ORGANIC LETTERS

2000
Vol. 2, No. 7
915−918

"Masterpiece" Copolymer Sequences by Targeted Equilibrium-Shifting

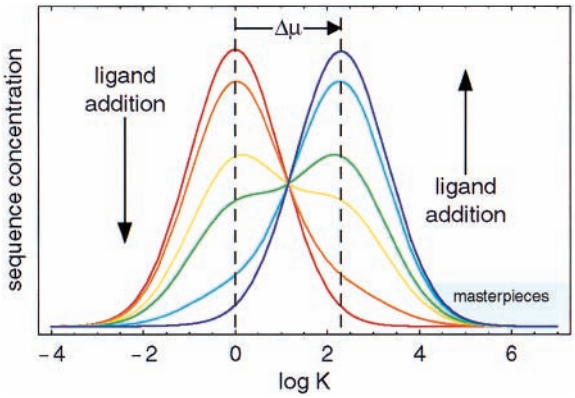
Jeffrey S. Moore* and Ned W. Zimmerman

School of Chemical Sciences, University of Illinois at Urbana-Champaign, 600 South Mathews Avenue, Urbana, Illinois 61801

moore@scs.uiuc.edu

Received January 20, 2000

ABSTRACT

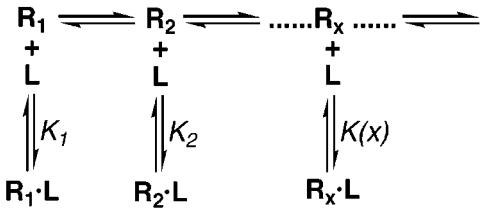


We describe an equilibrium model to determine whether a random population of dynamic copolymer sequences could be driven by molecular recognition to a subset of sequences that tightly bind a specific ligand. The model predicts that the population's mean binding constant can be shifted, but because of competitive binding, only to a limited degree (ca. 2 orders of magnitude larger than the original mean). True chemical evolution will require a mechanism for selection *and* amplification.

The success of combinatorial chemistry demonstrates the practical utility of a scientific approach in which synthesis is linked directly to function. Using the brute strength of parallel synthesis and screening, combinatorial methods bypass the need for understanding how structure gives rise to function. The payoff thus comes not from an immediate gain in structure-based knowledge but from the speed at which discoveries of a practical nature are made.

Recent chemical literature has begun to take these ideas one step further, by addressing systems in which the diversity is chemically dynamic.[1] As shown in Scheme 1, the concept is to replace parallel synthesis with processes involving reversible connections between components to generate a dynamic population of potential receptors ($\mathbf{R_1}$...$\mathbf{R_x}$). The population is then biased by the addition of a specific ligand, **L**, and a separation method is used to remove the high-affinity members of the population. Several variations on

Scheme 1. Targeted Equilibrium-Shifting



this general approach have appeared in the literature,[2] falling under headings such as receptor-driven ligand evolution, dynamic combinatorial chemistry, chemical evolution, and adaptive chemistry. In all cases the essential step is targeted equilibrium-shifting, most familiar to chemists as Le Châtelier's principle. So far this approach has been tested on simple systems consisting of just a few components—systems that could have easily been investigated by traditional synthesis or combinatorial methods. Although these studies demonstrated that receptors could be formed in high yield

(1) Reviews: (a) Lehn, J.-M. *Chem. Eur. J.* **1999**, *5*, 2455−2463. (b) Klekota, B.; Miller, B. L. *Trends Biotechnol.* **1999**, *17*, 205−209. (c) Ganesan, A. *Angew. Chem., Int. Ed.* **1998**, *37*, 2828−2831.

10.1021/ol0055723 CCC: $19.00
Published on Web 03/17/2000

by equilibrium-shifting, the true potential of dynamic diversity lies in the possibility of exploring systems of far greater complexity.

We became interested in knowing how useful targeted equilibrium-shifting would be in enhancing the fraction of "masterpiece" sequences that is present in a random population of dynamic copolymers. The masterpiece sequences, as Orgel has referred to them,[3] are the outliers of the population—the individual molecules whose properties fall on the extreme high-end tail of the distribution. For large populations such as those of random copolymer sequences, complexity arises because of competitive binding by the various members. The question arises as to whether the large ensemble of weakly binding "average sequences" will overwhelm the strong properties of the relatively small fraction of desirable masterpiece sequences. We are not aware of any attempts in the literature to address this question, either experimentally or theoretically.[4] The equilibrium model described here can thus serve as the basis on which further discussion and research may be approached.

As shown in Scheme 1, the model assumes a population of sequences that can reversibly interchange with one another. The individual members of the population, $\mathbf{R_x}$, bind to **L** with binding constant $K(x)$, defined in the usual way (eq 1). A distribution function, $f(x)$, characterizes the frequency of occurrence of sequences having a binding constant $K(x)$. The functional form of this distribution, as

$$K(x) = \frac{[\mathbf{R_x}\cdot\mathbf{L}]}{[\mathbf{R_x}][\mathbf{L}]} \tag{1}$$

well as the relationship between x and $K(x)$, will be discussed below. When **L** is added to the system, those sequences that become complexed will be removed from the original population. Following this perturbation, the system will return to equilibrium, meaning that those sequences left uncomplexed will reestablish the original distribution.

It is possible to quantify how the equilibrium shifts as a function of ligand concentration. Each sequence is characterized by a mass balance expression (eq 2)

$$[\mathbf{R_x}]_T = [\mathbf{R_x}] + [\mathbf{R_x}\cdot\mathbf{L}] \tag{2}$$

that sums the unbound and bound components. Here $[\mathbf{R}_x]_T$ is the total concentration of sequence $\mathbf{R_x}$, $[\mathbf{R_x}]$ is the concentration of sequence $\mathbf{R_x}$ that is uncomplexed, and $[\mathbf{R_x}\cdot\mathbf{L}]$ is the concentration of the complex formed between sequence $\mathbf{R_x}$ and **L**. Substituting from eq 1 gives the new mass balance relationship, eq 3,

$$[\mathbf{R_x}]_T = [\mathbf{R_x}](1 + K(x)[\mathbf{L}]) \tag{3}$$

where [**L**] is the concentration of uncomplexed ligand. Finally, we specify [T] as the total concentration of all sequences (uncomplexed and complexed) and [t] as the total concentration of all uncomplexed sequences as defined in eqs 4 and 5, respectively. From the definitions given above,

$$[\mathrm{T}] = \int [\mathbf{R_x}]_T \, dx \tag{4}$$

$$[\mathrm{t}] = \int [\mathbf{R_x}] \, dx \tag{5}$$

we can express $[\mathbf{R}_x]$ in terms of the distribution function $[\mathbf{R_x}] = [\mathrm{t}]\cdot f(x)$. Substitution of this relation into eq 3 gives the new mass balance relationship, eq 6. Equation 7 then

$$[\mathbf{R_x}]_T = [\mathrm{t}]f(x)(1 + K(x)[\mathbf{L}]) \tag{6}$$

$$g(x) = \frac{[\mathbf{R_x}]_T}{[\mathrm{T}]} = \frac{f(x)(1 + K(x)[\mathbf{L}])}{\int f(x)(1 + K(x)[\mathbf{L}]) \, dx} \tag{7}$$

provides $g(x)$, which gives the proportion of $\mathbf{R_x}$ once the system has been allowed to reequilibrate after adding **L**. Importantly, $g(x)$ is the new distribution function that results from shifting the equilibrium. Another useful quantity is the fraction of sequences with a binding constant greater than a particular cutoff value, K_{cutoff}. This would be obtained by integration of $g(x)$ as shown in eq 8.

$$G(x) = \frac{\int_{K>K_{\mathrm{cutoff}}} f(x)(1 + K(x)[\mathbf{L}]) \, dx}{\int f(x)(1 + K(x)[\mathbf{L}]) \, dx} \tag{8}$$

To proceed further we need a chemically sensible description of how the binding constants $K(x)$ might vary over a population of random sequences. From a compilation of cyclodextrin complex stabilities, Connors[5] has concluded that binding affinities are reasonably described as being normally distributed in log K. Of particular relevance to our discussion here, he logically argued that this behavior is universal and can be applied generally to any other noncovalent system. Connors writes, "... any chemically reasonably defined population of a noncovalent association process will have a maximum typical stability range of 5−6 orders of magnitude in the equilibrium constant, resulting in a standard deviation of about 1 log K unit. The mean value of the distribution will be determined by the inherent defining features of the population; *the standard deviation, however, is presumably controlled by the range of forces available from the noncovalent interactions*." Given this premise, $f(x)$ becomes

(2) (a) Goodwin, J. T.; Lynn, D. G. *J. Am. Chem. Soc.* **1992**, *114*, 9197−9198. (b) Swann, P. G.; Casanova, R. A.; Desai, A.; Frauenhoff, M. M.; Urbancic, M.; Slomczynska, U.; Hopfinger, A. J.; Le Breton, G. C.; Venton, D. L. *Biopolymers* **1996**, *40*, 617−625. (c) Huc, I.; Lehn, J.-M. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 2106−2110. (d) Brady, P. A.; Sanders, J. K. M. *J. Chem. Soc., Perkin Trans 1* **1997**, 3237−3253. (e) Hioki, H.; Still, W. C. *J. Org. Chem.* **1998**, *63*, 904−905. (f) Calama, M. C.; Hulst, R.; Fokkens, R.; Nibbering, N. M. M.; Timmerman, P.; Reinhoudt, D. N. *Chem. Commun.* **1998**, 1021−1022. (g) Eliseev, A. V.; Nelen, M. I. *Chem. Eur. J.* **1998**, *4*, 825−834. (h) Klekota, B.; Miller, B. L. *Tetrahedron* **1999**, *55*, 11687−11697.

(3) Orgel, L. E. *Acc. Chem. Res.* **1995**, *28*, 109−118.

(4) Eliseev and Nelen (ref 2g) considered the case of a two-state receptor population with ligand−receptor association constants of K_{strong} and K_{weak}.

(5) Connors, K. A. *Chem. Rev.* **1997**, *97*, 1325−1357.

the normal distribution function (eq 9) with standard deviation $\sigma = 1$ and $x = \log K$ (i.e., $K(x) = 10^x$).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{(-(x-\mu)^2/2\sigma^2)} \qquad (9)$$

Next, we estimate what one might expect for a typical value of the mean, μ, using a well-studied family of copolymers. Aptamers are random sequences of RNA that can be selected and amplified for binding and catalytic activity. In the early 1990s Szostak and co-workers showed[6] that approximately 1 in 10^{10} random sequence RNA 100-mers can fold into structures capable of binding specifically to a targeted ligand (e.g., organic dyes) with $K = 10^5 - 10^7$ M^{-1}. We will assume that Connor's premise applies to aptamers. For a normal distribution with a standard deviation of 1, an upper tail area corresponding to $1/10^{10}$ of the total above $\log K = 6$ fixes μ to a value very close to zero (i.e., $\langle K \rangle = 1$ M^{-1}). Thus, for the purposes of the discussion that follows, we will use a value of $\mu = 0$. However, it should be noted that the main conclusions are not altered by the choice of μ.

The main conclusion from the model presented here is that the distribution of equilibrium constants can be shifted, but only to a limited degree. Substitution of eq 9 into eq 7 gives the equilibrium-shifted distribution, plotted in Figure 1 at different values of free ligand concentration. At

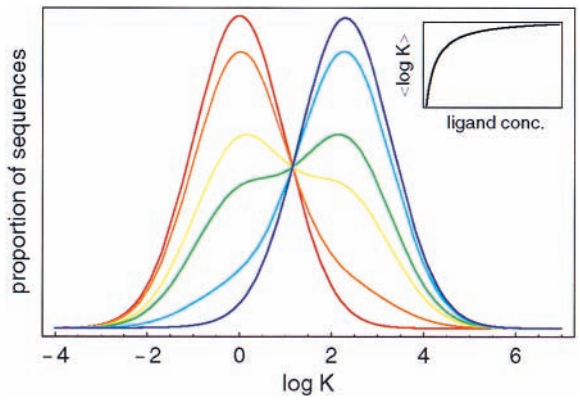


Figure 1. A random population of dynamic copolymer sequences is assumed to complex ligand **L** with binding constants that are normally distributed in $\log K$ ($\mu = 0$, $\sigma = 1$). Addition of **L** shifts the initial distribution as indicated by the series of curves. The distribution curves are plots of $g(x)$ vs $\log K$ for values of [**L**] = 0.0, 0.01, 0.05, 0.1, 0.5, and 10.0 M. The inset shows that $\langle \log K \rangle$ approaches a limiting value as a function of [**L**] (this plot covers the range $0.0 \leq$ [**L**] ≤ 1.0 M.

intermediate values of [**L**] the distribution is bimodal, reflecting the relative fractions of complexed and uncomplexed sequences. In the limit of a large excess of ligand, the distribution is again a single Gaussian but with a mean higher than that of the original. The degree to which the mean shifts depends on the standard deviation of the initial population (Figure 2). For example, an initial population

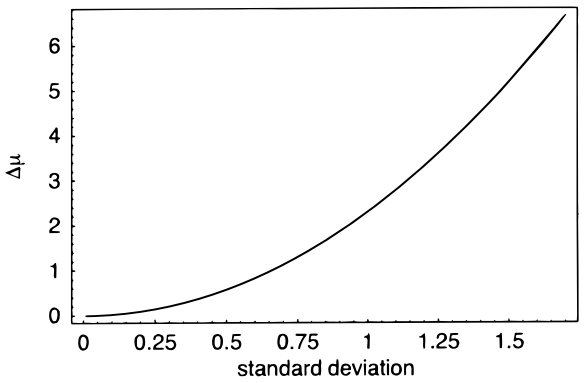


Figure 2. In the limit of a large excess of **L**, the mean value $\langle \log K \rangle$ of the new population shifts by an amount $\Delta\mu = \sigma^2 \cdot \ln(10)$.

having $\mu = 0$ and $\sigma = 1$ results in a new population whose mean binding constant is roughly 2 orders of magnitude higher than the mean of the original population.

A separate conclusion that emerges from the model is that a significant fraction (ca. 5%) of all sequences in the new population will have a binding constant greater than 10^4 times larger than the original mean. This result can best be seen from Figure 3 which plots $G(x)$ as a function of $\log K_{cutoff}$.

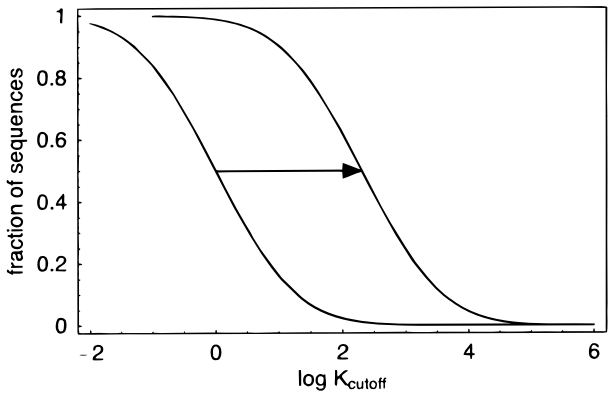


Figure 3. The fraction of sequences, $G(x)$, having binding constant $K \geq K_{cutoff}$ plotted as a function of $\log K_{cutoff}$. The two curves represent the extreme limits of ligand concentration ([**L**] = 0 and a large excess of **L**).

This figure gives an indication of the yield that would be obtained by an affinity chromatography experiment in which the targeted ligand is immobilized. With appropriate washing protocol, it should be possible to isolate and analyze only the tight-binding sequences with K above K_{cutoff}. It would be very interesting and useful to determine whether the comonomer composition of these sequences differs signifi-

(6) Lorsch, J. R.; Szostak, J. W. *Acc. Chem. Res.* **1996**, *29*, 103−110.

cantly from that of the initial population. If such is the case, then at the least this experimental approach may be a useful way to discover lead directions in terms of the comonomer composition best suited for producing tight-binding sequences.

The model developed here is obviously oversimplified in a number of ways. Most significantly, we have assumed that the sequences can be completely transformed from one into another. Strictly speaking, this would only be possible if the individual sequences are isomers of one another and if the equilibrium reaction that interconverts the sequences is an isomerization process. A more typical situation will be sequences consisting of comonomers of different chemical compositions and constitutions. The reversible reaction used to randomize these sequences would likely involve monomer catenation; thus, it would not be possible to transform a sequence of one composition into another of a different composition. Consequently, the theoretical yields resulting from this model should be viewed as an upper limit.

This model leaves several unanswered questions such as the possibility of evolving the selected population to a new set of sequences having higher affinity still. This might be achieved, for example, by cycling through an iterative process involving selection and isolation of the best receptors, followed by reequilibration of this chosen fraction. Among other considerations, this approach assumes that sequence interconversion reactions can be quenched and re-started on demand, perhaps by the addition of a catalyst. Otherwise, of course, it would not be possible to lock-in the structure of the selected sequences. The rationale behind this iterative approach is that, as suggested above, the comonomer composition of the selected population will probably be significantly different than that of the initial population. Thus, during successive equilibration steps, a more optimal pool of comonomers would be used to create the new sequences. Consequently, the large fraction of weakly binding average sequences would be absent and no longer competing for ligand binding.

Obviously, there must be diminishing returns for subsequent cycles of selection and reequilibration. This notion is borne out by the model, since the mean value shifts by an amount proportional to σ^2. Selecting that subpopulation of sequences falling above K_{cutoff} would likely result in a new population whose distribution is narrower than the original's. Consequently, in successive cycles of iteration, the mean will shift to an increasingly smaller degree. One of the practical factors that will significantly contribute to diminishing returns is that the overall yield will plummet exponentially with respect to the number of cycles of selection and reequilibration, making it impractical to carry out all but a few rounds of iteration. Another factor is that the kinetics of the sequence redistribution reaction may become too slow to adequately explore sequence space, perhaps especially as the binding affinity rises. Determining the practical limits of this approach is an aspect that is best left for experimentation.

Is there hope that targeted equilibrium-shifting can be used to synthesize practical quantities of masterpiece sequences? On the basis of the model described here, the answer is probably no. The model showed that, in the face of competitive binding, the distribution of a large population could be shifted only to a rather limited degree. We have also raised issues of a practical nature, such as the possibility of slow redistribution kinetics and exponentially diminishing yields for iterative cycles of selection and reequilibration. Thus, while the approach may find value as a tool for identifying lead compounds, it is unlikely to be of practical utility for synthesizing significant quantities of masterpiece sequences.

If this conclusion is borne out by experimentation, it heightens the need for continued research on nonenzymatic molecular replication[3,7] or other kinetic-based amplification processes. Chain molecules that can somehow be selected *and* amplified meet all the requirements for true chemical evolution. The process of selection may be coupled to a function (e.g., catalysis or recognition) while amplification could come from sequences that catalyze their own replication. For such a system, a subpopulation enriched in molecules having the desired properties could be expanded to the size of the original by exponential replication, thus overcoming the synthetic limitations of targeted equilibrium-shifting.

Acknowledgment. This research was supported by NSF Grant CHE 97-27172. The authors acknowledge Professor Steve Zimmerman for conversations that motivated this work. This material is also supported by the U.S. Department of Energy, Division of Materials Sciences, under Award No. DEFG02-96ER45439, through the Frederick Seitz Materials Research Laboratory at the University of Illinois at Urbana−Champaign.

OL0055723

(7) Representative examples: (a) Nowick, J. S.; Feng, Q.; Tjivikua, T.; Ballester, P.; Rebek, J., Jr. *J. Am. Chem. Soc.* **1991**, *113*, 8831−8839. (b) Terfort, A.; von Kiedrowski, G. *Angew. Chem., Int. Ed. Engl.* **1992**, *31*, 654−656. (c) Bag, B. G.; von Kiedrowski, G. *Pure Appl. Chem.* **1996**, *68*, 2145−2152. (d) Lee, D. H.; Granja, J. R.; Martinez, J. A.; Severin, K.; Ghadiri, M. R. *Nature* **1996**, *382*, 525−528. (e) Wang, B.; Sutherland, I. O. *Chem. Commun.* **1997**, 1495−1496. (f) Yao, S.; Ghosh, I.; Zutshi, R.; Chmielewski, J. *Nature* **1998**, *396*, 447−450. (g) Bag, B. G.; von Kiedrowski, G. *Angew Chem., Int. Ed.* **1999**, *38*, 3713−3714.