UniMaP: finding unique mass and peptide signatures in the human proteome

Anastasia Alexandridou1,2, George Th. Tsangaris1, Konstantinos Vougas1, Konstantina Nikita2 and George Spyrou1,*

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

ABSTRACT

Summary: The uniqueness of a measured molecular mass or peptide sequence plays a very important role in the fields of protein identification and peptide/protein-biomarker investigation. We present a publicly available web application that offers information concerning the uniqueness of one or more molecular masses and one or more peptide sequences in the human proteome. When a sequence is found to be unique in humans, the application is able to search across all species querying whether this sequence is unique, not only in humans but also in other species found in the Swiss-Prot Database. The application is also able to search for unique protein fragments derived computationally from enzymatic digestion driven by certain enzymes. Furthermore, the application can list all the unique masses and peptides of a given protein. Through this application, researchers are able to find unique tags, either on a molecular mass level or on a sequence level. These unique tags are remarkably important in research related to protein identification or biomarker discovery and measurements.

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

Contact: gspyrou@bioacademy.gr

Supplementary information: Supplementary data are available at Bioinformatics online.

The precise identification of peptide sequences and proteins is of great significance in proteomics. Specifically, in mass spectrometry, isolated proteins or protein mixtures are digested by enzymes into peptides. These peptides are then ionized by different techniques (TOF, ESI) and detected producing mass spectra that are computationally analyzed. The peptides identified with statistical significance are used for protein identification (Liebler, 2002; Marcotte, 2007). In proteomics technology of selected reaction monitoring (SRM) or multiple reaction monitoring (MRM), a predefined precursor ion and one of its fragments are selected by a filtering device and monitored over time for precise quantification (Lange et al., 2008).

The determination of unique characteristics for a protein is profoundly interesting in the application of the new concepts of emerging targeted proteomics workflows. Peptide Atlas is a major resource for target selection (Deutsch et al., 2008), since it facilitates the researcher to identify suitable proteotypic peptides to target and estimate approximate retention time for the target peptides. Peptide Atlas contains experimental data and it does not offer direct information regarding the ‘uniqueness’ of peptide sequences or masses. Furthermore, there are tools that have been reported to perform some unique peptide (UP) sequence calculations, namely ReMUS which identifies UP segments as epitopes (Pai et al., 2006) and UPF which calculates UP sequences for identification of highly homologous proteins (Kohl et al., 2008). However, these tools do not investigate the uniqueness of a given peptide or a group of given peptides, and they do not find unique molecular masses (UMMs). Furthermore, they do not find the shortest UPs (i.e. the key reference sequences for each protein) and they do not check peptide uniqueness across all species of the Swiss-Prot. Moreover, ReMUS does not find UPs derived from a set of enzymatic digestions whereas UPF is not a web service and thus it requires from the user to download and install UPF and requires also access to an adequate database management system.

We developed a web service that facilitates the investigation of the uniqueness of one or more molecular masses and one or more peptide sequences in the human proteome, covering all the above aspects.

Before the description of the methodology, we schematically give the definitions and relations concerning the data presented in the application (Fig. 1). Core UPs (CUPs) are the shortest peptides that are unique (i.e. a peptide exists only in one protein) in the set of human proteins registered in the Swiss-Prot. Any peptide sequence that contains a CUP is a UP as well. Some of the UPs can be derived from enzymatic digestion (UPed). Moreover, there are UMMs which individually can be found in only one protein. UniMaP application finds and organizes the CUPs and UMMs which act as key reference sequences and masses currently identified with known proteins, and thus they characterize and identify the proteins.

We have created a file repository containing indexed information that relates molecular masses with an accuracy of 0.01 Da to the corresponding peptides existing in human proteins that have been deposited in a completely digested protein database (Swiss-Prot) providing independence from any specific enzyme/digestion method (Alexandridou et al., 2008). From this repository, we have mined...
Therefore, for the simplicity sake and without losing significance given in the Supplementary Material Figure 1. The user is able to searching available as shown schematically in the information flow CUPs in human is 7,338,522. 20,163 proteins contain unique tryptic peptides. The total number of found that 1020 proteins have UMMs, while 20,282 out of 20,333 mined information is distributed among the database and a set of flat one protein (7783 UMMs), along with their corresponding molecular calculated molecular masses which individually can be found in only current version does not handle post-translational modifications. we took peptide fragments of up to 10 kDa resulting in peptide peptide fragments with molecular masses between 900 Da to 3 kDa.

It is common practice for a proteomics experiment to measure second small peak between 27 and 40 amino acids. In addition, are mainly distributed between 4 and 10 amino acids with a are very useful to identify the initial proteins. knowledge of the protein digestion, enzyme-independent peptides are useful in studies concerning Natural Occurring Peptides as well as in other peptidomics investigations. In cases where there is no prior knowledge of the protein digestion, enzyme-independent peptides are very useful to identify the initial proteins.

The system consists of three parts, namely a database, a set of flat files and the web interface. There is also a protein sequence processing procedure that updates the database and the repository as well as a reporting procedure based on dynamically generated html files. Our analysis has revealed that the lengths of CUPS are mainly distributed between 4 and 10 amino acids with a second small peak between 27 and 40 amino acids. In addition, it is common practice for a proteomics experiment to measure peptide fragments with molecular masses between 900 Da to 3 kDa. Therefore, for the simplicity sake and without losing significance we took peptide fragments of up to 10kDa resulting in peptide sequences of approximately 100 amino acids. The system at its current version does not handle post-translational modifications.

The database is developed in MySQL platform and contains all the calculated molecular masses which individually can be found in only one protein (7783 UMMs), along with their corresponding molecular mass IR. In order to avoid creating and handling a huge database, the mined information is distributed among the database and a set of flat files that contain protein/peptide sequence information. It has been found that 1020 proteins have UMMs, while 20,282 out of 20,333 human proteins (Swiss-Prot 57) contain UP sequences. From them, 20,163 proteins contain unique tryptic peptides. The total number of CUPS in human is 7,338,522.

In the web application (PHP language), there are three modes of searching available as shown schematically in the information flow given in the Supplementary Material Figure 1. The user is able to (i) search whether a given molecular mass (or a list of molecular masses) can be found exclusively in one human protein, i.e. it is a UMM; (ii) search whether a given peptide sequence (or a list of peptide sequences) can be found exclusively in one human protein, i.e. it is a CUP and (iii) search a specific protein for UMMs or UP sequences.

In the first searching mode, the user may investigate a specific molecular mass (or a list of molecular masses), with or without an estimated error in Dalton. In this case, the system answers whether or not each submitted mass can be found exclusively in a human protein and provides the user with the corresponding peptide and protein as well as the corresponding smallest range (IR) where no other molecular mass can be measured in human proteome. Upon user request, when at least one unique mass from the given list has been found, a dynamically driven reporting procedure (using CGI PERL scripts) provides the user with the masses from the given list that are not unique but can be measured in the corresponding protein(s).

In the second searching mode, the user may investigate whether a given peptide sequence (or a list of peptide sequences) can be found exclusively in one protein. Actually, the system answers whether or not each peptide sequence is a CUP sequence or whether it contains or belongs to a CUP sequence. In case that a peptide sequence is found to be a CUP, the user is provided with the corresponding peptide mass, protein and ‘container’-peptide sequence that can be found through enzymatic digestion. Additionally, upon user request, the peptide sequence is further investigated for uniqueness across all species in Swiss-Prot.

If a given sequence is not among the CUPS, it is digested into subsequences in order to investigate whether or not a subsequence of the initial sequence is a CUP. If it is, then the given sequence is a UP but not a CUP. If it is not, the given sequence is matched to the set of CUPS to investigate whether or not it is a part of a larger sequence characterized as a CUP. In the case of a positive match, the given sequence is a fragment of one or more CUPS. In all other cases, the given sequence is determined as not unique in human.

In the third searching mode, the user may investigate whether or not a specific protein has UMMs and UP sequences. In this case, the system provides the user with a list of UMMs with their corresponding peptide sequences as well as with a list of CUPS and their corresponding molecular masses. The user can also find out which unique protein fragments can be computationally derived by the use of different enzymes. This is a very useful search mode that provides unique tags for a specific protein. For example, in the case of the candidate biomarkers Von Willebrand Factor (P04275) and Cathepsin B (P07858) (Polanski et al., 2007), the application finds three and six UMMs, respectively, along with a plethora of UPs that can be targeted to identify these proteins.

UniMaP is hosted by an Apache server on a Linux platform and incorporates a script-based curation protocol of the database and file archive. The updating procedure has been scheduled to run after each major release of the Swiss-Prot. UniMaP is part of a group of tools, databases and web services developed in the Biomedical Research Foundation, Academy of Athens, called ‘BioServer’. UniMaP can be complementarily used with the PeptideFinder tool (Alexandridou et al., 2008) hosted by the same server. It is expected to be very useful to proteomics research activities since it can lead, under certain circumstances, to single peptide protein identification when a peptide is unique for a protein. This tool can also be used...

Fig. 1. Data definitions and relations in the application.
UniMaP

in the emerging proteomics technology of SRM as well as in biomarker investigation and feasibility studies. Certainly, sequence or mass uniqueness is dependent on the completeness of the database. However, since the rate of new discoveries concerning human proteins has become very low, it allows for a very good approximation of the completeness of the database. Statistics, structural aspects, as well as signatures specificity on certain diseases are some of the subjects that are currently under investigation by our group and comprise our future work in this matter (Alexandridou et al., manuscript in preparation).

Conflict of Interest: none declared.

REFERENCES


