Genes, Genomes and Genomics ©2012 Global Science Books

Peptide Identification by Searching Large-Scale Tandem Mass Spectra against Large Databases: Bioinformatics Methods in Proteogenomics

Mohamed Helmy[1,2] • Masaru Tomita[1] • Yasushi Ishihama[1,3*]

[1] Institute for Advanced Biosciences, Keio University, Japan
[2] Systems Biology Program, Graduate School of Media and Governance, Keio University, Japan
[3] Graduate School of Pharmaceutical Sciences, Kyoto University, Japan

Corresponding author: * **y-ishi@ttck.keio.ac.jp**

ABSTRACT

Mass spectrometry-based shotgun proteomics approaches are currently considered as the technology-of-choice for large-scale proteogenomics due to high throughput, good availability and relative ease of use. Protein mixtures are firstly digested with protease, *e. g.* trypsin, and the resultant peptides are analyzed using liquid chromatography - tandem mass spectrometry. Proteins and peptides are identified from the resultant tandem mass spectra by *de novo* interpretation of the spectra or by searching databases of putative sequences. Since this data represents the expressed proteins in the sample, it can be used to infer novel proteogenomic features when mapped to the genome. However, high-throughput mass spectrometry instruments can readily generate hundreds of thousands, perhaps millions, of spectra and the size of genomic databases, such as six-frame translated genome databases, is enormous. Therefore, computational demands are very high, and there is potential inaccuracy in peptide identification due to the large search space. These issues are considered the main challenges that limit the utilization of this approach. In this review, we highlight the efforts of the proteomics and bioinformatics communities to develop methods, algorithms and software tools that facilitate peptide sequence identification from databases in large-scale proteogenomic studies.

Keywords: bioinformatics, genome annotation, mass spectrometry, proteogenomics, proteomics
Abbreviations: ABLCP, algorithm based on longest common prefix; **EST**, expressed sequence tag; **IEF**, isoelectric focusing; **LC-MS/MS**, liquid chromatography – tandem mass spectrometry; **LSH**, locality-sensitive hashing; **ORF**, open reading frame; **p*I***, Isoelectric point; **PST**, peptide sequence tag; **PTM**, post-translational modifications; **QTOF**, quadrupole time-of-flight; **SNP** single-nucleotide polymorphism; **SCX**, strong cation exchange

CONTENTS

INTRODUCTION

Proteomics aims to characterize the expressed proteins and the corresponding peptides in a given sample, including elucidation of their sequence, structure and function (Tyers and Mann 2003). Thus, proteome level investigation represents a rich source of information that complements the traditional genome analysis process. It become generally acknowledged that the inclusion of proteome data in the genome analysis results in better genome annotation in so-called proteogenomics (Ansong *et al.* 2008a; Armengaud 2009; de Groot *et al.* 2009; Armengaud 2010; Castellana and Bafna 2010). Proteogenomics is the utilization of large-scale proteome data in genome annotation refinement (Ansong *et al.* 2008a). Due to their high throughput and accurate measurement of the peptides, high-throughput mass spectrometry-based proteomics methods, such as liquid chromatography – tandem mass spectrometry (LC-MS/MS), can provide a rich source of translational-level expression evidence to support the predicted protein-coding genes. This approach seems the best option for identification and confirmation of the protein-coding genes, or at least significant portion of them, in an independent and unambiguous way (Ansong *et al.* 2008a). This can be achieved by detecting the naturally occurring proteins (proteomics) and mapping them back to the genome sequence (genomics) in a systematic analysis, as presented in several recent reports (Ishino *et al.* 2007; Baerenfaller *et al.* 2008; Merrihew *et al.*

Received: 26 January, 2011. Accepted: 19 May, 2011

Invited Review

2008; Bringans *et al.* 2009; Payne *et al.* 2010).

For instance, *Arabidopsis thaliana* is the most studied plant and has the most thoroughly sequenced and annotated genome among plants. However, proteogenomics provided significant additions and corrections to its genome annotation (Baerenfaller *et al.* 2008; Castellana *et al.* 2008). A genome-scale proteomics study, with intensive sampling of several organs and life stages and over 1,300 MS/MS runs, added 57 new gene models to the *Arabidopsis* annotation, providing expression evidence for all of them. Moreover, the same study provided functional annotation by flagging the proteins that were expressed in one organ only as biomarkers (Baerenfaller *et al.* 2008). In another proteogenomics study, nearly 13% of the *Arabidopsis* proteome was deemed incorrect or missing due to incorrect or missing gene models through the identification of 778 new protein-coding genes and correction of 695 gene models (Castellana *et al.* 2008). These significant improvements in the best annotated plant genome, the *Arabidopsis* genome, demonstrate the value of proteogenomics approaches in improving genome annotation and indicate their potential for expanding incompletely annotated genomes.

The utility of proteogenomics in achieving significant improvements in genome annotation has been shown in various eukaryotic and prokaryotic genomes. Several reports have presented novel genomic information obtained via large-scale mass spectrometry-based proteogenomics in human. Desiere *et al*. (2005) demonstrated large-scale integration between peptides obtained through high-throughput proteomics and the human genome (Desiere *et al.* 2005). Power *et al*. (2009) found novel splice isoforms in human platelets by using a proteomics approach to identify the exon-skipping events (Power *et al.* 2009). The *Caenorhabditis elegans* (*C. elegans*) genome annotation was identified, corrected and confirmed using shotgun proteomics by identifying 429 unannotated coding sequences (including 33 pseudogenes), 151 errors in gene models and 254 novel gene models (Merrihew *et al.* 2008). The same approach is also applicable to the genomes of major plant crops such as rice (Helmy *et al.* 2011), fungi such as *Aspergillus niger* (the black mold fungus) (Wright *et al.* 2009), parasites such as *Plasmodium falsiparum* (the malaria parasite) and *Toxoplasma gondii* (Lasonder *et al.* 2002; Xia *et al.* 2008), insects such as *Drosophila melanogaster* (Tress *et al.* 2008; Loevenich *et al.* 2009), nematodes such as *Pristionchus pacificus* (Borchert *et al.* 2010) and Archaea such as *Thermococcus gammatolerans* (Zivanovic *et al.* 2009).

The basic outcome of a proteogenomic analysis is to validate the predicted gene models at the translational level, as presented in several reports (Jaffe *et al.* 2004a, 2004b; Wang *et al.* 2005). In addition, proteogenomics has been utilized to reveal many other significant genomic features, e.g., finding new gene models (Jaffe *et al.* 2004b; Baerenfaller *et al.* 2008; Castellana *et al.* 2008; Merrihew *et al.* 2008), determination of the protein start and termination sites (Nielsen and Krogh 2005; Mattow *et al.* 2007; Tanner *et al.* 2007), finding and verifying splice isoforms at the protein level (Tanner *et al.* 2007; Mo *et al.* 2008; Power *et al.* 2009) and verifying hypothetical and conserved hypothetical genes/proteins (Kolker *et al.* 2004; Hixson *et al.* 2006; Tanner *et al.* 2007; Ansong *et al.* 2008b). With such efficiency in the improvement of genome annotation, proteogenomics represents a promising approach to be applied to newly sequenced genomes, as well as for use in the primary annotation of the genome, rather than only to improve the annotation at a later stage (de Groot *et al.* 2009).

In addition to finding novel genomic features to refine the genome annotation, proteogenomics can be applied in biomarker discovery (Baerenfaller *et al.* 2008; Sigdel and Sarwal 2008), for identification of antibody targets (Huang *et al.* 2004), to provide a better understanding of the host-parasite relationship (Lasonder *et al.* 2002; Bindschedler *et al.* 2009; Delmotte *et al.* 2009) and to understand the mechanisms of ecological diversity and environmental adaptation (de Groot *et al.* 2009; Denef *et al.* 2010). Further, proteogenomic studies have been performed on several genomes of related species to identify rare post-translational modifications (Gupta *et al.* 2008), to investigate adaptive mutation capabilities among species (Bechah *et al.* 2010) and to understand diversity-shaping events between species (Nouvel *et al.* 2010). Proteogenomics also provides novel insight in cancer research, either in finding biomarkers, related somatic mutations or diagnosis (Jacob *et al.* 2009; Helmy *et al.* 2010).

LARGE-SCALE PROTEOGENOMICS WORKFLOW

Usually, large-scale proteome analyses are required for proteogenomics projects. In a typical proteogenomics project (**Fig. 1**), high-throughput proteomics (usually mass spectrometry-based) is used to obtain the sequences of the peptides of the sample in question (Castellana and Bafna 2010). The peptides are obtained through proteolytic digestion of the extracted proteins with proteases such as trypsin and Lys-C. The peptides are later pre-fractionated using several fractionation methods, such as strong cation exchange-StageTips (SCX-StageTips) (Ishihama *et al.* 2006) and isoelectric focusing (IEF) (Cargile *et al.* 2004), then the fractions are processed and prepared for mass spectrometry (MS/MS) analysis. The MS/MS analysis results in thousands or even millions of MS/MS spectra that hold the peptide fingerprints. The peptide sequences corresponding to the MS/MS spectra are later identified using i) *de novo* interpretation of the spectra (Seidler *et al.* 2010) or ii) database search of putative sequences (Sadygov *et al.* 2004; Kapp and Schutz 2007).

After obtaining the peptide sequences, proteogenomic analysis is conducted, but the details of the analysis are different according to the availability of the genome annotation (newly sequenced genome or annotated genome), the genome complexity (prokaryotic or eukaryotic genome) and the available informatics tools. However, here we will describe general steps that are shared in a wide range of proteogenomics projects.

1) If the genome sequence and genome annotation of the organism are available: the peptides are mapped to the genome using several computational methods, mostly sequence alignment tools such as different types of BLAST (Altschul *et al.* 1990, 1997; Tatusova and Madden 1999). Then, the alignment results can be compared with the current annotation to confirm and improve the annotated gene models (Kalume *et al.* 2005; Xia *et al.* 2008; Payne *et al.* 2010) or to perform whole genome re-annotation using gene-finding tools that use the peptide information as hints to improve the predictive capability, such as AUGUSTUS (Stanke *et al.* 2008).

2) If the genome is newly sequenced: the proteome information (peptide sequences) is included in the primary genome annotation process. So far, there are two examples of proteogenomics integration in the primary annotation of a newly sequenced genome, the *Mycoplasma mobile* genome annotation (Jaffe *et al.* 2004b) and the *Deinococcus deserti* genome annotation (de Groot *et al.* 2009).

One of the key points in the workflow is to identify the amino acid sequences corresponding to the MS/MS spectra accurately and efficiently, since the accuracy of the remaining steps is highly dependent on these sequences. Although database searching is considered the most reliable approach to identify peptide sequences from MS/MS spectra, and is the most widely used, it represents a bottleneck in large-scale proteogenomics, especially when large databases are employed (Zhou *et al.* 2010). Most proteogenomic studies have used the six-frame translation of the genome database for peptide sequence identification. Searching large-scale MS/MS data against such database is not a trivial task due to the enormous size of both the MS/MS data set and the database, and the linear relationship between search time and database size (Edwards 2007). Thus, the computational demands for such search are in some cases

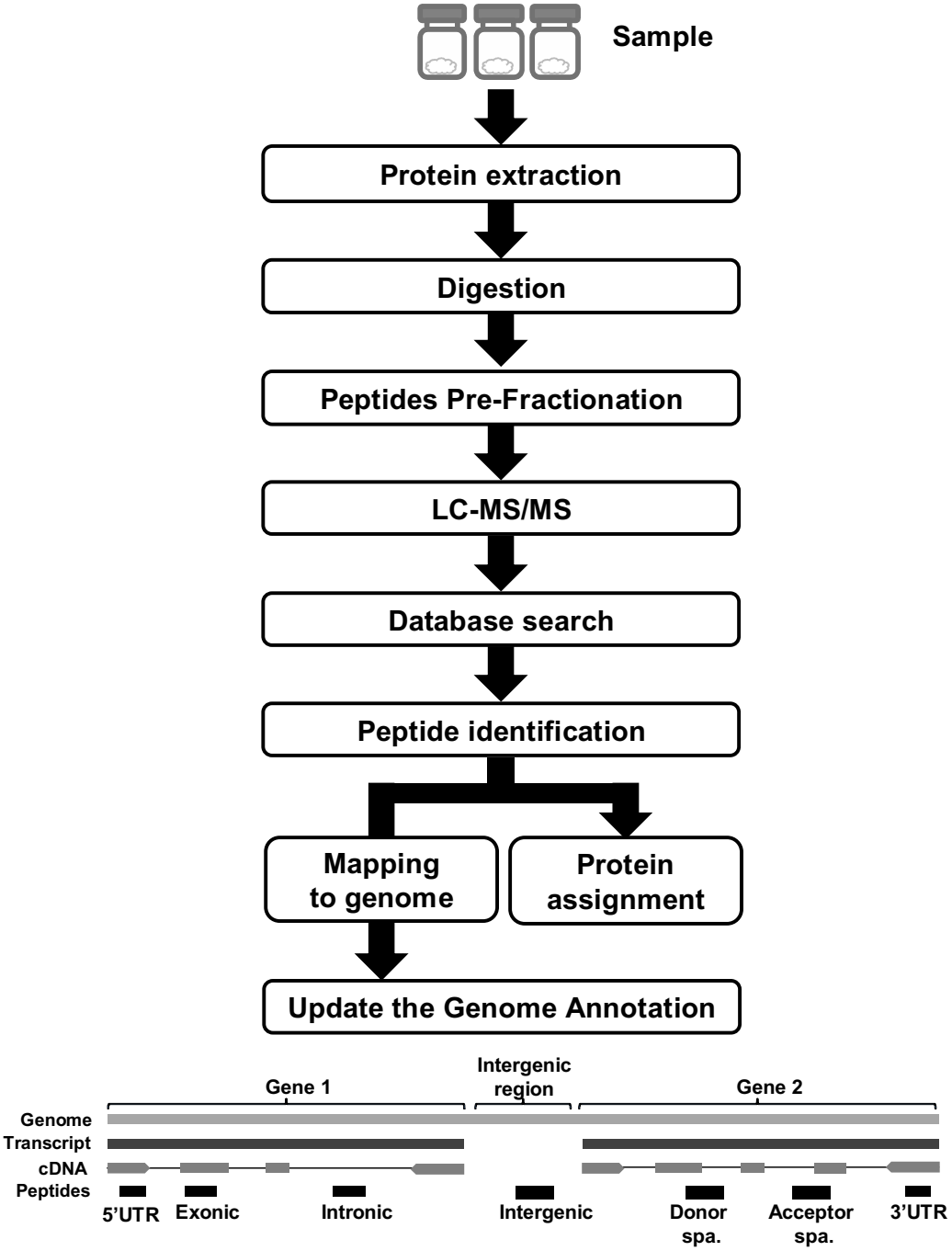

Fig. 1 Simplified representation of large-scale proteogenomics workflow. Proteins are extracted from the sample and digested using suitable protease(s), e.g. trypsin, then the resultant peptides are pre-fractionated. The fractions are then processed and submitted to LC-MS/MS. The obtained MS/MS spectra are searched against putative sequence databases using search engines such as Mascot and SEQUEST to identify peptide sequences. The sequences are later mapped to the genome to confirm or update the genome annotation.

so high as to be unaffordable. For example, 260 days of CPU time was required to run over 4,000 X!Tandem (Craig and Beavis 2003; Craig and Beavis 2004) searches against the *Shewanella* genome database (Turse *et al.* 2010), even though a PRISM computing cluster with 32 processing nodes was used (Kiebel *et al.* 2006).

THE NEED FOR FASTER DATABASE SEARCHING METHODS

Several database searching programs and algorithms are currently available, including SEQUEST (Eng *et al.* 1994), Mascot (Perkins *et al.* 1999), X!Tandem (Craig and Beavis 2003; Craig and Beavis 2004) and several other tools and algorithms, including pFind (Wang *et al.* 2007), OMSSA (Geer *et al.* 2004), PepSplice (Roos *et al.* 2007), Phenyx (Colinge *et al.* 2003), PEAKS (SPIDER) (Ma *et al.* 2003) SpectrumMill (Agilent Technologies, CA), ProteinPilot (AB-Sciex, CA) and Crux (Park *et al.* 2008). These programs use different approaches to facilitate the peptide sequences identification from databases. However, these tools seem to be insufficient for proteogenomics application for the following reasons:

1) The continuous expansion of the protein databases: for instance, the protein sequences in the IPI.Human database increased by ~30% from IPI.Human V3.22 to IPI.Human V3.49 (Li *et al.* 2010), while the size of the NCBInr protein sequence database doubled in about 18 months (Zhou *et al.* 2010).

2) Using the six-frame translation of the genome database and the genome-translated proteins: these genomic databases are more suitable for proteogenomics and recent ad-

vances in genome sequencing techniques have made such databases easily available. However, the size of the six-frame translation or the EST library that results from such database limits its utility. For example, the six-frame translation of the human genome database is over 6 Gbp and the EST library that results from its translation contains over 8 million protein sequences (over 100 times more than the human proteome) (Li *et al.* 2010).

3) The inclusion of chemical and post-translational modifications (PTMs): these modifications produce more peptides. For example, Zhou *et al*. (2010) showed that the inclusion of three variable post-translational modifications and up to two missed cleavage sites increased the number of peptides by ~ 38-fold, compared with the number that would result from a fully specific digestion of the IPI.Human database (Zhou *et al.* 2010).

4) Using semi-specific or non-specific digestion: in these cases, the number of peptides to be considered is increased by 10- to over-100 fold, respectively, comparing with specific digestion. For instance, the fully non-specific digestion of the IPI.Human database V3.65 (Kersey *et al.* 2004) resulted in a 170-fold increase in the non-redundant peptides, compared with fully specific digestion (Zhou *et al.* 2010).

5) The continuous development in mass spectrometry instruments: this has resulted in a remarkable increase in the generation rate of tandem mass spectra. An LTQ mass spectrometer from Thermo Fisher Scientific, for instance, can generate over 430,000 spectra per day while an LTQ Velos instrument, the newest LTQ model, can generate double this amount per day (Li *et al.* 2010; Zhou *et al.* 2010).

6) The steady development of computer hardware: this still remains a step behind the development of mass spectrometry and genome sequencing techniques. We can see a common pattern in the development of the pioneering database-searching programs, in that they follow the computer hardware development and try to make use of newly presented features to speed up database search (Kapp and Schutz 2007). For instance, the original implementation of SEQUEST performed the analysis sequentially, but was not multi-threaded and could not take advantage of multi-core CPUs (Eng *et al.* 1994). However, in later versions, SEQUEST was developed to be able to include modifications (Yates *et al.* 1995b), search EST databases (Yates *et al.* 1995a), and search high-energy CID data (Yates *et al.* 1996); in parallel to these proteomics-related updates, a cluster version was introduced to make use of multi-core/ multi-computer systems (Kapp and Schutz 2007). Mascot and X!Tandem are multi-threaded and can take advantage of multiple CPUs with multiple cores in the same machine. Further, they both have cluster versions that can make use of multiple computers to perform the database search (Kapp and Schutz 2007). However, all these computational and hardware developments have failed to keep pace with the rapid developments of the analytical instruments.

METHOD DEVELOPMENT TO SPEED UP DATABASE SEARCHING

The need for continuous method development to improve the efficiency of peptide identification at reasonable computational cost is driven by the factors mentioned above. Thus, researchers in the proteomics and bioinformatics communities have developed several methods to facilitate the searching process. However, four aspects should be taken into consideration while developing new methods. 1) Significant reduction of the search time and computational demand is needed, since this is the main aim of developing improved methods. 2) The capability for identifying peptides should be similar to or better than that of the current methods. 3) The accuracy should not be affected, unless it is improved. 4) The method should be flexible and have the ability to be integrated in different analysis workflows (Zhou *et al.* 2010).

In this survey, we review the efforts of scientists to speed up peptide identification from large databases during the last decade. Since the identification process involves three main players, the database, the searching algorithm and the experimental proteome data, methods are usually focused on one of these players to speed up the search process. A wide range of methods has been proposed, including database preprocessing (indexing, reduction or splitting), developing faster search algorithms, reducing the number of spectra to be searched, using hybrid methods that combine *de novo* spectra interpretation with database searching and even directly involving computer hardware and low-level programming. Within the scope and space of this review, we present several methods from each category, trying to cover all approaches.

DATABASE PREPROCESSING METHODS

Methods in this category are mainly concerned with preprocessing the databases in order to restrict the peptide search space. Restricting the search space leads to a reduction of the search time and the required resources to perform the search. Database preprocessing methods can be categorized into three sub-categories based on the type of preprocessing.

Database indexing

Database indexing, also known as peptide indexing, is one of the most widely used methods to facilitate peptide identification (Li *et al.* 2010). Several search engines use database indexing approaches, such as SEQUEST (Eng *et al.* 1994), pFind (Wang *et al.* 2007) and Crux (Park *et al.* 2008). Through the normal peptide identification process, the database is preprocessed by performing *in silico* digestion of the entire database contents and the resultant *in silico* peptides are then matched to the spectra (Dutta and Chen 2007). However, the digested proteins produce redundant peptides which increase the total number of peptides and result in redundant matching and scoring. Recently, it was shown that redundant peptides represent about 50% of the total peptides (Li *et al.* 2010). Therefore, database indexing methods remove the redundancy and index the peptides with their mass. Then, for a given spectrum with precursor ion mass m and precursor ion tolerance Δ, the searching program selects from index peptides within the range $[m-\Delta \sim m+\Delta]$. Thus, indexing allows the program to avoid the *one-against-all* comparison, thereby reducing search time and computational resources (Dutta and Chen 2007). Database indexing methods can be divided into three main models.

1. Off-line indexing

In this approach, the digestion and indexing are performed once and the index is saved to the disk. The search program only needs to load the index from the disk and start searching. Obviously, this reduces the search time and the required processing power. However, the index is static and any change in the search parameters, e.g. modifications, requires reconstruction of the index (Zhou *et al.* 2010).

2. On-line indexing

In *on-line* indexing, the index is created *on-the-fly* after the input of the search parameters. This dynamic indexing overcomes the disadvantage of *off-line* indexing, but it requires a longer time, since the search program needs to construct the index ahead of each search. Further, if the spectra are submitted in batches, a new index is constructed each time, even though the searching parameters are the same, and this redundancy increases search time and required processing power (Li *et al.* 2010).

3. Hybrid indexing

There is a hybrid method that combines *off-line* and *on-line* indexing, that is used by pFind (Wang *et al.* 2007). pFind constructs an *off-line* index for the digested peptides and an *on-line* index for modified peptides. The construction of the *on-line* index with pFind was shown to take only 5% of the total identification time (Li *et al.* 2010). Recently, Li *et al*. (2010) performed systematic investigation of all the peptide sequence identification steps performed by the first generation of search engines (engines that perform direct mapping of the MS/MS spectra to peptides without any interpretation of the spectra), such as SEQUEST, Mascot, X!Tandem and pFind. These steps include the *in silico* digestion of the proteins, peptide modification, peptide-precursor matching and fragment ion-peak matching. They were able to get a 5-fold increase in the identification speed compared to SEQUEST 2.7 by constructing two indexes: i) peptide index and ii) precursor and fragment ion index. SEQUEST was used for the comparison of index structure, construction and querying, since Mascot and X!Tandem do not use database indexing, while the efficiency was compared among the three of them. The new approach was implemented using pFind. The first index, the peptide index, speed up the identification by 2~3 times, while the other index, precursor and fragment ion index, added another two times (Li *et al.* 2010).

Database reduction

One of the most widely used methods to reduce the peptide identification time is database reduction. In these methods, a certain part of the database (that contains certain genomic features such as exons or open reading frames (ORFs)) is used for the identification, while the remaining portion is omitted. Database reduction methods significantly reduce the size of the database and, consequently, the search time. However, notable features that may be included in the omitted portion of the databases are lost. Several implementations of this approach have been presented in recent proteogenomic projects.

1. Exon graph

Exon graph methods aim to construct a compact representation of the database while covering all splice variants in all genes. Tanner *et al*. (2007) used exons and introns derived from GeneID and ESTmapper and since the putative exons and Expressed Sequence Tags (EST) predicted by gene prediction algorithms are from different lengths with overlaps, they compared them and merged them into larger intervals. If the interval overlaps an intron, the interval is split into two sub-intervals at the junction point. Then, edges are added between the adjacent intervals. Polymorphism was also incorporated in the graph by adding an interval for each allele if the interval contains a single-nucleotide polymorphism (SNP). To remove nodes corresponding to wrong mappings of reading frames, nodes and edges that are not part of a coding sequence of length 50 or more were removed. Thus, each exon graph node has a protein sequence and possibly an untranslated prefix or suffix. The final exon graph database size was significantly reduced from the original two billion amino acid residues (EST database) and 630 million residues (GeneID predicted exons) to 134 million residues. Searching the exon graph database using 18.5 million MS/MS spectra, the authors were able to validate exons and introns, confirm hypothetical proteins and discover alternative splicing events and extensions in known genes (Tanner *et al.* 2007).

Castellana *et al*. (2008) used the exon graph method to create an exon splice-graph database for *Arabidopsis*. The exon splice-graph database together with the six-frame translation of the *Arabidopsis* genome database and the annotated protein database (TAIR7 database) were used to identify 18,024 novel peptides from 21 million MS/MS spectra obtained from *Arabidopsis* proteins. The novel peptides were used to refine the *Arabidopsis* genome annotation, yielding 778 new protein-coding genes and updating 695 known gene models (Castellana *et al.* 2008).

2. Sophisticated sequence database comparison strategy to search EST databases

Aiming to identify peptides from alternative splicing isoforms and coding SNP proteins, Edwards suggested searching the ESTs. However, the enormous size of the EST database makes this proposal computationally infeasible. Edwards developed a sophisticated sequence database comparison strategy that resulted in a 35-fold reduction of the database size, making the identification of high-throughput MS/MS data from the EST database possible. This strategy requires the EST sequence to be mapped to the vicinity of a known gene, while the peptides are required to be contained in a 30-amino-acid ORF. Further, all peptides should be confirmed by at least two ESTs and the peptide sequence representation should avoid repetition. Applying this strategy to the human EST database reduced its size to less than 3% of the six-frame translation of the human genome database, thereby making this search possible using standard methods. The search time for the same MS/MS dataset and the same engine (Mascot) against the six-frame translation of the human genome database requires 22 hours, while this was reduced to 15 min with the reduced EST database. Furthermore, the results were similar, with some noticeable improvements in the second search (Edwards 2007).

3. Metric embedding and fast near-neighbor search approach

To avoid the *one-against-all* comparison that is used by most search engines, such as SEQUEST and Mascot, Dutta and Chen (2007) developed a novel method to preprocess the databases and create a limited subset of candidate peptides for a given query spectrum. This was achieved through designing a set of hash functions, where a random spectrum is used for the construction of the hash function and the normalized shared peak count score between the random spectrum and the hypothetical spectrum of a peptide is used as a peptide value. High–dimensional metric space (Euclidian space) and set of hash functions, constructed using random vectors, called locality-sensitive hashing (LSH), were used to implement the preprocessing, filtration and mapping. The method showed good accuracy: more than 95.6% of the spectra were filtered without missing any correct sequence. The filtration percentage reached 99.6% with minor loss of correct sequences (0.19%). The speed was increased 111 times in the case of 99.6% filtration, representing a remarkable speeding-up capability of this method. Further, the authors demonstrated additional applications for this method, such as accurate and efficient clustering of the MS/MS spectra (Dutta and Chen 2007).

4. SkipE

The SkipE method was developed to identify novel alternative splicing isoforms through identifying exon-skipping events, which are considered the most common form of alternative splicing. An exon-skipping event can be identified from peptides spanning the exon-exon junction of non-contiguous exons. The SkipE database was constructed through the creation of a list of all theoretical non-contiguous junction peptides from the human genome database based on the full-length transcript. Only the junction peptides were kept, while the preceding and following sequences were removed based on the last and first tryptic site upstream and downstream from the junction, respectively. Finally, duplications were removed, leaving only ~300,000 peptides, which is a suitable size for searching with standard methods. The MS/MS data of human platelets was used against the SkipE database and the International Protein Index Database (IPI.Human) (Kersey *et al.* 2004), yielding 89 genes with

alternative splicing isoforms; many of them were confirmed at the mRNA level using RT-PCR and sequencing of the products (Power *et al.* 2009).

5. Exon-exon junction database

Mo *et al*. (2008) presented the Exon-Exon junction method, which is similar to the SkipE method, to identify peptides spanning the exon-exon junctions. The identification of these peptides from the genome database is not possible, since the exon-exon junctions are separated by introns. Thus, Mo *et al*. (2008) used the Ensemble core database and its APIs and wrote scripts using perl, Bioperl and mySQL to construct a database of all putative exon-exon junction proteins covering all possible combinations of exons for each gene. The duplications and the previously described exon-exon junction events were removed, resulting in a final database with 873,024 entries and total size of 132 Mb. Using the MS/MS dataset of the human liver and two search engines (SEQUEST and X!Tandem), they were able to identify 488 non-redundant putative exon-skipping events (Mo *et al.* 2008).

Database splitting

A simple and straight forward approach that allows searching the whole of large databases, including all features without reduction, is the database splitting approach. In this approach, a large database, such as a six-frame translation of the human genome database, is simply split into a set of smaller databases (*e.g.* one database per chromosome), resulting in 24 separate databases in the case of the human genome, for instance. Then, the MS/MS data is searched against each database separately. Clearly, this consumes more time and resources. However, it is the only method that allows searching large dataset of peptide spectra against the whole of a large-sized database with reasonable computational resources and without prior reduction of the database or interpretation of the MS/MS spectra. Database splitting has been used to find novel genomic features in several proteogenomic projects.

To identify novel ORFs, Fermin *et al*. (2006) developed an approach based on creating a library of all possible ORFs in the human genome. The sequences of the ORFs in the library were obtained through complete six-frame translation of each chromosome of the human genome. The final library contained 217,305,234 putative ORFs, increasing the database size by an order of magnitude. Although the authors used a cluster of 106 nodes to perform X!Tandem searches, it was not possible to use it as one unit, especially after adding a decoy version of the database to calculate the false-positive rate (FPR). Therefore, they split the database per chromosome, creating 24 databases, and searched them one by one using the MS/MS data from the Human Proteome Organization Plasma Proteome Project (HUPO PPP). Using this approach, followed by several steps of analysis and confirmation, they were able to identify 282 significant ORFs using 2,314 peptides, of which 627 were novel (Fermin *et al.* 2006).

In a recent project, Bitton *et al*. (2010) compared proteome data obtained from human breast epithelial cell lines against the six-frame translation of the human genome database with a concatenated reverse database for false positive rate (FPR) calculation. Since the decoy database is equivalent in size to the target database (Elias and Gygi 2007), it was again not possible to search the whole database as one unit. Therefore, they used an approach similar to the method described by Fermin *et al.* (2006). The database was split by chromosome into 23 databases, each of which contains target and decoy versions, and the search was performed against each database separately. In this work, Bitton *et al*. were able to identify 346 putative novel peptides; of which two correspond to novel isoforms, while the remainder correspond to novel loci, and many of them were confirmed using several methods (Bitton *et al.* 2010).

METHODS BASED ON NEW SEARCH ALGORITHMS

Methods in this category are based on developing novel database-searching algorithms that speed up the search process.

GENQUEST

Sevinsky *et al*. (2008) developed the GENQUEST method, which made searching the human whole genome possible with common desktop computers. In GENQUEST, six-frame translation and *in silico* trypsin digestion for the whole genome database are performed and the molecular weight (MW) and the peptide isoelectric focusing value (p*I*) are calculated for each peptide (forward and reverse) with a MW between 800 and 3000. The resultant library of peptides was called the human genome peptide database (HGPdb). Next, to make the search possible with common desktop computers, narrow-range peptide fasta files were created for SEQUEST search by sorting the peptides into files based on the p*I*, with each fasta file representing 0.01 p*I*. When the p*I* range is determined, the files in this range are concatenated, indexed (using BioWorks Browser – Thermo Fisher Scientific) and searched. This significantly reduces the size of the database to be searched. Using GENQUEST, almost all exonic peptides identified from the protein database were identified from the genome database and 540 peptides were uniquely identified from the genome database. The whole analysis was done in a common desktop computer with a Pentium 4 2.8 GHz processor and 3 GB RAM (Sevinsky *et al.* 2008).

InsPecT

Identification of posttranslational modifications (PTMs) is crucial for understanding cellular regulation processes. However, PTMs identification from databases that contain all possible mutations (modifications) using the normal search tools is computationally demanding. In order to accurately identify posttranslational modifications with reasonable cost, Tanner *et al*. (2005) developed *InsPecT*. *InsPecT* identifies PTMs from MS/MS data and genomic databases using database filters. The basic principle in *InsPecT* is to filter the databases aggressively and accurately, using the given spectrum and to return a small fraction that contains the candidate peptides able to produce the given spectrum with high probability. This allows applying more sophisticated and intensive analysis to the remaining fraction of the databases by considering a rich set of PTMs for each peptide. Further, the reduction of the number of candidates reduces the probability of false positives and high score achievement by chance. Four datasets were used to evaluate the performance of *InsPecT*, resulting in the identification of number of novel PTMs in the employed datasets, including phosphopeptides. In addition, *InsPecT* was two orders of magnitude faster than SEQUEST and significantly faster than X!Tandem on a complex mixture (Tanner *et al.* 2005).

ABLCP

Zhou *et al*. (2010) presented the new Algorithm Based on Longest Common Prefix (ABLCP) method for speeding up database search by efficient organization of the database. ABLCP uses an *on-line* digestion method to create an index of all peptides, then removes duplicates, and uses methods to ensure that no candidate peptides are missed. Thus, the identification time was noticeably improved using ABLCP, compared with methods that use peptide indexing, while accuracy was not affected. Further, the time and disk space required for index creation were less than those required by pFind (Wang *et al.* 2007) (pFind was chosen because it had been proven to have better performance than SEQUEST, Mascot and X!Tandem (Li *et al.* 2010)) in the case of a nor-

mal database. ABLCP performance was compared with other approaches that use either peptide indexing or no special data structures, and all the analysis was done using affordable computational resources (desktop PC with 2 CPUs each with 2 1.6 GHz cores and 4 GB RAM). However, ABLCP is implemented on pFind and designed to work with protein sequence databases, not genomic sequence databases (Zhou *et al.* 2010).

PepSplice

The PepSplice search algorithm, presented by Roos *et al*. (2007), uses cache-optimization and restriction of combined search spaces to speed up large-scale peptide identification tasks. The algorithm is a cache-aware algorithm, since it was designed to take into account the different storage levels with different speed and size (CPU, RAM, hard disk…). The search spaces that can be combined and searched using PepSplice are the non-tryptic peptides, whole genome, several posttranslational modifications, un-annotated point mutations and un-annotated splice sites. However, PepSplice allows restricting the number of variations that can co-occur per peptide. Then, the search is carried out by the CPU as a single job, and the final result merges all results from all combined search spaces. The cache-optimization and restriction of combined search spaces improved the search speed to reach the theoretical hardware limit. The authors demonstrated outstanding performance of PepSplice by searching over 1.4 million spectra obtained from *Arabidopsis* culture cell against the *Arabidopsis* protein database and the *Arabidopsis* genome database, considering a variety of search spaces simultaneously (such as semi- and non-tryptic peptides, various posttranslational modifications, point mutations and a huge number of potential splice sites). Interestingly, the search was carried out with single CPU with a throughput of 8 spectra per second, a speed that exceeds the measurement speed of most recent mass spectrometers (for instance, the throughput of the current LTQ instruments is 2 spectra per second) (Roos *et al.* 2007).

Integrating the peptide sequences with the human genome

Desiere *et al*. (2005) described a pipeline for mapping peptides obtained from large-scale MS/MS analysis to the genome and built an expandable resource for integrating peptide data obtained from different proteomics experiments without searching the genome database. This strategy depends on searching several protein databases to obtain the peptide sequences correspond to the peptide spectra, then using several computational steps to map these peptides to the genome to confirm the expression of the proteins and the corresponding genes. They applied this pipeline to the human and *Drosophila melanogaster* genomes, resulting in validation rates of 27% (9,747 proteins and 3,107 genes) and 14% (6,423 proteins and 1,876 genes) in human and *Drosophila*, respectively (Desiere *et al.* 2005).

HYBRID METHODS

Methods in this category combine spectral *de novo* interpretation and database searching approaches in order to retain the advantages of both approaches, while overcoming the limitations.

PepLine

Ferro *et al*. (2008) presented a data processing pipeline, *PepLine*, which automates the process of mapping large-scale MS/MS spectra datasets derived from tryptic peptides to the genomic sequence without preprocessing of the database, allowing the identification of novel genomic features and refinement of the genome annotation. However, since direct mapping to the genome requires identification of the peptide sequence corresponds to the spectrum, a process that is computationally expensive with large-scale datasets if database search methods are employed, *PepLine* uses peptide sequence tags (PSTs) to perform spectrum interpretation. The data is processed using three sequential modules optimized for working with large-scale datasets. The first module, Taggor, interprets the MS/MS spectra and generates the PSTs, the second module, PMMatch, maps the PSTs to the genome sequence, while the third module, PMClust, clusters closely located genomic hits. The pipeline performance was tested against database searching programs using a standard proteins dataset and an *Arabidopsis thaliana* envelop chloroplast sample, and it shows outstanding performance with large-scale datasets and genomes. Further, it allows for accurate identification of the exon-intron boundaries, which make it suitable for eukaryotic genomes. However, it should be noted that the current Taggor module of *PepLine* is specially designed to handle quadrupole time-of-flight (QTOF) MS/MS data. Nevertheless, the *PepLine* modularity makes it possible to use another program instead of Taggor when using MS/MS from an ion-trap-like instrument, then the analysis can be continued using *PepLine* (Ferro *et al.* 2008).

Lookup Peaks

Lookup Peaks is another hybrid method that uses partial *de novo* analysis then database search. It applies a small *de novo* interpretation to the spectrum to identify the *b*- and *y*-ion peaks (the *lookup peaks*). The lookup peaks are used to extract candidate peptides from the database. The limited number of candidate peptides, compared with the total number of peptides in the whole database, makes the identification computationally affordable. Further, the authors developed a software tool, ByOnic, that implements the Lookup Peaks method. The method performance and sensitivity were assessed using several datasets and performance was better than that of sequence tagging methods, Mascot, SEQUEST and X!Tandem, at both the peptide and protein levels. ByOnic was able to find low-concentration spiked human peptides in a mouse blood plasma sample, while these peptides were missed by other tools (Bern *et al.* 2007).

Spectral Dictionaries

Spectral Dictionaries is a hybrid method that combines *de novo* interpretation of the MS/MS spectrum and database search, but in a novel way. Hybrid methods, in general, are based on the sequence tagging approach that was proposed in 1994 by Mann and Wilm (1994). In these methods, small fractions of the peptide sequence (usually limited to a length of three) are inferred from the spectrum and these tags are later used to search the database. Spectral Dictionaries goes a step further by generating all possible full-length peptide reconstructs and insures that one of the generated reconstructs is correct. Although the idea is not new (Taylor and Johnson 1997), it was implemented once with a software tool based on a slow searching approach (Alves and Yu 2005) that limited its usability with large-scale datasets. Kim *et al*. presented a new implementation with superior performance when using datasets of over 20,000 peptides. The new implementation makes searching the six-frame translation of the human genome possible with a large-scale proteome dataset for proteogenomics, and it can also be modified to search for mutations and polymorphisms by using error-tolerant pattern matching when searching the database (Kim *et al.* 2009).

Genomic Peptide Finder (GPF)

As mentioned in the previous section, exon-exon junction peptides, also known as intron-split peptides, cannot be identified from the genome database, though it has been estimated that these peptides represent 20~25% of the total tryptic peptides deduced from the genome (Choudhary *et al.*

2001). These peptides help in the correct determination of the exon-intron boundaries and, consequently, in accurate annotation of the protein-coding regions. Therefore, methods like SkipE (Power *et al.* 2009) and the exon-exon junction database (Mo *et al.* 2008) were developed to identify such peptides using *in silico* intron-split peptides databases. However, Allmer *et al*. (2004) presented a novel method to identify intron-split peptides directly from the genome database by developing the GenomicPeptideFinder (GPF) algorithm. GPF uses *de novo* amino acid sequence prediction to infer the peptide sequence information together with the molecular weight (MW) of the precursor ion. A short fragment of the predicted peptide sequence is aligned with the six-frame translation of the genome and the MW of the precursor ion is used to assemble the full peptide using the sequence as a matrix. To speed up the search process, GPF performs two types of search, one using a long stretch of the peptide sequence (five amino acids) and the second using a shorter stretch (three amino acids). However, the second search is invoked only if the first resulted in matches with the genomic data. Next, GPF triggers four sequential processes aiming to identify peptides that match the search criteria and the resultant peptides are saved in a database of potential intron-split peptides. Finally, normal peptide identification tools such as Mascot (Perkins *et al.* 1999) or SEQUEST (Eng *et al* 1994) can be used to search the original spectra against the intron-split peptides database for the actual identification of the correct peptides (Allmer *et al.* 2004).

Fast spectra profile comparison

Aiming to speed up peptide sequence identification from large databases, Liu *et al*. (2005) proposed the Fast Spectra Profile Comparison method. Like GPF, this method speeds up the search by reducing the number of peptides to be searched with normal peptide identification tools such as Mascot or SEQUEST. However, the main principle here is to perform coarse comparison between the experimental spectrum and the theoretical spectra in order to exclude peptides with spectra showing little similarity to the experimental spectra. The peptides that pass the comparison are subjected to preliminary evaluation and finally sorted in ascending order and saved to a database of candidate peptides. The next step is similar to GPF, where Mascot or SEQUEST is used to search the constructed database, which is significantly smaller than the original. For the evaluation of the method, three datasets from three different sources, with different accuracy, and obtained with different instruments were used. The positive peptides (correct matches) of each dataset were already known. Applying the methods to the three datasets, the positive peptides were ranked in the top 10%, limiting the second stage, Mascot/ SEQUEST search, to a very small number of candidates. Further, the identification time was two times faster, on average, than the control (Liu *et al.* 2005).

METHODS INVOLVING COMPUTER HARDWARE

Methods in this section use an extraordinary approach to speed up the search process, such as embedding the search program in the computer hardware. Such approaches improve performance significantly, but have the drawback of limiting the usability of the method due to the requirement of particular hardware or infrastructure, for instance.

SEQUEST sorcerer

A unique speeding-up approach was implemented in Sorcerer. Sorcerer is an implementation of SEQUEST in firmware (embedded software that runs directly on the hardware) with a web-based interface that is similar to the Mascot interface. This makes the search extremely fast compared with other search programs, including the normal SEQUEST itself. Further, it reduces the required informatics skills, administrative tasks and power consumption. However, it suffers from the major technical limitation with large databases that requires consideration of six-frame translation, such as EST and genomic sequence databases, since Sorcerer uses indexed databases (Kapp and Schutz 2007). Recently, Sorcerer became available in two versions, Sorcerer 2 and Sorcerer Enterprise. Sorcerer 2 is designed for laboratories with frequent high-throughput requirement, and can handle data from modern instruments at up to 30,000 spectra/hour. Sorcerer Enterprise is designed for laboratories with intensive high-throughput needs, e.g., more than one high-throughput mass spectrometer or multiple core facilities. The Enterprise version is 2.5-fold faster than the normal version and can even scale up an additional 10 times [www.sagenresearch.com].

SPECTRUM REDUCTION METHOD

Spectrum reduction is a well known approach that is optionally used by several search engines to reduce the number of spectra to be processed. In this approach, a quality threshold can be set to exclude all spectra below it, since these spectra are considered of low quality (Kapp and Schutz 2007). We can call this type of reduction *positive reduction,* as it reduces the effort that could be wasted in processing such low-quality spectra. Therefore, any reduction of high-quality spectra can be considered *negative reduction*. Although *positive reduction* reduces the number of spectra to be processed, the number of remaining spectra is still high, especially when using recent high-accuracy mass spectrometers. The method presented in this category proposes a novel approach to reduce the number of spectra, while avoiding *negative reduction*.

Mass Spectrum Sequential Subtraction (MSSS)

Mass Spectrum Sequential Subtraction (MSSS) is a bioinformatics method developed especially to facilitate the comparison of large-scale MS/MS data with large databases. MSSS uses a novel subtraction method to identify large numbers of spectra from sequence databases within a reasonable time and with affordable computational demands. The basic idea of MSSS is to compare the whole large-scale data with a reference database, usually a protein database, then to *subtract* all spectra corresponding to the identified peptides and to create new files containing the unidentified spectra. Next, the new files can be compared with the large database or with another reference database to subtract more spectra. With the described subtraction approach, MSSS reduces the number of high-quality spectra to be processed while avoiding *negative reduction*. This approach should reduce the search time and save computational resources. In addition, it can be used to find modifications and mutations by using databases of normal proteins and mutated or modified proteins, by performing subtraction after searching the database of normal proteins.

In a recent preliminary study, MSSS was used to identify modifications and disease-related mutations using four databases from normal individuals (protein, cDNA, transcript and genome databases) and one cancer patient database to identify a list of candidate onco-peptides, including phosphopeptides (Helmy *et al.* 2010). This tool promises to have further applications in cancer, such as identifying cancer-related mutations, new drug targets and new biomarkers.

CONCLUSION

Proteogenomics is an emerging research approach that utilizes current advances in proteomics and genomics, as well as creating its own tools and technologies. Among several challenges facing proteogenomics, the process of comparing large-scale MS/MS data with large databases remains the major obstacle to full utilization of recent high-throughput proteomics techniques. Various methods have been developed to facilitate this comparison, including deve-

loping new algorithms, methods for reducing the database size or reducing the number of spectra to be processed, and methods involving computer hardware to speed up the search process. However, despite these great efforts, most of the developed methods still suffer from problems such as having been tailored to solve certain problem(s), to work with data from certain instruments, to be applicable only with protein sequence databases, or to require special hardware infrastructure. Further, some methods are implemented to work with particular search engine(s), while others have not yet been implemented in any publicly available commercial or open-source tool. Therefore, there is still room for new methods, or improvements of the current methods, that would be more generalized, flexible and easy to integrate into existing data-processing workflow.

REFERENCES

Allmer J, Markert C, Stauber EJ, Hippler M (2004) A new approach that allows identification of intron-split peptides from mass spectrometric data in genomic databases. *FEBS Letters* **562 (1-3)**, 202-206

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215 (3)**, 403-410

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* **25 (17)**, 3389-3402

Alves G, Yu YK (2005) Robust accurate identification of peptides (RAId): Deciphering MS2 data using a structured library search with *de novo* based statistics. *Bioinformatics (Oxford, England)* **21** (19), 3726-3732

Ansong C, Purvine SO, Adkins JN, Lipton MS, Smith RD (2008a) Proteogenomics: Needs and roles to be filled by proteomics in genome annotation. *Briefings in Functional Genomics and Proteomics* **7 (1)**, 50-62

Ansong C, Yoon H, Norbeck AD, Gustin JK, McDermott JE, Mottaz HM, Rue J, Adkins JN, Heffron F, Smith RD (2008b) Proteomics analysis of the causative agent of typhoid fever. *Journal of Proteome Research* **7 (2)**, 546-557

Armengaud J (2009) A perfect genome annotation is within reach with the proteomics and genomics alliance. *Current Opinion in Microbiology* **12 (3)**, 292-300

Armengaud J (2010) Proteogenomics and systems biology: Quest for the ultimate missing parts. *Expert Reviews Proteomics* **7 (1)**, 65-77

Baerenfaller K, Grossmann J, Grobei MA, Hull R, Hirsch-Hoffmann M, Yalovsky S, Zimmermann P, Grossniklaus U, Gruissem W, Baginsky S (2008) Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science (NY)* **320 (5878)**, 938-941

Bechah Y, El Karkouri K, Mediannikov O, Leroy Q, Pelletier N, Robert C, Medigue C, Mege JL, Raoult D (2010) Genomic, proteomic, and transcriptomic analysis of virulent and avirulent *Rickettsia prowazekii* reveals its adaptive mutation capabilities. *Genome Research* **20 (5)**, 655-663

Bern M, Cai Y, Goldberg D (2007) Lookup peaks: A hybrid of *de novo* sequencing and database search for protein identification by tandem mass spectrometry. *Analytical Chemistry* **79 (4)**, 1393-1400

Bindschedler LV, Burgis TA, Mills DJ, Ho JT, Cramer R, Spanu PD (2009) *In planta* proteomics and proteogenomics of the biotrophic barley fungal pathogen *Blumeria graminis* f. sp. *hordei*. *Molecular and Cellular Proteomics* **8 (10)**, 2368-2381

Bitton DA, Smith DL, Connolly Y, Scutt PJ, Miller CJ (2010) An integrated mass-spectrometry pipeline identifies novel protein coding-regions in the human genome. *PLoS One* **5 (1)**, e8949

Borchert N, Dieterich C, Krug K, Schutz W, Jung S, Nordheim A, Sommer RJ, Macek B (2010) Proteogenomics of *Pristionchus pacificus* reveals distinct proteome structure of nematode models. *Genome Research* **20 (6)**, 837-846

Bringans S, Hane JK, Casey T, Tan KC, Lipscombe R, Solomon PS, Oliver RP (2009) Deep proteogenomics; high throughput gene validation by multidimensional liquid chromatography and mass spectrometry of proteins from the fungal wheat pathogen *Stagonospora nodorum*. *BMC Bioinformatics* **10**, 301

Cargile BJ, Bundy JL, Freeman TW, Stephenson JL Jr. (2004) Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification. *Journal of Proteome Research* **3 (1)**, 112-119

Castellana N, Bafna V (2010) Proteogenomics to discover the full coding content of genomes: A computational perspective. *Journal of Proteomics* **73 (11)**, 2124-2135

Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, Briggs SP (2008) Discovery and revision of *Arabidopsis* genes by proteogenomics. *Proceedings of the National Academy of Sciences USA* **105 (52)**, 21034-21038

Choudhary JS, Blackstock WP, Creasy DM, Cottrell JS (2001) Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics* **1 (5)**, 651-667

Colinge J, Masselot A, Giron M, Dessingy T, Magnin J (2003) OLAV: Towards high-throughput tandem mass spectrometry data identification. *Proteomics* **3 (8)**, 1454-1463

Craig R, Beavis RC (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Communications in Mass Spectrometry* **17 (20)**, 2310-2316

Craig R, Beavis RC (2004) TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics (Oxford, England)* **20 (9)**, 1466-1467

de Groot A, Dulermo R, Ortet P, Blanchard L, Guerin P, Fernandez B, Vacherie B, Dossat C, Jolivet E, Siguier P, Chandler M, Barakat M, Dedieu A, Barbe V, Heulin T, Sommer S, Achouak W, Armengaud J (2009) Alliance of proteomics and genomics to unravel the specificities of Sahara bacterium *Deinococcus deserti*. *PLoS Genetics* **5 (3)**, e1000434

Delmotte N, Knief C, Chaffron S, Innerebner G, Roschitzki B, Schlapbach R, von Mering C, Vorholt JA (2009) Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proceedings of the National Academy of Sciences USA* **106 (38)**, 16428-16433

Denef VJ, Kalnejais LH, Mueller RS, Wilmes P, Baker BJ, Thomas BC, VerBerkmoes NC, Hettich RL, Banfield JF (2010) Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *Proceedings of the National Academy of Sciences USA* **107 (6)**, 2383-2390

Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, Eng JK, Aderem A, Boyle R, Brunner E, Donohoe S, Fausto N, Hafen E, Hood L, Katze MG, Kennedy KA, Kregenow F, Lee H, Lin B, Martin D, Ranish JA, Rawlings DJ, Samelson LE, Shiio Y, Watts JD, Wollscheid B, Wright ME, Yan W, Yang L, Yi EC, Zhang H, Aebersold R (2005) Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biology* **6 (1)**, R9

Dutta D, Chen T (2007) Speeding up tandem mass spectrometry database search: Metric embeddings and fast near neighbor search. *Bioinformatics (Oxford, England)* **23 (5)**, 612-618

Edwards NJ (2007) Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. *Molecular Systems Biology* **3**, 102

Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4 (3)**, 207-214

Eng JK, McCormack AL, Yates JR (1994) An approach to correlate MS/MS data to amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **5**

Fermin D, Allen BB, Blackwell TW, Menon R, Adamski M, Xu Y, Ulintz P, Omenn GS, States DJ (2006) Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biology* **7 (4)**, R35

Ferro M, Tardif M, Reguer E, Cahuzac R, Bruley C, Vermat T, Nugues E, Vigouroux M, Vandenbrouck Y, Garin J, Viari A (2008) PepLine: A software pipeline for high-throughput direct mapping of tandem mass spectrometry data on genomic sequences. *Journal of Proteome Research* **7 (5)**, 1873-1883

Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH (2004) Open mass spectrometry search algorithm. *Journal of Proteome Research* **3 (5)**, 958-964

Gupta N, Benhamida J, Bhargava V, Goodman D, Kain E, Kerman I, Nguyen N, Ollikainen N, Rodriguez J, Wang J, Lipton MS, Romine M, Bafna V, Smith RD, Pevzner PA (2008) Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. *Genome Research* **18 (7)**, 1133-1142

Helmy M, Sugiyama N, Tomita M, Ishihama Y (2010) Onco-proteogenomics: A novel approach to identify cancer-specific mutations combining proteomics and transcriptome deep sequencing. *Presented at Beyond the Genome: The True Gene Count, Human Evolution and Disease Genomics*, Harvard Medical School, Boston, MA, USA, 17 pp

Helmy M, Tomita M, Ishihama Y (2011) OryzaPG-DB: Rice proteome database based on shotgun proteogenomics. *BMC Plant Biology* **11 (1)**, 63

Hixson KK, Adkins JN, Baker SE, Moore RJ, Chromy BA, Smith RD, McCutchen-Maloney SL, Lipton MS (2006) Biomarker candidate identification in *Yersinia pestis* using organism-wide semiquantitative proteomics. *Journal of Proteome Research* **5 (11)**, 3008-3017

Huang Y, Franklin J, Gifford K, Roberts BL, Nicolette CA (2004) A high-throughput proteo-genomics method to identify antibody targets associated with malignant disease. *Clinical Immunology* **111 (2)**, 202-209

Ishihama Y, Rappsilber J, Mann M (2006) Modular stop and go extraction tips with stacked disks for parallel and multidimensional peptide fractionation in proteomics. *Journal of Proteome Research* **5 (4)**, 988-994

Ishino Y, Okada H, Ikeuchi M, Taniguchi H (2007) Mass spectrometry-based prokaryote gene annotation. *Proteomics* **7 (22)**, 4053-4065

Jacob F, Goldstein DR, Fink D, Heinzelmann-Schwarz V (2009) Proteogenomic studies in epithelial ovarian cancer: Established knowledge and future needs. *Biomarkers in Medicine* **3 (6)**, 743-756

Jaffe JD, Berg HC, Church GM (2004a) Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **4 (1)**, 59-77

Jaffe JD, Stange-Thomann N, Smith C, DeCaprio D, Fisher S, Butler J, Calvo S, Elkins T, FitzGerald MG, Hafez N, Kodira CD, Major J, Wang

S, Wilkinson J, Nicol R, Nusbaum C, Birren B, Berg HC, Church GM (2004b) The complete genome and proteome of *Mycoplasma mobile*. *Genome Research* **14 (8)**, 1447-1461

Kalume DE, Peri S, Reddy R, Zhong J, Okulate M, Kumar N, Pandey A (2005) Genome annotation of *Anopheles gambiae* using mass spectrometry-derived data. *BMC Genomics* **6**, 128

Kapp E, Schutz F (2007) Overview of tandem mass spectrometry (MS/MS) database search algorithms. *Current Protocols in Protein Science* **Chapter 25**, Unit 25, 2.1-2.19

Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4 (7)**, 1985-1988

Kiebel GR, Auberry KJ, Jaitly N, Clark DA, Monroe ME, Peterson ES, Tolic N, Anderson GA, Smith RD (2006) PRISM: A data management system for high-throughput proteomics. *Proteomics* **6 (6)**, 1783-1790

Kim S, Gupta N, Bandeira N, Pevzner PA (2009) Spectral dictionaries: Integrating *de novo* peptide sequencing with database search of tandem mass spectra. *Molecular and Cellular Proteomics* **8 (1)**, 53-69

Kolker E, Makarova KS, Shabalina S, Picone AF, Purvine S, Holzman T, Cherny T, Armbruster D, Munson RS Jr., Kolesov G, Frishman D, Galperin MY (2004) Identification and functional analysis of 'hypothetical' genes expressed in *Haemophilus influenzae*. *Nucleic Acids Research* **32 (8)**, 2353-2361

Lasonder E, Ishihama Y, Andersen JS, Vermunt AM, Pain A, Sauerwein RW, Eling WM, Hall N, Waters AP, Stunnenberg HG, Mann M (2002) Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* **419 (6906)**, 537-542

Li Y, Chi H, Wang LH, Wang HP, Fu Y, Yuan ZF, Li SJ, Liu YS, Sun RX, Zeng R, He SM (2010) Speeding up tandem mass spectrometry based database searching by peptide and spectrum indexing. *Rapid Communications in Mass Spectrometry* **24 (6)**, 807-814

Liu J, Carrillo B, Yanofsky C, Beaudrie C, Morales F, Kearney R (2005) A novel approach to speed up peptide sequencing via MS/MS spectra analysis. *Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society* **4**, 4441-4444

Loevenich SN, Brunner E, King NL, Deutsch EW, Stein SE, Aebersold R, Hafen E (2009) The *Drosophila melanogaster* PeptideAtlas facilitates the use of peptide data for improved fly proteomics and genome annotation. *BMC Bioinformatics* **10**, 59

Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G (2003) PEAKS: Powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry* **17 (20)**, 2337-2342

Mann M, Wilm M (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry* **66 (24)**, 4390-4399

Mattow J, Siejak F, Hagens K, Schmidt F, Koehler C, Treumann A, Schaible UE, Kaufmann SH (2007) An improved strategy for selective and efficient enrichment of integral plasma membrane proteins of *mycobacteria*. *Proteomics* **7 (10)**, 1687-1701

Merrihew GE, Davis C, Ewing B, Williams G, Kall L, Frewen BE, Noble WS, Green P, Thomas JH, MacCoss MJ (2008) Use of shotgun proteomics for the identification, confirmation, and correction of *C. elegans* gene annotations. *Genome Research* **18 (10)**, 1660-1669

Mo F, Hong X, Gao F, Du L, Wang J, Omenn GS, Lin B (2008) A compatible exon-exon junction database for the identification of exon skipping events using tandem mass spectrum data. *BMC Bioinformatics* **9**, 537

Nielsen P, Krogh A (2005) Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics (Oxford, England)* **21 (24)**, 4322-4329

Nouvel LX, Sirand-Pugnet P, Marenda MS, Sagne E, Barbe V, Mangenot S, Schenowitz C, Jacob D, Barre A, Claverol S, Blanchard A, Citti C (2010) Comparative genomic and proteomic analyses of two *Mycoplasma agalactiae* strains: Clues to the macro- and micro-events that are shaping mycoplasma diversity. *BMC Genomics* **11**, 86

Park CY, Klammer AA, Kall L, MacCoss MJ, Noble WS (2008) Rapid and accurate peptide identification from tandem mass spectra. *Journal of Proteome Research* **7 (7)**, 3022-3027

Payne SH, Huang ST, Pieper R (2010) A proteogenomic update to *Yersinia*: Enhancing genome annotation. *BMC Genomics* **11**, 460

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20 (18)**, 3551-3567

Power KA, McRedmond JP, de Stefani A, Gallagher WM, Gaora PO (2009) High-throughput proteomics detection of novel splice isoforms in human platelets. *PLoS One* **4 (3)**, e5001

Roos FF, Jacob R, Grossmann J, Fischer B, Buhmann JM, Gruissem W, Baginsky S, Widmayer P (2007) PepSplice: Cache-efficient search algorithms for comprehensive identification of tandem mass spectra. *Bioinformatics (Oxford, England)* **23 (22)**, 3016-3023

Sadygov RG, Cociorva D, Yates JR 3rd (2004) Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. *Nature Methods* **1 (3)**, 195-202

Seidler J, Zinn N, Boehm ME, Lehmann WD (2010) *De novo* sequencing of peptides by MS/MS. *Proteomics* **10 (4)**, 634-649

Sevinsky JR, Cargile BJ, Bunger MK, Meng F, Yates NA, Hendrickson RC, Stephenson JL Jr. (2008) Whole genome searching with shotgun proteomic data: Applications for genome annotation. *Journal of Proteome Research* **7 (1)**, 80-88

Sigdel TK, Sarwal MM (2008) The proteogenomic path towards biomarker discovery. *Pediatric Transplantation* **12 (7)**, 737-747

Stanke M, Diekhans M, Baertsch R, Haussler D (2008) Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics (Oxford, England)* **24 (5)**, 637-644

Tanner S, Shen Z, Ng J, Florea L, Guigo R, Briggs SP, Bafna V (2007) Improving gene annotation using peptide mass spectrometry. *Genome Research* **17 (2)**, 231-239

Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V (2005) InsPecT: Identification of posttranslationally modified peptides from tandem mass spectra. *Analytical Chemistry* **77 (14)**, 4626-4639

Tatusova TA, Madden TL (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiology Letters* **174 (2)**, 247-250

Taylor JA, Johnson RS (1997) Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry* **11 (9)**, 1067-1075

Tress ML, Bodenmiller B, Aebersold R, Valencia A (2008) Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biology* **9 (11)**, R162

Turse JE, Marshall MJ, Fredrickson JK, Lipton MS, Callister SJ (2010) An empirical strategy for characterizing bacterial proteomes across species in the absence of genomic sequences. *PLoS One* **5 (11)**, e13968

Tyers M, Mann M (2003) From genomics to proteomics. *Nature* **422 (6928)**, 193-197

Wang LH, Li DQ, Fu Y, Wang HP, Zhang JF, Yuan ZF, Sun RX, Zeng R, He SM, Gao W (2007) pFind 2.0: A software package for peptide and protein identification via tandem mass spectrometry. *Rapid Communications in Mass Spectrometry* **21 (18)**, 2985-2991

Wang R, Prince JT, Marcotte EM (2005) Mass spectrometry of the *M. smegmatis* proteome: Protein expression levels correlate with function, operons, and codon bias. *Genome Research* **15 (8)**, 1118-1126

Wright JC, Sugden D, Francis-McIntyre S, Riba-Garcia I, Gaskell SJ, Grigoriev IV, Baker SE, Beynon RJ, Hubbard SJ (2009) Exploiting proteomic data for genome annotation and gene model validation in *Aspergillus niger*. *BMC Genomics* **10**, 61

Xia D, Sanderson SJ, Jones AR, Prieto JH, Yates JR, Bromley E, Tomley FM, Lal K, Sinden RE, Brunk BP, Roos DS, Wastling JM (2008) The proteome of *Toxoplasma gondii*: Integration with the genome provides novel insights into gene expression and annotation. *Genome Biology* **9 (7)**, R116

Yates JR 3rd, Eng JK, McCormack AL (1995a) Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Analytical Chemistry* **67 (18)**, 3202-3210

Yates JR 3rd, Eng JK, McCormack AL, Schieltz D (1995b) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Analytical Chemistry* **67 (8)**, 1426-1436

Yates JR, Eng JK, Clauser KR, Burlingame AL (1996) Search of sequence databases with uninterpreted high-energy collision-induced dissociation spectra of peptides. *Journal of the American Society for Mass Spectrometry* **7**, 1089-1098

Zhou C, Chi H, Wang LH, Li Y, Wu YJ, Fu Y, Sun RX, He SM (2010) Speeding up tandem mass spectrometry-based database searching by longest common prefix. *BMC Bioinformatics* **11**, 577

Zivanovic Y, Armengaud J, Lagorce A, Leplat C, Guerin P, Dutertre M, Anthouard V, Forterre P, Wincker P, Confalonieri F (2009) Genome analysis and genome-wide proteomics of *Thermococcus gammatolerans*, the most radioresistant organism known amongst the Archaea. *Genome Biology* **10 (6)**, R70