JMB

A Mathematical Analysis of *in Vitro* Molecular Selection–Amplification

Fengzhu Sun[1]*, David Galas[2] and Michael S. Waterman[1,3]

[1]*Department of Mathematics University of Southern California, Los Angeles CA 90089-1113, USA*

[2]*Darwin Molecular Corporation, 1631 220th St. SE, Bothell, WA 98021, USA*

[3]*Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089-1113 USA*

*Corresponding author

We construct a mathematical model for *in vitro* molecular selection with amplification. Using DNA-protein binding as the illustrative example, we obtain an expression for the probability that a randomly selected molecule from the final *in vitro* selection products is a molecule with the highest binding affinity. Experiments of this type have been reported for several examples of DNA binding proteins. Our study requires a model of the DNA-protein binding constant between DNA molecules and the target protein. The relationship between binding constants and selection probabilities is presented under simplifying but reasonable assumptions. From our analysis, we find that for successful *in vitro* selection experiments there should be a certain relationship between the number of polymerase chain reaction cycles and the concentration of free protein. The results obtained should be widely applicable to a variety of selection–amplification procedures.

© 1996 Academic Press Limited

Keywords: in vitro selection; DNA-protein binding; PCR; binding constant; probability

Introduction

In vitro selection of molecules is a developing technology that is being used in a wide range of biological studies such as protein–DNA interactions (Kinzler & Vogelstein, 1989, 1990; Murtin *et al.*, 1989; Blackwell & Weintraub, 1990; Thiesen & Bach, 1990; Pollock & Treisman, 1990; Rebar & Pabo, 1994), protein binding sites on RNA (Tuerk & Gold, 1990; Lin *et al.*, 1994; Peterson *et al.*, 1994), catalytic properties of RNA molecules (Joyce 1989a,b; Beaudry & Joyce, 1992; Lehman & Joyce, 1993; Bartel & Szostak, 1993; Lin *et al.*, 1994; Lorsch & Szostak, 1994; Prudent *et al.*, 1994; Wilson & Szostak, 1995), and catalytic properties of single-stranded DNA molecules (Breaker & Joyce, 1994; Cuenoud & Szostak, 1995). The basic principle of *in vitro* selection can be summarized as follows. First a library of random sequences (DNA, RNA or protein sequences) is constructed. Some of the molecules in the library are assumed to have a specific function in which we are interested, and a selection procedure is used to isolate those molecules. These molecules are then amplified by

Present address: Fengzhu Sun, Department of Genetics and Molecular Medicine, Emory University School of Medicine, Atlanta, GA 30329, USA.

Abbreviations used: PCR, polymerase chain reaction; IHF, integration host factor.
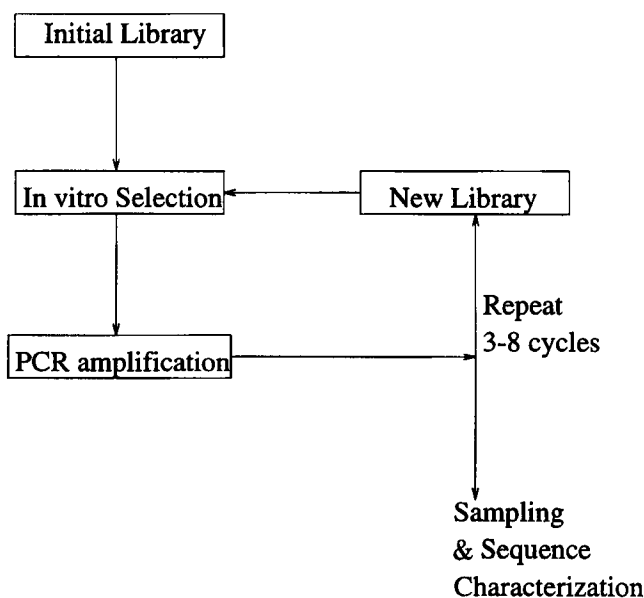
some means, and this population is subjected to selection. This cycle is then repeated.

In this paper, the goal is to characterize one class of these problems, namely finding DNA molecules that bind to a given protein. This protein will be called the target protein. The intention is to select DNA molecules that strongly bind to the target protein and to determine the contribution of each individual base involved in the protein–DNA interactions under some reasonable assumptions. In order to study this, we first synthesize a library of random sequences consisting of all DNA molecules that might bind to the protein. The DNA molecules are flanked by two primers for amplification by the polymerase chain reaction (PCR). These DNA in the library are presented to the target protein. Some of the DNA molecules will bind to the target protein. The DNA–protein complexes are separated from unbound DNA molecules by gel electrophoresis. We refer to this as the selection step. The small amount of DNA present in the DNA–protein complexes is eluted from the protein and then amplified by PCR. We refer to this as the amplification step. These two steps, selection and amplification, constitute one selection–amplification cycle. The experiment is repeated for several cycles. Figure 1 shows the mechanism of *in vitro* selection. The idea is that each cycle selectively enriches the DNA molecule population in proportion to their binding affinity. After many cycles, the DNA molecules with the

Figure 1. Mechanism of *in vitro* selection.

highest binding affinity will dominate. Thus, the DNA molecules with highest affinity should be selected by the process. We construct a mathematical model for *in vitro* selection and present some results using the theory of branching processes; we model the binding constant between a DNA molecule and the target protein. Then we study the probability that the best binding molecules can be selected after *m* selection–amplification cycles. Our goal here is a mathematical understanding of selection–amplification experiments. We derive a relationship between the number of PCR cycles and the concentration of free protein needed for successful *in vitro* selection experiments. We defer all our proofs to the Appendix. A discussion of the extension of our results to other classes of *in vitro* molecular selection–amplification problems appears in the Discussion.

A Mathematical Model

Consider an experiment in which many DNA molecules with different sequences are allowed to bind to a single type of protein molecule. Using equilibrium binding affinity of the DNA molecules to a target protein, we can divide the DNA library into several groups. Each group is composed of molecules with equal binding affinity. In our analysis, we also assume that there is only one binding site in each DNA molecule. Suppose that in the initial library we have N different groups of molecules. Let n_i and K_i, $i = 1, 2, \ldots, N$ be the number of molecules and the binding constant of *i*th group, respectively. The binding constant K of a DNA molecule to a target protein is defined as follows. In a DNA–protein binding experiment at equilibrium (DNA molecules with a single sequence), let p_T and p be the concentration of total

and free protein in the binding reaction. Let D_T and D be the concentration of total and free DNA molecules, respectively. Let D_P be the concentration of DNA–protein complexes. Then:

$$D_P = K \times D \times p, \quad D_P + D = D_T.$$

It follows that:

$$D_P = \frac{K \times p \times D_T}{1 + K \times p}.$$

Thus, the fraction of the DNA molecules that are bound to the target protein is:

$$f = \frac{K \times p}{1 + K \times p}. \tag{1}$$

If we assume that DNA molecules bind to the protein independently, then each DNA molecule binds to the target protein with probability f. For group i molecules, we define $f_i = K_i \times p / (1 + K_i \times p)$. We will discuss the limitations of this assumption below: Implications and Limitations. Therefore, in the selection step of a selection–amplification cycle, a group i molecule binds to the target protein with probability f_i. We select only DNA molecules that bind to the target protein in the selection step. These probabilities, $f_1, f_2, \ldots, f_N$, then describe the distribution of the selected DNA molecules.

Amplification is the result of PCR, which can be modeled by a branching process (Sun, 1995). Suppose the efficiency of the PCR reaction cycle is λ. Then, in each PCR cycle, a molecule generates two copies with probability λ and remains one copy with probability $1 - \lambda$. We assume that the amplification by PCR is completely accurate. We carry out l PCR cycles in the PCR amplification step. The experiment of selection and PCR amplification is repeated for m cycles.

We define the following parameters: N, number of different groups of DNA molecules; n_i, number of group i DNA molecules; K_i, binding constant of a group i DNA molecule; f_i, probability of a group i DNA molecule binding to the target protein; λ, efficiency of PCR; l, number of cycles in the PCR amplification step; m, number of experimental cycles.

First let us just consider one group of molecules, all with the same sequence, and denote the corresponding parameters without subscripts. Under the above assumptions we have the following theorem. This theorem gives the generating function, expectation, variance and the limit behavior of the number of DNA molecules after m selection–amplification cycles.

Theorem (1)

Suppose initially we have one molecule. Let N_m be the number of DNA molecules after m selection-amplification cycles (including selection and l PCR amplification cycles). Then we have

(1) The generating function $g_m(x)$ of N_m satisfies the following recursive equations:

$$g_0(x) = x,$$

$$g_{m+1}(x) = g_m(fh_l(x) + \bar{f}), \quad m = 0, 1, 2, \ldots,$$

and h_l is defined recursively by:

$$h_0(x) = x,$$

$$h_{l+1}(x) = h_l(\lambda x^2 + \bar{\lambda} x), \quad l = 0, 1, 2, 3, \ldots,$$

where $\bar{f} = 1 - f$, $\bar{\lambda} = 1 - \lambda$.

(2) The expectation and variance of N_m are:

$$E(N_m) = A^m,$$

$$Var(N_m) = \frac{Var(N_1)A^{m-1}(A^m - 1)}{A - 1},$$

where:

$$A = f(1 + \lambda)^l,$$

$$Var(N_1) = f(1 + \lambda)^{l-1}[(2 - f - f\lambda)(1 + \lambda)^l + \lambda - 1].$$

(3) If $A = f(1 + \lambda)^l > 1$, then N_m/A^m converges with probability 1 and in mean square to a random variable W_l as m tends to infinity, and:

$$E(W_l) = 1, \quad Var(W_l) = Var(N_1)/(A^2 - A).$$

(4) $P(W_l = 0) = q_l$ decreases with respect to l and $\lim_{l \to \infty} q_l = \bar{f}$. That is when l is sufficiently large, Z_m vanishes with probability approximately $\bar{f}$.

Now we consider the general case. The next corollary describes the results of letting the number m of selection-amplification cycles tend to infinity. The first result gives the probability that DNA molecules from groups $i = 1, 2, \ldots, \mathscr{I}$, ($[1, \mathscr{I}]$) vanish from the subsequent population of molecules, and the second result gives the probability that a randomly chosen molecule from this population is from groups $i \in [1, \mathscr{I}]$, given that DNA molecules do not vanish after the experiment as first m and then l tend to infinity. Suppose, without loss of generality, that $f_1 > f_2 > \ldots > f_N$. Let $N_m^{(i)}$ be the number of group i molecules after m selection-amplification cycles and for any $\mathscr{I}$,

$$S_m^{(\mathscr{I})} = \sum_{i=0}^{\mathscr{I}} N_m^{(i)}.$$

Then we have the corollary:

Corollary (1)

Let n_i be the initial number of group i molecules, f_i be the probability that an group i molecule being selected in one selection step, and λ be the efficiency of PCR. Then

(1) The probability that DNA molecules from groups $[1, \mathscr{I}]$ vanish in the subsequent population satisfies:

$$\lim_{l \to \infty} \lim_{m \to \infty} P\{S_m^{(\mathscr{I})} = 0\} = \prod_{i=1}^{\mathscr{I}} (1 - f_i)^{n_i}.$$

(2) Under the condition that $S_m^{(N)} \neq 0$ for all m:

$$\lim_{m \to \infty} S_m^{(\mathscr{I})} / S_m^{(N)} = \begin{cases} 1, & S_m^{(\mathscr{I})} \neq 0, \text{ for all } m. \\ 0, & \text{otherwise.} \end{cases}$$

Thus given $S_m^{(N)} \neq 0$ for all m, the probability that a randomly chosen molecule is from groups $[1, \mathscr{I}]$ tends to

$$\frac{1 - \prod_{i=1}^{\mathscr{I}} (1 - f_i)^{n_i}}{1 - \prod_{i=1}^{N} (1 - f_i)^{n_i}}$$

as m and then l tend to infinity.

The Case of a Large Initial Number of Molecules

In practice, biologists have usually done three to eight selection-amplification cycles (Kinzler & Vogelstein, 1989, 1990; Murtin et al., 1989; Blackwell & Weintraub, 1990; Thiesen & Bach, 1990; Pollock & Treisman, 1990; Rebar & Pabo, 1994). Theoretical analysis as m tends to infinity can not be applied in this case. In this section, we will study the case where the initial number of molecules in each group is large and the number of cycles is relatively small. We will also model the binding constant of a DNA molecule and analyze the connection between the binding constants and the selection process.

In theorem (1) we obtained the expectation and the variance of the number of molecules generated from one initial molecule after m selection-amplification cycles. Each molecule is selected and amplified independently by our assumption. Therefore, if we start from a large number n_i of group i molecules, from the strong law of large numbers, we have:

$$\lim_{n_i \to \infty} \frac{N_m^{(i)}}{n_i} = [f_i(1 + \lambda)^l]^m.$$

Therefore, when n_i is large, we can approximate $N_m^{(i)}$ by $n_i[f_i(1 + \lambda)^l]^m$. This approximation is good only when n_i is relatively large compared to

$$\frac{f_i(1 + \lambda)^{l-1}[(2 - f_i - f_i\lambda)(1 + \lambda)^l + \lambda - 1]A_i^{m-1}(A_i^m - 1)}{A_i - 1}$$

where $A_i = f_i(1 + \lambda)^l$. This condition is hard to obtain in practice.

If the above conditions are satisfied for each group of molecules, we can use $n_i[f_i(1 + \lambda)^l]^m$ to approximate $N_m^{(i)}$. Then the probability that a randomly chosen molecule in the final product after

m selection–amplification cycles is from groups $[1, \mathcal{I}]$ is approximately:

$$P(\mathcal{I}) = E \frac{\sum\limits_{i=1}^{\mathcal{I}} N_m^{(i)}}{\sum\limits_{i=1}^{N} N_m^{(i)}} \approx \frac{\sum\limits_{i=1}^{\mathcal{I}} n_i [f_i(1 + \lambda)^l]^m}{\sum\limits_{i=1}^{N} n_i [f_i(1 + \lambda)^l]^m} = \frac{\sum\limits_{i=1}^{\mathcal{I}} n_i f_i^m}{\sum\limits_{i=1}^{N} n_i f_i^m}. \quad (2)$$

From now on, we will use the following assumption.

Assumption (A)

The probability that a randomly chosen molecule in the final product is from groups $[1, \mathcal{I}]$ is given by equation (2).

From equation (1), we have the relationship between f_i and the binding constant K_i:

$$f_i = \frac{K_i p}{1 + K_i p},$$

where p is the concentration of free proteins at equilibrium. Substituting the above equation into equation (2), we have the probability that a randomly chosen molecule in the final product is in groups $[1, \mathcal{I}]$:

$$P(\mathcal{I}, p) = \frac{\sum\limits_{i=1}^{\mathcal{I}} n_i (K_i p / (1 + K_i p))^m}{\sum\limits_{i=1}^{N} n_i (K_i p / (1 + K_i p))^m}. \quad (3)$$

Differentiating with respect to p, we have:

$$(P(\mathcal{I}, p))'_p$$

$$= C \left[\sum_{i=1}^{\mathcal{I}} n_i \left(\frac{K_i p}{1 + K_i p} \right)^m \frac{1}{1 + K_i p} \sum_{j=1}^{N} n_j \left(\frac{K_j p}{1 + K_j p} \right)^m \right.$$

$$- \sum_{i=1}^{\mathcal{I}} n_i \left(\frac{K_i p}{1 + K_i p} \right)^m \sum_{j=1}^{N} n_j \left(\frac{K_j p}{1 + K_j p} \right)^m \frac{1}{1 + K_j p} \right]$$

$$= C \left[\sum_{i=1}^{\mathcal{I}} \sum_{j=\mathcal{I}+1}^{N} n_i n_j \left(\frac{K_i p}{1 + K_i p} \right)^m \left(\frac{K_j p}{1 + K_j p} \right)^m \right.$$

$$\times \left. \left[\frac{1}{1 + K_i p} - \frac{1}{1 + K_j p} \right] \right]$$

$$\leqslant 0,$$

where

$$C = \frac{m}{p} \left(\sum_{i=1}^{N} n_i \left(\frac{K_i p}{1 + K_i p} \right)^m \right)^{-2}.$$

Therefore, $P(\mathcal{I}, p)$ is a decreasing function of p and approaches its maximum of

$$\frac{\sum\limits_{i=1}^{\mathcal{I}} n_i K_i^m}{\sum\limits_{i=1}^{N} n_i K_i^m}$$

as p tends to 0. Our goal in a selection experiment is to enrich the DNA molecules with high binding affinity. Because $P(\mathcal{I}, p)$ is maximized as p goes to 0, in a binding experiment experimental conditions should be carefully designed so that the concentration p of free proteins is small. Letting p tend to 0 in equation (3), it follows that:

$$P(\mathcal{I}) = \lim_{p \to 0} P(\mathcal{I}, p) = \frac{\sum\limits_{i=1}^{\mathcal{I}} n_i K_i^m}{\sum\limits_{i=1}^{N} n_i K_i^m}. \quad (4)$$

In the following, we will use equation (4) as the probability that a randomly chosen sequence in the final product is from groups $[1, \mathcal{I}]$.

Modeling the Binding Constant

For a specific target protein, let us suppose that a DNA sequence $I = i_1 i_2 \ldots i_k$ of length k in base-pairs has an equilibrium binding constant (Cantor & Schimmel, 1980; Koblan *et al.*, 1992):

$$K_I = \exp(-G(I)/(kT)) = \exp(-\beta G(I)),$$

where T is the temperature and $G(I)$ is the total free binding energy. If $I^{(0)} = i_1^{(0)} i_2^{(0)} \ldots i_k^{(0)}$ and $G(I^{(0)}) = \min_I G(I)$, then $I^{(0)}$ is said to belong to the consensus sequence for the target protein. We use this terminology because binding sites are often of the form $ARYGR \ldots T$, but note that here consensus sequence just means the set of sequences with minimum free energy of binding. In the following, we assume that in the initial library, every DNA molecule is represented an equal number of times; that is, if n_I and n_J are the numbers of DNA molecules with sequences I and J, respectively, $n_I = n_J$ for any I and J. Under the above assumptions, we have the next theorem.

Theorem (2)

Suppose there is a unique consensus sequence $I^{(0)} = i_1^{(0)} i_2^{(0)} \ldots i_k^{(0)}$. Let $P_{con}^{(m)}$ be the probability that a randomly chosen molecule whose sequence is the

unique consensus sequence after m selection–amplification cycles. Then the following hold:

(1) Assume that all the positions contribute independently in the sense that the energy function $G(I)$ satisfies:

$$G(I) - G(I^{(0)}) = \sum_{\mu-1}^{k} \sigma_\mu(i_\mu),$$

where $\sigma_\mu(i)$ is a function on $\{A, C, G, T\}$. Then:

$$(P_{con}^{(m)})^{-1} = \prod_{\mu=1}^{k} (\exp(-m\beta\sigma_\mu(A)) + \exp(-m\beta\sigma_\mu(C))$$

$$+\exp(-m\beta\sigma_\mu(G)) + \exp(-m\beta\sigma_\mu(T))).$$

In particular, if the binding energy is independent of state i different from the consensus, that is $\sigma_\mu(i) = \delta/(m\beta)I(i \neq i_\mu^{(0)})$, then:

$$(P_{con}^{(m)})^{-1} = (1 + 3\exp(-\delta))^k.$$

(2) Suppose that the bases in DNA molecules have nearest-neighbor interactions in the sense that the energy function satisfies:

$$G(I) - G(I^{(0)}) = \sum_{\mu=1}^{k} \sigma_\mu(i_\mu) + \sum_{\mu=1}^{k-1} \sigma_\mu(i_\mu i_{\mu+1}).$$

Further we assume that:

$$\sigma_\mu(i_\mu) = \delta/(m\beta)I(i_\mu \neq i_\mu^{(0)}),$$

$$\sigma_\mu(i_\mu i_{\mu+1}) = \begin{cases} (\delta_1 - \delta)/(m\beta), & i_\mu \neq i_\mu^{(0)}, i_{\mu+1} \neq i_{\mu+1}^{(0)}, \\ (\delta_2 - \delta/2)/(m\beta), & \begin{cases} i_\mu = i_\mu^{(0)}, i_{\mu+1} \neq i_{\mu+1}^{(0)}, \text{ or} \\ i_\mu \neq i_\mu^{(0)}, i_{\mu+1} = i_{\mu+1}^{(0)}, \end{cases} \\ 0, & i_\mu = i_\mu^{(0)}, \quad i_{\mu+1} = i_{\mu+1}^{(0)}. \end{cases}$$

Then:

$$(P_{con}^{(m)})^{-1} = \frac{1}{a_2 - a_1}[(1 + a_2\exp(-\delta_2))^{k-2}(g_1 + a_2\hat{g}_1)$$

$$\times (a_1 - 3\exp(-\delta/2)) - (1 + a_1\exp(-\delta_2))^{k-2}$$

$$\times (g_1 + a_1\hat{g}_1)(a_2 - 3\exp(-\delta/2))].$$

where $a_1 < a_2$ are the two solutions of:

$$a^2 + \exp(\delta_2)(1 - 3\exp(-\delta_1))a - 3 = 0,$$

and:

$$g_1 = 1 + 3\exp(-(\delta/2)),$$

$$\hat{g}_1 = \exp(-\delta_2) + 3\exp(-(\delta_1 + \delta/2)).$$

Remark (1)

The conditions in part (1) of theorem (2) assume that each base contributes independently, and this particular condition means that if a base in a sequence differs from the consensus sequence, the energy is increased by a constant $\delta/(m\beta)$. The conditions in part (2) assume that, apart from the independent contributions, there are nearest-neighbor interactions. For nearest-neighbor interactions,

we assume that if the two bases differ from the consensus by one base, the energy is increased by a constant $(\delta_2 - \delta/2)/(m\beta)$ and if the two bases are both different from the consensus sequence, the energy is increased by a constant $(\delta_1 - \delta)/(m\beta)$. We choose this parameterization to simplify the final formula.

Remark (2)

Although in reality the base sequence contributes to the free energy in a complex fashion, we choose here to make a strongly simplifying assumption to permit a closed expression to be derived. Further, more realistic modes of contribution to the free energy can be taken into account by extending the above treatment.

Implications and Limitations

Many *in vitro* selection experiments have been performed to study the DNA binding properties of different proteins (for example, see Kinzler & Vogelstein, 1989, 1990; Blackwell & Weintraub, 1990; Thiesen & Bach, 1990; Pollock & Treisman, 1990; Rebar & Pabo, 1994). The experimental conditions depend on the properties of the protein to be investigated and are determined empirically. In this section, we examine the effects of experimental conditions according to our model in order to understand the power and limitations of *in vitro* selection–amplification techniques. For simplicity, we consider here only two groups of molecules. The molecules in the first group have binding probability f_1 and the molecules in the second group have binding probability f_2, where $f_1 > f_2$. We will use the notation and parameters introduced above.

First let us examine the effect of the number of PCR cycles in the PCR amplification step. Without any amplification, i.e. $l = 0$, a initial group i molecule is retained with probability f_i. It is retained after m selection–amplification cycles if, and only if, it is retained after each cycle. Therefore, the probability is f_i^m. The molecule vanishes after m selection cycles with probability $1 - f_i^m$. If the initial number of group i molecules is n_i, all the molecules, both group (1) and group (2), vanish after m selection cycles with probability $(1 - f_1^m)^{n_1}(1 - f_2^m)^{n_2}$. As we have noted above, in order to have high probability of selecting the best binding molecules during *in vitro* selection–amplification, we should let the concentration of free proteins be small. Then, from equation (1), the corresponding binding probability f_i will also be small. From the above we see that, with high probability, no molecules are retained. PCR therefore must play an essential role in the *in vitro* selection–amplification techniques. With PCR, the selected molecules are amplified exponentially between selections. Because we have many molecules after amplification, the probability that they all vanish is small. Therefore, whether we can obtain any molecules that bind the protein after

the experiment depends heavily on the first selection cycle. If the initial number of molecules is small, the probability that no molecules are selected can still be close to 1. We see that in order that the experiments be successful, the initial number of molecules should be large and that PCR must be used to amplify the selected molecules.

How can the protein concentration best be set in *in vitro* selection experiments? On one hand, we want to be able to select the best binding molecules with high probability. On the other hand, we want a high degree of discrimination and we do not want all the molecules to bind to the protein. Let us determine the concentration of free proteins as a function of the number of PCR cycles through a series of selection–amplification stages. Because the DNA molecules are amplified on average $(1 + \lambda)^l$ times after l PCR cycles and group (1) molecules are selected with probability f_1, the expected number of selected group (1) molecules is proportional to $f_1(1 + \lambda)^l$ after a single selection–amplification cycle. We should design the experimental conditions so that f_1 is on the same scale as $(1 + \lambda)^{-l}$. If f_1 is too small, the group (1) molecules are lost with high probability. Furthermore, background effects will dominate and specificity will be lost. If f_1 is too large, many DNA molecules will be selected, including many of those not in group (1). Therefore, f_1 should be proportional to $(1 + \lambda)^{-l}$. As noted above, we should let the concentration of free proteins be small and thus f_1 be small. Therefore, we should use as many PCR cycles as possible, but not too many to avoid PCR artifacts. Because of the other constraints, such as convenience of experimental conditions and mutations produced during PCR, biologists often do 20 to 30 PCR cycles.

We did Monte Carlo simulations to study the effects of the number of PCR cycles (l), the number of selection–amplification cycles (m), and the ratio between the two binding probabilities of the two groups of molecules (f_1/f_2). In all these simulations we chose f_1 such that $f_1(1 + \lambda)^l = 2$ so that each selection–amplification cycle essentially doubles the group (1) molecules. We chose $\lambda = 0.9$ and $n_1 = n_2 = (1 + \lambda)^l$ so that the expected number of selected group (1) molecules is two after the first selection. For each choice of (l, m, f_1, f_2), we did 5000 simulated selection–amplification experiments by Monte Carlo. After each Monte Carlo simulated experiment, we observe the number of group (1) and group (2) molecules.

First we studied the effect of the number of PCR cycles. We chose $m = 2$ selection–amplification cycles and $f_1/f_2 = 5$. (For each value of l, $f_1 = 2 \times (1 + 0.9)^{-l}$ and $f_2 = f_1/5$.) Figure 2(a) and (b) show the histogram for the proportion of the first group molecules for $l = 5$ (a) and $l = 20$ (b). These two Figures are almost the same. This shows that the PCR amplification step has the principal effect of rescuing the selected sequences. But we note that the f_1 corresponding to $l = 5$ is much larger than that corresponding to $l = 20$. From the relationship between binding probability f and the concentration

of free proteins (1), we see the free protein concentration for $l = 5$ is much higher than that for $l = 20$.

Next we examined the number of selection–amplification cycles. We chose $l = 20$ as has often been done in published experiments. Other conditions were the same as in the above simulations except that we chose $m = 1$. Figure 2(c) shows the histogram for the proportion of the first group sequences. Comparing Figure 2(c) with Figure 2(a) we see that doing one more selection greatly increased the efficiency of the experiment. Under our simulation conditions, the average proportion of the group (1) molecules after one selection–amplification cycle is only $0.836 \approx f_1/(f_1 + f_2)$ with a standard deviation 0.27. One more cycle increased the average to $0.940 \approx f_1^2/(f_1^2 + f_2^2)$ with a standard deviation 0.19.

Finally, we studied the effect of the ratio f_1/f_2. We chose $l = 20$, $m = 2$, and $f_1/f_2 = 2$. Figure 2(d) shows the histogram for the proportion of the group (1) molecules. Comparing Figure 2(d) with Figure 2(a), we see that the ratio f_1/f_2 plays a sensitive role in the final results of the experiment. If f_1/f_2 is large, then it is relatively easy to select the best binding molecules. Under our simulation conditions, after two selection–amplification cycles, the average proportion of the group (1) molecules is 0.758 with a standard deviation 0.32 for $f_1/f_2 = 2$, compared to the average proportion of the group (1) molecules 0.94 with a standard deviation 0.19 for $f_1/f_2 = 5$.

In practice, f_1/f_2 is usually between 10 and 100. In this range, we can almost surely select the group (1) molecules after *in vitro* selection–amplification experiments. We choose the above parameters to make our analysis much more clear.

Discussion

We have analyzed here the quantitative relationships defining the often repeated experimental procedure of using binding and amplification cycles to find the specific DNA sequences that bind well to a particular protein. There are many published examples of DNA–protein interactions that have been analyzed in this fashion. The analysis is also applicable to a number of potential procedures that are based on the binding–amplification cycle to find one or a small number of double-stranded DNA, RNA or single-stranded DNA molecules having a specific selected property. The relationship presented here should provide the basis for a quantitative analysis of these experiments.

To some extent, the present analysis was inspired by experiments in which we have previously examined the IHF (integration host factor) DNA consensus binding patterns by selection experiments. The protein IHF of *Escherichia coli* is a histone-like protein which was discovered through its requirement for the integration of bacteriophage λ into the bacterial chromosome (Williams *et al.*, 1977). Unlike other histone-like proteins, IHF binds to DNA in a sequence-specific fashion making
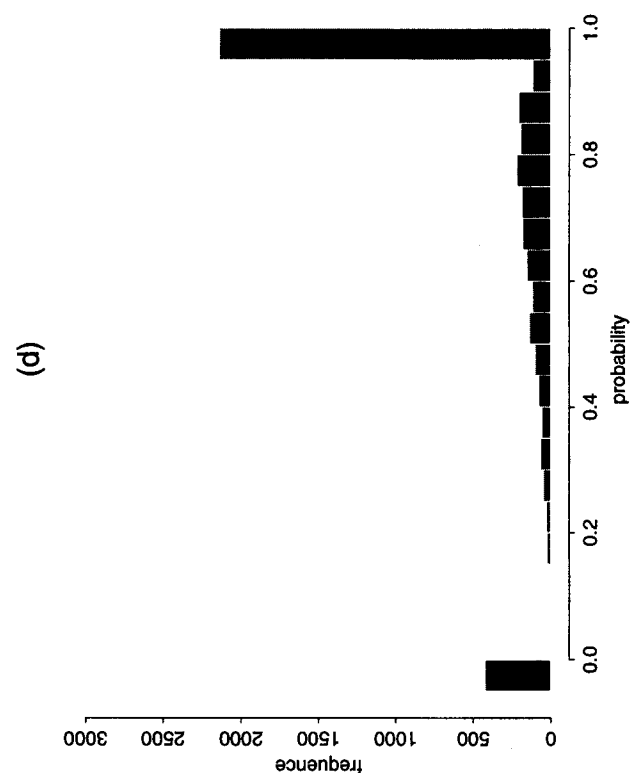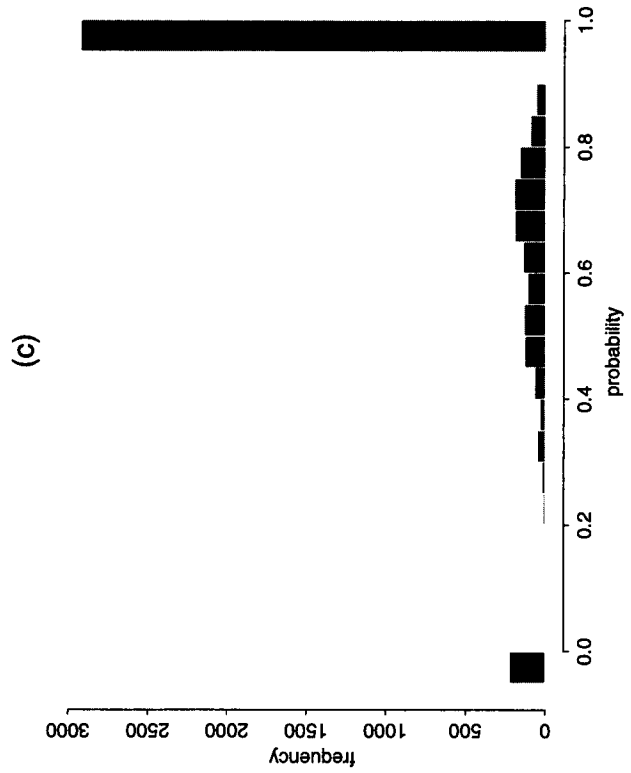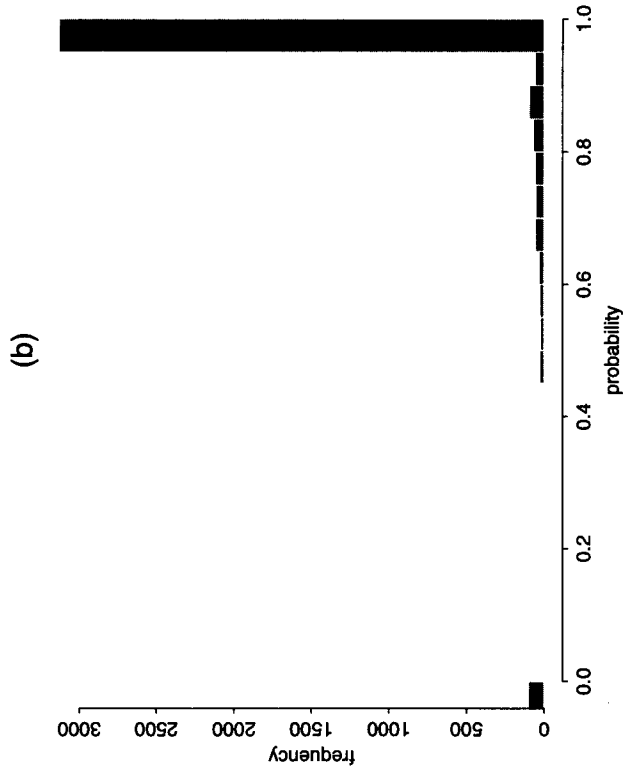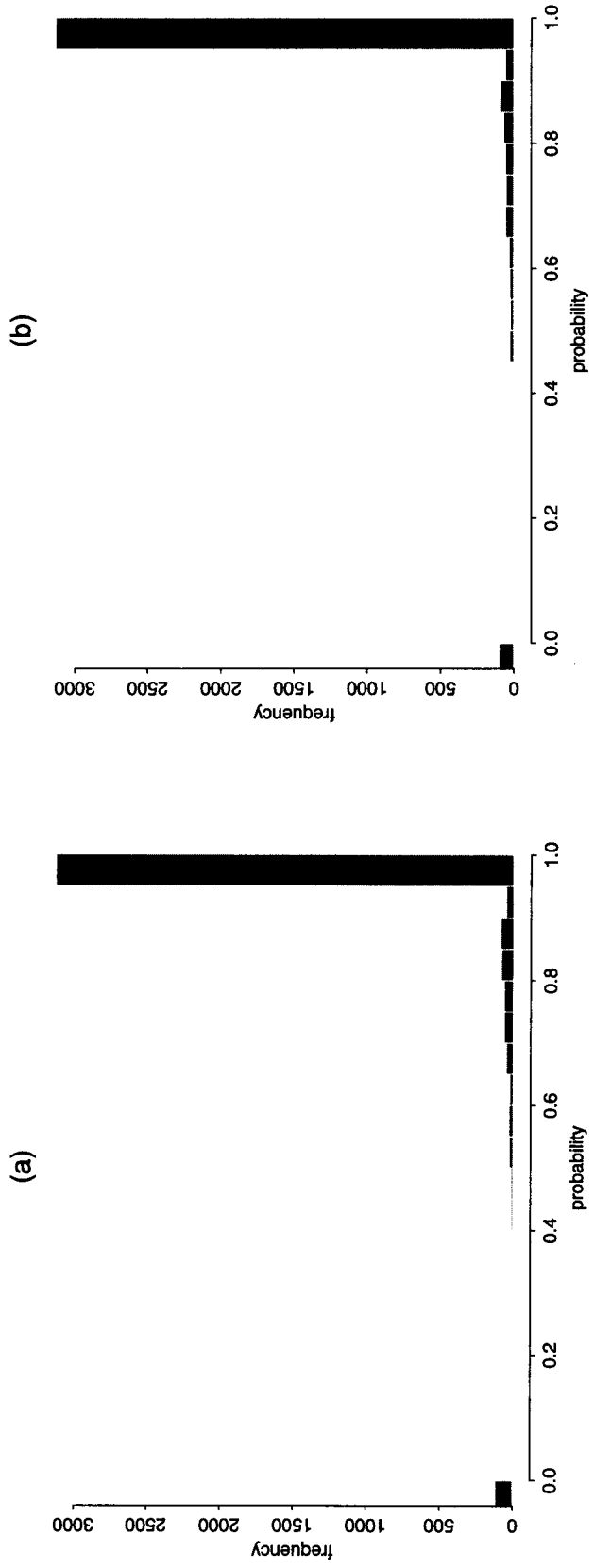
Figure 2. Histogram of the fraction of the first group molecules with: (a) $l = 20$, $m = 2$, $f_1/f_2 = 5$; (b) $l = 5$, $m = 2$, $f_1/f_2 = 5$; (c) $l = 20$, $m = 1$, $f_1/f_2 = 5$; (d) $l = 20$, $m = 2$, $f_1/f_2 = 2$ (5000 replications).

most of its sequence-specific contacts in the minor groove of DNA. DNA molecules that bound IHF were selected and consensus sequences were determined using the *r-tide* program (Galas *et al.*, 1985). We were surprised at the wide range of sequence variations present among the selected molecules, but found the consensus sequences were as expected from naturally occurring variants of the IHF binding sites. A preliminary account of this work was presented at the Keystone UCLA workshop on "The Polymerase Chain Reaction" (Murtin *et al.*, 1989).

As indicated in the Introduction, there are several distinct types of molecular selection–amplification experiments reported in the literature. The specific type of experiment described and analyzed here, sequence-specific protein–DNA interactions, is not by any means the only type to which the results derived here can be applied. These different types of experiment fall generally into three classes: (1) DNA and RNA equilibrium binding experiments, in which selection is determined by the differential binding properties at equilibrium (for example see Murtin *et al.*, 1989; Kinzler & Vogelstein, 1989, 1990; Blackwell & Weintraub, 1990; Pollock & Treisman, 1990; Rebar & Pabo, 1994); (2) DNA and RNA binding experiments in which selection is based on differential on and/or off rates not in equilibrium, as in a column binding experiment (Tuerk & Gold, 1990; Lin *et al.*, 1994); and (3) DNA and RNA cleavage-release experiments in which selection is determined by the reaction rates of the DNA or RNA catalyst (Joyce, 1989a,b; Beaudry & Joyce, 1992; Bartel & Szostak, 1993; Breaker & Joyce, 1994). There are two experimental characteristics that effectively define these different classes of selection–amplification experiments, and it is these characteristics that determine how our results can be applied. The first of these is equilibrium. Our analysis uses the simplifications made possible by the equilibrium assumption to express the probability of isolating the strongest binding molecules in terms of the set of equilibrium binding constants. These considerations will apply equally to DNA or RNA, double or single-stranded. When we shift to a non-equilibrium situation, however, as in class (2), above, the formula for K no longer applies, and f must be calculated for the specific experimental situation, but the remainder of the analysis carries through.

For the class (3) cleavage-release type of experiment f must now be interpreted to be the probability that, during the reaction time allowed in the experiment, a molecule in group i is released. This is in turn directly related to the catalytic constant specified by the sequence of group i. This constant plays essentially the same role as the specific binding affinity, but the time interval allowed for the cleavage reaction plays the role of the protein concentration in the equilibrium binding experiments. To get the maximum specificity, the time interval must be kept short, so that relatively few molecules achieve cleavage during the interval.

The experiment is not in equilibrium, of course, so that the relationship has the same monotonicity, but is not precisely the same relation. This similarity can be best seen in the extremes. If the protein concentration or the time interval is extremely long, then a large fraction of the molecules are selected whatever the spectrum of specificities. If the interval is extremely short, then very few molecules are selected, making the experiment difficult or impossible, but those are selected primarily by their specific binding or catalytic properties.

The relationships derived in this paper for the relatively simple case of equilibrium selection can be extended to encompass all of the above experimental situations and probably many others for which the selection–amplification process is central to the experiment. This extension of course depends on constructing a model of specific experimental situation, such as we have given for protein–DNA interactions.

References

Bartel, D. P. & Szostak, J. W. (1993). Isolation of new ribozymes from a large pool of random sequences. *Science*, **261**, 1411–1418.

Beaudry, A. A. & Joyce, G. F. (1992). Directed evolution of an RNA enzyme. *Science*, **257**, 635–641.

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.

Breaker, R. R. & Joyce, G. F. (1994). Emergence of a replicating species from an *in vitro* RNA evolution reaction. *Proc. Natl Acad. Sci. USA*, **91**, 6093–6097.

Cantor, C. R. & Schimmel, P. R. (1980). *Biophysical Chemistry*. W. H. Freeman, San Francisco.

Cuenoud, B. & Szostak, J. (1995). A DNA metallo-enzyme with DNA ligase activity. *Nature*, **375**, 611–614.

Galas, D. J., Eggert, M. & Waterman, M. S. (1985). Rigorous pattern-recognition methods for DNA sequences. Analysis of promoter sequences from *Escherichia coli*. *J. Mol. Biol.* **186**, 117–128.

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

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

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.

Koblan, K. S., Bain, D. L., Beckett, D., Shea, M. D. & Ackers, G. K. (1992). Analysis of site-specific

interaction parameters in protein–DNA complexes. *Methods Enzymol.* **210**, 405–425.

Lehman, N. & Joyce, G. F. (1993). Evolution *in vitro* of an RNA enzyme with altered matal dependence. *Nature*, **361**, 182–185.

Lin C. W., Hanna, O. & Szostak, J. W. (1994). Evidence that the guanosine substrate of the *Tetrahymena* ribozyme is bound in the anti conformation and that N7 contributes to binding. *Biochemistry*, **33**, 2703–2707.

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

Murtin, M., Galas, D. J., Arnheim, N. & Prentki, P. (1989). *In vitro* selection of protein binding sites: IHF-DNA interaction. *UCLA workshop: the Polymerase Chain Reaction*, April 1989, Keystone, Colorado.

Peterson, R. D., Bartel, D. P., Szostak, J. W., Horvath, S. J. & Feigon, J. (1994). ¹H NMR studies of the high-affinity Rev binding site of the Rev responsive element of HIV-1 mRNA: base pairing in the core binding element. *Biochemistry*, **33**, 5357–5366.

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

Prudent, J. R, Uno, T. & Schultz P. G. (1994). Expanding the scope of RNA catalysis. *Science*, **264**, 1924–1927.

Rebar, E. J. & Pabo. C. O. (1994). Zinc finger phage: affinity selection of fingers with new DNA-binding specificities. *Science*, **163**, 671–673.

Sun, F. (1995). The polymerase chain reaction and branching processes. *J. Comput. Biol.* **2**, 63–86.

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*, **250**, 1149–1151.

Williams, J. G. K., Wulff, D. L. & Nash, H. A. (1977). In *DNA Insertion Elements, Plasmids and Episomes.* (Bukhari, A., Shapiro, J. & Adhya, S., eds), pp. 357–361. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Wilson, C. & Szostak, J. W. (1995). *In vitro* evolution of a self-alkylating ribozyme. *Nature*, **374**, 777–782.

Appendix

Proof of theorem (1)

After the first selection step, a DNA molecule is selected with probability f and not selected with probability $\bar{f} = 1 - f$. Let S_1 be the number of molecules selected. Then the generating function of S_1 is:

$$Ex^{S_1} = fx + \bar{f},$$

and

$$E(S_1) = f, \quad Var(S_1) = f\bar{f}.$$

In the PCR amplification step, the number of DNA molecules T_l generated from one molecule after l PCR cycles is a branching process. After one PCR cycle, a DNA molecule generates two copies with probability λ and one copy with probability

$1 - \lambda$. Therefore, the generating function of T_1 is $\lambda x^2 + \bar{\lambda}x$. From the general theory of branching processes, we know the generating function $h_1(x)$ of T_1 satisfies the following recursive equation:

$$h_0(x) = x,$$

$$h_{l+1}(x) = h_l(\lambda x^2 + \bar{\lambda}x) = \lambda h_l^2(x) + \bar{\lambda}h_l(x),$$

$$l = 0, 1, 2, \dots . \tag{A1}$$

The expectation and variance of T_l are:

$$E(T_l) = (1 + \lambda)^l,$$

$$Var(T_l) = (1 - \lambda)(1 + \lambda)^{l-1}((1 + \lambda)^l - 1).$$

Therefore, the number of DNA molecules after the first experimental cycle, N_1, is:

$$N_1 = \sum_{i=1}^{S_1} T_1^{(i)},$$

where $T_1^{(i)}$, $i = 1, 2, \dots$ are *i.i.d.* and have the same distribution as T_l. Thus, the generating function of N_1 is:

$$g_1(x) = E(x^{N_1}) = E\left(x^{\sum_{i=1}^{S_1} T_l^{(i)}}\right) = fh_l(x) + \bar{f}.$$

The expectation and variance of N_1 are:

$$E(N_1) = E(S_1)E(T_l) = f(1 + \lambda)^l,$$

and

$$Var(N_1) = E(S_1)Var(T_l) + Var(S_1)(E(T_l))^2$$

$$= f(1 - \lambda)(1 + \lambda)^{l-1}((1 + \lambda)^l - 1) + f\bar{f}(1 + \lambda)^{2l}$$

$$= f(1 + \lambda)^{l-1}[(2 - f - f\lambda)(1 + \lambda)^l + \lambda - 1].$$

After m selection–amplification cycles, the process of *in vitro* selection–amplification produces a random number of DNA molecules, N_m. Note that $\{N_m, m \geq 1\}$ forms a branching process. From the general theory of branching processes, the generating function $g_m(x)$ of T_m satisfies the following recursive equation:

$$g_0(x) = x,$$

$$g_{m+1}(x) = g_m(g_1(x)) = g_m(fh_l(x) + \bar{f}),$$

$$m = 0, 1, 2, \dots .$$

The expectation and variance of N_m are:

$$E(N_m) = (E(N_1))^m = f^m(1 + \lambda)^m,$$

$$Var(N_m) = \frac{Var(N_1)A^{m-1}(A^m - 1)}{A - 1}.$$

(1) and (2) of the theorem is proven.

Part (3) of the theorem follows directly from the standard theory of branching processes (Harris, 1963).

(4) For any l such that $f(1 + \lambda)^l \leq 1$, we have $q_l = 1$. If $f(1 + \lambda)^l > 1$, then $q_l < 1$ and q_l is the unique solution of $x = g_1^{(l)}(x) = \bar{f} + fh_l(x)$ in $(0,1)$, $x < g_1(x)$ when $x < q_l$ and $x > g_1(x)$ when $x > q_l$ (Harris, 1963).

Since $h_l(x)$ is increasing in x, we have:

$$h_{l+1}(x) = h_l(\lambda x^2 + \bar{\lambda}x) \leqslant h_l(\lambda x + \bar{\lambda}x) = h_l(x).$$

Thus:

$$q_{l+1} = g_1^{(l+1)}(q_{l+1}) = \bar{f} + f h_{l+1}(q_{l+1})$$

$$< \bar{f} + f h_l(q_{l+1}) = g_1^{(l)}(q_{l+1}). \quad (A2)$$

Therefore, $q_{l+1} < q_l$ and $\lim_{l \to \infty} q_l = c < 1$ exists. From equation (A2) we have:

$$c = \bar{f} + f \lim_{l \to \infty} h_l(q_l).$$

Next we prove $\lim_{l \to \infty} h_l(q_l) = 1$. First we prove $\lim_{l \to \infty} h_l(x) = 0$, for any $0 < x < 1$.

By the recursive equation for $h_l(x)$, $h_l(x)$ is decreasing in l and is bounded. Therefore, $\lim_{l \to \infty} h_l(x) = d < 1$ exist and from equation (A1) we have:

$$d = \lambda d^2 + \bar{\lambda}d.$$

Therefore, $d = 0$.

Now choose $\epsilon > 0$ such that $c + \epsilon < 1$. Then:

$$\lim_{l \to \infty} h_l(q_l) \leqslant \lim_{l \to \infty} h_l(c + \epsilon) = 0.$$

Therefore, we have $c = \bar{f}$ and theorem (1) is proven.

Proof of corollary (1)

(1) For each molecule in group i, it vanishes with probability $\bar{f}_i$ as m then l tends to infinity. Because all the molecules are selected and amplified independently, all the group $[1, \mathcal{I}]$ molecules vanish in subsequent population with probability

$$\prod_{i=1}^{\mathcal{I}} \bar{f}_i^{n_i}$$

as m then l tends to infinity.

(2) On the set that $S_m^{(\mathcal{I})} \neq 0$ for all m, there exists a minimum $i_0 \leqslant \mathcal{I}$ such that $N_m^{(i_0)} \neq 0$ for all m. From theorem (1) we have:

$$\frac{N_m^{(i_0)}}{(f_{i_0}(1 + \lambda)^l)^m} \to W_{i_0},$$

$$\frac{N_m^{(i)}}{(f_{i_0}(1 + \lambda)^l)^m)} \to 0, \quad i \neq i_0,$$

as m tends to infinity and $W_{i_0} \neq 0$ on the set that $N_m^{(i_0)}$ does not vanish. Therefore:

$$\lim_{m \to \infty} S_m^{(\mathcal{I})} / S_m^{(N)} = 1$$

i.e. on the set where $\{S_m^{(\mathcal{I})} \neq 0\}$, a randomly chosen molecule is from groups $[1, \mathcal{I}]$ almost surely as m tends to infinity. From part (1) of this corollary we have:

$$\lim_{l \to \infty} P\{S_m^{(\mathcal{I})} \neq 0 | S_m^{(N)} \neq 0\}$$

$$= \lim_{l \to \infty} \frac{P\{S_m^{(\mathcal{I})} \neq 0\}}{P\{S_m^{(N)} \neq 0\}} = \frac{1 - \prod_{i=1}^{\mathcal{I}} \bar{f}_i^{n_i}}{1 - \prod_{i=1}^{N} \bar{f}_i}{}_{n_i}.$$

Corollary (1) is proven.

Proof of theorem (2)

(1) From equation (4), in the main text, we have:

$$(P_{con}^{(m)})^{-1} = \frac{\sum_I K_I^m}{K_{con}^m}$$

$$= \sum_I \exp(-m\beta(G(I) - G(I^{(0)})))$$

$$= \sum_I \exp\left(-m\beta \sum_{\mu=1}^{k} \sigma_\mu(i_\mu)\right)$$

$$= \sum_I \prod_{\mu=1}^{k} \exp(-m\beta\sigma_\mu(i_\mu))$$

$$= \prod_{\mu=1}^{k} (\exp(-m\beta\sigma_\mu(A)) + \exp(-m\beta\sigma_\mu(C))$$

$$+ \exp(-m\beta\sigma_\mu(G)) + \exp(-m\beta\sigma_\mu(T))),$$

which proves (1).

(2) Let

$$\gamma_\mu(i_\mu i_{\mu+1}) = \frac{\sigma\mu(i_\mu) + \sigma_{\mu+1}(i_{\mu+1})}{2} + \sigma_\mu(i_\mu i_{\mu-1}),$$

$$\mu = 1, 2, \ldots, k - 1.$$

It is easy to check from the conditions of the theorem that:

$$\gamma_\mu(i_\mu i_{\mu+1}) = \begin{cases} \delta_1/(m\beta), & i_\mu \neq i_\mu^{(0)}, \quad i_{\mu+1} \neq i_{\mu+1}^{(0)}, \\ \delta_2/(m\beta), & \begin{cases} i_\mu = i_\mu^{(0)}, & i_{\mu+1} \neq i_{\mu+1}^{(0)}, \quad \text{or} \\ i_\mu \neq i_\mu^{(0)}, & i_{\mu+1} = i_{\mu+1}^{(0)}, \end{cases} \\ 0, & i_\mu = i_\mu^{(0)}, \quad i_{\mu+1} = i_{\mu+1}^{(0)}. \end{cases}$$

Then, from equation (4), in the main text, we have:

$$(P_{con}^{(m)})^{-1} = \sum_I \exp(-m\beta(G(I) - G(I^{(0)})))$$

$$= \sum_I \exp(m\beta((\sigma_1(i_1) + \sigma_k(i_k))/2 + \sum_{\mu=1}^{k-1} \gamma_\mu(i_\mu i_{\mu+1}))).$$

(A3)

In order to calculate $(P_{con}^{(m)})^{-1}$, we define, for any $1 \leqslant l \leqslant k - 1$:

$$g_l(i_{k-l}) = \sum_{i_{k-l+1}, \ldots, i_k} \exp(-m\beta(\sigma_k(i_k)/2$$

$$+ \sum_{\mu=k-l}^{k-1} \gamma_\mu(i_\mu i_{\mu+1}))), \quad (A4)$$

and

$$g_l = g_l(i_{k-l}^{(0)}), \quad \hat{g}_l = g_l(i_{k-l}), \quad i_{k-l} \neq i_{k-l}^{(0)}.$$

Then, we have:

$$g_1 = \sum_{i_k} \exp(-m\beta(\sigma_k(i_k)/2 + \gamma_\mu(i_{k-1}^{(0)}i_k)))$$

$$= \sum_{i_k \neq i_k^{(0)}} \exp(-m\beta(\sigma_k(i_k)/2 + \gamma_\mu(i_{k-1}^{(0)}i_k)))$$

$$+ \exp(-m\beta(\sigma_k(i_k^{(0)})/2 + \gamma_\mu(i_{k-1}^{(0)}i_k^{(0)})))$$

$$= 3\exp(-(\delta_2 + \delta/2)) + 1.$$

Similarly, we can prove:

$$\hat{g}_1 = 3\exp(-(\delta_1 + \delta/2)) + \exp(-\delta_2).$$

From equation (A4), we have for $2 \leq l \leq k - 1$:

$$g_l = \sum_{i_{k-l+1}} \exp(-m\beta(\gamma_{k-l}(i_{k-l}^{(0)}i_{k-l+1})g_{l-1}(i_{k-l+1})))$$

$$= 3\exp(-\delta_2)\hat{g}_{l-1} + g_{l-1}. \quad \text{(A5)}$$

Similarly,

$$\hat{g}_l = 3\exp(-\delta_1)\hat{g}_{l-1} + \exp(-\delta_2)g_{l-1}. \quad \text{A6}$$

In order to solve the system of difference equations (A5) and (A6), we want to find a number a and a corresponding number c such that:

$$g_l + a\hat{g}_l = c(g_{l-1} + a\hat{g}_{l-1}). \quad \text{(A7)}$$

Using equations (A5), (A6) and (A7) and setting the coefficients of g_{l-1} and $\hat{g}_{l-1}$ equal, we have:

$$\begin{cases} 1 + a\exp(-\delta_2) = c, \\ 3\exp(-\delta_2) + a3\exp(-\delta_1) = ac. \end{cases}$$

Therefore:

$$3\exp(-\delta_2) + 3a\exp(-\delta_1) = a(1 + a\exp(-\delta_2));$$

that is:

$$a^2 + \exp(\delta_2)(1 - 3\exp(\delta_1))a - 3 = 0. \quad \text{(A8)}$$

Equation (A8) has two solutions. Let $a_1 < a_1$ be the two solutions of equation (A8). Then from equation (A7) we have:

$$\begin{cases} g_l + a_1\hat{g}_l = (1 + a_1\exp(-\delta_2))(g_{l-1} + a_1\hat{g}_{l-1}) = (1 + a_1\exp(-\delta_2))^{l-1}(g_1 + a_1\hat{g}_1), \\ g_l + a_2\hat{g}_l = (1 + a_2\exp(-\delta_2))(g_{l-1} + a_2\hat{g}_{l-1}) = (1 + a_2\exp(-\delta_2))^{l-1}(g_1 + a_2\hat{g}_1). \end{cases}$$

Solving this system of equations, we have:

$$\begin{cases} g_l = \dfrac{1}{a_2 - a_1}[a_1(1 + a_2\exp(-\delta_2))^{l-1}(g_1 + a_2\hat{g}_1) - a_2(1 + a_1\exp(-\delta_2))^{l-1}(g_1 + a_1\hat{g}_1)], \\ \hat{g}_l = \dfrac{1}{a_2 - a_1}[(1 + a_1\exp(-\delta_2))^{l-1}(g_1 + a_1\hat{g}_1) - (1 + a_2\exp(-\delta_2))^{l-1}(g_1 + a_2\hat{g}_1)]. \end{cases} \quad \text{(A9)}$$

From equation (A3), we have:

$$(P_{con}^{(m)})^{-1} = \sum_{i_1} g_{k-1}(i_1)\exp(-m\beta\sigma_1(i_1)/2)$$

$$= g_{k-1} + 3\hat{g}_{k-1}\exp(-\delta/2). \quad \text{(A10)}$$

Combining equations (A9) and (A10), we complete the proof of theorem (2).

Reference

Harris, T. E. (1963). The Theory of Branching Process. Springer, Berlin.