Data and text mining

Peptide Finder: mapping measured molecular masses to peptides and proteins

Anastasia Alexandridou¹,², George Th. Tsangaris¹, Konstantinos Vougás¹, Konstantina Nikita² and George Spyrou¹,*

¹Biomedical Research Foundation, Academy of Athens, 4 Soranou Ephesiou, 115 27 Athens and
²School of Electrical and Computer Engineering, National Technical University of Athens, 9 Irio Polytechniou Str., 15780, Zografos, Athens, Greece

Received on April 18, 2008; revised on July 16, 2008; accepted on August 1, 2008

Associate Editor: Burkhard Rost

ABSTRACT

Summary: The identification of unknown amino acid sequences of peptides as well as protein identification is of great significance in proteomics. Here, we present a publicly available web application that facilitates a high resolution mapping of measured molecular masses to peptides and proteins, irrespectively of the enzyme/digestion method used. Furthermore, multi-filtering may be applied in terms of measured mass tolerance, molecular mass and isoelectric point range as well as pattern matching to refine the results. This approach serves complementary to the existing solutions for protein identification and gives insights in novel peptides discovery and protein identification at the cases where the identification scores from the other approaches may be below significance threshold. Peptide Finder has been proven useful in proteomics procedures with experimental data from MALDI-TOF.

Availability: Peptide Finder web-application is available at http://bioserver-1.bioacademy.gr/Bioserver/PeptideFinder/.

Contact: gspyrou@bioacademy.gr

In proteomic techniques followed by mass spectrometry, isolated proteins or protein mixtures are digested by enzymes, forming peptides that are subsequently ionized by different techniques (TOF, ESI) and detected. The detected ions with further processing produce mass spectra (MS spectra) (Liebler, 2002) where the mass peaks are noted, selected and served as an input to computer applications that search in large biological databases. Each experimental peptide mass is compared with the mass of the theoretical peptide produced by the digestion of the protein from a selected database (Marcotte, 2007). The occurrence of a match depends mainly on input parameters, scoring functions and associated thresholds. Some applications that perform in silico protein identification with similar requested input are Mascot (Perkins et al., 1999), X!Tandem (Craig and Beavis, 2004) and pFind (Li et al., 2005) that produce comparable results.

We developed a web service that facilitates mapping of any molecular mass measured through an MS procedure to the corresponding peptides and finally to the proteins that include the peptides found. The search is done through completely digested proteomic sets from protein databases (e.g. Swiss-Prot, Trembl) corresponding to the species included in the database of this tool and thus it is not necessary to define specific enzyme/digestion method. It can count for any type of digestion (enzyme driven, random, etc.). Furthermore, through a proper web interface the user may upload complete peak lists with measured molecular masses and ask for mapping them to the proteins of the database. Additionally, various filters are available in order to let the user have a more refined list of peptides and proteins with the requested molecular masses.

The system consists of three parts, namely a database, a file repository and the web interface. There are also a protein sequence processing procedure that produces and registers data in both the database and the repository and a reporting procedure based on dynamically generated html files. For example, suppose we have an ACEM fragment. All the possible combinations AC, ACE, ACES, CEM and EM are derived and their molecular mass is calculated. This procedure is applied for the whole protein and finally for all proteins in the selected database producing a list of fragments with known sequences having a particular molecular mass. In our calculations, we have used molecular mass precision of 0.01 Da. The system registers peptide fragments with molecular masses up to 10 kDa. It is common practice for a Protein Mass Fingerprint (PMF) experiment to measure peptide fragments with molecular masses between 900 Da and 3 kDa. Thus, 0.01 Da precision corresponds to 11.11 p.p.m. for the lowest considered peptide fragment and 3.3 p.p.m. for the maximum one. According to recent publications (Chen et al., 2006; Stead et al., 2006; Taylor et al., 2007) the accepted error tolerance for a PMF experiment ranges from 25 to 50 p.p.m. On the other hand, the tool provides the user with the potentiality to use more rough estimations utilizing the “estimated error” option. This precision value makes the application computationally demanding since it requires a plethora of data to be processed and analyzed. Nevertheless, according to our calculations, the number of human peptides with molecular masses from 900 Da to 3000 Da follows a Gaussian like distribution with a mean of 1950 Da and a standard deviation of ~170 Da. For example, for the current compilation of the database, in the range [2000.0, 2000.1] Da there are 38731 peptides instead of 4228 that are found in the [2000.00, 2000.01] Da range. On the contrary, in the range [1000.0, 0.01 Da.

*To whom correspondence should be addressed.
A.Alexandridou et al.

Fig. 1. Snapshots from Peptide Finder use to map either a single molecular mass (left part) or a list of molecular masses (right part of the image).

1000.3) Da there are only 37 peptides instead of ‘nothing found’ in the range [1000.00, 1000.01) Da or even in the range of [1000.0, 1000.1) Da.

Thus, with 0.01 precision the potential users would have a smaller set of candidate peptides for molecular masses around 2000 Da. On the other hand, if they handle molecular masses near 1000 Da or 3000 Da they should use the ‘estimated error’ option to collect a more representative set of candidate peptides.

The database called OREA (mOleculaR wEight of peptide frAgments) and developed in MySQL platform, contains all the calculated molecular masses and their corresponding frequencies in the proteins sequence database (for Human, Swiss-Prot 55.4 release, there exist 879 689 molecular mass entries, with frequencies from 1 to 338 807). In order to avoid creating and handling a huge database, the mined information is distributed among OREA database and a repository archive of files. The repository archive is separated in sets of text files according to the species. Each file corresponds to one molecular mass and contains information about the matching peptide sequences. In the web application (PHP language) there are three modes of searching available: (i) search for all existing peptides having a particular molecular mass, (ii) search for the existing peptides in a range of molecular masses and (iii) search for existing proteins that contain peptides corresponding to all or part of a set of molecular masses (peak list) either from a PMF or from an MS/MS experiment.

When a user is interested in finding all possible peptide sequences for a particular molecular mass (Fig. 1, left part), the user enters the species and the molecular mass. Other parameters that can be used for searching is the estimated error of the measurement, and an amino acid sequence (or a regular expression) for pattern matching purposes. The system at its current version does not handle post translational modifications.

The use of the estimated error gives the system the flexibility to count for any possible shift in the molecular mass value due to experimental errors. The amino acid sequence is used for pattern matching in peptide sequences. Especially the expressions dealing with the start or end of the peptide sequences is of great importance since they may simulate the specificity of the digesting enzyme applying at one end of the peptide but not the other in spectrometry analysis. Thus, although the molecular mass mapping via the system is enzyme-independent, it can be easily converted to enzyme dependent when it is needed. A range of protein molecular mass and isoelectric point are also available as search parameters and work as a filter in the searching procedure. Upon user request, there is a dynamically driven reporting procedure (using CGI PERL scripts), providing in a dynamic way the user with the following type of information: (i) list of peptides per protein having the requested molecular mass, (ii) list of proteins containing at least one fragment with the specified molecular mass and (iii) combination view of the previous two lists. When a user has a set of experimental molecular
The proteins are sorted and displayed according to the number of molecular masses he/she may use the Protein Identification form (Fig. 1, right part). A threshold is applied to the number of proteins to be displayed. The proteins are sorted and displayed according to the number of their matches, accompanied with other protein-specific information. Peptide Finder uses also a heuristic scoring algorithm, based on the statistics of the molecular masses distribution. It suggests that a good identification should count on the number of molecular masses matched, the frequency they have inside the Swiss-Prot database for each species (i.e. their randomness index), the molecular mass (indicating the size) of the suggested protein. The suggested score calculation formula is:

$$\text{score} = \frac{\sum_{i=1}^{N} \text{mm}_i/f_i}{\text{MM}}$$

where, $N$ is the number of molecular mass matched ($\text{mm}_i$) with frequency $f_i$ inside the Swiss-Prot database and MM is the molecular mass of the suggested protein.

Peptide Finder is hosted at an Apache server on a Linux platform and it incorporates a script-based curation protocol of the database and the file archive. It is part of a newly established group of tools, data bases and web services developed in the Biomedical Research Foundation, Academy of Athens, called ‘BioServer’. Initially, Peptide Finder had been designed to serve human proteomics studies. However, we have started to build the tables of the database as well as the flat file repositories for other species (e.g. Tetrahymena thermophila). In the near future we plan to include data for Mouse and Rat. Also, any user should ask to include the species he/she is interested in.

The tool has been proven useful to our Proteomics research activities since it managed to give us insights of molecular masses mapping on peptides and proteins especially where there was uncertainty from the standard protein identification programs used (A.Xanthopoulou et al., 2008, personal communication). It has been mainly developed to map measured molecular masses to peptides and proteins, using a method that is less dependent on enzymes. As far as the mapping is concerned, since the described method is fully deterministic, Peptide Finder finds all possible peptides with the requested molecular masses and subsequently the corresponding proteins that contain them. We do not claim that this tool is suitable to replace other well-working protein identification software, commercial or not. However, we believe it is a tool that will help researchers to have a complete view for the mapping of the measured molecular masses, giving them insights in cases where the other software do not return any results (cases below threshold), as well as in cases where they need to have all the possible peptide and protein candidates in an enzyme (protease) independent manner. Furthermore, we believe it will be useful in clear peptidomics studies. Peptide Finder could potentially be implicated to the analysis of proteomes derived from genomes which, de facto, include many hypothetical proteins (Maillet et al., 2007). Additionally, because of the reason that OREA database is constructed based on peptide molecular masses, Peptide Finder is extremely useful for the identification of sequenced peptides of known molecular masses by MS/MS and subsequently the identification of the corresponding proteins. Among our future plans is the inclusion of proteomes from more organisms, the development of a more thorough scoring algorithm in order to provide ranked lists of matches and finally the adaptation of the whole system to operate under a distributed computing environment.

**Conflict of Interest**: none declared.

**REFERENCES**


