J.Natn.Sci.Foundation Sri Lanka 2021 49 (2): 255 - 271
DOI: http://dx.doi.org/10.4038/jnsfsr.v49i2.8845

RESEARCH ARTICLE

An improved method with higher efficiency for protein biomarker discovery and verification workflow

SM Vidanagamachchi*
Department of Computer Science, Faculty of Science, University of Ruhuna, Matara.

Submitted: 28 February 2019; Revised: 02 January 2021; Accepted: 25 February 2021

Abstract: Disease-specific protein biomarkers can assist in the disease identification process utilizing protein inference. These biomarkers play a major role in the drug discovery process. Biomarker discovery consists of a set of phases starting from mass spectrum files of peptides or proteins and ending with some significantly expressed proteins of a particular disease condition. Different techniques and tools have been introduced to perform protein inference and biomarker identification, and it still requires improvements in the accuracy of the protein and biomarker identification process. Further, it requires improvements in speed as it consumes hours or days to carry out the processes of protein biomarker discovery. In this paper, we thoroughly present and validate the Open Pipeline for Biomarker Identification (OPBI) on six different datasets and show how the pipeline fits into the process of protein identification with hardware acceleration. OPBI uses the information of tandem mass spectrometry (MS2) and the first stage of mass spectrometry (MS1). It achieved 0.0003–0.0004 false discovery rate and 2–3 times of speed-up with respect to existing MaxQuant software in different contexts on a general purpose computer with Intel Core i7 processor of 3.4 GHz frequency and 12 GB memory. Furthermore, the identified biomarkers can be utilized with the FPGA accelerated protein identification framework. According to the results observed, a considerable speed-up is achieved in the whole process of protein inference as well as peptide matching and peptide-protein mapping process. It further provides a methodology for the downstream analysis of protein biomarkers.

Keywords: Accuracy, biomarker-discovery, efficiency, fold-change-ratio, mass-spectrometry, shotgun-proteomics.

INTRODUCTION

The main objective of the article is to discuss the performance analysis of improved R-based open source pipeline (Open Pipeline for Biomarker Identification/ OPBI) (Vidanagamachchi & Niranjan, 2017) for protein biomarker discovery in shotgun proteomics. Here the main focus is on validation of the OPBI pipeline by extending the work of Vidanagamachchi and Niranjan (2017), compare the performance of the pipeline with MaxQuant pipeline, analyse the results for three different fold change ratios, and integrate the results of the pipeline to hardware accelerated protein identification framework (Vidanagamachchi *et al*., 2014).

There is an importance of the discovery of biomarkers to accelerate the diagnosis process based on previous knowledge when there is a disease spreading (eg: Dengue in Sri Lanka in 2017). These biomarkers should be accurate in order to identify the disease correctly and improve the reliability of identification. The algorithms currently used for biomarker discovery use peptide fragmentation search. Its search space is very large due to millions of fragments caused by proteins (eg: Humans). For example, for a peptide with 10 amino acids, 18 fragments might be used in the search and for around 100,000 proteins of humans, it increases the search space by more than three million peptides available and the

* Corresponding author (smv@dcs.ruh.ac.lk; https://orcid.org/0000-0002-2245-4527)

search space may become incredibly large. Although redundant peptides were removed, the fragmented database is large. Although most of the tools use peptide fragmentation fingerprinting (eg: Mascot, SEQUEST, ProteinPilot, ProteomeDiscoverer, Peaks, MaxQuant, ProteinProphet, IDPicker, TPP, X!Tandem), the tools such as MS-FIT, ProFound and Mascot use peptide mass fingerprinting. However, they can identify one particular protein as they were designed to use mass spectra of earlier low resolution mass spectrometers. Further, in the current methods of resolution, the requirement of primary memory is high as the number of entries being matched is high. Here, intermediate results have to be stored in MS/MS matching process; therefore demand for the secondary storage is high. It may take hours, days or months depending on the size of the dataset. Protein identification based on the peptides identified in the mass spectrometry has to be performed efficiently in the labs that process thousands of data every day. In any experiment, searching thousands of peptides from a database of peptides or a set of peak-lists, and mapping them with thousands of proteins are required and could not be accomplished manually. A normal workstation may require several days or months to process these proteomics data. For example, to analyse large datasets of proteins the processing time in a cluster environment (336 virtual cores, 2.53 GHz and 24 GB of RAM) was 5 days and on a desktop computer (Intel Core i7 2600 processor with 3.4. GHz, 16 GB of RAM, 460 GB hard disk space and 4 virtual cores) it was 20 days (Neuhauser *et al*., 2013). This requires sufficient accelerations in the laboratories that perform these tasks every day for analysing high-throughput proteomics experiments. Therefore, acceleration of protein identification by peptide matching would be advantageous for biologists. It will become more advantageous if the order of magnitude of the processing speed-up is high, which could be achieved by the hardware acceleration. However, the latter part of the biomarker identification process is accelerated here. The process is started from the identified biomarkers for protein identification. The algorithm reduces the search space of peptide identification by using previously known properties of peptides in the database (eg: theoretical peptide mass, pKa values of amino acids) and using the current pH value of the sample or mixture. pKa value is the pH at which the amino acid or group is neutral. The pipeline has been modified using peptide-protein mapping to identify proteins instead of searching through the whole database of proteins (eg: peptide 1: protein 10, protein 20) in order to reduce the time to be searched. Further, the sum of the intensities technique for calculating the fold change ratios was applied instead of using average or median fold change ratios. It has already proved that error propagation in the sum of the intensities methodology while calculating the fold change with real intensities is low (Vidanagamachchi & Niranjan, 2017). Here fold change ratio is calculated by the ratio between patient protein intensity and control protein intensity, where intensity is the expression of a particular protein in the sample. Therefore, utilizing the sum of the peptide intensities method is expected to provide users with a more accurate set of biomarkers.

Shotgun proteomics mostly relies on tandem mass spectrometry-based protein identification and quantification. Over the past few years, researchers have made different attempts to identify proteins in a complex mixture by digesting them into peptides as well as quantifying them for the discovery of biomarkers. In proteomics, biomarkers are indicators of a particular disease or a condition, such as proteins and peptides. The protein biomarkers have the ability to reduce the search space of the disease identification process as they act as disease-specific proteins. Biomarkers have the potential to reduce healthcare costs, improve treatments (Drucker & Krapfenbauer, 2013), early detection of a disease, and monitoring disease progression (He & Chiu, 2003). Biomarker discovery in proteomics provides insights into developing new drugs as well, bringing about a virtual revolution in the drug development process (Srinivas *et al*., 2002; Aebersold & Mann, 2003).

In shotgun proteomics, a mixture of digested peptides from a sample or multiple samples is the input for the protein identification process. Thousands of proteins could be seen in one sample (Krásný *et al*., 2013). Then the labelling can be applied for the peptides coming from different samples. Digested peptides are then ionized and fragmented through Liquid Chromatography (LC) and Tandem Mass Spectrometry (MS-MS). The peptide identification has been performed by mapping these fragmented spectra into peptide spectra. Tools such as Mascot, SEQUEST, and Proteome Discoverer are some popular tools for this mapping purpose. Assembling the identified sets of peptides is the final stage of protein identification. This is also known as protein inference. In the protein biomarker discovery process, the quantified expression levels (intensity/abundance) and significance values of protein were gained via mass spectrum data over different fractions. This information is processed further to identify significant biomarkers (Bai *et al*., 2012; Zhou *et al.*, 2012; McDermott *et al*., 2013). As designed here, a biomarker pipeline included as a pre-processing stage to the general shotgun proteomics pipelines could then bring-significant reductions to the time complexity of the integrated process.

The R-based open source pipeline (OPBI) for biomarker discovery which is described in detail in this paper targets high throughput, high accuracy mass spectrometry data, and allows all possible static modifications of peptides, up to one missed cleavage, and user defined mass tolerances in the search. Moreover, the performance of the pipeline was compared with the MaxQuant tool and integrated the outcome into the hardware accelerated protein identification framework. The pipeline is currently designed to work with iTRAQ and TMT Labelling and which is extendable to handle all types of other labelled experiments after straightforward adjustments.

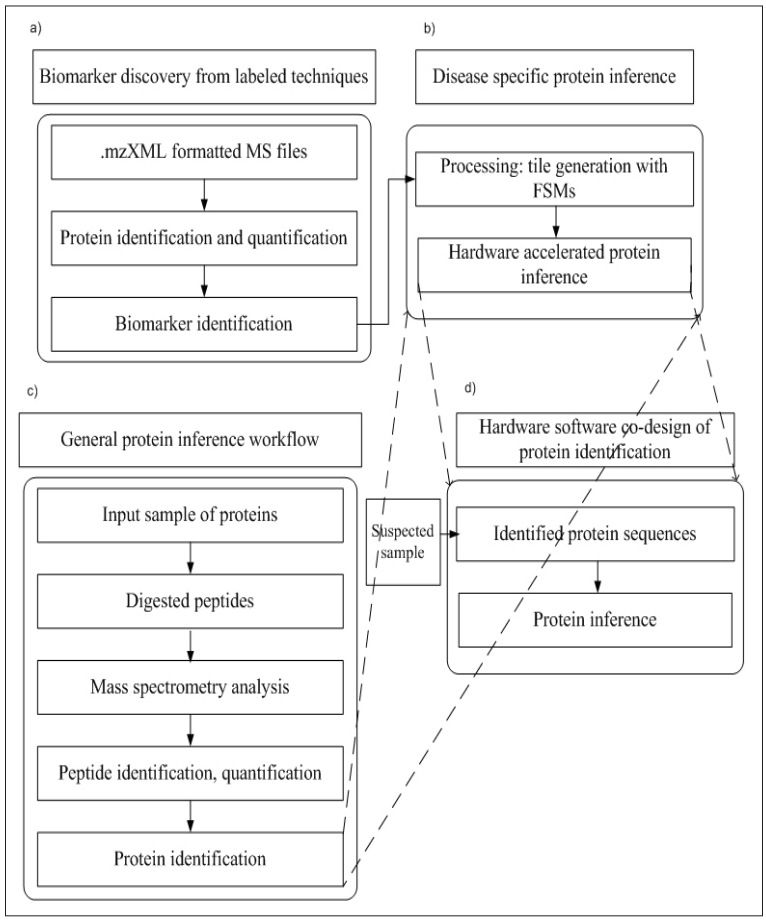

Figure 1: The integrated process of OPBI pipeline and hardware acceleration with FPGA: (a) R-based pipeline in identification and quantification of proteins of labelled experiments; (b) disease specific protein inference process; c) general protein inference process; d) hardware-software co-designed protein identification.

The integrated process of OPBI pipeline and hardware acceleration with FPGA is shown in Figure (1). The input to the biomarker discovery pipeline in the process (a) should be .mzXML files in labelled proteomics experiments and an R-based pipeline is used in the identification and quantification of proteins there. The identified biomarkers (proteins) from disease specific protein inference in the process (a) are subjected to digestion and construction of finite state machines in the process (b), which keeps a peptide pattern database in order to identify proteins in the hardware accelerated protein inference process shown in process (d). The process in (b) can also be used with general protein inference in the process (c).

During the protein inference process, one has to search through the whole list of proteins in a given database to identify the proteins from identified peptides. This can be done utilizing assembly algorithms or locating all identified peptides with string matching algorithms. Both these techniques require considerable time in identifying proteins. In general, to reduce computational time, cluster based databases of proteins could be utilized (Uniprot, 2017). However, for disease-specific protein inference, clusters are not very meaningful, whereas biomarkers are. These biomarkers can then be included in the hardware accelerated protein inference pipeline and it will reduce the time for protein assembly/ mapping significantly. The problem then is efficiently (in terms of computation time) and accurately (in terms of less false positives and false negatives) discovering these biomarkers for a given context.

Several commercial and non-commercial protein identification pipelines can be found in the literature (Perkins *et al.*, 1999; Jiang *et al*., 2007; Cox & Mann, 2008). Since commercial tools are very expensive, non-commercial tools have high demand among general users. In addition, there is no way to formally ascertain the efficiency claims of commercial tool chains other than through empirical evidence after using them. However, it still consumes days/hours to carry out the processes of protein biomarker discovery using these tools. Hence, the requirement of demonstrably faster procedures, which are accurate. Further, an R-based open-source pipeline will allow users to utilize the tool, access and modify the code based on the preference.

In protein inference, protein identification from identified peptides is the final step. Different tools and techniques are used for this process based on the requirement type (protein identification or quantification or both) and output mass spectrum file type of the experiment. Some tools start the processing of mass spectrum data with different available formats and identify only the peptides. Some identify proteins from identified peptide sequences, and some tools are for the identification of proteins starting from the mass spectrum data. Some researchers have introduced tools for further analysis (to calculate fold change ratios, compare different types of ratios, extract significant markers, etc.) of identified proteins. There exist some packages in R (eg: MSnbase, rTANDEM, iq, etc.) for the identification and quantification of proteins and statistical packages (eg: MSstats) for the above mentioned functions. Identification of biomarkers has to be performed separately in most of the tools since they do not include such a facility. In this case, biologists require some other tools or the support of programmers for further analysis of protein identifications. The biomarker discovery pipeline in the present study could be utilized to satisfy this requirement.

The searching process of peptides and proteins can be implemented using string matching algorithms. Among the available algorithms, Aho-Corasick algorithm is known to be better in time performance than the other multiple pattern identification algorithms, but it is not the best for memory utilization (known as a memory hungry algorithm (Commentz-Walter, 1979). Even so, with the ever-growing sizes of proteomics databases, there still remains much to be done in reducing the search times. Literature shows that a considerable speed-up could be achieved by implementing the sequence (e.g. genes, proteins or peptides) search and alignment processes on a hardware platform such as Field Programmable Gate Arrays (FPGAs) (Oliver *et al*., 2005; Wienbrandt, 2014). Hardware accelerated protein inference system can immensely benefit from biomarkers by reducing the scope of the search in databases, since the capacity of reconfigurable hardware is limited (in the experiments a DE2 Cyclone II 2C35 FPGA was used which is limited to 33, 216 logic elements).

The biomarker discovery pipeline is a totally separate workflow to assist the hardware accelerated protein inference. However, one could deploy some stages of a biomarker discovery pipeline: viz., peptide and protein identification stages on FPGA hardware, if the required capacity to generate FSM logic is available. Therefore, this biomarker discovery pipeline runs once, and the outputs observed are used in the hardware accelerated protein inference system.

The biomarker discovery problem has been formulated as a multi-criteria optimization problem and solved using Data Envelopment Analysis (DEA) (Lorenzo *et al*., 2015). In this paper, a simple strategy based on threshold values set for fold change ratios was adopted. It is straightforward to replace the latter with different algorithms like the optimization routine of Lorenzo *et al*. (2015) in the present pipeline.

METHODOLOGY

Isobaric tagged labelled experiments include iTRAQ, TMT and ICAT experiments, and of these experiments iTRAQ and TMT are more frequently used than ICAT. In the PRIDE database, most of the experiments are

SILAC based and iTRAQ and TMT take the next places respectively. In the current work for the biomarker discovery, the bottom-up approaches that utilize iTRAQ and TMT isobarically labelled experiments were considered. Here, the major advantage observed is the use of more than one sample at a time in the process of protein identification and quantification to identify biomarkers.

The steps of the biomarker discovery process in isobarically labelled proteomics experiment of OPBI pipeline has been described in the next paragraphs. OBPI is solely based on peptide mass fingerprinting for high resolution mass spectrum data and that makes it significant from other existing workflows. Further, as contributions of OBPI; the utilization of new techniques throughout the protein identification process such as pH value based m/z calculations, peptide database reduction methodology (unique peptide sequence and unique m/z), peptide mass fingerprinting with more information provided from MS2 spectrum analysis (precursor ion and reporter ion intensities), and real precursor intensity based protein fold change ratio calculations, can be identified. OPBI pipeline consists of 10 different stages except the peptide database generation after the in-silico trypsin digestion process: processing raw MS files, quantification of reporter ions, purity correction, missing value imputation, peptide identification, protein identification, peptide fold change ratio calculations, FDR calculations, P value calculations and biomarker identification (Nadin *et al*., 2013).

Different types of raw files are generated from mass spectrometers in the laboratories and most of them cannot be directly analysed by existing computational methods and tools due to the difficulty in reading and parsing their binary formats. Furthermore, the binary file formats become obsolete over the years with new software and become unreadable since the new tools only have the ability to read newer formats (Deutsch, 2012; Shah *et al*., 2010). Mass spectra generated from experiments are normally represented in two major modes: profile mode and peak lists (centroided). Profile mode is the most common type of mode the data is collected in mass spectrometers and it is continuous. In the peak list mode, only the picked peaks by an automated computerized system are written into the file and it is not continuous (Deutsch, 2012). The output format of mass spectrum files from mass spectrometer is vendor specific. For example, the AB SCIEX instruments use the .wiff extension as one of the extensions for storing the output spectra files (since they use AB SCIEX Analyst Software for mass spectrometry data acquisition) and these files might contain meta-data/all the information in a specific run. They are comprised of a .wiff.scan file that contains the spectra (Deutsch, 2012). In almost every biomarker discovery pipeline in proteomics, the above explained mass spectrum data are used. The developed pipeline allows .mzXML, .mzML and .mzData as the input, thus requires conversions of input files in some experiments. Here MSConvert tool was used as the converter.

The converted files are then subjected to quantification, which was done using the MSnbase package in R platform (version 3.1 or later). The major task of this process is to pre-process fragmented mass spectrometry scan files (MS level2 scans) in order to extract reporter ion abundances for protein quantification. A scan is a fraction from a particular sample. Therefore, a particular precursor m/z could be seen in several fractions as well. These details can be used in the process of protein identification later. The reporter ion intensities, precursor ion m/z, and intensities were also extracted. These are used for performing the peptide mass fingerprinting based protein identification. This process is used in most protein quantification tools. In OPBI, the extracted precursor intensities were used to improve the results of protein identifications, which is based on Peptide Mass Fingerprinting approach.

In this article, the results of analyses of some datasets selected from PRIDE database with the pipeline, which uses MSnbase package are discussed. The quantification results of the iTRAQ and TMT experiments were observed, which correspond to precursor ions with other important details of precursor ions such as corresponding reporter ions and precursor ion m/z and intensity. The intensities of reporter ions or precursor ions, which are already quantified, can be used in calculating the protein fold change ratio of proteins. Here the intermediate results of biomarker discovery in the stages such as quantification of peptides, purity correction, missing value imputation of reporter ion intensities, and peptide and protein ratio calculation are not explained, but they all contributed to the final results. The functions used for iTRAQ/TMT quantification process, purity correction, and missing value imputations are already proved as having consistent management and reproducible with iTRAQ and TMT labelled experiment data (Laurent *et al*., 2014; Gatto & Lilley, 2010).

In this peptide mapping of precursor ions utilizing the peptide mass fingerprinting approach, the observed mass to charge ratios of charged peptide ions (precursor ions) are used to identify peptides by matching theoretically calculated mass to charge ratios of precursor ions. This

approach is the basic approach for the identification of proteins and matches theoretical m/z values of peptides to experimental values (Carrillo *et al*., 2010; Risk *et al*., 2013). This is a powerful technique and has been used in many protein identification applications that use mixtures up to few samples or with a mixture of proteins (Pratt *et al*., 2002; Licker *et al*., 2007). It is used in the tools such as Mascot PMF, MS-FIT, and Profound. Mascot peptide mass fingerprinting is already developed for identifying protein mixtures up to 6 components (Matrix-Science, 2003). An optimization strategy for the results of peptide mass fingerprinting in mixture identification was introduced by He *et al*. (2009). There are limitations and advantages in this method. The advantage is that the molecular weight and pI or pH values can be incorporated into the peptide matching process (Giardina *et al*., 2010), and peptides that seem to be weak to perform MS2 fragmentation can be identified (He *et al*., 2009). The limitation is the identification of peptides that have identical mass to charge ratios. As a solution for this limitation peptide fragmentation fingerprinting (peptide sequencing method) can be utilized, but it is still unable to perform tandem mass spectrometry scanning on each and every ion (this results in an incomplete identification of peptides as mentioned earlier). Therefore, in analysing low abundance proteins, peptide mass fingerprinting could support whenever the samples are pure and more optimization to this method is done (He *et al*., 2009). However, with the unique m/z peptides, which include non-degenerate peptides, it was possible to obtain a set of proteins that are also unique from a given database of proteins. Therefore, in OPBI, it was first attempted to identify proteins with the peptide mass fingerprinting method. Then further identification is performed with peptide fragmentation fingerprinting if it is required. In most of the PMF approaches the user has to define a mass tolerance in the identification of peptides from m/z ratios. This mass tolerance defines the maximum mass difference between the theoretical mass and the experimental mass of the peptide. Assuming MS level 1 results are accurate in high resolution mass spectrometry, the peptide mass fingerprinting was performed to identify peptides allowing suitable static modifications. Here binary search was utilized as it can match the peptide mass with minimum tolerance value. For the purpose of the reanalysis of raw mass spectrometric data from isobarically labelled experiments, iTRAQ datasets were analysed first. These datasets are commonly used in biomarker discovery and are convenient to process with provided functionalities in R (e.g. reading mzXML data, MSnbase package). The use of PMF and the subsequent methodologies in protein identifications for high-resolution mass spectrum data are the main contributions to the whole biomarker discovery pipeline.

The pipeline could perform the digestion on any FASTA formatted database based on their format. The pipeline was developed for three databases; UniprotSwissProt (Expasy Bioinformatics Resource Portal, 2011), RefSeq (Pruitt *et al*., 2005), and TAIR (Phoenix Bioinformatics, 2011) database (release 10), which were used in the validation process of the pipeline. This digestion is performed online in some tools such as PeptideMass (Wilkins *et al*., 1997) and PeptideCutter (Gasteiger *et al*., 2005) and offline in others such as Protein Digestion Simulator (Pacific Northwest National Laboratory, 2011).

In the database generation step, mass to charge ratios of peptides were calculated based on their pH values. This is a new idea used in the pipeline. In the first step of peptide identification from mass spectrometry data (MS level 1), the peptides with unique mass to charge values were extracted from all the proteins since it might cause confusion while more than one peptide share the same mass to charge ratio. For peptide identification from fragmentation ions (MS level 2), the unique peptides were selected from the digested protein list based on the sequence of peptides.

In OPBI, while performing the digestion, the corresponding proteins per each peptide are recorded. Therefore, no more effort is needed to identify proteins for each peptide except just extracting the corresponding protein already mapped.

For peptide identification from fragmented precursor ions in MS/MS fragmentation, spectral matching based approaches (known as peptide fragmentation fingerprinting or PFF) are mostly used in the commercial tools of protein identification by mass spectrometry. In peptide mass fingerprinting, the peptides with identical masses cannot be identified and should deal with only unique peptides in the searching phase to overcome this problem as described in the previous section. The only solution is to look at more details of the peptides. To achieve this, the utilization of tandem mass spectrometry (MS2 level) is conducted in the mass spectrometry experiments. Most of the scoring methods and techniques used in different proteomics tools are not free and open source. The applicability of the sum of the intensities method in labelled experiments were demonstrated with some selected large and real datasets of proteins in PRIDE archive for confident determination of biomarkers

(Blizzard *et al*., 2010). Furthermore, one scoring method (Peppy's score) was selected that could compete with existing scoring techniques used in commercial tools for the MS2 level protein identification (Berger, 2010). However, this was integrated as an additional function in the OPBI pipeline as it takes considerable time in matching with large datasets. Figure (2) shows the design of biomarker discovery including the main contributions of the OPBI pipeline, which targets a minimal protein list in hardware assisted disease-specific protein inference. This figure summarizes the whole process of protein identification and quantification that aims to identify biomarkers starting from mass spectrometry raw files collected from proteomics experiments. The figure shows how the database of peptides is made starting from extracting protein data from (as FASTA files) existing databases of proteins. The m/z calculations are based on the pH value of the solution used in the experiment since the chemical breakdown, and the charge state (z) of the precursor ion depends on pH value (Syed *et al*., 2015). The high-resolution mass spectrometry provides accurate masses in the first level of mass spectra (Syed *et al*., 2015). PI values or pH values can be integrated into the search of peptides in this method of identification (Giardina *et al*., 2010). The application of peptide mass fingerprinting can be seen only in a single or few proteins in protein identification tools. The reason for this is known as the difficulty of identification of identical mass peptides and mass inaccuracy (Matrix Science, 2003). This work demonstrates that the advantage of peptide mass fingerprinting by the integration of both MS level 1 data (precursor mass and intensity) and MS level 2 data (reporter ion intensities) in identifying confident biomarkers, which consist of the intensity based fold change ratios of proteins and unique peptides up to 8 samples (8-plex). The significance of proteins was calculated based on the intensities of proteins from the two groups: patient and control. The sum of the intensities method is then used as the main protein ratio calculation method due to its minimum error propagation while calculating protein ratios from peptide ratios (Carrillo *et al*., 2010).

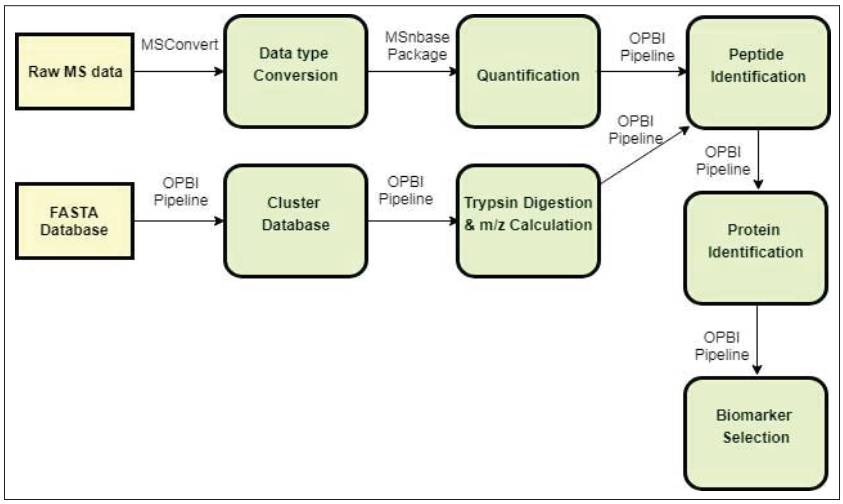

Figure 2: The design of biomarker discovery process including main contributions of OPBI pipeline

In most of the biomarker discovery and protein quantification tools, the abundance of proteins is measured by calculating the protein ration first. Then the protein ratio or protein fold change is calculated. There exist several methods of calculating fold change of proteins in labelled experiments; average of peptide ratios, libra ratio, linear regression, principal component analysis and the sum of the peptide intensities (Carrillo *et al*., 2010). Average of peptide ratios method calculates the intensity ratios of all peptides (in isobarically labelled experiments it is reporter ion ratios). Libra ratio is calculated based on applying an outlier rejection scheme to the average of the peptide ratios. The linear regression method attempts to fit a line over peak intensities based on standard linear regression and the slope of this line is considered to be the relative abundance/protein ratio.

In the method of principal component analysis, a principal axis is selected by projecting the highest variance of the data and the slope of this axis is used as the relative abundance of the corresponding protein. In the sum of peptide intensities method, the intensities of all peaks of the same label are summed together to calculate the protein ratio. Although the total least square method is bit similar to linear regression method, it is more complex than linear regression method. In this method, a line is fit over the peak intensities of two types of labels compared (Carrillo *et al*., 2010). In the current pipeline, the sum of peptide intensities method was utilized, which has mathematically and experimentally been proven as having the best performance and being better than the mostly used average of ratios method. The propagation of errors in calculating protein ratios from this sum of peptide intensities method has been proven as less than in other methods (Carrillo *et al*., 2010). The average of the peptides can be further calculated as a method of imputing missing values in the sum of peptide intensities method. Here the peptide ratio is the ratio of two labels (e.g. patient/healthy or 115/114), which is calculated from the reporter ion intensities since it is proportional to the corresponding peptide intensities. In this present pipeline of biomarker discovery for labelled experiments (TMT, iTRAQ), the combined intensity for a selected precursor ion was used from differently labelled identical peptides from the same/different proteins. Following equations 1, 2, 3 and 4 show the steps of calculating the intensity of a precursor ion from a specifically labelled peptide for a simple 2-plex TMT/ iTRAQ experiment. Here r1 and r2 represent the reporter ion intensity of patient sample and control sample, respectively. P1 and P2 represent precursor ion intensity of patient sample and control sample, respectively. R1 and R2 represent the total intensity of reporter ions in patient and control sample, respectively and k is the number of peptides in each sample (generally assumed to be 1). Therefore, obviously P1 = kr1, P2 = kr2, R1 = kr1 and R2 = kr2. This sum of intensities method has not been used for iTRAQ and TMT in the currently developed tools and here it has been contributed by considering this ratio, combining both reporter ratios and precursor intensities. Here, i and j in equation (4) are finite numbers and can be equal as well.

$$P1 = (P1+P2)\, kr1/(kr1+kr2) \quad \ldots (1)$$

$$P1 = (P1+P2)\, R1/(R1+R2) \quad \ldots (2)$$

$$P2 = (P1+P2)\, R2/(R1+R2) \quad \ldots (3)$$

$$\text{Protein ratio} = \sum Pi / \sum Pj \quad \ldots (4)$$

In this pipeline, reporter ion ratios were calculated from the already validated MSnbase package and the sum of the intensities methodology for fold-change ratio calculations proved more accurate than the other existing techniques (Carrillo *et al*., 2010; Gatto & Lilley, 2012). Further, unique precursor masses were considered in the peptide identification and it improved the accuracy of identification. Accuracy of the discovery process can be further measured by the False Discovery Rate (Walhout *et al*., 2012). The matching peptides from the database searches can consist of false positives. Therefore, a target-decoy search is generally utilized for increasing the confidence of the peptide searches by performing a target database (forward database) search against a reversed, shuffled or random decoy database (Elias, 2007; Elias & Gygi, 2009; Gupta & Pevzner, 2009). Reverse database search allows in finding confidential target-decoy search (Elias, 2007) and it is automatically performed by reversing a given FASTA formatted database. Therefore, it will always generate the same reverse sequence whereas shuffling or randomization generates different sequences at different times (Jeong *et al*., 2012). Then, this decoy database can be concatenated with the target database or use separately in the decoy search (after digesting peptides or for later digestion) (Elias, 2007; Jeong *et al*., 2012). The concatenated strategy is convenient to use than performing a separate search and it is widely used (separate strategy is conservative, but some groups prefer that strategy) (Jeong *et al*., 2012). False Discovery Rate (FDR) can be calculated by doubling the number of decoy matches (false positives) and dividing this by the total number of matches (false positives + true positives) (Elias *et al*., 2005). The advantages of this method are confidence selection of peptides, no requirement of prior knowledge of mixture composition, and independence from both search algorithm and the instrument (Elias *et al*., 2005).

There are two types of replicates in a mass spectrometry experiment: biological replicates and technical replicates. Biological replicates can be from different patients and healthy people. Technical replicates are the ones from separate fractions from the same sample or same sample extracted multiple times. In most of the iTRAQ and TMT experiments, the researchers use technical replicates by taking fractions of the same sample. Therefore, it leads to a different set of proteins identified from each fraction. The final set of proteins is the union of the protein identifications of these fractions.

In order to check the possibility of getting significant biomarkers from the proteins identified, Wilcoxon test based p value is calculated. For this t-test based p value

could be applied if the distribution of protein abundances among fractions is normal. Further, the p value can be used in rejecting a null-hypothesis. Here, the p value is compared with the probability of rejecting the null hypothesis when the null hypothesis is true (known as α= significant level). In all the experiments, the significance criteria considered was 0.05, in which the null hypothesis was rejected when the p value < 0.05 (corresponding to 95 % confidence level). Here a distribution of abundance is made using all fractions as a matrix from all patient and control samples. For a properly designed experiment, it is known that a small p value is given for a false of null hypothesis and not because of random sampling (Philip, 2008).

The up-regulated and down-regulated proteins could be identified based on the results of protein abundance ratios or fold change (typically threshold is around 1.3 for up-regulated and 0.7 for down-regulated proteins, here up-regulated proteins increase in their expression and down-regulated proteins decrease in their expression). These proteins can be taken as biomarkers and it is considered good if the significance of these proteins is high (based on the p value calculated as mentioned in the previous section). Here a contribution is made by calculating protein fold change ratios based on real intensities of precursor ions.

RESULTS AND DISCUSSION

Firstly, for the validation of OPBI pipeline, a dataset of breast cancer was selected for biomarker discovery, which was previously analysed by another method by Chang *et al*. (2015). This dataset was analysed based on the peptide mass fingerprinting methodology, which was developed in R. Breast cancer is a major type of cancer, and it is reported that more than 18,000 humans are identified as having breast cancer annually in Nordic countries that lead to 15% of female deaths (Thivierge *et al*., 2006). Over-expression of proteins can be seen while developing a particular disease condition (Hojo *et al*., 1998; Thivierge *et al*., 2006). It refers to making multiple copies of the proteins and increasing the concentration of that. With this dataset, two over-expressed (fold change >= 1.1 or <= 0.8) and significant proteins were noticed: DCAF8L1 and RPTOR. Three methods of protein ratio calculation: average of the peptide ratios, median of the peptide ratios and sum of the intensities methods were utilized here. No up-regulated protein was found common in the three methods. The threshold for the up-regulated proteins was taken as above 1.1 fold change, and it was below 0.8 fold change for down-regulated proteins. These proteins are already reported as the biomarkers of breast cancer (Chang *et al*., 2015; Chen *et al*., 2015), and also recorded in the protein atlas database as highly expressed in breast cancer. Figure 3 is the Venn diagram of over-expressed proteins of breast cancer based on three different protein ratios mentioned earlier. Here, only 2 proteins are common in breast cancer patients based on the calculations of two methods: average and median of the peptide ratios methods. Two and one protein can only be seen in the results of the sum of the intensities method and average of peptide ratios method, respectively. Further analysis of accuracy was achieved by calculating False Discovery Rates (FDR) by utilizing target-decoy-searching approach and it was 0.0003 in the current approach, whereas it was 0.01 (strict FDR) in the previous approach (Chang *et al*., 2015).

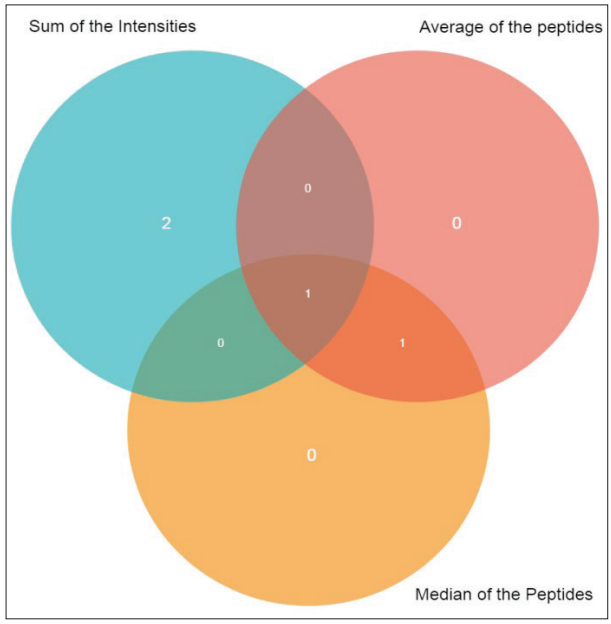

Figure 3: Venn diagram of over-expressed proteins of breast cancer based on three different protein ratios

In order to differentiate the malignant and non-malignant biliary stenosis, one could perform biomarker identification of bile cancers. Here, a dataset in PRIDE was used for identifying common biomarkers of biliary stenoses: pancreatic cancer and cholangiocarcinoma (Farina *et al*., 2014). These two diseases can be caused by bile stones or surgical injuries, and these can be seen as common types of cancers in developed countries. Annually, more than 200,000 people die because of pancreatic cancer as reported in 2004 (Michaud, 2004).

According to the results obtained, 10 proteins were common among three calculated ratios for the two disease types, cholangiocarcinoma, and pancreatic cancer. From these proteins, 9 proteins are highly significant ($p <= 0.05$) in the pancreatic cancer group (NOTCH4, PKHD1L1, CPN1, PRKRIR, ASB18, HS6ST1, PTPN9, NXPH2, CDC40, and CTR9). Among these proteins, there were no common proteins for the three ratios calculated. Therefore, it can be decided that the proteins reported as common are not significant and cannot be considered as good biomarkers. Although there are high fold change (fold change >= 1.2) proteins in the group of cholangiocarcinoma, none of them are very significant. An isoform of one of the proteins in the list used in Farina *et al*. (2014) is observed here, and it is significant in pancreatic cancer (Ankyrin repeat and SOCS box protein 18/ ASB18). Most of the proteins reported from the sum of the intensities method have previously reported evidence. From the up-regulated proteins both in pancreatic cancer and cholangiocarcinoma, the proteins NOTCH4, PKHD1L1, HS6ST1, PTPN9, CDC40, CTR9, CPN1, PRKRIR and ASB18 are noticed as significant and has evidence for pancreatic cancer (Elnemr *et al*., 2001; Zhang *et al*., 2001; Chaudhary *et al*., 2007; Hagiwara, 2011; Joseph *et al*., 2011; Lu *et al*., 2014; Xu, 2015). Although NXPH2 and DOCK2 are not significant, they are up-regulated and already identified for their association with pancreatic cancer (Pilarsky *et al*., 2008; Mahoney *et al*., 2015). From the down-regulated proteins (fold change <= 0.7), it was observed that some proteins such as HOXC13, LTBP4 and LTBP4 are already identified as biomarkers in previous research (Thalappilly *et al*., 2008; Cantile *et al*., 2009; Leconet *et al*., 2009). Some of the proteins were already classified as highly expressed in cancer patients in the human protein atlas database (e.g. PRKRIR, ITFG1, and PTPN9 from the up-regulated proteins and FAM117A and C2orf69 from down-regulated proteins). Figure 4 shows the comparison of over-expressed proteins resulting from three ratios; the sum of the peptide intensities, the average of the peptide ratios, and the median of the peptide ratios methods. Further analysis of accuracy was achieved by calculating FDR by utilizing target-decoy-searching approach and it was 0.0004 in the present approach whereas it was 0.05 in the previous approach (Farina *et al*., 2014).

Thirdly, a dataset of chikungunya serum samples was selected for biomarker discovery (Puttamallesh *et al*., 2013) with different techniques developed in R. Chikungunya is a mosquito transmitted disease, which can be seen in the countries of Europe, America, Asia and Africa. The symptoms of this disease are fever and severe joint pain. Samples from 5 chikungunya patients and 4 controls were used in this experiment.

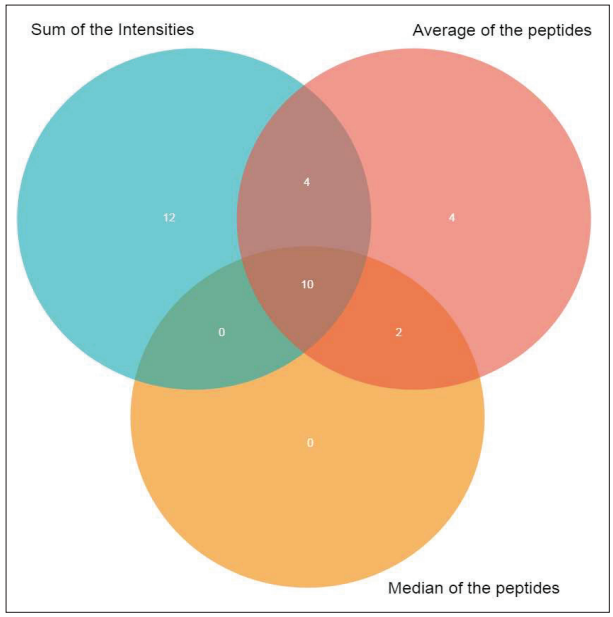

Figure 4: Venn diagram of over-expressed proteins of pancreatic cancer based on three different protein ratios

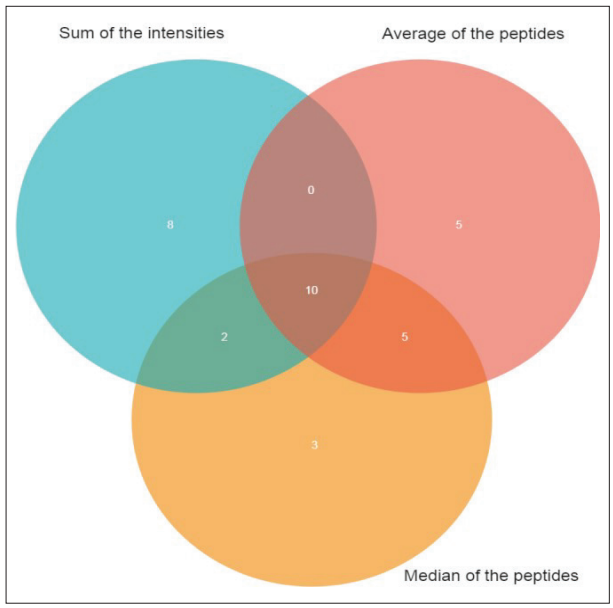

Figure 5: Venn diagram of over-expressed proteins of chikungunya based on three different protein ratios

Initially, equal amounts of each sample was obtained and proteins were separated (in-silico). Then proteins were digested and in vitro labelling was performed with iTRAQ reagents. Finally, the labelled peptides were subjected to SCX fractionation (resulted in 22 fractions) and high-resolution Fourier transformation mass spectrometry. While mapping precursor ions to peptides in a database, some tolerance should be allowed in order to find a matching for that particular ion, since the measurements might have some noise. First, matching was performed allowing different mass tolerance to show how the identified peptides and proteins could differ based on the user allowed mass tolerance in a particular proteomics experiment. Figure 5 shows the identified proteins and peptides based on mass tolerance distribution. The decision of which tolerance to be used is based on the mass spectrometer and its frequency of calibration. Figure 5 shows the comparison of over-expressed protein results from three ratios: the sum of the peptide intensities, the average of the peptide ratios and the median of the peptide ratios methods. Here, only 10 proteins are common in chikungunya patients based on the calculations of the three methods. Sum of the intensities method and average of peptide ratios method have shown twelve and four proteins, respectively and the details are presented in the next few paragraphs.

The results of protein identification depend on the parameters used. Therefore, the results with some parameter tuning were performed first. Initially, one static modification was used allowing no (0) cleavages. Then, two and three modifications with no allowed cleavage were explored. Finally, three modifications with one cleavage were used in the biomarker discovery process. The results of two static modifications (iTRAQ labels at the peptide N-terminus and Lysine residues) allowing zero missed cleavage are presented first.

In the paper Puttamallesh *et al*. (2013) have calculated the fold change protein ratios only; they do not give any measure for the statistical significance of the output proteins. It was observed in this study that the output proteins are not statistically significant. However, based on fold-changes, it was possible to find a set of biomarkers (these are proteins identified from unique peptides) that relate to host response. In this study 10 common proteins were found. As mentioned in Puttamallesh *et al*. (2013) proteins having high/low fold change (AJUBA, POU4F2, NDUFA11, P2RX4, GBA, NCF2, CYB561D1, and DPPA3) were found with some previous evidence of dengue/acute dengue and chikungunya. Since profiling of serum samples was attempted from chikungunya-infected patients recently, the biomarkers as evidence were not found for chikungunya in the literature for the present study.

From the results observed, protein glucosylceramidase (GBA), previously reported as a biomarker for chikungunya and dengue, is down-regulated in the current results (Nguetcheu *et al*., 2010). Proteins having more than 1.2 fold change are considered as up-regulated and less than 0.7 are considered as down-regulated in the pipeline. Since there are many over-expressed proteins, the first 10 were selected from each up-regulated and down-regulated protein sets. Here p value of the proteins are not that significant, but some proteins like P2RX4 have p values of 0.09 and previously reported as associated with acute dengue (Kwissa *et al*., 2014). NUT family member 2A (NUTM2A), Actin-binding LIM protein 2 (ABLIM2), chymotrypsin-like protease CTRL-1 (CTRL), POU domain-class 4 and transcription factor 2 are observed as up-regulated proteins that are already reported as associated with dengue disease (POU2AF1) (Nascimento *et al*., 2009; Kwissa *et al*., 2014; Saisawang *et al*., 2015). From the other down-regulated proteins, NADH dehydrogenase (ubiquinone) 1 alpha sub complex subunit 11 (NDUFA11), P2X purinoceptor 4 (P2RX4), developmental pluripotency-associated protein 3 (DPPA3), platelet factor 4 variant (PF4V1), kinesin-like protein KIF2A (KIF2A) and proheparin-binding EGF-like growth factor (HBEGF) are reported as associated with dengue in the literature (Chiu *et al*., 2014; Kwissa *et al*., 2014). Platelet factor 4 variant (PF4V1) and kinesin-like protein KIF2A (KIF2A) are reported as related to both dengue and malaria (Min-Oo *et al*., 2003; Muehlenbachs *et al*., 2007; Kwissa *et al*., 2014); protein neutrophil cytosol factor 2 (NCF2) was reported as associated to malaria disease, and it is down-regulated in the results of the current study (Sercundes *et al*., 2013). Figure 6 shows the Venn diagram of chikungunya results in the present study with reported dengue and malaria proteins. Further analysis of accuracy was achieved by calculating FDR by utilizing target-decoy-searching approach and it was 0.0004 in this approach whereas it was 0.01 in the previous approach (Puttamallesh *et al*., 2013). The parameters of peptide and protein search were then tuned with more static modifications and missed cleavages of peptides: methylthiol modification of cysteine, iTRAQ labels at the peptide N-terminus and lysine residues as static modifications. Protein S100-A7 was found significant and slightly over-expressed and was also previously reported as associated as a host response factor during an infection in the previous research (this is one out of the 3 proteins/protein families: clusterin, apolipoproteins and S100A family of proteins) (O'Farrell, 1975). From the

other 5 over-expressed proteins, peroxisomalcarnitine O-octanoyltransferase (CROT), trafficking protein particle complex subunit 9 (TRAPPC9), receptor-type tyrosine-protein kinase FLT3 (FLT3) and sodium-dependent phosphate transporter 1 (SLC20A1) were reported as over-expressed and significant with the fold-change thresholds, 1.2 for up-regulated proteins and 0.7 for down-regulated proteins. All these proteins were previously reported to be associated with the malaria disease (GEO profiles, Refgene database) (Yuda *et al*., 2011; Tamura *et al*., 2014). Protein C-type lectin domain family 4 member K (CD207) was also reported as over-expressed and significant, but previously reported as associated with dengue. Figure 7 shows the output comparison of three ratios where no over-expressed proteins were common to all the three ratios and 3, 3 and, 4 proteins were over-expressed only based on the sum of the intensities ratios, the average of the peptide ratios and the median of the peptide ratios, respectively. When the number of allowed missed cleavages were increased, the time consumed for each peptide and protein search gets increased. For example, the time gets doubled (approximately- for the chikungunya dataset) for the peptide and protein search if the allowed cleavages are increased from 0 to 1. For example, time consumption on the personnel computer of Intel core i7 machine with 12 GB memory for chikungunya dataset was around 40 min, 160 min, 160 min, 30 min, 20 min, and 50 min for quantification, trypsin digestion, m/z database generation, peptide identification, protein identification, and biomarker identification, respectively. From these, the functions trypsin digestion and m/z database generation can be performed once for a particular organism in the pre-processing stage. Therefore, total time spent on a PC could be considered as 90 min (except for the biomarker discovery function as it cannot be found in MaxQuant). The speed-up varies between 2–4 for different datasets.

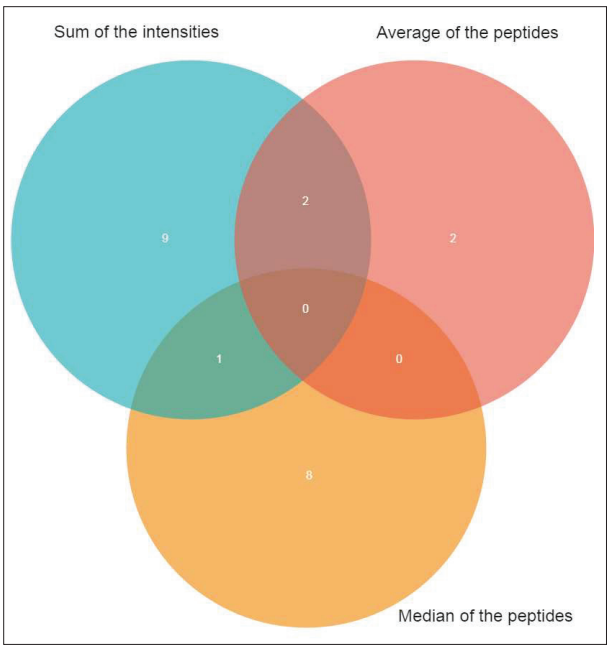

Figure 6: Venn diagram of chikungunya results with reported dengue and malaria proteins

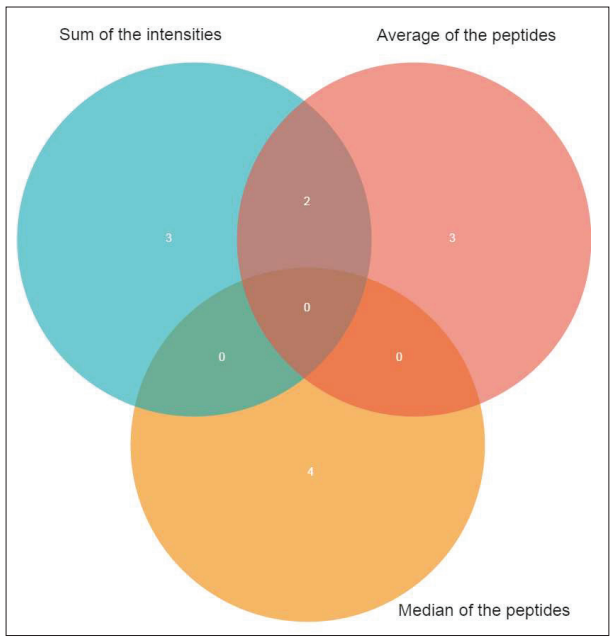

Figure 7: Venn diagram of chikungunya results with reported dengue and malaria proteins with more static modifications and missed cleavages of peptides

Head and neck squamous cell carcinoma is a common type of cancer, and it can be seen in 4 % of cancers in the United States and 5 % of cancers in the United Kingdom. For the purpose of identifying biomarkers of head and neck squamous cell carcinoma (HNSCC), HNSCC cell lines and normal oral keratinocyte cell lines can be used. A reanalysis of the dataset used in Sahasrabuddhe *et al*. (2015) was performed with the pipeline and a comparison was performed with the protein ratios observed. Three ratios were calculated as previously performed with other datasets. The threshold for the up-regulated proteins was set to be 1.2, and it was 0.7 for the down-regulated proteins. Only proteins from the intensity-based ratio method was observed,

from which proteins SCO1 and LRP1 have already been reported in head and neck squamous cell carcinoma (Langlois *et al*., 2010; Sandulache & Skinner, 2012; Paulette *et al*., 2014), and protein TAS2R43 in colon cancer (Nehrt *et al*., 2012). Further analysis of accuracy has been achieved by calculating FDR by utilizing the target-decoy-searching approach, and it was 0.0003 in the present approach whereas it was 0.05 in the previous approach (Sahasrabuddhe *et al*., 2015).

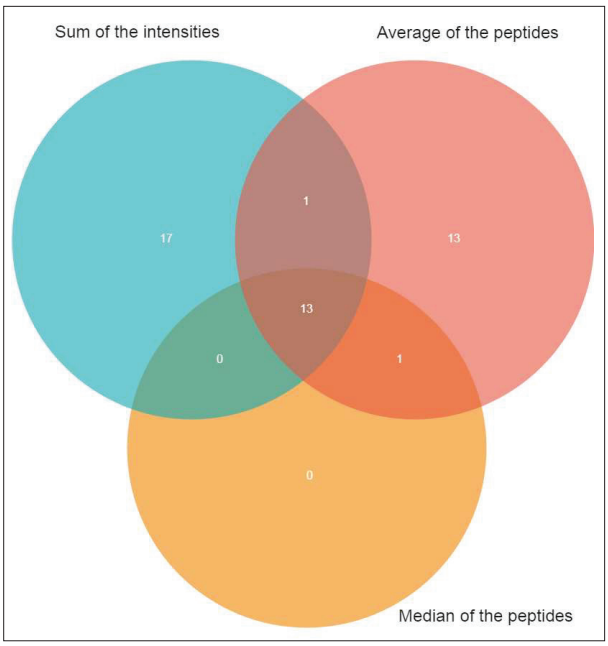

Figure 8: Venn diagram of over-expressed proteins of gallbladder cancer based on three different protein ratios

Gallbladder cancer is known as bile duct cancer and not very common in most parts of the world except Chile, India, China, and Japan. The raw data used by Sahasrabuddhe *et al*. (2014) was analysed with the present method, that utilized peptide mass fingerprinting with unique peptides and observed one significant ($p <= 0.05$) protein FBXO2, which is down-regulated (fold change <= 0.7). It is reported as an over-expressed protein several times in the past for gallbladder cancer (Liang *et al*., 2013; Zhang *et al*., 2013) and thyroid cancer (Giulia *et al*., 2010). Other proteins that have a very high or low fold-change are CD207 and BMP7. These could also be seen in gallbladder/bile duct cancer as associated proteins (Numata & Oshima, 2012; Yang *et al*., 2014). Figure 8 is the Venn diagram for overlapping of proteins based on the three ratios calculated. Further analysis of accuracy has been achieved by calculating FDR by utilizing the target-decoy-searching approach and it was 0.0004 in the present approach whereas it was 0.01 in the previous approach (Sahasrabuddhe *et al*., 2014).

Dugesia japonica is a planarian/flatworm that can regenerate a head within less than one week (Sandmann *et al*., 2011). The data of Geng *et al*. (2015) was analysed with the present pipeline. The experiment is designed as a time- course experiment and the *D. japonica* has been kept for 0 hours, 2 hours and 6 hours. The present study calculated the ratios with respect to 0 hour proteomics measurements. Here we have a small database, and few up- (fold change >= 1.1) or down- (fold change <= 0.8) regulated proteins from the sum of the intensities method and the median of the ratios method, except the average of the protein ratios method. The ratios illustrate that the average of the peptide method can have many false positives as outputs due to error propagation of protein ratio calculation. None of the output is significant as well. However, here it is difficult to find out proteins that could be the only output by the sum of the intensities method, since there is not much difference before and after regeneration of *D. japonica*. Here we consider actual intensities and output of the peptides only in the proteins of peptides with unique m/z. Figure 9 illustrates how the proteins get overlapped for three ratios calculated, in which we have one down-regulated protein in common (Q8MXW3/Djfkh7). Out of the 4 proteins expressed in the sum of the intensities and the average method, one protein (O76154/DjTRPMb) is also identified as a biomarker in the planarian head regeneration process (Beane *et al*., 2011). Further analysis of accuracy was achieved by calculating FDR by utilizing target-decoy-searching approach and it was 0.0009 in the present approach whereas no FDR was reported in the previous approach (Geng *et al*., 2015).

Table 1: FDR values in OBPI and previous method

Dataset	FDR (present study)	FDR (previous study)
Breast cancer	0.0003	0.01
Pancreatic cancer	0.0004	0.05
Chikungunya	0.0004	0.01
HNSC cancer	0.0003	0.05
Gallbladder cancer	0.0004	0.01
Head regeneration (Planarians)	0.0009	Not reported

Considering the total time in quantification, peptide identification and protein identification, we have achieved 2–4 times speed up with respect to MaxQuant for different datasets (breast cancer, pancreatic cancer, chikungunya, head and neck squamous cell carcinoma, gallbladder cancer, head re-generation of planarians) on Intel Core i7 processor, 3.4 GHz and 12GB RAM personnel computer. Table 1 shows the comparison of FDR values of the present method with the previously reported method. The higher the FDR, the lower the accuracy.

CONCLUSION

Although different techniques and tools have been introduced for the identification process of biomarkers, it still requires improvements in accuracy and speed. Here, the OPBI pipeline has been presented and validated our for 5 different datasets. We observed 2–4 times speed-up with respect to existing MaxQuant software in different contexts on a general purpose computer with Intel Core i7 processor of 3.4 GHz frequency and 12 GB memory and from 0.0003 to 0.0009 False Discovery Rates, which shows the improvements of accuracies in the open-source biomarker discovery process. OPBI provides a methodology for the downstream analysis of protein biomarkers.

Acknowledgements

The author acknowledges Prof. M. Niranjan and Dr S.D. Dewasurendra; partners in the main project group where this research work was carried out, for early work with them led the author to undertake the current investigation.

REFERENCES

Aebersold R. & Mann M. (2003). Mass spectrometry-based proteomics. *Nature* **422**: 198–127.
DOI: https://doi.org/10.1038/nature01511

Bai U.V., Hwang O., Divine G.W., Barrack E.R., Menon M., Reddy G.P. & Hwang C. (2012). Averaged Differential Expression for the Discovery of Biomarkers in the Blood of Patients with Prostate Cancer. *PLoS ONE* **7**: 37–51.
DOI: https://doi.org/10.1371/journal.pone.0034875

Beane W.S., Morokuma J., Adams D.S. & Levin M. (2011). A chemical genetics approach reveals H,K-ATPase-mediated membrane voltage is required for planarian head regeneration. *Chemistry and Biology* **18**(1): 77–89.
DOI https://doi.org/10.1016/j.chembiol.2010.11.012

Berger B. (ed.) (2010). MRF for NMR side-chain assignment. Research in Computational Molecular Biology. *14th Annual International Conference RECOMB 2010*, Lisbon, Portugal, pp. 551–552.

Blizzard E.L., Davis C.D., Scott H., Long D.B., Hall C.A. & Yabsley M.J. (2010). Distribution, prevalence, and genetic characterization of baylisascarisprocyonis in selected areas of Georgia. *Journal of Parasitology* **96**(6): 1128–1133.
DOI: https://doi.org/10.1645/GE-2518.1

Cantile M. *et al*. (14 authors) (2009). Hox d13 expressionacross 79 tumor tissue types. *International Journal in Cancer* **125**: 1532–1541.
DOI: https://doi.org/10.1002/ijc.24438

Carrillo B., Yanofsky C., Laboissiere S., Nadon R. & Kearney R.E. (2010). Methods for combining peptide intensities to estímate relative protein abundance. *Bioinformatics* **26**: 98–103.
DOI: https://doi.org/10.1093/bioinformatics/btp610

Chang H., Li M., Huang T., Hsu C., Tsai S., Lee S., Huang H. & Juan H. (2015). Quantitative proteomics reveals middle infrared radiation-interfered networks in breast cancer cells. *Journal of Proteome Research* **14**(2): 1250–1262.
DOI: https://doi.org/10.1021/pr5011873

Chaudhary K., Deb S., Moniaux N., Ponnusamy M.P. & Batra S.K. (2007). Human RNA polymerase II-associated factor complex: dysregulation in cancer. *Oncogene* **26**(54): 7499–7507.
DOI: https://doi.org/10.1038/sj.onc.1210582

Chen Z., Sui J., Zhang F. & Zhang C. (2015). Cullin family proteins and tumorigenesis: Genetic association and molecular mechanisms. *Journal of Cancer* **6**(3): 233–242.
DOI: https://doi.org/10.7150/jca.11076

Chiu H.C., Hannemann H., Heesom K.J., Matthews D.A. & Davidson A.D. (2014). High-throughput quantitative proteomic analysis of dengue virus type 2 infected a549 cells. *PLoS ONE* **9**(3): e93305.
DOI: https://doi.org/10.1371/journal.pone.0093305

Commentz-Walter B. (1979). A string matching algorithm fast on the average. In: *Automata, Languages and Programming* (ed. H.A. Maurer). ICALP 1979. Lecture Notes in Computer Science, volume 71, pp. 118–132. Springer, Heidelberg, Berlin, Germany.
DOI: https://doi.org/10.1007/3-540-09510-1_10

Cox J. & Mann M. (2008). Maxquant enables high peptide identification rates: individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Natures Biotechnology* **26**(12): 1367–1372.
DOI: https://doi.org/10.1038/nbt.1511

Deutsch E.W. (2012). File formats commonly used in mass spectrometry proteomics. *Molecular and Cellular Proteomics* **11**(12): 1612–1621.
DOI: https://doi.org/10.1074/mcp.R112.019695

Drucker E. & Krapfenbauer K. (2013). Pitfalls and limitations in translation from biomarker discovery to clinical utility in pre-dictive and personalised medicine. *EPMA Journal* **4**: 4–7.
DOI: https://doi.org/10.1186/1878-5085-4-7

Elias J.E. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4**(3): 207–214.
DOI: https://doi.org/10.1038/nmeth1019

Elias J.E. & Gygi S.P. (2009). Target-decoy search strategy for mass spectrometry-based proteomics. *Methods in Molecular Biology* 6**04**: 55–71.
DOI: https://doi.org/10.1007/978-1-60761-444-9_5

Elias J.E., Haas W., Faherty B.K. & Gygi S.P. (2005). Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nature Methods* **2**(9): 667–675.
DOI: https://doi.org/10.1038/nmeth785

Elnemr A. *et al*. (11 authors) (2001). Human pancreatic cancer cells disable function of Fas receptors at several levels in Fas signal transduction pathway. *International Journal of Oncology* **18**: 311–6.
DOI: https://doi.org/10.3892/ijo.18.2.311

Expasy Bioinformatics Resource Portal (2011). Uniprotkb/swiss-prot. Available at *http://web.expasy.org/docs/swiss-prot_guideline.html*

Farina A., Dumonceau J., Antinori P., Annessi-Ramseyer I., Frossard J., Hochstrasser D.F., Delhaye M. & Lescuyer P. (2014). Bile carcinoembryonic cell adhesion molecule 6 (CEAM6) as a biomarker of malignant biliary stenoses. *Biochimicaet Biophysica Acta (BBA) - Proteins and Proteomics* **1844**(5): 1018 – 1025.
DOI: https://doi.org/10.1016/j.bbapap.2013.06.010

Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D. & Bairoch A. (2005). Protein identification and analysis tools on the expasy server. In: *The Proteomics Protocols Handbook*, pp. 531–552. Humana Press, USA.

Gatto L. *et al*. (11 authors) (2014). A foundation for reliable spatial proteomics data analysis. *Molecular and Cellular Proteomics* **13**(8): 1937–1952.
DOI: https://doi.org/10.1074/mcp.M113.036350

Gatto L. & Lilley K.S. (2010). Towards reproducible msms data preprocessing, quality control and quantification. *Nature Proceedings*.
DOI: http://dx.doi.org/10.1038/npre.2010.5010.1

Gatto L. & Lilley K.S. (2012). MSNbase-an R/bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics* **28**(2): 288–289.
DOI: https://doi.org/10.1093/bioinformatics/btr645

Geng X. *et al*. (11 authors) (2015). iTRAQ-based quantitative proteomic analysis of the initiation of head regeneration in planarians. *PLoS ONE* **10**(7): e0132045.
DOI: https://doi.org/10.1371/journal.pone.0132045

Giardina M., Artaev S. & Ott L.W. (2010). Peptide mass fingerprinting using high resolution time of flight (hrt) mass spectrometry: An integrated approach. *Journal of Biomolecular Techniques* **21**: S57–S57.

Giulia K., Bonnie A., Darya C., Eric W., Hui W., Moraima P., Nusrat R. & Jonathan W. (2010). Methods and compositions of molecular profiling for disease diagnostics. United States patent no US20100131432A1. Available at https://patents.google.com/patent/US20100131432A1/en

Gupta N. & Pevzner P.A. (2009). False discovery rates of protein identifications: A strike against the two-peptide rule. *Journal of Proteome Research* **8**(9): 4173–4181.
DOI: https://doi.org/10.1021/pr9004794

He Q. & Chiu J. (2003). Proteomics in biomarker discovery and drug development. *Journal of Cellular Biochemistry* **89**: 868–886.
DOI: https://doi.org/10.1002/jcb.10576

Hagiwara T., Saito Y., Nakamura Y., Tomonaga T., Murakami Y. & Kondo T. (2011). Combined use of a solid-phase hexapeptide ligand library with liquid chromatography and two-dimensional difference gel electrophoresis for intact plasma proteomics. *International Journal of Proteomic*s **2011**: Article ID 739615.
DOI: https://doi.org/10.1155/2011/739615

He Z., Yang C., Yang C., Qi R.Z., Tam J.PM. & Yu W. (2009). Optimization-based peptide mass fingerprinting for protein mixture identification. In: *Research in Computational Molecular Biology* (ed. S. Batzoglou). RECOMB 2009. Lecture Notes in Computer Science, volume, pp. 5541. Springer, Berlin, Heidelberg, Germany.
DOI: https://doi.org/10.1007/978-3-642-02008-7_2

Hojo S., Fujita J., Yamadori I., Kamei T., Yoshinouchi T., Ohtsuki Y., Yamaji Y. & Takahara J. (1998). Overexpression of p53 protein in interstitial lung diseases. *Respiratory Medicine* **92**(2): 184–190.
DOI: https://doi.org/10.1016/S0954-6111(98)90093-2

Jeong K., Kim S. & Bandeira N. (2012). False discovery rates in spectral identification. *BMC Bioinformatics* **13**(16): Article number S2.
DOI: https://doi.org/10.1186/1471-2105-13-S16-S2

Jiang X., Jiang X., Han G., Ye M. & Zou H. (2007). Optimisation of filtering criterion for sequest database searching to improve proteome coverage in shotgun proteomics. *BMC Bioinformatics* **8**: Article number 323.
DOI: https://doi.org/10.1186/1471-2105-8-323

Joseph S.D. (2011). Examining the role of hedgehog signaling in the pancreatic tumor microenvironment. *PhD thesis*, University of Michigan, USA.

Krásný L., Hynek R. & Hochel I. (2013). Identification of bacteria using mass spectrometry techniques. *International Journal of Mass Spectrometry* **353**: 67–79.
DOI: https://doi.org/10.1016/j.ijms.2013.04.016

Kwissa M. *et al*. (11 authors) (2014). Dengue virus infection induces expansion of a CD14(+)CD16(+) monocyte population that stimulates plasmablast differentiation. *Cell Host Microbe* **16**(1): 115–127.
DOI: https://doi.org/10.1016/j.chom.2014.06.001

Langlois B., Perrot G., Schneider C., Henriet P., Emonard H., Martiny L. & Dedieu S. (2010). LRP-1 promotes cancer cell invasion by supporting ERK and inhibiting JNK signaling pathways. *PLoS ONE* 5(7): e11584.
DOI: https://doi.org/10.1371/journal.pone.0011584

Leconet W. *et al*. (14 authors) (2009). Preclinical validation of AXL receptor as a target for antibody-based pancreatic cancer immunotherapy. *Oncogene* **33**(47): 5405–5414.
DOI: https://doi.org/10.1038/onc.2013.487

Liang Z.L., Kim M., Huang S.M., Lee H.J. & Kim J.M. (2013). Expression of carboxyl terminus of hsp70-interacting protein (chip) indicates poor prognosis in human gallbladder carcinoma. *Oncology Letters* **5**(3): 813–818.
DOI: https://doi.org/10.3892/ol.2013.1138

Licker V., Patel V. & Ward M. (2007). Characterisation of human cerebrospinal fluid (CSF) after tandem mass tag (tmt0) labelling. *MSc thesis*, King's College London, UK.

Lorenzo E. *et al*. (11 authors) (2015). An optimisation-driven analysis pipeline to uncover biomarkers and signaling paths. *Cervix Cancer* **4**(2): 287.
DOI: https://doi.org/10.3390/microarrays4020287

Lu J., Auduong L., White E.S. & Yue X. (2014). Up-regulation of heparan sulfate 6-0-sulfation in idiopathic pulmonary fibrosis. *American Journal of Respiratory Cell and Molecular Biology* **50**(1): 106–114.
DOI: https://doi.org/10.1165/rcmb.2013-0204OC

Mahoney K.M., Rennert P.D. & Freeman G.J. (2015). Combination cancer immunotherapy and new immunomodulatory targets. *Nature Reviews on Drug Discovery* **14**(8): 1474–1776.
DOI: https://doi.org/10.1038/nrd4591

Matrix Science (2003). Mixture mode for peptide mass fingerprinting. Available at *http://www.matrixscience.com/pdf/2003WKSHP2.pdf*

McDermott J.E., Wang J., Mitchell H., Webb-Robertson B., Hafen R., Ramey J. & Rodland K.D. (2013). Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. *Expert opinion on medical diagnostics* **7**: 37–51.
DOI: https://doi.org/10.1517/17530059.2012.718329

Michaud D.S. (2004). Epidemiology of pancreatic cancer. *Journal of Biomolecular Techniques* **59**(2): 99–111.

Min-Oo G., Fortin A., Tam M., Nantel A., Stevenson M.M. & Gros P. (2003). Pyruvate kinase deficiency in mice protects against malaria. *Nature Genetics* **35**(4): 357–362.
DOI: https://doi.org/10.1038/ng1260

Muehlenbachs A., Fried M., Lachowitzer J., Mutabingwa T.K. & Duffy P.E. (2007). Genome-wide expression analysis of placental malaria reveals features of lymphoid neogenesis during chronic infection. *The Journal of Immunology* **179**(1): 557–565.
DOI: https://doi.org/10.4049/jimmunol.179.1.557

Nascimento E.J.M. (12 authors) *et al*. (2009). Gene expression profiling during early acute febrile stage of dengue infection can predict the disease outcome. *PLoS ONE* **4**(11): e7892.
DOI: https://doi.org/10.1371/journal.pone.0007892

Nehrt N.L., Peterson T.A., Park D. & Kann M.G. (2012). Domain landscapes of somatic mutations in cancer. *BMC Genomics* **13**: S9.
DOI: https://doi.org/10.1186/1471-2164-13-S4-S9

Neuhauser N., Nagaraj N., McHardy P., Zanivan S., Scheltema R., Cox J. & Mann M. (2013). High performance computational analysis of large-scaleproteome datasets to assess incremental contribution to coverage of the human genome. *Journal of Proteome Research* **12**(6): 2858–2868.
DOI: https://doi.org/10.1021/pr400181q

Nguetcheu S.C., Khun H., Pincet L., Roux P., Bahut M., Huerre M., Guette C. & Choumet V. (2010). Differential protein modulation in midguts of *Aedes aegypti* infected with chikungunya and dengue 2 viruses. *PLoS ONE* **5**(10): 99–111.
DOI: https://doi.org/10.1371/journal.pone.0013149

Numata M. & Oshima T. (2012). Significance of regenerating islet-derived type iv gene expression in gastroenterological cancers. *World Journal of Gastroenterology* **18**(27): 3502–3510.
DOI: https://doi.org/10.3748/wjg.v18.i27.3502

O'Farrell P.H. (1975). High resolution two-dimensional electrophoresis of proteins. *Journal of Biological Chemistry* **250**(10): 4007–4021.
DOI: https://doi.org/10.1016/S0021-9258(19)41496-8

Oliver T., Schmidt B., Nathan D., Clemens R. & Maskell D. (2005). Using reconfigurable hardware to accelerate multiple sequence alignment with ClustaIW. *Bioinformatics* **21**(16): 3431–3432.
DOI: https://doi.org/10.1093/bioinformatics/bti508

Pacific Northwest National Laboratory (2011). Protein digestion simulator. Available at *https://omics.pnl.gov/software/protein-digestion-simulator*

Paulette M.T., Hideki F., Yoshiko S. & Toshihiko K. (2014). Sphingosine kinase 1 mediates head and neck squamous cell carcinoma invasion through sphingosine 1 - phosphate receptor 1. *Cancer Cell International* **14**: 76.
DOI: https://doi.org/10.1186/s12935-014-0076-x

Perkins D.N., Pappin D.J., Creasy D.M. & Cottrell J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**(18): 3551–3567.

Philip H.W. (2008). Evaluating cancer protein identification from mass spectroscopy data. *PhD thesis*, University of Arkansas at Little Rock, USA.

Phoenix Bioinformatics (2011). The arabidopsis information resource. Available at *https://www.arabidopsis.org/*

Pilarsky C. *et al*. (20 authors) (2008). Activation of Wnt signalling in stroma from pancreatic cancer identified by gene expression profiling. *Journal of Cellular and Molecular Medicine* **12**(6B): 2823–2835.
DOI: https://doi.org/10.1111/j.1582-4934.2008.00289.x

Pratt J.M. *et al*. (11 authors) (2002). Stable isotope labelling in vivo as an aid to protein identification in peptide mass fingerprinting. *Proteomics* **2**(2): 157–163.

Pruitt K. D., Tatusova T. & Maglott D. R. (2005). NCBI reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **33**: D501–D504.
DOI: https://doi.org/10.1093/nar/gki025

Puttamallesh V.N., Sreenivasamurthy S.K., Singh P.K., Harsha H.C., Ganjiwale A., Broor S., Pandey A., Narayana J. & Prasad T.S.K. (2013). Proteomic profiling of serum samples from chikungunya-infected patients provides insights into host response. *Clinical Proteomcs* **10**: Article number 14.
DOI: https://doi.org/10.1186/1559-0275-10-14

Risk B.A., Edwards N.J. & Giddings M.C. (2013). A peptide-spectrum scoring system based on ion alignment, intensity, and pair probabilities. *Journal of Proteome Research* **12**(9): 4240–4247.
DOI: https://doi.org/10.1021/pr400286p

Sahasrabuddhe N.A. *et al*. (15 authors) (2014). Identification of prosaposin and transgelin as potential biomarkers

for gallbladder cancer using quantitative proteomics. *Biochemical and Biophysical Research Communications* **446**(4): 863–869.
DOI: https://doi.org/10.1016/j.bbrc.2014.03.017

Saisawang C., Sillapee P., Sinsirimongkol K., Ubol S., Smith D.R. & Ketterman A.J. (2015). Full length and protease domain activity of chikungunya virus nsp2 differ from other alphavirus nsP2 proteases in recognition of small peptide substrates. *Bioscience Reports* **35**(3): e00196.
DOI: https://doi.org/10.1042/BSR20150086

Sandulache V.C. *et al*. (10 authors) (2012). Individualizing antimetabolic treatment strategies for head and neck squamous cell carcinoma based on TP53 mutational status. *Cancer* **118**(3): 711–21.
DOI: https://doi.org/10.1002/cncr.26321

Sercundes M.K., Ortolan L.D., Debone D., Aitken E.H., Alvarez J.M., Russo M., Marinho C.R. & Epiphanio S. (2013). Inflammatory factors and leucocytes are involved in the pathogenesis of malaria associated acute lung injury/acute respiratory distress syndrome in murine model. *Frontiers in Immunology Conference Abstract: 15th International Congress of Immunology (ICI)*.
DOI: https://doi.org/10.3389/conf.fimmu.2013.02.00004

Shah A.R. *et al*. (11 authors) (2010). An efficient data format for mass spectrometry-based proteomics. *Journal of the American Society for Mass Spectrometry* **21**(10): 1784 – 1788.
DOI: https://doi.org/10.1016/j.jasms.2010.06.014

Srinivas P.R., Verma M., Zhao Y. & Srivastava S. (2002). Proteomics for cancer biomarker discovery. *Clinical Chemistry* **48**: 1160–1169.
DOI: https://doi.org/10.1093/clinchem/48.8.1160

Sandmann T., Vogg M.C., Owlarn S., Boutros M. & Bartscherer K. (2011). The head-regeneration transcriptome of the planarian Schmidtea mediterranea. *Genome Biology* **12**(8).
DOI: https://doi.org/10.1186/gb-2011-12-8-r76

Syed N. *et al*. (18 authors) (2015). Silencing of high-mobilitygroup box 2 (HMGB2) modulates cisplatin and 5-fluorouracil sensitivity in head and neck squamous cell carcinoma. *Proteomics* **15**: 383–393.
DOI: https://doi.org/10.1002/pmic.201400338

Tamura T., Akbari M., Kimura K., Kimura D. & Yui K. (2014). Flt3 ligand treatment modulates parasitemia during infection with rodent malaria parasites via myd88and ifn-γ- dependent mechanisms. *Parasite Immunology* **36**(2): 87–99.
DOI: https://doi.org/10.1111/pim.12085

Thalappilly S., Suliman M., Gayet O., Soubeyran P., Aurélie Hermant, Lecine P., Iovanna J.L. & Dusetti N.J. (2008). Identification of multi-sh3 domain-containing protein interactome in pancreatic cancer: A yeast two-hybrid approach. *Proteomics* **8**(15): 3071–3081.
DOI: https://doi.org/10.1002/pmic.200701157

Thivierge C., Kurbegovic A., Couillard M., Guillaume R., Coté O. & Trudel M. (2006). Overexpression of PKD1 causes polycystic kidney disease. *Molecular and Cellular Biology* **26**(4): 1538–1548.
DOI: https://doi.org/10.1128/MCB.26.4.1538-1548.2006

Vidanagamachchi S.M., Dewasurendra S.D., Ragel R.G. & Niranjan M. (2014). A structured hardware software architecture for peptide based diagnosis-sub-string matching problem with limited tolerance. *7th International Conference on Information and Automation for Sustainability*, Sri Lanka, pp. 1–7.
DOI: https://doi.org/10.1109/ICIAFS.2014.7069624

Vidanagamachchi S.M. & Niranjan M. (2017). OPBI: An Open Pipeline for Biomarker Identification. *IEEE International Conference on Industrial Engineering and Engineering Management*, Singapore, pp. 1510–1514.
DOI: https://doi.org/10.1109/IEEM.2017.8290145

UniProt (2017). UniRef. Available at http://www.uniprot.org/help/uniref , Accessed 15 January 2018.

Walhout A.J.M, Vidal M. & Dekker J. (2012). *Handbook of Systems Biology: Concepts and Insights*. Academic Press, USA.

Wienbrandt L. (2014). The FPGA-based high-performance computer RIVYERA for Applications in bioinformatics. *Lecture Notes in Computer Science* **8493**: 383–392.
DOI: https://doi.org/10.1007/978-3-319-08019-2_40

Wilkins M.R., Lindskog I., Gasteiger E., Bairoch A., Sanchez J.C., Hochstrasser D.F. & Appel R.D. (1997). Detailed peptide characterisation using peptidemass - a world-wide web accessible tool. *Electrophoresis* **18**: 403–408.
DOI: https://doi.org/10.1002/elps.1150180314

Xu Y., Zhu F., Xu S. & Liu L. (2015). Anti-tumor effect of the extract from qingyihuaji formula on pancreatic cancer by down-regulating notch-4 and jagged-1. *Journal of Traditional Chinese Medicine* **35**(1): 77–83.
DOI: https://doi.org/10.1016/S0254-6272(15)30012-1

Yang Z., Yang Z., Zou Q., Yuan Y., Li J., Li D., Liang L., Zeng G. & Chen S. (2014). A comparative study of clinicopathological significance, FGFBP1, and WISP-2 expression between squamous cell/adenosquamous carcinomas and adenocarcinoma of the gallbladder. *International Journal of Clinical Oncology* **19**(2): 325–335.
DOI: https://doi.org/10.1007/s10147-013-0550-9

Yuda M., Yui K., Urban J.F., Tamura T. & Kimura K. (2011). Prevention of experimental cerebral malaria by flt3 ligand during infection with plasmodium bergheianka. *Infection and Immunity* **79**(10): 3947–3956.
DOI: https://doi.org/10.1128/IAI.01337-10

Zhang B., Ji L.H., Liu W., Zhao G. & Wu Z-Y. (2013). Skp2-RNAi suppresses proliferation and migration of gallbladder carcinoma cells by enhancing p27 expression. *World Journal of Gastroenterology* **19**(30): 4917–4924.
DOI: https://doi.org/10.3748/wjg.v19.i30.4917

Zhang J. *et al*. (12 authors) (2001). The SOCS box of suppressor of cytokine signaling-1 is important for inhibition of cytokine actionin vivo. *Proceedings of the National Academy of Sciences* **98**(23): 13261–13265.
DOI: https://doi.org/10.1073/pnas.231486498

Zhou C. *et al*. (11 authors) (2012). Statistical considerations of optimal study design for human plasma proteomics and biomarker discovery. *Journal of Proteome Research* **11**: 2103–2113.
DOI: https://doi.org/10.1021/pr200636x