HDXWizard Operating Instructions

Version 1.0

Zachary Cohen

Cohen.za@northeastern.edu

Contents

11. Saving Your File

1.Introduction, Compatibility, and Contact Information

Welcome to HDXWizard, a tool designed with a simple, intuitive user interface for the best and fastest processing of HDX data into chiclet plots, difference plots, and colored peptide plots and peptide difference plots. All plots can be analyzed using “experimental” (maxD) mode or “theoretical” mode to allow for diverse applications to different data sets. Data is inputted by the user from Waters DynamX state data. If you have any questions or feature requests, please email them to cohen.za@northeastern.edu.

2.File Entry

2.1.1) Preparing State Data

After data has been processed in DynamX, it can be exported as “State Data” into a .csv file. This file, or a .xlsx file can be uploaded directly to be analyzed. As many files as desired can be analyzed at once, but please make sure that states that are not the same have different “states” and/or different “proteins”.

To create a mock DynamX state data, make an excel file with columns:

A: Protein

B: Start

C: End

D: Peptide Sequence

E:

F:

G: Filled

H: Filled

I: State

J: Exposure Time

K: Filled

L: Filled

M: Uptake

N: Uptake SD

Columns E and F may be left empty. Columns G, H, K, and L should all be filled with any character.

2.1.2) Preparing Cluster Data

Cluster data can also be exported from DynamX and used. This is less preferrable to state data, as it has undergone less extensive testing. Additionally, the display of errors is not possible.

2.2) Preparing Sequence

If you only want to create a chiclet or chiclet difference plot, you can select the "skip" button. If you select the skip button and try to make any peptide plots, the sequence will be read as AAA etc. If you are only analyzing one protein, it is possible to enter a plain text file containing only the sequence of the protein using the ".txt (p)" button. The user can also enter a .fasta file or a .txt file in the format of a .fasta file using the ".fasta" and ".txt (>)" buttons. As many of these as desired can be added, but please make sure the protein name in the fasta file (names formatted as ">Protein Protein" will be read as "Protein") is exactly the same as in the state data file, otherwise the sequence will be replaced with AAA etc.

3. Selecting a Preference for RFU Calculation or MaxD Correction

3.1.1) Theoretical Preference

Selecting the theoretical button will provide any chiclet and peptide plots as their relative fractional uptake, or RFU by dividing the measured uptake by the theoretical maximum uptake which is calculated by:

$$Max\ Theoretical\ Uptake = \ Length\ of\ Peptide - Prolines - 1$$

If the first residue is not proline, or:

$$Max\ Theoretical\ Uptake_{N\ terminal\ P} = Length\ of\ Peptide - Prolines$$

If the first residue is proline.

$$RFU = \frac{Measured\ Uptake}{Max\ Theoretical\ Uptake}$$

If duplicate peptides exist for a state, they will be averaged.

3.1.2) Customizing Back Exchange

It is possible to customize coloring of peptide plots for an amount of back exchange by entering a number into the box labeled "Insert Back Exchange". Numbers entered should be the percent back exchange you would like to select, such as "25". If no number is placed into the box, back exchange will

be set to 0%. Maximum theoretical uptakes are multiplied by $(1 - \frac{back\ exchange}{100})$, and peptide uptake values at timepoints are divided by this number to find the RFU.

3.1.3) Differences maps with Theoretical Preference

When the theoretical preference is on, all differences will be the raw uptake differences between states in Daltons.

3.2.1) Experimental Preference

If you have data for a maximally deuterated control (maxD) that you would like to be corrected for, then selecting the experimental preference is the correct option. Please make sure that your maxD data is the largest timepoint. You can, but do not have to have the maxD in every state if the maxD data is compatible with those states. If duplicate peptides exist for a state, they will be averaged.

3.2.2) Missing maxD peptides

If the value of the highest peptide is missing (or was set to -99999 by the user), then the average RFU for all maxD peptides in the state is calculated, and the theoretical maximum deuterium incorporation of the peptide is multiplied by the average RFU of maxD peptides in the state.

$$Corrected\ RFU = \frac{Uptake}{MaxD\ Uptake}$$

or (if no maxD peptide can be found)

$$Corrected\ RFU^{*} = \frac{Uptake}{Max\ Theoretical\ Uptake} * Average\ maxD\ theoretical\ RFU$$

For peptides where the assigned maxD value, preference is missing, the second equation is used, and peptides are marked with an asterisk on peptide plots. In chiclet plots, the whole row is marked with an asterisk.

3.2.3) MaxD State Selection

Automatically, the program will attempt to get the maxD of each peptide from the same state as itself, however, after selecting “Experimental”, many dropdown menus will show up, and allow you to change the state where the maxD for the peptide is looked for.

3.2.4) Creating Custom States

By clicking the create custom states button, you can create an average of two states from which to take the maxD measurements from. Select the two states you wish to average, and click save state, then apply it to any states you want to use it for. Average RFU is calculated for both states’ maxD measurements combined. If you are having trouble with your custom state, make sure there are no “|” symbols in the name of either state you want to average.

You cannot use custom states to directly generate figures, only to set as a maxD. If you want to create a map of the average of two states, simply put them into the file with the same state and protein name, and they will be averaged automatically.

4. Adding and Customizing Differences

4.1) Adding Differences

Differences can be added by clicking the “+Dif” button. Two dropdown lists with states from your file should become available. If there are no states here, the program was not able to read your file. Select the states you want to create a difference for. The second state will be subtracted from the first. Then, give a title to this state (less than 25 characters). This title will be the title of any sheets and chiclet plots created for this state, so make it descriptive. At this time, only 12 differences can be created at once. If you wish to create more, you will have to run the program multiple times.

4.2) Calculation of Differences Using Theoretical Preference

When the theoretical preference is on, all differences will be calculated using the absolute uptakes between states.

4.3) Calculation of Differences Using Experimental Preference

When the experimental preference is on, differences between states are calculated by subtracting the RFU of the second state from the RFU of the first state.

4.4) Differences with Missing Timepoint and Peptides

Missing peptides or timepoints will not affect the accuracy of output. If peptides are not present in one difference state but are in the other, they will not appear in the difference maps at all.

5. General Notes on Data Processing

5.1) Absent Peptides

If a peptide is not present for a timepoint but is present for other timepoints in the same state, it is considered absent, and will be colored peach (by default), and will not affect data. Any differences using absent peptides will produce an absent, peach colored peptide in the difference plot. If any average is taken between an absent peptide and a peptide with a value, the new value will ignore the absent value.

5.2) Duplicate Peptides

If duplicate peptides are present with the same state name, the uptake values will be averaged before anything else is done to the data.

6. Chiclet Plot Functionalities

6.1) Chiclet Plot

By selecting the “Chiclet” button before hitting run, the program will create a chiclet plot of the peptides from your file. The peptides in your state data do not need to be in order. They are sorted by start value,

with tiebreakers being decided by end value before being placed into the plot. Each chiclet plot is titled using the state name from the state data. The timepoint 0 is automatically removed. If you want to bypass this, you can replace the name of that timepoint with another nonzero timepoint that would also be the lowest timepoint and change the timepoint post processing in excel. Chiclet plots are colored according to section 8.1.

6.2) Chiclet Difference Plot

By selecting the "Chiclet Difference" button before hitting run, the program will create a difference plot in the style of a chiclet plot between any two states you have added for a difference. Peptides that do not exist for both states are omitted from the difference completely. Each difference plot will be titled as the difference name inputted for the difference. Difference plots are colored according to section 8.2. Please see section 6.1 for more information on chiclet plot creation.

7. Peptide Plot Functionalities

7.1) Peptide Plot

By selecting the "Peptide Plot" button before hitting run, the program will create a series of sheets, each titled as the appropriate state from the state data file. Each sheet contains a series of timepoints in order of colored peptides. Peptides are placed by finding the highest place on the sheet within the given timepoint where there would not be overlayed upon another peptide. Peptides are placed in the order of their start value, with the end value breaking any ties. For coloring details, see section 8.1.

7.2) Condensed Peptide Plot

By selecting the "Condensed Peptide" button before hitting run, the program will create a series of sheets, each titled as the state shown, along with "cond". Output shows uptake in percentage as described in section 3. Condensed peptide plots are similar to peptide plots (see section 7.1), but the peptides are shorter, and exchange values are overlayed over the peptides. For coloring details see section 8.1.

7.3 Difference Plot

By selecting the "Peptide Difference" button before hitting run, the program will create a sheet for each difference input. The sheets are titled as the difference name given. Differences are calculated as described in section 4. Difference plots are otherwise very similar to peptide plots (see section 7.1). For coloring details, see section 8.2.

7.4 Condensed Difference Plot

By selecting the "Condensed Difference" button before hitting run, the program will create a sheet for each difference input. Sheets are titled as the given difference name, along with "cond". Differences are calculated as described in section 4. When the theoretical preference is on, differences are in Daltons. When the experimental preference is on, differences are shown in percentage (out of 100).

8. Formatting Options

8.1) Coloring of Plots

Plot color is completely customizable with up to 10 colors for uptake plots and 10 colors for difference plots. Once custom colors are generated, they are saved as a .JSON file on your computer and can be used anytime in the future by selecting that file in the dropdown menu. For more information, click "Create Custom Colors" in the formatting section and then click "See Examples".

8.2) Other Formatting Options

Chiclet Difference Plots by default will have a white gap if a peptide is in one state but not the other. This function can be turned off if desired.

Cell widths can be adjusted as desired. Making the cell width narrower when trying to show errors in the condensed peptide plot may cause spacing issues.

For condensed difference plots, the difference values can be displayed for insignificant values or not. By default, they are.

Additionally, for all condensed peptide plots, error is not added by default, but can be. It is calculated using the uptake standard deviation in the state data file.

9. Uptake Plots

9.1) Corrected vs. Uncorrected and Show Last Timepoint

The default option for plots is to show the raw uncorrected deuterium uptake. This can be switched to the corrected uptake, either based on the percent back exchange for theoretical option preference or the maxD values for experimental preference. Additionally, the last timepoint can be removed from the plot, which may be useful for creating graphs without incorporating the maxD timepoint itself.

9.2) Show Button

For many changes, such as the axes, states, colors, and symbols, changes will not take affect until the graph is created again by clicking the show button (or any other button such as show, next peptide, corrected/uncorrected that re-creates the graph)

9.3) Linewidth and Dashed Line

The linewidth can be altered, as can the line being dashed or solid. This will affect both the graph on the page and the page with all graphs.

9.4) Hex Color and Unicode Symbol

Any hexadecimal color can be added (Formatted as FFFFFF) for the color of the graph. Additionally, the markers can be made to be any Unicode symbol (There are so many of these), formatted as U+XXXX. The size of these can also be scaled.

9.5) Legend Options

The figure legend can be placed in any corner of the graph using the corresponding button. Its size and linewidth can be adjusted. When exported to the full sheet, the legend will be provided separately in the

pdf, and will not be on top of the graphs. The titles of the legend entries can be edited in the “Title” set of boxes.

9.6) Search Function, Next Peptide, and Last Peptide

The set of peptides can be iterated through using the next and last peptide buttons. The set of peptides can also be searched using the “Search Peptides” button and the entry to the left of it. This will bring up the first peptide that contains that residue number.

9.7) Exporting Graph as a PNG

Individual Graphs can be exported as PNGs using the “Save as PNG” button.

9.8) Exporting all to PDF

Graphs can be exported to PDF in a 6x8 or 8x6 grid by selecting the “Uptake Plot” button in the choose scripts box and running the program. The horizontal and vertical button in the uptake plot box will determine whether this output is 6x8 or 8x6. After the program has finished running, there will be an option to save the PDF. The pdf will also contain the legend.

10. Creating and Editing Linear Maps, Export to PyMOL

10.1) Machine Learning Methodology in Brief

Output from difference maps (which is automatically enabled) is processed by replacing all cells where there is no peptide or an absent peptide with 0’s. The A top row of data is then added with a Boolean value for Coverage/No Coverage. The data is then processed into a 3-dimensional matrix, where the second layer is a Boolean value for whether a peptide is present in that location in the top layer. This data, now a 27:residues:2 matrix, is then sliced, for each residue, into a 27:21:2 matrix, with the 10 closest neighboring residues on each side of the residue in question. Any necessary padding is added. Following this, the first residue of every peptide is removed and replaced with a 0. Following this, if there is any area of the matrix where there is no peptide overlap at all, the peptides to the irrelevant side are replaced with padding. This is the format of the data that the model was trained on, and it is the format of input it takes. The model outputs a number 0-5. 2 is significant protection. 1 is questionable protection. 0 is an insignificant difference. 4 is questionable deprotection. 5 is significant deprotection. 3 is no coverage.

10.2) Linear Map Scaling

The model was trained on data where +/- 0.5 Da was deemed the cutoff for significance. The default coloring scheme also has 0.5 Da as the cutoff for significance. If a coloring scheme where a different number is used as the cutoff for significance is used, then data will be scaled first before being fed to the model, such that the cutoff for a significant difference will be read as 0.5 Da. 0.5 Da, however, is recommended.

10.3) Linear Map with maxD Differences

Creating the linear map when the experimental preference is enabled is not ideal, as differences are calculated as difference in corrected RFU, which is a different difference than absolute difference in

Daltons. However, since values are meant to be edited anyways, it should still often be a quite good starting point.

10.4) Linear Map Editor

The output of the machine learning model is not perfect, and depending on the complexity of the difference map, some edits may need to be made. Because of this, an interactive linear map editor has been created, that takes images of the condensed peptide maps and allows for editing of the linear map sequence. Left clicking on the numbers above the colored square representing the linear map allows for the changing of their value to any value desired. Right clicking on another number will then paste the most recently used value to that square. When switching states or timepoints in this interface, it is important to either save, or export to pymol (which also saves). This will both save the items in the excel sheet in a sheet ending in "_predicts" as well as keeping the same values if you return to that state.

10.5) Exporting to PyMOL

When the user clicks on the export to pymol button, they will be prompted to enter a .pdb file for the protein in use. Then, if there is no sequence available for the current difference state, a sequence will be generated based on the peptides in the state. Whichever sequence is used, it will be compared via a pairwise alignment to the sequence extracted from the .pdb file. The alignment is then corrected (as the sequence extracted from the .pdb file does not have the same numbering as are used for coloring specific residues in pymol) and used to color the pymol model of the protein. Colors are based on the 2 most minor protection and deprotection colors, as well as the color for absent peptides/no coverage and insignificant difference. From here, pymol can be used as normal.

11. Saving your File

When the program has finished running, you will be prompted to save the workbook and/or the PDF (for uptake plots).