University of California, San Francisco

DSFworld data analysis

Gestwicki Lab
March 18, 2020

This website and its analyses are presented in a publication currently under review (as of 2020-03-17): "Three Essential Resources to Improve Differential Scanning Fluorimetry (DSF) Experiments." What follows here is a combination of the main body text from "Section III: Data Analysis" of that paper, and its supplementary information. The full code for DSFworld, as well as stand-alone scripts and modularized R Shiny web apps for its various capabilities, are freely available on GitHub, at https://github.com/gestwicki-lab/dsfworld. If you use DSFworld to analyze your data, or build your own analysis pipelines, please spread the world by citing the paper! For more information, you can download the full paper and its supplementary information from the "About DSFworld" tab.

DSFworld analyzes raw DSF data. It accepts raw Temperature vs Fluorescence data, and exports visualizations and apparent melting temperatures. Most DSF data can be easily analyzed using either of two general approaches—first derivative or sigmoid fitting—and DSFworld supports both. These different analysis methods don't represent conflicting interpretations of DSF data; rather, they are different mathematical approaches which typically return the same Tm_a value and can be used interchangeably.

The fitting methods may not work for some systems (see caveats below). To support customization and modification of the DSFworld analyses in these cases, the full code for the DSFworld website, as well as stand-alone R scripts, and modularized applications for data uploading, data formatting, plotting, Tm_a determination, and downloading are available on Github.

The methods used to determine the apparent melting temperatures calculated on this website are as follows:

First derivative, single Tm

Using this method, a single Tm is calculated for each DSF curve in the following manner:

1. The first derivative of the input data are calculated using a Savistky-Golay filter with a filter length of three degrees.
2. A Loess smoothing function is then used to interpolate the first derivative data to 0.1 C increments.
3. The maximum of the interpolated first derivative data is returned as the Tm_a.

The script and all associated functions to implement and modify this analysis outside of DSFworld is available on GitHub.

Sigmoid fitting, four possible models.

To determine the number of models necessary to describe DSF data broadly, we first generated a representative dataset of DSF results. Briefly, we assembled a panel of X proteins diversified in molecular weight, biological activity, fold, and oligomeric state. We then performed DSF experiments in a variety of standard conditions, varying buffers, pH, concentrations of known ligands, SYPRO Orange concentrations, and heating rates. From these 347 DSF results, four archetypes of curves were visually identified: a single transition with no initial fluorescence (Model 1), a single transition with high initial, decaying fluorescence (Model 2), two transitions with no initial fluorescence (Model 3), and two transitions with high initial, decaying fluorescence (Model 4). These archetypes were mathematically defined as follows:

$$RFU(T) = Sig_1(T) \quad \text{(Model 1)}$$

$$RFU(T) = Sig_1(T) + Id(T) \quad \text{(Model 2)}$$

$$RFU(T) = Sig_1(T) + Sig_2(T) \quad \text{(Model 3)}$$

$$RFU(T) = Sig_1(T) + Sig_2(T) + Id(T) \quad \text{(Model 4)}$$

Where the general form of the decaying sigmoids, $Sig_i(T)$, is:

$$Sig_i(T) = \frac{A_i}{1 + e^{\frac{Tma_i - T}{scali}}} \times e^{d \times (T - Tma_i)}$$

- $Sig_i(T)$ is the RFU value at temperature T

- A_i is the scaling factor for the final sigmoid
- Tma_i is the Tm_a
- $scal_i$ controls the slope of the transition
- d is the magnitude of the temperature-dependent RFU decay

And where the general form of the initial decaying fluorescence, $Id(T)$, is:

$$Id(T) = C \times e^{(T \times id)}$$

- $Id(T)$ is the RFU of the initial fluorescence at temperature T
- C is the starting value of the initial fluorescence
- id defines the rate and linearity of the decay from C

Approximately 10 datasets of each visual archetype were extracted and used throughout development of the curve fitting scripts. The resulting script was tested by fitting the 347-curve dataset, after which no further modifications were made to the script to maximize the relevance of the test results to the performance of the exact procedures applied at DSFworld. A stand-alone script for the model fitting is available at GitHub, alongside the 347-curve sample dataset used to test it.

The final fitting procedure is as follows:

1. A normalized version of the full uploaded dataset is generated by individually normalizing both raw RFU data and temperatures to a 0 to 1 range. This step minimizes inconsistent curve fitting behavior from arbitrary differences in measured temperature ranges and magnitudes of RFUs reported from different qPCR instruments.

2. From the normalized raw data, first- and second-order derivatives are calculated with respect to temperature using a Savitsky-Golay filter over a three-degree window (the number of individual measurements in a three-degree window is calculated from the non-normalized temperatures). A smoothed, interpolated version of the first- and second-derivative data are then calculated using a Loess filter of span 0.1 (10 percent of the measured temperature range).

3. Starting parameters for the models are generated in the following manner:

   - *Tm_a of the major transition (All models)*
     In DSF, multiple transitions typically present as one high-magnitude (major) transition joined by a smaller (minor) one. In our experience, reasonable estimates for the Tm_a of the major transition can

typically be identified as the smooth major peak in the first derivative data; this is true for curves with both single and multiple transitions. Specifically, smooth peaks are identified in the Loess-smoothed first derivative data using the findPeaks() function from the quantmod package. Any peaks identified in the first or last five data points in the run are then removed because (i) in our experience, irreproducible, noise-like variations in DSF data are highest in these regions, meaning that such peaks are typically artifactual and (ii) reproducible transitions in DSF typically occur over at least a 10-measurement window, meaning that any transitions within the first or last five data points, if not artifactual, are likely partial. We have not yet encountered an application which required accurate fitting of partial curves in DSF (it is better to extend the measured temperature range, or optimize reaction conditions to bring the transition fully within the measured temperature range), so we did not optimize or test the performance of the DSFworld curve fitting in these cases. Finally, to eliminate peaks resulting from minor noise, any identified peaks with a maximum dRFU value of less than 0.0002 are removed as well. Peaks are then rank-ordered by their maximum dRFU, such that the largest peak in the dRFU is provided as the starting estimate for the Tm_a of the first sigmoid in subsequent model fitting.

- *Tm_a of the minor transition (Models 3 and 4)*
  Estimation of the Tm_a of the minor transition are often more challenging. In DSF data, minor transitions often occur close in temperature to the major peak, appearing in the first derivative data as a shoulder on the primary peak. None of the peak finding algorithms we tested for use at DSFworld consistently identified these shoulders. However, we found that the minor transitions were more consistently captured as small stretches of positive-slope linearity in the raw data. These stretches can be quantified as valleys in the second derivative of the raw data. Specifically, smooth valleys are identified in the Loess-smoothed second derivative data using the findValleys() function from the quantmod. From the identified valleys, any which occur when the first derivative is < 0 are discarded. This step is necessary to separate the desired estimated Tm_as of the minor transitions from the largely linear, negatively-sloped post-maximum regions which occur in most DSF data. Any valleys which occur in the first or last five measurements are discarded for the same reasons described in the previous paragraph.

  To generate the final starting estimates for model fitting, the temperatures at which the identified peaks and valleys occur are combined. If multiple Tm_as are estimated within a three-degree window, only the lowest-temperature Tm_a of the closely-spaced estimates are retained. If only one Tm_a estimate is returned, a second Tm_a is estimated as the measurement directly after the first estimated Tm_a.

- *Magnitude of initial, decaying fluorescence (Models 2 and 4)*
  The magnitude of the initial fluorescence is estimated as the first RFU value in the dataset.

- *Estimation of remaining parameters*
  The starting estimates for the remaining parameters are the same for all input data, because their empirical variation is, in our experience, relatively small. These are set as follows:

Parameter	Parameter description	Starting estimate	Used in models...
Asym1	Relative magnitude of major transition	1.0	1, 2, 3, 4
Asym2	Relative magnitude factor for minor transition	0.1	3, 4
xmid1	Tma of major transition	data-dependent	1, 2, 3, 4
xmid2	Tma of minor transition	data-dependent	3, 4
scal	Slope of major transition	0.03	1, 2, 3, 4
scal2	Slope of minot transition	0.03	3, 4
d	Temperature dependent RFU decay of major transition	-1	1, 2, 3, 4
d2	Temperature dependent RFU decay of minor transition	-2	3, 4
id_d	Relative magnitude of initial fluorescence	data-dependent	2, 4
id_b	Temperature dependent RUF decay of initial fluorescence	-5	2, 4

Table 1: Summary of starting estimates used in DSFworld model fitting.

4. From the resulting fits, the final parameters are then used to generate curves for each individual component of the full fit, multiplied by their individual temperature-dependent decay terms. These components are: the first sigmoid (all fits), the second sigmoid (fits 3 and 4), and high initial fluoresence (fits 2 and 4). The Tm_a values are then calculated by taking the maximum of the first derivative of each isolated sigmoid component, as described section 2.1 of this supplementary note. This approach is analogous to the separation of a complex melting transition into its most likely individual unfolding populations, and then the use of the currently-accepted methods of Tm_a, which do not account for the influence of temperature-dependent decay on the midpoint of the transition, to each isolated population. This way, the DSFworld analysis leverages the temperature-dependent fluorescence decays necessary for robust fitting of complex transitions, while remaining consistent with the existing practices for DSF data analysis. The impact of correction for fluorescence decay on

the model parameters themselves are discussed further in the Comments and Caveats section below.

5. Users of DSFworld can fit their data to any combination of the four models provided at DSFworld. The fit results of each model to all curves in the uploaded data can be downloaded individually. However, not all curves may be best described by the same model. By default, for summarized results in which only one fitted model per curve is desired, DSFworld will return the results from the fitted model with the lowest Bayesian Information Criterion (BIC) for each individual curve. The model of choice can also be manually selected by the user by clicking on the desired fit in the "select best model" plot in the analysis window.

Considerations and caveats

- *Replicate handling.* Even when replicates of a particular condition have been defined by the user, a Tm_a is calculated for every individual curve, and the Tm_a for the condition is reported as the mean of the Tm_as calculated for each individual replicate, +/- standard deviation. When model fitting is used, if different models are selected for replicates of the same condition, results from the selected models with the smallest number of free parameters for that condition are applied to all members of that condition. Models 1, 2, 3, and 4 have four, six, eight, and ten free parameters, respectively.

- *Temperature-dependent fluorescent decay, and ramifications for the interpretation of model parameters.* DSFworld model fitting incorporates a temperature-dependent fluorescence decay coefficient throughout the entire the curve. At this time, we believe this approach is appropriate for the following reasons:

  - Fluorescence decay is not observed in low-temperature readings in up-down mode experiments. In an assembled panel of seven diverse proteins, fluorescence was monitored at both the high and low temperatures of an up-down mode experiment (Figure S2), and the results were compared to the results from straight-ramp experiments presented in Figure 1. While the Tm_as determined for the high- and low-temperature readings were similar, fluorescence decay was observed only in the high-temperature readings, and not in the low-temperature readings. Furthermore, the observed decay in the high-temperature readings of the up-down mode experiments matched that of the straight-ramp experiments. Combined, this observation suggests that the observed fluorescence decay in these experiments may be largely temperature dependent.
  - The stereotypical linearity of the post-maximum decay in real DSF data conflicts with a model in which this decay is caused by the transition of the protein into a final aggregated state which excludes

the dye. While this dye-exclusion model may certainly be true for some proteins, given this consistent linearly, we do not think it is an appropriate assumption for general-purpose DSF fitting tools.

- The strength of the protein-small molecule interactions is often decreased with temperature. Therefore, if the current hypothesis that SYPRO-protein interactions are primarily hydrophobic in nature is correct, an associated temperature dependent loss of dye binding is expected, even if all dye binding sites are retained as temperature is increased. This is anecdotally supported by observations that when SYPRO fluorescence is induced by the folded state of a protein, this fluorescence exhibits a very similar decay during heating through temperatures at which the protein undergoes no conformational changes–and by extension, changes in dye binding sites–as measured by circular dichroism (unpublished data).

The application of a temperature-dependent decay to the full sigmoid creates the following behaviors, of which users interested in direct interpretation of the parameters returned from DSFworld fits (as opposed to the Tm_as returned from fits, which are robust to these effects)should be aware:

- Tm_as determined by sigmoid fits which incorporate temperature-dependence fluorescence decay will be higher than those calculated by methods which do not, such as the maximum of the first derivative. For the majority of DSF data, this difference is 1.5 C. Steeper decays produce larger differences.
- Two DSF curves may have the same Tm_a when analyzed using the maximum of the first derivative. However, if the two curves have different decay intensities, methods which incorporate temperature-dependent fluorescence decay, such as the sigmoid fits provided at DSFworld, will report different Tm_as for the two curves. This difference is typically small (< 2 C), but increases as the difference in decay intensities increases.

It is possible that the above effects manifest in real DSF data, and are systematically overlooked by current analysis methods. If so, the correction of temperature-dependent decays in standard Tm_a calculations may be appropriate. However, more information on the mechanisms and consistent behaviors of temperature-dependent decays in DSF will be necessary to answer this question broadly.