## TITLE

# DIMPL – DIsulfide Mapping PLanner software tool

## AUTHORS

Andreas M. Kist[1,2], Angelika Lampert[1,3], Andrias O. O'Reilly[1,4]*

1 Institute of Physiology and Pathophysiology Friedrich-Alexander Universität Erlangen-Nürnberg, Universitaetsstrasse 17, 91054 Erlangen, Germany

2 Max-Planck-Institute of Neurobiology, Am Klopferspitz 18, 82152 Martinsried, Germany

3 Institute of Physiology, RWTH Aachen University, Pauwelsstrasse 30, 52074 Aachen, Germany

4 School of Natural Sciences and Psychology, Liverpool John Moores University, Liverpool, UK

* Corresponding author

## KEY WORDS

Disulfide mapping, disulfide bridge, mass spectrometry, protein structure, web-tool

## AUTHOR CONTACT INFORMATION

**Andreas M. Kist**

Max-Planck-Institute of Neurobiology

Am Klopferspitz 18

82152 Martinsried

Germany

Email: anki@neuro.mpg.de

Tel: +49 (89) 8578 3488

Fax: +49 (89) 8578 3541

**Angelika Lampert**

Institute of Physiology

RWTH Aachen University

Pauwelsstr. 30

52074 Aachen

Germany

Email: alampert@ukaachen.de

Tel: +49 (241) 8088 810

Fax: +49 (241) 8082 434

**Andrias O. O'Reilly***

School of Natural Sciences and Psychology

Liverpool John Moores University

Liverpool

L3 3AF

UK

Email: a.o.oreilly@ljmu.ac.uk

Tel: +44 (0) 151 231 2330

Fax: +44 (0)151 231 2338

## ABSTRACT

Disulfide bridges are side-chain mediated covalent bonds between cysteines that stabilize many protein structures. Disulfide mapping experiments to resolve these linkages typically involve proteolytic cleavage of the protein of interest followed by mass spectroscopy to identify fragments corresponding to linked peptides. Here we report the sequence-based “DIMPL” web-tool to facilitate the planning and analysis steps of experimental mapping studies. The software tests permutations of user-selected proteases to determine an optimal peptic digest that produces cleavage between cysteine residues, thus separating each to an individual peptide fragment. The webserver returns fragment sequence and mass data that can be dynamically ordered to enable straightforward comparative analysis with mass spectroscopy results, facilitating dipeptide identification.

## INTRODUCTION

A disulfide bridge formed between the thiol groups of two closely-apposed cysteine residues is the most common covalent bond found in proteins after the peptide bond. The formation of the correct disulfide reticulation can be crucial for establishing the functional tertiary fold of a protein and moreover can contribute to its thermostability and proteolytic resistance (Fass, 2012). Identifying cysteines involved in disulfide bridges and determining their interconnectivity is an important step in biochemical characterization that can aid in classifying the protein fold, identifying family homologues and facilitating structure determination by indicating spatial restraints (Mouhat et al., 2004). Disulfide mapping can also reveal misfolding that results from the loss of native reticulation and therefore has application in molecular epidemiological studies (Gilchrist et al., 2013; Mossuto, 2013). Distinguishing different disulfide-mediated folds is also relevant for assessing recombinant protein expression studies, as the correct structural isomer in a misfolded ensemble can be identified by disulfide mapping of chromatographic fractions (Berkmen, 2012).

The experimental steps of a disulfide mapping study using the ‘bottom-up’ approach (Chait, 2006) involve protein digestion using proteolytic enzymes or chemical reagents, followed by mass determination of the fragments (Gorman et al., 2002; Tsai et al., 2013). Disulfide-linked peptides can be identified using the profile comparison approach, where differences in the mass spectra of reduced and unreduced aliquots of the protein digest corresponded to linked peptides. If disulfide reduction using reagents such as dithiothreitol (DTT) is inefficient or leads to sample loss due to the aggregation of unfolded protein, it is still possible to identify disulfide-linked peptides by comparing the mass spectrum of the unreduced peptic digest with a theoretical complete digest of the fully-reduced protein calculated using the same protease set.

A key decision in the planning of disulfide mapping studies is the choice of proteolytic enzymes or chemical reagents and the order of their application. An effective digestion for mapping studies produces cleavage of the substrate between each cysteine residue as this eliminates the possibility of fragments containing intra-peptide disulfides. However, testing permutations of proteases to achieve optimal cleavage can be arduous, particularly when there is a large number of cysteines in the target protein sequence. Here we report the DIsulfide Mapping PLanner (DIMPL) webtool for *in silico* protein digestion, which determines the minimal set of user-selected proteases that will produce segregation of cysteines in a user-provided protein sequence. In addition to reporting the optimal protease set, the software returns the protein fragments and the corresponding masses and provides a range of options to order and analyze this data.

## METHODS

The core of DIMPL was developed using the Python programming language (v.2.7) and the software is hosted on a Linux platform. The web interface (http://www.lampert-lab.com/dimpl) is written in server- and user-sided programming languages including PHP, MySQL, HTML5, CSS and the Javascript based jQuery framework.

### Input

The amino acid sequence of the protein of interest in FASTA format or as raw text is entered via the web interface. Alternatively, a UniProt accession number can be entered for dynamic sequence retrieval from the UniProt server following web-form submission. Next, the endo-peptidase enzymes (or chemical reagents: cyanogen bromide, iodosobenzoic acid, hydroxylamine) are selected for the digestion. For convenience, the proteases are grouped into

‘Common’ (e.g. trypsin, chymotrypsin and pepsin), ‘Less Common’ and ‘Uncommon’ sets, which can be selected via a drop-down menu. Each protease can also be selected or deselected individually.

To further extend the number of digestion options available, custom proteases can be created with a user-provided substrate specificity. The custom digestion pattern consists of four amino acid positions before and two after the cutting site and the user can provide amino acids in single letter code that are to be matched or avoided at each position. Unset amino acids positions are kept as non-specific.

Given that digestion reactions are not always 100% efficient, the user is provided with the option (via a drop-down menu) to include up to three ‘miscleavages’ in there results. These are intact peptides that contain an uncleaved substrate site.

A final input option is a text box for an email address to allow the user to receive results via email.

**Enzyme scoring scheme**

Following web-form submission, each of the selected proteases is tested for its ability to digest the input protein sequence using the specificity rules for endo-peptidase activity listed in the EXPASY PeptideCutter library (Gasteiger et al., 2005). Proteases unable to cleave the sequence at least once are rejected. The fragmentation patterns formed by the remaining proteases are then analyzed and scored (S) according to the following function:

$$S = \frac{C^2}{F}$$

where ‘F’ represents the total number of fragments produced and ‘C’ is the number of separated cysteines. The ‘C’ term is squared in order to increase the score of proteases that separate a greater number of cysteines. Similarly, when different digestions result in an equal number of separated cysteines, the proteases that produce fewer fragments are scored higher.

If individual proteases are unable to produce full cysteine segregation then the activity of multiple enzymes is assessed. However, determining the fragmentation pattern resulting from the simultaneous action of a combination of proteases is problematic as the activity of one protease may eliminate substrate sites for a second protease and vice versa. Therefore DIMPL adopts the more experimentally-relevant approach of digesting the protein substrate sequentially with proteases to produce a series of nested fragments. To determine whether full cysteine segregation is achievable, the software performs iterative testing of substrate digestion with protease permutations beginning with two enzymes and including additional proteases up to a maximum of five. If a satisfactory solution cannot be found with user-selected proteases, the substrate is then tested with other proteases in order to suggest an alternative digestion strategy to the user.

### Output

The results page from the DIMPL server is comprised of two sections. The first section gives the list of proteases and the chronological order for their application in order to produce the requisite inter-cysteine cleavage (or when full cysteine separation is not possible, the selection of proteases that gives the submaximal cysteine segregation).

The second section details the fragment data produced by the *in silico* digestion. Fragment sequences are listed with corresponding masses, the order with which they occur in the protein primary sequence and whether a cysteine is present in the fragment; the user can dynamically sort these results by their preferred field. The selection (by double-clicking) of two fragments produces the calculated sum of their masses.

Ambiguous digestion sites in the sequence are designated with multiple arrows “>>”. These sites occur when two or more cleavage sites are present concurrently in the primary sequence. For example, trypsin cleaves after lysine (K) and will generate fragments of varying lengths when multiple consecutive ‘K’s are encountered. To account for these ambiguous sites, a feature of the results section is that the mass for any sequence of contiguous amino acids can be calculated by highlighting the corresponding region in the full-length protein sequence. Indeed, this allows the mass for a fragment of any length to be calculated, thus further enabling comparative analysis with mass spectroscopy results.

The results are available for download in XML or CSV files and are also posted by email if requested via the input page. A list with all possible fragments bound together by a disulfide bridge is available for download as CSV file. In addition, the results page is accessible for 30 days via a uniquely-generated URL.

## DISCUSSION

The DIMPL webtool provides an *in silico* digest of a protein sequence using user-selected proteases (and chemical cleavage reagents) and reports the order of digestion steps that produces the maximum number of cysteines separated onto the fewest number of peptide fragments. The molecular mass, sequence and the order that each fragment occurs in the submitted protein are also reported.

A great variety of software tools exist for both the prediction and *in silico* study of disulfide connectivity in proteins (Craig and Dombkowski, 2013; O'Connor and Yeates, 2004; Yachdav et al., 2014). For example, the sequence-based software tools DISULFIND (Ceroni et al., 2006) and DiANNA (Ferrè and Clote, 2005) can accurately predict the disulfide connectivity of a correctly-folded protein. DIMPL provides a complementary software tool that can support experimental efforts to validate these predictions. The aim of DIMPL is to provide researchers with an optimized workflow for disulfide mapping that is both time effective – involving the fewest number of digestion steps – and resource efficient in terms of protein sample and available digestion enzymes and reagents.

DIMPL is targeted to researchers adopting the bottom-up mass spectroscopy approach for protein characterization (Chait, 2006), which involves an initial sample digestion step. In general, trypsin is the sole protease used for digestion prior to fractionation of peptides by liquid chromatography and application to a mass spectrometer (LC-MS). Unfortunately, the slightly alkaline conditions required for optimal trypsin activity can produce the phenomenon of disulfide shuffling (Ryle and Sanger, 1955), which can confound the mapping process. Other proteases (e.g. pepsin) can be used to circumvent this problem and so DIMPL provides the user with a wide range of digestion options, including the ability for users to define their own unique protease specificity. This feature is advantageous as it also provides a measure of future proofing for the software.

Depending on the mass spectroscopy technique and ionization source, a multitude of different ions can be generated per peptide. Prior to a mass spectroscopy run, software such as Skyline (MacLean et al., 2010) can be used to calculate all the masses of a complex ionic mixture and plot a predicted mass spectrum. In comparison, DIMPL software adopts the comparatively straightforward approach of reporting just the mass of each full peptide fragment with charged N- and C-terminii. This is consistent with the focus of DIMPL as an experimental planning software aid, as it aims to inform the user of the mass range of their digestion products, thus allowing them to adjust their digestion strategy so that a suitable fragment range for their mass spectrometer can be generated. Finally, for researchers using SDS-PAGE or other techniques instead of mass spectroscopy to size their proteolytic fragments, the results returned by DIMPL should prove similarly useful.

DIMPL is a user-friendly webtool that facilitates the planning and data analysis steps of disulfide mapping studies. The use of DIMPL can save time and labor and can make efficient use of protein samples as it can determine the minimum number of stepwise proteolytic reactions required for inter-cysteine cleavage. Its application can therefore support the many diverse experimental studies that assess the contribution of disulfide bridges to protein structural integrity.

## ACKNOWLEDGEMENTS

This research was supported by a MINT Excellence stipend of the Manfred Lautenschläger Stiftung to AMK, a German Research Foundation grant to AL (DFG LA2740/2-1) and a European Commission Marie Curie Research Fellowship to AOR (FP7-PEOPLE-2010-IEF, No. 275768).

## AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

## REFERENCES

Berkmen, M. 2012. Production of disulfide-bonded proteins in Escherichia coli. *Protein Expr. Purif.* 82, 240–251.

Ceroni, A., Passerini, A., Vullo, A., et al. 2006. DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. *Nucleic Acids Res*. 34, W177–W181.

Chait, B.T. 2006. Mass Spectrometry: Bottom-Up or Top-Down? *Science* 314, 65–66.

Craig, D.B., and Dombkowski, A.A. 2013. Disulfide by Design 2.0: a web-based tool for disulfide engineering in proteins. *BMC Bioinformatics* 14, 346.

Fass, D. 2012. Disulfide Bonding in Protein Biophysics. *Annu. Rev. Biophys*. 41, 63–79.

Ferrè, F., and Clote, P. 2005. DiANNA: a web server for disulfide connectivity prediction. *Nucleic Acids Res*. 33, W230–W232.

Gasteiger, E., Hoogland, C., Gattiker, A., et al. 2005. Protein Identification and Analysis Tools on the ExPASy Server. *In* The Proteomics Protocols Handbook, J.M. Walker, ed. (Humana Press), pp. 571–607.

Gilchrist, J., Das, S., Petegem, F.V., et al. 2013. Crystallographic insights into sodium-channel modulation by the β4 subunit. *Proc. Natl. Acad. Sci.* 201314557.

Gorman, J.J., Wallis, T.P., and Pitt, J.J. 2002. Protein disulfide bond determination by mass spectrometry. *Mass Spectrom. Rev.* 21, 183–216.

MacLean, B., Tomazela, D.M., Shulman, N., et al. 2010. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinforma. Oxf. Engl.* 26, 966–968.

Mossuto, M.F. 2013. Disulfide Bonding in Neurodegenerative Misfolding Diseases. *Int. J. Cell Biol.* 2013, e318319.

Mouhat, S., Jouirou, B., Mosbah, A., et al. 2004. Diversity of folds in animal toxins acting on ion channels. *Biochem. J.* 378, 717.

O’Connor, B.D., and Yeates, T.O. 2004. GDAP: a web tool for genome-wide protein disulfide bond prediction. *Nucleic Acids Res.* 32, W360–W364.

Ryle, A.P., and Sanger, F. 1955. Disulphide interchange reactions. *Biochem. J.* 60, 535–540.

Tsai, P.L., Chen, S.-F., and Huang, S.Y. 2013. Mass spectrometry-based strategies for protein disulfide bond identification. *Rev. Anal. Chem.* 32.

Yachdav, G., Kloppmann, E., Kajan, L., et al. 2014. PredictProtein--an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res*. 42, W337-343.