

SELEXION

Systematic Evolution of Ligands by Exponential Enrichment with Integrated Optimization by Non-linear Analysis

Doug Irvine†, Craig Tuerk and Larry Gold

Department of Molecular, Cellular and Developmental Biology
University of Colorado
Boulder, CO 80309–0347, U.S.A.

(Received 13 March 1991; accepted 23 July 1991)

Recently, novel technologies for isolation of nucleic acid molecules with specific biological activities have been reported. In each case, the enrichment process involves repeated rounds of selection from complex mixtures of nucleic acid sequences, followed by polymerase chain reaction (PCR) amplification of ligand sequences that function in the desired manner. Particular variations in experimental conditions can dramatically alter the outcome of these processes. In this study, we use mathematical analysis and computer simulation to predict which variations have the greatest impact and to develop strategies and guidelines for enhanced effectiveness. First, we perform reconstruction tests to demonstrate that a mathematical description based on equilibrium binding is sufficient to explain the high levels of enrichment attained in the laboratory after just a few rounds. Then, we show the expected enrichment for an extensive range of conditions; and, finally, we determine the optimum protein and nucleic acid concentrations to use for maximum enrichment, while also ensuring a high likelihood of recovering even the rare molecule that binds well. The strategies and guidelines for enhanced effectiveness are generally applicable to processes for systematic enrichment of DNA, RNA or peptide ligands and have been implemented in an interactive simulation program for integrated non-linear optimization of enrichment using any target of interest.

Keywords: nucleic acid binding; high-affinity ligands; selective enrichment; sequence evolution; optimized SELEX

1. Introduction

Authors from several laboratories (Abelson, 1990; Biedenkapp *et al.*, 1988; Blackwell *et al.*, 1990; Blackwell & Weintraub, 1990; Ellington & Szostak, 1990; Green *et al.*, 1990; Joyce, 1989*a*,*b*; Joyce & Inoue, 1989; Kinzler & Vogelstein, 1989, 1990; Mavrothalassitis *et al.*, 1990; North, 1990; Oliphant & Struhl, 1987, 1988; Oliphant *et al.*, 1989; Pollock & Treisman, 1990; Robertson & Joyce, 1990; Sompayrac & Danna, 1990; Theisen & Bach, 1990; Tuerk & Gold, 1990) have reported systematic methodologies for isolation of nucleic acid sequences with specific biological activities. Many of these processes utilize the capacity of DNA or RNA molecules for binding a target molecule, followed by amplification with polymerase chain reaction (PCR‡; Innis *et al.*, 1988) to enrich for sequences in a pool that bind well to the target. Tuerk & Gold (1990) call their implementation SELEX, for *Systematic Evolution of Ligands by EXponential enrichment*. Systematic methodologies also have been reported for isolation of peptide sequences with specific biological activities (Cwirla *et al.*, 1990; Devlin *et al.*, 1990; Scott & Smith, 1990; Tuerk & Gold, 1990). The following description of SELEX uses, as the illustrative example, binding of RNA ligand sequences to a purified protein with partitioning of bound and unbound RNA molecules by nitrocellulose filtration.

After mixing a protein with a pool of RNA molecules in the appropriate buffer solution, protein-bound RNA molecules are separated from unbound

† Present address: Department of Microbiology and Immunology, The University of Michigan, Ann Arbor, MI 48109–0620, U.S.A.

‡ Abbreviations used: PCR, polymerase chain reaction; gp43, DNA polymerase in phage T4.

RNA molecules by passing the solution through a nitrocellulose filter. Most protein sticks to the filter, including protein–RNA complexes, while most unbound RNA molecules wash through (Uhlenbeck *et al.*, 1983; Yarus, 1976; Yarus & Berg, 1967, 1970). RNA molecules collected on the filter are then eluted, and cDNA copies are amplified by PCR, followed by *in vitro* transcription to restore the RNA pool to a concentration high enough for the next round. These selection and amplification steps are repeated until the desired level of enrichment is attained for RNA sequences that bind well to the protein.

Any method of partitioning RNA, DNA or peptide sequences using any target can be utilized in an enrichment process like SELEX. In general, the three main steps involved in each round are: (1) selection of ligand sequences that bind to a target; (2) partitioning of bound and unbound ligand sequences; and (3) PCR amplification of ligand sequences in the desired fraction. When RNA ligands are enriched using SELEX, they are transcribed from PCR-amplified cDNA; or when peptide ligands are enriched using SPERT (*Systematic Peptide Evolution by Reverse Translation*; Tuerk & Gold, 1990), they are translated after the associated mRNA molecules are transcribed from PCR-amplified cDNA. The general steps involved in enrichment processes like SELEX are illustrated in Figure 1.

Particular variations in experimental conditions can dramatically alter the outcome of SELEX experiments. Since each round of SELEX with available methods is labor-intensive, and a typical SELEX experiment requires several rounds to complete, employing the current laboratory technology to test an extensive range of conditions would require an excessive amount of time and effort. Alternatively, by first applying mathematical analysis and computer simulation in order to understand better enrichment processes like SELEX, available laboratory results can be used to test critically predictions made from first principles. In this paper, we derive a mathematical description of SELEX, test it against available laboratory results, use it to predict the consequences of extensive variations in experimental conditions, and apply it to derive practical strategies and guidelines for enhanced effectiveness of any enrichment process like SELEX.

The first objective of this paper is to test whether the mechanisms proposed for SELEX are sufficient to explain the high levels of enrichment attained in the laboratory after just a few rounds; in particular, our first goal is to test whether a description based on equilibrium binding is sufficient. We accomplish this by using a computer program to simulate each round of SELEX in reconstruction tests. Then we use further mathematical analysis to predict the levels of enrichment to expect under different conditions: for example, different total RNA and protein concentrations, different strengths of binding, different starting fractions of each RNA species,

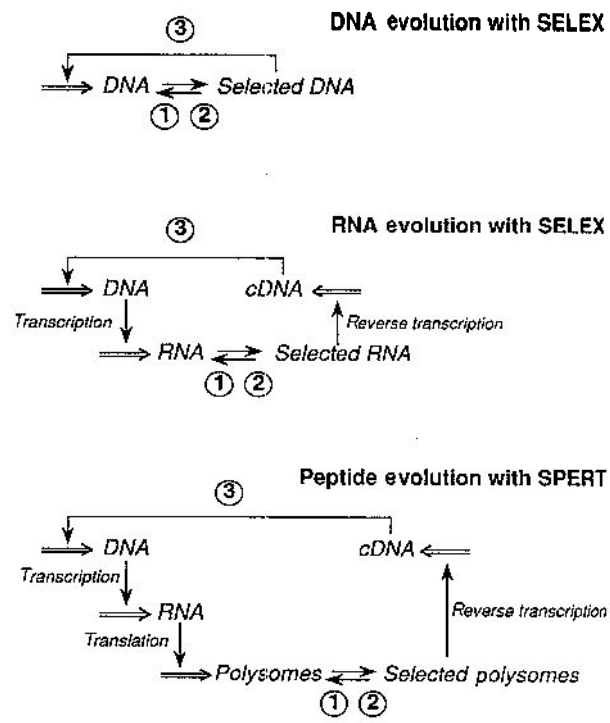


Figure 1. Enrichment of nucleic acid ligand sequences using SELEX (*Systematic Evolution of Ligands by EXponential enrichment*; Tuerk & Gold, 1990) or enrichment of peptide sequences using SPERT (*Systematic Peptide Evolution by Reverse Translation*; Tuerk & Gold, 1990). The 3 general steps involved in each process are: ① Selection of ligand sequences by binding to a target molecule, ② partitioning of bound and unbound ligand sequences, and ③ PCR amplification of ligand sequences in the desired fraction. Reverse transcription of an RNA molecule associated with a selected nascent peptide on a polysome gives the desired effect of reverse translation. Further details are given in the corresponding sections of the text.

different levels of background partitioning, different concentrations of non-specific competitor molecules and different sensitivities of PCR or *in vitro* transcription to the number of RNA or DNA molecules representing each species.

Our analysis demonstrates that with the right proportions of nucleic acids and protein, just a few rounds of SELEX are enough for enrichment of the best-binding ligands to form a predominant fraction of the pool. But when either too little or too much protein is used, the process can fail outright, or the number of rounds necessary for enrichment can become so large that the overall process becomes impractical. Knowledge of the ideal proportions of nucleic acids and protein to use can be applied to promote good enrichment in every round of SELEX. This knowledge will help make the overall enrichment process less time consuming, less expensive and less susceptible to technical errors.

The final objectives of this analysis are to determine the optimum protein and nucleic acid concentrations to use in SELEX for enrichment of the

best-binding nucleic acid molecules in as few rounds as possible, while ensuring a high likelihood of recovering the best-binding species, even when represented by only a single molecule in the first round. With the insight gained from this extensive mathematical and computer-assisted analysis, we are able to give guidelines on how to increase the likelihood of enrichment of RNA molecules containing sequences that bind well with any target of interest. Although the particular example chosen for detailed analysis is enrichment of RNA ligands using SELEX, the results can be generalized to include enrichment of DNA ligands with processes like SELEX, or enrichment of peptide ligands (e.g., see Cwirla *et al.*, 1990; Devlin *et al.*, 1990; Scott & Smith, 1990; Tuerk & Gold, 1990). Hence, the strategies and guidelines developed here can be employed for enrichment of DNA, RNA or peptide sequences selected with almost any target of interest.

2. Selection

A simple kinetic mechanism for reversible protein–RNA complex formation in a well-mixed solution is written as follows:

$$[Pf]+[RNAf_i]\underset{k_{-i}}{\overset{k_{+i}}{\rightleftarrows}}[P:RNA_i] \quad i=1,\ldots n, \qquad (1)$$

where $[Pf]$ is the free protein concentration, $[RNAf_i]$ is the free RNA species i concentration, $[P:RNA_i]$ is the protein–RNA species i complex concentration, k_{+i} is the rate constant for association of free protein and free RNA species i, k_{-i} is the rate constant for dissociation of protein–RNA species i complexes, and n is the number of RNA sequences with a unique set of rate constants. Alternative mechanisms, including multiple binding sites or cooperativity, could be considered in subsequent treatments with appropriate extensions of this simple scheme.

For any system represented by the above scheme, the fundamental chemical-kinetic or mass-action equations describing the change in concentration of each protein–RNA species i complex as a function of time are:

$$\frac{d[P:RNA_i]}{dt}=k_{+i}\cdot[Pf]\cdot[RNAf_i]-k_{-i}\cdot[P:RNA_i], \quad i=1,\ldots n, \qquad (2)$$

where $[Pf]$, $[RNAf_i]$, and $[P:RNA_i]$ are the concentrations of free protein, free RNA species i, and protein–RNA species i complex at time t.

The free protein concentration is the difference between the total protein concentration and the concentration of all protein–RNA complexes ($[P]-\Sigma[P:RNA_k]$); likewise, the free RNA species i concentration is the difference between the total RNA species i concentration and the protein–RNA species i complex concentration ($[RNA_i]-[P:RNA_i]$):

$$\frac{d[P:RNA_i]}{dt}=k_{+i}\cdot\left([P]-\sum_{k=1}^{n}[P:RNA_k]\right)\cdot([RNA_i]-[P:RNA_i])-k_{-i}\cdot[P:RNA_i], \quad i=1,\ldots n, \qquad (3)$$

These dynamic equations can be used for either kinetic or equilibrium analysis. The continuous differential form is valid whenever the mean rate of each process is large relative to the variance in that process or, in other words, equation (3) is accurate for the description of a pool of RNA with several molecules representing each unique set of rate constants. Whenever there is only one molecule, or just a few molecules, of the best-binding RNA present, a statistical description of binding is used to determine the conditions that give a high likelihood of recovering the best-binding RNA. These statistical formulae are derived in a subsequent section on the likelihood of success.

At equilibrium, the change in concentration of each protein–RNA species i complex equals zero:

$$\left([P]-\sum_{k=1}^{n}[P:RNA_k]\right)\cdot([RNA_i]-[P:RNA_i])-Kd_i\cdot[P:RNA_i]=0, \quad i=1,\ldots n, \qquad (4)$$

with symbols as defined in equations (2) and (3), and with Kd_i being the equilibrium dissociation constant for protein–RNA species i complex ($Kd_i = k_{-i}/k_{+i}$). Rearrangement of equation (4) gives the following implicit formula for the equilibrium concentration of each protein–RNA species i complex:

$$[P:RNA_i]=\frac{F_i^0\cdot[RNA]\cdot([P]-[P:RNA])}{Kd_i+([P]-[P:RNA])}=\frac{F_i^0\cdot[RNA]\cdot[Pf]}{Kd_i+[Pf]}, \quad i=1,\ldots n, \qquad (5)$$

with $[P:RNA]$ being the concentration of all protein–RNA complexes ($\Sigma[P:RNA_k]$), and with F_i^0 being the fraction of the RNA pool composed of RNA species i ($[RNA_i]/[RNA]$).

When only one RNA species is considered (i.e. $n=1$), an analytical solution for the equilibrium concentration of protein–RNA complexes is possible by solving the following quadratic equation:

$$[P:RNA_1]^2-([P]+[RNA_1]+Kd_1)\cdot[P:RNA_1]+[P]\cdot[RNA_1]=0, \qquad (6)$$

which has two real roots, one physically realizable:

$$[P:RNA_1]=\frac{2\cdot[P]\cdot[RNA_1]}{([P]+[RNA_1]+Kd_1)+\sqrt{\{([P]+[RNA_1]+Kd_1)^2-4\cdot[P]\cdot[RNA_1]\}}}. \qquad (7)$$

Of course there are numerous classical approximations for equilibrium or quasi-steady-state concentrations of complexes, like that in the Michaelis-Menten formalism, but none give sufficient accuracy over the range of total RNA and protein concentra-

tions used in SELEX. (For revealing discussions of some pitfalls and limitations of classical approximation see Savageau (1991), Straus & Goldstein (1943) and Webb (1963).) Although analytical solution of the quadratic equation for simple reversible association of a single RNA species with a single binding site on the protein is accurate over all RNA and protein concentrations used in SELEX, and although the bound concentrations of two competing species can be calculated by analytical solution of a cubic equation, iterative numerical methods are required to calculate equilibrium concentrations of protein–RNA complexes whenever three or more competing RNA species are considered.

We have developed a computer program to solve numerically for the equilibrium concentration of each protein–RNA species i complex, $[P:RNA_i]$, given any total protein concentration, $[P]$, any distribution of RNA species i concentrations, $[RNA_i]$, and any distribution of equilibrium dissociation constants, Kd_i. The Jacobian matrix for implicit solution of equation (4) by Newton's method (e.g. see Leunberger, 1973; Press *et al.*, 1988) is calculated with the following formula:

$$Fn_i^{eql} = \left([P] - \sum_{k=1}^{n} [P:RNA_k]\right) \cdot ([RNA_i] - [P:RNA_i]) - Kd_i \cdot [P:RNA_i] = 0, \quad (8)$$

$$a_{ij} = \frac{\partial Fn_i^{eql}}{\partial [P:RNA_j]} = -([RNA_i] - [P:RNA_i]) - \delta_{ij} \cdot \left([P] - \sum_{k=1}^{n} [P:RNA_k] + Kd_i\right), \quad i = 1, \ldots n, \; j = 1, \ldots n,$$

where a_{ij} is the element in row i, column j of the Jacobian matrix, with $\delta_{ii} = 1$ and $\delta_{ij} = 0$ for $i \neq j$.

Often the success of Newton's method depends on a good initial estimate for the solution (e.g. see Leunberger, 1973; Press *et al.*, 1988), in this case, the equilibrium concentration of each protein–RNA species i complex, $[P:RNA_i]$. We use equation (5) to estimate these values, after the concentration of all protein–RNA complexes, $[P:RNA]$, is calculated using an estimate for the bulk Kd of the RNA pool:

$$[P:RNA] = \frac{2 \cdot [P] \cdot [RNA]}{([P] + [RNA] + \langle Kd \rangle) + \sqrt{\{([P] + [RNA] + \langle Kd \rangle)^2 - 4 \cdot [P] \cdot [RNA]\}}}, \quad (9)$$

with $[RNA]$ being the concentration of the total RNA pool, and with $\langle Kd \rangle$ being the bulk equilibrium dissociation constant for the pool.

We derive $\langle Kd \rangle$ from the conventional definition for Kd, substituting equation (5) for each occurrence of $[P:RNA_i]$ as follows:

$$\langle Kd \rangle \equiv \frac{[RNAf] \cdot [Pf]}{[P:RNA]}$$

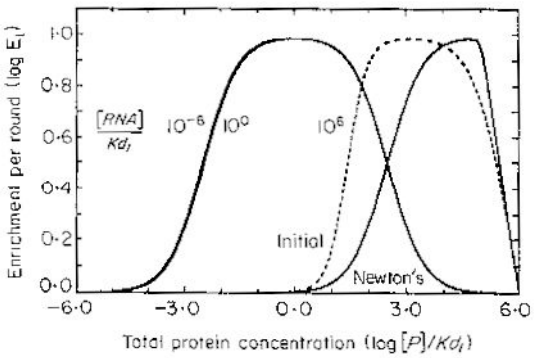
$$= \frac{\left([RNA] - \sum_{i=1}^{n} [P:RNA_i]\right) \cdot [Pf]}{\left(\sum_{i=1}^{n} [P:RNA_i]\right)}$$

$$= \frac{[RNA] \cdot [Pf] - \left(\sum_{i=1}^{n} \frac{F_i^0 \cdot [RNA] \cdot [Pf]}{Kd_i + [Pf]}\right) \cdot [Pf]}{\left(\sum_{i=1}^{n} \frac{F_i^0 \cdot [RNA] \cdot [Pf]}{Kd_i + [Pf]}\right)}, \quad (10)$$

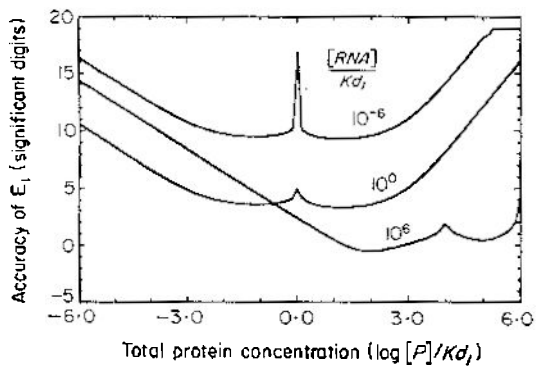
$$= \frac{1}{\left(\sum_{i=1}^{n} \frac{F_i^0}{Kd_i + [Pf]}\right)} - [Pf] \approx \frac{1}{\left(\sum_{i=1}^{n} \frac{F_i^0}{Kd_i}\right)},$$

with $F_i^0 = [RNA_i]/[RNA]$. The last part of equation (10) gives an approximation for $\langle Kd \rangle$ that is independent of $[Pf]$, whenever $[Pf]$ is much less than each Kd_i. Substituting equation (10) into equation (9) gives exact values for $[P:RNA]$ only when the value of $[Pf]$ used in equation (10) happens to equal the actual value, $[P] - [P:RNA]$. For any other estimate of $[Pf]$, the resulting $\langle Kd \rangle$ yields an estimate for $[P:RNA]$. In other words, equation (9) and (10) still are implicit formulae. These formulae could actually be solved numerically for $[P:RNA]$, and the result could be substituted into equation (5) to calculate each $[P:RNA_i]$, instead of numerically solving the implicit formulae in equation (4).

In this implementation, after initial approximations are made using equation (5), (9) and (10), accurate solutions for the complete set of $[P:RNA_i]$ that satisfy equation (4) are obtained by iterative application of Newton's method using equation (8). Solutions with at least 12 significant digits are attained in less than four or five iterations of Newton's method. This rapid convergence to an accurate solution is due to the initial approximations typically giving one or more significant digits at the onset, depending on the range of equilibrium dissociation constants and the abundance of each RNA species. As noted above, one reason for this level of accuracy is that equation (10) gives an accurate estimate for $\langle Kd \rangle$ whenever $[Pf]$ is much less than each Kd_i. The second reason is that errors in $[P:RNA_i]$ tend to cancel in equation (5) whenever $[Pf]$ is much greater than the corresponding Kd_i, for example, when $[RNA]$ is less than Kd_1 and Kd_i is less than $\langle Kd \rangle$. Interestingly, this means that accuracy tends to be high for any protein–RNA species i complex with better binding than the bulk RNA pool. Representative examples of the accuracy of approximations made using Equation (5), (9) and (10) are shown in Figure 2(a) and (b) over a wide range of relevant conditions. The overall accuracy for the enrichment calculation shown (defined as the relative increase in the amount of the best-binding RNA for each round)



(a)



(b)

Figure 2. Enrichment (defined as the increase in the fraction of the RNA pool composed of the best-binding RNA species) calculated using approximations for the concentration of each protein–RNA species i complex *versus* enrichment calculated using accurate solutions obtained by Newton's method. (a) The broken lines are the approximations calculated with eqns (9), (10) and (19), and the continuous lines are the accurate solutions obtained by Newton's method, each plotted *versus* the base-10 logarithm of the total protein concentration, normalized to Kd_1. The number next to each pair of curves represents the total RNA concentration, normalized to Kd_1. For total RNA concentrations less than or equal to Kd_1, differences are too small to be seen graphically (see (b) for calculated differences). Both $[P]/Kd_1$ and $[RNA]/Kd_1$ are varied over 12 orders of magnitude, and other parameters have been chosen (i.e. $Kd_1 = 1$, $\langle Kd \rangle = 1000$, $F_1^0 = 0{\cdot}10$, $BG/CP = 0{\cdot}1\%$ and $[Pf] \approx [P]/2$) to emphasize potential regions of inaccuracy. (b) The number of significant digits (calculated as the negative base-10 logarithm of the maximum relative difference in eqn (19) solved with the approximation for $\langle Kd \rangle$ from eqn (10) or with high-accuracy solutions refined by iterative application of Newton's method using eqn (8)) is plotted *versus* the base-10 logarithm of the total protein concentration, normalized to Kd_1. Parameter values are the same as in (a). Accuracy using eqn (10) is generally higher at lower total RNA concentrations, e.g. 9 to 20 digits accuracy for $[\mathrm{RNA}]/Kd_1 \leqslant 10^{-6}$, 3 to 16 digits for $[\mathrm{RNA}]/Kd_1 \leqslant 1$, 0 digits (i.e. 100% maximum error) to 14 digits accuracy for $[\mathrm{RNA}]/Kd_1 \leqslant 10^6$. Accuracy is greatest at both extremes of total protein concentration. Similar levels of accuracy are attained when several competing species of RNA are treated simultaneously (results not shown).

reflects the accuracy of every protein–RNA species i complex concentration calculated with equation (5), (9) and (10). In a subsequent section, we capitalize on the accuracy of the approximation for $\langle Kd \rangle$ in equation (10) to solve analytically for the optimum total protein concentration that gives maximum enrichment.

3. Partitioning

Any method of partitioning different species of nucleic acid sequences, including filter binding (Tuerk & Gold, 1990), gel-mobility shifts (Blackwell & Weintraub, 1990; Mavrothalassitis *et al.*, 1990), affinity chromatography (Ellington & Szostak, 1990; Oliphant *et al.*, 1989), antibody precipitation, phase partitions, protection from nucleolytic cleavage (Robertson & Joyce, 1990), or splicing activity (Green *et al.*, 1990), can be used to advantage with SELEX. For example, with filter binding most protein–RNA complexes stick to a nitrocellulose filter while most free RNA molecules wash through (Uhlenbeck *et al.*, 1983; Yarus, 1976; Yarus & Berg, 1967, 1970).

Since a fraction of free RNA molecules also partitions as non-specific background, the total amount of each RNA species i collected and then eluted from the filter for PCR amplification is calculated using the following formula, which accounts for both the desired *signal* from the best-binding RNA molecules that partition appropriately in protein–RNA complexes and the *noise* from competing RNA molecules that partition in protein–RNA complexes plus free RNA molecules that partition as non-specific background:

$$RNA_i^{part} = Vol \cdot \{CP \cdot [P : RNA_i] + BG \cdot ([RNA_i] - [P : RNA_i])\} \cdot 6{\cdot}02 \times 10^{23}, \quad (11)$$
$$i = 1, \ldots n,$$

where RNA_i^{part} is the number of molecules of RNA species i that partition for subsequent PCR amplification, Vol is the volume of the reaction mixture passed through the filter, CP is the fraction of protein–RNA complexes that partition appropriately, $[P : RNA_i]$ is the molar equilibrium concentration of protein–RNA species i complex calculated as described in the preceding section, BG is the fraction of free RNA molecules that partition as non-specific background, and $[RNA_i]$ is the total RNA species i concentration.

Any method of partitioning typically gives less than perfect separation of bound and unbound ligands and, hence, the mathematical description of SELEX requires an estimate for background partitioning in each round. With any effective partitioning technique, the fraction of free RNA molecules that partitions as non-specific background is obviously less than the fraction of protein–RNA complexes that partitions appropriately; the smaller the ratio of BG to CP the better. For the best partitioning techniques BG approaches zero, and CP approaches unity.

Tests have shown that with nitrocellulose filtration, background partitioning is essentially constant for any given RNA pool, whether the smallest or largest number of molecules available is passed through the filter. Further tests have shown that even after nitrocellulose filters are pretreated with the maximum amount of non-amplifiable RNA available (with the intention of reducing the fraction of PCR-amplifiable molecules that partitions as non-specific background) background partitioning for any given RNA pool remains essentially unchanged.

As is accounted for by the parameter CP in equation (11), not all protein–RNA complexes in solution may be retained on the filter. Furthermore, RNA in tightly bound complexes may be retained better on the filter than RNA in weakly bound complexes. Whenever this is true, enrichment for RNA molecules that bind tightly would be further enhanced in each round of SELEX. On the other hand, if some molecules were not eluted from the filter as well as others, their enrichment would be reduced.

4. Amplification and Renormalization

The amount of each RNA species i recovered from the filter and copied into cDNA for PCR amplification is calculated using the following formula:

$$cDNA_i^{pcr} = RT \cdot RNA_i^{part}, \qquad i = 1, \ldots n, \quad (12)$$

where RNA_i^{part} is the number of molecules of RNA species i that partitions for PCR, as calculated with equation (11), and RT is the fraction of partitioned RNA molecules that is copied by reverse transcription into cDNA for amplification by PCR. In this treatment, the value of RT is assumed to be constant. It is determined both by the fraction of RNA copied by reverse transcriptase to make cDNA and by the fraction of cDNA molecules that is effectively replicated in each cycle of PCR. The assumption that RT is constant for all species is a reasonable starting point, since, given sufficient time, when all molecules have the same primer sites for PCR and an excess of primer molecules is used, each species (whether rare or abundant) has virtually the same likelihood of annealing with a primer molecule. Also, since each cDNA molecule typically is the same length, there is virtually no differential rate of amplification on the basis of size. Of course, if any RNA species has a secondary structure that interferes with primer annealing for cDNA synthesis, or if the primary or secondary structure of the corresponding cDNA slows the rate of DNA polymerase during PCR amplification, enrichment of that species is reduced. We do not incorporate these effects, since there are no good rules for predicting what structures actually make a difference. When more is learned about these structures, any significant effects can be added to the mathematical description of SELEX.

The total amount of RNA amplified as cDNA by PCR is calculated by summing the number of molecules of each species amplified, as calculated with equation (12):

$$cDNA^{pcr} = \sum_{i=1}^{n} cDNA_i^{pcr}. \quad (13)$$

Affinity measurement protocols often include "carrier" or "non-specific competitor" RNA molecules. Whenever such molecules are used in SELEX, obviously these species should be non-amplifiable and thus are excluded from the total in equation (13). Interestingly, whenever non-specific competitor molecules interact with the protein at the same site as the best-binding ligand molecules, the main consequence of adding competitor molecules is a reduction in the number of specific sites available for selection. Hence, to determine the protein concentration that binds the desired amount of amplifiable ligand molecules with a high concentration of non-specific competitor molecules present, corrected binding curves must be generated by including the appropriate concentration of these molecules in each titration.

The advantages of using a high concentration of non-specific, non-amplifiable competitor molecules in each round of SELEX can include a reduction in adsorption of amplifiable ligand molecules to any non-specific sites on labware, a reduction in binding of amplifiable ligand molecules to any non-specific sites on the target protein, or a reduction in the fraction of free amplifiable molecules collected as non-specific background on "false-partitioning" sites, but only when such sites are effectively saturated by the amount of non-specific competitor molecules used. If these conditions are not met, the effect of adding non-specific competitor molecules is essentially the same as reducing the amount of protein used.

The amount of each amplifiable cDNA species i recovered after one round, relative to the total in equation (13), is calculated as follows:

$$F_i^1 = \frac{cDNA_i^{pcr}}{cDNA^{pcr}}, \qquad i = 1, \ldots n. \quad (14)$$

After renormalization of the RNA pool to its original concentration by *in vitro* transcription (from promoter sites with the same primary sequence in every correctly amplified cDNA molecule) the concentration of each RNA species after one round of SELEX is:

$$[RNA_i] = F_i^1 \cdot [RNA], \qquad i = 1, \ldots n, \quad (15)$$

where $[RNA]$ is the total concentration of the RNA pool after transcription. For each additional round of SELEX, the concentration of every RNA species can be computed by reiteration of equation (4) to (15), with F_i^1 for each RNA species from one round being the starting fraction F_i^0 in the next (see eqns (5) and (10)).

This completes the mathematical descriptions of the mechanisms proposed for SELEX. We utilize

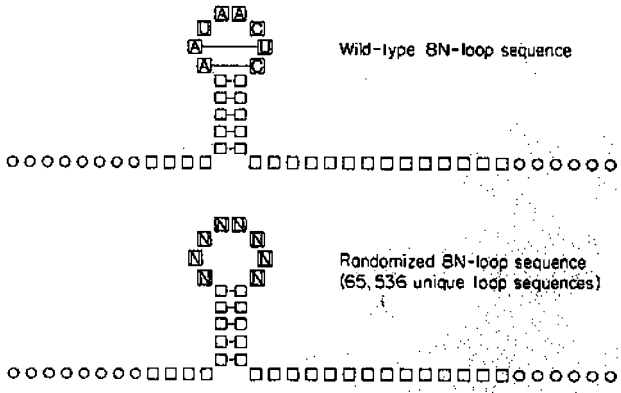


Figure 3. Predicted structure of the gene *43* translational operator in bacteriophage T4 (Tuerk *et al.*, 1990; Tuerk & Gold, 1990). The extent of the ribosome binding site on the messenger RNA for gene *43* is shown. Boxes indicate nucleotides protected by binding with gp43, and capital letters represent the wild-type loop sequence. Possible extension of the stem by 2 additional base-pairs in the wild-type loop sequence is represented as the dotted lines connecting the potential A·C and A·U pairs (see Tuerk & Gold, 1990). A pool of RNA molecules with each of 4 nucleotides substituted at every position in the 8-base loop was utilized for the experiments described in the text.

these descriptions in the next section to perform reconstruction tests.

5. Reconstruction Tests

The messenger RNA encoding the DNA polymerase in bacteriophage T4 (gp43) has an operator region for autogenous regulation of translation. gp43 binds to the operator region, occludes the ribosome binding site on the message, and blocks translational initiation by 30 S ribosomal subunits (Andrake *et al.*, 1988). The equilibrium dissociation constant for gp43–operator complexes is $4{\cdot}8 \times 10^{-9}$ M (Tuerk & Gold, 1990). The structural determinants of the operator required for recognition by gp43 include both the primary sequence (e.g., contacts with specific nucleotides in specific positions) and the secondary structure (e.g. stem–loop structure). The predicted structure of the RNA operator (Tuerk *et al.*, 1990; Tuerk & Gold, 1990) is given in Figure 3.

To determine precisely what the sequence of eight nucleotides in the loop of the RNA operator region contributes to gp43 binding, Tuerk & Gold (1990) synthesized a pool of RNA molecules with these eight nucleotides randomized. The rest of the operator region was left unchanged. Incorporation

Table 1

A. *Input parameters for simulations with the program SELEXION (our unpublished results)*

Input ligand pool	
Random nucleotide positions, N	
Total number of ligand molecules sampled, RNA	
Calculated saturation of sequence space, $RNA/4^N$	
Kd of bulk ligand pool, $\langle Kd \rangle$	
Fraction of bound ligands partitioned for PCR, CP	
Fraction of free ligands partitioned as background, BG	
Binding conditions for each round	
Round number, rnd	
Total reaction volume, Vol	
Total ligand concentration, $[RNA]$	
Total target concentration, $[P]$	
Estimated ligand properties	
Nucleic acid-information content in best ligand, I	
Kd of best-binding ligand, Kd_1	
Distribution of ligands by unique Kd values, n	

B. *Parameters for reconstruction test of 8N-loop experiment B (Tuerk & Gold, 1990)*

Input ligand pool	
Random nucleotide positions, N	= 8 random nucleotides
Total number of ligand molecules sampled, RNA	= $1{\cdot}8 \times 10^{15}$ molecules
Calculated saturation of sequence space, $RNA/4^N$	= $2{\cdot}7 \times 10^{10}$ molecules/sequence
Kd of bulk ligand pool, $\langle Kd \rangle$	= $3{\cdot}2 \times 10^{-7}$ M
Fraction of bound ligands partitioned for PCR, CP	= 80%
Fraction of free ligands partitioned as background, BG	= 0·1%
Binding conditions for each round	
Round number, rnd	= 1 through 8
Total reaction volume, Vol	= 10^{-4} l
Total ligand concentration, $[RNA]$	= 3×10^{-5} M
Total target concentration, $[P]$	= 3×10^{-8} M
Estimated ligand properties	
Nucleic acid information content in best ligand, I	= 15 bits
Kd of best-binding ligand, Kd_1	= $4{\cdot}8 \times 10^{-9}$ M
Distribution of ligands by unique Kd values, n	= 5 unique Kd values

of four different nucleotides in each of eight positions would generate 4^8 or 65,536 unique sequences of RNA molecules in the pool. The bulk *Kd* for the mixed pool is $3{\cdot}2 \times 10^{-7}$ M, or 67-fold weaker binding than the wild-type operator. The RNA pool with the eight-loop nucleotides randomized in the RNA operator has been used in a series of SELEX experiments (Tuerk & Gold, 1990) referred to here as the 8N-loop experiments. In the most extensively characterized 8N-loop experiment, the total RNA and protein concentrations were approximately 3×10^{-5} and 3×10^{-8} M in each round. We use these RNA and protein concentrations in the reconstruction tests described below.

The mathematical techniques derived in the preceding sections have been implemented in a computer program called SELEXION, for *Systematic Evolution of Ligands by EXponential enrichment with Integrated Optimization by Non-linear analysis* (our unpublished results). Table 1A shows the input parameters required for simulations with this program, and Table 1B shows the settings used for a reconstruction test of the most extensively characterized 8N-loop experiment (experiment B). Figure 4 shows the results predicted by the program for several rounds of SELEX. Each input parameter and the results are explained below.

(a) *Input RNA pool*

In the 8N-loop experiments, the possible number of unique sequences is 4^8. With 10^{-4} l of 3×10^{-5} M-RNA in the first round, the total number of RNA molecules sampled is $1{\cdot}8 \times 10^{15}$. This number of molecules gives a saturation of sequence space, which we define as the mean number of molecules of each unique sequence, of $2{\cdot}7 \times 10^{10}$ molecules per unique sequence. The *Kd* for the bulk input RNA pool was measured as $3{\cdot}2 \times 10^{-7}$ M. The fraction of RNA molecules in protein–RNA complexes that partition for subsequent PCR amplification was measured as 80%, and the fraction of free RNA molecules that partition as non-specific background was measured as 0·1%.

(b) *Binding conditions*

Although different total reaction volumes, RNA concentrations or protein concentrations can be used in each round of SELEX, the binding conditions in each round of 8N-loop experiment B were all the same. The total reaction volume was 10^{-4} l, the total RNA concentration was approximately 3×10^{-5} M, and the total protein concentration was approximately 3×10^{-8} M. These are the settings for each round of the reconstruction test.

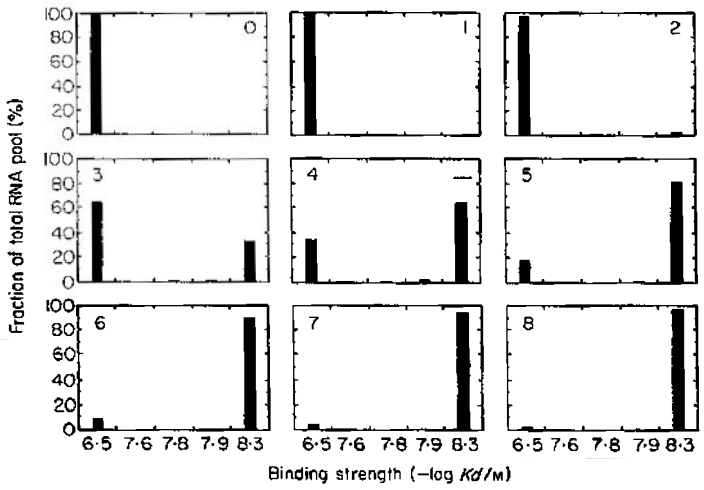
(c) *Estimated ligand properties*

If every nucleotide in the eight-base loop were involved in making specific contacts with gp43, or in positioning other contacts, the best-binding RNA sequence in the 8N-loop experiments would have an additional 16 bits of specific nucleic acid information content (Schneider *et al.*, 1986) relative to the bulk RNA pool. This is the most stringent reconstruction test for these experiments, since whenever less information is actually required the fraction of the bulk RNA pool made up of molecules with wild-type binding would be even greater. In the 8N-loop experiments, a single variant loop sequence was discovered that has a *Kd* indistinguishable from wild-type. Thus, the mean fraction of RNA molecules with the wild-type *Kd* is two out of every 65,536 molecules. Since this example is a reconstruction, we use a setting of 15 bits, which specifies one sequence with the wild-type *Kd* out of every 32,768 unique sequences. The *Kd* for the best-binding RNA is set equal to the value measured for both the wild-type and major-variant sequences, $4{\cdot}8 \times 10^{-9}$ M (Tuerk & Gold, 1990).

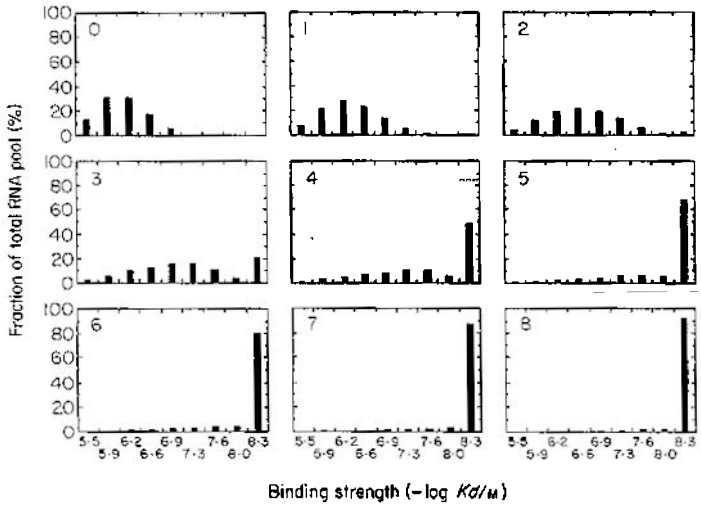
The number of species with unique *Kd* values in the total population is usually unknown, and an arbitrary choice often has to be made for this parameter. For example, this number can be set to a value of 2 when only the enrichment of the best-binding RNA *versus* all competing RNA molecules is of interest. On the other hand, whenever it is also important to follow one or more RNA species with intermediate binding, this number can be set to any value, up to the memory limits of the computer (33 in the current microcomputer implementation). Note that the number of unique *Kd* values equals n in the equations given in the previous sections; hence, the computer time required for solution increases with n^2 (e.g. see eqn (7)).

In this reconstruction test using conditions to simulate 8N-loop experiment B, we follow two RNA species each having the wild-type *Kd* (i.e. the wild-type sequence and the major-variant sequence; Tuerk & Gold, 1990), a mixture of species all having approximately the bulk *Kd* of the RNA pool, and three species each having a unique intermediate *Kd* (corresponding to the *Kd* values of the three minor-variant sequences isolated; Tuerk & Gold, 1990). Pools with every number of different RNA species between two and 33, and with several species having a *Kd* worse than the bulk pool (i.e. larger), have also been followed, with little change in the number of rounds for enrichment of the best-binding RNA to at least half the pool (e.g. see Fig. 4(b)).

Figure 4(a) shows the input distribution of RNA molecules having the bulk *Kd*, the best *Kd*, and three intermediate *Kd* values before the first round of SELEX (panel 0). This input distribution was generated by setting the best *Kd* equal to the *Kd* of the wild-type operator, and then by setting the abundance of the best-binding RNA to one in every 2^{15} molecules, the observed value for wild-type and major-variant sequences. The *Kd* values of the three intermediate species were set to match the minor-variants, and the abundance of each was set to one in every 2^{16} molecules, the observed value. Finally, the *Kd* values for all remaining RNA species, not clonally isolated after the fourth round of 8N-loop



(a)



(b)

Figure 4. Enrichment for 8 simulated rounds of SELEX. (a) Parameter values correspond to 8N-loop experiment B (see Table 1). The input distribution of RNA shown in panel 0 is based on observed values (Tuerk & Gold, 1990): $F_1^0 = 2/65536$, $F_2^0 = F_3^0 = F_4^0 = 1/65536$, $F_5^0 = 65531/65536$, $Kd_1 = 4{\cdot}8 \times 10^{-9}$, $Kd_2 = 1{\cdot}2 \times 10^{-8}$, $Kd_3 = 1{\cdot}7 \times 10^{-8}$, $Kd_4 = 2{\cdot}7^{-8}$ and $Kd_5 = 3{\cdot}2 \times 10^{-7}$ M. The predicted level of every RNA species after each of 8 rounds of SELEX is shown in panels 1 through 8. The relative concentration of RNA molecules having the wild-type Kd after 4 rounds of SELEX is represented as the broken line in panel 4. This reconstruction test predicts that the best-binding RNA will make up 64% of the RNA pool after 4 rounds of SELEX, which correlates reasonably well with the estimated level of 55% after 4 rounds in the laboratory. When the total RNA in each round is set at $1{\cdot}2 \times 10^{-5}$ M for the simulation, instead of the empirically estimated value of $3{\cdot}0 \times 10^{-5}$ M, the predicted level of the best-binding RNA after 4 rounds is 84%, and the predicted level of species around the bulk Kd is also more like that observed in the laboratory (results not shown). (b) Panel 0 shows a more complex input distribution of RNA, hypothetically made up of 9 RNA species with unique Kd values. As in (a), the wild-type Kd, the bulk Kd, and the abundance of RNA molecules having the wild-type Kd all correspond to measured values (Tuerk & Gold, 1990). The Kd of each intermediate species and its abundance have been chosen on the basis of the observed correlation between nucleic acid information content and free energy of binding (Berg & von Hippel, 1987; Stormo, 1988, 1990; Stormo & Yoshioka, 1990). The difference in Kd values from one species to the next is approximately 2·2-fold. Fifteen bits of nucleic acid information content in the best-binding RNA sequences is equivalent to 1 particular position in the 8N loop having either of 2 specific nucleotides (i.e. $F_1^0 = 2^{-15} = 2$ sequences out of 65,536). The abundance of intermediate species follows a random distribution for all permutations of base matches from 7 to 0 in the 7 remaining positions

$$\text{i.e. for } i = 2, 3, 4, \ldots, 9, \quad F_i^0 = \left(\frac{1}{4}\right)^{9-i} \cdot \left(\frac{3}{4}\right)^{i-2} \cdot \frac{7!}{(9-i)! \cdot (i-2)!}.$$

Even with this more complex input distribution, there is no change in the number of rounds required to enrich the best-binding RNA to at least half the pool, and the predicted level of species around the bulk Kd after 4 rounds also is more like that observed in the laboratory.

experiment B, were grouped together and set equal to the bulk Kd for the input pool. Other more complex input distributions that match the wild-type and bulk Kd values have also been used, again with little change in the number of rounds for enrichment of the best-binding RNA to at least half the pool (e.g. see Fig. 4(b)).

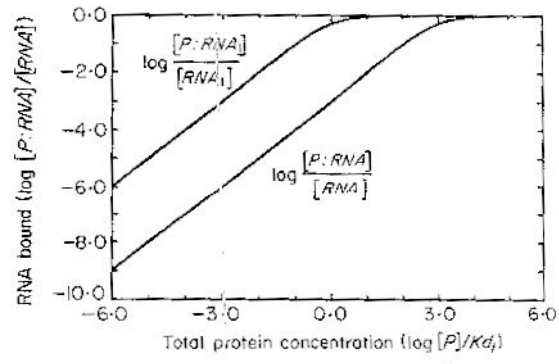
(d) *Predicted results*

The change in the distribution of RNA molecules predicted after each of eight rounds of SELEX with parameters corresponding to 8N-loop experiment B is shown in Figure 4(a), 1 through 8. The relative abundance of RNA molecules with the wild-type Kd obtained in the laboratory after four rounds of SELEX is shown for comparison as the broken line in panel 4. There is a good correlation between the laboratory results using SELEX and the results of simulations using SELEXION. The correlations are similar for different input distributions (e.g. see Fig. 4(b)) and for the conditions used in all of the 8N-loop experiments (results not shown), which indicates that *the equilibrium mechanism proposed for SELEX is sufficient to explain the high levels of enrichment attained in the laboratory after just a few rounds*. With good correlations between the laboratory results and the simulations, further analysis based on the mathematical description of SELEX should provide accurate predictions for the levels of enrichment to expect under different conditions.

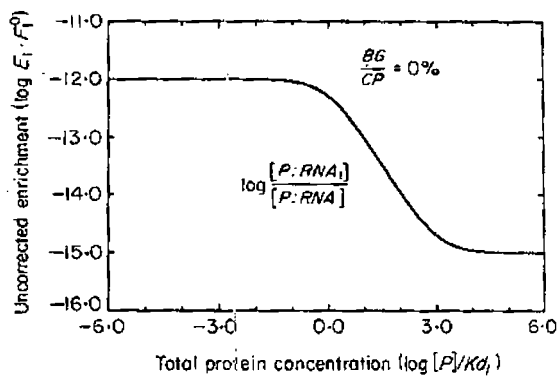
6. Predicted Enrichment for Different Conditions

We have developed computer programs to solve for and plot the predicted concentration of each protein–RNA species i complex as a function of different input conditions, including the total protein concentration, the total RNA concentration, the fraction of the RNA pool made up of the best-binding RNA, the bulk Kd of the total RNA pool relative to the Kd of the best-binding RNA, and the fraction of free RNA molecules that partitions as non-specific background *versus* the fraction of protein–RNA complexes that partitions appropriately ($[P]$, $[RNA]$, F_i^0, $\langle Kd\rangle/Kd_1$, and BG/CP, respectively). Note that the fraction of partitioned RNA molecules that is effectively copied into cDNA for amplification by PCR is not included here, since RT is cancelled entirely from enrichment calculations and therefore has no impact whenever there is adequate representation of each RNA species (e.g. see eqn (10)).

Figure 5(a) shows an example of the predicted enrichment for different conditions. The fraction of the best-binding RNA in complexes, $[P:RNA_1]/[RNA_1]$, and the fraction of the total RNA pool in complexes, $[P:RNA]/[RNA]$, are plotted on a $\log_{10}$ axis *versus* the total protein concentration, normalized to Kd_1 and plotted on a $\log_{10}$ axis. At equilibrium in solution, the relative amount of the best-binding RNA in protein–RNA



(a)



(b)

Fig. 5. (a) Fraction of the best-binding RNA molecules, RNA_1, and fraction of the total RNA pool, RNA, in protein–RNA complexes. (b) Uncorrected enrichment calculated as the ratio between the concentration of the best-binding RNA in protein–RNA complexes, $[P:RNA_1]$, and the concentration of all protein–RNA complexes, $[P:RNA]$, at equilibrium in solution *versus* the total protein concentration. $[P]/Kd_1$ is varied over 12 orders of magnitude, and results are shown for a representative set of parameter values (i.e. $Kd_1 = 1$, $\langle Kd\rangle = 1000$, $F_1^0 = 10^{-15}$, $[RNA] = 10^{-4}$ M, and $BG/CP = 0\%$). In this case, maximum discrimination between the best-binding RNA molecules and competing molecules occurs at total protein concentrations less than Kd_1.

complexes is greatest under conditions that maximize the vertical distance between these curves or as shown in Figure 5(b), the ratio between complexes with the best-binding RNA and all protein–RNA complexes is greatest at low total protein concentration and is smallest when nearly all RNA is bound at high total protein concentration. The upper limit for the ratio plotted in Figure 5(b) can be calculated using the following formula:

$$\frac{[P:RNA_1]}{[P:RNA]} = \frac{[RNA_1]}{[RNA]}\cdot\left(\frac{\langle Kd\rangle+[Pf]}{Kd_i+[Pf]}\right) < \frac{[RNA_1]}{[RNA]}\cdot\left(\frac{\langle Kd\rangle}{Kd_1}\right) = F_i^0\cdot\frac{\langle Kd\rangle}{Kd_1} \qquad (16)$$

For the particular parameter values used to generate Figure 5(b), the upper limit is 10^{-12}, a factor of $\langle Kd\rangle/Kd_1$ above F_i^0, the input fraction of

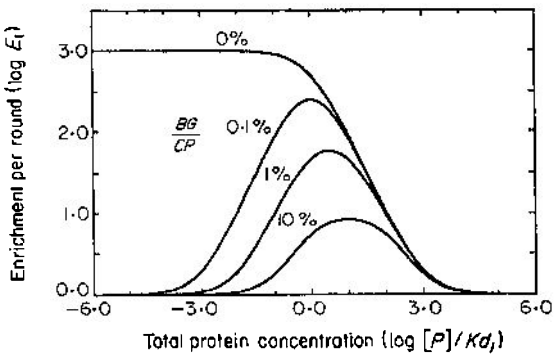


Fig. 6. Enrichment for the best-binding RNA in each round of SELEX *versus* the total protein concentration. Enrichment is calculated with eqn (18). Parameter values are the same as in Fig. 5. The number next to each curve is the percentage of free RNA molecules that partitions as non-specific background *versus* the fraction of protein–RNA complexes that partitions appropriately (BG/CP). Maximum enrichment decreases and occurs at higher total protein concentrations as background partitioning increases.

RNA_1. In this case, decreasing total protein to concentrations much less than Kd_1 does little to increase the enrichment. As illustrated below, when background partitioning is also considered, enrichment actually decreases again at low total protein concentrations.

Figure 5 gives the results to expect for $[RNA]=10^{-4}$ M, $F_i^0=10^{-15}$, and $\langle Kd\rangle/Kd_1=10^3$ but with no correction for background partitioning (i.e. $BG/CP=0\%$). Different values of F_i^0 and $\langle Kd\rangle/Kd_1$ just change the scales of each axis. Correcting for different values of BG/CP has the effects shown in Figure 6. As BG/CP increases, enrichment for the best-binding RNA decreases, and maximum enrichment occurs at higher total protein concentrations. This is because the relative amount of RNA molecules that partitions as non-specific background increases either as BG/CP increases or as the amount of the best-binding RNA molecules that partitions in protein–RNA complexes decreases when less protein is used.

Coupled equilibrium equations similar to equations (1) to (5) have also been used to predict *in vivo* response characteristics; for example, the fraction of *lac* operator sites bound with *lac* repressor in a population of uninduced *Escherichia coli* (Berg & von Hippel, 1987; von Hippel *et al.*, 1974; von Hippel & Berg, 1986). We have duplicated these results by solving equations (1) to (5) to predict a 99·9% fractional saturation of *lac* operator sites (i.e. consistent with an induction ratio of 1000 : 1) given 10^7 non-specific binding sites for *lac* on the *E. coli* chromosome, a non-specific binding constant of $1{\cdot}67\times10^{-3}$ M, a single *lac* operator site per cell, a specific binding constant of 10^{-12} M, and 10 *lac* molecules per cell. These kinds of mathematical treatments are just two examples of the type of non-linear analysis required to integrate detailed knowledge of molecular interactions into a robust, quantitative understanding of any complex process. The results demonstrate that non-linear mathematical analysis is not only a *practical* tool that can be used to enhance the effectiveness of a biotechnological process, but is also another essential tool for understanding complex processes such as regulatory mechanisms in intact organisms (e.g. see Berg, 1988; Berg & von Hippel, 1987, 1988; Savageau, 1976; Voit, 1991; von Hippel & Berg, 1986).

7. Optimum Protein Concentration with Estimates for Kd_1

The enrichment for each round of SELEX is defined as the increase in the fraction of the RNA pool composed of the RNA species with the smallest Kd_i (by convention Kd_1):

$$E_1=\frac{F_1^1}{F_1^0}, \tag{17}$$

where F_1^0 equals $[RNA_1]/[RNA]$ before the round, and F_1^1 equals $[RNA_1]/[RNA]$ after the round, as calculated with equation (14) from the previous section on amplification. After substitution and rearrangement, equation (17) becomes:

$$E_1=\frac{1}{F_i^0}\cdot\frac{[P:RNA_1]+BGb\cdot F_1^0\cdot[RNA]}{\left(\sum_{i=1}^{n}[P:RNA_i]\right)+BGb\cdot[RNA]}=\frac{[P:RNA_1]/[RNA_1]+BGb}{[P:RNA]/[RNA]+BGb}, \tag{18}$$

where $BGb=BG/(CP-BG)$, with BG being the fraction of free RNA molecules that partitions as non-specific background, and with CP being the fraction of protein–RNA complexes that partitions appropriately. The second half of equation (18) shows that enrichment is the fraction of RNA_1 bound plus BGb relative to the fraction of total RNA bound plus BGb.

By using equation (5) to calculate the concentration of complexes with the best-binding RNA, and then by using equations (5) and (10) to calculate the concentration of all protein–RNA complexes, enrichment can be written as a function of one variable ($[Pf]$) and four parameters (Kd_1, $\langle Kd\rangle$, BG and CP):

$$E_1=\frac{[Pf]/(Kd_1+[Pf])+BGb}{[Pf]/(\langle Kd\rangle+[Pf])+BGb}. \tag{19}$$

At low free protein concentrations the calculated value for $\langle Kd\rangle$ is for all practical purposes independent of $[Pf]$ (e.g. see eqn (10)). Typical examples of the accuracy of equation (19) using such approximations for $\langle Kd\rangle$ are shown in Figure 2(a) and (b) and were discussed earlier in the section on selection. The degree to which the approximations for $\langle Kd\rangle$ influence the value predicted for the optimum protein concentration is examined further in this section (e.g. see Figs 7 and 8).

Whenever the free protein concentration is low enough to make calculation of $\langle Kd \rangle$ using equation (10) virtually independent of $[Pf]$ (i.e. $\partial\langle Kd \rangle/\partial[Pf] \approx 0$), the optimum free protein concentration that gives maximum enrichment can be determined analytically by setting the partial derivative of equation (19) with respect to $[Pf]$ equal to zero and rearranging to obtain the following formula:

$$[Pf]\bigstar = \sqrt{\left(\frac{BG}{CP} \cdot Kd_1 \cdot \langle Kd \rangle\right)}. \qquad (20)$$

Hence, the optimum total protein concentration that gives maximum enrichment is calculated as follows:

$$\begin{aligned}
[P] &= [Pf] + [P:RNA] \\
&= [Pf] + \frac{[RNA] \cdot [Pf]}{(\langle Kd \rangle + [Pf])} \\
&= \left\{1 + \frac{[RNA]}{(\langle Kd \rangle + [Pf])}\right\} \cdot [Pf], \qquad (21)
\end{aligned}$$

$$\begin{aligned}
[P]\bigstar &= \left\{1 + \frac{[RNA]}{(\langle Kd \rangle + [Pf]\bigstar)}\right\} \cdot [Pf]\bigstar, \\
&= \left\{1 + \frac{[RNA]}{\left[\langle Kd \rangle + \sqrt{\left(\frac{BG}{CP} \cdot Kd_1 \cdot \langle Kd \rangle\right)}\right]}\right\} \cdot \sqrt{\left(\frac{BG}{CP} \cdot Kd_1 \cdot \langle Kd \rangle\right)}, \\
&= \left\{\langle Kd \rangle + \frac{[RNA]}{\left[1 + \sqrt{\left(\frac{BG}{CP} \cdot \frac{Kd_1}{\langle Kd \rangle}\right)}\right]}\right\} \cdot \sqrt{\left(\frac{BG}{CP} \cdot \frac{Kd_1}{\langle Kd \rangle}\right)}, \\
&\approx (\langle Kd \rangle + [RNA]) \cdot \sqrt{\left(\frac{BG}{CP} \cdot \frac{Kd_1}{\langle Kd \rangle}\right)},
\end{aligned}$$

where $\sqrt{\{(BG/CP) \cdot (Kd_1/\langle Kd \rangle)\}}$ is approximately equal to the fraction of the total RNA pool bound (i.e. $[P:RNA]/[RNA]$). By substituting $[Pf]\bigstar$ from equation (20) into equation (19) and rearranging, the following formula for the maximum enrichment per round is obtained:

$$\begin{aligned}
E_1\bigstar &= \left(\frac{\sqrt{(BG \cdot Kd_1)} + \sqrt{(CP \cdot \langle Kd \rangle)}}{\sqrt{(CP \cdot Kd_1)} + \sqrt{(BG \cdot \langle Kd \rangle)}}\right)^2 \\
&= \left(\frac{1 + \sqrt{\left(\frac{CP}{BG} \cdot \frac{\langle Kd \rangle}{Kd_1}\right)}}{\sqrt{\frac{CP}{BG}} + \sqrt{\frac{\langle Kd \rangle}{Kd_1}}}\right)^2. \qquad (22)
\end{aligned}$$

Equations (21) and (22) show that the optimum protein concentration and maximum enrichment per round are functions of three physical conditions: namely, the bulk Kd of the total RNA pool ($\langle Kd \rangle$), the Kd of the best-binding RNA (Kd_1), and background partitioning (BG/CP). For the mechanisms under consideration, maximum enrichment per round is never greater than the difference in binding ($\langle Kd \rangle/Kd_1$) or the fraction of protein–RNA complexes that partitions appropriately *versus* the fraction of free RNA molecules that partitions as non-specific background (CP/BG). Although the difference in binding by definition is never greater than the reciprocal of the starting fraction of the best-binding RNA (i.e. $\langle Kd \rangle/Kd_1 \leqslant 1/F_1^0$, see eqn (10)), for any round with a small starting fraction of the best-binding RNA, the optimum protein concentration and maximum enrichment are for all practical purposes independent of F_1^0.

Figure 7(a) shows optimum total protein concentrations calculated by independent numerical methods that include the actual dependence of $\langle Kd \rangle$ on $[Pf]$ (i.e. $\partial\langle Kd \rangle/\partial[Pf] \neq 0$, see eqn (10)). These numerically determined concentrations agree with equation (21), and the maximum levels of enrichment shown in Figure 7(b) also agree with equation (22). Therefore, $\partial\langle Kd \rangle/\partial[Pf]$ approaches zero at the optimum free protein concentration, which means that the assumption required to derive the analytical solution for $[Pf]\bigstar$ is valid.

In general, any total RNA concentration giving adequate representation of the best-binding RNA species can be chosen, and maximum enrichment can be attained, by using the total protein concentration calculated with equation (21). In the 8N-loop experiments (Tuerk & Gold, 1990), protein concentrations from 12-fold greater than optimum (experiment A) to fourfold less than optimum (experiment B) were used, and the experiment with the protein concentration closest to optimum (2·6-fold greater than optimum in experiment C) gave the best enrichment.

Figure 8(a) shows the minimum number of rounds for enrichment of the best-binding RNA from one molecule in 10^{15} to at least half the pool using optimum protein concentrations for $BG/CP = 1\%$, and Table 2 summarizes the minimum, maximum and average number of rounds for differences in binding between fourfold and 10^6-fold and for levels of background partitioning of 0·1%, 1% and 10%. The minimum number of rounds is five when there is a 10^6-fold difference in binding and background partitioning is as low as 0·1%, and the maximum number of rounds is 49 when there is just a fourfold difference in binding and background partitioning is as high as 10%. The average is 12 rounds for the full range of differences in binding between fourfold and 10^6-fold and for levels of background partitioning of 0·1%, 1% and 10%. Naturally, even fewer rounds are required when there is a higher starting fraction of the best-binding RNA; for example, with a starting fraction of one in a billion molecules, the

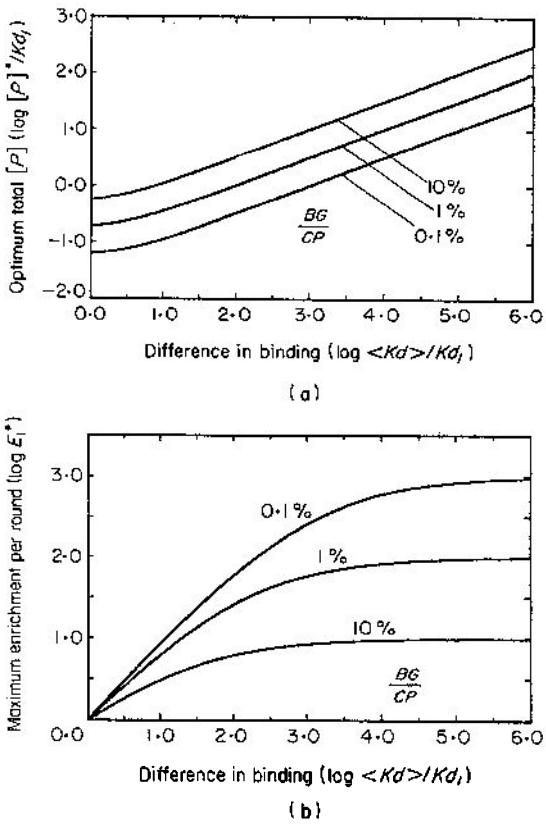


Fig. 7. (a)Optimum total protein concentration ($[P]\bigstar/Kd_1$. (b) Maximum enrichment per round ($E_1\bigstar$) each as a function of the relative difference in binding between best RNA and the bulk pool ($\langle Kd\rangle/Kd_1$). Numerically determined results with $\partial\langle Kd\rangle/\partial[Pf] \neq 0$ are shown for differences in binding between 4-fold and 10^6-fold. The number next to the each curve is the percentage of free RNA molecules that partitions as non-specific background *versus* the fraction of protein–RNA complexes that partitions appropriately (BG/CP). For the mechanisms described in the text, maximum enrichment per round is never greater than the relative difference in binding ($\langle Kd\rangle/Kd_1$) or the fraction of protein–RNA complexes that partitions appropriately *versus* the fraction of free RNA molecules that partitions as non-specific background (CP/BG).

average is just seven rounds to enrich to at least half the pool using optimum protein concentrations (e.g. see Table 2).

The program SELEXION supports the use of optimum total protein concentrations calculated in each round with equation (21). To confirm that these settings actually minimize the number of rounds, we have compared the predicted minimum number of rounds with the number of rounds determined independently in simulations using the optimum total protein concentration for each round. Figure 8(a) shows number of rounds needed to enrich the best-binding RNA from one molecule in 10^{15} to at least half the pool *versus* the initial difference in binding ($\langle Kd\rangle/Kd_1$). The optimum protein concentration set with equation (21) always gives enrichment within two or three rounds of the predicted minimum, and the reason why a few extra rounds are required is easy to understand. For a pool with one or more species having intermediate binding, the difference in binding between the bulk pool and the best-binding RNA becomes appreciably smaller after a few rounds than it was initially. Then, with smaller differences in binding, enrichment for each subsequent round is less than the maximum predicted on the basis of a pool with just two species. This reduction of the difference in binding explains why the minimum number of rounds measured in simulations with many unique species is always a few rounds higher than the minimum number of rounds predicted for just two species, as is borne out by simulations with just two species (results not shown).

The enrichment attained with optimum protein concentrations provides an ideal benchmark by which any strategy for enrichment can be evaluated objectively. Also, being able to calculate the optimum protein concentration, given any total RNA concentration, eliminates all trial and error and guesswork from finding conditions that give rapid enrichment with any protein or other binding agent of interest. The only parameters required are the bulk Kd of the total RNA pool ($\langle Kd\rangle$), the desired Kd for the best-binding RNA (Kd_1), and the fraction of free RNA molecules that partitions as non-specific background *versus* the fraction of protein–RNA complexes that partitions appropriately (BG/CP). In the next section we develop an alternative strategy that gives near-maximum enrichment even when the binding advantage for the best RNA is unknown.

8. Near-optimum Protein Concentration with No Estimate for Kd_1

In many circumstances, the Kd of the best-binding RNA is unknown, and the optimization strategy described in the preceding section would not be helpful unless a guess were made for Kd_1. For these situations, we have developed an alternative, near-optimum strategy that ensures good levels of enrichment over a broad range of possible Kd values for the best-binding RNA without requiring any estimate for Kd_1. The approach is to use the mathematical description of SELEX to find the protein concentration that minimizes the increase in the number of rounds, no matter what the actual difference is between the bulk Kd and Kd_1. Then, the only parameters required to determine near-optimum settings are the total RNA concentration, the bulk Kd, and the level of background partitioning, BG/CP, with BG being the fraction of free RNA molecules that partitions as non-specific background and CP being the fraction of protein–RNA complexes that partitions appropriately.

Figure 9(a) shows an example of the maximum enrichment possible whenever the ratio of $\langle Kd\rangle$ and Kd_1 is some value between unity (i.e. no difference) and 10^6 (i.e. a 10^6-fold difference), and it shows the

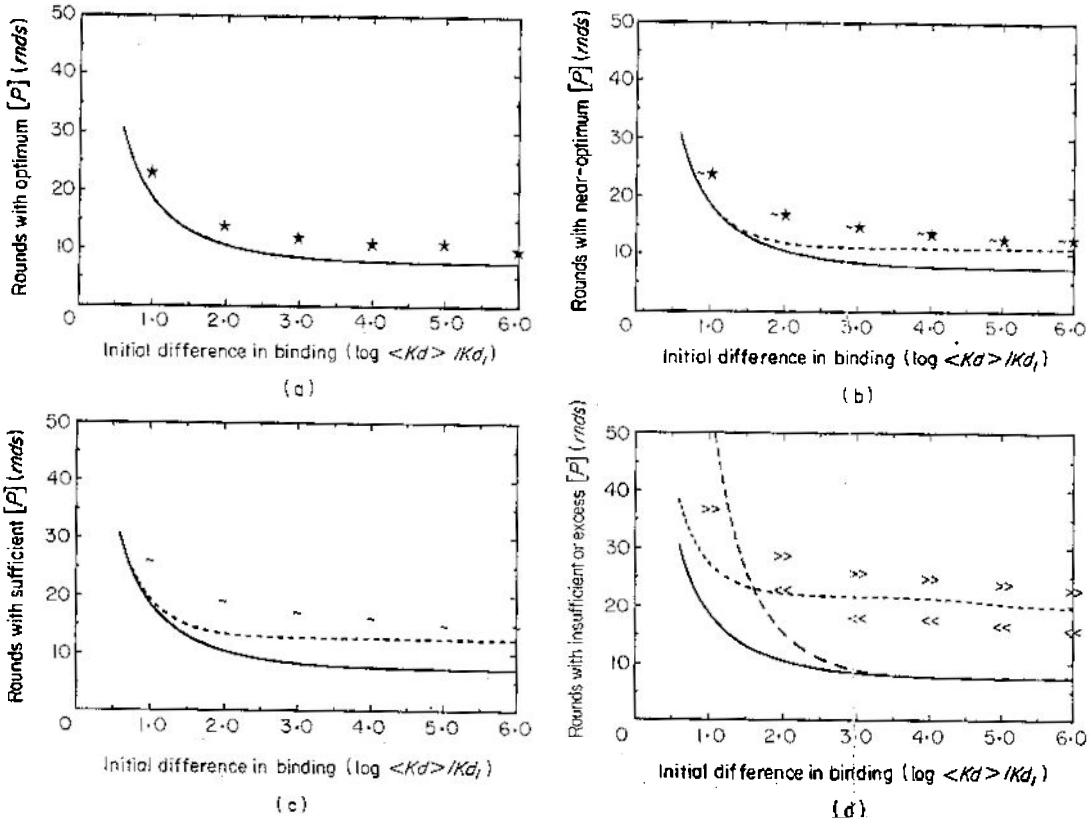


Fig. 8. Simulation tests of SELEX using: (a) optimum (★), (b) near-optimum (~★), (c) sufficient (~), (d) insufficient (<<) or excess (>>) total protein concentrations (see eqns (21), (23) and (25)). The number of rounds required to enrich the best-binding RNA from 1 molecule in 10^{15} to at least half the pool is plotted *versus* the initial difference in binding ($\langle Kd \rangle / Kd_1$). All results are for representative settings of background partitioning and total RNA ($BG/CP = 1\%$ and $[RNA] = 10^{-4}$ M). In each panel, the continuous line is the predicted minimum number of rounds with optimum protein concentrations and only 2 competing species of RNA. The broken lines are the predicted number of rounds with 2 RNA species and the relevant protein concentration, and the symbols represent the results of simulations with 10 unique species of RNA and the relevant protein concentration. The insufficient protein concentration used in (d) binds 0·1% of the total RNA in each round, and the excess protein concentration binds 20%. For any difference in binding, the optimum, near-optimum and sufficient protein concentrations all give enrichment in fewer rounds than either the insufficient or the excess protein concentrations. Also, with the insufficient protein concentration, the best-binding RNA is lost in the first round for an initial difference in binding of 10-fold or less. In each case, the extra rounds with near-optimum or sufficient protein concentrations *versus* the minimum number of rounds with the optimum protein concentration correspond to those predicted (see Figs 9 and 10).

enrichment to expect when different guesses for the value of Kd_1 are used in equation (21). There is no reduction in enrichment whenever the guess for $\langle Kd \rangle / Kd_1$ happens to equal the actual value for $\langle Kd \rangle / Kd_1$. Figure 9(b) shows the potential increase in the number of rounds for a more-focused range of different guesses. For overly optimistic guesses, the number of rounds goes up dramatically, but for overly pessimistic guesses the impact is not quite as severe. For guesses of one 4th, 8th, 16th, 32nd, 64th or 128th, actually using a setting for Kd_1 that is one 64th the bulk Kd of the total RNA pool minimizes the average increase in the number of rounds for all possible values of $\langle Kd \rangle / Kd_1$, or using a setting that is one 16th the bulk Kd minimizes the peak increase in the number of rounds. Since the peak increase in the number of rounds represents the maximum amount of extra work that could be encountered with any guess for $\langle Kd \rangle / Kd_1$, we have chosen to base our near-optimum strategy on minimizing this peak, rather than minimizing the average increase.

On the basis of several cases similar to those shown in Figure 9(a) and (b), the following protein concentrations minimize the peak increase in the number of rounds for all possible values of $\langle Kd \rangle / Kd_1$ between two and 10^6, and for all possible values of $\langle Kd \rangle$, $[RNA]$, and BG/CP:

$$[P]_{\sim\bigstar} = (\langle Kd \rangle + [RNA]) \cdot 0{\cdot}7 \cdot \left(\frac{BG}{CP} \right)^{2/3}. \quad (23)$$

Whenever the total RNA concentration is appreciably greater than the bulk Kd of the RNA pool, for example when a typical nucleic acid binding protein is used and the total nucleic acid concentration approaches 10^{-4} M, the bulk Kd can be neglected in equation (23):

$$[P]_{\sim\bigstar} \approx [RNA] \cdot 0{\cdot}7 \cdot \left(\frac{BG}{CP} \right)^{2/3}. \quad (24)$$

Table 2

Minimum, maximum and average number of rounds of SELEX required to enrich the best-binding ligands to at least half the pool, starting from five different input fractions ($F_1^0 = [LIG_1]/[LIG]$), using the optimum, the near-optimum or the sufficient total concentration of target molecules in each round

Total $[T]$‡		Number of rounds starting from five different fractions of LIG_1 (F_1^0)†				
		10^{-3}	10^{-6}	10^{-9}	10^{-12}	10^{-15}
$[T]$★§	Min	1	2	3	4	5
	Max	10	20	30	40	49
	Avg	2	4	7	9	12
$[T] \sim$ ★‖	Min	2	3	5	6	8
	Max	10	20	30	40	49
	Avg	3	6	8	11	14
$[T] \sim$ ¶	Min	3	5	8	10	12
	Max	12	23	34	45	56
	Avg	3	6	9	12	15

$[LIG]$ = total concentration of amplifiable ligands, $[LIG_1]$ = concentration of amplifiable ligand sequences that bind best, $[T]$ = total concentration of target molecules, $\langle Kd \rangle$ = bulk equilibrium dissociation constant for total amplifiable ligand pool, Kd_1 = equilibrium dissociation constant for amplifiable ligand sequences that bind best, BG = fraction of free amplifiable ligands that partitions as non-specific background, CP = fraction of target–ligand complexes that partitions appropriately. When accessible $\langle Kd \rangle$, Kd_1, BG and CP should be measured at the actual concentration of non-amplifiable, non-specific competitor molecules used in each round.

†At each target concentration, the minimum (Min) is for a 10^6-fold difference in binding and background partitioning of 0·1%, the maximum (Max) is for a 4-fold difference in binding and background partitioning of 10%, and the average (Avg) includes 4-fold to 10^6-fold differences in binding and levels of background partitioning of 0·1%, 1% and 10%.

‡Whenever the total concentration of ligands is less than $\langle Kd \rangle$ multiplied by the fraction of ligands bound, the total concentration of target molecules necessary to bind that fraction of ligands actually is greater than the total ligand concentration itself (e.g. see formulae below).

§Optimum $[T]$ $\approx (\langle Kd \rangle + [LIG]) \cdot \sqrt{\{(BG/CP) \cdot (Kd_1/\langle Kd \rangle)\}}$ (eqn (21))

‖Near-optimum $[T]$ $= (\langle Kd \rangle + [LIG]) \cdot 0{\cdot}7 \cdot (BG/CP)^{2/3}$ (eqn (23))

¶Sufficient $[T]$ $= (\langle Kd \rangle + [LIG]) \cdot 0{\cdot}06$ (eqn (25))

In these circumstances, the protein concentration that gives near-maximum enrichment can be determined simply by knowing the total RNA concentration and background partitioning.

An example of the performance of this near-optimum strategy is shown in Figure 8(b). For near-optimum protein concentrations set with equation (23), the number of rounds for enrichment of the best-binding RNA from one molecule in 10^{15} to at least half the pool is always within two or three of that for optimum protein concentrations set with equation (21). This corresponds to the number of extra rounds predicted for the setting used for background partitioning (i.e. between 2 and 3 extra rounds for $BG/CP = 0{\cdot}1\%$; see the plot for a guess of $\langle Kd \rangle / Kd_1 = 16$ in Fig. 9(b)). The performance of this strategy remains near optimum for all other sets of parameter values tested. Table 2 summarizes the minimum, maximum and average numbers of rounds to enrich the best-binding RNA from one molecule in 10^{15} to at least half the pool using near-optimum protein concentrations for differences in binding between fourfold and 10^6-fold and for levels of background partitioning of 0·1%, 1% and 10%. The minimum number of rounds is eight when there is a 10^6-fold difference in binding and background partitioning is as low as 0·1%, and the maximum number of rounds is 49 when there is just a fourfold difference in binding and background partitioning is as high as 10%. The average is 14 rounds for the full range of differences in binding between fourfold and 10^6-fold and for levels of background partitioning of 0·1%, 1% and 10%. Once again, even fewer rounds are required when there is a higher starting fraction of the best-binding RNA; for example, with a starting fraction of one in a billion molecules, the average with near-optimum protein concentrations is just eight rounds (e.g. see Table 2). Thus, even when the potential improvement in binding is unknown, equation (23) still gives an effective guideline for setting the total protein concentration to ensure rapid enrichment.

9. Sufficient Protein Concentration with No Estimate for Kd_1 or Background

Certain nucleic acid species may tend to partition more readily as background than others. Without extra steps to eliminate such molecules, they may be enriched along with molecules that bind well to the desired target, and the fraction of free RNA molecules that partitions as background actually might increase after each round. Since some estimate of background partitioning is necessary to calculate either the optimum or the near-optimum protein concentration for each round (e.g. see eqn (21) and (23)), a good estimate for the fraction of free RNA molecules that partitions as background compared with the fraction of protein–RNA complexes that partitions appropriately (BG/CP) must be made to apply either of these strategies. Whenever enough RNA is available, BG can be measured in each round simply as the fraction of total RNA molecules that partitions when no

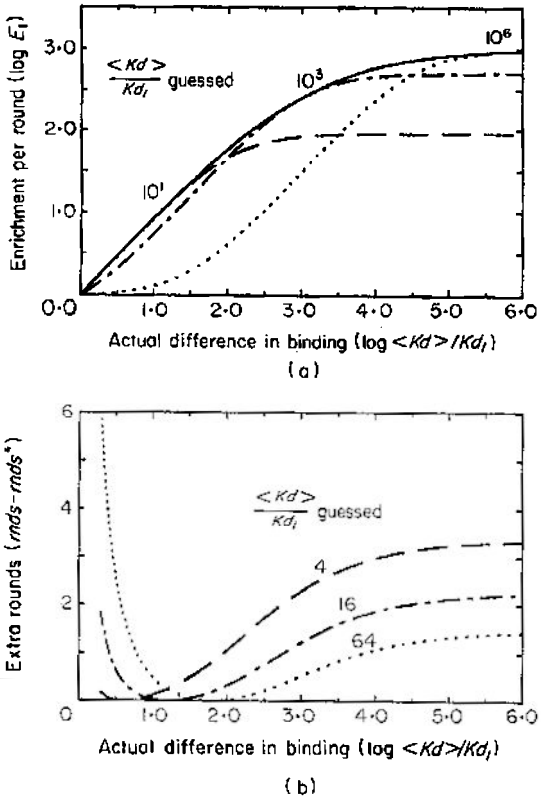


Fig. 9. (a) Enrichment per round of SELEX using various guesses for the relative difference in binding versus a broad range of actual values for the relative difference in binding, $\langle Kd \rangle / Kd_1$. The number next to each curve represents the guess used for $\langle Kd \rangle / Kd_1$. The continuous line is the maximum enrichment possible for the parameters used ($F_1^0 = 2^{-50} \approx 10^{-15}$, $BG/CP = 0{\cdot}1\%$, and $[RNA] = 10^{-4}$ M). Whenever the guess happens to equal the actual value of $\langle Kd \rangle / Kd_1$, the enrichment equals the maximum enrichment possible for the round. With a guess for $\langle Kd \rangle / Kd_1$ of 10, there is little difference between the enrichment attained and the maximum enrichment possible at values of $\langle Kd \rangle / Kd_1$ less than 10, but enrichment is as much as 10-fold below maximum at values of $\langle Kd \rangle / Kd_1$ approaching 10^6. With a guess for $\langle Kd \rangle / Kd_1$ of 10^6 the opposite is true; there is as much as a 10-fold reduction in enrichment when $\langle Kd \rangle / Kd_1$ is small and almost no reduction as it approaches a value of 10^6. With a guess for $\langle Kd \rangle / Kd_1$ of 10^3 there are smaller, but still significant, reductions in enrichment at values of $\langle Kd \rangle / Kd_1$ either less than or greater than 10^3. (b) Increase in the number of rounds required to enrich the best-binding RNA from 1 molecule in 10^{15} to at least half the pool. Again, the number next to each curve is the guess used for $\langle Kd \rangle / Kd_1$. With the particular setting used for background partitioning ($BG/CP = 0{\cdot}1\%$), and with guesses of 4, 8, 16, 32, 64 or 128, the guess for $\langle Kd \rangle / Kd_1$ of 64 minimizes the average increase in the number of rounds over the full range of actual values of $\langle Kd \rangle / Kd_1$ between 2 and 10^6, and the guess for $\langle Kd \rangle / Kd_1$ of 16 minimizes the peak increase in the number of rounds over the same range. Different starting levels for the best-binding RNA simply change the scale on the y axis, but the shape and the relative position of each curve remain the same. Several similar curves, generated using different values for BG/CP, were analyzed to derive a general formula for calculation of the near-optimum total protein concentration ($[P] \sim \bigstar$) giving the minimum peak increase in the number of rounds for any value of $\langle Kd \rangle / Kd_1$ between 2 and 10^6 and for any total RNA concentration or any relevant level of background partitioning (see eqn (23)).

protein is added, and CP can be measured as the fraction of all complexes that partitions at saturation.

However, there may be times when the available amount of RNA is too limited, and measuring background partitioning for every round might be impractical. For these situations, we have used the mathematical description of SELEX to derive a third strategy that ensures sufficient levels of enrichment for any difference in binding between fourfold and 10^6-fold (i.e. $\langle Kd \rangle / Kd_1 = 4$ to 10^6), and for any level of background partitioning between 0·1 and 10% (i.e. $BG/CP = 0{\cdot}1\%$ to 10%). As in the preceding section, the approach used is to find the protein concentration that minimizes the absolute increase in the number of rounds for the relevant range of parameters.

Figure 10(a) to (c) shows the absolute increase in the number of rounds above the minimum (see Fig. 8) for differences in binding between fourfold and 10^6-fold and levels of background partitioning of 0·1%, 1% and 10%, when enough protein is used to bind 10%, 6·3% or 4% of the total RNA. In each case, there are no extra rounds whenever the fraction of total RNA bound happens to equal $\sqrt{\{(BG/CP) \cdot (Kd_1/\langle Kd \rangle)\}}$ (e.g. see eqn (21)). With 10% of the RNA bound, the peak increase is ten extra rounds; with 6·3% bound, the peak drops to only seven extra rounds; and with 4% bound, the peak goes up again, in this case to 20 extra rounds. Figure 10(d) shows the peak increase and the average increase in the number of rounds when enough protein is added to bind from 0·2% to 20% of the total RNA. The minimum peak increase is seven extra rounds when approximately 6% of the total RNA is bound in each round. With this fraction of RNA bound, the average increase over the full range of differences in binding and background partitioning is four extra rounds.

The protein concentration that binds roughly 6% of the total RNA can be calculated as follows:

$$[P] \sim = (\langle Kd \rangle + [RNA]) \cdot 0{\cdot}06, \qquad (25)$$

or whenever $[RNA] \gg \langle Kd \rangle$ (see also eqn (24)):

$$[P] \sim \approx [RNA] \cdot 0{\cdot}06. \qquad (26)$$

By using this protein concentration, 12 rounds would be required to enrich from one molecule in 10^{15} to at least half the pool for a difference in binding of 10^6-fold and background partitioning of 0·1%, and 56 rounds would be required for a difference in binding as little as fourfold and background partitioning as high as 10% (see Table 2). The average is 15 rounds for all differences in binding from fourfold to 10^6-fold and for levels of background partitioning of 0·1%, 1% and 10%.

These minimum, maximum and average numbers can be considered upper limits based on enrichment

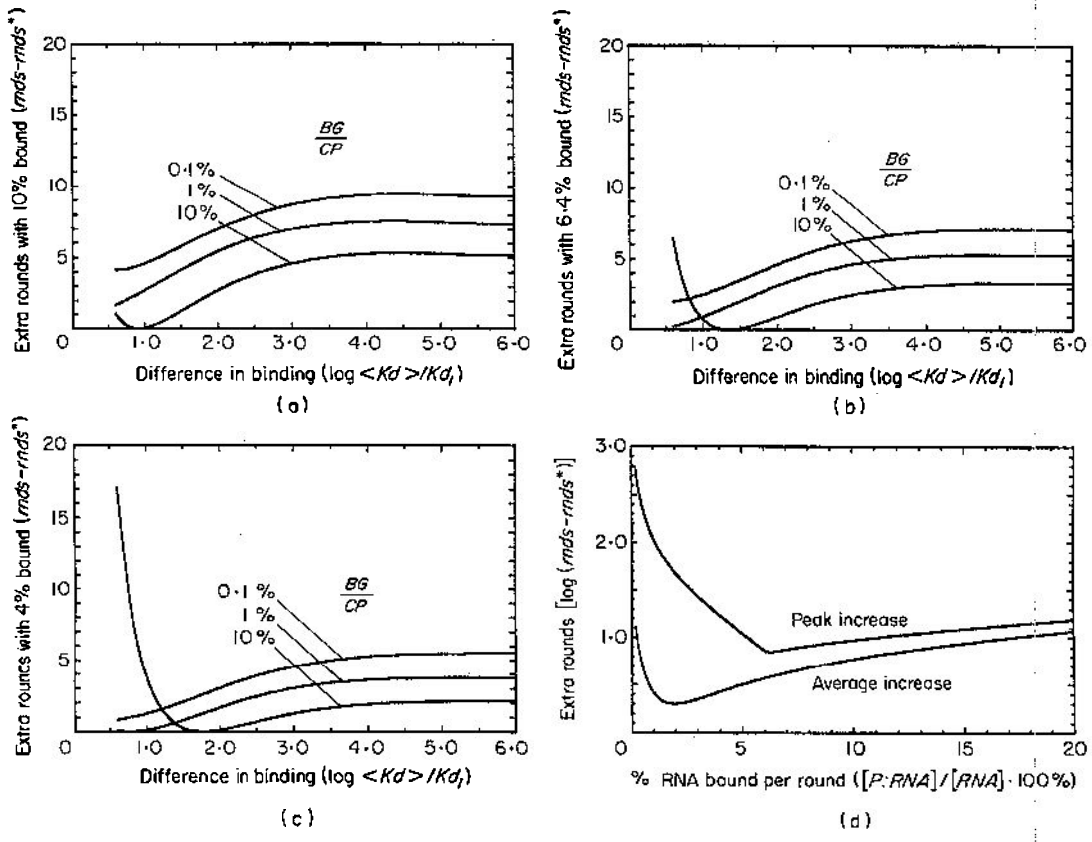


Fig. 10. Absolute increase in the number of rounds over the minimum number required to enrich from 1 molecule in 10^{15} to at least half the pool ($rnds - rnds^{\bigstar}$) with (a) 10%, (b) 6·3%, or (c) 4% of the total RNA bound in each round, *versus* the actual difference in binding ($\langle Kd \rangle / Kd_1$) and the actual background partitioning (BG/CP). The total RNA concentration equals 10^{-4} M, and results are shown for levels of background partitioning of 0·1%, 1% and 10%. (d) Peak and average increase in the number of rounds over the full range of differences in binding between 4-fold and 10^6-fold and for levels of background partitioning of 0·1%, 1% and 10% *versus* the fraction of total RNA bound. The minimum peak increase is 7 rounds with approximately 6% of the RNA bound in each round, and the average increase with that amount of RNA bound is 4 extra rounds.

from as little as one molecule in 10^{15} to at least half the pool. For higher starting fractions of the best-binding RNA (i.e. $F_1^0 > 10^{-15}$), the absolute increase in the number of rounds shown in Figure 10 can simply be scaled down by a factor of $-\log_{10}(F_1^0)/15$. Table 2 summarizes the minimum, maximum and average numbers of rounds to expect when 6% of the total RNA is bound in each round, and the starting fraction of the best-binding RNA is one molecule in 10^3, 10^6, 10^9, 10^{12} or 10^{15}.

Results of simulations using the sufficient protein concentration to bind 6% of the total RNA in each round are shown in Figure 8(c). The number of rounds required to enrich the best-binding RNA from one molecule in 10^{15} to at least half the pool is always within seven of that required with optimum protein concentrations. This peak increase corresponds to the maximum number of extra rounds predicted, and the strategy is equally effective for all other sets of parameter values tested (e.g. see Table 2). Figure 8(d) illustrates the consequences of using a protein concentration either significantly less than (i.e. insufficient) or significantly greater than (i.e. excess) the sufficient protein concentration. For these examples, the insufficient protein concentration binds 0·1% of the total RNA in each round, and the excess protein concentration binds 20%. In general, the sufficient protein concentration gives enrichment in fewer rounds than either the insufficient or the excess protein concentration, and unlike the case with the insufficient protein concentration, the best-binding RNA is never lost in the first round. Thus, even when both the actual difference in binding and the actual level of background partitioning are unknown, equation (25) still gives an effective guideline for setting the total protein concentration in each round to ensure adequate enrichment for any difference in binding between fourfold and 10^6-fold and for levels of background partitioning between 0·1% and 10%.

10. Likelihood of Success

The likelihood of recovering the best-binding RNA in each round of SELEX increases with the number of such molecules present, with their binding advantage *versus* the bulk RNA pool, and with the total amount of protein used. Although it

is difficult to know in advance how to maximize the difference in binding, the likelihood of recovering the best-binding RNA can still be increased by maximizing the number of RNA molecules sampled and by using a high protein concentration. Of course, using a high protein concentration can reduce the enrichment per round enough for the total number of rounds to increase dramatically. In this section we develop a strategy to balance the competing demands of using a high protein concentration to promote successful recovery (especially in the initial rounds when there might be only a single molecule or just a few molecules of the best-binding RNA as opposed to using a lower protein concentration to enrich the best-binding RNA in as few rounds as possible.

The likelihood of binding a single RNA molecule having an equilibrium dissociation constant of Kd_i can be approximated by dividing both sides of equation (5) (giving the protein–RNA species i complex concentration, $[P:RNA_i]$) by the total RNA species i concentration, $[RNA_i]$:

$$\mathscr{L}_i^b = \frac{[P:RNA_i]}{[RNA_i]} = \frac{([P]-[P:RNA])}{Kd_i+([P]-[P:RNA])}$$

$$= \frac{[Pf]}{Kf_i+[Pf]}, \qquad i = 1, \ldots n, \qquad (27)$$

where $[P]$ is the total protein concentration, and $[P:RNA]$ is the concentration of all protein–RNA complexes calculated with equation (9), and $[P:RNA_i]/[RNA_i]$ can be interpreted either as the mean fraction of RNA_i molecules bound or as the mean fraction of time that each individual RNA_i molecule is bound.

To calculate the total protein concentration that gives an $\mathscr{L}_i^b$ likelihood of binding each individual RNA_i molecule, first equation (27) is solved for $[Pf] = \mathscr{M}_i^b \cdot Kd_i$, where $\mathscr{M}_i^b$ equals $\mathscr{L}_i^b/(1-\mathscr{L}_i^b)$. Then this expression for $[Pf]$ is substituted into the first part of equation (21) to obtain the following formula for the total protein concentration:

$$P]_{\mathscr{L}_i^b} = \mathscr{M}_i^b \cdot Kd_i + \frac{\mathscr{M}_i^b}{\mathscr{M}_i^b + \frac{\langle Kd \rangle}{Kd_i}} \cdot [RNA],$$

$$i = 1, \ldots n, \quad (28)$$

where $\langle Kd \rangle$ is the bulk Kd of the total RNA pool calculated with equation (10) and $[RNA]$ is the total RNA concentration.

For example, the total protein concentration that gives a 99% likelihood of binding each individual molecule of the best-binding RNA can be calculated with the following formula:

$$P]_{0.99} = 99 \cdot Kd_1 + \frac{99}{99 + \frac{\langle Kd \rangle}{Kd_1}} \cdot [RNA]. \qquad (29)$$

This formula is derived by setting $\mathscr{L}_i^b = 0.99$ in equation (28) and by setting $i = 1$ for the best-binding RNA RNA_1 by convention). When this

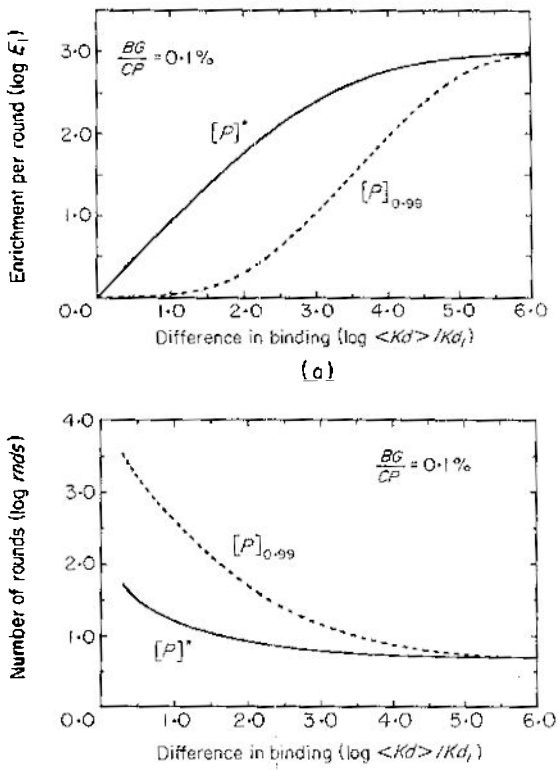


Fig. 11. (a) Reduction in enrichment using a total protein concentration that gives a 99% likelihood of binding each individual RNA_1 molecule ($[P]_{0.99}$ calculated with eqn (29)). (b) Increase in the number of rounds required to enrich the best-binding RNA to at least half the pool using $[P]_{0.99}$ *versus* a broad range of differences in binding, $\langle Kd \rangle / Kd_1$. The continuous line in each panel is the maximum enrichment possible for the parameters used ($F_1^0 = 2^{-50} \approx 10^{-15}$, $BG/CP = 0.1\%$ and $[RNA] = 10^{-4}$ M). Trends are similar for different parameter values.

formula is used to calculate the protein concentration to use in each round of SELEX, the likelihood of binding each individual RNA molecule having a Kd less than or equal to Kd_1 is 99% or better.

Of course, as already mentioned, using high protein concentrations dramatically increases the number of rounds required, whenever enrichment per round is reduced appreciably. Figure 11(a) shows the reduction in the enrichment per round as a function of $[P]_{0.99}$ over a wide range of differences in binding of the best RNA *versus* the bulk RNA pool, $\langle Kd \rangle / Kd_1$; and Figure 11(b) shows the resulting increase in the number of rounds required to enrich RNA_1 from one molecule in 10^{15} to at least half the pool, if $[P]_{0.99}$ were used in every round. The trends shown in Figure 11 are for one representative example of background partitioning ($BG/CP = 0.1\%$); similar trends exist for any other relevant level of background partitioning.

The total number of rounds shown in Figure 11(b) could be reduced significantly by switching to the near-optimum protein concentration {$[P] \sim \bigstar$ calculated with equation (23)} as soon as enough RNA_1 molecules are present to ensure a 99% likeli-

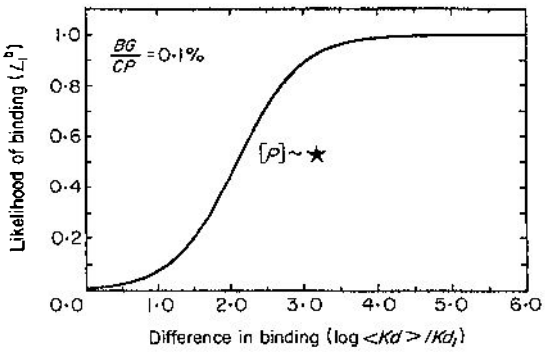


Fig. 12. Likelihood ($\mathscr{L}$) of binding each individual RNA_1 molecule using the total protein concentration that gives near-maximum enrichment ($[P] \sim ★$ calculated with eqn (23)) *versus* a broad range of differences in binding, $\langle Kd \rangle / Kd_1$. Parameters are $F_1^0 = 2^{-50} \approx 10^{-15}$, $BG/CP = 0{\cdot}1\%$ and $[RNA] = 10^{-4}$ M. The likelihood of binding each individual RNA_1 molecule is identical for different values of F_1^0 or $[RNA]$. Higher background partitioning results in a higher value for $[P] \sim ★$, as calculated with eqn (23). With these higher protein concentrations, the curve giving the likelihood of binding is shifted left.

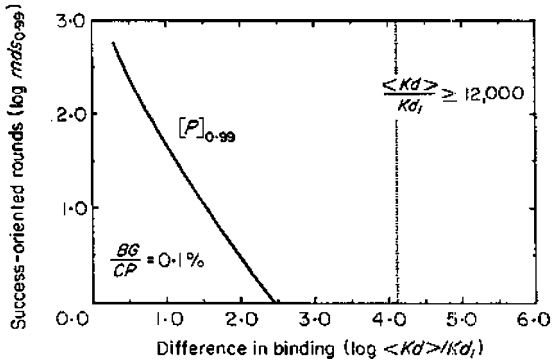


Fig. 13. Number of initial success-oriented rounds required with $[P]_{0·99}$ before reducing the concentration to $[P] \sim ★$ for near-maximum enrichment *versus* a broad range of differences in binding, $\langle Kd \rangle / Kd_1$. Parameters are $F_1^0 = 2^{-50} \approx = BG/CP = 0{\cdot}1\%$ and $[RNA] = 10^{-4}$ M. The number of initial success-oriented rounds is the same for different values of $[RNA]$ or F_1^0. However, higher background partitioning results in a higher value for $[P] \sim ★$, as calculated with eqn (23). With these higher protein concentrations the likelihood of binding each individual RNA_1 molecule is higher, and fewer initial success-oriented rounds are necessary, since fewer molecules of the best-binding RNA are required before the protein concentration can be dropped to $[P] \sim ★$.

hood of binding one or more of these molecules. To know when it is safe to switch to $[P] \sim ★$, the likelihood of binding each individual RNA_1 molecule using this amount of protein must be known. This likelihood can be calculated simply by substituting $[P] \sim ★$ and Kd_1 into equation (27):

$$\mathscr{L}_1^b \sim ★ = \frac{[P{:}RNA]}{[RNA_1]} \sim ★ = \frac{([P] \sim ★ - [P{:}RNA] \sim ★)}{Kd_i + ([P] \sim ★ - [P{:}RNA] \sim ★)} = \frac{[Pf] \sim ★}{Kd_1 + [Pf] \sim ★}, \quad (30)$$

where $[P{:}RNA] \sim ★$ is calculated by substituting $[P] \sim ★$ into equation (9).

The likelihood of binding each individual RNA_1 molecule with $[P] \sim ★$ is plotted in Figure 12 as a function of the difference in binding of the best RNA *versus* the bulk RNA pool, $\langle Kd \rangle / Kd_1$. As expected, the likelihood of binding increases with $\langle Kd \rangle / Kd_1$. Once again, these are likelihoods of binding for $[P] \sim ★$ calculated with 0·1% background partitioning (see eqn (23)); for $[P] \sim ★$ calculated with any other relevant level of background partitioning, the scales representing the difference in binding are numbered differently, but the shape of each curve remains the same.

To have a 99% likelihood of binding one or more RNA_1 molecule with $[P] \sim ★$, the product of the probabilities that each RNA_1 molecule is unbound, $(1 - \mathscr{L}_1^b \sim ★)$, using this amount of protein should be less than or equal to 1%.

$$\prod_{i=1}^{rna_1} (1 - \mathscr{L}_1^b \sim ★) = (1 - \mathscr{L}_1^b \sim ★)^{rna_1} \leqslant 0{\cdot}01 \quad (31)$$

$$rna_1 \geqslant \frac{2}{-\log_{10}(1 - \mathscr{L}_1^b \sim ★)},$$

where rna_1 is the number of RNA_1 molecules, and $\mathscr{L}_1^b \sim ★$ is calculated with equation (30).

The number of rounds required to enrich from one molecule of the best-binding RNA to rna_1 molecules using $[P]_{0·99}$ depends upon the enrichment attained per round with this amount of protein:

$$E_1([P]_{0·99}) = \frac{0{\cdot}99 + BGb}{\frac{[P{:}RNA]_{0·99}}{[RNA]} + BGb}, \quad (32)$$

where $[P]_{0·99}$ is given by equation (29), $[P{:}RNA]_{0·99}$ is calculated by substituting $[P]_{0·99}$ into equation (9), and BGb equals $BG/(CP - BG)$. With the above level of enrichment per round, it would take $rnds_{0·99}$ success-oriented rounds to enrich RNA_1 from one molecule to rna_1 molecules:

$$\{E_1([P]_{0·99})\}^{rnds_{0·99}} = rna_1 \quad (33)$$

$$rnds_{0·99} = \frac{\log_{10} rna_1}{\log_{10} E_1([P]_{0·99})},$$

where rna_1 is the number of RNA_1 molecules required before a total protein concentration as low as $[P] \sim ★$ can be used and still maintain a 99% likelihood of binding one or more RNA_1 molecules.

The number of success-oriented rounds required with $[P]_{0·99}$ before switching to enrichment-oriented rounds with $[P] \sim ★$ is plotted in Figure 13 as a function of the difference in binding of the best RNA *versus* the bulk RNA pool, $\langle Kd \rangle / Kd_1$. Note that any fraction of a round shown in this Figure is interpreted as a complete round. The vertical line in

Table 3

Relative difference in Kd *values sufficient to give a 99% likelihood of binding each individual ligand molecule as a function of the number of initial success-oriented rounds with* $[P]_{0.99}$ (see eqn (29))

$rnds_{0.99}$	$\frac{\langle Kd \rangle}{Kd_1}$
0	⩾ 12,000
1	⩾ 290
2	⩾ 150
4	⩾ 77
8	⩾ 42
16	⩾ 24

Parameters are $F_1^0 = 2^{-50} \approx 10^{-15}$, $BG/CP = 0{\cdot}1\%$, $[RNA] = 10^{-4}$ M, and $Vol = 10^{-4}$ l.

Figure 13 at $\langle Kd \rangle / Kd_1 = 12{,}000$ shows the threshold for the relative binding advantage above which no initial success-oriented rounds are required in this case because $[P] \sim \bigstar$ already gives a 99% or better likelihood of binding each individual RNA_1 molecule. Once again, the trend shown in Figure 13 is for both $E_1([P]_{0.99})$ and $[P] \sim \bigstar$ calculated with 0·1% background partitioning (see eqns (32) and (23)); similar trends exist when $E_1([P]_{0.99})$ and $[P] \sim \bigstar$ are calculated with any other relevant level of background partitioning.

Figure 13 demonstrates that as the number of success-oriented rounds is increased, the binding advantage necessary to attain a 99% likelihood of forming a complex with one or more RNA_1 molecules using $[P] \sim \bigstar$ is smaller. Table 3 gives the number of success-oriented rounds required to maintain a 99% likelihood of binding one or more RNA molecules with a binding advantage of either 24-, 42-, 77-, 150- or 290-fold over the bulk RNA pool. With no success-oriented rounds, and with $E_1([P]_{0.99})$ and $[P] \sim \bigstar$ calculated for background partitioning of 0·1% (see eqns (32) and (23)), again each individual RNA molecule having a 12,000-fold or larger binding advantage over the bulk pool has a 99% or better likelihood of being bound with $[P] \sim \bigstar$. After one success-oriented round, a 290-fold binding advantage is sufficient to give a 99% likelihood of binding one or more RNA_1 molecules with $[P] \sim \bigstar$, and after 16 success-oriented rounds, as little as a 24-fold binding advantage is sufficient.

The equations derived in this section provide a general strategy for calculating the number of success-oriented rounds required before switching to enrichment-oriented rounds with near-optimum protein concentrations, $[P] \sim \bigstar$, for any given set of conditions including the minimum binding advantage sufficient to ensure a 99% likelihood of success, the total RNA concentration, and the level of background partitioning. Similar strategies are also available for enrichment-oriented rounds with optimum or sufficient protein concentrations, $[P]\bigstar$ or $[P] \sim$. The appropriate strategy is activated in the program SELEXION whenever a desired threshold Kd for a 99% likelihood of success is set. Then, when the ideal protein concentration is calculated for each round of SELEX, this concentration is set not only to give enrichment of the best-binding RNA in as few rounds as possible with the available information, but also to ensure a 99% or better likelihood of binding every species of RNA having a Kd as good as or better than the desired threshold.

11. Sequence Representation

The mean number of RNA molecules representing each species in a population having a randomized sequence of length L is:

$$RNA_i = \frac{[RNA] \cdot Vol \cdot 6{\cdot}02 \times 10^{23}}{4^L}, \qquad (34)$$

where $[RNA]$ is the total RNA concentration, and Vol is the total volume. We refer to the ratio of the total number of RNA molecules sampled divided by the possible number of unique sequences of length L as the *saturation of sequence space* (e.g. see Table 1). When this ratio is less than unity it represents the upper limit for the fraction of unique sequences sampled out of all possible sequences of length L. For RNA transcripts of approximately 100 bases (which is a typical number for a SELEX experiment utilizing a total of 60 bases for 2 PCR primer sites and a transcriptional promoter site, and 30 or more bases within the protein binding region; Tuerk & Gold, 1990), the maximum RNA concentration generally is 10^{-4} M in a volume of 10^{-4} l. With this amount of RNA, for a randomized sequence of length 26, on average there would be about one molecule representing each of $4{\cdot}5 \times 10^{15}$ unique RNA sequences, and for a randomized sequence of length 24 the probability of complete representation of all $2{\cdot}8 \times 10^{14}$ possible sequences approaches 99%. Unless the total amount of RNA sampled is increased to more than 10^{-8} mol, all possible sequences of more than 26 nucleotides could not be represented in a single experiment, and complete representation of all possible sequences of more than 24 nucleotides is unlikely (e.g. see Table 4).

Whenever the program SELEXION is set to determine a sufficient RNA concentration, a three-tier strategy is activated to determine the total RNA concentration necessary to promote adequate sequence representation. Tier 1 increases the number of molecules of the best-binding RNA simply by increasing the total RNA concentration, up to a concentration as high as 10^{-4} M. Then, when necessary, tier 2 increases the number of molecules of the best-binding RNA recovered by increasing the total protein concentration, up to a concentration as high as $[P]_{0.99}$ (see eqn (29)). Since $[P]_{0.99}$ forms complexes with approximately 99% of the best-binding RNA molecules, tier 2 should work for L-length sequences up to the limits shown in Table 4. For sequences with more than 48 bits of specific information content, the total number of RNA molecules sampled must be increased, either by using a higher total concentration of RNA or, perhaps more practically, by using a greater volume as is done in tier 3. Of course, at some point using a

Table 4

Minimum number of nucleic acid molecules required to represent all possible sequences of length L, *and the number giving a 99% likelihood of complete sequence representation*

Length (L)	Information (bits)	Minimum (mols)	Complete (mols)
16	32	$7{\cdot}13 \times 10^{-15}$	$1{\cdot}91 \times 10^{-13}$
18	36	$1{\cdot}14 \times 10^{-13}$	$3{\cdot}37 \times 10^{-12}$
20	40	$1{\cdot}83 \times 10^{-12}$	$5{\cdot}91 \times 10^{-11}$
22	44	$2{\cdot}92 \times 10^{-11}$	$1{\cdot}03 \times 10^{-9}$
24	48	$4{\cdot}68 \times 10^{-10}$	$1{\cdot}77 \times 10^{-8}$
26	52	$7{\cdot}48 \times 10^{-9}$	$3{\cdot}04 \times 10^{-7}$
28	56	$1{\cdot}20 \times 10^{-7}$	$5{\cdot}20 \times 10^{-6}$
30	60	$1{\cdot}92 \times 10^{-6}$	$8{\cdot}85 \times 10^{-5}$
32	64	$3{\cdot}06 \times 10^{-5}$	$1{\cdot}50 \times 10^{-3}$

The information content in a sequence of length L is shown for comparison. The minimum number of nucleic acid molecules required is $1{\cdot}66 \times 10^{-24} \times 4^L$ mols, and the number giving a 99% likelihood of complete sequence representation is $(2{\cdot}30 \times L + 7{\cdot}65) \times 10^{-24} \times 4^L$ mols.

greater volume will not be practical either; for example, when the amount of RNA or protein, or the size of labware becomes limiting. In these cases, SELEXION still calculates the conditions necessary for promoting adequate sequence representation, but it is left for the user to judge when these conditions actually can be applied in the laboratory.

12. Conclusions

Enrichment processes like SELEX provide systematic, general means to isolate nucleic acid molecules with specific biological activities. In biotechnological or pharmacological applications, these enrichment processes can be used to evolve nucleic acid ligands for diagnostic or therapeutic purposes; for example, to use for assays or treatments in which specific binding agents (equivalent to antibodies) might be considered useful. A systematic process like SELEX has the advantage of rapid *in vitro* development of ligands with specific biological activities and the advantage of multifunctional training. Drug delivery, intracellular uptake and ligand stability are among the challenges facing the application of nucleic acid ligand sequences as therapeutic agents.

SELEX also provides a new probe for studying nucleic acid structure–function relationships and for studying nucleic acid–protein interactions. SELEX can be used to discover sequences more rapidly and exhaustively with intermediate *Kd* values or higher *Kd* values than wild-type; for example by using a target *Kd* either greater than or less than the wild-type *Kd* to calculate optimum total protein concentrations. For experiments in which a binding sequence can occupy any frame or register within a longer randomized sequence, sequence analysis algorithms (e.g., Stormo & Hartzell, 1989) and RNA folding algorithms (e.g. see Freier *et al.*, 1986; Gutell *et al.*, 1985; Zuker, 1989) will be helpful in finding a consensus sequence or functionally equivalent structures within an apparently diverse set of unique clones that all bind well.

In this study, we have utilized mathematical analysis and computer simulation to show that an equilibrium mechanism for binding in solution is sufficient to explain the high levels of enrichment attained in the laboratory after just a few rounds of SELEX. Beyond these reconstruction tests, we have performed additional analyses to show the levels of enrichment to expect under an extensive range of relevant conditions. Perhaps most importantly, for practical purposes, we have determined the ideal nucleic acid and protein concentrations to use in four pertinent situations: (1) The total nucleic acid and protein concentrations to use for maximum enrichment when both the *Kd* of the best-binding nucleic acid species and the fraction of free nucleic acid molecules that partitions as non-specific background *versus* the fraction of protein–nucleic acid complexes that partitions appropriately are known or can be fixed at desired values; (2) the total protein concentration to use for near-maximum enrichment even when the best *Kd* is unknown; (3) the total protein concentration sufficient for rapid enrichment even when the best *Kd* and the actual level of background partitioning are unknown; and (4) the total protein concentration to use in the initial rounds to ensure a high likelihood of recovering even the rare molecule that binds well.

The mathematical description of SELEX is based on formulae for accurate solution of coupled-equilibrium equations using Newton's method (see eqns (4) through (10)). These formulae have been implemented in an interactive simulation program called SELEXION: for *Systematic Evolution of Ligands by EXponential enrichment with Integrated Optimization by Non-linear analysis* (our unpublished results). This program can be used to predict the outcome of SELEX experiments, or to test the consistency of input conditions with the levels of enrichment attained in the laboratory. The formulae for the optimum, near-optimum or sufficient total protein concentrations, or for the total protein concentration giving a high likelihood of success and the amount of RNA giving adequate sequence representation, are also available in SELEXION. With these formulae, experimental conditions can be set to promote successful enrich-

ment of the best-binding species in as few rounds as possible.

The analysis and guidelines for enhanced effectiveness can be generalized to any enrichment process involving selection and partitioning, followed by amplification of ligand sequences from the desired fraction. Table 2 summarizes three practical guidelines for setting the target concentration in each round of SELEX. In every case, the total number of ligand molecules sampled should always be enough to give the greatest possible saturation of sequence space (see Table 4), and the target concentration should be set using either equation (21), (23) or (25), depending on which parameter values are known (see Table 2).

For example, when both the improvement in binding ($\langle Kd \rangle / Kd_1$) and the level of background partitioning (BG/CP) are known, the target concentration should bind the fraction of the total ligand pool equal to the geometric mean of BG/CP and $Kd_1/\langle Kd \rangle$, i.e. $\sqrt{\{(BG/CP) \cdot (Kd_1/\langle Kd \rangle)\}}$. Such optimum target concentrations will give enrichment in as few rounds of SELEX as possible. For differences in binding between fourfold and 10^6-fold and for levels of background partitioning between 0·1% and 10%, on average, 12 rounds are enough to enrich the best-binding ligand from one molecule in 10^{15} to at least half the pool.

Whenever the level of background partitioning is known but the improvement in binding is unknown, the target concentration should be set to bind a fraction of the total ligand pool equal to $0{\cdot}7 \times (BG/CP)^{2/3}$. This near-optimum target concentration will give enrichment in as few extra rounds of SELEX as possible, regardless of the actual difference in binding. For differences in binding between fourfold and 10^6-fold and for levels of background partitioning between 0·1% and 10%, on average, 14 rounds are enough to enrich the best-binding ligand from one molecule in 10^{15} to at least half the pool (i.e. only 2 rounds more on average than with the optimum target concentrations). These near-optimum target concentrations are probably the most generally useful, since (except in reconstruction tests with a well-characterized target molecule) the actual difference in binding is generally unknown, but the fraction of free ligands that partitions as non-specific background *versus* the fraction of protein–ligand complexes that partitions appropriately is usually easy to measure, and may not change appreciably from one round to the next.

Even when the level of background partitioning is unknown, using a target concentration that binds 6% of the total ligand pool will still give a sufficient level of enrichment in each round. With 6% of the total ligand pool bound in each round, the following simple rule of thumb applies: *for differences in binding between fourfold and 10^6-fold and for levels of background partitioning between 0·1% and 10%, on average, approximately* X *rounds of SELEX are enough to enrich the best-binding ligand from one molecule in 10^X to at least half the pool* (see average for $[T] \sim$ in Table 2).

When any of the guidelines given in Table 2 are followed precisely, but no enrichment is attained after an extended number of rounds (and there are no technical complications) either the input ligand pool had too few molecules of the best-binding species to give a high likelihood of success in the initial rounds, or the input ligand pool had no molecules with a Kd appreciably better than the bulk Kd. Possible remedies to such problems include increasing the total protein concentration in the initial rounds, as is discussed in the section on the likelihood of success (see also eqns (27) to (33) and Table 3), and increasing the total number of unique sequences sampled, as is discussed in the section on sequence representation (see also Table 4).

We thank Sean Eddy, Gary Stormo and Peter von Hippel for critically reviewing the manuscript and offering constructive suggestions for improvements. We also thank Britta Singer, Barney Whitman and Mike Yarus for helpful discussions. D.I. was supported by fellowship DRG-1031 from the Damon Runyon-Walter Winchell Cancer Research Fund. Additional support came from NIH grants GM28685 and GM19963 to L.G. We also thank the W. M. Keck Foundation for their generous support of RNA science on the Boulder campus.

References

Abelson, J. (1990). Directed evolution of nucleic acids by independent replication and selection. *Science*, **249**, 488–489.

Andrake, M., Guild, N., Hsu, T., Gold, L., Tuerk, C. & Karam, J. (1988). DNA polymerase of bacteriophage T4 is an autogenous translational repressor. *Proc. Nat. Acad. Sci., U.S.A.* **85**, 7942–7946.

Berg, O. G. (1988). Selection of DNA binding sites by regulatory proteins: the LexA protein and the arginine repressor use different strategies for functional specificity. *Nucl. Acids Res.* **16**, 5089–5105.

Berg, O. G. & 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.

Berg, O. G. & von Hippel, P. H. (1988). Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. *J. Mol. Biol.* **200**, 709–723.

Biedenkapp, H., Borgmeyer, U., Sippel, A. E. & Klempnauer, K-H. (1988). Viral myb oncogene encodes a sequence-specific DNA-binding activity. *Nature (London)*, **335**, 835–837.

Blackwell, T. K. & 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.

Blackwell, T. K., Kretzner, L., Blackwood, E. M., Eisenman, R. N. & Weintraub, H. (1990). Sequence-specific DNA binding by the c-Myc protein. *Science*, **250**, 1149–1151.

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. Nat. Acad. Sci., U.S.A.* **87**, 6378–6382.

Devlin, J. J., Panganiban, L. C. & Devlin, P. E. (1990). Random peptide libraries: a source of specific protein binding molecules. *Science*, **249**, 404–406.

Ellington, A. D. & Szostak, J. W. (1990). In vitro selec-

tion of RNA molecules that bind specific ligands. *Nature (London)*, **346**, 818–822.

Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T. & Turner, D. H. (1986). Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Nat. Acad. Sci., U.S.A.* **83**, 9373–9377.

Green, R., Ellington, A. D. & Szostak, J. W. (1990). *In vitro* genetic analysis of the Tetrahymena self-splicing intron. *Nature (London)*, **347**, 406–408.

Gutell, R. R., Weiser, B., Woese, C. R. & Noller, H. F. (1985). Comparative anatomy of 16-S-like ribosomal RNA. *Progr. Nucl. Acids Res.* **32**, 155–216.

Innis, M. A., Myambo, K. B., Gelfand, D. H. & Brow, M. A. D. (1988). DNA sequencing with *Thermus aquaticus* DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA. *Proc. Nat. Acad. Sci., U.S.A.* **85**, 9436–9440.

Joyce, G. F. (1989*a*). Amplification, mutation and selection of catalytic RNA. *Gene*, **82**, 83–87.

Joyce, G. F. (1989*b*). RNA evolution and the origins of life. *Nature (London)*, **338**, 217–224.

Joyce, G. F. & Inoue, T. (1989). A novel technique for the rapid preparation of mutant RNAs. *Nucl. Acids Res.* **17**, 711–723.

Kinzler, K. W. & Vogelstein, B. (1989). Whole genome PCR: application to the identification of sequences bound by gene regulatory proteins. *Nucl. Acids Res.* **17**, 3645–3653.

Kinzler, K. W. & Vogelstein, B. (1990). The GLI gene encodes a nuclear protein which binds specific sequences in the human genome. *Mol. Cell. Biol.* **10**, 634–642.

Leunberger, D. G. (1973). *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA.

Mavrothalassitis, G., Beal, G. & Papas, T. S. (1990). Defining target sequences of DNA-binding proteins by random selection and PCR: determination of the GCN4 binding sequence repertoire. *DNA Cell Biol.* **9**, 783–788.

North, G. (1990). Expanding the RNA repertoire. *Nature (London)*, **345**, 576–578.

Oliphant, A. R. & Struhl, K. (1987). The use of random-sequence oligonucleotides for determining consensus sequences. *Methods Enzymol.* **155**, 568–582.

Oliphant, A. R. & Struhl, K. (1988). Defining the consensus sequences of *E. coli* promoter elements by random selection. *Nucl. Acids Res.* **16**, 7673–7683.

Oliphant, A. R., Brandl, C. J. & Struhl, K. (1989). Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Mol. Cell. Biol.* **9**, 2944–2949.

Pollock, R. & Treisman, R. (1990). A sensitive method for the determination of protein–DNA binding specificities. *Nucl. Acids Res.* **18**, 6197–6204.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1988). *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, New York.

Robertson, D. L. & Joyce, G. F. (1990). Selection *in vitro* of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature (London)*, **344**, 467–468.

Savageau, M. A. (1976). *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology*, Addison-Wesley, Reading, Mass.

Savageau, M. A. (1991). A critique of the enzymologist's test tube. In *Fundamentals of Medical Cell Biology*, vol. 34, *Chemistry of the Living Cell* (E. E. Bittar, ed.), pp. 45–108, JAI Press Inc., Greenwich, Connecticut.

Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188**, 415–431.

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

Sompayrac, L. & Danna, K. J. (1990). Method to identify genomic targets of DNA binding proteins. *Proc. Nat. Acad. Sci., U.S.A.* **87**, 3274–3278.

Stormo, G. D. (1988). Computer methods for analyzing sequence recognition of nucleic acids. *Annu. Rev. Biophys. Biophys. Chem.* **17**, 241–263.

Stormo, G. D. (1990). Consensus patterns in DNA. *Methods Enzymol.* **183**, 211–221.

Stormo, G. D. & Hartzell, G. W. (1989). Identifying protein-binding sites from unaligned DNA fragments. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 1183–1187.

Stormo, G. D. & Yoshioka, M. (1991). Specificity of the mnt protein determined by binding to randomized operators. *Proc. Nat. Acad. Sci., U.S.A.* **88**, 5699–5703.

Straus, O. H. & Goldstein, A. (1943). Zone behavior of enzymes illustrated by the effect of dissociation constant and dilution on the system cholinesterase-physostigmine. *J. Gen. Physiol.* **26**, 559–585.

Thiesen, H.-J. & Bach, C. (1990). Target Detection Assay (TDA): a versatile procedure to determine DNA binding sites as demonstrated on SP1 protein. *Nucl. Acids Res.* **18**, 3203–3209.

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

Tuerk, C., Eddy, S., Parma, D. & Gold, L. (1990). Autogenous translational operator recognized by bacteriophage T4 DNA polymerase. *J. Mol. Biol.* **213**, 749–761.

Uhlenbeck, O. C., Carey, J., Romaniuk, P. J., Lowary, P. T. & Becket, D. (1983). Interaction of R17 coat protein with its RNA binding site for translational repression. *J. Biomol. Struct. Dynam.* 1, 539–552.

Voit, E. O. (1991). *Canonical Nonlinear Modelling: S-system Approach to Understanding Complexity*, (Voit, E. O., ed.), Van Nostrand & Reinhold, New York.

von Hippel, P. H. & Berg, O. G. (1986). On the specificity of DNA-protein interactions. *Proc. Nat. Acad. Sci., U.S.A.* **83**, 1608–1612.

von Hippel, P. H., Revzin, A., Gross, C. A. & Wang, A. C. (1974). Non-specific DNA binding of genome regulating proteins as a biological control mechanism: I. The *lac* operon: equilibrium aspects. *Proc. Nat. Acad. Sci., U.S.A.* **71**, 4808–4812.

Webb, J. L. (1963). *Enzyme and Metabolic Inhibitors*, vol. 1, *General Principles of Inhibition*, pp. 66–78, Academic Press, New York.

Yarus, M. (1976). Adsorbent filters: a new technique for microexperimentation on nucleic acid. *Anal. Biochem.* **70**, 346–353.

Yarus, M. & Berg, P. (1967). Recognition of tRNA by aminoacyl-tRNA synthetases. *J. Mol. Biol.* **28**, 479–490.

Yarus, M. & Berg, P. (1970). On the properties and utility of a membrane filter assay in the study of isoleucyl-tRNA synthetase. *Anal. Biochem.* **35**, 450–465.

Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.