|  |
| --- |
| Application Note  Spaghetti: Visualization of Observed Peptides in Tandem Mass Spectrometry  Co-responding Steven Lewis\*, Terry Farrah, Robert L. Moritz, Eric W. Deutsch and John Boyle  1Institute for Systems Biology, Seattle, WA, 98109, USA  Received on XXXXX; revised on XXXXX; accepted on XXXXX  Associate Editor: XXXXXXX |

[[1]](#footnote-2)\*abstract

**Summary:** We present a visualization and analysis tool, call Spaghetti, for exploring the relationship of peptide fragments collected in proteomic tandem mass spectrometry studies to their inferred parent protein. The tool can be used to overlay fragments on both linear sequence and 3D structures, so that patterns and biases in the data can be explored. Studying patterns of peptide location across a protein can be used for many purposes: exploring PTM (post translational modification) locations with respect to putative protein active sites and protein-protein interaction sites; verifying the reliability of a protein inference by exploring peptide coverage and biases to specific regions; and studying biases in the distribution of detected peptides across protein secondary structure elements, structural domains, and protein chains. Analysis, and coverage statistics are also provided through the tool, and it has been coupled with the PeptideAtlas database to allow the inclusion of as large a collection of data as possible. Spaghetti provides the information about the underlying structure and its relationship to specific fragments as an integrated web based view application.

**Availability and Implementation:** The full source code, documentation and test code is available at <http://code.google.com/p/hydra-proteomics/>. A live demonstration site is available for access at <http://db.systemsbiology.net:8080/Spaghett>i.

# introduction

Tandem Mass spectrometry has proven to be a powerful tool for identifying and quantifying proteins. However there is uncertainty at all levels of the process. Fragmenting an enzymatically cleaved protein into smaller peptides, and detecting the masses of these fragmented and parent ions, can result in multiple matches and numerous unassigned spectra. As the majority of detected spectra do not map to a specific peptide, and a fraction of the amino acids in known proteins are actually detected, it is important to gain an understanding as to why this occurs. The multiple possibilities and poor sampling also means that there is often uncertainty and biases in experiments.

To help overcome these problems we have developed Spaghetti, a visual analytics tool for exploring the positions and overlaps of the peptides that actually map to proteins and to integrate the data with known structure information to gain a better understanding of how this affects detectability. Recently there have been a number of approaches developed to predict the behavior of peptides within the instruments and separation columns; these have generally been analytical methods rather than visualizations (Siepen 2007, Tang 2006, Webb 2008). The advantage of visual analytic approaches is that the human visual system can often efficiently identify numerous subtle phenomena such as over representation of peptides (in specific domains or in particular regions), analysis problems (e.g. resulting from multiple peptides mapping to the same location due to overlapping partial digests), or the identification of biological relevant factors (e.g. where and when PTMs occur) that are difficult to detect algorithmically.

The Spaghetti tool can use the protXML (Keller 2005) file format directly or a Uniprot id and a list or detected fragments and is integrated into the widely used Trans-Proteomics Pipeline software (Keller 2005), so that the identified peptides fragments and inferred protein relationships for a specific experiment can easily be viewed.

# RESULTS

The web application can be initiated directly from the TPP package or by reading in a protXML file. The tool also has a user interface that allows for the input of a series of Uniprot IDs and corresponding lists of detected fragments. Once the data has been read into the application, relevant sequence and structural information is automatically retrieved. The sequence is compared with the 3D model to identify matching segments, and the analyses and view construction is then undertaken.

The fragments are displayed in a sequence view and, if available, a 3D view. By default the most suitable 3D model to display is determined to be that which matches the highest number of peptide fragment sequences for a particular protein. A number of statistics are also generated for each model, and these can be overlaid on the structural view to allow for further analysis (these include hydrophobicity (Lesser 1990), solvent access (Shrake 1973)), as well as information relevant to the fragmentation (e.g. coverage and missed cleavages). As this information is displayed through JMol (Hansen 2010) several additional visualization and selection criteria are available (ribbon display, toggling molecule granularity, coloring options etc).

A sequence based view is also provided, which provides some of the information available in the structural view. The detected fragments are shown as a series of colored rectangles in the sequence view, with the colors corresponding to the coloring of any three dimensional view. Coverage information and feature (e.g. secondary structure, missed cleavages) are also overlaid on the sequence view.

The base web application is accessible at <http://db.systemsbiology.net:8080/Spaghetti/Index.html>, which will show the fragments aligned relative to the protein structure and any detected and useful 3D models. A selected set of proteins and fragments from Peptide \Atlas (Desiere 2006) are also available.

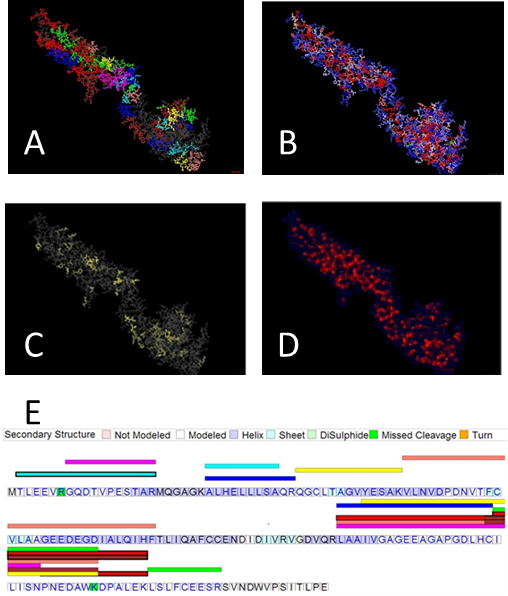


Fig 1. Visualizations used in Spaghetti to show the location of the mapped peptide fragments. Both linear and, if available, 3D representations can be displayed. The linear view (E) is shown on the bottom, while the top four images show different 3D representations (A-D). For the 3D view a variety of options are available to highlight fragments by secondary structure (B), hydrophobicity view coverage view (C) and solvent exposure (D). As the tool uses Jmol, other controls are ribbon structure of proteins and to show or not show specific chains in the molecule. .

# DISCUSSION

Spaghetti is a tool that can be used to integrate information about detected peptides from proteins with existing structural and sequence data. In addition to providing a tool for visualization, the tool can be used to help determine why certain peptides are detected and certain region of protein molecules are found in many detected peptides and others are not.

Proteins are normally denatured prior to digestion for proteomic analysis. This denaturation raises questions as to the relevance of any prior three dimensional structure. Visualization of the relationship between known three dimensional structure and detected fragments may prove useful in seeing where denaturation may be incomplete and the dimensional structure may affect digestion and detected peptides. By integrating several sources of information about protein structure with information about peptides detected by proteomic studies spaghetti allows examination of regions of proteins which have many detected fragments as well as the characteristics of regions where few or no peptides are detected.

acknowledgements

This project is supported by Award Number R01GM087221 from NIGMS, major research instrumentation grant 0923536 (to RM) and we also thank the Luxembourg Centre for Systems Biomedicine and the University of Luxembourg for support. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R. (2006). The Peptide Atlas Project. Nucleic Acids Research, 34, 2006.

Keller A, Eng J, Zhang N, Li XJ. Aebersold R. (2005) "A uniform proteomics MS/MS analysis platform utilizing open XML file formats." Molecular Systems Biology 2:1–8.

Hanson, R (2010). Jmol -- A Paradigm Shift in Crystallographic Visualization. J. Appl. Cryst, 43, 2010.

Lesser G, and Rose G (1990). Hydrophobicity of Amino Acid Subgroups in Proteins. Proteins: Structure, Function, and Bioinformatics 8 (1), 1990.

Shrake A, Rupley JA. (1973). "Environment and exposure to solvent of protein atoms. Lysozyme and insulin". J Mol Biol 79(2), 1973.

Siepen J, Keevil E, Knight D, and Hubbard S. (2007) “Prediction of Missed Cleavage Sites in Tryptic Peptides Aids Protein Identification in Proteomics.” Journal of Proteome Research 6 (1), 2007.

Tang H, Arnold R, Alves P, Xun Z, Clemmer D, Novotny M, Reilly J, and Radivojac P. A Computational Approach Toward Label-free Protein Quantification Using Predicted Peptide Detectability. Bioinformatics 22 (14), 2006.

Webb-Robertson B, Cannon W, Oehmen C, Shah A, Gurumoorthi V, Lipton M, and Waters K. (2008). “A Support Vector Machine Model for the Prediction of Proteotypic Peptides for Accurate Mass and Time Proteomics.” Bioinformatics 24 (13), 2008.

1. \*To whom correspondence should be addressed. [↑](#footnote-ref-2)