ELSEVIER

Biophysical Chemistry 94 (2001) 185–199

Biophysical Chemistry

www.elsevier.com/locate/bpc

Maximum-entropy determination of self-association distribution functions; daunorubicin and ATP

Douglas Poland*

Department of Chemistry, The Johns Hopkins University, Baltimore, MD 21218, USA

Received 3 August 2001; received in revised form 12 September 2001; accepted 12 September 2001

Abstract

In the present paper we show how one can use the perturbation of some molecular optical property (for example circular dichroism or chemical shift) as a function of concentration to construct cluster distribution functions describing the self-association of molecules in solution. The optical data are first converted into data giving the variation of the average extent of clustering as a function of the total concentration and then, using straightforward thermodynamics, a set of moments of the cluster distribution function can be obtained. Utilizing the maximum-entropy method, the moments are then used to calculate approximate distribution functions, where the more moments that are used the better the approximation obtained. Given the probability distribution for clusters of different sizes one can then calculate the equilibrium constant for each stage of association. Thus one converts average degree of association into equilibrium constants without having to use any specific model. By this method one can clearly tell whether the equilibrium constants remain constant, increase, or decrease with the number of molecules in a cluster. We apply the method to literature data for two systems, namely daunorubicin, which has a strong tendency to cluster in solution, and $Mg(ATP)^{2-}$ which forms weaker clusters. We find that the successive equilibrium constants for adding a monomer to a cluster are approximately constant for daunorubicin but clearly decrease as a function of increasing cluster size for $Mg(ATP)^{2-}$. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Self-association; Cluster distribution function; Moments; Maximum-entropy method

1. Introduction

Many more or less flat molecules of biological origin, such as the nucleic acid bases and other similar molecules, tend to self-associate in solution to form roughly linear stacks of varying sizes. Examples of these phenomena are given in the work of Martin [1] on the antibiotic daunorubicin and Scheller et al. [2] on nucleotides, which we will use as examples. These workers analyzed their data using the assumption that the equilibrium constant for adding a monomer to a stack was independent of the size of the stack (equal-K model). Martin [3] has reviewed a great deal of the data in this field and has compared the results obtained from the equal-K model with a model where the equilibrium constant for adding a mon-

* Corresponding author. Tel.: +1-410-516-7441; fax: +1-410-516-8420.

E-mail address: poland@jhunix.hcf.jhu.edu (D. Poland).

PII: S0301-4622(01)00241-1

omer gets smaller as the size of the stack increases (attenuated-*K* model). Using the review of Martin as a starting point we will develop a method for treating these systems whereby one can obtain information about the stack-size dependence of the equilibrium constant without having to introduce any specific model. The approach uses the maximum-entropy method whereby moments of a distribution function are used to obtain an approximate distribution. Knowledge of the distribution function (giving the probabilities of various cluster species) then directly gives information about the equilibrium constants for different stages of association. We have previously used this method to treat enthalpy and free energy distributions in proteins and have applied it (Poland [4–6]), using techniques very similar to those employed in the present application, to the treatment of titration's and ligand binding properties in proteins and nucleic acids.

We begin with the basic clustering reaction (where *A* indicates a basic monomer unit and the subscripts indicate how many such units are in a cluster) where we add a monomer to a chain of (*n*-1) units

$$A_1 + A_{n-1} \leftrightarrow A_n \tag{1}$$

The equilibrium constant for this reaction will be designated as K_n which we will take in general as a function of *n*. Alternatively, for equilibrium, one can consider the direct formation of A_n relative to monomer

$$n\ A_1 \leftrightarrow A_n \tag{2}$$

This reaction will have equilibrium constant Q_n which is related to the K_n as follows (one adds up the stepwise reactions of Eq. (1) to get Eq. (2); the net equilibrium constant is the product of the equilibrium constants for the individual steps)

$$Q_n = \prod_{m=2}^{n} K_m \tag{3}$$

For completeness we define Q_1, the equilibrium constant for the identity reaction, $A_1 \leftrightarrow A_1$, as $Q_1 = 1$.

The total amount of monomer, *c*, in the system is constant (conservation relation)

$$c = \sum_{n=1}^{\infty} n\ [A_n] \tag{4}$$

We will assume that we have experimental data giving the average extent of clustering $\langle n \rangle$ as a function of the total concentration of monomer in the system (*c* of Eq. (4)). We then construct an empirical Taylor series expansion of $\langle n(c) \rangle$ about a given total concentration c_o.

$$\langle n(c) \rangle = n_o + n'(c-c_o) + \frac{1}{2} n''(c-c_o)^2 \tag{5}$$

where the primed quantities are given by

$$n_o = \langle n(c_o) \rangle,\ n' = \left(\frac{\partial \langle n \rangle}{\partial c} \right)_{c_o}, \quad n'' = \left(\frac{\partial^2 \langle n \rangle}{\partial c^2} \right)_{c_o} \tag{6}$$

The point of this exercise is that the three numbers given in Eq. (6), representing simply a local Taylor series expansion of the experimental data (with no models assumed), give enough information to calculate the first three moments of the cluster distribution function at $c = c_o$ and the first three moments are in turn enough information to construct an approximation to the entire cluster distribution function from which the equilibrium constants of the reaction given in Eq. (1) can be determined. In the next section we show how the parameters in Eq. (6) are related to the first three moments of the cluster distribution function and then, in the following section, we show how the moments can be actually used to construct the distribution function.

2. Moments of the distribution

In this section we show that the numbers in Eq. (6) give sufficient information to allow us to calculate the first three moments of the cluster distribution function without the requirement of any additional assumptions or models. We begin with the basic association reaction of Eq. (2). The standard equilibrium constant expression for this reaction gives

$$[A_n] = [A_1]^n Q_n \tag{7}$$

We then define an association polynomial

$$\Gamma = \sum_{n=1}^{\infty} [A_n] = \sum_{n=1}^{\infty} [A_1]^n Q_n \tag{8}$$

which represents the total concentration of clusters of any size. Γ is the analog for self-association of the binding polynomial for the binding of ligands to macromolecules [4–6]. The probability of an n-mer is then given by

$$P_n = [A_1]^n Q_n / \Gamma \tag{9}$$

while the average extent of association is

$$\langle n\rangle = \sum_{n=1}^{\infty} n\ P_n = \sum_{n=1}^{\infty} n\ [A_1]^n Q_n / \sum_{n=1}^{\infty} [A_1]^n Q_n \tag{10}$$

which can be generated as a derivative of Γ

$$\langle n\rangle = \frac{A_1}{\Gamma}\frac{\partial\Gamma}{\partial A_1} \tag{11}$$

Higher moments are generated in a similar manner

$$\langle n^m\rangle = \sum_{n=1}^{\infty} n^m P_n = \frac{1}{\Gamma}\frac{\partial^m\Gamma}{\partial y^m} \tag{12}$$

where

$$y = \ln A_1 \tag{13}$$

If one takes further derivatives of $\langle n\rangle$ using either Eq. (10) or Eq. (11) one obtains

$$\frac{\partial\langle n\rangle}{\partial y} = \langle n^2\rangle - \langle n\rangle^2 \tag{14}$$

and

$$\frac{\partial^2\langle n\rangle}{\partial y^2} = \langle n^3\rangle - 3\langle n\rangle\langle n^2\rangle + 2\langle n\rangle^3 \tag{15}$$

These relations are given in terms of A_1 which is the concentration of free monomer in solution. Since we have $\langle n\rangle$ as a function of the total monomer concentration in any form, we must convert Eq. (14) and Eq. (15) into relations involving derivatives with respect to c rather than A_1. To accomplish this we use the following relations

$$\frac{\partial\langle n\rangle}{\partial y} = n^{(1)}\frac{\partial w}{\partial y} \tag{16}$$

and

$$\frac{\partial^2\langle n\rangle}{\partial y^2} = n^{(2)}\left(\frac{\partial w}{\partial y}\right)^2 + n^{(1)}\frac{\partial^2 w}{\partial y^2} \tag{17}$$

where we define

$$w = \ln c \tag{18}$$

We then have

$$n^{(1)} = \frac{\partial\langle n\rangle}{\partial \ln c} = c\frac{\partial\langle n\rangle}{\partial c} = cn' \tag{19}$$

and

$$\begin{aligned} n^{(2)} &= \frac{\partial^2\langle n\rangle}{\partial(\ln c)^2} = \frac{\partial\langle n\rangle}{\partial c} + c^2\frac{\partial^2\langle n\rangle}{\partial c^2} \\ &= cn' + c^2 n'' \end{aligned} \tag{20}$$

where n' and n'' are given by Eq. (6). The relation between A_1 and c (or between y and w) is given through the conservation relation of Eq. (4)

$$c = \sum_{n=1}^{\infty} n\ [A_n] = \sum_{n=1}^{\infty} n[A_1]^n Q_n \tag{21}$$

We define the first three moments of the cluster distribution function as

$$M_1 = \langle n\rangle, \quad M_2 = \langle n^2\rangle, \quad M_3 = \langle n^3\rangle \tag{22}$$

then from Eq. (21) we obtain the relations

$$\begin{aligned} \frac{\partial w}{\partial y} &= \frac{\langle n^2\rangle}{\langle n\rangle} \text{ and} \\ \frac{\partial^2 w}{\partial y^2} &= \frac{\langle n^3\rangle}{\langle n\rangle} - \left(\frac{\langle n^2\rangle}{\langle n\rangle}\right)^2 \end{aligned} \tag{23}$$

Combining Eqs. (19)–(23) one finally obtains the following simple relations for the first three moments in terms of the variables $n^{(1)}$ and $n^{(2)}$ which through Eq. (19) and Eq. (20) are obtained from the experimental quantities n_o, n' and n'' of Eq. (5)

$$M_1 = n_o,\ M_2 = \frac{M_1^2}{[1 - n^{(1)}/M_1]} \tag{24}$$

$$M_3 = \frac{[M_2/M_1]^2(n^{(2)} - n^{(1)}) + 3M_1M_2 - 2M_1^3}{[1 - n^{(1)}/M_1]} \tag{25}$$

The above equations have the form of a set of recursion relations for the moments

$$M_1 = n_o, M_2 = f_1(M_1, n') \qquad (26)$$
$$M_3 = f_2(M_1, M_2, n', n'')$$

3. Probability distributions from moments

We now use the maximum-entropy method as treated by Tagliani [7] to turn knowledge of moments into knowledge of the cluster distribution function. One begins by defining an entropy-like quantity, $-\int P(x)\ln P(x)dx$, in terms of the probability distribution $P(x)$ and then maximizes this quantity subject to a finite set of constraints, namely the moment equations $M_n = \int x^n P(x)dx$.

The result of this procedure is the maximum-entropy form of the distribution function which is found to be

$$P(x) = \exp[-g(x)] \qquad (27)$$

where

$$g(x) = \sum_{n=0}^{L} \lambda_n x^n \qquad (28)$$

The quantity $g(x)$ is a finite polynomial with the limit L being the number of moments that one has (which, in our case, is three). If one has only a single moment, the value of the average x, then the distribution function has the form of a simple exponential; if one has two moments, the average x and the average x^2, then the distribution has the form of a standard Gaussian distribution. When one has more than two moments then one has a distribution function that goes beyond a simple Gaussian distribution. In constructing $P(x)$ one starts with the numerical values of a finite set of moments and uses a simple iteration scheme to calculate the values of the λ_n as given in Eq. (28).The constant λ_o is a normalization factor. Thus one trades a finite set of values of the moments for a set of the same size of values of the λ_n.

The probability distribution in Eq. (27) is given in terms of a continuous variable x while the cluster distribution function is in terms of the discrete number of particles in a cluster. The correspondence between the continuous and discrete functions (discussed in [5]) is as follows. To obtain a discrete set of probabilities we use integrals centered at the integers. For example, to obtain P_1 we use the integral

$$P_1 = \int_{1/2}^{3/2} P(x)dx \qquad (29)$$

and so on. The continuous function $P(x)$ given in Eq. (27) is defined on a finite interval starting with $x=0$. Thus in order to treat P_o we must shift the origin of the function $P(x)$ by $1/2$ introducing the new moments for the shifted distribution

$$M'_m = \sum_{n=0}^{N} (n+1/2)^m P_n \qquad (30)$$

giving

$$M'_1 = M_1 + \frac{1}{2}$$

$$M'_2 = M_2 + M_1 + \frac{1}{4}$$

$$M'_3 = M_3 + 1\frac{1}{2}M_2 + \frac{3}{4}M_1 + \frac{1}{8} \qquad (31)$$

and so on. The cluster probabilities are then (using $P(x)$ obtained with the shifted moments)

$$P_o = \int_0^1 P(x)dx, \quad P_1 = \int_1^2 P(x)dx \qquad (32)$$

and so on. This represents a small change, but it gives the best correspondence between the continuous distribution of Eq. (27) and the discrete cluster distribution.

4. Equal and attenuated *K* models

As an example of the above methods we treat the two models used by Martin [3] and show that for model data this method can easily distinguish between these two models. The equal-*K* model (*KE* model) assumes that

$$K_n = K_e \tag{33}$$

where K_e is independent of cluster size. The attenuated-K model (KA model) assumes that successive K's get smaller in a particular manner

$$K_n = \left(\frac{1}{n-1}\right) K_a \tag{34}$$

where K_a is the equilibrium constant for dimerization (this gives $K_2 = K_a$, $K_3 = K_a/2$, etc.).

The corresponding Q_n (from Eq. (3)) are then

$$\begin{aligned} Q_{n+1} &= K_e^n \quad (KE \text{ model}) \\ Q_{n+1} &= \frac{1}{n!} K_a^n \quad (KA \text{ model}) \end{aligned} \tag{35}$$

The general form of the association polynomial introduced in Eq. (8) is

$$\Gamma = \sum_{n=1}^{\infty} X^n Q_n \tag{36}$$

where x is the concentration of free monomer

$$X = [A_1] \tag{37}$$

which is not the same as the total concentration of c given in Eq. (21). Using the forms for Q_n given in Eq. (35) in Eq. (36) one obtains the closed forms

$$\begin{aligned} \Gamma_e &= X \sum_{n=0}^{\infty} (XK_e)^n = \frac{X}{1 - XK_e} \\ \Gamma_a &= X \sum_{n=0}^{\infty} \frac{1}{n!} (XK_a)^n = Xe^{XK_a} \end{aligned} \tag{38}$$

The fact that these two models lead to closed forms for Γ is, of course, the reason for choosing them. Our general approach is not restricted to these two simple forms but we use them here as examples of cluster distribution functions that can be treated exactly. Given Γ one then immediately has closed forms for the moments [ln $X = y$ as defined in Eq. (13)]

$$\begin{aligned} &\Gamma^{(1)} = \partial\Gamma/(\ln X);\ \Gamma^{(2)} = \partial^2\Gamma/(\ln X)^2; \\ &\Gamma^{(3)} = \partial^3\Gamma/(\ln X)^3 \end{aligned} \tag{39}$$

$$\begin{aligned} &M_1 = \Gamma^{(1)}/\Gamma;\ M_2 = \Gamma^{(2)}/\Gamma; \\ &M_3 = \Gamma^{(3)}/\Gamma \end{aligned} \tag{40}$$

In particular it is useful to have the details of the KE model as a reference. For convenience we define the following variable

$$u_e = [A_1]\ K_e = X\ K_e \tag{41}$$

the value of which is determined by the conservation relation of Eq. (21)

$$u_e = 1 - \left[\sqrt{1 + 4cK_e} - 1\right]/2cK_e \tag{42}$$

where c is the total concentration. Notice that the physical solution for u must be a real positive number less than one. The experimental data one has is $\langle n \rangle = M_1$ as a function of c which in the KE model is explicitly

$$\langle n \rangle = \frac{2cK_e}{\sqrt{1 + 4cK_e} - 1} \tag{43}$$

while the moments are given by

$$\begin{aligned} &M_1 = \frac{1}{1 - u_e}, \quad M_2 = \frac{1 + u_e}{(1 - u_e)^2}, \\ &M_3 = \frac{1 + 4u_e + u_e^2}{(1 - u_e)^3} \end{aligned} \tag{44}$$

For the KA model the corresponding quantities are

$$\begin{aligned} M_1 &= \langle n \rangle = 1 + u_a, \\ M_2 &= \langle n^2 \rangle = 1 + 3u_a + u_a^2, \\ M_3 &= \langle n^3 \rangle = 1 + 7u_a + 6u_a^2 + u_a^3 \end{aligned} \tag{45}$$

where now

$$u_a = x\ K_a \tag{46}$$

Note that in Eq. (45) the exponential that occurs in Eq. (38) cancels [see Eq. (40)]. The quantity u_a is obtained from the conservation relation of Eq. (21) which, using Eq. (38), is transcendental and must be solved numerically

$$c = u_a(1 + u_a)e^{u_a}/K_a \tag{47}$$

For the KE model u_e is given by Eq. (41) and is obtained from Eq. (42). This quantity must be less than one for the KE model since as u_e approaches one the moments given by Eq. (44) become infinite. For the KA model u_a is given by Eq. (46) and is obtained by Eq. (47). For the KA model u_a can be greater than one since clearly the

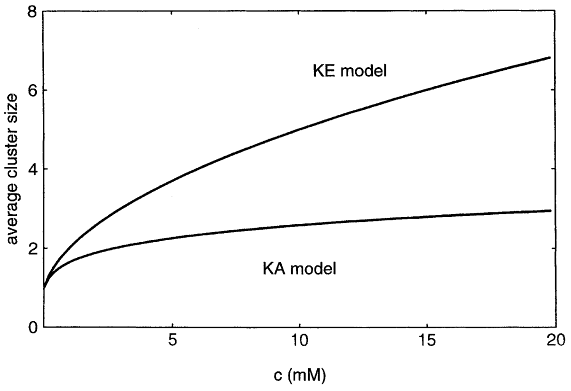

Fig. 1. The variation of the average cluster size, $\langle n \rangle$, as a function of c, the total concentration of monomer in any form for the KE and KA models described by Eq. (35) using the parameters given in Eq. (53).

first moment given by Eq. (45) can be greater than two.

Of considerable interest are the probability distributions. We focus attention on two probability distributions which we define as follows:

P_n = probality a cluster picked at random contains n monomers (48)

R_n = probality a molecule picked at random is in a cluster of n particles. (49)

These quantities are given respectively by the following relations

$$P_n = X^n Q_n / \sum_{n=1}^{\infty} X^n Q_n,$$

$$R_n = n\ X^n Q_n / \sum_{n=1}^{\infty} n\ X^n Q_n \tag{50}$$

For the KE model these relations simplify to give

$$P_n = (1-u_e)\ u_e^{n-1}$$
$$R_n = n\ u_e^{n-1}(1-ue)^2 \tag{51}$$

where both distributions are normalized. For the KA model the corresponding functions are

$$P_n = \frac{1}{(n-1)!}\ u_a^{n-1} e^{-u_a}$$
$$R_n = \frac{n}{(n-1)!}\ u_a^{n-1} e^{-u_a}/(1+u_a) \tag{52}$$

Note that since $u_e < 1$ for the KE model, P_n must be a monotonically decreasing function (it is simple exponential). For the KA model u_a can be greater than one so u_a^{n-1} can be an increasing function while $1/(n-1)!$ is decreasing; thus in this model P_n can have a maximum at an intermediate value of n.

Fig. 1 compares $\langle n \rangle$ as a function of c for the two models with $K_a = K_e$. Since the K's are bigger for the KE model (except at the first step), the clustering is greater in the KE model as seen in Fig. 1. Comparing the two models at the same value of c one has

$$\begin{aligned}&KE: \quad c=10,\ K_e=2, \quad \langle n \rangle = 5.0\\ &X=[A_1]=0.4, \quad u_e=0.8\\ &KA: \quad c=10,\ K_a=2, \quad \langle n \rangle = 2.59\\ &X=[A_1]=0.793,\ u_a=1.585\end{aligned} \tag{53}$$

In the above tabulation the units of concentration are mM and the K's are given in units of mM^{-1}. The probability distributions given in Eq. (51) for the KE model are shown in Fig. 2 for the parameters given in Eq. (53). While the P_n distribution drops off exponentially (with monomer always the most probable cluster size), the distribution R_n has a maximum at approximately $n = 4$. The R_n distribution illustrates the fact that most of the molecules are in clusters greater than one, and hence from this point of view monomer is not the most important molecular environment. The same prob-

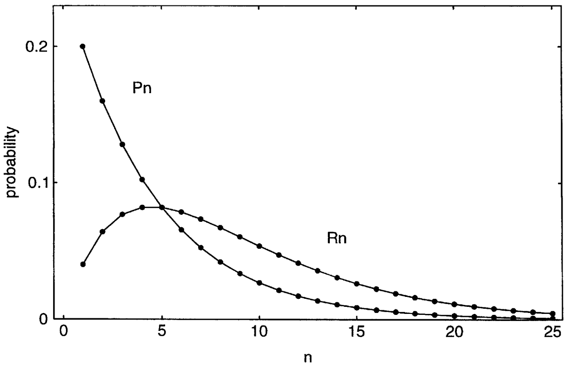

Fig. 2. The probability that a cluster contains n units, P_n, and the probability that a monomer is in a cluster containing n units, R_n, for the KE model. The solid points are obtained from Eq. (51) using the parameters of Eq. (53).

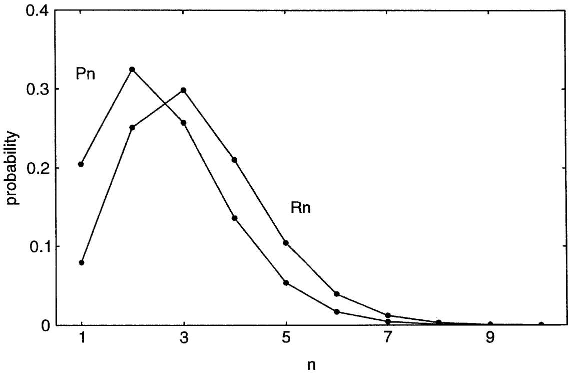

Fig. 3. The probability that a cluster contains n units, P_n, and the probability that a monomer is in a cluster containing n units, R_n, for the KA model. The solid points are obtained from Eq. (52) using the parameters of Eq. (53).

ability distributions are shown in Fig. 3 for the KA model using the equations of Eq. (52) and the parameters of Eq. (53). Again we see that R_n has a maximum at a value of n greater than that for P_n. In this case monomer is not the most probable cluster size in the P_n distribution. Even though the average cluster size is 2.6 units, one sees that most of the molecules are in clusters larger than monomer.

Using the numbers given in Eq. (53) and the general relations of Eq. (44) and Eq. (45) one finds the following values of the first three moments for the KE and KA models (these are all dimensionless numbers):

$$KE:\quad M_1=5,\ M_2=45,\ M_3=605$$

$$KA:\quad M_1=2.6,\ M_2=8.3,\ M_3=31.2 \qquad (54)$$

In this example we have calculated these quantities from Eqs. (44) and (45), but given the data of Fig. 1 as experimental data, one can expand the $\langle n\rangle$ as a function of c, as given in Eq. (5) and then obtain the moments as indicated in Eq. (23) and Eq. (24). The KE model serves as a useful check for our moment equations. Given c and K_e one obtains the quantity u from Eq. (42) and then directly the first three moments from Eq. (44); for $c=10$ mM and $K_e=2$ mM^{-1} one has $u_e=4/5$ and the values of the moments given in Eq. (54). Alternatively one can use Eq. (43) to calculate the derivatives $n^{(1)}$ and $n^{(2)}$ defined in Eqs. (19) and (20). For the values of c and K_e already used one finds $n^{(1)}=20/9$ and $n^{(2)}=820/729$ which when used in Eqs. (24) and (25) again yield the values of the first three moments given in Eq. (54).

Given the sets of three moments given in Eq. (54) one can then immediately construct the maximum-entropy distributions. Fig. 4 shows the P_n calculated from Eq. (32) using the three moments for the KE model given in Eq. (54) based on the parameters given in Eq. (53). The results obtained using one, two, and three moments are shown by solid lines (that are essentially superimposable) and are compared with the exact results (shown by solid dots) given by Eq. (51) and shown in Fig. 2. In this case the distribution function is exactly exponential and so the maximum-entropy form with $g(x)$ of Eq. (28) linear in x (requiring only one moment of the distribution) gives the exact result. Using the second and third moments does not change the distribution. Note that because the distribution is exactly exponential does not mean that the moments beyond the first are zero.

Given the results of Fig. 4, that is P_n as a function of n determined from moments (which in the general case will be obtained from experimental data), one can then use the values of the P_n to calculate the K_n. The equilibrium constant expression for the general reaction of Eq. (1) is

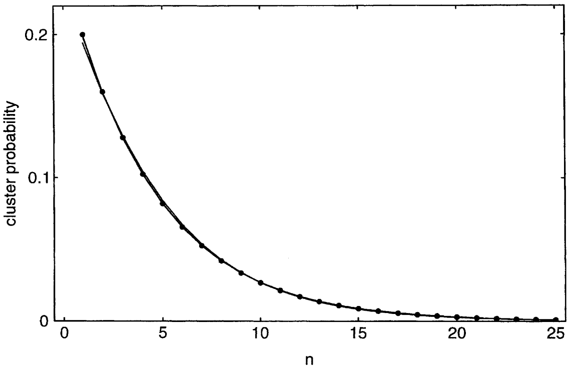

Fig. 4. The probability that a cluster contains n units, P_n, for the KE model using the parameters of Eq. (53) as obtained from the maximum-entropy method using one, two and three moments (solid lines) compared with the exact result obtained from Eq. (52) and shown in Fig. 2. The three maximum-entropy results are essentially identical.

$$K_n = \frac{[A_n]}{[A_{n-1}]\ [A_1]} = \frac{P_n}{P_{n-1}\ [A_1]} \tag{55}$$

Now $[A_1]$ is the concentration of free monomer and is given by

$$[A_1] = P_1 \sum_n [A_n] \tag{56}$$

which gives [using Eq. (4) for c and Eq. (11) for M_1]

$$[A_1] = c\ P_1/M_1 \tag{57}$$

Then given the P_n shown in Fig. 4 one can calculate the K_n. In this manner one finds the arrays (giving the K_n for K_2 through K_6)

one moment: 2.016, 2.016, 2.016, 2.016, 2.016

two moments: 2.006, 2.006, 2.007, 2.007, 2.007

three moments: 2.106, 2.094, 2.082, 2.073, 2.06 (58)

We recall from Eq. (53) that $K_e = 2$ exactly and that is the result that one is getting from all three levels of approximation here. Thus the use of three moments clearly would tell us if a model is obeying the *KE* form.

The *KA* model is more interesting and in Fig. 5 we show the results of using successively one, two and three moments to obtain the P_n from the maximum-entropy method and compare these results with the exact results of Fig. 3 obtained from Eq. (52). The parameters used are given in Eq. (53) and the three moments thus obtained are given in Eq. (54). In Fig. 5a (using one moment) the maximum-entropy distribution is exponential and does not fit the exact distribution well. In Fig. 5b (using two moments) the maximum-entropy $g(x)$ is quadratic and now fits the exact distribution much better. Finally, in Fig. 5c (using three moments) the maximum-entropy $g(x)$ is cubic and the fit with the exact distribution is very good.

Using the data of Fig. 5c with Eqs. (55) and (57) we obtain estimates of the K_n [which are given by Eq. (34) with $K_a = 2$] and are shown in Fig. 6. These results illustrate that the use of empirical sets of three moments such as those given in Eq. (54) can easily distinguish between the *KE* and *KA* models.

We now want to apply this procedure to real data. But before applying this method to actual cases we must address the problem of converting experimental data on the concentration dependence of some optical property into a graph of average cluster size, $\langle n \rangle$, as a function of c, the total concentration.

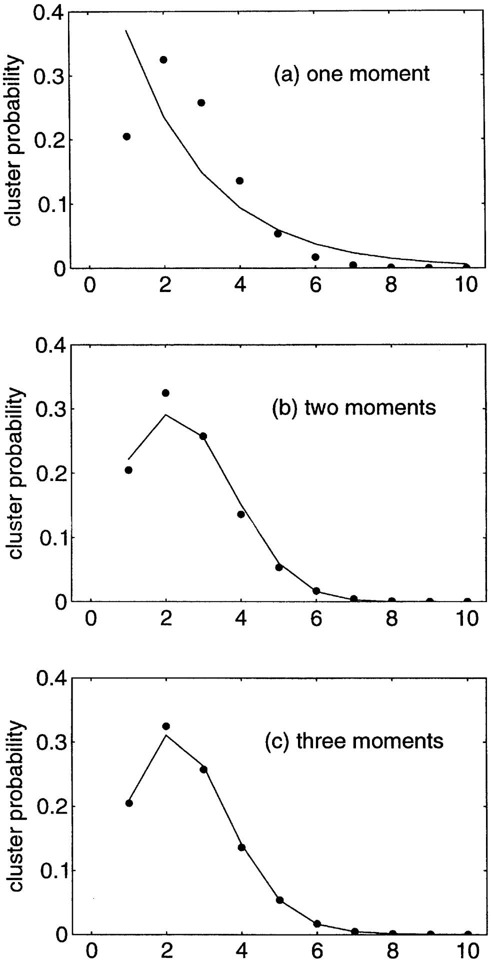

Fig. 5. The probability that a cluster contains n units, P_n, for the *KA* model using the parameters of Eq. (53) as obtained from the maximum-entropy method using (a) one, (b) two and (c) three moments (solid lines) compared with the exact results obtained from Eq. (52) and shown in Fig. 3.

5. Experimental average cluster size

In general the experimental data on self-association gives some optical property of the system as a function of the total concentration of monomers

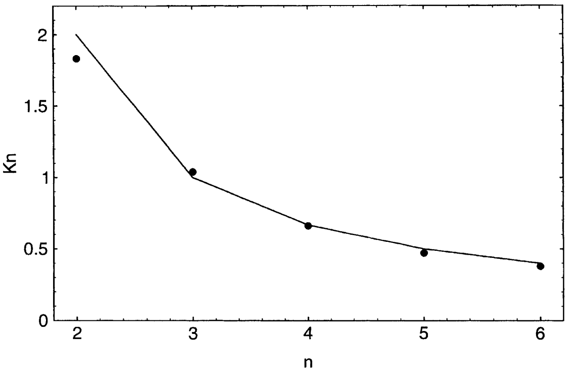

Fig. 6. The K_n (solid dots) obtained from the probability distribution shown in Fig. 5c for the *KA* model. The K_n are obtained from the P_n using equation Eq. (55). The solid line joins the correct values as given by Eq. (34) with $K_a=2$.

used. Examples of the optical properties used, among many, are chemical shift and circular dichroism. We will use the example of chemical shift here. In Fig. 7a we schematically represent a flat molecule by the large rectangle. A single hydrogen in each is further indicated by the smaller rectangle containing an H symbol. In Fig. 7a we show the monomers coming together to form a dimer where we indicate that the chemical shifts of the hydrogens illustrated in each monomer are in general perturbed by a different amount, labeled here α and β. When the extent of stacking is increased, as illustrated in Fig. 7b, we assume that the same perturbations apply (and we also assume that the structure of the stack remains constant as a function of the size of the stack). If δ_o is the unperturbed chemical shift of a particular proton, then in a cluster of n units the chemical shift of the cluster, δ_{cl}, will be

$$\delta_{cl}=n\ \delta_o+(n-1)\ (\alpha+\beta) \tag{59}$$

In a long stack of monomers, the chemical shift of each unit will be

$$\delta_8=\delta_o+(\alpha+\beta) \tag{60}$$

Then Eq. (59) gives

$$\delta_{cl}=\delta_o+(n-1)\ \delta_8=(\delta_o-\delta_8)+n\ \delta_8 \tag{61}$$

The chemical shift per unit in the stack then becomes

$$\delta_{obs}=\delta_{cl}/n \tag{62}$$

or for the average cluster size $\langle n\rangle$

$$\delta_{obs}=\delta_8+(\delta_o-\delta_8)/\langle n\rangle \tag{63}$$

Solving Eq. (63) for $\langle n\rangle$ one has

$$\langle n(c)\rangle \ = \ \frac{\delta_o\ -\ \delta_\infty}{\delta_{obs}(c)-\delta_\infty} \tag{64}$$

Combining Eq. (64) with Eq. (43) for the *KE* model gives Eq. (3) of Scheller et al [2]. Eq. (64) is a simple equation, but it is at the heart of our method for in this equation the experimental data, here given by the observed chemical shift as a function of concentration, $\delta_{obs}(c)$, is converted into the average cluster size as a function of concentration $\langle n(c)\rangle$. And it is the latter quantity that gives us directly the first three moments of the cluster probability distribution and hence the probability distribution itself and the size dependence of the binding constants. The essential assumptions are seen to be that the stack is regular and the perturbations on the optical property depend in a regular way on nearest-neighbor inter-

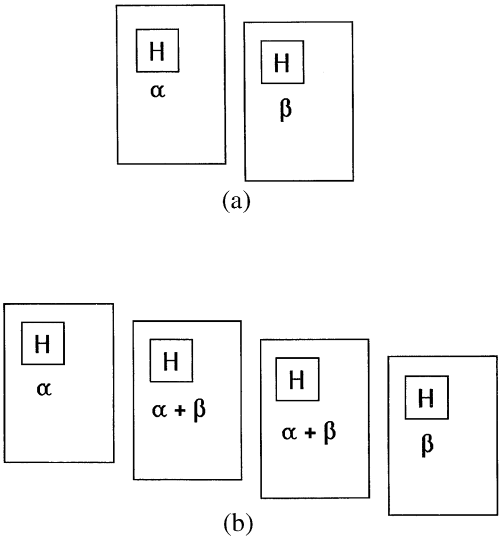

Fig. 7. Schematic illustration of the perturbation of the chemical shift on a given hydrogen in a monomer (large rectangle) that forms stacks. In (a) a dimer is shown with different perturbations α and β for the first and second hydrogen's in the stack. In (b) a regular stack is shown with the assignment of the respective perturbations.

actions. Martin [3] has discussed the case where one takes into account next-nearest neighbor interactions.

There are then two problems here: one is the cooperativity and regularity of the observed optical properties that are needed to give the average cluster size as a function of concentration and the other is the cooperativity and regularity of the thermodynamic parameters of stacking. The two could be completely independent of one another or not. But clearly the conversion of the optical property into the average cluster size as a function of concentration is the crucial step.

6. Daunorubicin

Daunorubicin is an antibiotic that strongly self-associates in aqueous solution. We have chosen it as our first example because in Martin's extensive review of self-associating systems [3], this compound had the largest association constant (in the mM^{-1} range) of all those described. This means that the extent of stacking for this compound is large so that we will be testing our method on a case where there is a big effect. In addition it means that extensive stacking can be achieved for total concentrations in the mM range where we can expect non-ideal effects to be minimal.

We will use the data of Martin [1] who studied association in daunorubicin using circular dichroism giving data obtained at several different wavelengths. The different sets of data closely give the same results so we have picked a single optimal set, namely, the data taken at 450 nm. These data are shown in his Fig. 3b and reproduced in our Fig. 8; the optical property measured is $\Delta\varepsilon$ as a function of the total concentration of daunorubicin in mM. In this case there is an equation analogous to Eq. (64) with $\Delta\varepsilon$ replacing δ_{obs}. Martin [3] notes that the optical properties for this system are well-represented by the assumption of nearest-neighbor interactions only as illustrated in Fig. 7 and described by Eq. (59). The asymptotes are given in Martin's Table 1 [1] and they are

$$\Delta\varepsilon_{\text{o}} = 1.11 \Delta\varepsilon_{\infty} = 5.72/2 \tag{65}$$

Martin interprets his data in terms of a dimerization model and hence his $\Delta\varepsilon_8$ is the limit for pairs of molecules. For a per molecule limit we take his number and divide by two as in Eq. (65). The quantity $\Delta\varepsilon_{\text{o}}$ is the lower dashed line in Fig. 8 while $\Delta\varepsilon_8$ is the upper dashed line.

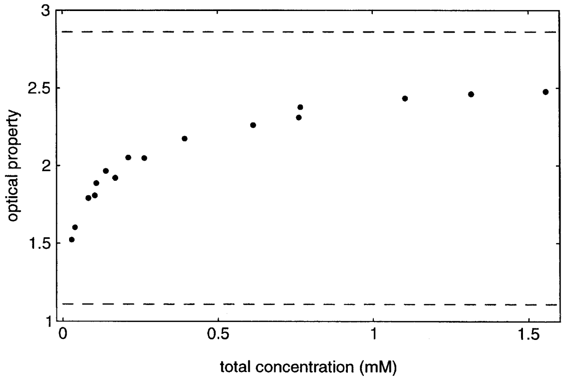

Fig. 8. The circular dichroism as a function of total monomer concentration, c, for daunorubicin as given by Martin [1]. The lower and upper dashed curves, respectively, give the limits $\Delta\varepsilon_{\text{o}}$ and $\Delta\varepsilon_8$ given by Eq. (65).

Given the basic experimental data, usually some optical property as a function of concentration, the next crucial step in our method is to convert this data into the average cluster size as a function of total concentration using the analog of Eq. (64). Using the limits given in Eq. (65) and the data points shown in Fig. 8, the results of this conversion are shown in Fig. 9 where the solid dots represent the same points shown in Fig. 8 transformed into $\langle n \rangle$ as a function of c. The solid curve is a best cubic fit to the data in the range covered by the curve. This curve can then be used to give the expansion shown in Eq. (5) where we take the expansion center as $c_o = 0.75$ mM Given the expansion constants n_{o}, n' and n'' in Eq. (5) one immediately obtains three moments of the cluster probability distribution evaluated at $c_o = 0.75$ using Eqs. (22)–(25). One finds: $M_1 = 3.39$, $M_2 = 21.4$ and $M_3 = 194.3$. The quantity M_1 is the average cluster size for the value of c_o used.

With the three moments given above one can then use the maximum-entropy method we have described earlier to give three successive degrees of approximation (using one, two and three moments) to the cluster distribution function. The results of this calculation are shown in Fig. 10

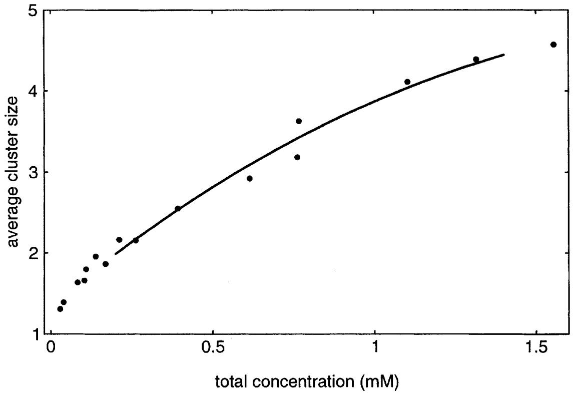

Fig. 9. The conversion of the data given in Fig. 8 for daunorubicin to give the average cluster size, $\langle n\rangle$, as function of total concentration, c using the analog of Eq. (64) and the asymptotic values given in Eq. (65). The solid curve shows a cubic fit to the data in the region covered by the curve that is used to obtain the fit of Eqs. (5) and (6) and hence the first three moments of the cluster probability distribution as given by Eqs. (22)–(25).

where we plot the probability that a cluster has n units, P_n, as a function of n for the distribution function obtained using one, two and three moments. The three approximate distributions are seen to be very similar. At $n = 1$ the top to bottom points are, respectively, for the three, two and one-moment distributions. The qualitative shape of the cluster probability distribution functions shown in Fig. 10 is similar to exponential decrease, so we anticipate that this system is well-described by the *KE* model.

Having the P_n as given in Fig. 10 we use Eq. (55) to calculate the respectively K_n. The results of this calculation are given in Fig. 11 showing K_n as a function of n obtained from the one, two and three-moment distributions. The horizontal constant curve is obtained from the one-moment distribution. Since using one moment gives a probability distribution, from Eq. (28), that is exactly simple exponential, this level of approximation exactly gives the *KE* model. The dashed curve gives the K_n obtained from using the two-moment distribution function. In this case the K_n need not be constant and indeed they are not (increasing slightly). The curved solid line gives the results obtained from the three-moment distribution function and in this case the K's increase slightly and then decrease with n.

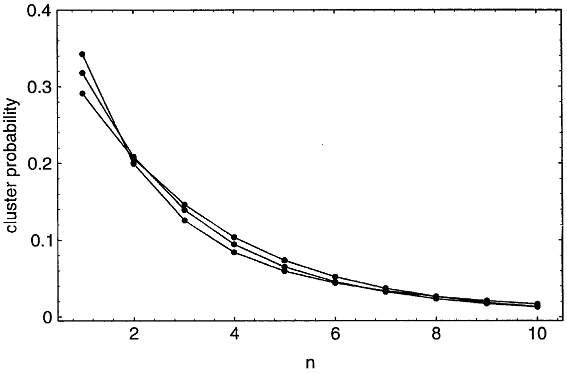

Fig. 10. The cluster probabilities, P_n, for daunorubicin obtained from the maximum-entropy method using one, two and three moments obtained from the solid curve shown in Fig. 9. At $n = 1$ the curves represent, from top to bottom, the results obtained using three, two and one moments.

We observed that the P_n curves in Fig. 10 were very similar in appearance. It is also true that the K_n curves obtained from the data given in Fig. 10 also are very similar, there being no marked difference in the K_n obtained on using one, two or three moments. Thus we conclude that the self-association of daunorubicin is well described by the *KE* model. What is significant about the results

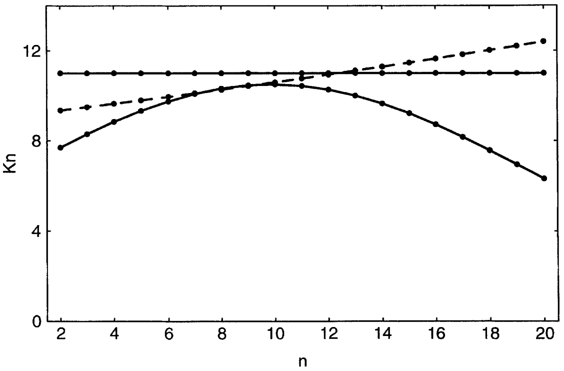

Fig. 11. The K_n for daunorubicin obtained from the P_n of Fig. 10 using Eq. (55) for one, two and three moments. The constant horizontal line gives the results obtained using one moment, the dashed curve gives the results obtained from using two moments, while the curve that goes through a maximum gives the results obtained from using three moments.

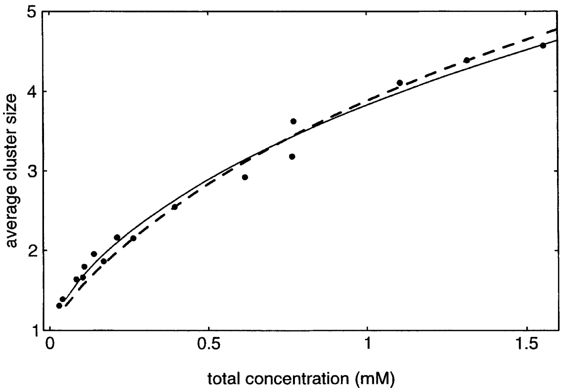

Fig. 12. The data of Fig. 9 for daunorubicin compared with the results obtained from the general Γ of Eq. (8) with the K_n obtained from using one moment (solid curve) and from using three moments (dashed curve).

of Figs. 10 and 11 is that this conclusion is a final result of our analysis and is not an initial assumption.

It is always useful to see if the final results of a calculation agree with the initial data. For this case we take the K_n given in Fig. 11 and use them in the basic quantity Γ of Eq. (8) [calculating the Q_n from the K_n via Eq. (3)]. We then calculate $\langle n \rangle$ as a function of c and compare this with the experimental data given in Fig. 9. The results are shown in Fig. 12 where the solid dots are the original experimental data; the solid curve is the result obtained using the one-moment K's while the dashed curve results from using the three-moment K's. Clearly either set of K's gives a very good fit to the original data, enforcing our conclusion that this system is well-approximated by the *KE* model.

Finally we use the same two sets of K_n to calculate the probability that a monomer is in a cluster of n units. This is the function R_n defined in Eq. (49). The results of this calculation for c_o= 0.75 are shown in Fig. 13 where the solid curve is based on the one-moment K's and the dashed curve is based on the three-moment K's. From Fig. 10 one sees that the most probable cluster is free monomer (n=1). But from Fig. 13 one sees that most of the molecules are in clusters with cluster sizes ranging from two to twenty. Thus the molecular environment most molecules see in this case is the interior of a cluster.

7. $Mg(ATP)^{2-}$

The self-association of ATP is typical of the many systems reviewed by Martin [3], this compound having an association constant in the M^{-1} range and an average cluster size of approximately 2. The system $Mg(ATP)^{2-}$ has a greater tendency to cluster than ATP^{4-} itself, so we have chosen it as our second example. We use the data of Scheller et al. [2] who measured the chemical shifts of several protons as a function of total $Mg(ATP)^{2-}$ concentration. In particular we use the data points given in their Fig. 1 for the proton labeled H2. These data, giving the observed chemical shift, δ_{obs}, as a function of total concentration, c, are reproduced in Fig. 14. In Fig. 14 the dashed lines give the zero and infinite concentration asymptotes the values of which are given in Table 3 and are

$$\delta_o = 8.278 \pm 0.002 \delta_\infty = 6.95 \pm 0.09 \qquad (66)$$

In Fig. 14 the upper and lower dashed curves represent δ_o and δ_8, respectively. Using the data given in Fig. 14, the above limits, and Eq. (64) we can convert the $\delta_{obs}(c)$ data of Fig. 14 into data giving the variation of the average cluster size as a function of c, as shown in Fig. 15. The solid curve in Fig. 15 shows a cubic fit to the data points covered. From this fit we obtain n_o, n' and

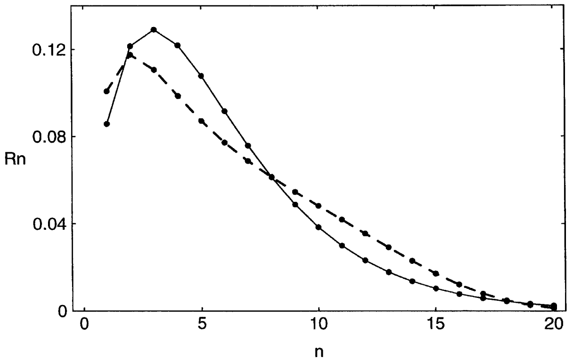

Fig. 13. The probability, R_n, that a monomer is in a cluster containing n units for daunorubicin. The solid curve is the result obtained from the one-moment K_n while the dashed curve is the result obtained from the three-moment K_n.

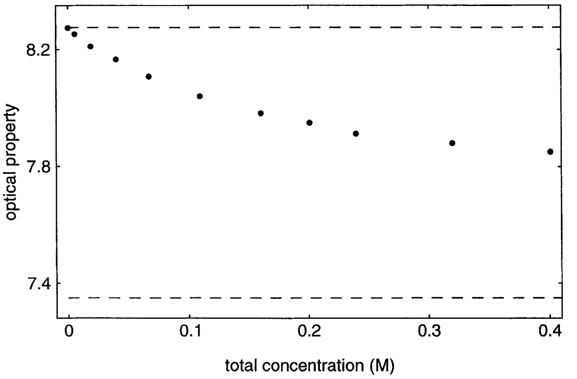

Fig. 14. The chemical shift of H2 for $Mg(ATP)^{2-}$ as a function of total concentration as given by Scheller et al. [2]. The upper and lower dashed lines give respectively the asymptotes δ_o and δ_8 given in Eq. (66).

n'' of Eq. (6) using $c_o = 0.20$ M for our expansion point. These values can be immediately converted into the first three moments of the cluster distribution using Eqs. (22)–(25). The values of the first three moments obtained in this manner are $M_1 = 1.55$, $M_2 = 3.28$ and $M_3 = 9.03$ where M_1 is the average cluster size for the given value of c_o.

Given the moments M_1, M_2 and M_3 we can then calculate the maximum-entropy cluster distributions based on the successive use of one, two and three moments. The cluster distribution functions, P_n, so obtained are shown in Fig. 16. The one-moment distribution gives the lowest value of P_n at $n = 2$; the P_n for the two and three-moment distributions are essentially identical.

As before, given the cluster probability distribution, P_n, one can then calculate the corresponding K_n using Eq. (55). The results of this calculation are shown in Fig. 17 where the constant horizontal line gives the K_n from the one-moment distribution (where this level of approximation

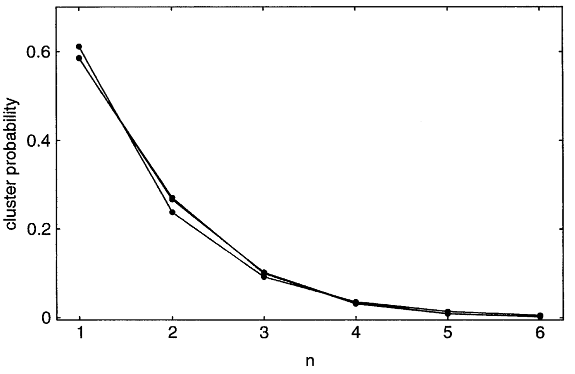

Fig. 16. The probability that a monomer is in a cluster containing n monomers, P_n, for $Mg(ATP)^{2-}$ as obtained from the maximum-entropy method using one, two and three moments. The curves obtained using two and three moments are essentially identical; the one-moment curve has the lowest value shown at $n = 2$.

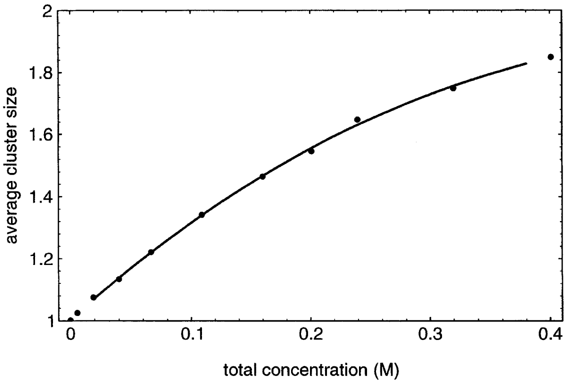

Fig. 15. The conversion of the data given in Fig. 14 for $Mg(ATP)^{2-}$ into the average cluster size, $\langle n \rangle$, as a function of total concentration, c. The solid curve is a cubic fit for use in obtaining the results of Eqs. (5) and (6) and hence the moments given in Eqs. (22)–(25).

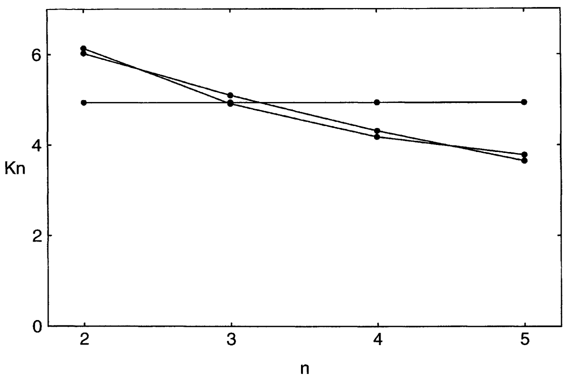

Fig. 17. The K_n for $Mg(ATP)^{2-}$ as a function of n as obtained from the maximum-entropy method P_n given in Fig. 16 using Eq. (55). The constant, horizontal line is the one-moment result; the three moment result is the lower curve in the middle range of variation of n.

gives the *KE* model). The other two curves, that are almost superimposable, are the results obtained from the two and three-moment distributions (the three-moment result is the lower curve in the middle range of variation of n). In this case both the two and three-moment results indicate a deviation from the *KE* model: the K_n do tend to decrease with n but we note that the extent of decrease is not as large as that given by the *KA* model [as in Eq. (34)].

8. Conclusion

We have shown in this paper that experimental data on the variation of some optical property (chemical shift, circular dichroism, etc.) as a function of concentration can be converted into approximate cluster distribution functions using the maximum-entropy method (the approximation improving with the number of moments used) and this information in turn can be used to calculate the K_n. No model concerning the variation of the K_n is used: the variation of the K_n emerges at the end as a final result. We note that this is a general method for any system that self-associates, the only requirement being data on the average extent of clustering as a function of concentration.

Schematically the calculation proceeds with the following steps:

1. Obtain experimental data (usually some optical property such as chemical shift or circular dichroism) that is a measure of the extent of clustering as a function of concentration.
2. Convert the optical data into data giving the variation of the average cluster size as a function of concentration. This process requires a conversion equation [analog of Eq. (64)] and asymptotic values of the optical property at low and high concentrations
3. Fit the $\langle n\rangle$ data obtained above to a quadratic in the concentration and thereby obtain the parameters n_o, n', n'' of Eq. (5) evaluated at c_o.
4. Use Eqs. (22)–(25) to convert the parameters n_o, n', n'' obtained above into the first three moments, M_1, M_2 and M_3, of the cluster probability distribution.
5. Use the moments obtained above with maximum-entropy method to calculate P_n, the cluster probability distribution.
6. Use the mass action relation of Eq. (55) to convert the cluster probabilities given by the maximum-entropy distribution functions into the values of the K_n.

The first step, obtaining reliable experimental data, of course is where the real work is involved. All of the remaining steps in the above scheme, except the second, involve straightforward mathematics and do not involve any model assumptions about the system. The second step is the weak link in the chain, requiring assumptions about the conversion of optical properties into average cluster size. The reliability of the whole approach then rests on the soundness of the second step.

If one has only a single moment of the cluster distribution function, $M_1 = \langle n\rangle$, this is equivalent to using the *KE* model (exponential distribution function). If in addition one knows the slope of the curve $\langle n\rangle$ as a function of concentration then this is enough information to calculate the first and second moments of the distribution function which is, in turn, enough information to resolve deviations from the *KE* model.

We note that the treatment of self-association we have given here is very similar to our treatment of ligand binding in biopolymers [4–6]. The major difference between the two cases is that in the treatment of binding isotherms the number of binding sites is constant and in addition one measures extent of binding (analog of average cluster size) as a function of the concentration of free monomer in solution (this is true even if the independent variable is the total monomer concentration since the conservation relation for ligand binding gives the concentration of free monomer).

References

[1] S.R. Martin, Absorption and circular dichroic studies on the self-association of daunorubicin, Biopolymers 19 (1980) 713–721.

[2] K.H. Scheller, F. Hofstetter, P.R. Mitchell, B. Prijs, H. Sigel, Macrochelate formation in monomeric metal ion complexes of nucleotide 5′-triphosphates and the promotion of stacking by metal ions. Comparison of the self-association of purine and pyrimidine 5′-triphosphates using proton nuclear magnetic resonance, J. Am. Chem. Soc. 103 (1981) 247–260.

[3] R.B. Martin, Comparisons of indefinite self-association models, Chem. Rev. 96 (1996) 3043–3064.

[4] D. Poland, Ligand-binding distributions in biopolymers, J. Chem. Phys. 113 (2000) 4774–4784.

[5] D. Poland, Ligand binding distributions in nucleic acids, Biopolymers 58 (2001) 477–490.

[6] D. Poland, Protein-binding polynomials, J. Protein Chem. 90 (2001) 91–97.

[7] A. Tagliani, On the application of maximum entropy to the moment's problem, J. Math. Phys. 34 (1993) 326–337.