Protein and Peptide Letters **6**, 429-436 (2001)

AN EFFICIENT, FLEXIBLE-MODEL PROGRAM FOR THE ANALYSIS OF DIFFERENTIAL SCANNING CALORIMETRY PROTEIN DENATURATION DATA

Sasha B. Grek, John K. Davis, and Michael Blaber*

Department of Chemistry and Institute of Molecular Biophysics
Florida State University, Tallahassee FL 32306-4380

Although thermodynamic formalisms for protein denaturation have been established for some time, available software programs for deconvolution of DSC data exhibit various limitations. These include enforcing a constant $\Delta C_p(T)$, linear heat capacity functions, and so on. We have developed a Windows™ based program that allows greater flexibility, speed and accuracy than previously available programs for the analysis of DSC data. One novel feature of the program is the inclusion of, and ability to refine, a concentration-dependent term.

Introduction

Monitoring the temperature-dependent phase transitions of proteins in aqueous solution can provide important details regarding the non-covalent interactions that stabilize the native state. Knowledge of the thermodynamic parameters of unfolding for globular proteins, in conjunction with mutagenesis, has advanced our understanding of protein structure, stability and folding [1-3]. As dilute solutes in aqueous solution, the heat capacity contribution of protein molecules is small compared to the overwhelming contribution of the solvent. The development of differential scanning calorimetry (DSC) has provided a method to subtract the heat capacity contribution of the solvent and characterize the thermodynamic properties of the protein solute (for review see [4]).

Statistical mechanics-based models describing the thermodynamic properties of macromolecular phase transitions have been described [5-7]. However, there is a very limited availability of software for analysis of DSC data. In addition, recent state-of-the-art instrumentation places additional demands upon speed, accuracy, model flexibility and implementation on a common hardware platform for such software. For example, available programs typically require operator intervention to define pre- and post-transition baseline functions. Coupled with a non-convergent least-squares fitting proceedure, refined values for the thermodynamic parameters will vary depending upon the choices of the particular operator. Some programs force the pre- and post-transition baselines to be parallel (i.e. ΔC_p is a constant with temperature). While this simplifies the computation (the definition of the enthalpy associated with the phase transition includes an integration term for

ΔC_p), such a model will erroneously identify transitions with non-parallel baselines as being non-2-state. Additionally, although the denatured state heat capacity function of polypeptides is generally acknowledged to be non-linear [8], current analysis programs utilize computationally expedient linear functions for such baselines. To address these issues we have developed a highly customizable program, named *DSCFit*, for the analysis of DSC data, implemented on the Windows™ platform.

Materials and Methods

Thermodynamic functions

All temperatures are absolute (K), and heat capacities are $Jmol^{-1}K^{-1}$. However, since the temperature range to be analyzed for soluble proteins is limited to the liquid phase of water, the temperature value for the y intercept (heat capacity) of all functions is not referenced at 0K, but rather, the melting temperature (*TD*) of the data. The native state heat capacity function, $C_N(T)$, and denatured state heat capacity function, $C_D(T)$, are defined as second order polynomials with coefficients A_0, B_0 and C_0 and A_1, B_1 and C_1 for the $C_N(T)$ and $C_D(T)$ functions, respectively:

$$C_N(T) = 3A_0(T - TD)^2 + 2B_0(T - TD) + C_0 \quad (1)$$

$$C_D(T) = 3A_1(T - TD)^2 + 2B_1(T - TD) + C_1 \quad (2)$$

From these definitions, it follows that the $\Delta C_p(T)$ function ($\Delta C_p = C_D(T) - C_N(T)$) is also a second order polynomial:

$$\Delta C_p(T) = 3(A_1 - A_0)(T - TD)^2 + 2(B_1 - B_0)(T - TD) + (C_1 - C_0) \quad (3)$$

The $\Delta H_{sys}(T)$ function is defined as a third-order polynomial with coefficients *DA*, *DB*, *DC* and *DD*:

$$\Delta H_{sys}(T) = DA(T - TD)^3 + DB(T - TD)^2 + DC(T - TD) + DD \quad (4)$$

From (4) it can be seen that at T = TD, the value of the $\Delta H_{sys}(T)$ function reduces to *DD*. Since the derivative of the $\Delta H_{sys}(T)$ function is $\Delta C_p(T)$, it follows from (4) that:

$$\Delta C_p(T) = 3DA(T - TD)^2 + 2DB(T - TD) + DC \quad (5)$$

Thus:

$$\begin{aligned} DA &= (A_1 - A_0) \\ DB &= (B_1 - B_0) \\ DC &= (C_1 - C_0) \end{aligned} \quad (6)$$

Given that the entropy function, $\Delta S(T)$, is formally:

$$\Delta S(T) = \frac{\Delta H(TD)}{TD} + \int_{TD}^{T} \frac{\Delta C_p(T)}{T} dT \quad (7)$$

all thermodynamic functions are now described in terms of parameters A_0, B_0, C_0, A_1, B_1, C_1, DD and TD. From (3) and (5) it can be seen that the value of the $\Delta C_p(T)$ function at the melting temperature, TD, is equal to:

$$\Delta C_p(TD) = (C_1 - C_0) = DC \quad (8)$$

Therefore:

$$C_1 = DC + C_0 \quad (9)$$

Substituting (9) into (2) gives:

$$C_D(T) = 3A_1(T - TD)^2 + 2B_1(T - TD) + (DC + C_0) \quad (10)$$

Thus, $\Delta C_p(T)$ is now:

$$\Delta C_p(T) = 3(A_1 - A_0)(T - TD)^2 + 2(B_1 - B_0)(T - TD) + DC \quad (11)$$

$\Delta H_{sys}(T)$ is:

$$\Delta H(T) = (A_1 - A_0)(T - TD)^3 + (B_1 - B_0)(T - TD)^2 + DC(T - TD) + DD \quad (12)$$

$\Delta S(T)$ is:

$$\Delta S(T) = \left[\left(\ln(T) - \ln(TD) + \frac{3}{2}\right)TD^2 - 2T(TD) + \frac{1}{2}T^2\right]3(A_1 - A_0) + [(-1 + \ln(TD) - \ln(T))TD + T]2(B_1 - B_0) + (\ln(T) - \ln(TD))DC + \frac{DD}{TD} \quad (13)$$

All thermodynamic functions are now described in terms of parameters A_0, B_0, C_0, A_1, B_1, DC, DD and TD, where DC is the value of the $\Delta C_p(T)$ function at the melting temperature (TD), and DD is the value of the $\Delta H_{sys}(T)$ function at the melting temperature.

$\Delta G(T)$ is:

$$\Delta G = \Delta H(T) - T(\Delta S(T)) \quad (14)$$

The fractional component of native state as a function of $\Delta G(T)$ is given as:

$$F_N(T) = \left[1 + \exp\left(-\frac{\Delta G(T)}{RT}\right)\right]^{-1} \quad (15)$$

The raw data is thus modeled by the $C(T)$ function [5-7]:

$$C(T) = C_N(T) + \Delta C_p(T)(1 - F_N(T)) + \left(\frac{H_{sys}(T)^2 F_N(T)(1 - F_N(T))}{RT^2}\right) \quad (16)$$

The first part of this equation describes the heat capacity contribution associated with the mole fraction of the native and denatured states. The second part describes the contribution to the observed heat capacity from the enthalpy associated with the phase transition.

The introduction of a scalar, *k*, allows for evaluation of concentration-dependent effects, such as concentration errors in the sample, and the value of the *van't Hoff* to calorimetric enthalpy ratio ($\Delta H_{vH}/\Delta H_{cal}$) (see discussion below):

$$C(T) = \frac{\left(C_N(T) + \Delta C_p(T)\left(1 - F_N(T)\right) + \left(\frac{H_{sys}(T)^2 F_N(T)(1 - F_N(T))}{RT^2} \right) \right)}{k} \qquad (17)$$

A graphical representation of the model parameters is shown in Fig. 1 and the implications upon the fitting model when constraining the various parameters are listed in table I.

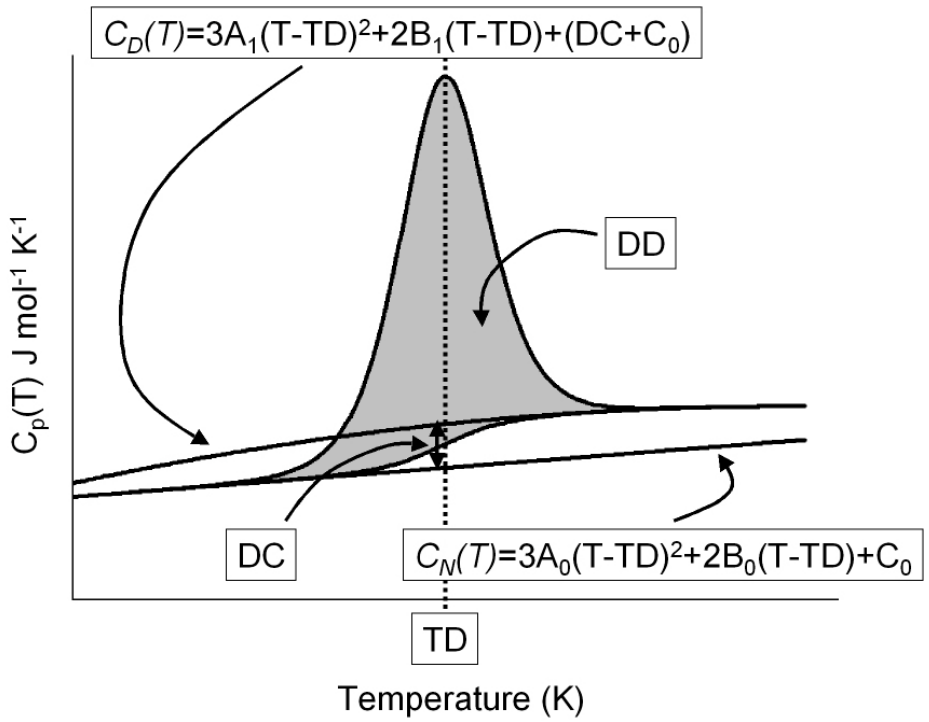

Figure 1. Graphical representation of model parameters A_0, B_0, C_0, A_1, B_1, DC, DD and TD for the thermal denaturation of a hypothetical protein.

Table 1. Parameter values and resulting restraints of the fitting model

Parameter Value	Model
$A_0 \neq 0$	Native state baseline second-order polynomial
$A_0 = 0$	Native state baseline linear
$A_0 = 0$, $B_0 = 0$	Native state baseline constant
$A_1 \neq 0$	Denatured state baseline second-order polynomial
$A_1 = 0$	Denatured state baseline linear
$A_1 = 0$, $B_1 = 0$	Denatured state baseline constant
DC = constant	ΔC_p @ TD fixed during refinement
$k = 1$	$\Delta H_{vH} = \Delta H_{cal}$
k = float	1) Evaluation of concentration errors/extinction coefficient ($\Delta H_{vH} = \Delta H_{cal}$) 2) Evalution of $\Delta H_{vH}/\Delta H_{cal}$ (with known concentration)

Parameter Refinement

Parameter fitting utilizes the Levenberg-Marquardt nonlinear regression method [9, 10]. Initial guesses for the parameters are determined without operator intervention using the following simple algorithm. The value of TD is assigned to be the temperature of the largest y value (heat capacity) in the input data set. The parameters A_0, A_1 and DC are assigned default values of 0, and parameter *k* is assigned a default value of 1.0. A linear function is determined using the first and last data points. Parameters B_0, B_1 are set equal to the slope of this linear function and C_0 is set equal to the y intercept referenced at the TD value. The initial guess for the DD term is determined by taking the difference between the value of the data set at TD and the above described baseline function and multiplying by 10K (we have found this to be a generally useful heuristic for proteins). The program then performs a single iteration of least-squares refinement and presents the resulting values as appropriate initial guesses for all parameters. The merit function for refinement, χ^2, is defined as follows:

$$\chi 2 = \sum_i \left(y_i - y\left(x_i, a_j\right)\right) \quad (18)$$

where *i* ranges from 1 to the number of input data points, and a_j is the parameter vector with *j* parameters (i.e. 9 in the present model). The gradient of the merit function is determined by calculating the partial derivatives of the fitting equation (17) with respect to each of the parameters and solving for the minimum. Thus, the model includes nine partial derivatives. These equations are solved simultaneously using the method of Gauss-Jordan elimination. Fit convergence is achieved when a predetermined value for the standard deviation is reached, $\Delta\chi^2$ is sufficiently small, a specified number of iterations are completed, or χ^2 does not improve for a certain number of iterations. The user determines the choice of convergence criteria.

Analysis of Test Data

Analysis of test data involved the temperature-dependent phase transition for human acidic fibroblast growth factor (FGF-1). FGF-1 has been extensively characterized with regard to DSC, circular dichroism (CD) and fluorescence data for a wide range of temperature and denaturant conditions [11]. The denaturation of this macromolecule has been demonstrated to be two-state and reversible under the conditions employed. Experimental determination of the denatured state heat capacity function, $C_D(T)$, over the temperature range 273-363K was achieved by analysis in the presence of 3.95M GuHCl, 20mM *N*-(2-Acetamido)iminodiacetic acid, 100Mm NaCl, pH 6.60. The temperature of maximum stability of FGF-1 is approximately 290K [11]. In 3.95M GuHCl at 290K the $\Delta G_{unfolding}$ extrapolates to -47kJ/mol [11], to give a $K_{eq} = 2.35 \times 10^8$. Thus, in 3.95M GuHCl there is no temperature at which the native state is significantly populated, and the experimentally observed heat capacity function represents $C_D(T)$ throughout the entire temperature range.

Discussion

Program Speed

The program utilizes algebraically-derived derivatives, as opposed to using numerical methods, to calculate the partial derivatives required for the Levenberg-Marquardt non-linear least squares refinement method. The speed advantage of such an approach was evaluated by comparison to an identical model implemented within a general-purpose non-linear least squares fitting program (*DataFit*, Oakdale Engineering) that relies upon numerical methods to determine partial derivatives. Notwithstanding any graphics-related overhead, the results showed that the implementation of algebraically derived derivatives is approximately 25 times faster for an equivalent number of refinement cycles.

Program Accuracy

The accuracy of the general purpose non-linear least squares fitting program *DataFit* has been verified with the *Statistical Reference Datasets Project of the National Institute of Standards and Technology* (NIST). The refined parameters from our program, for every data set analyzed, are in excellent agreement (i.e. to 8 digits of precision) with those from *DataFit* and typically exhibit slightly improved values for the fitting residual.

Implementation of Independent Baselines

Native and denatured state baselines in our fitting model are implemented as independent, second-order polynomial functions. Thus, $\Delta C_p(T)$ is likewise a second-order polynomial. This is in contrast to other DSC analysis programs (e.g. *DSC for Origin*, MicroCal Software) that constrain $\Delta C_p(T)$ to be a constant throughout the temperature range. Other than ease of calculation, there is no basis to assume that the heat capacity functions of different phases of a macromolecule will be parallel. For those proteins where the native and denatured state heat capacity functions are clearly not parallel, forcing $\Delta C_p(T)$ to be a constant will result in systematic residual scatter indicative of non-two-state behavior (Fig. 2).

Ability to use second-order polynomial $C_N(T)$ and $C_D(T)$ functions

Analysis of the DSC data for denatured FGF-1 indicates that the $C_D(T)$ function exhibits a noticeable positive curvature (Fig. 3). A second-order polynomial is a more accurate model to this data than a linear function. Available experimental data for *Cp(T)* functions of other proteins over such an extensive temperature range is limited, but indicates that negative curvatures are more commonly observed [12, 13]. The implementation of second-order polynomials still allows the use of linear functions, since the coefficient for the second-order term can be set to zero. For some proteins with generally low thermal stability (like FGF-1) the ability to perform DSC analysis in the presence of denaturing conditions allows the entire $C_D(T)$ function to be determined and to evaluate whether a second-order polynomial fit is appropriate.

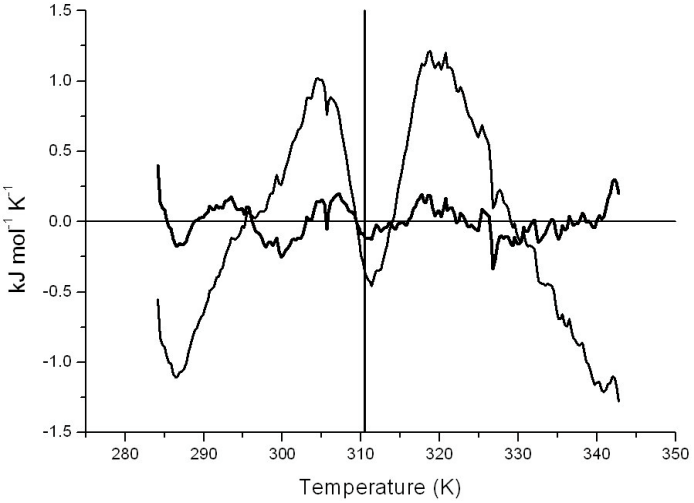

Figure 2. Residual scatter for the fitting of DSC data of human acidic fibroblast growth factor [11]**. The light line indicates the residual scatter using a model with constant $\Delta C_p(T)$. The heavy line indicates the residual scatter using a model where $C_N(T)$ and $C_D(T)$ are independent functions.**

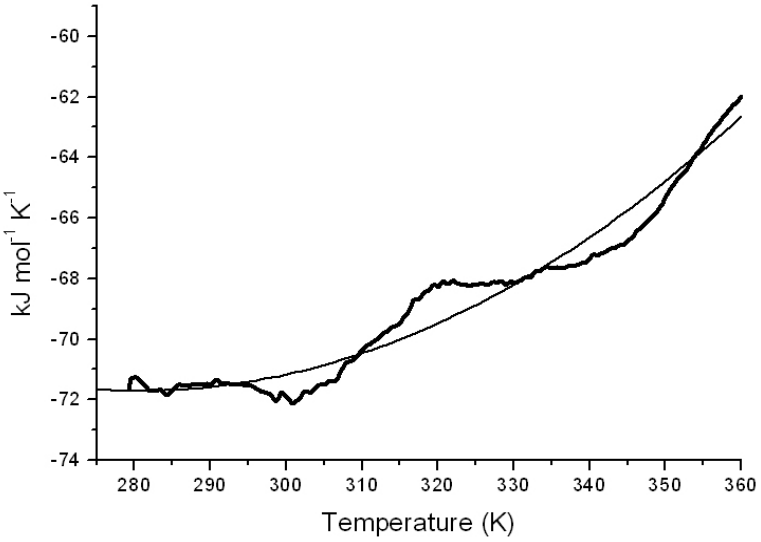

Figure 3. *$C_D(T)$* function for human FGF-1 (DSC analysis performed in 3.95M GuHCl buffer). The thin line shows a second-order polynomial fit to the data.

Inclusion of Concentration-dependent Parameter, *k*

Raw DSC data for FGF-1 at a concentration of 0.04mM (as determined from the spectroscopic extinction coefficient), were normalized prior to analysis using values of either 0.02mM, 0.04mM or 0.06mM in order to evaluate concentration errors. If the concentration parameter *k* was allowed to float during the fit it converged to a value of 0.489, 0.979 and 1.47 for the data normalized as 0.02, 0.04 and 0.06mM, respectively. Using these values for *k* and the respective normalized concentrations, the corrected concentration of the protein was determined to be (0.02mM/0.489), (0.04mM/0.979), (0.06mM/1.47), or 0.0409mM in each case. Using *k* in this manner, the actual concentration of FGF-1 in the sample was determined to be 0.0409 mM, or within 2% of the concentration determined by spectroscopic means. The parameter *k* can thus be used in two different ways

to provide additional useful information from DSC data; it can be interpreted as either 1) a concentration parameter (assuming that the transition is purely two-state), or 2) the value of *van't Hoff* to calorimetric enthalpy (i.e. $k = \Delta H_{vH}/\Delta H_{cal}$) if the concentration has been determined accurately.

In summary, we have developed a special-purpose flexible-model program, *DSCfit*, designed for the analysis of DSC data. Because of algebraically-derived derivatives, the program is significantly faster and more accurate than other DSC analysis programs that rely upon general purpose non-linear least square routines utilizing numerical methods (e.g. *DSC for Origin,* and *DDCL* [6, 14]). The program can also model more accurately situations in which the native and denatured state heat capacity functions are independent and non-linear (for polypeptides the denatured state heat capacity function is decidedly non-linear [15]). With the current generation of instrumentation [13, 16] such features of the heat capacity functions are often demonstrable. A completely novel feature found in the program is the inclusion of the concentration parameter, *k*. In this regard, it is now feasible to use DSC as a method to determine the extinction coefficient of macromolecules. To our knowledge, this use of DSC data to determine an extinction coefficient has not previously been described.

Acknowledgements

We would like to thank Drs. S. Kidokoro and K. Chiba-Kamoshda for helpful discussions and use of the *DDCL* program. We are also grateful to Ms. Sachiko Blaber for her translation of the *DDCL* user's manual. The *DSCfit* program and user's manual are freely available at *http://wine1.sb.fsu.edu/DSCfit/*. This work was supported by grants from the Florida Space Grant Consortium and the Florida division of the American Cancer Society.

References

[1] de Prat Gray G. Johnson C.M. and Fersht A.R. (**1994**) *Prot. Eng.*, *7,* 103-108.
[2] Matthews B.W. (**1996**) *FASEB J.*, *10,* 35-41.
[3] Johnson C.M. Oliveberg M. Clarke J. and Fersht A.R. (**1997**) *J. Mol. Biol.*, *268,* 198-208.
[4] Privalov P.L. (**1980**) *Pure and Applied Chemistry*, *52,* 479-497.
[5] Freire E. and Biltonen R.L. (**1978**) *Biopolymers*, *17,* 463-479.
[6] Kidokoro S.-I. and Wada A. (**1987**) *Biopolymers*, *26,* 213-229.
[7] Kidokoro S.-I. Uedaira H. and Wada A. (**1988**) *Biopolymers*, *27,* 271-297.
[8] Privalov P.L. and Makhatadze G.I. (**1990**) *J. Mol. Biol.*, *213*
[9] Levenberg K. (**1944**) *Applied Mathematics*, *2,* 164-168.
[10] Marquardt D.W. (**1963**) *SIAM Journal*, *11,* 431-441.
[11] Blaber S.I. Culajay J.F. Khurana A. and Blaber M. (**1999**) *Biophys. J.*, *77,* 470-477.
[12] Makhatadze G.I. Clore G.M. Gronenborn A.M. and Privalov P.L. (**1994**) *Biochemistry*, *33,* 9327-9332.
[13] Privalov G. Kavina V. Freire E. and Privalov P.L. (**1995**) *Anal. Bioch.*, *232,* 79-85.
[14] Nakagawa T. and Oyanagi Y. (1980) In: Matsushita K (ed) Recent developments in statistical inference and data analysis . North Holland Publishing Co. p 221-225
[15] Gomez J. Hilser V.J. Xie D. and Freire E. (**1995**) *Proteins*, *22,* 404-412.
[16] Plotnikov V.V. Brandts J.M. Lin L.N. and Brandts J.F. (**1997**) *Anal. Bioch.*, *250,* 237-244.