ELSEVIER

Biophysical Chemistry 108 (2004) 141–163

Biophysical Chemistry

www.elsevier.com/locate/bpc

Characterization of macromolecular heterogeneity by equilibrium sedimentation techniques☆

Yujia Xu*,[1]

Department of Chemistry, Rutgers University, Piscataway, NJ 08854, USA

Abstract

New graphical procedures have been developed to investigate the heterogeneity of protein preparations using sedimentation equilibrium. The heterogeneous systems that can be studied include self-associating systems contaminated by incompetent monomer, self-associating systems contaminated by non-dissociating oligomer and simple non-interacting monomer–oligmer disperse systems. The new procedures are based on the concentration dependence of the apparent association constants estimated by a non-linear least square fitting program (NONLIN), on the assumption of conservation of mass during sedimentation and on the applications of several standard techniques for statistical inferences of NONLIN estimations. The procedures outlined here can detect various types of heterogeneity, discriminate amongst different types of heterogeneity, estimate the amount of contaminant causing heterogeneity and determine the true equilibrium constant of the self-associating components. The procedures appear to be sensitive, accurate and easily applicable when tested using both protein samples and computer simulated data.

Keywords: Heterogeneity in self-associating system; Detection of heterogeneity by sedimentation equilibrium; Incompetent monomer; Incompetent dimer; Secondary plots

1. Introduction

Heterogeneity corresponds to the presence of more than one thermodynamic component as solute [1–3]. Preparations of self-associating macromolecule often contain adventitious macromolecules. These contaminating molecules may take on several different forms and can become difficult to detect if the contaminants have a size similar to that of one of the main components of the self-associating system [3]. This type of heterogeneity may occur during protein purification procedures when some of the monomer or polymer molecules become damaged and lose their ability to self-associate (or to dissociate) or if their association constant(s) are otherwise altered [1,2]. The presence of such contaminants can significantly affect estimates of equilibrium constants and of the corresponding standard free energies of

☆ All the work presented here was accomplished while I was a graduate student in David Yphantis's lab. He was also actively involved in the drafting and editing of part of this manuscript. He subsequently encouraged me to publish it myself. I could not have done any of it without his long time encouragement, guidance and trust, for that, I am deeply indebted to him.

*Corresponding author. Tel.: +1-732-445-5846; fax: +1-732-445-5312.

E-mail address: yujiaxu@rutchem.rutgers.edu (Y. Xu).

[1] Present address: Department of Chemistry, Hunter College, the City University of New York, 695 Park Ave., New York, NY 10021, USA. Tel.: +1-212-772-4310; fax: +1-212-772-5332; e-mail address: yujia.xu@hunter.cuny.edu.

0301-4622/04/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2003.10.015

self-association. Such damaged molecules may be strongly allergenic and may cause severe side effects when preparations containing these macromolecules are used as pharmaceuticals [2,4]. Few methods exist to detect this type of heterogeneity. Quantitative estimation of the level of contaminants in such preparations and precise estimation of the equilibrium constants of the interacting components can be difficult.

In this paper, new methods will be presented for the detection and the characterization of three different types of heterogeneous self-associating systems based on sedimentation equilibrium and the estimation procedures of NONLIN, a nonlinear least squares fitting program used extensively for the study of self-associating behavior of macromolecules [5]. It was first pointed out by Squire and Li [1] that heterogeneous self-associating systems consists of more than one thermodynamic component[2] will exhibit dependence of the apparent self-association constant on the loading concentrations and on ultracentrifuge speeds. Variations in the observed self-association constant for samples examined at different loading concentrations and/or at different centrifugation speed thus can be used to detect heterogeneity in self-associating systems, provided that no other sources of variation of the apparent association constants exist other than heterogeneity [2–4,6].

The sensitivity and accuracy of this procedure depend critically on the uncertainties of the estimated parameters, especially the apparent association constant(s), and on the rate of change of these apparent equilibrium constants with variations in loading concentration [2,3] and/or speed. In the first part of the paper, the reliability and precision of the estimations of NONLIN will be analyzed. The applicability of standard statistical inference techniques for the detection of the heterogeneity based on the variations of the observed association constant will be investigated. The performance of such detection scheme has been evaluated using computer simulation data for different heterogeneous systems and under different experimental settings. The second part of the paper will focus on the analysis of the effects of different heterogeneity on the observed association constant(s). Based on this analysis, a set of graphic procedures has been developed to further characterize the heterogeneous self-associating systems. Although the procedures and results presented here are based on the estimations of NONLIN, some of the approaches are fairly general in nature and should be applicable to other estimation procedures for sedimentation equilibrium.

2. Material and methods

2.1. Computer simulation

Computer simulations were conducted using data sets generated with known parameters using the program EQIGEN.[3] All data sets were truncated, as necessary, to ensure that each simulated data set excluded data points with refractive index gradients exceeding the effective resolution of the Rayleigh optical system.[4] The concentration scale used throughout this work was mg/ml (g/l). Fringe displacements (*fr*) are closely proportional to solute concentration, for example, using interferometry at 785 nμ in 12-mm thick centerpieces, 1 *fr* = 0.36 mg/ml [10]. Pseudo-random noise with known mean and variance was added to each data set. The standard deviation of the noise was taken to be 0.01 to 0.015 *fr* for equilibrium sedimentation data, a range similar to the typical fitting RMS in routine interferometric sedimentation equilibrium experiments. The mean was taken to be zero.

[2] Strictly speaking this approach requires observations at constant temperature, constant pressure and constant solvent composition [3]. In most cases the solvent can safely be considered to be a single component and the requirement of constant pressure may be relaxed if there is no volume change on association. (Should there be a significant volume change on association its effects can be compensated for by calculation.)

[3] An updated PC-DOS version of the data generator program is available as NONEQ4.EXE on both the FTP RASMB server site (bbri.eri.harvard.edu) and our FTP server site (spin6.mcb.uconn.edu). This version provides concentration distributions for both ideal and non-ideal self-associating systems.

[4] A maximum refractive index gradient of about 100 *fr* cm^{-1} was used. It has been our experience with both the Model E and the XLI ultracentrifuges that data with higher refractive index gradients cannot be considered reliable.

The protocols for Monte Carlo simulation described in 'Numerical Recipes' [7] were used for this work (the compiled programs for VAX/VMS are available up on request). Detailed descriptions of the programs and of their proper use are given elsewhere [8].

2.2. Curve fitting procedures using NONLIN

A few features of data analysis using NONLIN will be re-emphasized here. They are crucial for application of the hypothesis testing technique and secondary plots. The functional form of NONLIN for a self-associating system [5] is:

$$Y(r)=\delta y+\sum_{j=1}^{n}\left\{\exp\left[\ln\ K_j/L(j)+\ln\ C_1(r')+\sigma_1\left(r^2/2-r'^2/2\right)\right]-B(Y(r)-\delta y)\right\}^{L(j)} \quad (1)$$

where $Y(r)$, the observable or dependent variable, is the fringe displacement or absorbance (assumed to be proportional to the total concentration); δy is a constant offset in Y; r is the radial position in the ultracentrifuge cell; r' refers to an arbitrary reference point (usually taken to be the radius of the first data point of a data set in NONLIN); $C(r')_1$ is the concentration of monomer at the reference position (in units of Y); the $L(j)$[5] are the degrees of association of the j-th oligomer; n is the number of molecular species (oligomers) present in the sample (the maximum value of n in NONLIN is five, including the monomer); the K_j's are the association constant(s) with K_1 taken as unity; B is the second virial coefficient (in units of Y^{-1}); and $\sigma_1=M_1(1-\bar{V}\rho)\times\omega^2/RT$ is the effective reduced molecular weight of the monomer [9]. Here M_1 is the monomer molecular weight; ω the angular velocity; T the absolute temperature; R the gas constant; $\bar{V}$ the partial specific volume and ρ the density of the solution.

The most frequently used current version of NONLIN simultaneously fits up to 15 data sets with individual concentration scales and/or speeds using global parameters—the joint fit. In a joint fit the estimation vector ($\mathbf{\Phi}$) consists of the individual values of δy and of $C_1(r')$ for each data set, in addition to the common parameters of σ_1, the $L(j)$ and B. The logarithms of the association constants, the $\ln K_j$'s, can be treated as common parameters for homogeneous systems; these $\ln K$'s are referred to as the common $\ln K$.

One can also conduct joint fits with individual values of $\ln K$'s for each data set; these individual $\ln K$'s are treated as independent, separate parameters (for each data set) and are referred to as the separate $\ln K$'s. Since the apparent values of the $\ln K$ are no longer assumed to be identical this type of joint fit is useful in characterizing heterogeneous systems. In addition to the joint fit one can fit the data of each channel separately, as an individual fit. The estimations from the individual fits are usually lack of precision but they may provide more detail and more intuitive information helpful in the interpretation of the fitting results.

2.3. Detection of heterogeneity and the ROCs plots

The application of hypothesis testing to detect heterogeneity of the sample is described in Section 3. To conduct the hypothesis testing the data has to be fitted twice: the original fit, the fit that chooses the best model for the data, and the H-fit, the fit based on the hypothesis to be tested. In the original fit more than two data sets corresponding to different loading concentrations and/or speeds are fitted jointly with separate values of the $\ln K$ and the minimized sum of residuals squared ($S(\mathbf{\Phi})$ in Eq. (4) is calculated. To test the hypothesis that the separated values of the $\ln K$ are equal to each other in a statistical sense, in the H-fit these separate $\ln K$'s (the $\mathbf{\Phi}_2$) were set equal to a constant C (the $\mathbf{\Phi}_2^c$) and a value of $S(\mathbf{\Phi}^c)$ is obtained. In the next step, the test statistic (F) is calculated using Eq. (4) and the probability of H_0 being true (the value of false alarm rate α) is determined from the table for standard F-statistics.

The performance of the test of hypothesis approach for the detection of heterogeneity is evaluated using the ROCs plots (details in Section 3.2) based on Monte Carlo simulations. First, the

[5] It is usually assumed that $L(j)=j$, but for pressure-dependent associations (where the partial specific volume of the j-mer, V_j, differs from $\bar{V}$ of the monomer), the $L(j)$ are defined as: $\bar{V}_j(1-\bar{V}_{j\rho})/(1-\bar{V}_{1\rho})$.

Monte Carlo simulation of 100–500 test of hypothesis was conducted on a given sample under specified experimental conditions. Then, the number of tests having a false alarm rate smaller than or equal to a selected level of significance was counted. This number is proportional to the probability of detection. For the convenience of conducting automated Monte Carlo simulations the value of the common ln K estimated from samples with different loading concentrations was used as the constant C (details see Section 3.2). For homogeneous samples, the constants C for the tests were set to equal to the true (known) values of ln K's used for data generation.

2.4. The secondary plots

For each heterogeneous system, six sets of simulated data corresponding to different loading concentrations were generated as described above with selected values for the radial position of the meniscus (r_m) and the base (r_b). The data were then fitted to an ideal monomer–dimer model of NONLIN and the estimation parameters, together with the values of r_m and r_b were used to construct the secondary plots. The linear fits of the secondary plots were conducted using program Origin 6.0 by Microal®. The error bars were propagated from the confidence intervals of 67% confidence probability of NONLIN using Taylor expansion to the first order.

2.5. Characterization of the BSA sample

A commercial sample of BSA from Armor Pharmaceutical company was used, Lot No W69312. The BSA was examined in the following isoelectric buffer consisting of 0.15 M sodium chloride, 0.02 M sodium acetate, 0.03 M acetic acid and 2 mM ethylenediamine tetraacetic acid disodium salt (EDTA), pH 4.4 (ionic strength $I \sim 0.185$). The density of the buffer was taken to be 1.006 according to previous work [9]. The specific volume of the protein is calculated to be 0.734.

Equilibrium sedimentation runs were conducted on a Model E analytical ultracentrifugation machine equipped with Rayleigh interference optics and a real-time data acquisition device [10]. The radial position of the meniscus (r_m) and the base (r_b) of the solution column were determined using a graphic software IMAGES run on IBM PC, which allows the fringe pattern to be examined pixel by pixel on a gray scale. To avoid possible complication caused by high concentration gradient, the fringe patterns used for determination of r_m and r_b were taken at a very low speed (8000 rev./min). Three to four readings of the radial position were taken from the graphic image for each measurement and the average values were used to construct the secondary plots. The initial loading concentration of the samples was estimated from the absorbance at 280 nm.

3. Detection of heterogeneity

3.1. Asymptotic properties of non-linear estimations

It is generally difficult to attain the accurate statistical information about a non-linear fitting process by theoretical analysis, since the covariance matrix is not directly obtainable [5,11]. However, when the available number of data points is large, non-linear estimation procedures may become asymptotically unbiased, asymptotically efficient [11,12] and the covariance matrices can be approximated. Under certain regularity conditions, estimation parameters by a non-linear least-squares algorithm are also likely to be asymptotically normally distributed, i.e. asymptotically Gaussian [12]. With these asymptotic properties, several standard statistical techniques for linear regression can be applied to make statistical inferences about the non-linear estimation parameters.

The asymptotic property of a non-linear estimation procedure can be explored using the Rao–Cramér theorem, which establishes a theoretical lower bound on the attainable covariance matrix of a non-linear estimator [11,12]. Under the condition that the random Gaussian noise of a data set is independently and identically distributed with zero mean and unknown but equal variance, the lower bound (also called the minimum variance bound: **MVB**) can be estimated by [11]:

$$\mathbf{MVB} = s^2(\boldsymbol{P}^T\boldsymbol{P})^{-1} \quad (2)$$

Here s^2 is the variance of fit and the matrix $(\boldsymbol{P}^T\boldsymbol{P})$ is also known as the information matrix [5,13], which is directly available from the fitting procedure. An (non-linear) estimator is said to be asymptotically efficient if it satisfies the following condition with probability one [11]:

$$\lim_{m\to\infty} m(\boldsymbol{\Sigma} - \mathbf{MVB}) = 0 \quad (3)$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the distribution of the estimation parameters, available only through Monte Carlo approach [14], and m is the number of data points used for estimation. According to Eq. (3) the covariance matrix of an asymptotically efficient non-linear estimation procedure can be approximated using the **MVB** matrix.

3.2. Hypothesis testing

Hypothesis testing is generally used to draw conclusions about the estimation parameters based on the knowledge of the probability distribution of the parameters. The goal of the test is to decide which one of the two complementary hypotheses termed the null hypothesis (H_0) and the alternative hypothesis (H_1), respectively, is likely to be true and to determine the degree of accuracy in accepting/rejecting the hypothesis. According to Seber and Wild [12], a maximum likelihood-ratio test can be constructed for an asymptotically efficient non-linear least-squares estimation procedure with asymptotically Gaussian (normally) distributed estimates:

H_0 (null hypothesis): $\boldsymbol{\Phi}_2 = \boldsymbol{\Phi}_2^c$

H_1 (alternative hypothesis): $\boldsymbol{\Phi}_2 \neq \boldsymbol{\Phi}_2^c$

$$\text{Test statistics:} \quad F = \frac{\left[S(\boldsymbol{\Phi}^c) - S(\boldsymbol{\Phi})\right]}{S(\boldsymbol{\Phi})} \frac{m-p}{p_2} \quad (4)$$

Here $\boldsymbol{\Phi}_2$ (with dimension p_2) is a subset of the estimation parameter vector $\boldsymbol{\Phi}$ (with dimension $p > p_2$), $\boldsymbol{\Phi}_2^c$ is a known vector of constants with the same dimension as $\boldsymbol{\Phi}_2$; $S(\boldsymbol{\Phi})$ is the minimized sum of residuals squared of the estimation, $S(\boldsymbol{\Phi}^c)$ is the minimum value of $S(\boldsymbol{\Phi})$ under H_0 (i.e. under the constrain that the parameters in $\boldsymbol{\Phi}_2$ equal to a set of constants); m is the number of data points and p is the number of estimation parameters. The test statistics F follows F-statistics asymptotically with p_2 and $(m-p)$ degrees of freedom when H_0 is true. Thus, for a given value of F the probability of H_0 being true can be obtained from the statistics table, denoted α (also known as the 'false alarm rate' or the 'significance level' of the test), and the probability of H_1 being true is equal to $(1-\alpha)$.

The decision making of a test (of hypothesis) is based on the value of F and the corresponding false alarm rate α. It is clear from Eq. (4) that when H_0 is true the value of F would be close to 0 ($S(\boldsymbol{\Phi}^c) \approx S(\boldsymbol{\Phi})$) and the corresponding probability α close to 1. Accepting H_1 with a large value of α would indicate a high risk of making a false alarm, i.e. accepting H_1 when it is false (the probability of H_1 being true is $(1-\alpha)$). When F is large, however, it is an indication that $S(\boldsymbol{\Phi}^c) \neq S(\boldsymbol{\Phi})$ ($F \geq 0$ since $S(\boldsymbol{\Phi}^c) \geq S(\boldsymbol{\Phi})$ by definition) and the corresponding α is close to 0 (the probability of H_1 being true close to 1). A small value of α of 0.01 or 0.05 is generally considered suitable for accepting H_1 (rejecting H_0) with a little chance (1–5%) of making a false alarm.

The false alarm rate of a test only tells one side of the story. As in any decision making process, the acceptance or rejection of the null hypothesis (H_0) will run a risk of making one of two types of mistakes. The false alarm is considered the type I error, i.e. to reject the H_0 when it is true (or accepting H_1 when it is false). A type II error is to accept the H_0 when it is false (reject H_1 when it is true). The power of a test is the probability of accepting H_1 when it is true, which can be evaluated from Monte Carlo approach. The performance of a test is judged by its power given a level of false alarm rate α. Such performance is often evaluated using a parametric plot of probability of detection (the power of the test) against the false alarm rate, also known as the ROCs plot (borrowed from radar terminology for the receiver operating characteristics) [15].

3.3. Detection of heterogeneity

For the purpose of detecting heterogeneity of a sample, the null hypothesis can be described as:

$$H_0\colon \boldsymbol{\Phi}_2 = [\ln K_n(1),\ \ln K_n(2),\ldots \ln K_n(p_2)] = C\boldsymbol{I}_{p2} \tag{5}$$

Here $\boldsymbol{I}_{p2}$ is an identity vector with dimension p_2, K_n is the apparent association constant for monomer–n-mer association estimated with different loading concentration specified by (1), (2),…(p_2) (the test is on $\ln K_n$ not on K_n, since the former is directly obtainable form NONLIN fitting (Eq. (1))), and C is a constant independent of loading concentration. Thus, the test is about the (null) hypothesis that the sample is homogeneous and any variations in the estimated $\ln K_n$ are caused solely by the random noise. Rejection of this hypothesis would indicate the sample is heterogeneous with the probability $(1-\alpha)$.

4. Characterization of heterogeneous system

Having demonstrated that the inherent statistical uncertainties of NONLIN estimation parameters can be described by the asymptotic properties of a non-linear least square estimation procedure, the focus of this section is to show how different heterogeneity of the sample will affect the estimated values of the apparent association constant. It is important to note that the function used in NONLIN for self-associating systems assumes a single thermodynamic component [5]. Much of the work presented here, then, is testing what happens when the data do not conform to this assumption.

4.1. The estimation parameters of NONLIN

As it is for any estimation procedures, the physical meanings of the estimation parameters are dictated by the functions used for fitting. Two features of NONLIN functional form are of particular interests. First, the concentration of n-mer of a self-associating system at equilibrium in a centrifuge cell was expressed in NONLIN as [5]:

$$C_n(r) = K_n C_1(r)^n \tag{6}$$

here $C_n(r)$ and $C_1(r)$ are, respectively, the concentration of n-mer and monomer at radial position r, and K_n is the association constant. When used to fit data of homogeneous self-associating systems (system with a single component), NONLIN provides direct estimation of the association constant(s). However, the expression of Eq. (6) is also applicable to solutes not at equilibrium with each other, since the ratio shown in Eq. (7) below is a constant inside a centrifuge cell for any two components with molecular weight ratio of n,

$$q_n = \frac{C_B(r)}{C_A(r)^n} = \frac{C_B(r')}{C_A(r')^n}\frac{\exp\left[\sigma_B\left(r^2/2 - r'^2/2\right)\right]}{\exp\left[n\sigma_A\left(r^2/2 - r'^2/2\right)\right]} = \frac{C_B(r')}{C_A(r')^n}. \tag{7}$$

here σ_A and σ_B are the reduced molecular weight of components A and B, respectively, and $\sigma_B = n\sigma_A$. In case of a homogeneous self-associating system governed by mass action law, the q_n in Eq. (7) is the association constant and can be use to describe the original solution directly, i.e.:

$$\frac{C_n(r)}{C_1(r)^n} = \frac{C_n^0}{C_1^{0n}} = K_n \tag{8}$$

Here C_1^0 and C_n^0 are the concentration of monomer and n-mer, respectively, of the original solution. For a non-interacting system or a heterogeneous system, however, the value of q_n varies with loading concentration and sedimentation speed and Eq. (8) does not hold in general.

The second feature of NONLIN relates to the expression of the concentration distribution in NONLIN functional form (Eq. (1)). Each component in Eq. (1) is identified by the molecular weight (more precisely the reduced molecular weight σ) in the exponent. Therefore, components with similar reduced molecular weight σ would appear to be the same solutes to NONLIN, except the pre-exponential coefficient in NONLIN fitting

function which would be the sum of the pre-exponential terms of all the species:

$$C_t(r) = \sum A_i \exp\left[\sigma_i\left(r^2/2 - r'^2/2\right)\right] = \left(\sum A_i\right)\exp\left[\sigma_A\left(r^2/2 - r'^2/2\right)\right] \quad (9)$$

when $\sigma_i = \sigma_A$ for all i.

These properties of NONLIN functional form allow the program to be directly applicable to an array of systems with multi-components, except in each case the estimation parameters will take on different meanings depending on the composition of the system under investigation. Most noticeably, the apparent values of association constant(s) will vary with loading concentration and/or speed unless the sample is a homogeneous (pure) self-associating system. However, useful ways of characterizing various heterogeneous systems by sedimentation equilibrium can be devised based on the dependence of the apparent association constant on the loading concentrations and/or speeds. A few such examples will be presented below.

The following discussion will be restricted to heterogeneous systems containing two components: the self-associating component A of the primary interest and a small amount of contaminant B, which has molecular weight and other molecular properties nearly identical or closely related to that of A. For simplicity, only the case involving monomer–dimer association will be presented in detail, generalization of some of the treatment to a monomer–n-mer association will be pointed out when applicable. The degree of heterogeneity of each system will be characterized by the parameter, θ, defined as the weight fraction of the contaminant component B in the original solution:

$$\theta = \frac{C_B^0}{C_A^0 + C_B^0} \quad (10)$$

here C_A^0 and C_B^0 are the (total) weight concentration of components A and B, respectively, in the original solution, and the weight fraction of species A is $\lambda = (1-\theta)$.

4.2. Case 1: a disperse system containing non-interacting monomer and dimer

A disperse system normally refers to a multi-component system with non-interacting solutes. Such a system become interesting when one component appear to be the oligomer of the other, such as in a solution containing A and B with the molecular weight $M_B = 2M_A$ (more precisely, $\sigma_B = 2\sigma_A$). The sedimentation equilibrium data of this disperse sample at a single loading concentration and speed will fit equally well to the monomer–dimer model of NONLIN as that of a homogeneous self-associating system. Only when experiments are carried out with different loading concentrations and/or at different sedimentation speed, it becomes possible to reveal that the two components in a disperse system are not at equilibrium with each other.

At sedimentation equilibrium, the distribution of the total concentration of the disperse system can be expressed by:

$$C_t(r) = C_A(r')\exp\left[\sigma_1\left(r^2/2 - r'^2/2\right)\right] + q_2\left\{C_A(r')\exp\left[\sigma_1\left(r^2/2 - r'^2/2\right)\right]\right\}^2 \quad (11)$$

here q_2 is as defined in Eq. (7) with $n = 2$. Comparing Eq. (11) with Eq. (1) of NONLIN functional form for $n = 2$ (assuming the second virial coefficient B negligible), one obtain the following NONLIN estimation parameters: $C_1(r')^{\mathrm{NL}} = C_A(r')$ the concentration of component A at the reference position r'; $\sigma_1^{\mathrm{NL}} = \sigma_A$ reduced molecular weight of component A; $2 \times \sigma_1^{\mathrm{NL}} = \sigma_B$ the reduced molecular weight of component B; $K_2^{\mathrm{NL}} = q_2$ a value that varies with loading concentration and speed. The superscript 'NL' denotes NONLIN estimation parameters.

Based on the principle of conservation of mass [16] it can be written for component A:

$$C_A^0 = \frac{2}{r_b^2 - r_m^2}\int_{r_m}^{r_b} C_A(r)\, r \; \mathrm{d}r = C_A(r')I_A \quad (12)$$

here r_m and r_b are the radial position of the meniscus and the base of the cell, respectively,

and I_A is defined as:

$$I_A \equiv \frac{2}{\sigma_A(r_b^2 - r_m^2)}\left[\exp\left(\sigma_A r_b^2/2\right) - \exp\left(\sigma_A r_m^2/2\right)\right] \tag{13}$$

Similar expressions can also be derived for component B and defining $\gamma = I_B/I_A^2$, and the NONLIN estimate of K_2^{NL} can be relate to the compositions of the original solution by:

$$K_2^{\mathrm{NL}} = \frac{C_B(r')}{C_A(r')^2} = \frac{\theta}{(1-\theta)}\frac{1}{C_t^0\gamma} \tag{14}$$

Clearly the apparent association constant of NONLIN (K_2^{NL}) for a disperse system will change with the loading concentration C_t^0 and/or speed (γ is a function of speed). As shown in Eq. (14) the plot of K_2^{NL} vs. $1/C_t^0\gamma$ is a straight line for a disperse system, and the weight fraction of component B in the original solution (the θ) can be determined from the slope of the plot. This characteristic linear plot of a disperse system will be termed the DISP plot.

4.3. Monomer–dimer association system contaminated by incompetent monomer—the heterogeneous I system

The heterogeneous I system will be used to refer to a system containing the self-association component (A) with a sub-population of monomer that lacks the ability to associate (the component B), designated the 'incompetent monomer' [2]. After equilibrium is reached the observed concentration as a function of radial position r, with $\sigma_B = \sigma_{A1}$, can be described as:

$$C_t(r) = \{C_{A_1}(r') + C_B(r')\}\exp\left[\sigma_{A_1}\left(r^2/2 - r'^2/2\right)\right] + K_2 C_{A_1}(r')^2 \exp\left[2\sigma_{A_1}\left(r^2/2 - r'^2/2\right)\right] \tag{15}$$

Comparing Eq. (15) with Eq. (1) for a monomer–dimer model ($n=2$), it is clear that this heterogeneous I system, can be fit by NONLIN with the following estimation parameters:

$$C_1(r')^{\mathrm{NL}} = (C_{A_1}(r') + C_B(r')) \tag{16}$$

$$K_2^{\mathrm{NL}} = \frac{C_{A_2}(r')}{\left[C_{A_1}(r') + C_B(r')\right]^2} \tag{17}$$

and $\sigma_1^{\mathrm{NL}} = \sigma_{A_1}$ the reduced molecular weight of the monomer.

The Eq. (12) is applicable to the conservation of mass of the incompetent monomer B. While for the self-associating component A, the total mass of the monomer and the dimer is conserved:

$$\begin{aligned} C_{A,t}^0 &= \frac{2}{\left(r_b^2/2 - r_m^2/2\right)}\int_{rm}^{rb}\left[C_{A_1}(r) + C_{A_2}(r)\right]r\ \mathrm{d}r \\ &= C_{A_1}(r')I_A + K_2 C_{A_1}(r')^2 I_{A_2} \end{aligned} \tag{18}$$

where I_A and I_{A_2} are defined by Eq. (13) for A and A_2, respectively. Solving for $C_{A_1}(r')$ from Eq. (18), and taking use of the following expressions $C_A^0 = (1-\theta)C_t^0$, $C_B^0 = \theta C_t^0$, $I_A = I_B$ and $\gamma = I_{A_2}/I_{\mathrm{A}}^2$, it is clear that:

$$\frac{C_B(r')}{C_{A_1}(r')} = \frac{C_B^0}{I_B}\frac{2K_2 I_{A_2}}{\sqrt{I_A^2 + 4K_2 I_{A_2} C_{A,t}^0} - I_A} = \frac{(\theta)}{2(1-\theta)}\left(\sqrt{1 + 4(1-\theta)K_2\gamma C_t^0} + 1\right). \tag{19}$$

Substituting Eq. (19) into Eq. (17) and rearranging, the NONLIN estimate K_2^{NL} can be expressed as a function of the total loading concentration (C_t^0), the true self-association constant of component A (K_2) and the degree of heterogeneity (θ) of the original solution:

$$\frac{1}{\sqrt{K_{2\mathrm{app}}}} = \frac{1}{\sqrt{K_2}}\cdot\left[1 + \frac{(1-\alpha)}{2\alpha}\left(\sqrt{1 + 4K_2\alpha\gamma C_t^0} + I\right)\right]. \tag{20}$$

It can be demonstrated that the value of K_2^{NL}

approaches to $(1-\theta)^2(K_2)^{1/2}$ as C_t^0 approaches to zero:

$$\underset{C_t^0\to 0}{Lim}\frac{1}{\sqrt{K_2^{\mathrm{NL}}}}=\frac{1}{\sqrt{K_2}}\left[1+\frac{\theta}{(1-\theta)}\right]=\frac{1}{\sqrt{K_2}}\frac{1}{(1-\theta)}\ \equiv\ t$$

and the limiting slope of the plot $(1/K_2^{\mathrm{NL}})^{1/2}$ vs. γC_t^0 at $\gamma C_t^0\to 0$ is:

$$\underset{\mathrm{C}_t^0\to 0}{\mathrm{Lim}}\frac{\partial\left(1/\sqrt{K_2^{\mathrm{NL}}}\right)}{\partial \mathrm{C}_t^0}=\theta\ \sqrt{K_2}\ \equiv\ s.$$

Thus, potentially, the true equilibrium constant (K_2) and the weight fraction of the contaminant incompetent monomer (θ) can be calculated from the plot of $1/(K_2^{\mathrm{NL}})^{1/2}$ vs. $C_t^0\gamma$ shown in Eq. (20). However, since the curve of $(1/K_2^{\mathrm{NL}})^{1/2}$ vs. γC_t^0 is often non-linear in the region of $\gamma C_t^0\to 0$ and it usually consists of only limited number of data points, the precise estimation of t and s from the plot can be difficult in practice.

For a system with reasonable strength of association (the dissociation constant K_d in the range of a few mM or smaller) the value of $4(1-\theta)K_2C_t^0\gamma$ is often significantly greater than unity $(4(1-\theta)K_2C_t^0\gamma\gg 1)$ at moderately high loading concentrations $(C_t^0\sim 1$ mg/ml) and speeds ($\sigma_1>2$, $\sigma_1\propto\omega^2$ Eq. (1)). Under these conditions, Eq. (20) can be approximate by:

$$\frac{1}{\sqrt{K_2^{\mathrm{NL}}}}=\left[\frac{\theta}{2(1-\theta)}\frac{1}{\sqrt{K_2}}+\frac{\theta}{\sqrt{(1-\theta)}}\sqrt{C_t^o\gamma}\right] \qquad (21)$$

Eq. (21) is a more practical expression than Eq. (20). It defines a distinctive linear plot of $1/(K_2^{\mathrm{NL}})^{1/2}$ vs. $(C_t^0\gamma)^{1/2}$ for a heterogeneous I system, designated the HETERO I plot, and the K_2 and the θ can be estimated from the intercept and the slope of this plot.

4.4. Monomer–dimer association system contaminated by incompetent dimer—the heterogeneous II system

The second type of heterogeneous self-associating systems, the heterogeneous II system, corresponds to a monomer–dimer self-associating systems contaminated by tightly bound dimer that does not dissociate during the course of the experiments—the 'incompetent dimer' (the component B). The concentration distribution of such a system at sedimentation equilibrium is of the form:

$$\begin{aligned}C_t(r)=&C_{A_1}(r')\exp\left[\sigma_{A_1}\left(r^2/2-r'^2/2\right)\right]\\&+\{K_2C_{A_1}(r')^2\\&+C_B(r')\}\exp\left[2\sigma_{A_1}\left(r^2/2-r'^2/2\right)\right]\end{aligned} \qquad (22)$$

here $\sigma_B=2\sigma_{A_1}$. By fitting to the monomer–dimer model of NONLIN, the following estimation parameter will result:

$$C_1(r')^{\mathrm{NL}}=C_{A_1}(r') \qquad (23)$$

$$K_2^{\mathrm{NL}}\ =\ \frac{C_{A_2}(r')\ +\ C_B(r')}{C_{A_1}(r')^2}. \qquad (24)$$

and $\sigma_1^{\mathrm{NL}}=\sigma_{A_1}$ the reduced molecular weight of the monomer of A.

Applying Eq. (12) for conservation of mass of the incompetent dimer B, and combining Eqs. (23) and (24), the K_2^{NL} can be expressed as a function of the loading concentration C_t^0 and the true association constant K_2:

$$K_2^{\mathrm{NL}}=K_2+\frac{C_B^0/I_B}{C_{A_1}(r')^2}\ =\ K_2+\frac{\theta C_t^0}{I_B\left[C_1(r')^{\mathrm{NL}}\right]^2}. \qquad (25)$$

here $C_1(r')^{\mathrm{NL}}$, the concentration of monomer A_1 at the reference position, is a NONLIN fitting parameter (Eq. (23)) directly obtainable from NONLIN fitting. The value of $C_1(r')^{\mathrm{NL}}$ will vary with loading concentration and speed.

As shown in Eq. (25) the plot of K_2^{NL} vs. $\left(C_t^0/C_1(r')^2I_B\right)$ of a heterogeneous II system is a straight line. This linear plot, defined as the HETERO II plot, can be used to distinguish the heterogeneous II system from the other systems containing two species. The intercept of the HETERO II plot at $C_t^0=0$ gives the direct estimation of the K_2, and the slope equals θ.

It is interesting to note that both the HETERO I plot (Eq. (21)) and the HETERO II plot (Eq. (25)) of a disperse system containing two non-interacting components are also straight lines, except in both cases the line goes though the origin (since $K_2=0$):

$$\frac{1}{\sqrt{K_2^{\mathrm{NL}}}}=\left(\sqrt{\frac{1-\theta}{\theta}}\right)\sqrt{C_t^0\gamma} \tag{26a}$$

and

$$K_2^{\mathrm{NL}}=\frac{C_B(r')}{C_A(r')^2}=\frac{\theta\ C_t^0}{I_B\left[C_1(r')^{\mathrm{NL}}\right]^2}. \tag{26b}$$

The HETERO II plot of a disperse system (Eq. (26b)) is especially useful since it provide direct estimation of θ from the slope.

4.5. Other possible heterogeneous systems

It is likely that a monomer–dimer self-associating system contains both the incompetent monomer (B_1) and the incompetent dimer (B_2), designated heterogeneous III system. Such a system can be fit equally well by NONLIN monomer–dimer model with the following estimation parameters:

$$C_1(r')^{\mathrm{NL}}=C_{A_1}(r')+C_{B_1}(r') \tag{27}$$

$$K_2^{\mathrm{NL}}=\frac{C_{A_2}(r')+C_B(r')}{\left[C_{A_1}(r')+C_{B_1}(r')\right]^2} \tag{28}$$

and $\sigma_1^{\mathrm{NL}}=\sigma_{A_1}$ the reduced molecular weight of the monomer.

To relate the NONLIN apparent estimation parameters to the composition of the original solution will involve complicated arithmetic operations, and there appear to be no simple mathematical relation(s) for the derivation of K_2 and/or θ from the NONLIN fitting parameters. A characteristic plot for the identification of this system is also difficult to derive. Such a heterogeneous III system may be identified based on the lacking of any of the characteristic plots of the other three systems mentioned above. This approach is applicable, however, only when there is reason to rule out more complex association behaviors (more in Section 6).

4.6. The secondary plots

Eqs. (15), (21) and (25) define a set of linear plots based on secondary calculations of NONLIN estimation parameters as functions of loading concentrations and/or speeds. These plots are designated the Secondary plots including: DISP plot: K_2^{NL} vs. $1/C_t^0\gamma$ (Eq. (15)), HETERO I plot: $1/(\mathrm{K}_2^{\mathrm{NL}})^{1/2}$ vs. $(C_t^0\gamma)^{1/2}$ (Eq. (21)), HETERO II plot: K_2^{NL} vs. $C_t^0/C_1(r')^2I_B$(Eq. (25)).

These secondary plots are very helpful for identify three types of the heterogeneous systems and can be used to further characterize these systems. The applications of the secondary plots are summarized in Table 1.

While derived for monomer–dimer association systems, the DISP plot and the HETERO II plot can be generalized to treat disperse systems containing components A and B with $\sigma_B=n\sigma_A$, and heterogeneous II systems of a monomer–n-mer association. The situation of heterogeneous I system with monomer–n-mer association will be more difficult to analyze since it requires solving Eq. (18) with degree n ($n>2$).

5. Results

5.1. The asymptotic properties of NONLIN

The asymptotic properties of NONLIN have never been investigated before. The Rao–Cramér lower bound (the **MVB**) of NONLIN for a monomer–dimer self-associating system and the corresponding covariance matrix obtained by 500 Monte Carlo simulation are shown in Table 2. The Rao–Cramér lower bound almost coincides with the covariance matrix. The small differences between the two matrices are likely to arise from random noise. As a control experiment, the Rao–Cramér lower bound of a linear model evaluated by NONLIN is compared with the covariance matrix obtained from linear regression (Table 3).

Table 1
Characterization of different heterogeneous self-associating systems by secondary plots

Sample	Identification	Estimation	
		K_2	θ
Disperse: non-interacting monomer and dimer	linear DISP plot linear HETERO II plot through the origin linear HETERO I plot through the origin		from the slope of DISP plot θ = slope of HETERO II plot
Heterogeneous I: $A_1+A_1=A_2$ with incompetent monomer	linear HETERO I plot with finite intercept at $C_t^0=0$	from the slope and the intercept of HETERO I plot	from the slope and the intercept of HETERO I plot
Heterogeneous II: $A_I+A_I=A_2$ with incompetent dimer	linear HETERO II plot with finite intercept at $C_t^0=0$	K_2 = intercept of HETERO II plot	θ = slope of HETERO II plot
Heterogeneous III: $A_1+A_1=A_2$ with both incompetent monomer and incompetent dimer	Lack of linearity in all three secondary plots[a]		

[a]Assume no other more complicated association behavior.

Table 2
Comparison of the Rao–Cramér lower bound and the covariance matrix of NONLIN fitting parameters of a simulated monomer–dimer self-associating system

Fitting parameters	The covariance matrix[a]						
	Estimated Rao–Cramér lower bound						
$\delta\ y(1)$	8.72×10^{-7}						
$\delta\ y(2)$	3.38×10^{-7}	6.21×10^{-7}					
$\delta\ y(3)$	1.61×10^{-7}	1.14×10^{-7}	4.34×10^{-7}				
$\ln C_1(1)$	-2.60×10^{-6}	-1.81×10^{-6}	-8.62×10^{-6}	1.38×10^{-8}			
$\ln C_1(2)$	-2.52×10^{-6}	-1.83×10^{-6}	-8.45×10^{-7}	1.35×10^{-8}	1.32×10^{-8}		
$\ln C_1(3)$	-2.34×10^{-6}	-1.65×10^{-6}	-9.52×10^{-7}	1.25×10^{-5}	1.23×10^{-5}	1.17×10^{-5}	
$\ln K_2$	5.52×10^{-6}	3.89×10^{-6}	1.85×10^{-6}	-2.95×10^{-5}	-2.89×10^{-5}	-2.70×10^{-5}	6.35×10^{-5}
	Estimated by 500 Monte Carlo simulations						
$\delta\ y(1)$	7.71×10^{-7}						
$\delta\ y(2)$	2.63×10^{-7}	5.22×10^{-7}					
$\delta\ y(3)$	6.70×10^{-8}	2.40×10^{-7}	3.53×10^{-7}				
$\ln C_1(1)$	-2.52×10^{-6}	-1.84×10^{-6}	-8.35×10^{-7}	1.35×10^{-5}			
$\ln C_1(2)$	-2.44×10^{-6}	-1.86×10^{-6}	-8.21×10^{-7}	1.32×10^{-5}	1.30×10^{-5}		
$\ln C_1(3)$	-2.32×10^{-6}	-1.70×10^{-6}	-9.22×10^{-7}	1.24×10^{-5}	1.22×10^{-5}	1.17×10^{-5}	
$\ln K_2$	5.35×10^{-6}	3.96×10^{-6}	1.79×10^{-6}	-2.89×10^{-5}	-2.84×10^{-5}	-2.67×10^{-5}	6.35×10^{-5}

[a]Estimations for joint fit of a monomer–dimer system with three loading concentrations 1.2 mg/ml, 0.76 mg/ml and 0.27 mg/ml, respectively, for data set (1), (2) and (3), association constant $K_2=5\ fr^{-1}$ ($K_d\sim60$ μM for a species with monomer molecular weight of approx. 20 kDa), and $\sigma_1=2\ \mathrm{cm}^{-1}$.

Table 3
Comparison of the Rao–Cramér lower bound and the covariance matrix of fitting parameters for linear models

Fitting parameters	The covariance matrix		
A[a]	Estimated by linear regression		
a_1	3.84×10^{-6}		
a_2	0.000000	4.53×10^{-5}	
	Estimated by Rao–Cramer lower bound		
a_1	3.88×10^{-6}		
a_2	3.80×10^{-20}	4.51×10^{-5}	
B[a]	Estimated by linear regression		
a_0	3.06×10^{-5}		
a_1	-1.22×10^{-4}	6.55×10^{-4}	
a_2	1.01×10^{-4}	-6.13×10^{-4}	6.13×10^{-4}
	Estimated by Rao–Cramer lower bound		
a_0	3.11×10^{-5}		
a_1	-1.24×10^{-4}	6.64×10^{-4}	
a_2	1.03×10^{-4}	-6.20×10^{-4}	6.20×10^{-4}

[a]Covariance matrix estimated for polynomials of first degree (A) and second degree (B). The parameters used were $a_1=1$, $a_2=2$ in (A) and $a_0=1$, $a_1=2$, $a_2=1$ in (B). Note a_1 and a_2 in (A) are independent of each other.

The exact agreement of the two matrices is expected since, by definition, the linear least square estimation is an efficient estimator [11]. The variations between the two matrices, thus, reflect the random noises inherent in NONLIN estimations and in the evaluation of the covariance matrices.

Similar observations were also found for several other sedimentation equilibrium models with the value of σ_1 ranging from 1 to 5 cm^{-2} and, for a wide range of association constant K for both monomer–dimer and monomer–trimer associations [8]. Based on these findings, it can be concluded that NONLIN is asymptotically efficient under typical experiment settings, and the covariance matrix can be approximated by the Rao–Cramér lower bound. In addition, it can be demonstrated [8] that the NONLIN estimation procedures satisfy the regularity conditions to be asymptotically Gaussian [17]. These asymptotic properties of NONLIN provide a solid base to use the maximum likelihood ratio test to assess variations of the NONLIN estimated parameters, specifically the K_2^{NL}, in order to detect heterogeneity.

5.2. Hypothesis testing: is the sample homogeneous?

The applicability of hypothesis testing for detection of heterogeneity is demonstrated using simulated sedimentation equilibrium data for a monomer–dimer self-associating system containing varying amount of incompetent monomer (heterogeneous I system). Previous studies have indicated that the detection of heterogeneity in such a system is most challenging [7], especially when the association is weak. Only the results of a system with a relatively small association constant, $K_2=0.5\,fr^{-1}$ (equivalent to K_d in mM range of a protein with monomer molecular weight $\sim$20 kDa), are presented here to demonstrate the lower limit of the sensitivity of detection scheme. A more complete recording of studies on different model systems and with a range of different values of association constants is given elsewhere [8].

The (null) hypothesis to be tested is that the sample is homogeneous and the estimated association constants of samples with different loading concentrations are statistically the same and equal a constant C. The outcome of the test is expressed by the value of false alarm rate α (the probability that the null hypothesis is true). In practice, the constant C can be specified as the value of the common $\ln K$ (see Section 2), as one of the separate $\ln K$'s of samples with different loading concentrations, as the average of the separate values of $\ln K$ or as another known constant(s) (Table 4). Regardless the choice of C, the values

Table 4
Detection of heterogeneity based on the test of hypothesis

θ	ln K_2[a]		False alarm rate (α) for different choices of C[b]			
			ln K_{2com}	ln K_{2avg}	ln $K_2(1)$	ln $K_2(3)$
10%	(1) −1.135	(−1.171 −1.098)	<0.001	<0.001	<0.001	<0.001
	(2) −1.087	(−1.122 −1.051)				
	(3) −0.997	(−1.030 −0.964)				
5%	(1) −0.933	(−0.969 −0.897)	<0.001	<0.001	<0.001	<0.001
	(2) −0.909	(−0.943 −0.875)				
	(3) −0.860	(−0.891 −0.828)				
4%	(1) −0.891	(−0.926 −0.854)	0.003	0.001	0.002	0.003
	(2) −0.871	(−0.906 −0.838)				
	(3) −0.831	(−0.862 −0.800)				
3%	(1) −0.847	(−0.883 −0.811)	0.019	0.008	0.002	0.018
	(2) −0.833	(−0.867 −0.799)				
	(3) −0.794	(−0.825 −0.763)				
2%	(1) −0.802	(−0.838 −0.766)	0.381	0.3	0.213	0.375
	(2) −0.794	(−0.828 −0.760)				
	(3) −0.773	(−0.803 −0.742)				
0	(1) −0.710	(−0.746 −0.674)	0.931	0.973		
	(2) −0.714	(−0.742 −0.682)				
	(3) −0.712	(−0.742 −0.682)				

[a]NONLIN estimation of ln K_2 from three sets of computer simulated data generated using ln $K_2 = -0.693$ ($K_2 = 0.5\,fr^{-1}$, equivalent to $K_d \sim 1$ mM for a species with monomer molecular weight of approx. 20 kDa) with varying amount of heterogeneity (θ, Eq. (10)). Three sets of data corresponding to loading concentrations of 1.2 mg/ml (1), 0.75 mg/ml (2) and 0.25 mg/ml (3) were used for fitting. The confidence intervals were evaluated by Rao–Cramér lower bound. [b]The values of C used for the test are the common ln K_2 (ln K_{2comm}), the average of the separated ln K_2 of the three samples (ln K_{2avg}), or the values of the separated ln K_2 estimated for the 1.2 mg/ml sample (ln $K_2(1)$) or that of the 0.2 mg/ml sample (ln $K_2(3)$).

of α for samples with heterogeneity (θ) of 10% and 5% are all very small, so small that accepting the sample being heterogeneous will run a very small risk (<0.1%) of making a false alarm. Decision making for these examples based on the test of hypothesis is quite clear-cut.

For samples with lower heterogeneity the false alarm rate of test increases accordingly. When the θ is as low as 2%, the value of α increase to 38.1%, 30%, 21.3% and 37.5%, respectively, depending on the choice of C. The conclusion of the solution being heterogeneous at this point will face approximately 20–38% probability of being a false alarm. However, accepting the sample of being pure would make the type I error, since the sample is heterogeneous. In a difficult situation like this, the decision-making would likely be dependent on the consideration of other factors, such as the tolerance of the experiments or products to sample impurity, the cost of the sample, etc.

As a control, the test was also conducted on simulated data of a homogeneous sample ($\theta = 0$), and the probability of false alarm is found to be as high as 93–97%. The decision making for this sample is, again, clear-cut; the purity of the sample is unmistakable.

The detection of heterogeneity using hypothesis testing is more sensitive and effective than the approach based on the confidence intervals. For example, with $\theta = 10\%$ the value of ln K_2^{NL} of loading concentration 1.2 mg/ml falls outside the confidence interval estimated for samples with loading concentration of 0.75 and/or 0.26 mg/ml (Table 4). It is generally safe to conclude, for such situation, that the values of ln K_2^{NL} of different loading concentrations are different and the sample heterogeneous. The situation soon becomes ambig-

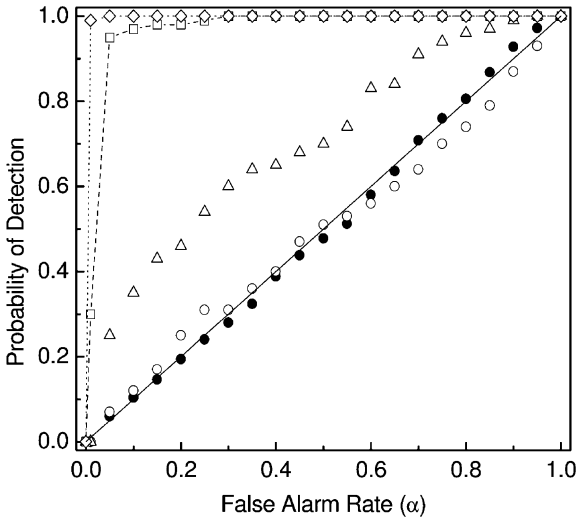

Fig. 1. ROCs plots for the detection of heterogeneity in a monomer–dimer self-associating system containing incompetent monomer. The test was conducted using three sets of simulation data generated for samples with $\sigma_1 = 2$ cm^{-2}, $K_2 = 0.5$ fr^{-1} and corresponding loading concentrations of 1.2 mg/ml, 0.76 mg/ml and 0.27 mg/ml, respectively. Data shown for heterogeneity (θ) of 0% (●); 1% (○); 2% (△); 3% (□); and 4% (◇). The probability of detection was determined using 100–500 Monte Carlo simulations per data point.

uous. There are significant level of overlapping of the confidence intervals when $\theta = 4$–3% and the values of ln K_2^{NL} estimated for loading concentration of 0.75 mg/ml fell inside the confidence interval estimated for loading concentration of 1.2 mg/ml. When θ is as low as 2%, the estimated values for the three loading concentration are mutually inclusive. Although, one may argue that the slight trend of the confidence intervals moving toward a smaller value with the decreasing of loading concentration is likely an indication of the heterogeneity, the decision-making based on these observations is generally difficult.

5.3. The ROCs curves: the probability of detection of heterogeneity

The probability of detection of heterogeneity based on the test of hypothesis is demonstrated in the ROCs curve (Fig. 1) of the same heterogeneous I system shown in Table 4. There are approximately 98% and 100% probability of detecting 4% of heterogeneity in this sample at the false alarm rate of 1% and 5%, respectively. For samples with $\theta > 4\%$ the probability of detection is nearly 100% with the false alarm $\alpha < 1\%$ (data not shown). The probability of detecting 3% of heterogeneity is ~30% at false alarm rate of 1%, but increases to 95% when the tolerable risk of false alarm is relaxed to 5%. The performance of the test for sample with $\theta = 2\%$ or less is, however, rather poor. At the false alarm rate of 5% the probability of detecting 2% heterogeneity only is approximately 20%. While the situation of $\theta = 1\%$ is almost identical to that of a homogeneous sample with $\theta = 0$, marked by the characteristic diagonal line, and the detection of such low heterogeneity is not possible.

It is interesting to refer back to Table 4 at this point. When a single test is conducted on the sample with θ of 2%, the estimated false alarm rate is 20–38%. The ROCs plots reveal that even if we are willing to run a high risk of false alarm for this sample, the probability of detecting such a small amount of heterogeneity is barely over 55%. There is ~45% of the chance that the three estimates of ln K_2^{NL} with differing loading concentrations would appear statistically the same, and the hypothesis test will not even detect the existence of the incompetent monomer.

The sensitivity of the test of hypothesis for the detection of heterogeneity improves with the increase of the loading concentration. When the test was conducted on the same heterogeneous system of Fig. 1 but with loading concentrations increased by 10-fold, the sensitivity of detection increases dramatically (Fig. 2). The probability of detecting 1% of heterogeneity is close to 93% with false alarm rate of 1%. At the false alarm rate of 5%, there is more than 75% chance of detecting the presence of 0.5% heterogeneity. For cases of $\theta \geq 2\%$, the probability of detection is nearly 100% with false alarm < 0.01 (data not shown).

The performance of the detection scheme is sensitive to the value of the reduced molecular weight of monomer σ_1, a parameter depending on the speed of sedimentation, the density of the solution and the specific volume ($\bar{V}$) as well as the molecular weight of the monomer. Fig. 3 is

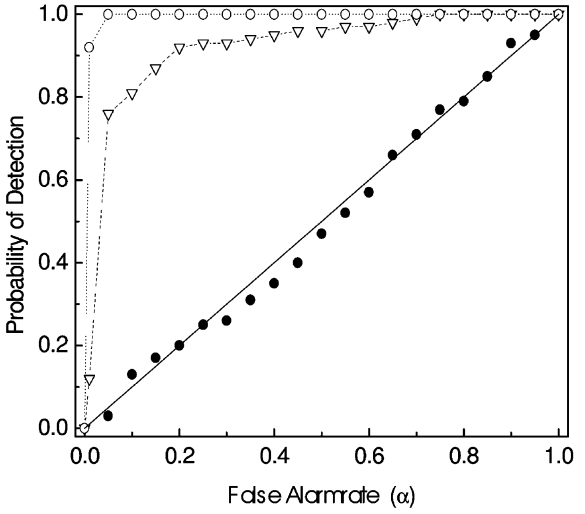

Fig. 2. ROCs plots for the detection of heterogeneity of the same heterogeneous self-associating system as Fig. 1, but with high loading concentrations: 12.1 mg/ml, 7.58 mg/ml and 2.67 mg/ml, respectively, for the three sets of data used for test. Data shown for heterogeneity (θ) of 0% (●); 0.5% (△); 1% (○). Other conditions are the same as in Fig. 1.

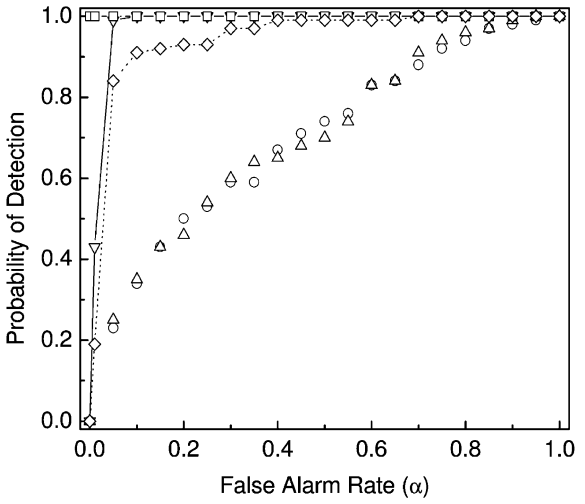

Fig. 3. Effects of the reduced molecular weight σ_1 on the ROCs plots for the detection of heterogeneity of a monomer–dimer system containing 2% incompetent monomer. The maximum likelihood ratio test was conducted for simulation samples with σ_1 equals 1 cm^{-1} (○); 2 cm^{-1} (△); 3 cm^{-1} (▽); 6 cm^{-1} (□); and 8 cm^{-1} (◇). Other conditions are the same as in Fig. 1.

the ROCs plots of the heterogeneous I system of Fig. 1 with $\theta = 2\%$ at five different values of σ_1. The probability of detection at the false alarm rate of 5% increased from approximately 25% at $\sigma_1 =$ 1 or 2 to nearly 100% at $\sigma_1 = 3$ and 6, but dipped back a bit to 84% at $\sigma_1 = 8$. Equilibrium sedimentation experiments at high σ_1 ($\sigma_1 = 8$) often have limited accuracy that may account for the decreased sensitivity of detection.

The detection scheme is more sensitive to systems with stronger association. The ROCs plots of another heterogeneous I system with $K_2 = 500$ fr^{-1} (K_d in μM range for a protein with molecular weight of approx. 20 kDa) are shown in Fig. 4. At the false alarm rate of 5%, the probability of detecting 2% heterogeneity increases to 96%. The detection of 1% of heterogeneity, however, is still difficult.

5.4. Applications of the secondary plots for the characterization of the heterogeneous self-associating systems

The applications of the secondary plots are demonstrated using computer simulated data gen-

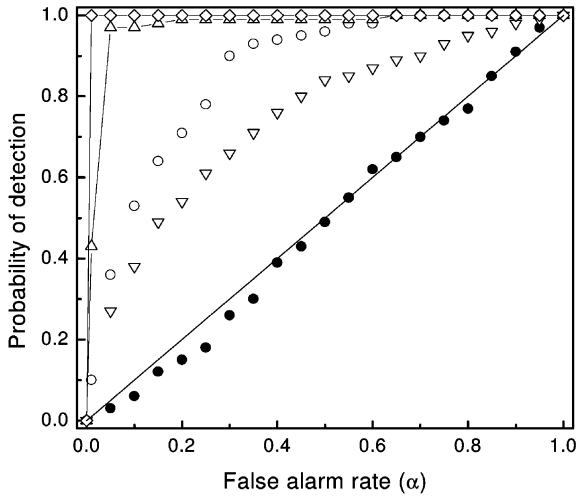

Fig. 4. ROCs plots for the detection of heterogeneity of a strongly dimerizing solute contaminated by incompetent monomer. Conditions as for Fig. 1 except $K_2 = 500$ fr^{-1} ($K_d \sim 1$ μM for a species with monomer molecular weight of ~20 kDa). Data shown for heterogeneity (θ) of 0% (●); 0.5% (▽); 1% (○); 2% (△); and 4% (◇).

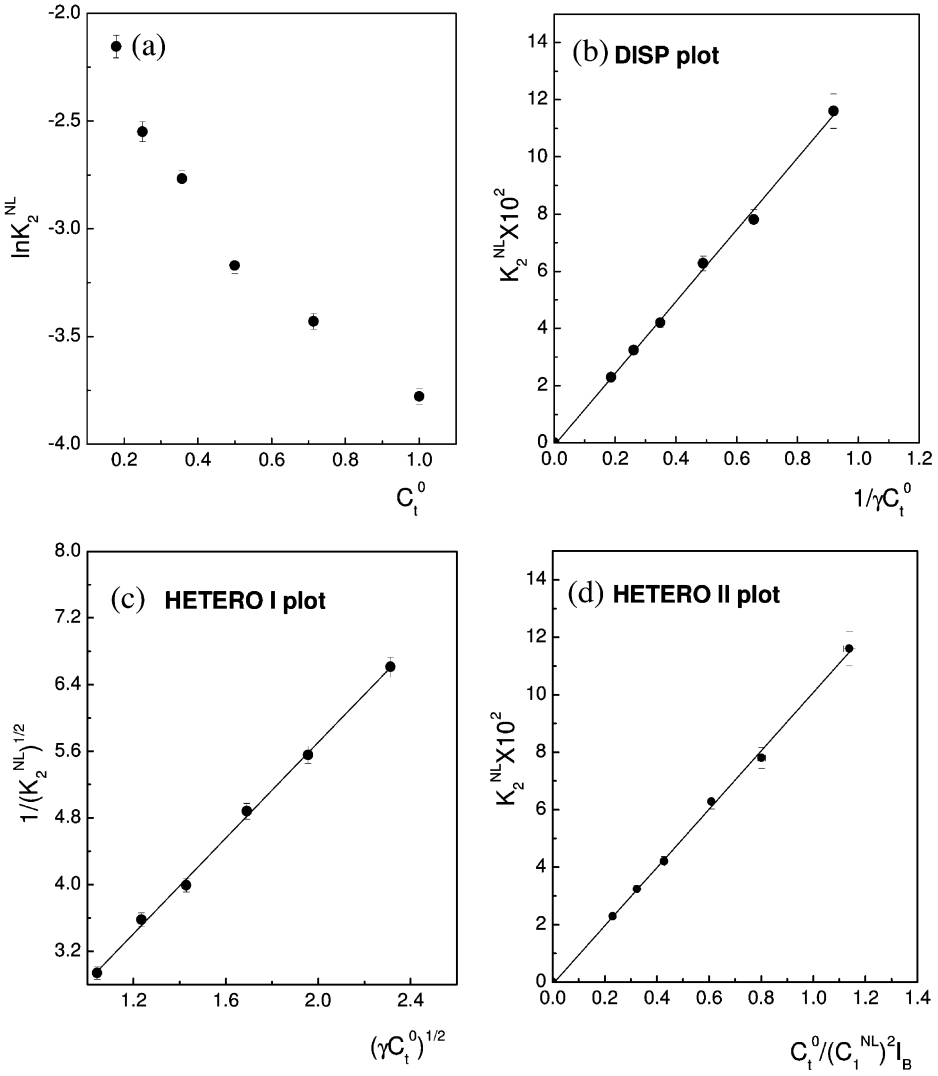

Fig. 5. Secondary plots of a disperse system. The plots were constructed based on NONLIN fitting results of six sets of simulation data generated for an disperse system with $\sigma_A=2$, $\sigma_B=4$, $\theta=10\%$ and corresponding loading concentrations of 1.0 mg/ml, 0.72 mg/ml, 0.5 mg/ml, 0.36 mg/ml, 0.25 mg/ml and 0.18 mg/ml, respectively. The error bars are propagated from the confidence interval of 67% confidence probability of NONLIN. The line in DISP plot is the linear fit to Eq. (15), with the estimated slope of 0.124 ± 0.003 and the intercept of $-5.1\times10^{-4}\pm1\times10^{-3}$. The line in HETERO I plot is the linear fit to Eq. (21) with estimated slope of 2.87 ± 0.06 and the intercept -0.03 ± 0.09. The line in HETERO II plot is the fit to Eq. (25) with the estimated slope of 0.102 ± 0.002, and the intercept $-4.6\times10^{-4}\pm1\times10^{-3}$.

erated with parameters that are typical to routine sedimentation equilibrium experiments. Only selected cases are shown (Figs. 5 and 8); a more extensive record can be found elsewhere [8]. When the maximum likelihood ratio test was conducted on these data, all systems (except the ones for homogeneous monomer–dimer association) were found to be heterogeneous with a false alarm rate $<1\%$, and will not be discussed further in each case.

Case 1: the disperse system. The NONLIN estimate of $\ln K_2^{NL}$ of a disperse system with non-interacting monomer and dimer with $\theta = 10\%$ clearly depends on the loading concentration (Fig. 5a). The $\ln K_{2app}$ decreases with an increase of C_t^0 and the presence of heterogeneity is obvious. All three secondary plots of this sample appear to be straight lines going through the origin (Fig. 5b, c and d). These distinctive features of the secondary plots indicate a disperse nature of this system containing non-interacting monomer and dimer. The weight fraction of the dimer (component B) calculated from the slope of the HETERO II plot is $\theta = 10.2 \pm 0.2\%$ (Fig. 5d). Alternatively, the θ can be calculated from the slope of the DISP plot using Eq. (14), and the estimated value is 10.1%. Both are in good agreement with the true values of $\theta = 10\%$ used to generate the data.

Case 2: self-associating system with incompetent monomer. The NONLIN estimated $\ln K_2^{NL}$ of a heterogeneous I system decreases with the increase of loading concentration (Fig. 6a) and the values of $\ln K_2^{NL}$ are all smaller than the true value of $\ln K_2$ marked by the dashed line. The identification of a heterogeneous I system can be made based on characteristic straight-line appearance of the HETERO I plot (Fig. 6c), while both the DISP plot and the HETERO II plot are curved (Fig. 6b,d). The linear fit of the HETERO I plot result in the intercept and the slope of 1.55 ± 0.02 and 0.089 ± 0.011, respectively. The θ calculated from these estimates is 9.5%, which is very close to the true value of 10%, and the calculated K_2 is 0.46, about 8% off from the true value of 0.5.

Case 3: self-associating system with non-dissociating dimer. The secondary plots of a heterogeneous II system with relatively weak association are demonstrated in Fig. 7. Similar to the heterogeneous I system in Fig. 6, the values of $\ln K_2^{NL}$ decreases with the loading concentration (Fig. 7a), except the values of $\ln K_2^{NL}$ are all greater than the true value of the association constant ($\ln K_2 = -2.996$). The heterogeneous II system can be identified by the linear HETERO II plot with a finite intercept on the y-axis (Fig. 7d). The θ and the K_2 determined from the slope and the intercept of the linear HETERO II plot are $11.3 \pm 0.4\%$ and 0.051 ± 0.003, respectively. Both are quite accurate comparing to the true values of 10% and 0.05.

The DISP plot for this heterogeneous II system also appears to be a straight line because of the small values of K_2 used for the simulated data (Fig. 7b). However, the estimations of K_2 and θ obtained from this linear DISP plot are not very accurate. The K_2 from the intercept of the plot is 0.0635 approximately 27% higher than the true value, and the θ from the slope is 0.15, nearly 46% too high. The DISP plot of heterogeneous II systems with reasonably strong association constant (greater K_2) is not linear in general [8].

For the comparison, the secondary plots of a homogeneous monomer–dimer self-associating system are shown in Fig. 8. The $\ln K_2^{NL}$ is close to a constant with no obvious trend of changing with the loading concentration, and the three secondary plots are all flat. It appears that there was a certain degree of bias in the NONLIN estimates for this particular simulation system; the values of $\ln K_2^{NL}$ are slightly lower than the expected value marked by the dashed line in Fig. 8a.

5.5. The characterization of a BSA sample by sedimentation equilibrium

The secondary plots were used to characterize a bovine serum albumin (BSA) sample. Although BSA is generally considered as monomer in solution and has been widely used as molecular weight marker, there has been controversy regarding its purity and state of association in solution [18], which is also found to be dependent on specific preps (Yphantis, personal communications).

The BSA sample used here appeared to contain species larger than the monomer based on sedimentation equilibrium experiments with different loading concentrations and at different speeds. When treated as a single component system, the estimated apparent values of the molecular weight (M_{app}) by NONLIN were approximately 4–8% higher than the formula molecular weight of monomer BSA (data not shown). In addition, the value of M_{app} decreases with the increase of loading concentration, and decreases slightly at higher speeds. These features are typical for a heteroge-

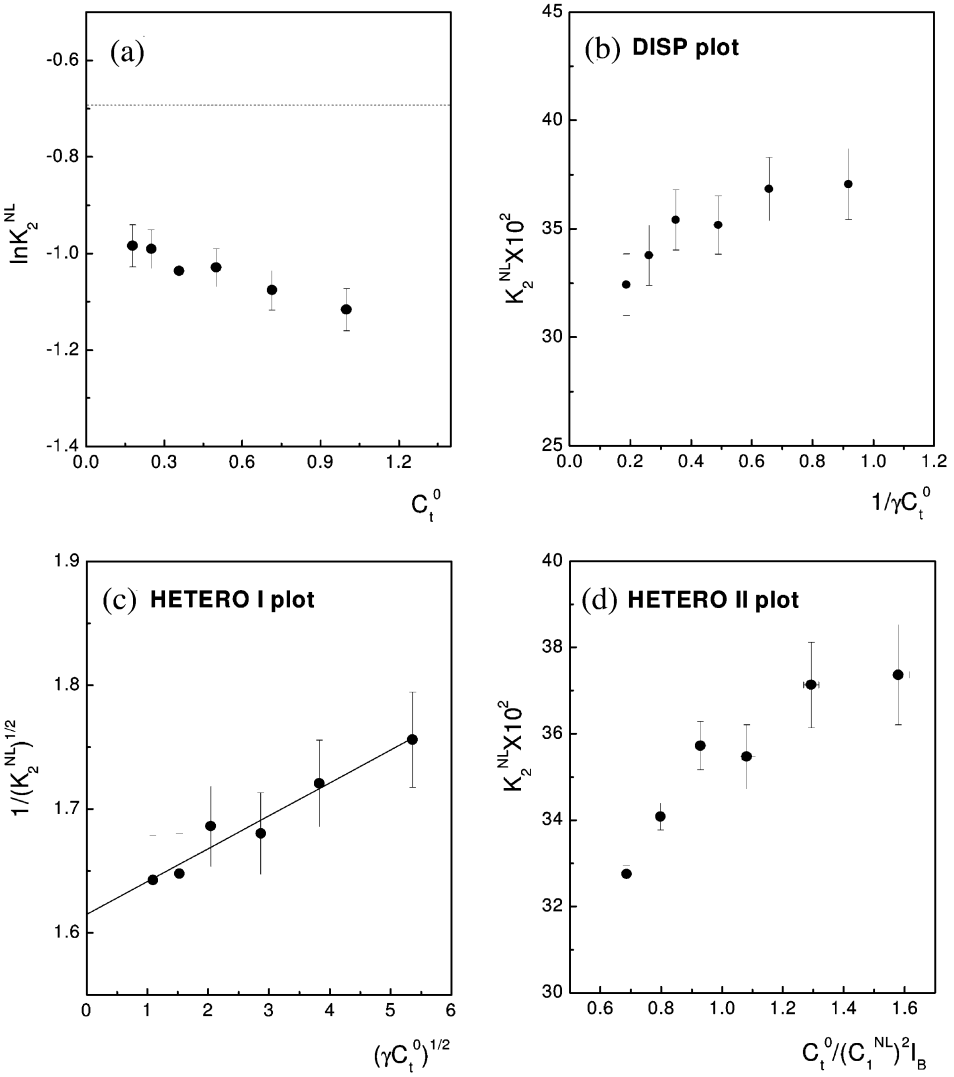

Fig. 6. Secondary plots for a heterogeneous I system based on simulation data generated with $K_2=0.5\ fr^{-1}$, $\sigma_{A_1}=2$, $\sigma_B=2$ and $\theta=10\%$, and six loading concentrations specified in Fig. 5. The line in HETERO I plot is the linear fit to Eq. (21) with the estimated slope of 0.0892 ± 0.011 and intercept of 1.5450 ± 0.018. The dashed line in (a) marked the true association constant.

neous system containing monomer and higher oligomer(s).

The NONLIN monomer–dimer model appears to fit all the data of this BSA sample well, judging by the low fitting RMS values ($0.8–1.2\times10^{-3}\ fr$) and the random residual plots (not shown). The average apparent monomer molecular weight estimated from the monomer–dimer model at different loading concentrations and speeds is $M_{1\mathrm{app}}=$ 67.182, only approximately 1.2% higher than the formula molecular weight of 66.288. Nevertheless, the sample appears to be heterogeneous. The values of ln K_2^{NL} decrease with the loading concentration and/or speeds (Fig. 9a). The test of hypothesis indicates the sample is heterogeneous with false alarm rate less than 1%.

The secondary plots of the BSA sample were constructed using the data collected at speed of

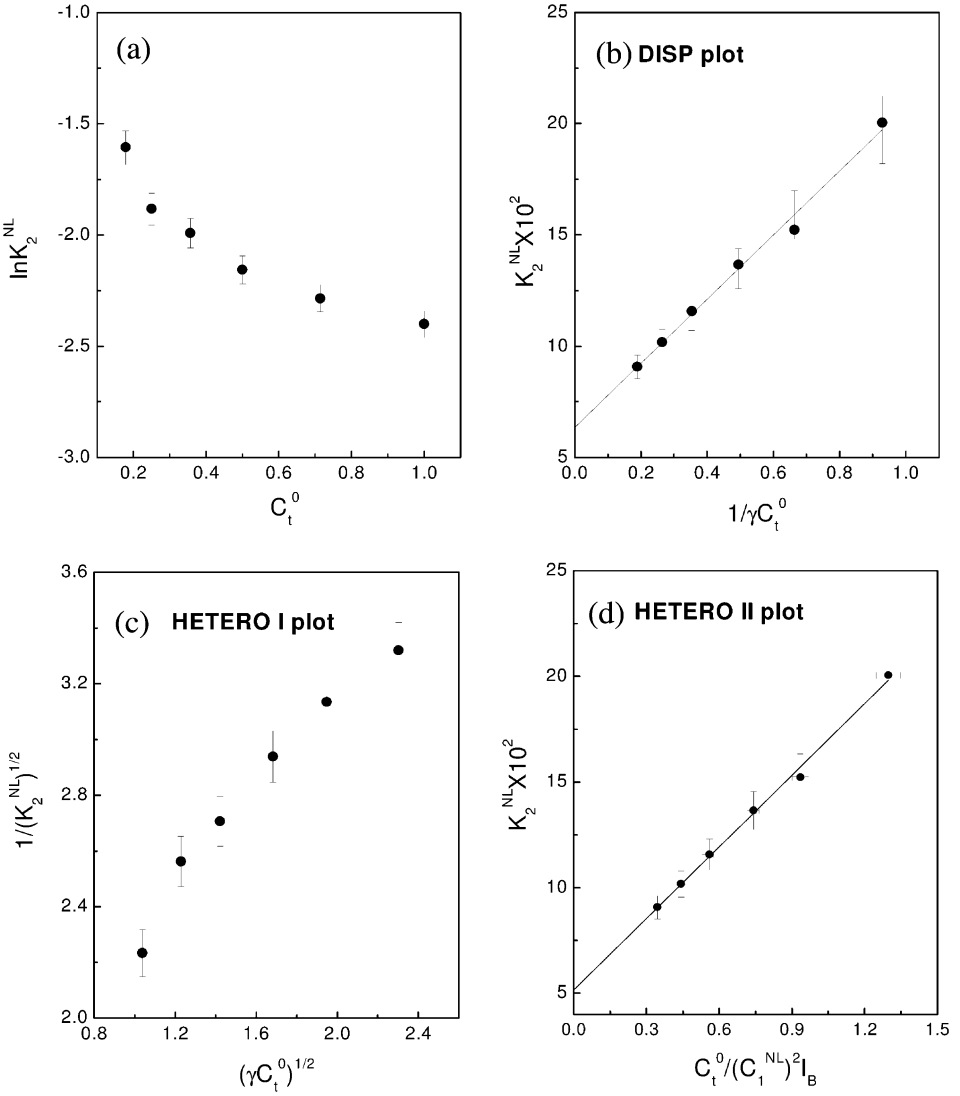

Fig. 7. Secondary plot of a heterogeneous II system based on simulation data generated with $K_2 = 0.05\ fr^{-1}$ ($K_d \sim 10$ mM for a species with monomer molecular weight of approx. 20 kDa), $\sigma_{A_1} = 2$, $\sigma_B = 4$, $\theta = 10\%$ and the six different loading concentrations specified in Fig. 5. The line in DISP plot (b) is the fit using Eq. (15) with the estimated slope of 0.144 ± 0.06, and intercept of 0.0064 ± 0.003. The line in HETERO II plot (d) is the fit using Eq. (25) with the estimated slope of 0.107 ± 0.004, and intercept of 0.051 ± 0.003.

16 000 rev./min with six different loading concentrations (Fig. 9). All three secondary plots (Fig. 9b, c and d) appeared to be straight lines going through the origin, indicating a disperse system with non-interacting monomer and dimer. The amount of dimer (the θ) in the original solution is 7% estimated form the slope of the DISP plot (Fig. 9b), and 6.8% from the slope of the HETERO II plot with the exclusion of the last data point (Fig. 9d). The last data point in the HETERO II plot, corresponding to the sample with the lowest loading concentration (0.4 mg/ml), appeared to

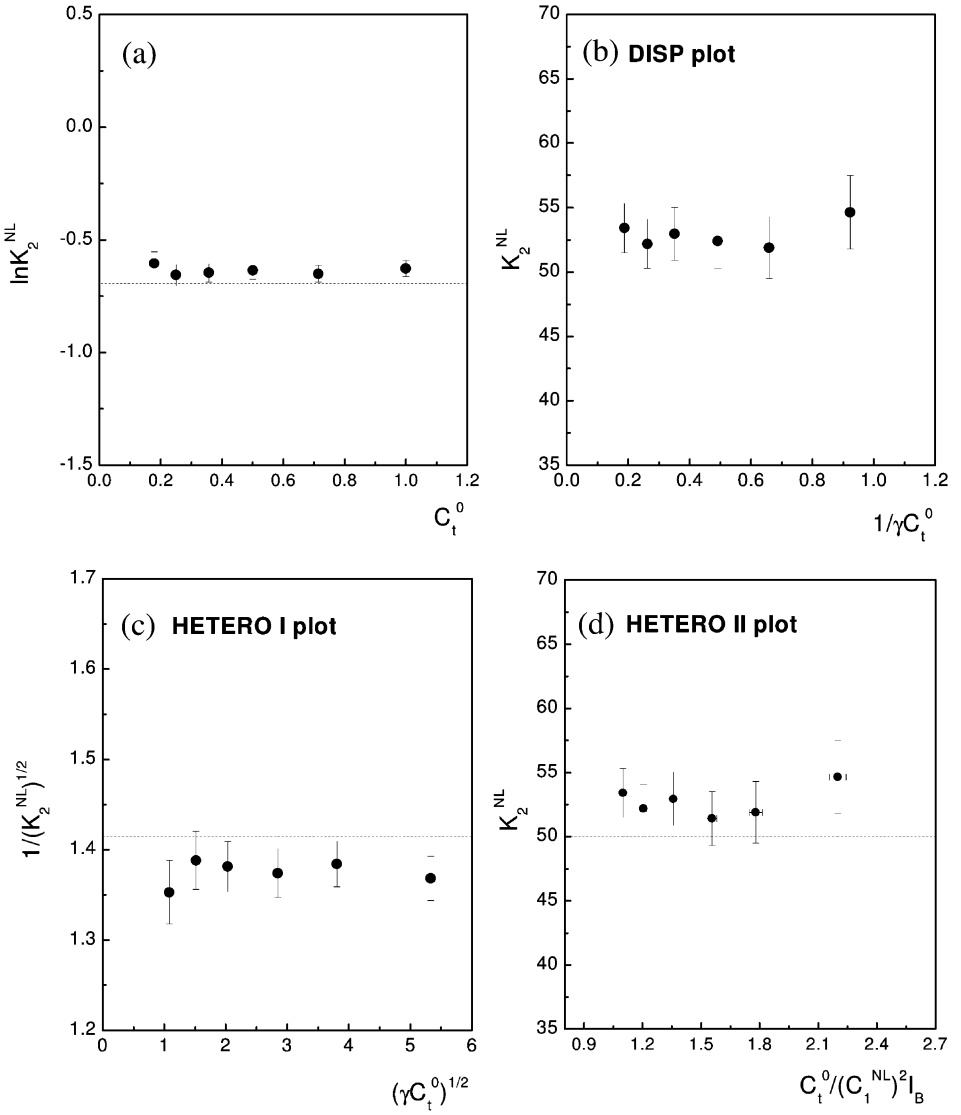

Fig. 8. Secondary plot of a homogeneous system with $K_2 = 0.5\ fr^{-1}$, $\sigma_{A_1} = 2$, and six different loading concentrations specified in Fig. 5. The dashed lines in each plot marked the expected value of the true association constant.

be an 'outlier'. The deviation was likely caused by the error in the NONLIN estimation of $C_1(r')^{NL}$, the concentration of monomer at the reference position (chosen to be at the meniscus). The value of $C_1(r')^{NL}$ is expected to be small at the lowest loading concentration and carry relatively larger uncertainties in its estimation. In addition, the HETERO II plot depends on the square of $C_1(r')^{NL}$ (Eq. (25)), further magnifying the effects of any error on the estimate of $C_1(r')^{N}$. Another factor may affect the accuracy of the plot is the violation of the conservation of mass which, again, would have more significant effect on data with low loading concentration. Nonetheless, the estimation of θ from the HETERO II plot appears to be in a good agreement with that estimated

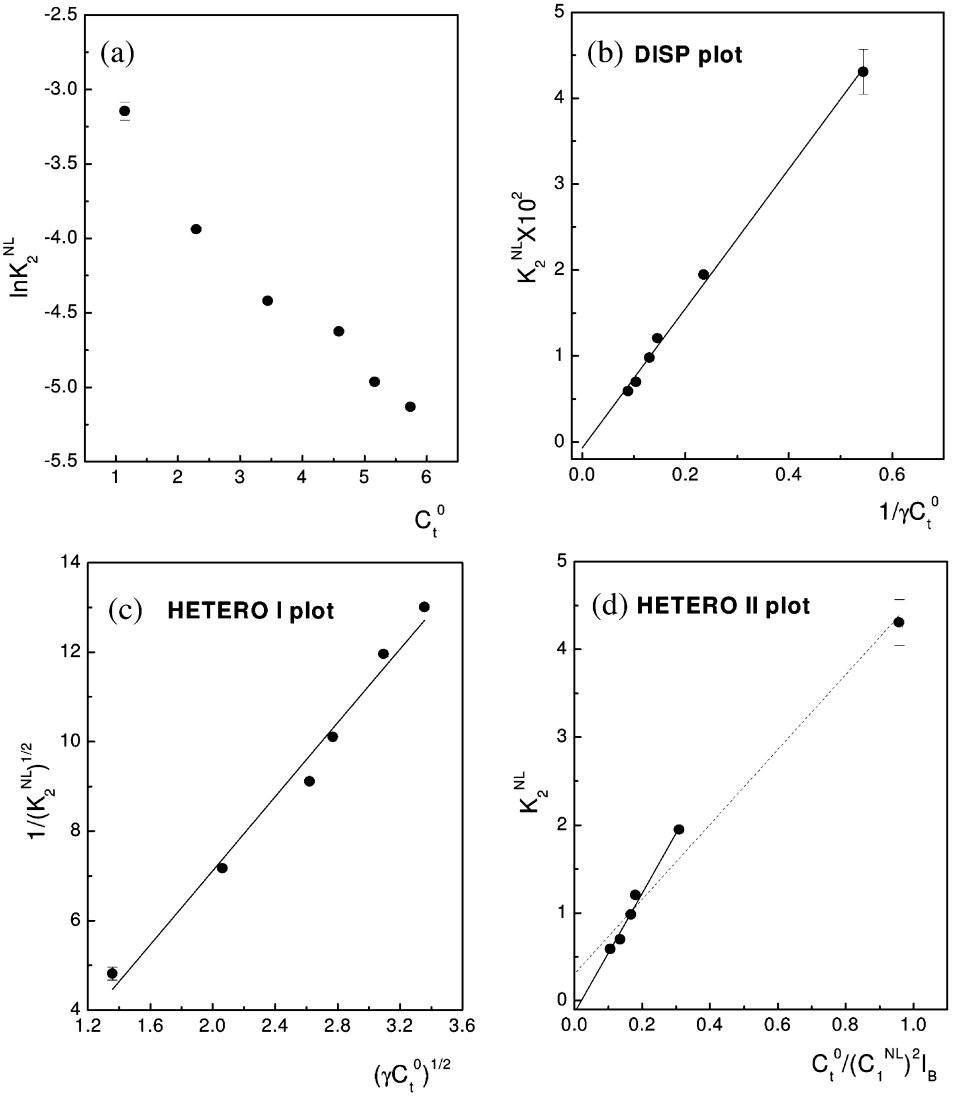

Fig. 9. Secondary plot of BSA sample based on NONLIN estimations of data obtained at 16 000 rev./min and loading concentrations of 0.41, 0.83, 1.24, 1.65 and 2.1 mg/ml. The line in the DISP plot (b) is the linear fit of Eq. (15), with the estimated slope of 0.081 ± 0.002 and the intercept of -0.0007 ± 0.0006. The solid line in HETERO II plot (d) is the fit to Eq. (25) excluding the last data point with the estimated slope of 0.068 ± 0.005, and the intercept -0.0014 ± 0.0010. The dashed line indicates the fit to Eq. (25) including the last data point. The line in HETERO I plot is the linear fit to Eq. (21) with estimated slope of 4.13 ± 0.26 and the intercept -1.14 ± 0.69.

from the DISP plot. Thus, it was concluded that this BSA sample was a disperse system containing 93% monomer and 7% dimer that are not in fast-equilibrium with the monomer.

6. Discussion

It often happens in sedimentation equilibrium experiments that a group of data collected at

different conditions can fit satisfactorily to a NONLIN model, but only with the association constants treated as variables for each data set (fitting with separate ln K's). Here, attempts were made to provide some insight into the variations of the apparent association constant(s) caused by the uncertainties inherent in estimation procedures and/or by the heterogeneity in the sample. Based on this analysis a new approach, the secondary plots, has been developed that can help to identify three types of heterogeneity and be used to estimate the amount of contaminant causing heterogeneity and determine the true equilibrium constant of the self-associating components.

The application of hypothesis testing for statistical inferences transforms the process of assessing the variations of the apparent association constants caused by solute heterogeneity into a typical signal detection problem, where quantitative detection procedures have been worked out. The sensitivity and accuracy of the detection procedure has been evaluated using the ROCs curves. The test of hypothesis and the confidence intervals are closely related concepts of statistical inference of the estimates and the same results are often expected for the two approaches [11,17]. The use of hypothesis testing, however, provides a quantitative statement about the accuracy of the conclusions that were made. The outcome of the tests is easier to interpret and appear to be less ambiguous, especially for the so-called 'border line' situations where the confidence intervals of K_2^{NL} of two samples overlap to some degree.

Both the detection procedure and the correct application of the secondary plots assume that heterogeneity is the sole source of variation of the apparent values of the equilibrium constants. The apparent values of the parameters for a homogeneous self-associating sample will exhibit strong loading concentration dependence if the fitting model is incorrect [3]. Choosing the correct model for data fitting is a common concern for non-linear estimation procedures. The joint fit of a group of data obtained at different conditions often appear 'improved' when the association constant(s) were treated as individual values of ln$K$'s for each data set (see Section 2). This 'improvement', however, does not necessarily mean the heterogeneous model is a 'better' fitting model; rather, it could merely be the consequence of introducing more fitting parameters.

It may prove extremely helpful to incorporate the earlier approach of detecting heterogeneity using the point average molecular weight averages [1–4] as part of the detection procedure. The chief advantage of this earlier graphical procedure is that it does not require knowledge of the correct model of the association system [3]. The sensitivity and accuracy of such an approach is limited since the estimation of the point averages of molecular weight requires numerical differentiation of the data, which is often noisy and of limited precision. Such procedure can nevertheless be used to rule out certain wrong conclusions about heterogeneity caused by fitting with a wrong model. In addition, conducting equilibrium sedimentation runs under a series of different conditions, utilizing different speeds and a range of loading concentrations is highly recommended. Not only is it essential for detection and characterization of heterogeneous samples, such practice often will provide more information for determining the correct fitting model and for accurate estimations of the thermodynamic parameters.

Both the accuracy of estimation and the sensitivity of detection will be limited by the weak association of the self-associating component A. Small values of equilibrium constant were used in most of the cases here to demonstrate the lower limits of the secondary plots and of the detection scheme. The secondary plots are generally applicable for association systems with K_d (the dissociation constant of A) values up to mM range. For heterogeneous II system the secondary plots appear to be able to provide reliable estimations for K_d in the ranges of 10s of mM. When the association is too weak, all three secondary plots will approach a straight line for all heterogeneous systems discussed here and the secondary plot approach will, thus, lose its power of discriminating and characterizing different heterogeneous systems.

Heterogeneous association systems discussed here are limited to those that contain only one prominent equilibrium of the solute (that of component A). Treatment of systems involving more complex association behavior, such as the dimeri-

zation of B (with association constant $K_{B2} \neq K_2$) and/or the mixed association between A_1 and B_1 (with yet, another association constant K_{AB}) will have to wait for future studies.

Acknowledgments

I would like to thank Jeff Lary and Jiawen Wu for technical support of running the machine, and Dr Cun-Ming Han for comments on the statistical analysis part of this manuscript. This work was supported by National Science Foundation grants (to David Yphantis) BBS-8612159, BIR-9218679 and DBI-9603153.

References

[1] P.G. Squire, C.H. Li, J. Am. Chem. Soc. 83 (1961) 3521–3528.
[2] D.C. Teller, T.A. Horbett, E.C.a. Richard, H.K. Schachman, Ann. NY Acad. Sci. 164 (1969) 66–101.
[3] D.A. Yphantis, J.J. Correia, M.L. Johnson, G. Wu, in: N. Catsimpoolas (Ed.), Physical Aspects of Protein Interactions, Elsevier/North Holland, Inc, Amsterdam, 1979.
[4] V.D. Hoagland Jr, D.C. Teller, Biochemistry 8 (1969) 594–602.
[5] M.L. Johnson, J.J. Correia, D.A. Yphantis, H.R. Halvorson, Biophys. J. 36 (1981) 575–588.
[6] D.E. Roark, D.A. Yphantis, Ann. NY Acad. Sci. 164 (1969) 245–278.
[7] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Cambridge University Press, Cambridge, New York and Oakleigh, Victoria, Australia, 1986.
[8] Y. Xu, (1995) Ph.D. Thesis. University of Connecticut, Storrs, Conn.
[9] D.A. Yphantis, Biochemistry 3 (1964) 297–317.
[10] J. Larry, J.W. Wu, D.A. Yphantis, Biophys. J. (1990) 57, 337a (ABSTRACT).
[11] Y. Bard, Academic Press, New York, London, 1974.
[12] G.A.F. Seber, C.J. Wild, (1989).
[13] M.L. Johnson, L.M. Faunt, in: L. Brand, M.L. Johnson (Eds.), Methods in Enzymology, 1992, 210, 1–37.
[14] M. Straume, M.L. Johnson, in: L. Brand, M.L. Johnson (Eds.), Methods in Enzymology, Numerical Computer Methods, 210, Academic Press, San Diego, New York, Boston, London, Sydney, Tokyo, Toronto, 1992, pp. 117–129.
[15] H.V. Poor, Springer-Verlag, Heidelberg, 1988.
[16] J.W. Williams, K.E. Van Hold, R.L. Baldwin, H. Fujita, Chem. Rev. 58 (1958) 715–806.
[17] G.A.F. Seber, John Wiley and Sons, Inc. New York, Chichester, Brisbane, Toronto, Singapore, 1987.
[18] J.F. Foster, in: F.W. Putnam (Ed.), The Plasma Protein: Isolation, Characterization and Function, 1, Academic Press, New York and London, 1960.