

# Search-space Reduction of a Non-redundant Peptide Database

Ian Shadforth  
Cranfield University  
i.p.shadforth.s01@  
cranfield.ac.uk

Daniel Crowther  
GlaxoSmithKline  
daniel.j.crowther@  
gsk.com

Conrad Bessant  
Cranfield University  
c.bessant@  
cranfield.ac.uk

## Abstract

*Peptide mass fingerprinting and database searching with tandem mass spectrometry are two methods commonly employed to identify proteins in a sample. However, up to 90% of peptides can remain unidentified. In this paper, a search-space filter using amino acids identified by a novel de novo methodology is presented. This provides a high-accuracy set of amino acid predictions through exploiting the internal fragmentation of amino acid chains during tandem mass spectrometry. These predictions are used to reduce the number of peptides considered from a non-redundant peptide database. The presence of one confirmed amino acid can be used to reduce the search-database size by between 33% (Leucine) and 83% (Tryptophan). One or more accurate amino acid identifications are made in 18% of simulated and 9% of experimental peptide spectra considered. Given the large proportion of currently unidentified peptides, this method represents a useful tool for increasing peptide identification rates.*

## 1. Introduction

There are currently a number of systems available that allow fast and accurate identification of proteins using mass spectra, such as Mascot [1] and Sequest [2] amongst others. Often, database searching with MS/MS spectra is used to identify proteins in a sample, but whilst this provides a good degree of confidence in the protein assignment, many peptides remain unidentified. Reasons for this include noise in the spectra, contamination, unexpected post-translational modifications (PTMs), unexpected splice variants and the protein not occurring in the search-database [3].

*De novo* sequencing allows the generation of a proposed sequence from a tandem MS spectrum

without reference to a search database. However, the same factors that cause proteomic database searching with tools such as Mascot to fail will often also prevent full *de novo* sequencing of that peptide.

The system presented here uses *de novo* algorithms developed in-house to identify isolated amino acids within a peptide from the tandem MS spectrum. These amino acids are used alongside protein grouping as filters on a non-redundant peptide database to reduce the size of the search-space to be considered. The reduced search-space can then be queried iteratively using PTMs to find a match.

## 2. Materials and Methods

### 2.1 Data

A simulated dataset of tandem MS spectra was created based on the following two proteins (Swiss-Prot accession numbers): O75346 and P13569. These proteins consist of 39 and 168 tryptic peptides respectively.

The simulation algorithm was configured to phosphorylate Serine and/or sulphate Tyrosine 50% of the time.

The two simulated tandem MS spectra were combined to form a notional protein mixture consisting of the first 20 peptides from each protein spectrum.

A non-redundant database containing the peptides resulting from a tryptic digest of the IPI-Human-FASTA protein database (downloaded 05/03/2003 from <http://www.ebi.ac.uk/IPI/IPIhelp.html>) was created. The tryptic digest was performed using the publicly available *Proteogest* program [4]. The subsequent output was entered into a non-redundant relational database using software developed in-house.

## 2.2 Process

The combined spectrum was first submitted to the web-hosted version of Mascot (to be found at <http://www.matrixscience.com>). The search options used were: Database, Swiss-Prot; Taxonomy, *Homo sapiens*; Enzyme, Trypsin; no missed cleavages; no fixed or variable modifications; no protein mass entered; peptide tolerance and MS/MS tolerance left at the default values of 2.0 Da and 0.8 Da respectively; Data file type, Micromass (.PKL); Monoisotopic masses.

All peptides were also submitted to the in-house *de novo* amino acid identification system. Amino acid identifications from this phase were used to search the local non-redundant peptide database. The resulting sub-group of valid peptides can then be iteratively searched for potential PTMs by comparing the mass differences between the calculated and experimental peptide masses.

The grouped database is currently queried using SQL statements, with the process being automated as part of a high-throughput pipeline for peptide identification.

## 3. Results

Querying the non-redundant peptide database can reveal new information about the proteome, such as the frequencies of amino acids in peptides shown in Table 1.

**Table 1. Amino acid prevalence**

AA	%	AA	%	AA	%	AA	%
L	67	Q	45	R	52	H	30
S	60	D	44	V	51	Y	28
A	55	I	41	P	50	C	26
E	54	F	37	K	49	M	25
G	53	N	36	T	48	W	17

The proportion of peptides (%) in the search database containing each amino acid (AA).

The Mascot search against the simulated dataset correctly identified seven peptides from O75346 and four from P13569, ten peptides were assigned to other proteins and the remaining 19 were unassigned. Both proteins were correctly identified with high confidence scores.

Five peptides, three from O75346 and two from P13569, had amino acid predictions made against them, all of which were true. Of these peptides, three were correctly identified by Mascot, leaving one unassigned peptide from each protein for which amino acid predictions were available: a Proline in a peptide

from O75346, possessed by seven out of the 39 peptides in this protein, and both Asparagine and Aspartic Acid from P13569, a combination present in just 16 out of 168 peptides. The search-space for PTMs in these proteins is therefore reduced by between 80% and 90% in these cases, thus allowing for efficient, iterative, processing.

## 4. Discussion

Although database searching using fixed and variable modifications is possible in Mascot, and in other database searching tools, increasing the number of modifications to be considered drastically increases search times, thus preventing their direct use in high-throughput systems.

Once automated, this method will provide the ability to identify a significant number of peptides beyond those provided by Mascot. This will allow for increased annotation of genomes in terms of experimentally proven PTMs and, potentially, polymorphisms. Further development of the amino acid assignment algorithms is ongoing, with the aim being to increase the identification rates in experimental data. Sufficient predictions will enable protein grouping without the need to start from prior Mascot identifications.

Analysis of the non-redundant peptide database could pave the way to a statistically significant scoring method to quantify the increased confidence gained through grouping peptides by protein.

## 5. References

- [1] Pappin, D.J.C., Hojrup, P., Bleasby, A.J., *Bioinformatics Applications UK HGMP Resource Centre* 2002, 1-9.
- [2] Eng, J.K., McCormack, A.L., Yates, I.I.I., *Journal of the American Society for Mass Spectrometry* 1994, 5, 976-989.
- [3] P.Dainese and P.James, in: P.James(Eds.), *Proteome Research: Mass Spectrometry*, Springer, 2001, pp. 103-124.
- [4] Cagney, G., Amiri, S., Premawaradena, T., Lindo, M. *et al.*, *Proteome Science* 2003, 1, 1-15.

## 6. Acknowledgements

We are grateful to GlaxoSmithKline for conjointly funding this research alongside the EPSRC as part of the Engineering Doctorate (EngD) programme.