DSF Analysis

Excel script for the analysis of protein unfolding data acquired by Differential Scanning Fluorimetry (DSF)

Manual

Version 3.0

(Release date: 11th November 2010)

Frank Niesen, PhD

Structural Genomics Consortium, Oxford

ORCRB, Roosevelt Drive, OX3 7DQ, United Kingdom

This manual and all files presented and discussed in it can be found in ftp://ftp.sgc.ox.ac.uk/pub/biophysics

Dear user,

Thank you for downloading this version of DSF Analysis, an MS Excel-based tool to analyze and visualize data from fluorescence-based thermal-shift screens. This type of screen, used to identify stabilizing conditions including ligands/additives for proteins, is known as Differential Scanning Fluorimetry (DSF) as well as a number of other names, e.g., Thermofluor or Temperature-dependent fluorescence (TdF).

DSF is sturdy, easy-to-use and doesn't require specialized equipment other than a standard real-time PCR machine. Since I began using the method it has advanced to the most frequently used generic method to determine protein stability. With the development of *DSF Analysis* during the last five years my main aim was to create a tool for the fast analysis and visualization of protein unfolding data that is accomplished without using any specialized software. The goal of the Structural Genomics Consortium is the determination of X-ray structures for proteins from humans and from Apicomplexan organisms causing human disease. Frequently, the addition of a ligand or change to more favourable solvent conditions is necessary to achieve diffracting crystals for our targets, and *DSF* is the main method used by us to detect such 'molecular helpers'. The need for screening of a multitude of target proteins every day, against many compounds and solvents was the motivation for me to create tools to facilitate quick, user-friendly and error-preventing analysis of the results from *DSF*.
Compared to earlier versions of DSF Analysis the version 3.0 contains a number of changes, most of which were suggested by users whom I thank very much for their input! The most noticeable difference is that there are no longer instrument output-specific versions of DSF Analysis; it became hard to ensure the identity of all functions between the different files every time a change or bug-fix was introduced. The customization for each of the available instruments is now accomplished by the use of specific "transformer" files: each of these generate an output from the data that has an identical layout, to be imported into the main DSF Analysis.

However, it took quite some time to fully achieve this goal: although previous versions of *DSF Analysis* offered a to transfer and fit a number of curves in the fit software, there was so far no possibility to process all 96 curves from an experiment in parallel. I am happy that the version I am presenting to the community today has achieved such 'full functionality': all that is necessary to analyze a dataset is Microsoft Excel and mathematical software that accomplishes fitting to the Boltzmann equation, such as GraphPad.
Protein unfolding can be quite a complicated matter, and the *DSF* curves may contain a wealth of information about how a protein unfolds (e.g. if it forms oligomers or consists of different domains). My aim with *DSF Analysis* is to facilitate annotation and recording of all parameters that may become important for the interpretation in combination with techniques such as CD, DSC, X-ray crystallography and many others. The calculation – and transfer into the database – of the slope within the T_m (indicative for cooperativity of unfolding) and of the temperature at which the maximum of the transition is reached (may indicate potential stabilization by a ligand whose effective concentration is

decreased due to, e.g., low solubility in the solvent used) reflect this aim and, hopefully, will be beneficial towards different biological questions the user may address with *DSF*.
I wish to emphasize that I am not the developer of the method itself; my contribution is merely in refining the concept and make it accessible to a wider audience. In addition, I hope to have contributed to the understanding of the phenomenon as well as to the assessment of the applicability of *DSF* to biological questions. As described in detail within our summary of the success thus far with the application of *DSF* to detect conditions and ligands that improve crystallizability (Vedadi, Niesen, et al., 2006), testing of proteins for stabilizing conditions using fluorescent, hydrophobicity-sensing dyes was first reported by Gorazd Vesnaver and coworkers (Poklar et al., 1997). It was the work of F. Raymond Salemme and his coworkers (Pantoliano et al., 2001) that showed the applicability of the method to screening of compounds in high-throughput, while, however, using a special and expensive, self-developed instrument. A number of people, among them Mei-Chu Lo (Lo et al., 2004), then realized that a non-proprietary instrument common to biology laboratories, namely a real-time (RT-) PCR, can be used for *DSF* screening.

I am grateful to all my co-workers at the SGC, and to the increasing number of people in the 'DSF community' who supplied their feedback and ideas that were necessary to accomplish my goal. Please continue supplying these valuable comments via email to me: dr.frank.niesen@gmail.com. I acknowledge the support for the SGC from our body of funding organizations: The Structural Genomics Consortium is a registered charity (number 1097737) and receives funds from the Canadian Institutes for Health Research, the Canadian Foundation for Innovation, Genome Canada through the Ontario Genomics Institute, GlaxoSmithKline, Karolinska Institutet, the Knut and Alice Wallenberg Foundation, the Ontario Innovation Trust, the Ontario Ministry for Research and Innovation, Merck and Co., Inc., the Novartis Research Foundation, the Swedish Agency for Innovation Systems, the Swedish Foundation for Strategic Research and the Wellcome Trust.

At the end, I would like to mention that *DSF Analysis* and all other tools and templates from our ftp site are supplied entirely without restriction of use! The release of information and any tools that may be of help to the scientific community is the aim and purpose of the SGC, and I am happy to comply with it. I would, however, be grateful if results analyzed with the help of *DSF Analysis* would be referenced to our work*. Finally, as a disclaimer, I would like to point out that I cannot accept responsibility for any damage, misinterpretation or faulty analysis; the use is at anyone's own risk.

Many thanks, and my best wishes,

Frank Niesen Oxford, 11th Nov, 2010

*

- M. Vedadi, F. Niesen, A. Allali-Hassani, O. Fedorov, P. Finerty jr., G. Wasney, R. Yeung, C. Arrowsmith, L. Ball, H. Berglund, R. Hui, B. Marsden, P. Nordlund, M. Sundstrom, J. Weigelt & A. Edwards (2006): Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination. *Proc Natl Acad Sci USA* 103(43):15835-15840.
- F. Niesen, H. Berglund & M. Vedadi (2007): The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nature Protocols* 2(9):2212-2221.

About this manual

The aim of this manual is to explain how to use *DSF Analysis* - a Microsoft Excel-based worksheet to analyze, visualize and prepare for database processing data acquired from protein denaturation screens performed using an RT-PCR. For instructions on how to run *DSF* screens and a review of the method, the user is referred to the detailed description available (Niesen et al., Nature Protocols, 2007).
The manual as well as the script, examples, templates and additional information can be found on our ftp folder (ftp://ftp.sgc.ox.ac.uk/pub/biophysics). The best way to explain steps and functions is with an example. I used a dataset acquired with pig Citrate Synthase under a set of several generic conditions. The raw data (from an Agilent Mx3005p instrument), as well as ready processed files for the *Transformer* and *DSF Analysis* for this example are part of the DSF directory.

Formatting

Bold writing is used throughout the manual to denote tabs/worksheets within *DSF Analysis*. Furthermore, cells within worksheets are denoted in **bold letters**.

Tab colours

The colour-coding of the tabs of *DSF Analysis* indicates the extent of input asked from the user (Table 1).

Colour of tab	Type of worksheet	Worksheets of this type
Green	User input largely required or supported for annotations etc. For transfer of results between Excel and fit software	**Experiment Parameters** **All Graphs** **Processed Data**
Light Blue	No input from the user; location from which to paste or print	**Plate Layout for Printing** **Result at a Glance** **Database Table**
Dark Blue	Input of entire table from other Excel table	**Paste in transformed Data**
Orange	Customized visualization of selected curves	**Custom Graphs** **Custom Normalized Graphs** **Custom Referenced Graphs**
Grey	Administrator only	**Helptable Database Table**

Furthermore, a general color-coding for fields has been followed such that green fields denote entries requested from the user while light blue fields denote information that the user is supposed to transfer into other places such as a database or the software used for fitting of the curves.

Drop-down menus

In several places *DSF Analysis* contains fields requesting a user entry (indicated by the green colour) that only accept selected contents. For example, the unit for "protein concentration" (**F10** in **Experiment Parameters**) can be either "µM", "mg/ml", "mM", "µg/ml", "x" (times) or "%" [the list can, however, be extended with other entries into the source table for the entry, **I143-I151**]. Another example is the individual annotation of curves with respect to an apparent irregularity, to be found in **All Graphs**. The drop-down menu here comprises the entries from the list on the top of the worksheet. So, although the entries are custom the drop-down menu makes sure that they are not all different, which with respect to database processing would cause inconvenience.

Handling of templates

It is important to emphasize that many fields within *DSF Analysis* allow custom changes to otherwise automatically generated entries. Once changed by the user, however, the automatic entry generation is disabled. Using the same copy of *DSF Analysis* subsequently for another set of data would, therefore, result in improper function and could have faulty analysis as a consequence! Therefore, it is good practice to save *DSF Analysis* under a different file name as soon as the template is opened, and to keep a safe copy of the template as a backup (although it can be re-downloaded from our ftp site at any time, of course).

Data output from different RT-PCR machines

All available RT-PCR instruments that can be used for *DSF* (e.g., that allow the use of appropriate filters for *DSF* dyes such as SYPRO orange) have different outputs. It is, in principle, possible to accommodate all of these different outputs into *DSF Analysis* and, as pointed out earlier, this is accomplished by means of output-specific *Transformer* files. Users are encouraged to supply a typical output (.xls or .csv) of their machine to me, therefore, and I will look into creating a *Transformer* accordingly. Many thanks, at this point, for all users that have done so in the past!

Step-by-Step instructions using CS dataset as example

We will be using the data of a well-behaved control protein, pig Citrate Synthase (CS) as an example for how to work your way through the different functions of *DSF Analysis*. To generate the dataset, the protein was screened under three different buffer conditions against a set of common compounds. The data was recorded using an Agilent/Stratagene Mx3005p RT-PCR (file with raw data: "CS example Mx3005p.xlsx"). However, while the first step, i.e., of transforming the data into the common input format for *DSF Analysis*, differs from how other machine data outputs would be handled, all the steps concerning the *DSF Analysis* 'core' are similar no matter what machine has been used.

Transforming the file

- Open the file "CS example Mx3005p.xlsx" containing the raw data, i.e., one worksheet containing "**Chart Data horizontal**".
- Click on the upper left corner of the table, thus highlighting its entire content. Press CTRL+C, to copy to clipboard.
- Open the file "Transform Mx3005p data for DSF Analysis v3.0.xlsx".
- Go to worksheet **Chart Data horizontal**, click on the top left corner and press "CTRL+V". Note: The progress up to this point can be taken up by opening the file "CS example Transform Mx3005p.xlsx", for example for the benefit of users of a machine from a manufacturer other than Agilent.
- Change over to the worksheet **Output**. Make sure that the right values for the start and end temperatures of your experiment are filled into **C5** and **F5**, respectively.
- You will see that the entirety of your data is now transferred into a table that is headed by a row of numbers denoting the wells of the plate 1-96. In our example all 96 wells are used in the experiment. {Note: Most instruments generate an output from experiments where not all wells are being used with the empty wells omitted from the output: you will find that the transformer file, however, assorts the data to the correct well, leaving empty columns where appropriate.}
- It is of note that all *Transformers* limit the amount of data for processing to one point per degree Celsius. For experimental data comprising a higher reading frequency the intensity only of the "first" read at each temperature (e.g., for experiments with readings at 37.1, 37.3, 37.5, 37.7 and 37.9 °C only the reading at 37.1 °C) is taken further.

DSF Analysis processing 1: Parameters

- Within the *Transformer* file, worksheet **Output**, click on the top left corner, thus highlighting the entirety of the table. Press CTRL+C to transfer to clipboard.
- Open the file called "DSF Analysis v3.0.xlsx".
- Change over to the worksheet **Paste in transformed Data**. Click on the top left corner and press CTRL+V. No more input is needed in this worksheet (hence the dark blue color).
- Change over to the worksheet **Experiment Parameters**. Note that all fields in green are user entry fields. Start by choosing an experiment ID (**B6**), if you so wish. This ID

will be used to annotate each experiment row (i.e., each condition) within the **Database Table**.

- An entry in the field for Plate ID (**E6**) is useful as it can then be combined with the well ID of the experiment, respectively, within the **Database Table**.
- The entry choices in cells **I4**, **I6**, **L4** and **L6** are meant as basic selectors streamlining the process of filling in the parameters and removing the need for repetitive entries.
- Keep the default selection "yes" in **I4**; our screen comprised only a single protein and all wells contained it at a single concentration only. Because of this selection, we are only required to fill in information into **E11** and **F11** which are then copied into the rest of the table. If the selection is "no" in **I4** the user needs to fill out all table entries in columns **E** and **F**.
- It is possible to select custom units for the concentrations of protein and added compounds. Please select from the drop-down menu next to the cells in row **10**, highlighted in pink, respectively. If necessary, other units can be entered into the table **I143**-**I151**.
- In a similar manner "no" is the selection made in **I6** (the compounds in the present example were not tested at the same concentration; we need to fill in the information row by row), as well as in **L4** (the example contained no compound that was added to all 96 wells; it is sufficient to delete the entries in **K11** and **L11**), and in **L6** (different buffers were used in the example, so the description has to be filled into column **M** for each of the rows).
- For users that perform subsequent database processing there are entries for the IDs of each compound (columns **G** and **J**, respectively), as well as of the buffer (column **N**) that will be transferred into the **Database Table** worksheet at the end.
- For the calculation of effects on the stability of the tested protein ΔT_m values are calculated by subtracting the averaged T_m of all wells annotated as "Reference" (not case sensitive) from all other T_m values. To avoid inclusion of any reference well value, for example due to bad data quality in that well, simply rename the entry (as in the present example, as "not used Reference").
- It is possible to print out, from the worksheet **Plate Layout for Printing**, the configuration of the experiment. You will find that specific entries are omitted from the list dependent on the selections made in Experiment Parameters. For example, if "yes" is selected in **I4** the protein names and concentrations (same for entire experiment) are not shown, etc.

DSF Analysis processing 2: All Graphs

- Change over to the worksheet **All Graphs** and find the curves recorded from all 96 wells of the plate displayed as dark blue traces. Automatically highlighted, in turquoise color, are the parts of the curves with the unfolding transitions for the protein. It is useful to understand that this is done by determination of the maximum intensity within a selected temperature range, followed by determination of the minimum intensity at temperatures below the one with the maximum intensity. The temperature range between these minimum and maximum intensities are then extended by three degrees, so that the transition would contain two asymptotes for fitting to the Boltzmann equation.

- For difficult datasets it can be useful or desired to restrict the range of temperatures for fitting for all curves, by modifying the entries in **E2** and **E3** (default: start and end temperatures of the experiment, as given by the user during data transformation). In addition to this global modification, the user can also adjust the start and end temperatures of each dataset by replacing the appropriate numbers (e.g., in rows **17** and **18** for experiments in row **A** of the plate).
- Note that the entries for the slope (e.g., in row **19** for experimental row **A**) and T_m will not be filled until fitting of the datasets is performed, see below.
- Annotations of irregularities found with each of the experiments can be made, by selecting the appropriate entry from the drop-down menu (e.g., row **21** for experimental row **A**), respectively. These entries are transferred into the **Database Table**. Additional entries or amends can be made in/to the list, by editing the table in **C2-C12**. It is important to know that annotations starting with "ambiguous" will result in T_m values being omitted from the downstream analysis and visualization.

DSF Analysis processing 3: Processed Data

- As mentioned above, a second piece of software is necessary to determine the T_m values for the data, from fitting to the Boltzmann equation. In our laboratory the use of *GraphPad Prism* is standard, but many other programs can potentially be used if they allow on-the-fly processing based on a table comprising 96 columns, i.e., datasets, and have as output a table of T_m values (96 columns) that can be pasted back into *DSF Analysis*.
- To simplify the transfer of the data into the fitting software, a table is already highlighted within the worksheet **Processed Data**. Press CTRL+C to copy this table to the clipboard, from where it can be pasted into the fitting software (for *Prism*, a template is available on the SGC ftp site), followed by back-transfer of the resultant T_m table, into **D6-CU6** of **Processed Data**. No further input is needed in this worksheet unless the data processing (e.g., temperature ranges for fitting of curves) is re-edited. In such case, the transition data table (in grey between **C10** and **CU120**) needs to be highlighted using the computer mouse and re-copied into the fitting software (see above).

Data Visualization

- Change over to the worksheet **Result at a Glance**. A bar diagram displays the entirety of the results either as T_m or ΔT_m values; switching between the two displays is accomplished by selection in **N2** (drop-down menu; whether or not referencing is to be applied to the data). In some cases the user may find it necessary to re-visit and re-edit a few curves based on the display, by changing over to **All Graphs** and, subsequently, re-fitting the data as it is compiled in **Processed Data**.
- In the worksheet **Custom Graphs** the user may each fill in the plate coordinates of ten experiments to display them in the two graphs. The selection is supported by the display of the at-a-glance bar diagram (ΔT_m) on the left side.
- In the worksheet **Custom Normalized Graphs** the user may fill in the plate coordinates of ten experiments to display them in a graph after normalization to the

transition amplitudes (i.e., the difference between the fluorescence intensity at the start and at the maximum of the unfolding transition, respectively).

- In the worksheet **Custom Referenced Graphs** the user has the possibility to select the coordinates of up to nine experiments to be subtracted from a tenth (**I-J31**) experiment.

Database processing

- After all processing is finished, a large body of information about the experiment is automatically compiled and can be observed in the worksheet **Database Table**. From here it can conveniently transferred into a database or printed out. Note that empty wells, i.e., wells without annotated experimental content, will be automatically omitted from the list.

Appendix

Version no.	Released	Changes applied
1-2	01-08-2007	- Fixed bug in **All Graphs**: T_m transfer from T_m table and to **Result at a Glance** - Fixed bug in **Result at a Glance**: Transfer of comments from **All Graphs** - Fixed bug in **Database Table**: Transfer of comments from **Result at a Glance**
1-3	22-08-2007	- Fixed bug in **Processed Data** that led to error in ΔT_m calculations
1-4	17-09-2007	- Fixed error in T_m calculation from 2^{nd} derivatives (max of curves instead of min) in BioRad version
1-4	01-05-2008	- Introduced chart data help sheet to fix problem processing changed output from Stratagene PCR software (e.g. numbers preceded by spaces, table shifted down by one line)
2-5	11-09-2008	- Applied complete makeover including, e.g., introduction of a parallel preparation of curves for the fitting software - Added worksheet **Custom Normalized Graphs** - Created versions for Mx3005p and IQ5 output
2-6	12-01-2010	- Created versions for Applied Biosystems (ABi) outputs
2-7	29-03-2010	- Refined amplitude calculation - Added reporting of temperature at peak of intensity and amplitude to **Database Table**
3-0	11-11-2010	- Divided tool into instrument-specific *Transformers* and instrument-independent *DSF Analysis* - Extended allowed temperature range to 0 to 110 °C - Added customization for units - Refined procedure for automated detection of the transition temperature range - Added globally customizable transition temperature range selection - Added display of at-a-glance T_m or ΔT_m - Added worksheet **Custom Referenced Graphs**