HDXWizard Operating Instructions

Contents

1.Introduction, Compatibility, and Contact Information

Welcome to HDXWizard, a tool designed with a simple, intuitive user interface the processing of HDX data into chiclet plots, difference plots, and colored peptide plots, peptide difference plots, uptake plots, and localized difference plots. For all inquiries, email Thomas Wales at t.wales@northeastern.edu. The program is available on GitHub at https://github.com/ZacharyACohen/HDXWizard.

2.File Entry

2.1 Data Entry Types

HDXWizard supports the use of DynamX State and Cluster data, as well as tabular datasets exported from HDExaminer and HDXWorkbench

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 otherwise the sequence will not be matched. Lines in the file beginning with ">" will be read until the first space or the first "|" to determine the name of the protein.

3. Selecting a Preference for RFU Calculation or MaxD Correction

3.1.1) No maxD

Selecting the no maxD 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 can be specified by selecting "N-1", "N-2" or the "Englander Rates" options. The first two options set maximum uptake equal to the number of non-proline residues minus 1 or 2 (except for N terminal prolines). The Englander rates option allows for the calculation of the theoretical deuterium remaining in a peptide given it begins maximally deuterated according to various parameters: pH, temperature, time in water, acid, base, and water catalyzed reference rates, and NMR derived constants for how each amino acid side chain affects the exchange rate of the adjacent backbone amides (Nguyen et al. 2018). These values can all be determined by the user and used in place of the N-1 or N-2 options as the denominator to calculate RFU. Additionally for all options, a set percentage of back exchange can be corrected for by multiplying this theoretical maximum by $(1 - \frac{back\ exchange}{100})$.

3.1.2) Differences maps with No maxD Preference

When the no maxD preference is on, if back exchange is left blank or set to 0, all differences will be the raw uptake differences between states in Daltons. If back exchange is corrected for (by typing a number into the box), both corrected RFUs converted to Daltons according to Zhang and Smith, Protein Science 1992 before differences are taken.

3.2.1) MaxD Corrected Preference

If you have data for a maximally deuterated control (maxD) that you would like to be corrected for, then selecting the maxD corrected preference is the correct option. Please make sure that your maxD data is the highest 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, 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)

$$Corrected\ RFU^{*} = \frac{Uptake}{Max\ Theoretical\ Uptake} * \frac{\sum \frac{maxD\ Uptake}{\max theoretical\ uptake}}{number\ of\ maxD\ peptides}$$

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.

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 “maxD corrected”, 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) Differences maps with maxD Preference

Differences with the maxD preference can either be shown as the difference in relative fractional uptake between two states, or a difference in absolute uptake in Daltons. Please see 3.1.1 and 3.1.2 for more information.

4. Adding and Customizing Differences

Differences can be added by clicking the “+Dif” button. Two dropdown lists with states from your file should become available. 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 difference. This title will be the title of any sheets and chiclet plots created for this state.

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) Replicates and Standard Error Calculations for State Data and HDExaminer Data

For state data and HDExaminer data, if there is only one instance of each state/peptide/timepoint combination, then the standard deviation in uptake will be read from file and used as is (if the user elects to show standard error). The user can select the number of replicates used to calculate the standard error from the standard deviation. If there is more than one replicate for any peptide in the file, no standard deviation measurements will be taken directly form the data files. Instead, for instances with 3 or more replicates, the standard deviation will be calculated and used for further calculation and error propagation.

5.3) Replicates and Standard Error Calculations for Cluster Data and HDXWorkbench

For Cluster data and HDXWorkbench data, standard error will only be calculated for instances with 3 replicates (i.e. from 3 different files) of undeuterated treatment and treatment at specific timepoints. For instances with only one charge state, centroids are converted to mass by $mass = (\frac{m}{z} * z) - z$ and then averaged. The standard deviation of the mass of the deuterated and undeuterated species is then used to calculate the uptake standard deviation $std_{uptake} = \sqrt{std_{und}{}^2 + std_{t=x}{}^2}$. For peptides with more than one charge state, the mass calculated from each file is first averaged, and then the standard deviation is calculated from the 3 (or more) completely separate replicate masses and then converted to an uptake standard deviation as above.

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. A second sheet with only peptides shared between all states of a protein is also created. 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 (or can be added as a blank white space by selecting “add gaps if pep in only one state” in the formatting options. Each difference plot will be titled as the difference name inputted for the difference. A second sheet with only peptides shared between all states of a protein is also created. Difference plots are colored according to section 8.2.

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 no maxD preference is on, differences are in Daltons. When the maxD correction preference is on, differences are shown in percentage or Daltons depending on user selection.

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. Custom schemes can be generated by clicking the “create custom colors” button and following instructions there. 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

Add gaps if pep in only one state - 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.

Reorder states – Do you want to change the order chiclet plots are displayed in? Click this and drag states

Enumerate All Residues – This option makes peptide plots come out with every single residue numbered, which can be nice for looking closely at data but is much less readable as a figure for publication

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

Show standard error - 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 divided by the square root of the given n.

Exclude Exchange at Prolines – For localized difference plots, this option will run after the neural network to set all prolines to have an insignificant difference, as there is no exchange at prolines.

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 no maxD option preference or the maxD values for maxD corrected 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. If there are multiple proteins shown, a dropdown will pop up asking which protein to search in.

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 Localized Difference Plots, 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) Localized Difference Plot 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) Localized Difference Plot with MaxD Differences

Creating localized difference plots when the maxD corrected preference is enabled is not ideal if the RFU selection has also been enabled, 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) Localized Difference Plot 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 localized difference plot editor has been created, that takes images of the condensed peptide maps and allows for editing of the localized difference plot sequence. Left clicking on the numbers above the colored square representing the localized difference 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. In addition to

the colors correlating to classifications predicted by the machine learning matrix, the user can select additional colors to use for coloring.

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. If the protein has multiple chains, the program will prompt a yes or no questions for aligning and coloring each chain. A sequence will be automatically generated from the DynamX file and aligned pairwise with 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. From here, pymol can be used as normal.

11. Principal Component Analysis

PCA plots are a new feature to HDXWizard in 1.2.3 and have limited functionality. For each state of a protein, the uptake for each timepoint for each peptide is a value in the PCA. Missing values are imputed as column means and PCA is performed. There is no customizability at this time.

12. Butterfly Difference Plots

Butterfly Difference plots are a new feature to HDXWizard in 1.2.3 and have limited functionality. Differences are shown in butterfly format. Note that figure legends are not necessarily consistent between plots so please double check each figure legend. There is no customizability at this time.

13. Other Tips and Tricks

13.1) Reducing the Number of Peptide Plots to be Made

In the formatting options box, there is a check-button that says: “only show states from uptake plot box”. When this is selected, if peptide plot or condensed peptide plot scripts are selected, only states that have been selected in the dropdown menus of the uptake plot box will be made, allowing to skip over irrelevant states.

14. Saving your File

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