BIOPOLYMERS VOL. 6, PP. 575–584 (1968)

Calculation of Binding Isotherms for Heterogeneous Polymers

DONALD M. CROTHERS, *Department of Chemistry, Yale University, New Haven, Connecticut 06520*

Synopsis

The matrix method of statistical mechanics is used to calculate equilibria for the binding of small molecules to polymers. When there is only one kind of binding site the problem is simple; some examples are given for illustrative purposes. If, however, the binding sites are not all equivalent and the bound molecules interact or interfere with each other, the problem is no longer trivial, being formally analogous with calculation of the helix–coil transition equilibrium in a heterogeneous polypeptide. Particular difficulties arise when the sequence of binding sites is aperiodic; most naturally occurring materials fall in this class. The purpose of this paper is to point out that problems of this type are readily solved with good accuracy by use of random-number methods on a high-speed digital computer. One such calculation is presented for illustration. The methods developed are applicable to such systems as the binding of actinomycin, Hg^{++}, and acridine dyes to DNA.

Introduction

The binding of small molecules such as dyes, antibiotics, and transition metal ions to nucleic acids is a commonly observed phenomenon.[1] In many such cases there seems to be a preference for binding to one or another of the bases or base pairs. For example, it is well established that actinomycin is highly selective for dG,[2–5] whereas the acridine dyes seem to bind more strongly to AT-rich than to GC-rich DNA.[6,7] Similarly, Hg^{++} binds more strongly to dAT than to guanine-containing nucleic acids.[8] If the binding sites are completely independent, this heterogeneity in their affinity for the small molecule presents no special problems of analysis, since it is described by two independent chemical equilibria with different equilibrium constants. On the other hand, if the bound molecules interfere with each other or interact in any way, the problem of calculating the adsorption equilibrium is greatly complicated. The model for this latter case is formally analogous to that encountered in calculating helix–coil transition curves for heterogeneous polymers, where the stability of the helical form depends on the nature of the residue present at each position. For such systems involving a heterogeneous lattice, analytical solutions are readily obtained if the arrangement of residues is periodic.[9] In the most interesting case, which includes naturally occurring nucleic acids, the sequence is not periodic, and the only information concerning it is statistical. This problem has

been the object of considerable attention,[10–16] but a fully satisfactory general solution, easily applicable to the kinds of binding equilibria commonly encountered, is still lacking.

Any thorough study of the adsorption of a small molecule by a polymer includes a study of the binding equilibrium. These binding curves contain considerable potential information concerning the nature and size of the binding site and the magnitude of the intrinsic binding constant, but it is difficult to extract this information without a theory for calculating the curves expected for various models of the binding site. This paper is concerned with the pragmatic aim of calculating accurate binding isotherms for the adsorption of small molecules by heterogeneous polymers. The method involves the use of random-number techniques with the expenditure of quite modest amounts of computation time on a high-speed digital computer. The curves calculated for a given model are accurate within statistical limits which are subject to rigorous analysis. Analytic solutions are given only when they can be written without questionable approximations.

Description of the Model

Consider a very long polymer (DNA) molecule to which small molecules can bind at specific sites. The chemical potential μ of the monomer in solution is given by

$$\mu = \mu^\circ + RT \ln m \tag{1}$$

where μ° is the chemical potential at unit activity, and m is the concentration of the small molecule free in solution. (Solutions are assumed to be ideal.) Let μ_α' be the chemical potential of the monomer when bound to a site of kind α but isolated from other bound monomers. (The index α is used to distinguish the binding to AT pairs from binding to GC pairs, for example.) Thus the free energy change ΔG_α accompanying the binding of a single monomer to a specific site on the polymer is

$$\begin{aligned} \Delta G_\alpha &= \mu_\alpha' - \mu \\ &= \mu_\alpha' - \mu^\circ - RT \ln m \\ &= \Delta G_\alpha^\circ - RT \ln m \end{aligned} \tag{2}$$

where ΔG_α° is the standard free energy change for this process. Utilizing a familiar thermodynamic relation, there results

$$K_\alpha = \exp\{-\Delta G_\alpha^\circ / RT\} \tag{3}$$

whereby we define K_α to be the intrinsic binding constant for site α.

The affinity of a given binding site for monomer may depend on the presence of other bound monomers in the vicinity. It will be assumed that the free energy of binding close to other monomers depends only on the distance away of the nearest neighbors on each side. Consider the process

of binding a monomer a distance of n lattice sites away from another monomer on its left, with no neighbor within a finite distance on the right side. Let $\Delta G(n)$, the free energy change for this process, be expressed as

$$\Delta G(n) = \Delta G_\alpha^\circ - RT \ln m + \Delta G^\circ(n) \tag{4}$$

where the terms which depend on the presence of the neighbor have been included in $\Delta G^\circ(n)$. (For simplicity it will be assumed that this term does not depend on α.) Let $\tau(n)$ be defined by an equation analogous to eq. (3):

$$\tau(n) = \exp \{-\Delta G^\circ(n)/RT\} \tag{5}$$

The heterogeneous lattice, which for illustrative purposes may be thought of as a certain sequence of base pairs, can be represented by a sequence of indices α_j specifying the nature of the lattice unit at each site j. If the lattice contains a total of N units, then j varies from 1 to N. In addition, let the symbol 1 represent a lattice site to which a monomer is bound, and 0 an empty site. Let a particular configuration of the system be defined as a specific sequence of zeroes and ones, written in register with the sequence of indices α_j, by which the state of each residue in the lattice is specified.

The probability P_i that the system takes on configuration i is given by the basic equation

$$P_i = \exp \{-G_i/RT\}/Q \tag{6}$$

where

$$Q = \sum_i \exp \{-G_i/RT\} \tag{7}$$

is the configurational partition function for the system, and G_i is the free energy of forming configuration i from some reference state. The average $\langle a \rangle$ of a quantity which has a value a_i in configuration i is then readily seen to be

$$\langle a \rangle = \sum_i a_i P_i = \sum_i a_i \exp \{- G_i/RT\}/Q \tag{8}$$

A convenient reference state for calculating free energies is the polymer to which no monomer is adsorbed. The term $\exp \{-G_i/RT\}$ can then be generated as a product of the form

$$\exp \{-G_i/RT\} = \prod_{j=1}^{M} \exp \{-G_j/RT\} \tag{9}$$

where G_j is the free energy change when the jth monomer, out of a total of M, is bound to the polymer. Taking note of eqs. (2)–(5), it is apparent that $\exp \{-G_i/RT\}$ can be generated by multiplying together factors $K_\alpha m$ for each monomer bound alone with factors $K_\alpha m\tau(n)$ for each monomer bound n lattice units away from its left-hand neighbor. These factors will be referred to as "statistical weighting factors" in what follows. The partition function Q is obtained by summing such products for all possible configurations i.

Matrix Method for Some Simple Cases

One of the common techniques for solving problems in lattice statistics is the matrix method, applied, for example, to the helix–coil transition problem by Zimm and Bragg.[17] The method involves operations on a "statistical weight vector" by a matrix operator to form the partition function. Suppose δ units in the lattice must be considered in order to take account of the range of interactions in the system. A group of δ units can have up to 2^δ possible states, since each unit can, in principle, be in one of the forms 0 and 1. These states will be referred to as the "joint configurations" of the lattice units $k - \delta + 1$ through k. The statistical weight vector $\mathbf{a}_k$ has one component for each possible joint configuration. Each such component is a sum of the product of statistical weighting factors corresponding to all configurations of the lattice sites 1 through k which are consistent with the particular joint configuration. The vector $\mathbf{a}_{k+i}$ is generated from $\mathbf{a}_k$ by use of the matrix operator **M:**

$$\mathbf{a}_{k+1}{}^\dagger = \mathbf{M}\mathbf{a}_k{}^\dagger \tag{10}$$

where the superscript dagger (†) indicates the transposed or column vector. The element M_{ij} of the matrix **M** is the appropriate statistical weighting factor when the $(k + 1)$th segment is added, thereby producing joint configuration i for the units $k - \delta + 2$ through $k + 1$ from joint configuration j for the units $k - \delta + 1$ through k. If the lattice contains N units, the partition function is the sum of the elements of $\mathbf{a}_N$, or

$$\begin{aligned} Q &= \mathbf{e}\mathbf{a}_N{}^\dagger \\ &= \mathbf{e}\mathbf{M}^{(N-\mu)}\mathbf{a}_\delta{}^\dagger \end{aligned} \tag{11}$$

where **e** is a unit vector. The vector $\mathbf{a}_\delta$ can usually be written down by inspection. For reasons discussed thoroughly by other authors,[17] Q may be approximated by

$$Q = \lambda_{\max}^N \tag{12}$$

when N is large, where $\lambda_{\max}$ is the largest eigenvalue of the matrix **M.** The average quantity of interest, the fraction r of lattice sites occupied by monomers is

$$\begin{aligned} r &= N^{-1}\,\partial \ln Q/\partial \ln m \\ &= \partial \ln \lambda_{\max}/\partial \ln m \end{aligned} \tag{13}$$

This formalism will be clarified by the examples which follow.

Consider first the simple case in which binding at each site is independent of its neighbors, with a single intrinsic binding constant K. Suppose that there are $1/B_0$ base pairs per binding site, and that N is associated with the number of base pairs in the lattice. In this case δ is 1, and the two joint configurations are 1 and 0. The initial vector $\mathbf{a}_1$ is $(1, Km)$, and the matrix **M** is

$$\mathbf{M} = \begin{pmatrix} 1 & 1 \\ Km & Km \end{pmatrix} \tag{14}$$

The eigenvalues of $\mathbf{M}$ are 0 and $Km + 1$, so that

$$Q = (Km + 1)^{NB_0} \tag{15}$$

Equation (13) gives

$$r = KmB_0/(Km + 1) \tag{16}$$

which can be rearranged to

$$r/m = K(B_0 - r) \tag{17}$$

Equation (17) is the form appropriate to a Scatchard[18] plot of the binding isotherm, in which r/m is plotted against r, with slope K and intercept B_0 on the r axis. Values of equilibrium constants determined in this way will be referred to as apparent binding constants K_{ap}, with a similar convention for B_{ap}. It is clear that in this case K_{ap} is identical with the intrinsic binding constant K, and that B_{ap} is the same as B_0, the number of binding sites per base pair. More complicated cases will not show the same identity.

Suppose now that there are two classes of binding sites with intrinsic binding constants K_1 and K_2, but that binding at each site is still independent of neighboring sites. For simplicity, let $B_0 = 1$. The matrix $\mathbf{M}_\alpha$ associated with binding sites of class α is

$$\mathbf{M}_\alpha = \begin{pmatrix} 1 & 1 \\ K_\alpha m & K_\alpha m \end{pmatrix} \tag{18}$$

The vector $\mathbf{a}_1$ is either $(1,K_1m)$ or $(1,K_2m)$, depending on whether the first lattice site is of kind 1 or 2. The partition function Q is

$$Q = \mathbf{e} \prod_{j=2}^{N} \mathbf{M}_{\alpha_j} \mathbf{a}_1^\dagger \tag{19}$$

Inspection reveals that the matrix operator $\mathbf{M}_\alpha$ simply multiplies both components of $\mathbf{a}$ by $(1 + K_\alpha m)$, and that Q is therefore independent of the order of the matrix multiplications in eq. (19). Letting N_α be the number of units of type α, with $N_\alpha/N \equiv \beta_\alpha$,

$$Q = (1 + K_1m)^{N_1}(1 + K_2m)^{N_2} \tag{20}$$

From which there results

$$r_\alpha/m = K_\alpha(\beta_\alpha - r_\alpha) \tag{21}$$

consistent with the independence of the binding sites.

The simple cases treated so far are characterized by independent binding sites, expressed in the formalism described above by setting $\tau(n) = 1$ for all values of $n \geq 1$. If the binding sites are not independent, analytic results can be easily derived if there is only one kind of binding site. Consider the situation in which binding of a small molecule to a given base pair prevents binding of another monomer closer than n' base pairs from the first, leaving $n' - 1$ empty potential binding sites between

the two monomers. (The difference between this and the first case considered is that then there were B_0^{-1} base pairs per binding site, and the monomer was assumed to bind in only one way to the binding site. In the present case, any base pair can serve as a binding site, thereby excluding binding at adjacent pairs.) As an example, let $n' = 2$. Then $\tau(n) = 0$ for $n = 1$, and $\tau(n) = 1$ for $n \geq 2$. δ is 2, and the permitted joint configurations are 00, 01, and 10. The initial vector $\mathbf{a}_2$ is $(1, Km, Km)$, and the matrix operator $\mathbf{M}$ is

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 1 \\ Km & 0 & Km \\ 0 & 1 & 0 \end{pmatrix} \tag{22}$$

It may readily be verified that for arbitrary n', the characteristic equation is

$$Km = \lambda^{n'} - \lambda^{n'-1} \tag{23}$$

from which r may be calculated by using eq. (13). For this purpose it is convenient to use a parametric representation; eqs. (13) and (23) give

$$r = (\lambda - 1)/(n'\lambda - n' + 1) \tag{24}$$

Insertion of a suitable set of values of $\lambda(\lambda > 1)$ into eqs. (23) and (24) yields r as a function of Km without the necessity for solving the nth order eq. (23).

Computational Method for Heterogeneous Polymers

In the general case of monomer binding to a heterogeneous polymer with binding sites α_i, the partition function Q is given by

$$Q = \mathbf{e} \prod_{j=\delta+1}^{N} \mathbf{M}_{\alpha_j} \mathbf{a}_\delta \tag{25}$$

Only in special circumstances, such as independent binding sites as discussed above, can the matrix product in eq. (25) be evaluated in simple terms. The matrices $\mathbf{M}_{\alpha_j}$ usually do not commute with each other, so the value of Q depends on the sequence in which the binding sites are arranged. Approximate methods developed by ourselves and others to deal with this kind of problem require substantial amounts of computation, and suffer from the additional disadvantage that rapid convergence to the correct solution is not necessarily assured.[16] The central point of this communication is to observe that with present-day computers, direct calculations of Q from eq. (25) are easily made for most cases of interest. In the unlikely event that the sequence of binding sites $\{\alpha_j\}$ is known, the computer can be given this information. Otherwise, the most appropriate model is usually a random sequence of binding sites. The machine can readily generate such a sequence of arbitrary length, containing close to preselected compositions of the various kinds of sites. Repeated multiplication of the vector $\mathbf{a}_k$ by the matrix M_α a total of $N - \delta$ times, followed by summation

of the elements of $\mathbf{a}_N$, generates Q. Differentiation of Q to obtain r [eq. (13)] can be done numerically by calculating Q for two values of m which differ by a small quantity δm. Satisfactory statistics for binding isotherms of the kind reported here and in a forthcoming paper[19] were obtained with computation times varying from a few seconds up to about a minute (IBM 7094) for the complete binding isotherm.

As a specific example, consider a problem of the kind encountered in the binding of actinomycin analogs to DNA. Suppose that the small molecule is able to insert[20] between two neighboring base pairs if one of the pairs is GC, but that there must be at least two base pairs between adjacent bound monomers. (The intercalation model of binding is taken purely for illustrative purposes and does not influence any of the general conclusions which follow. Conversely, it is not possible to distinguish between intercalation and external binding on the basis of the shape of binding isotherms.) The binding sites may be identified with planes between the base pairs, which means, in the language of the final example of the proceding section, that there must be at least one empty site between adjacent monomers and n' is therefore 2. Again, $\tau(n) = 0$ for $n = 1$, and $\tau(n) = 1$ for $n \geq 2$. The computer, using a random number generator, can store a sequence of symbols representing AT and GC pairs. There are two kinds of sites, specifically, the space between two AT pairs, which does not bind the monomer, and the spaces between all other nearest neighbor combinations, which bind with intrinsic binding constant K. If the first two base pairs in the sequence are AT, then $\mathbf{a}_2$ is (1, 0, 0); otherwise it is (1, Km, Km). For each successive lattice point the computer ascertains the nature of the site and then multiplies the vector $\mathbf{a}$ by one of the two matrices $\mathbf{M}_\alpha$. For a GC-type site, $\mathbf{M}_1$ is as given by eq. (22), and for a nonbonding site $\mathbf{M}_2$ is

$$\mathbf{M}_2 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \tag{26}$$

The large number of zeroes in $\mathbf{M}_\alpha$ greatly facilitates the iterative procedure. r is determined by numerical differentiation as described above. $N = 5000$ was found a convenient lattice size for most problems, a value that permits advance specification of the base composition of the random sequence within roughly 2%. Statistical fluctuations in calculated binding isotherms for the same base composition were around 2%, which is better than isotherms can usually be measured experimentally.

Results and Discussion

Results are of interest primarily in connection with specific cases, for which our study of the binding of actinomycin and analogous compounds to DNA[19] may be taken as an example. This discussion will be restricted to a single case study, from which some general observations can be made concerning the meaning of K_{ap} and B_{ap} determined from Scatchard plots.

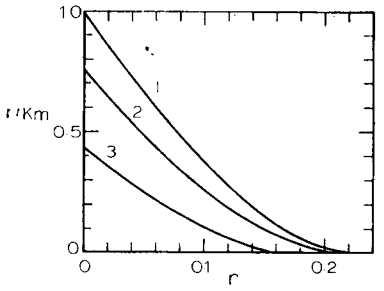

Fig. 1. Binding isotherms for a model in which a small molecule intercalates between two base pairs when at least one of them is GC. In addition, there must be at least four base pairs between adjacent bound monomers (r/Km vs. r) where K is the intrinsic binding constant, r the ratio of bound monomers to base pairs, and m the concentration of monomers free in the solution. The curves are for different base compositions: The fractional GC contents are (1) 1.0; (2) 0.505; (3) 0.257. The base sequence is assumed to be random. Curve *1* was calculated from eqs. (23) and (24), and curves *2* and *3* from eq. (25) by the random-number method described in the text.

Figure 1 shows binding isotherms calculated for an intercalation model in which a small molecule can insert between any nearest neighbor pair of which at least one member is GC, but there must be at least four base pairs between adjacent bound monomers ($n' = 4$). Curves are shown for various values of GC content.

The first general observation which can be made is that the binding isotherms are not linear, even when there is only a single class of binding sites. The small slope of the isotherm at large values of r, implying a smaller apparent binding constant, results from the large reduction in the number of ways of achieving a given degree of binding as saturation is approached.

Secondly, the slope of the first part of the isotherm, which is nearly linear, is not closely related to the intrinsic binding constant K. (Curves are normallized by plotting r/Km; if K were equal to K_{ap}, the slope of this plot would be 1.) For the present case, K_{ap} (the slope of the isotherm) exceeds K by a factor of about 6.

If the initial nearly linear region of the isotherm is extrapolated through the r axis, the intercept is B_{ap}. For the pure GC polymer B_{ap} is about 0.16, whereas the polymer can actually bind up to 1 monomer per four base pairs. Hence in this case r at saturation is 0.25. It is clear that B_{ap} has no firm physical significance.

On the other hand, the intercept on the r/m axis of a binding isotherm is well defined. In the limit as r approaches zero the binding can be described by a simple equilibrium since bound monomers no longer interfere with each other. Equation (17) is therefore valid in this limit, and one can write

$$\lim_{r \to 0} r/m = KB_0 \tag{27}$$

where K is the intrinsic binding constant and B_0 is the number of potential binding sites per base pair. In the present example, B_0 is the fraction of nearest neighbor pairs which contain at least one GC, or

$$B_0 = 2\beta_{GC} - \beta^2_{GC} \tag{28}$$

where β_{GC} is the fractional GC content. If K is truly independent of GC content, eqs. (27) and (28) give (again for the specific example at hand)

$$\lim_{r\to 0} r/m = K(2\beta_{GC} - \beta^2_{GC}) \tag{29}$$

A plot of the limiting value of r/m, determined for a variety of DNA's, versus a specific expression for B_0 in terms of GC content could be of general utility in testing different models for the nature of the binding site. If there are several kinds of binding sites α, eq. (27) is modified to read

$$\lim_{r\to 0} r/m = \sum_{\alpha} K_\alpha B_{0,\alpha} \tag{30}$$

A relation like eq. (30) is discussed in connection with independent binding sites by Edsal and Wyman.[21]

Recently, in a paper which appeared after the work reported here had been completed, Vedenov et al.[22] used a method related to that described here to perform calculations on an analogous problem, that of the helix–coil transition in heterogeneous polymers.

This research was supported by grant GB 4083 from the National Science Foundation.

References

1. R. F. Steiner and R. F. Beers, Jr., *Polynucleotides*, Elsevier, Amsterdam, 1961, Chap. 10.
2. W. Kersten, *Biochim. Biophys. Acta*, **47**, 610 (1961).
3. E. Reich, J. H. Goldberg, and M. Rabinowitz, *Nature*, **196**, 743 (1962).
4. E. Kahan, F. M. Kahan, and J. Hurwitz, *J. Biol. Chem.*, **238**, 2491 (1963).
5. M. Gellert, C. E. Smith, D. Neville, and G. Felsenfeld, *J. Mol. Biol.*, **11**, 445 (1965).
6. A. R. Peacocke and J. N. H. Skerett, *Trans. Faraday Soc.*, **52**, 261 (1956).
7. V. Kleinwächter and J. Koudelka, *Biochim. Biophys. Acta*, **91**, 539 (1964).
8. U. S. Nandi, J. C. Wang, and N. Davidson, *Biochemistry*, **4**, 1687 (1965).
9. D. M. Crothers and N. R. Kallenbach, *J. Chem. Phys.*, **45**, 917 (1966).
10. S. Lifson, *Biopolymers*, **1**, 25 (1963).
11. S. Lifson and G. Allegra, *Biopolymers*, **2**, 65 (1964).
12. D. M. Crothers, N. R. Kallenbach, and B. H. Zimm, *J. Mol. Biol.*, **11**, 802 (1965).
13. E. Montrol and N. Goel, *Biopolymers*, **4**, 855 (1966).
14. Y. Kawai, M. Ozaki, M. Tanaka, and E. Teramoto, *J. Phys. Soc. Japan*, **20**, 1457 (1965).
15. H. Reiss, D. A. McQuarrie, J. P. McTague, and E. R. Cohen, *J. Chem. Phys.*, **44**, 4567 (1966).
16. D. M. Crothers, *Biopolymers*, **4**, 1025 (1966).
17. B. H. Zimm and J. K. Bargg, *J. Chem. Phys.*, **31**, 526 (1959).
18. G. Scatchard, *Ann. N.Y. Acad. Sci.*, **51**, 660 (1949).

19. W. Müller and D. M. Crothers, in preparation.
20. L. S. Lerman, *J. Mol. Biol.*, **3**, 18 (1961).
21. J. T. Edsal and J. Wyman, *Biophysical Chemistry*, Vol. 1, Academic Press, New York, 1958, Chap. 11.
22. A. A. Vedenov, A. M. Dykhne, A. D. Frank-Kamenetskii, and M. D. Frank-Kamenetskii, *Molekul. Biol.*, **1**, 313 (1967).

Received October 20, 1967