Systems biology approaches to improve genome annotations through development of bioinformatics pipeline for large-scale mass spectrometry-based proteogenomics

KEIO UNIVERSITY
Graduate School of Media and Governance

Mohamed Helmy

Doctoral Dissertation Academic Year 2012

Systems biology approaches to improve genome annotations through development of bioinformatics pipeline for large-scale mass spectrometry-based proteogenomics

Mohamed Helmy

KEIO UNIVERSITY
Graduate School of Media and Governance
Institute for Advanced Biosciences

The thesis Submitted for the fulfillment of the requirements of the degree of Doctor of Philosophy in the Graduate School of Media and Governance (Systems Biology Program)

2012

Abstract

Systems biology is an interdisciplinary field aims to create holistic understanding of the biological systems through employing different experimental and computational techniques. Proteogenomics is a systems biology approach makes alliance between proteomics and genomics to utilize the powerfulness of the modern proteomics techniques, namely mass spectrometry (MS)-based shotgun proteomics, in revealing novel genomic information. MS-based shotgun proteomic analysis generates MS/MS peptide spectra that hold the peptide fingerprints. To identify the peptide sequences corresponds to each MS/MS spectrum, peptide spectra are compared against databases of putative protein or nucleotide sequences. Mapping the identified peptide sequences to the genomic data reveals valuable information about its genomic origin that can lead to improve, correct or confirm the genome annotation. However, peptide sequence identification from large-scale proteomic analysis and using large-sized databases remains major challenge for proteogenomics, due to the capabilities of the modern high-throughput mass spectrometers that can generate millions of spectra and the growing size of protein and genomic sequence databases. In this thesis, I present an intensive survey for the efforts done by the proteomics and bioinformatics societies in the last decade to facilitate the peptide identification process in large-scale studies (chapter 1). Then, I describe the design and development of a novel bioinformatics pipeline for large-scale proteogenomics (chapters 2~5). The developed bioinformatics pipeline consists of 1.Mass Spectrum Sequential Subtraction (MSSS) method (chapter 2), 2.The Rice Proteogenomics Database (OryzaPG-DB v1.0 and v1.1) (chapter 3 and 4, respectively), and 3.The ProteoGenomic Features Evaluator (PGFeval) software tool (chapter 5). The developed pipeline employed in the analysis of rice proteome and phosphoproteome data (61 LC-MS/MS runs) revealing 98 novel genomic feature in 62 rice genes.

Key Words: Proteomics, Genomics, Proteogenomics, Bioinformatics, Rice, Mass Spectrometry

Table of Content

List of Publications

1. Helmy M, Tomita M, Ishihama Y: **OryzaPG-DB: Rice Proteome Database based on Shotgun Proteogenomics**. *BMC Plant Biology* (2011), **11**(1):63. doi:10.1186/1471-2229-1111-1163.
2. Helmy M, Tomita M, Ishihama Y: **Peptide Identification by Searching Large-Scale Tandem Mass Spectra against Large Databases: Bioinformatics Methods in Proteogenomics**. *Genes Genomes and Genomics* (2012), **6**(Special Issue 1):76-85.
3. Helmy M, Sugiyama N, Tomita M, Ishihama Y: **The rice proteogenomics database OryzaPG-DB: development, expansion and new features** *Frontiers in plant proteomics* (2012), **3**:65. doi: 10.3389/fpls.2012.00065.
4. Helmy M, Sugiyama N, Tomita M, Ishihama Y: **MSSS: Mass spectrum sequential subtraction speeds up searching large peptide MS/MS spectra datasets against large nucleotide databases for proteogenomics**. *Genes to Cells,* (In Press).

List of Tables

List of Figures

Acknowledgment

Completing a PhD was truly the most challenging activity in my life and I would not have been able to complete it without the aid and support of countless people over the past five years since I joined the systems biology program at Keio University.

First and foremost, I want to thank my acting advisor Professor Yasushi Ishihama – Kyoto University and Keio University (IAB). It has been an honor to be his student for the past four years. He has taught me how excellent research is done. I appreciate all his contributions of time, ideas, and advise to make my PhD experience stimulating and fruitful. The joy and enthusiasm he has for his research was contagious and motivational for me, even during tough times in the PhD chase. Further, his hardworking attitude shaped up my research attitude and resulted in a productive PhD experience.

Professor Masaru Tomita, my main advisor and the director of the Institute for Advanced Biosciences (IAB), is a very special advisor and director with his deep understanding of the student's mentality, non-stop and generous support for students. I believe I was very lucky having him as my main advisor. I would like also the express my gratitude to my co-advisors Professor Akio Kanai and Professor Tomoyoshi Soga for their beneficial comments, feedbacks and ideas on my research.

The members of the proteomics group of IAB contributed immensely to my personal and professional time at Keio. The group has been a source of friendships as well as good advice and collaboration. I am especially grateful for Dr. Naoyuki Sugiyama, Dr. Takeshi Masuda and Assistant Professor Natsumi Saito. Further, my close friends and former members of the group Dr. Koshi Imami – University of British Colombia - Canada, Mai Tsukahara – Nara University, Haruna Imamura – Kyoto University, Mio Iwasaki – Kyoto University and Tsuyoshi Akuzawa. I had very joyful and fruitful time with them either in doing research or social life in Tsuruoka.

I am thankful to all faculty members and researchers of IAB past and present. I benefited a lot from the classes, lectures, seminars and discussions I had with them. I especially thank, Professor Mitsuhiro Itaya for allowing me to perform my teaching experience with him, Assistant Professor Martin Robert, Assistant Professor Douglas Murray, Assistant Professor Kumar Selvarajoo, Associate Professor Masa Tsuchiya and Assistant Professor Arun Krishnan.

I am also grateful to Dr. Naoyuki Sugiyama and Mr. Yasuyuki Igarashi for helping me performing my experiments and for performing Hela phosphoproteome analysis and the rice in-gel extraction and digestion, respectively.

I highly appreciate the help provided by the administrative staff at IAB and SFC. I especially thank, Akiko Shiozawa, Yoko Tsuchiya, Ayumi Mikami, Miho Sato, Shin Sato, Yuma Kuga, Maki Oike, Kaori Tomita and Ayami Mizukami.

I am grateful to all my friends in IAB for both friendship and helpful research advices. I especially mention my mentor and very close friends Kentaro Hayashi, Kalesh Sasidharan and Cornelia Amariei. Also, Dr. Kosuke Fujishima – Astrobiology Lab - NASA - USA, Dr. Nozomu Yachie – Toronto University - Canada, Dr. Vincent Piras, Kaori Takane and Dr. Kian-Kai Cheng.

This research was conducted using generous funding from the Ministry of Higher Education and Scientific research - Egypt and its representative in Japan, the Culture, Education and Science Bureau in The Embassy of Egypt in Japan, Yamagata prefecture and Tsuruoka City grant to Keio, The Global-Center of Excellence (G-COE) - Keio University and Japan Society for Promotion of Science (JSPS) Grant-in-aid (Grant ID 236127) – Japan.

I would like to thank my dearly beloved people, Helmy and Mariam, my parents, first fans and continuous supporters. Heba, my wife, Omar and Yusof, my kids who left Egypt where the family and friend and joined me at Japan. I cannot express how I am happy and proud having those great people near to me all my life giving me their love and support.

Finally, I dedicate this work, with all my gratitude and respect, to the martyrs and the injured of the January 25th Revolution in Egypt, hopping to contribute to the development of our nation using what I had learned.

Mohamed Helmy
June 25th 2012
Tsuruoka City
Japan

Chapter 1

Survey of bioinformatics methods and algorithms facilitate peptide sequence identification from databases in large-scale proteogenomics.

1.1 Chapter 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 chapter, I 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.

1.2 Introduction

Proteomics aims to characterize the expressed proteins and the corresponding peptides in a given sample, including elucidation of their sequence, structure and function. [1] 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. [2-5] Proteogenomics is the utilization of large-scale proteome data in genome annotation refinement. [2] 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. [2] 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. [6-10]

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. [7, 11] 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. [7] 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. [11] 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. [12] Power *et al*. (2009) found novel splice isoforms in human platelets by using a proteomics approach to identify the exon-skipping events. [13] 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. [6] The same approach is also applicable to the genomes of major plant crops such as rice [14], fungi such as *Aspergillus niger* (the black mold fungus) [15], parasites such as *Plasmodium falsiparum* (the malaria parasite) and *Toxoplasma gondii* [16, 17], insects such as *Drosophila melanogaster* [18, 19], nematodes such as *Pristionchus pacificus* [20] and Archaea such as *Thermococcus gammatolerans*.[21]

The basic outcome of a proteogenomic analysis is to validate the predicted gene models at the translational level, as presented in several reports. [22-24] In addition, proteogenomics has been utilized to reveal many other significant genomic features, e.g., finding new gene models [6, 7, 11, 23], determination of the protein start and termination sites [25-27], finding and verifying splice isoforms at the protein level [13, 25, 28] and verifying hypothetical and conserved hypothetical genes/proteins. [25, 29-31] 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. [3]

In addition to finding novel genomic features to refine the genome annotation, proteogenomics can be applied in biomarker discovery [7, 32], for identification of antibody targets [33], to provide a better understanding of the host-parasite relationship [16, 34, 35] and to understand the mechanisms of ecological diversity and environmental adaptation. [3, 36] Further, proteogenomic studies have been performed on several genomes of related species to identify rare post-translational modifications [37], to investigate adaptive mutation capabilities among species [38] and to understand diversity-shaping events between species. [39] Proteogenomics also provides novel insight in cancer research, either in finding biomarkers, related somatic mutations or diagnosis. [40, 41]

1.3 Large-scale proteogenomics workflow

Usually, large-scale proteome analyses are required for proteogenomics projects. In a typical proteogenomics project (Figure 1.1), high-throughput proteomics (usually mass spectrometry-based) is used to obtain the sequences of the peptides of the sample in question. [5] 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) [42] and isoelectric focusing (IEF) [43], 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 [44] or ii) database search of putative sequences. [45, 46]

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.

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. [47-49] Then, the alignment results can be compared with the current annotation to confirm and improve the annotated gene models [10, 17, 50] 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. [51]

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 [23] and the *Deinococcus deserti* genome annotation. [3]

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. [52] 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. [53] Thus, the computational demands for such search are in some cases so high as to be unaffordable. For example, 260 days of CPU time was required to run over 4,000 X!Tandem [54, 55] searches against the *Shewanella* genome database [56], even though a PRISM computing cluster with 32 processing nodes was used. [57]

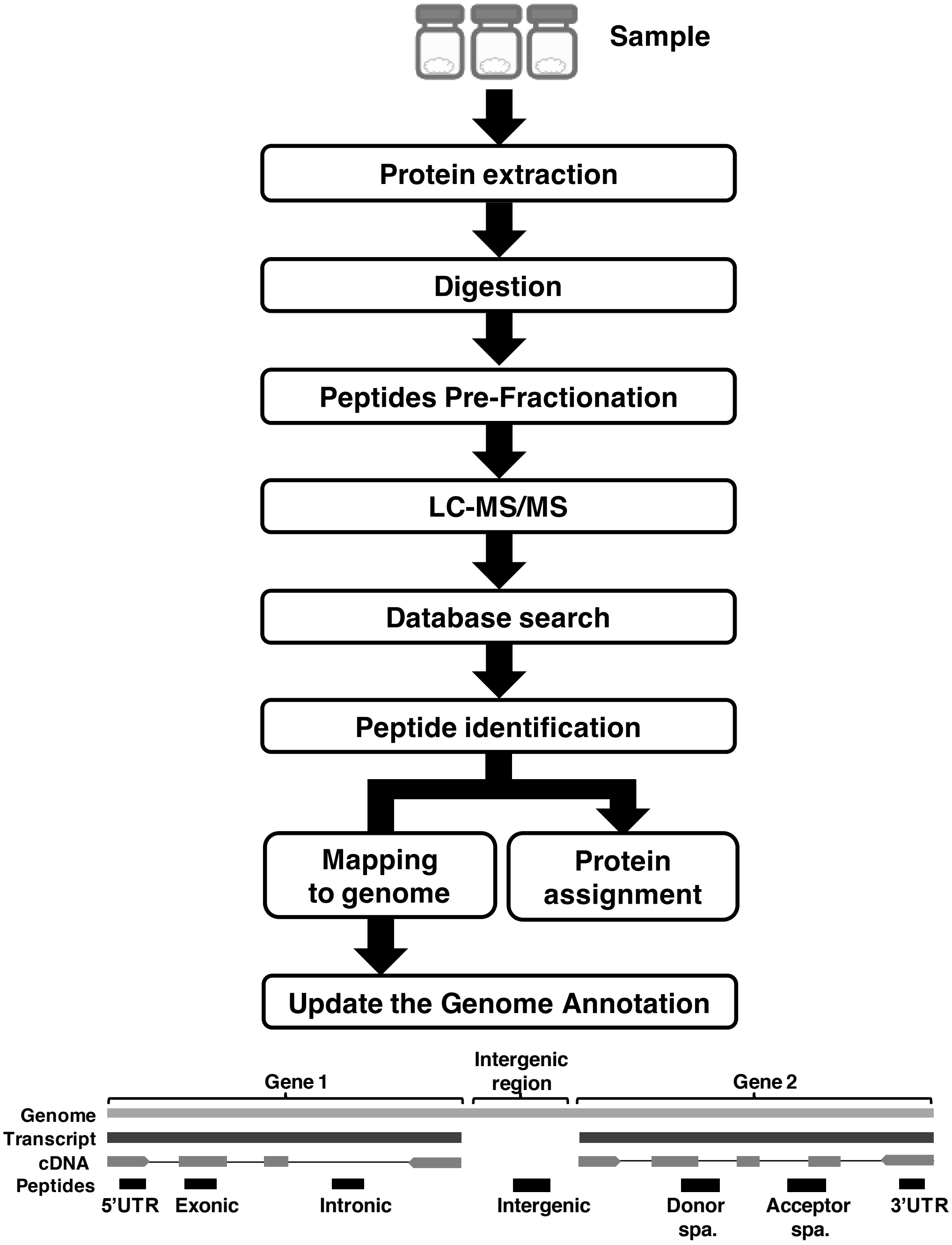


Figure1.1 Simplified representation of large-scale proteogenomics workflow. Proteins are extracted from the sample and digested using suitable protease(s), e.g. trypsin, then the resultant peptides are pre-fractionated. The fractions are then processed and submitted to LC-MS/MS. The obtained MS/MS spectra 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.

1.4 The need for faster database searching methods

Several database searching programs and algorithms are currently available, including SEQUEST [58], Mascot [59], X!Tandem [54, 55] and several other tools and algorithms, including pFind [60], OMSSA [61], PepSplice [62], Phenyx [63], PEAKS (SPIDER) [64] SpectrumMill (Agilent Technologies, CA), ProteinPilot (AB-Sciex, CA) and Crux. [65] 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 [66], while the size of the NCBInr protein sequence database doubled in about 18 months. [52]

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 advances 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). [66]

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 fully specific digestion of the IPI.Human database. [52]

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 [67] resulted in a 170-fold increase in the non-redundant peptides, compared with fully specific digestion. [52]

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. [52, 66]

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. [45] 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. [58] However, in later versions, SEQUEST was developed to be able to include modifications [68], search EST databases [69], and search high-energy CID data [70]; in parallel

to these proteomics-related updates, a cluster version was introduced to make use of multi-core/multi-computer systems. [45] 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. [45] However, all these computational and hardware developments have failed to keep pace with the rapid developments of the analytical instruments.

1.5 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. [52]

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.

1.5.1 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.

1.5.1.1 Database indexing

Database indexing, also known as peptide indexing, is one of the most widely used methods to facilitate peptide identification. [66] Several search engines use database indexing approaches, such as SEQUEST [58], pFind [60] and Crux. [65] 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. [71] 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. [66] 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-Δ ~ m+Δ]. Thus, indexing allows the program to avoid the *one-against-all* comparison, thereby reducing search time and computational resources. [71] 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. [52]

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 [66].

3. Hybrid indexing

There is a hybrid method that combines *off-line* and *on-line* indexing, that is used by pFind [60]. 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. [66] 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. [66]

1.5.1.2 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. [25]

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. [11]

2. Sophisticated sequence database comparison strategy to search EST databases

Aiming to identify peptides from alternative splicing isoforms and coding SNP proteins, Edwards (2007) 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. [53]

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. [71]

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) [67], yielding 89 genes with alternative splicing isoforms; many of them were confirmed at the mRNA level using RT-PCR and sequencing of the products. [13]

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. [28]

1.5.1.3 Database splitting

A simple and straightforward 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 proteogenomics 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. [72]

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 [73], 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. [74]

1.5.2 Methods based on new search algorithms

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

1. 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 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. [75]

2. 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. [76]

3. 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 [60] (pFind was chosen because it had been proven to have better performance than SEQUEST, Mascot and X!Tandem [66]) in the case of a normal 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. [52]

4. 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). [62]

5. 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. [12]

1.5.3 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.

1. 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*. [77]

2. 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 other tools missed these peptides. [78]

3. 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). [79] 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 [80], it was implemented once with a software tool based on a slow searching approach [81] that limited its usability with large-scale datasets. Kim *et al*. (2009) 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. [82]

4. GenomicPeptideFinder (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 [83]. 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 [13] and the exon-exon junction database [28] 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 or SEQUEST can be used to search the original spectra against the intron-split peptides database for the actual identification of the correct peptides. [84]

5. 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. [85]

1.5.4 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.

1. 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 [45]. 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].

1.5.5 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. [45] 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*.

1. 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 (see chapter 2). 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. [86]

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 [40]. This tool promises to have further applications in cancer, such as identifying cancer-related mutations, new drug targets and new biomarkers.

1.6 Chapter 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 developing 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.

Chapter 2

Developing the bioinformatics pipeline (1) Mass Spectrum Sequential Subtraction (MSSS) Method.

2.1 Chapter Abstract

In this chapter, I describe the development of the novel bioinformatics method Mass Spectrum Sequential Subtraction (MSSS) to search large peptide spectra datasets produced by liquid chromatography-mass spectrometry (LC-MS/MS) against protein and large-sized nucleotide sequence databases. The main principle in MSSS is to search the peptide spectra set against the protein database, followed by removal of the spectra corresponding to the identified peptides to create a smaller set of the remaining peptide spectra for searching against the nucleotide sequences database. Therefore, we reduce the number of spectra to be searched to limit the peptide search space. Comparing MSSS and conventional search approach using a dataset of 27 LC-MS/MS runs of rice culture cells indicated that MSSS reduced the search queries to 50% and the search time to 75% on average. In addition, MSSS had no effect on the identification false-positive rate (FPR) or the novel peptide sequences identification ability. We used MSSS to analyze another dataset of 34 LC-MS/MS runs, resulting in identifying additional 74 novel peptides. Proteogenomic analysis with these additional peptides yielded 47 new genomic features in 24 rice genes plus 24 intergenic peptides. These results demonstrate the utility of MSSS in searching large databases with large MS/MS datasets for proteogenomics.

2.2 Introduction

Recent advances in genome sequencing have resulted in an enormous expansion of sequenced genomes. The genome online database (http://genomesonline.org/) currently exceeds 18,040 completed and ongoing genome-sequencing projects (June 2012). However, the genome sequence alone is not sufficient to elucidate biological functions and, therefore, the primary task in connection with any newly sequenced genome is to annotate the genomic sequence in order to attach biological meaning to it. [2, 15] The genome annotation has two levels: i) the structural level, identifying the gene structure, and ii) the functional level, identifying the biological or biochemical function of the gene product.[6, 87]

To perform genome annotation, genome sequencing projects usually relay on transcriptional evidence, such as expressed sequence tags (EST), and a variety of *de novo* tools for gene finding and protein prediction.[11, 15, 51] Although cDNA and EST can provide evidence for expression of a predicted gene, they still rely on the untranslated mRNA. Thus, they cannot confirm expression at the protein level.[15] Consequently, though the *de novo* tools for gene finding and protein prediction vary in their algorithms and accuracy, they predict large numbers of gene models that suffer from errors in reading frames and exon definition, and are sometimes highly redundant.[2, 26, 88-90] These limitations indicated the need of another source of information to help correct and confirm the predicted gene models.

Shotgun proteomics approaches using liquid chromatography-tandem mass spectrometry (LC-MS/MS) directly measure the protease-digested peptides derived from the expressed proteins and therefore, allows confirmation/correction of the expressed coding regions.[2, 87] Typically, the MS/MS spectra are searched against a reference protein database to identify peptides and proteins using algorithms such as Mascot [59], SEQUEST [58], X!Tandem [55], PepSplice [62] and pFind, [60, 91, 92] which identify the peptides and their parent proteins. This limits the identification to the sequences available in the employed database. Thus, the incompleteness of the protein databases causes many high-quality spectra to remain unidentified due to the absence of the corresponding amino acid sequence, and limits the identification to known and predicted proteins.[7, 13, 28, 74]

Searching the MS/MS spectra against the genomic database is a well-known approach that permits wider identification of peptide sequences, as the database will include almost all possible sequences generated by the organism.[2, 69, 93] Since the drafting of the human and *Arabidopsis* genomes, this approach has been widely employed in finding novel genomic features using MS/MS data.[7, 83, 93, 94] In early work, the number of MS/MS spectra in the sample employed for searching against the whole genome database was limited. For instance, the whole draft of the human genome (3.3 Gbp) was searched using a test sample containing a total of 169 MS/MS spectra from 22 proteins [83], which made identification possible even with limited computational resources. However, searching large-scale MS/MS datasets against the genome database remains a major challenge due to the enormous size of the MS/MS data and the databases, and due to the linear relationship between search time and database size.[53]

For instance, the human proteome database is about 25 Mbp, while the six-frame translation of the genome database is about 6 Gbp.[30] The six-frame translated genome database of rice is over 25 times the size of the rice protein database (Figure 2.1). Further, in the different updates of the genomic databases, the database size remains almost the same, e.g., the first draft of the human genome database and the current version (HG19) are 3.3 and 3.12 Gbp, respectively. Thus, the principal factor that affects the search time and the required computing resources will be the number of MS/MS spectra to be identified. With large numbers of MS/MS spectra and large numbers of database searches, the time required to perform the search can be very long. For example, 260 days of CPU time was required to run 4,261 X!Tandem searches against the *Shewanella* genome, even though a PRISM computing cluster with 32 processing nodes was used.[56, 57] Thus, several methods have been developed to facilitate this kind of search. [95]

Sevinsky *et al*. (2008) made it possible to search the whole human genome using common desktop computers though the GENQUEST method, which utilizes isoelectric focusing of peptides and accurate peptide mass to reduce the search space.[75] Bitton *et al*. (2010) developed an integrated method to search the whole human six-frame-translated genome database by splitting the database per chromosome, creating 23 target and decoy databases, then eliminating non-matching peptides; this significantly reduced the search space. [74]

Edwards (2007) successfully reduced the human EST database by 35-fold by means of a sophisticated database compression strategy requiring the EST to be mapped to the vicinity of a known gene, and the peptide to be contained in a 30-amino-acid open reading frame (ORF), in which the peptide sequence is confirmed by at least two ESTs,, followed by elimination of peptide sequence repetition. This makes search against the human ESTs database possible with affordable computational resources.[53] The exon-graph method, proposed by Tanner and colleagues, was successfully used to identify novel genomic features in human and *Arabidopsis*.[11, 25] Mo *et al*. (2008) and Power *et al*. (2009) presented methods to identify peptides that overlap the exon-exon junction or exon-skipping events in human, respectively, by creating databases containing only the features that they are targeting.[13, 28] Fermin and colleagues identified 282 significant ORFs in human by creating an ORFs library for all possible reading frames of the human genome after splitting the genome per chromosome.[72]

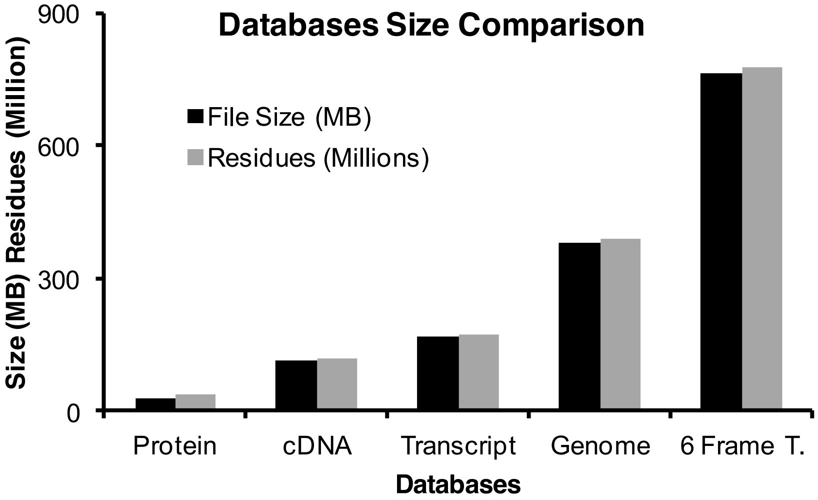


Figure 2.1 Comparison of different rice databases in terms of file size and number of residues.

Further, in order to speed up tandem mass spectra identification, some search engines, such as SEQUEST, pFind and Crux, [58, 60, 65] use peptide indexing.[66] To create a peptide index, the proteins are digested *in silico* and information such as mass, length and position are calculated for each peptide. The list is filtered later to remove redundancy and the final list is saved to disk, from which the engine can load it and match it with the spectra.[95] However, there are several drawbacks, such as the time required to construct the index and the need to re-construct the whole index if there is any change in the searching parameters. Further, for non-specific digestion, the time required for constructing the index can be increased by 100-fold. [66]

Although these efforts have made it possible to search a large-sized database using MS/MS spectra, all of the procedures are dependent on either preprocessing of the database before searching or searching only a selected portion of the database which contains certain features. The preprocessing, such as splitting the database per chromosome, increases the overall processing time and effort, while searching against only certain features such as ESTs or exons causes loss of a significant part of the genomic information included in the omitted portion. Therefore, a new bioinformatics method that allows searching large datasets of MS/MS spectra of peptides against large databases, with reasonable computational cost and search time, is required.

In this work, I propose a novel bioinformatics approach, Mass Spectrum Sequential Subtraction (MSSS), which facilitates the identification of novel peptide sequences from large-scale MS/MS peptide spectra datasets and protein sequence and genomic databases. In MSSS, we search the MS/MS spectra against a reference database, e.g. a database containing putative protein sequences, then remove the spectra corresponding to all identified peptides, creating a new file that contains only the unidentified spectra. Then, we use the new file to search a nucleotide sequence database to identify novel peptides, which cannot be derived from any annotated protein.

The spectra subtraction approach is well known approach that is optionally used by several database searching tools to exclude the low-quality spectra or as step in complex process aims to reduce the number of the unassigned spectra. [45, 96] Further, Mascot allows successive rounds of searching the unassigned spectra against different databases with different search parameters. [59] However, we adopted this approach in MSSS to increase the novel peptides identification for applications in proteogenomics. MSSS avoids the redundancy when searching the same MS/MS dataset against several databases and works an integrated step that is independent from the employed search engine or bioinformatics pipeline, so that, it can be integrated within any existing proteomics data-analysis pipeline. Following this approach, we can reduce the number of spectra to be compared with the database instead of reducing the database size, and consequently reduce the number of the queries to be performed by the search engine, thereby reducing the search time and computational demand.

2.3 Results and Discussion

2.3.1 Assessment of the MSSS approach

For evaluating the MSSS approach shown in Figure 2.2A, our published data from shotgun MS-based proteome analyses of rice cultured cells was employed, as a typical testing dataset. [14] The dataset consisting of 152,908 MS/MS spectra were searched against the rice protein, cDNA, transcript and genome databases of Michigan State University (MSU), formerly known as the database of The Institute of Genome Research (TIGR),[97] using Mascot 2.3.[59] The four databases and the searching order were carefully selected to provide a novel outcome from each database. The protein database was used as the reference database, since it contains all annotated proteins and peptide sequences. While the content of the cDNA database corresponds to the protein database content, it allows searching different frame translations of the nucleotide sequence (frame translation is implicitly done by Mascot). The transcript database includes the introns, so we can identify intronic and exon-intron spanning peptides. Finally, the genome database includes the intergenic regions. Thus, each of the four databases offers the possibility of identifying unique features.

Four points should be taken into consideration during the assessment of any new method that aims to speed up the peptide sequence identification process 1) improvement of identification time, as the main purpose of the method, 2) the peptide sequence identification capability should be similar to that of the current methods, 3) the accuracy should not be impaired, and 4) the method should be flexible and easy to integrate into current data analysis workflows.[66]

In order to assess the performance of MSSS, its performance was compared with that of the normal peptide identification approach in three respects: 1) the search time required for the identification, 2) the peptide identification capability and 3) the false-positive rate (FPR) of the identification. The fourth feature, flexibility, is already present since MSSS is an intermediate step that can be easily integrated into any workflow that supports the MGF file format. Figure 2.2C illustrates the two approaches. We compared the two approaches with four different peptide acceptance criteria based on the identity score of the peptides (Table 2.1), to find the most suitable condition for MSSS.

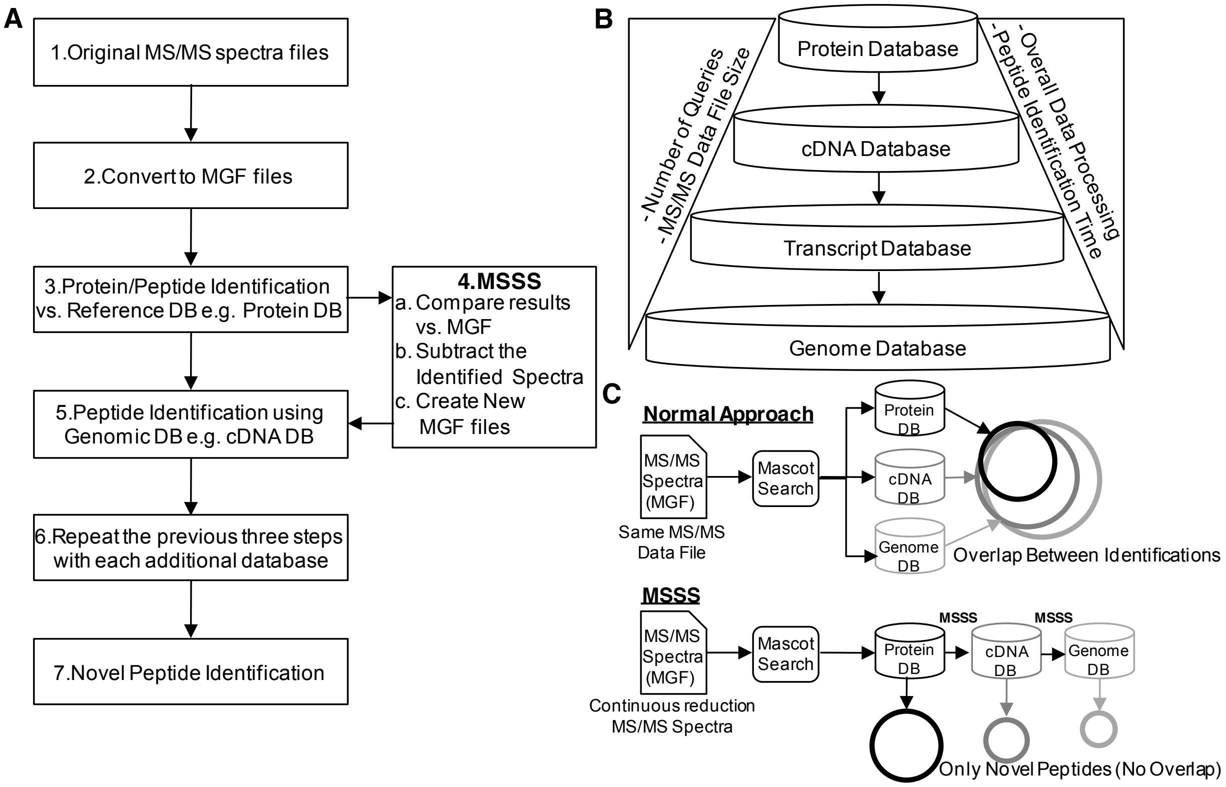

A
1.Original MS/MS spectra files
2.Convert to MGF files
3.Protein/Peptide Identification vs. Reference DB e.g. Protein DB
4.MSSS
a. Compare results vs. MGF
b. Subtract the Identified Spectra
c. Create New MGF files
5.Peptide Identification using Genomic DB e.g. cDNA DB
6.Repeat the previous three steps with each additional database
7.Novel Peptide Identification
B
- Number of Queries
- MS/MS Data File Size
Protein Database
cDNA Database
Transcript Database
Genome Database
- Overall Data Processing
- Peptide Identification Time
C
Normal Approach
MS/MS Spectra (MGF)
Same MS/MS Data File
Mascot Search
Protein DB
cDNA DB
Genome DB
Overlap Between Identifications
MSSS
MS/MS Spectra (MGF)
Continuous reduction MS/MS Spectra
Mascot Search
Protein DB
MSSS
cDNA DB
MSSS
Genome DB
Only Novel Peptides (No Overlap)

Figure 2.2 MSSS method. A) Flowchart of the data analysis in the MSSS method. B) Advantages of applying MSSS in the peptide identification process (increased peptide search space, decreased search time and decreased overall data processing requirement). C) Comparison between MSSS and the normal approach. In this figure, I use three databases for demonstration, while in the actual work we used four databases in the same sequence as shown in B.

Table 2.1 Peptide acceptance criteria used in the MSSS evaluation

Criteria	Identification Confidence	*p* value	Description
C1	95%	P<= 0.05	Mascot score confidence >=95%
C2	99%	P<= 0.01	Mascot score confidence >=99%
C3	99.9%	P<= 0.001	Mascot score confidence >=99.9%
C4	99.99%	P<= 0.0001	Mascot score confidence >=99.99%

Since the main goal of this method is to provide a new strategy that makes peptide identification from large-scale MS/MS datasets with large-sized databases affordable (i.e, with reasonable computational demands and in a shorter time), it is indispensable to demonstrate the influence of the MSSS approach on the size of MS/MS data files and the number of MS/MS spectra to be identified, since these are the two key factors that determine the time and computational resources required for searching a database.

In contrast to the normal approach, in MSSS the file size is reduced after each search due to subtraction of the identified spectra. This is reflected in the number of MS/MS spectra per file and, therefore, the number of search queries to be performed by the search engine. In MSSS, the total file size was reduced by 50% on average due to the sequential subtraction of the identified MS/MS spectra after each database search (Figure 2.3A). The reduction in size was proportional to the number of MS/MS spectra remaining in the files, which was reduced by 45% on average (Figure 2.3B). This means that the total number of search queries to be performed by Mascot was reduced by 45%, resulting in a decrease of search time and required computational resources by 25% on average (Figure 2.3C).

The second comparison between MSSS and the normal approach is the peptide identification capability for i) total non-redundant peptides and ii) novel non-redundant peptides. We defined a non-redundant peptide as a unique combination of sequence plus modifications. Therefore, total peptides are all peptides identified from the four databases and novel peptides are peptides that cannot be derived from any annotated protein (peptides identified from the cDNA, transcript and genome databases). Both total and novel non-redundant peptides identified through MSSS were the same as those identified through the normal approach with various peptide-acceptance criteria (Figure 2.4, Figure 2.5A-B).

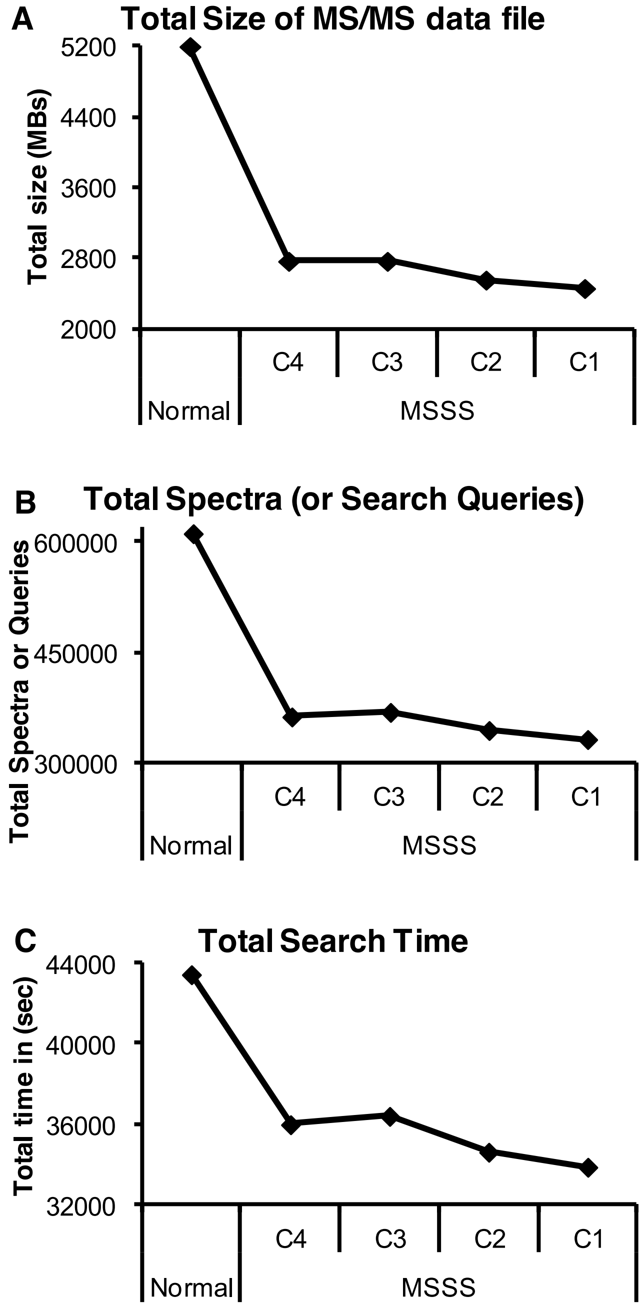


Figure 2.3 MSSS facilitates the identification of peptides from a large-scale MS/MS peptide spectra dataset and large-sized databases. A) MS/MS data file size, B) the number of MS/MS spectra (or the number of search queries performed by Mascot) and C) the search time required for peptide identification was reduced significantly after applying MSSS.

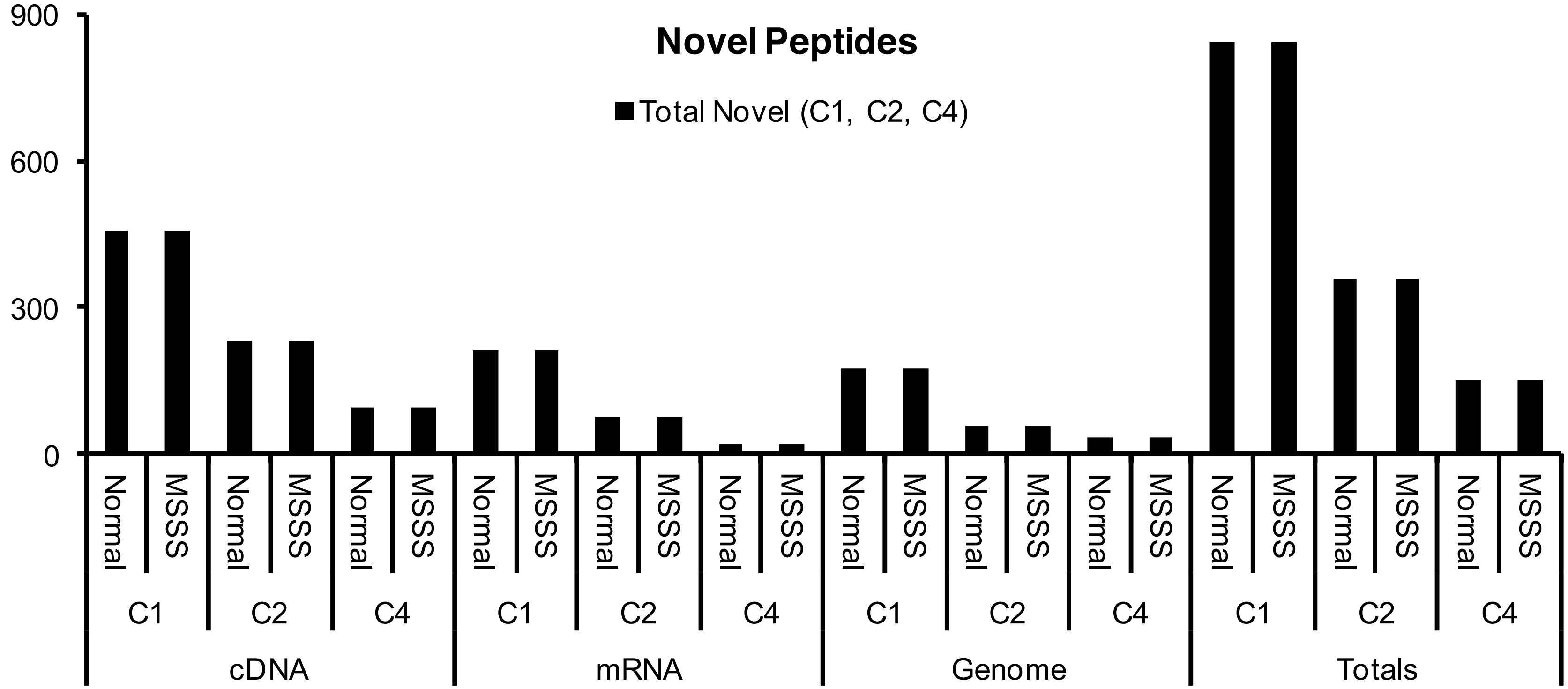


Figure 2.4 Identification of novel non-redundant peptide sequences by MSSS and the normal approach. (C1, C2 and C4 acceptance criteria)

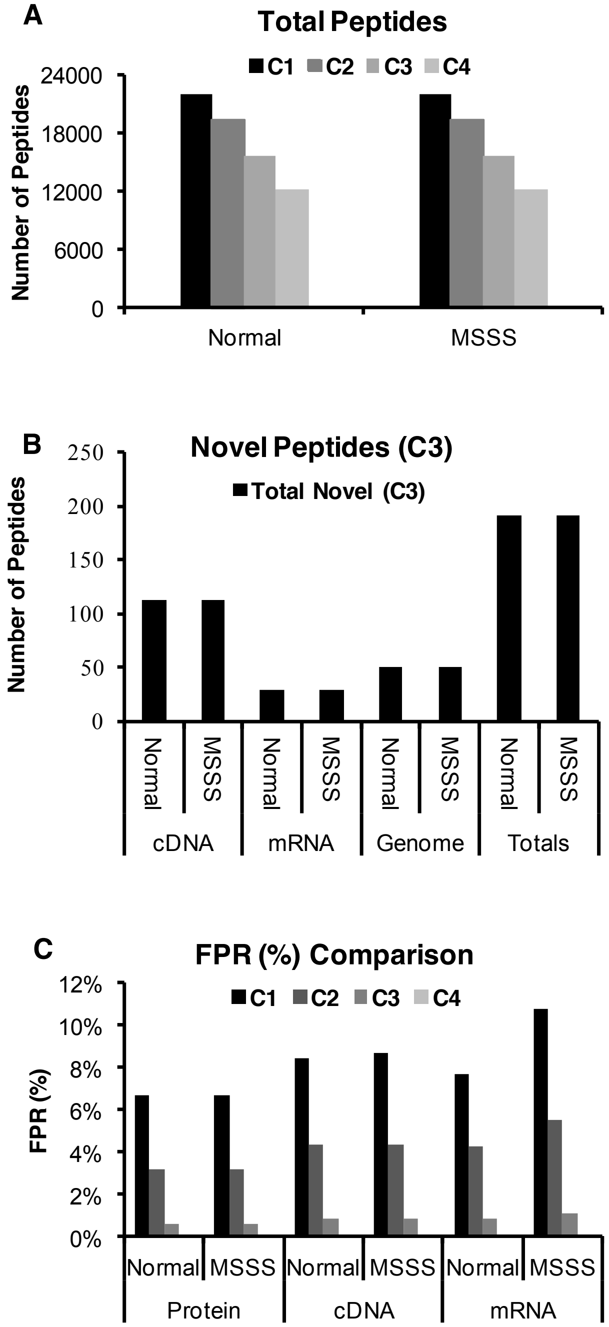


Figure 2.5 Assessment of MSSS performance. A) Total non-redundant peptides identification in MSSS and the normal approach. B) Novel non-redundant peptide sequences identification in MSSS versus the normal approach in C3 (99.9% peptide acceptance criteria). C) FPR (%) in MSSS versus the normal approach. C1~C4 represent the four different peptide acceptance criteria (Table 2.1).

Further, we compared the sources of novel peptide identifications at each acceptance level in MSSS and the normal approach to evaluate the contribution of each database. In all cases, the contribution of each database to the novel peptides was similar in both methods (Figure 2.6). Furthermore, we compared the overlap between the peptides identified in both approaches and we found the same matches in all cases. These results demonstrate that the peptide identification capabilities of MSSS and the normal approach are comparable, regardless of the selected peptide-acceptance criteria.

Finally, we compared MSSS and the normal approach in terms of the false-positive rate (FPR) of peptide identification (see Materials and Methods). Since a decoy database is equal in size to the target database, it was practically not possible to append a decoy version to the genome database, due to its large size (Figure 2.1). Therefore, FPR was calculated for the protein, cDNA and transcript databases only (see Materials and Methods). For the protein database and cDNA database identifications, the false-positive rate was the same in MSSS and in the normal approach with all four peptide acceptance criteria (Figure 2.5C). In the case of the transcript database identification, the false-positive rate was slightly increased in MSSS. However, this increase in FPR was negligible with the third and fourth criteria (Figure 2.5C), which were the two criteria with acceptable FPR (FPR (%) <=1%). Thus, MSSS has a slight effect on the FPR at lower peptide score confidence, but has a negligible effect on the FPR at higher peptide score confidence.

The assessment of MSSS performance thus indicates that MSSS is comparable with the normal identification approach in terms of FPR and peptide identification. However, MSSS offers advantages in terms of reducing the search time (comparing with the normal approach and using the same computational resources), avoids redundant identification of the same spectra when searching multiple databases and facilitating peptide identification from large-scale MS/MS datasets and large-sized databases. Further, MSSS is an intermediate step that can easily be integrated in any existing proteomics data-analysis pipeline that supports MGF file format.

2.3.2 Application of the MSSS approach on phosphopeptide-enriched rice samples

Next, we applied MSSS to another dataset of phosphopeptide-enriched rice samples (with total of 185,126 spectra). [98] The dataset of phosphopeptide-enriched samples was shown to extend the peptides coverage in proteogenomic application. [11] These MS/MS spectra from 34 rice phosphoproteomic samples were searched against the same set of databases used in the above section (protein, cDNA, transcript and genome databases of MSU v6.1) using the MSSS approach. The MSSS search resulted in 5,175, 237, 27 and 31 non-redundant peptides from the protein, cDNA, transcript and genome databases, respectively, when we employed the third criterion of Table 2.1 for peptide identification. Note that the FPR(%) for identified non-redundant phosphopeptides was less than 0.1% (~ 0.08%) as was the case of the first test dataset, since we used more stringent filer with the phosphoproteomic samples (see Materials and Methods). The identified peptides were compared with all peptides of the rice proteogenomics database (OryzaPG-DB) [14] to exclude the peptides that already exist in the database. The comparison resulted in 3,095, 48, 6 and 26 non-redundant peptides identified from the protein, cDNA, transcript and genome databases, respectively, and not existing in OryzaPG-DB.

The 80 novel peptides identified from the cDNA, transcript and genome databases are a useful source of new genomic information, which can be used to refine the genome annotation by applying proteogenomic approaches.[2] Since we used very strict peptide acceptance criteria (99.9%), all peptides passed the statistical quality filters such as score, *e.value* and delta score. In order to further confirm the identified spectra, the spectral quality was manually verified in terms of peak annotation and identified *b* and *y* ions. As a result, the total of 74 novel peptides (42, 6, and 26 peptides from the cDNA, transcript and genome databases, respectively) were accepted. The final dataset was processed using the following steps to find new genomic features that would help in the genome annotation refinement (Figure 2.7A).

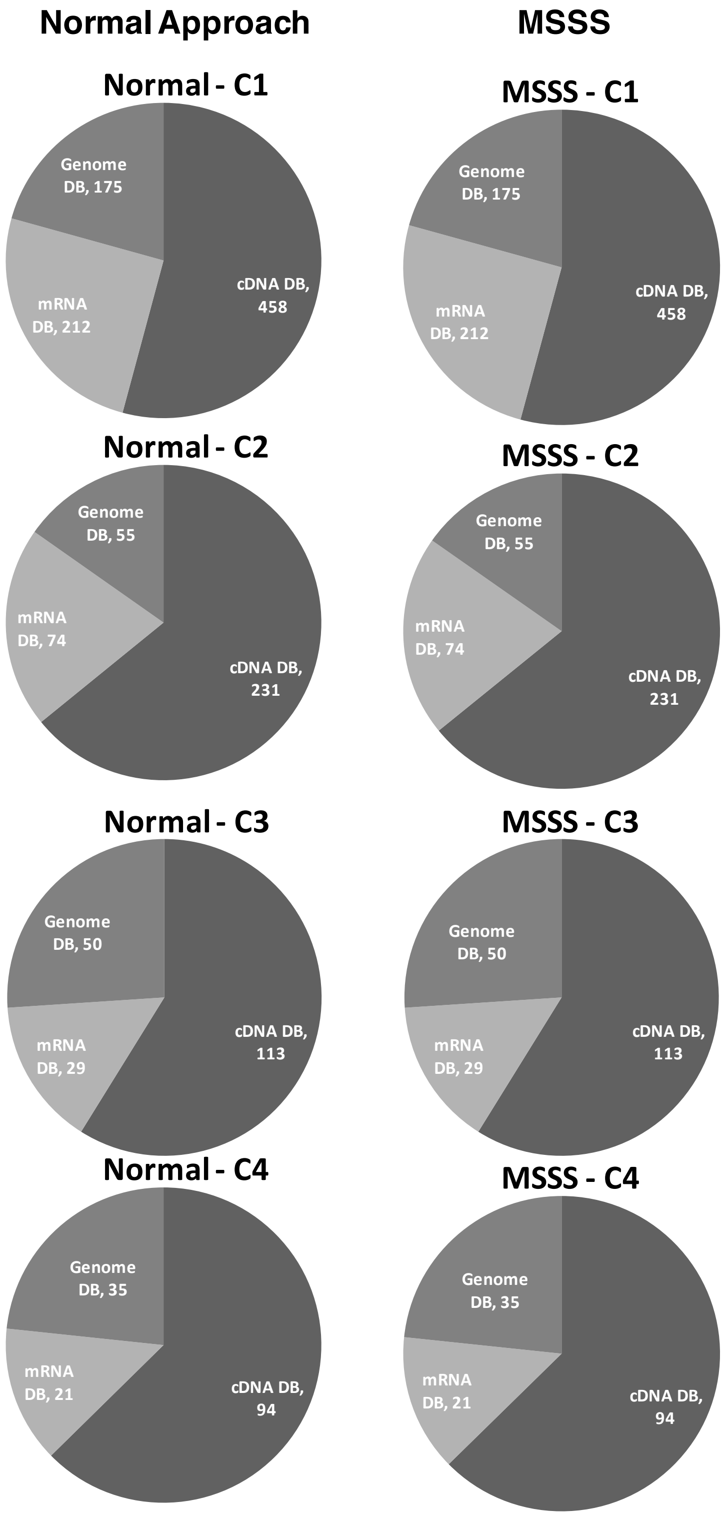


Figure 2.6 Comparison of the sources of the novel peptide sequences identified in this study. The left line shows peptides identified from each database with the normal approach. The right line shows peptides identified from each database with MSSS.

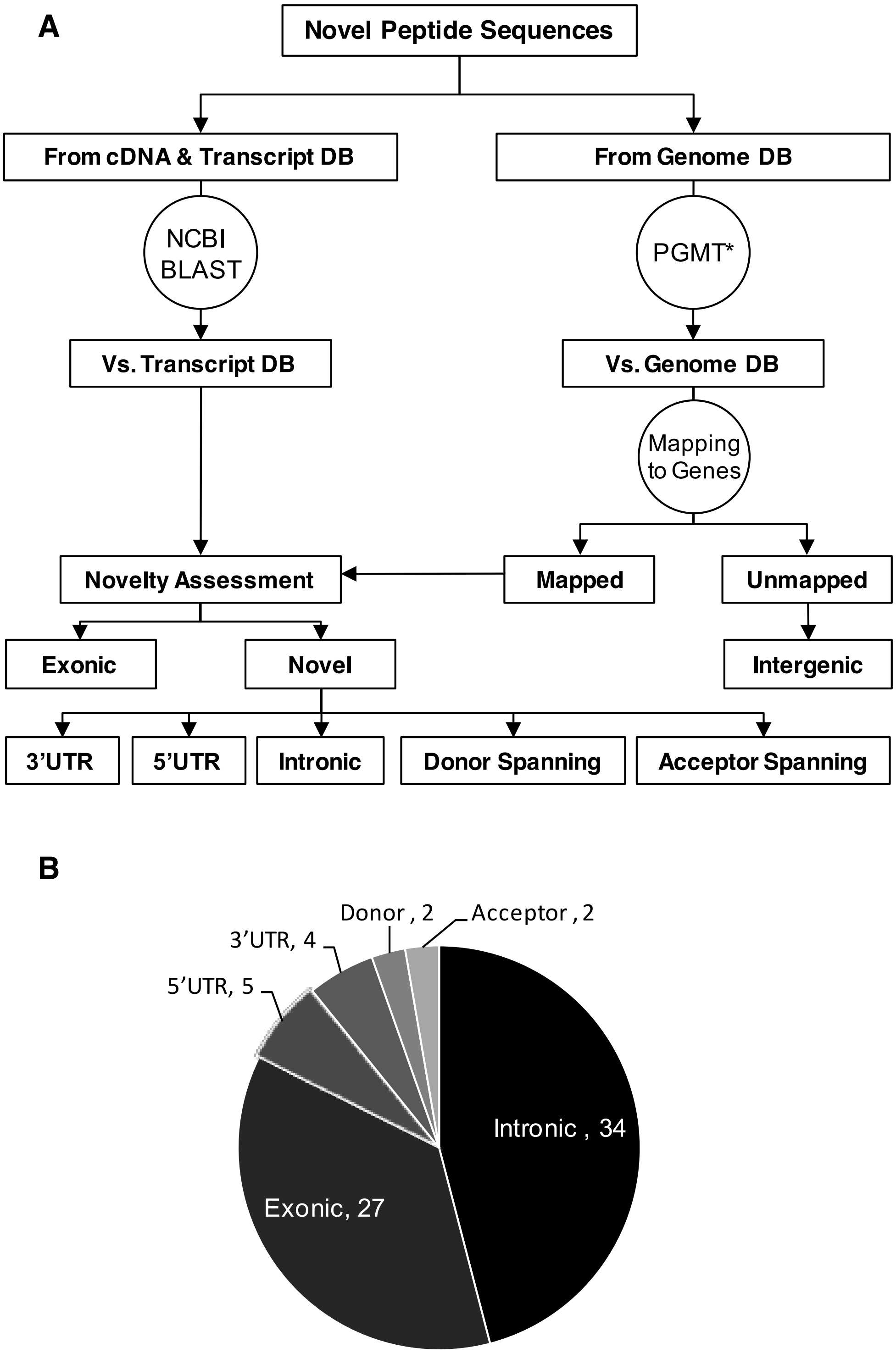


Figure 2.7 Proteogenomic analyses performed using the novel peptides identified. A) Flowchart of the proteogenomic analysis. B) Novel peptides per category. (*) PGMT stands for the proteogenomic mapping tool.[99]

The peptides identified from the cDNA and transcript databases are from 39 genes (see Materials and Methods). Thus, we aligned each peptide to the corresponding unspliced genomic mRNA (transcript). Peptides identified from the genome database were mapped to the genome directly using the proteogenomic mapping tool [99] and the mapping coordinates (start and end) were compared with the gene coordinates to map the peptides identified from the genome database to known genes. The mapping resulted in 24 peptides mapped to intergenic regions; the remaining peptides were mapped to known genes (Tables 2.2 and 2.3). Peptides mapped to intergenic regions can potentially point to new unannotated genes or coding regions.[72, 100]

Peptides mapped to known genes can be either from known coding regions, such as exons, or from novel regions, such as introns or untranslated regions (UTR). Peptides mapped to known coding regions are confirmatory peptides, that can be used to validate the current annotation, while peptides mapped to novel regions can be used to improve the current annotation by adding novel gene features, novel alternative splicing isoforms or new genes.[5] To assess the novelty of each peptide, we used an updated version of our novelty assessment algorithm previously implemented in the ProteoGenomics Features Evaluator (PGFeval) software tool (Chapter 3).[14] The novelty categories of the newer version include intronic, acceptor spanning, donor spanning, exonic and 3'UTR, 5'UTR.

The proteogenomic analysis revealed 47 novel genomic features in 24 genes 22 of them not existing in OryzaPG-DB (Figure 2.8). The majority of the novel features were intronic peptides (34) and UTR peptides (9) (Figure 2.7B) (Tables 2.2 and 2.3). Figure 2.9 shows an example of an intronic peptide with its MS/MS spectra. Table 2.4 shows the comparison between the output of this study and the currently available in OryzaPG-DB.

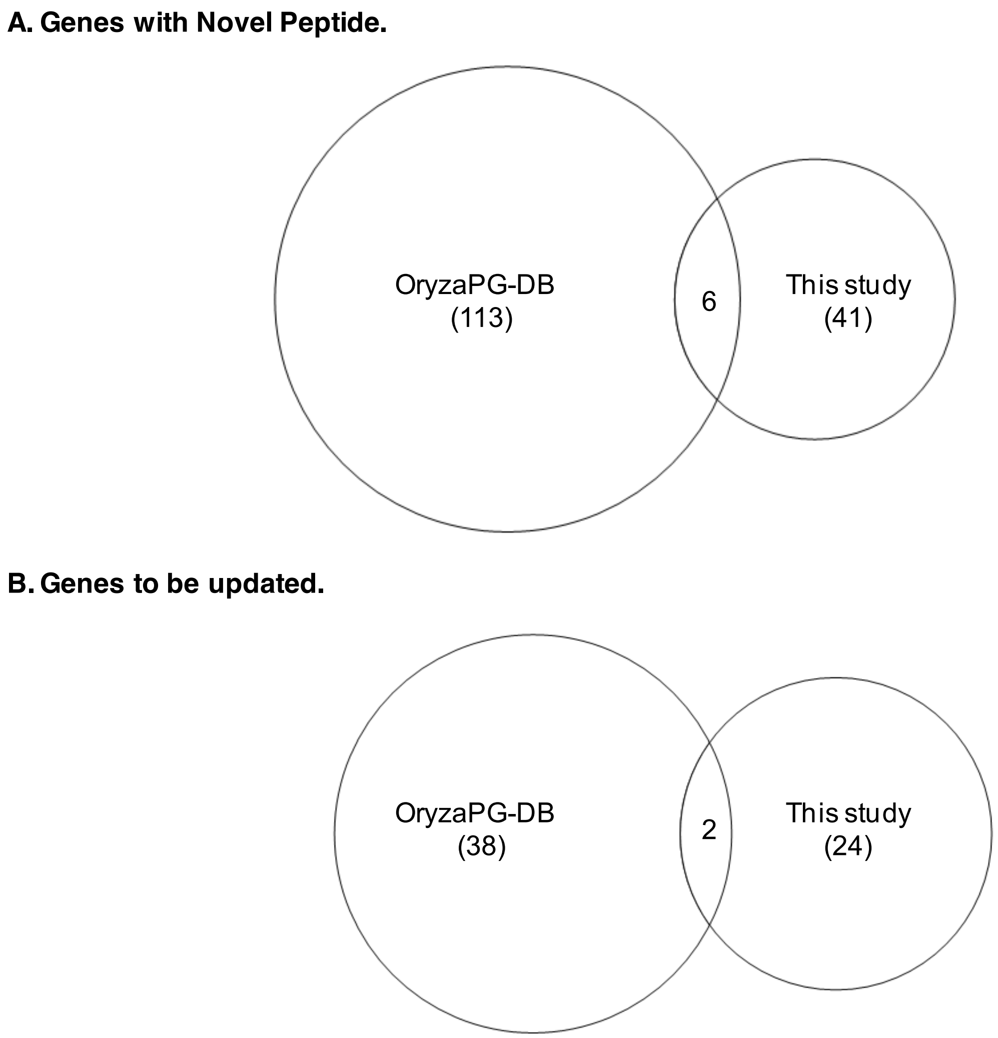
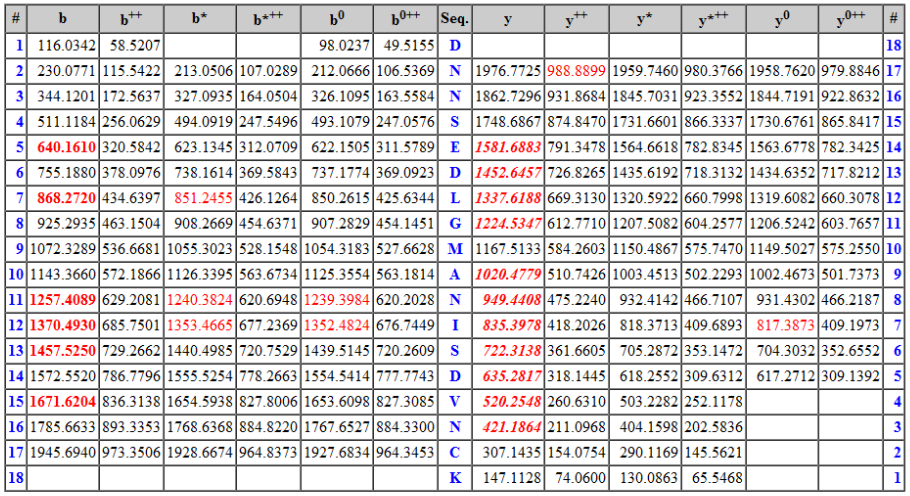
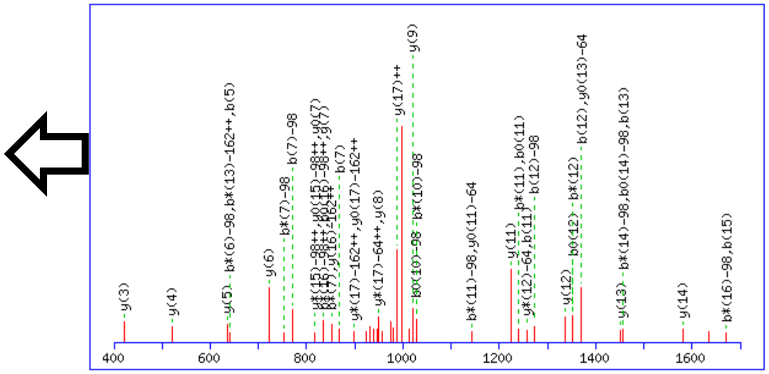


Figure 2.8 Comparison between the output of this study and the OryzaPG-DB content (see chapter 4). A) Genes with novel peptides, including peptides mapped to known coding regions. B) Genes to be updated, with peptides mapped to novel regions. (*) Numbers of genes with novel peptides/features in this study resulted from a proteogenomic analysis using peptides identified solely in the rice phosphoproteome after excluding all peptides shared with OryzaPG-DB. Note: the output of this study is currently included in OryzaPG-DB (v1.1).



#	b	b++	b*	b*++	b0	b0++	Seq.	y	y++	y*	y*++	y0	y0++	#
1	116.0342	58.5207			98.0237	49.5155	D							18
2	230.0771	115.5422	213.0506	107.0289	212.0666	106.5369	N	1976.7725	988.8899	1959.7460	980.3766	1958.7620	979.8846	17
3	344.1201	172.5637	327.0935	164.0504	326.1095	163.5584	N	1862.7296	931.8684	1845.7031	923.3552	1844.7191	922.8632	16
4	511.1184	256.0629	494.0919	247.5496	493.1079	247.0576	S	1748.6867	874.8470	1731.6601	866.3337	1730.6761	865.8417	15
5	**640.1610**	320.5842	623.1345	312.0709	622.1505	311.5789	E	***1581.6883***	791.3478	1564.6618	782.8345	1563.6778	782.3425	14
6	755.1880	378.0976	738.1614	369.5843	737.1774	369.0923	D	***1452.6457***	726.8265	1435.6192	718.3132	1434.6352	717.8212	13
7	**868.2720**	434.6397	851.2455	426.1264	850.2615	425.6344	L	***1337.6188***	669.3130	1320.5922	660.7998	1319.6082	660.3078	12
8	925.2935	463.1504	908.2669	454.6371	907.2829	454.1451	G	***1224.5347***	612.7710	1207.5082	604.2577	1206.5242	603.7657	11
9	1072.3289	536.6681	1055.3023	528.1548	1054.3183	527.6628	M	1167.5133	584.2603	1150.4867	575.7470	1149.5027	575.2550	10
10	1143.3660	572.1866	1126.3395	563.6734	1125.3554	563.1814	A	***1020.4779***	510.7426	1003.4513	502.2293	1002.4673	501.7373	9
11	**1257.4089**	629.2081	1240.3824	620.6948	1239.3984	620.2028	N	***949.4408***	475.2240	932.4142	466.7107	931.4302	466.2187	8
12	**1370.4930**	685.7501	1353.4665	677.2369	1352.4824	676.7449	I	***835.3978***	418.2026	818.3713	409.6893	817.3873	409.1973	7
13	**1457.5250**	729.2662	1440.4985	720.7529	1439.5145	720.2609	S	***722.3138***	361.6605	705.2872	353.1472	704.3032	352.6552	6
14	1572.5520	786.7796	1555.5254	778.2663	1554.5414	777.7743	D	***635.2817***	318.1445	618.2552	309.6312	617.2712	309.1392	5
15	**1671.6204**	836.3138	1654.5938	827.8006	1653.6098	827.3085	V	***520.2548***	260.6310	503.2282	252.1178			4
16	1785.6633	893.3353	1768.6368	884.8220	1767.6527	884.3300	N	***421.1864***	211.0968	404.1598	202.5836			3
17	1945.6940	973.3506	1928.6674	964.8373	1927.6834	964.3453	C	307.1435	154.0754	290.1169	145.5621			2
18							K	147.1128	74.0600	130.0863	65.5468			1



Monoisotopic mass of neutral peptide Mr(calc): 2090.7922
Fixed modifications: Carbamidomethyl (C) (apply to specified residues or termini only)
Variable modifications: S4 : Phospho (ST), with neutral losses 0.0000(shown in table), 97.9769 **M9 :** Oxidation (M), with neutral losses 0.0000(shown in table), 63.9983
Ions Score: 81 **Expect:** 1.6e-006
Matches : 46/470 fragment ions using 36 most intense peaks

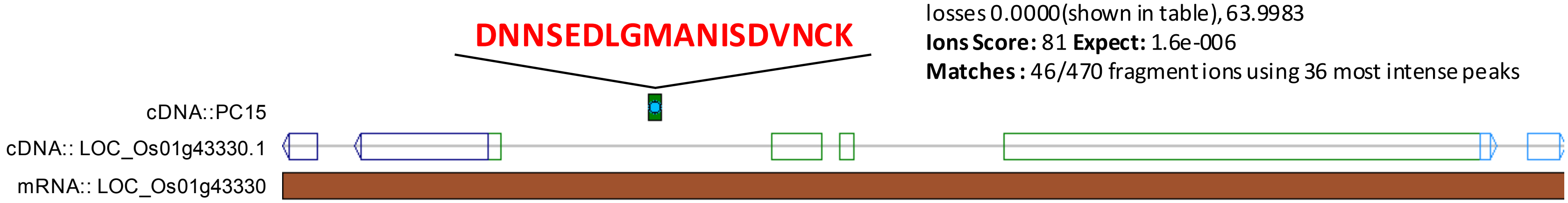


Figure 2.9 Example of a novel peptide with MS/MS spectra. The peptide was identified from the cDNA database and mapped using PGFeval [14] to an intronic region on its genomic mRNA. The peptide's amino acid sequence and spectra are shown in the upper right panel while the matched ions are shown in the upper left panel.

These results demonstrate the utility of MSSS as a novel method to maximize the utility of the MS/MS spectra in proteogenomic studies. For instance, in the above-mentioned *Arabidopsis* study,[7] 1,354 MS/MS runs were performed and identified using 261 novel peptides, while in our MSSS study we have 34 nano LC-MS/MS runs (~0.025% of the *Arabidopsis* study MS/MS runs) and identified 74 novel peptides (~0.3% of the *Arabidopsis* study novel peptides), although we have to consider that different MS instruments used in the two studies (ion trap in the previous study and ion trap-orbitrap in our study).

In addition to its utility as demonstrated above, MSSS can be used to find mutated or abnormal peptides related to diseases that cause somatic mutations, such as cancer.[40] For example, the cancer proteome can be compared against the "normal" protein database, then the genome database using MSSS. Next, the remaining spectra can be compared against a cancer-driven database, e.g., the cancer transcriptome database. In this case, the identified peptides will be related to the disease condition, e.g., mutations caused by the disease, and, therefore, should be useful to find new biomarkers or new drug targets.

2.4 Materials and Methods

2.4.1 Datasets for method development and application

Our published dataset from 27 LC-MS/MS analyses of rice cultured cells[14] was used for the method development, while another published dataset from 34 LC-MS/MS analyses of phosphopeptide-enriched rice tryptic peptides [98] was used for the proteogenomic application of the method.

2.4.2 Database search

Peptides and proteins were identified by Mascot v2.3 (Matrix Science, London, U.K.) [59] against the Michigan State University (MSU) rice protein, cDNA, transcript (unspliced genomic mRNA) and genome databases.[97, 101] Mascot identification parameters were; carbamidomethyl (C) as a fixed modification and acetyl (protein N-term), Gln->pyro-Glu (N-term Q), Glu->pyro-Glu (N-term E) and oxidation (M) as partial modifications for the method development dataset. Phosphorylation (S, T and Y) as additional partial modifications were employed for the application dataset. The product ion mass tolerance was 0.80 Da, while the precursor ion mass tolerance was 3 ppm and strict trypsin specificity was employed, allowing for 2 missed cleavages only. In all Mascot searches, peptides were rejected if the Mascot score was below the 95%, 99%, 99.9% or 99.99% confidence limit based on the identity score of each peptide (Table 2.1) (see Results and Discussion). To increase the identification accuracy and peptide specificity, we accepted peptides with at least seven amino acids[93] and rejected the peptide if the delta score between the first and second hits was less than 10. For the phosphopeptides identification, we require at least three successive *y*- or *b*- ions with a further two or more *y*-, *b*-, and/or precursor-origin neutral loss ions to be observed. In cases where different identification results were obtained from two databases for the same spectrum, i.e., one from the protein database and the second from the cDNA, transcript or genome database, we selected the hit with higher significance (smaller *e.value*).

2.4.3 Mass Spectrum Sequential Subtraction (MSSS)

After obtaining the MS/MS spectra by means of the Materials and Methods described above, we converted the raw data files to Mascot Generic Format (MGF) (Figure2.2A step 2). Next, we performed Mascot search against the protein database (reference database) (Figure 2.2A step 3), then we compared the identification results with the original MGF files, as each identified peptide corresponds to certain MS/MS spectra. To automate this step, we created an in-house web-based tool written in PHP that performs the comparison, subtracts the identified MS/MS spectra and creates new MGF files containing only the unidentified MS/MS spectra (a basic version of the program, written in perl, is also available upon request). The subtraction significantly reduces the file size and the number of MS/MS spectra in the file (Figure 2.2A steps 4a~4c). The new MGF files can be searched against another database such as a cDNA, transcript or genome database (Figure 2.2A step 5). For each database, we repeat the steps of identification, comparison, MS/MS spectra subtraction and new MGF file construction (Figure 2.2A step 6). We end up with novel peptide sequences that do not exist in any of the annotated proteins, identified from searching of multiple genomic databases, with affordable computational demands, reduced search time and reduced overall data processing requirement (Figure 2.2B).

2.4.4 MSSS evaluation scheme

To evaluate the performance of MSSS we used the normal peptide identification approach as a control (shown in Figure 2.2C, top). In the normal approach, the whole MS/MS peptide spectra dataset was searched separately and respectively against the protein, cDNA, transcript and genome databases to obtain the peptide sequences, using Mascot 2.3 (Figure 2.2C, top), then the results of the different searches were combined in an accumulative way (only novel non-redundant peptides from each genomic database are added to the final list). In MSSS, Mascot search was performed against the same four databases, but after each search the identified MS/MS spectra were subtracted and new MGF files were created, then used to search the next database. The searching order in MSSS was protein database, cDNA database, transcript database then genome database (Figure 2.2C, bottom).

2.4.5 Calculating the false-positive rate (FPR)

To calculate the FPR, each of the protein, cDNA and transcript databases was appended with a decoy database since the use of concatenated target-decoy databases is preferable to separated database searches.[73] For each database search result, we calculated the false positives (FP) and the true positives (TP) of the nonredundant set of identified peptides. Next, to calculate the false-positive rate (FPR) of the protein database search result, we used its own FP and TP ($FPR_{protein} = FP_{protein} / (FP_{protein}+TP_{protein})$).[73] For the cDNA and mRNA databases search, we calculated an accumulative FPR. The FPR of the cDNA database search result was calculated from $FPR_{cDNA} = FP_{protein+cDNA} / (FP_{protein+cDNA} + TP_{protein+cDNA})$, while the FPR of the transcript database search result was calculated from $FPR_{transcript} = FP_{protein+cDNA+transcript} / (FP_{protein+cDNA+transcript} + TP_{protein+cDNA+transcript})$. Therefore, we calculated unbiased FPR for both MSSS and the normal approach, avoiding the effect of any anomalous FPR value.

2.4.6 Bioinformatics analysis

The evaluation of a peptide's novelty and visualization of the genomic features were performed using "ProteoGenomic Features Evaluator" (PGFeval) (see results).[14] Sequence alignment was performed using a local version of NCBI BLAST and BLS2SEQ Windows version with the default parameters [47, 49] and perl script. Mapping the peptides identified from the genome database to the genome was done using the proteogenomic mapping tool.[99].

Table 2.2 The genes with novel peptides

#	Gene	PepID	Database	Category	Sequence	Score	Modifications
1	LOC_Os12g32986	PC7	cDNA	Intronic	AQSMGDMSSLDFMR	80.5	15.994915@M:7,15.994915@M:13,79.966331@S:3
2	LOC_Os12g32986	PC8	cDNA	Intronic	AQSMGDMSSLDFMR	75.8	15.994915@M:7,79.966331@S:3
4	LOC_Os12g31800	PC31	cDNA	--	KNDELMDDLFKDDEPDNYANK	71.7	15.994915@M:6
5	LOC_Os12g21890	PC40	cDNA	Intronic	NDKDDVFSDSEAEDGSSK	75.8	79.966331@S:8,79.966331@S:10
6	LOC_Os12g05920	PC41	cDNA	5'UTR	NGSAESALVNPISLDGSDGSDK	73.4	79.966331@S:3
7	LOC_Os11g47760	PC39	cDNA	--	NALENYAYNMR	65.6	15.994915@M:10
11	LOC_Os09g30412	PC25	cDNA	--	GYEVLYMVDAIDEYAVGQLK	82.9	15.994915@M:7
12	LOC_Os08g45110	PC33	cDNA	--	LLLLLLLLLL	50.4	
13	LOC_Os08g23360	PC24	cDNA	Intronic	GVSTLSLGER	76	79.966331@S:6
14	LOC_Os08g02340	PC29	cDNA	Acceptor Sp.	KEEAKEESDDDMGFSLFD	72.3	
15	LOC_Os07g42300	PC6	cDNA	--	APAADDDDDDDVDLFGEETEEEK	73	79.966332@T:19
16	LOC_Os07g08330	PC36	cDNA	--	MANADLGR	65.4	15.994915@M:1
17	LOC_Os07g08330	PC37	cDNA	--	MATLAEAAR	73.3	15.994915@M:1
18	LOC_Os07g01920	PC1	cDNA	Intronic	AASDGDDMDIDGQQSSK	75.6	15.994915@M:8,79.966331@S:3
19	LOC_Os07g01020	PC27	cDNA	--	IAAPYDLVMQTK	79.3	15.994915@M:9
20	LOC_Os06g46000	PC10	cDNA	3'UTR	AVLMDLEPGTMDSVR	65.2	15.994915@M:11
21	LOC_Os06g34710	PC20	cDNA	--	GGEVVLEVSDMDEEDGEDDTDVK	81.5	15.994915@M:11,79.966331@S:9
23	LOC_Os06g14406	PC43	cDNA	3'UTR	SADEGDEDLSEQDDLPLSPPK	69.9	79.966331@S:10
24	LOC_Os06g08280	PC23	cDNA	--	GVGMGVGDPSSPSAR	67.2	15.994915@M:4,79.966331@S:10
26	LOC_Os05g48880	PC45	cDNA	Acceptor Sp.	SSGPSAPDAENIEDLTDDEDDSNE	91.6	79.966332@T:16
27	LOC_Os05g08970	PC18	cDNA	--	GAAAMDVDSGPASD	66.8	15.994915@M:5,79.966331@S:13
28	LOC_Os04g58280	PC48	cDNA	5'UTR	VDSEGVMCGATFK	74.3	57.021465@C:8,15.994915@M:7,79.966331@S:3
29	LOC_Os04g58280	PC49	cDNA	5'UTR	VDSEGVMCGATFK	92.3	57.021465@C:8,79.966331@S:3
31	LOC_Os04g31620	PC38	cDNA	5'UTR	MLQSGLPLDDRPEGARSPSPEPVYDNLGIR	65.3	79.966331@S:17,79.966331@S:19
32	LOC_Os04g27950	PC26	cDNA	--	HTHYSTTTTTTDNGASGDDNNR	95.6	
33	LOC_Os04g27950	PC30	cDNA	--	KHTHYSTTTTTTDNGASGDDNNR	97.8	
34	LOC_Os04g02820	PC46	cDNA	--	STGISLFYEMSDESLK	68.2	15.994915@M:10

35	LOC_Os03g22350	PC4	cDNA	--	AGYGSESEVDDEAATVSLASDVDK	81.9	79.966331@S:7,79.966331@Y:3
36	LOC_Os03g19530	PC9	cDNA	--	AVDAGMMEYDSDDNPIVVDK	74.7	15.994915@M:6,15.994915@M:7,79.966331@S:11
37	LOC_Os02g58220	PC47	cDNA	Intronic	TYSAMGSSSSNGFSEMTTPTSVK	80.8	15.994915@M:16,79.966331@S:9
38	LOC_Os02g54700	PC14	cDNA	--	DLAPQTEAFVDRPPTADGSGPPGGAVK	68.2	79.966332@T:15
39	LOC_Os02g38920	PC44	cDNA	Donor Sp.	SDVNIVSNASCTTNCLAPLAK	75.6	57.021465@C:11,57.021465@C:15,79.966331@S:10
40	LOC_Os02g33490	PC5	cDNA	Intronic	ANSYSAMMEADMENGR	77	79.966331@S:3
41	LOC_Os02g12794	PC2	cDNA	5'UTR	ADNPLYGSSLIEYAHIEQWNDFSATEVDANIGK	98.5	
42	LOC_Os02g10970	PC21	cDNA	3'UTR	GLVSYGDGSPDSAGK	64.9	79.966331@S:9
43	LOC_Os02g10970	PC22	cDNA	3'UTR	GLVSYGDGSPDSAGK	69.8	79.966331@S:9,79.966331@S:12
44	LOC_Os02g07260	PC35	cDNA	--	LSELLGVDVVMANDCIGEEVEK	83	57.021465@C:15,15.994915@M:11
45	LOC_Os01g67126	PC34	cDNA	--	LNALNSSAGADDDDEEEDDE	96.4	79.966331@S:7
47	LOC_Os01g43330	PC15	cDNA	Intronic	DNNSEDLGMANISDVNCK	81.5	57.021465@C:17,15.994915@M:9,79.966331@S:4
22	LOC_Os06g22550	PG3	Genome	--	AQVVPVAEGAAVEGDSPR	98.8	79.966331@S:16
46	LOC_Os01g43639	PG12	Genome	--	KGSPAADEEQSTAAAAVR	98.9	79.966331@S:3
47	LOC_Os12g32986	PM1	us-mRNA	Intronic	AQSMGDMSSLDFMR	77.9	15.994915@M:4,15.994915@M:13,79.966331@S:3
48	LOC_Os11g07850	PM6	us-mRNA	Intronic	VLDPDPDPDAAAAAATGNSPLGGR	86.9	79.966331@S:19
49	LOC_Os10g38489	PM3	us-mRNA	Intronic	FGELDVAEK	67.1	
80	LOC_Os10g32550	PM5	us-mRNA	--	SVAAGMNAMDLR	68.3	15.994915@M:9
51	LOC_Os06g03676	PM2	us-mRNA	Donor Sp.	FDLSDSEDATR	73.7	79.966331@S:4
52	LOC_Os04g54930	PM4	us-mRNA	Intronic	NLSYQYSTGANGR	65.9	79.966331@S:7

Table 2.3 The novel peptides mapped to intergenic regions

#	PepID	Chromosome	Database	Novelty	Sequence	Length	PepScore	Modifications
1	PG1	chr07	Genome	Intergenic	ADSGGYNLNEK	11	64.5	79.966331@S:3
2	PG10	chr05	Genome	Intergenic	GSPSAGDAGGDAPVR	15	86.8	79.966331@S:2
3	PG11	chr02	Genome	Intergenic	ILSGLQSDGDESR	13	79.9	79.966331@S:3
4	PG13	chr01	Genome	Intergenic	KTSFQTDASSLGK	13	95.3	79.966331@S:3
5	PG14	chr12	Genome	Intergenic	KVEPETSSPLANDSQQDAAVGDVDDSR	27	101.1	79.966332@T:6
6	PG15	chr05	Genome	Intergenic	LALTPAGGTDNDGEGTVERPSK	22	99	79.966332@T:9
7	PG16	chr03	Genome	Intergenic	LHNGHNDMDNISEGGSMANDLNDAGELNNGR	31	110.5	79.966331@S:12
8	PG17	chr12	Genome	Intergenic	MTSGDMALGSPVSSQGK	17	67.9	79.966331@S:10
9	PG18	chr02	Genome	Intergenic	NVDESCDMDGSPEEVVSK	18	64.9	57.021465@C:6,79.966331@S:11
10	PG19	chr07	Genome	Intergenic	QISALNISSPSTK	13	70	79.966331@S:9
11	PG2	chr07	Genome	Intergenic	AEDAYHSDEEQYDGGR	16	70.9	79.966331@S:7
12	PG20	chr12	Genome	Intergenic	SELDGSESADPAPPSNALQR	20	83.5	79.966331@S:6
13	PG21	chr05	Genome	Intergenic	SHSISNDLHGVQPDPVAADILR	22	83.7	79.966331@S:3
14	PG22	chr01	Genome	Intergenic	SPCQPSYIDDSLR	13	66.6	57.021465@C:3,79.966331@S:1
15	PG23	chr09	Genome	Intergenic	TLYGVVDGGSSDEDESTSK	19	84.6	79.966331@S:10,79.966331@S:11
16	PG24	chr09	Genome	Intergenic	TLYGVVDGGSSDEDESTSK	19	75.6	79.966331@S:11
17	PG25	chr06	Genome	Intergenic	TNDETLSESGPSNQGESVEMIK	22	91.4	15.994915@M:20,79.966331@S:7
18	PG26	chr06	Genome	Intergenic	TNDETLSESGPSNQGESVEMIK	22	97.5	79.966331@S:7
19	PG4	chr11	Genome	Intergenic	DADGRSDDEVADEYWANHIAR	21	80.3	79.966331@S:6
20	PG5	chr06	Genome	Intergenic	DAESSASSSPFVASSSSPR	19	88.2	79.966331@S:17
21	PG6	chr10	Genome	Intergenic	DIDFSSAGASENEEDDDEPLEK	22	97.9	79.966331@S:10
22	PG7	chr01	Genome	Intergenic	DNDDDDDGNDDDESELAR	18	99.8	
23	PG8	chr12	Genome	Intergenic	EVSGASEHATEMQYER	16	68.3	15.994915@M:12,79.966331@S:3
24	PG9	chr05	Genome	Intergenic	GSFAFAASPR	10	76.5	79.966331@S:8

Table 2.4 The output of this study [1] versus the content of OryzaPG-DB

	OryzaPG-DB	**This Study**
Main goal	Build the rice proteogenomic database	Report the Mass Spectrum Sequential Subtraction Method
Employed dataset(s)	27(rice proteome) samples.	34 (rice phosphoproteome) samples[3].
Confirmatory peptides [2]	15,125	5,175 (3,095 new)
Novel peptides	166	74
Novel genomic features	51	47
Genes with novel peptides	119	41
Genes to be updated	40	24
Intergenic peptides	112	24
Peptides novelty categories	4	6

[1)] Numbers related to this study resulted from a proteogenomic analysis using peptides identified solely in the rice phosphoproteome and after excluding all peptides shared with OryzaPG-DB.

[2)] Peptides identified from the protein databases.

[3)] 27 rice proteome samples were also used only for method development.

2.5 Chapter Conclusion

We have developed MSSS as a new bioinformatics method to facilitate the identification of peptides from large-scale MS/MS datasets and large-sized databases, and demonstrated that MSSS is useful in maximizing the utility of high-throughput mass spectrometry-based shotgun proteomics and phosphoproteomics data in proteogenomics. MSSS decreased the required search time and computational demands without affecting the accuracy or the peptide identification capability, comparing with the normal approach. Further, it makes searching the whole genome database possible without extra preprocessing, reduction of the database features or splitting the database into smaller databases. Although additional improvements may be needed to further optimize MSSS to work with lower peptide score confidence, it nevertheless represents a promising approach with a range of potential applications in proteogenomics and cancer research.

Chapter 3

Developing the bioinformatics pipeline (2) The ProteoGenomic Features Evaluator (PGFeval).

3.1 Chapter Abstract

Integrating the proteome-level information for the refinement of genome annotation is the basic application of proteogenomics. Such integration requires the identification of the peptide/protein sequences, mapping the identified sequences to its genomic origin, evaluating the genomic novelty of the identified sequence and updating the current gene annotation. These steps are considered the shared part in all proteogenomic projects. The peptide/protein sequence identification is usually done using a high-throughput and high performance proteomics approach such as Liquid chromatography–mass spectrometry (LC-MS/MS). However, the other steps remain highly manual and, in some cases, low-throughput. In this chapter, I am presenting the ProteoGenomic Features evaluator (PGFeval) a software tool designed I developed for the evaluation of the proteogenomic novelty of the peptides identified using means of high-throughput proteomic approach. Further, PGFeval visualize the gene annotation including the identified peptides, its positions and its proteogenomic novelty. The input of PGFeval is the current genome annotation and the peptide mapping information to the current genes, both in the standard GFF3 format. PGFeval updates the annotation and evaluate the proteogenomic novelty of each peptide. The output of PGFeval is the updated annotation in standard GFF3 format, the visualization of the gene annotation including the peptides and colored annotations for its novelty and two reports in CSV format. PGFeval represent the first attempt to automate the process of peptide's proteogenomic novelty evaluation.

3.2 Introduction

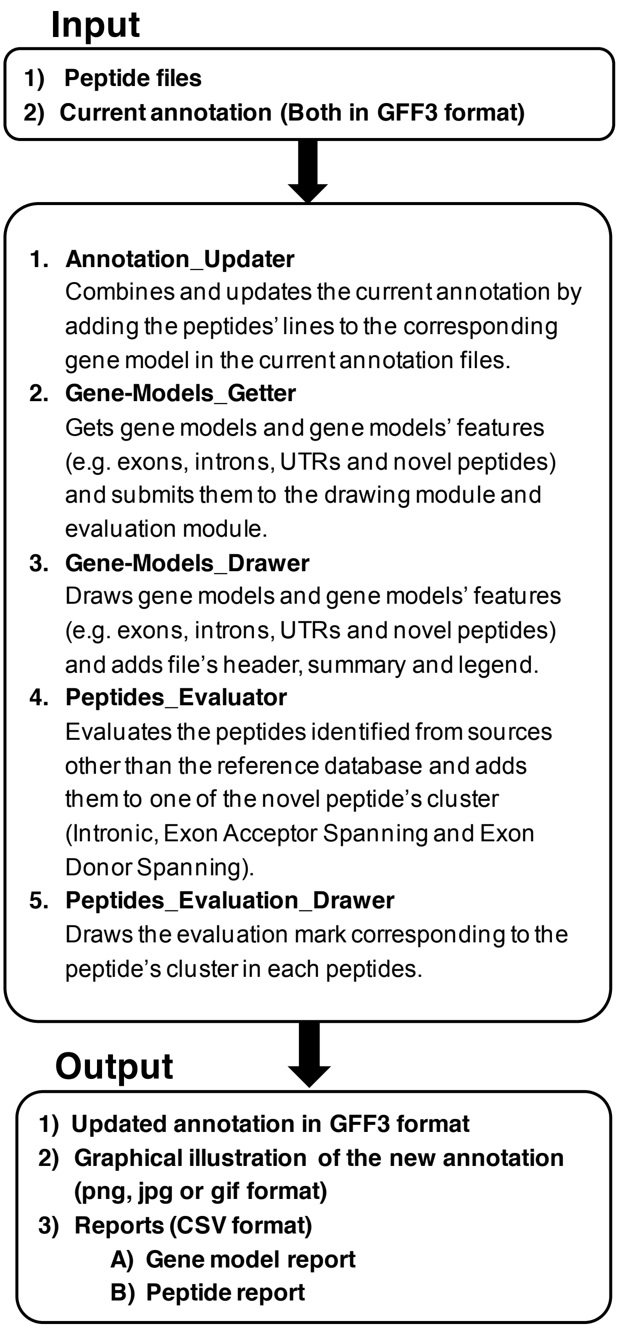
Proteogenomics approaches, which utilize the proteome information for the improvement of the genome annotation, became well appreciated as a relatively easy way to have experimental evidence for the expression of the predicted genes. [2, 4]Recently, proteogenomics applications in genome annotation extended to be integrated in the primary genome annotation process for newly sequenced genome. [3, 95]The main benefit from utilizing the proteome information in the primary genome annotation of a newly sequenced genome or in the improvement of an existing annotation is the existence of an expression evidence for the predicted gene or protein, that is the MS/MS measured peptides. Thus, it became possible to confirm several putative genes, hypothetical genes and conserved hypothetical genes using this approach. [2] Further, it gives the ability of to identify novel genomic features in the annotated genes such as new exons or new alternative splicing isoforms. [2, 13] Furthermore, it is possible, using proteogenomics approaches, to find new non-annotated genes that were never been found using the conventional annotation approaches (computational prediction with transcriptional-based confirmation). [7, 11, 100] Therefore, the details of a proteogenomic analysis/project varied based on the available data, tools and aim of the project (see chapter 1). [95]

Despite the diversities between different proteogenomic analysis/projects, they always contain three proteogenomic modules that differentiate the proteogenomic analysis from other. These three modules are 1. mapping the MS/MS identified peptides to the genome, 2. evaluating proteogenomic novelty of the mapped peptide, and 3. update the current genome annotation or confirm/modify the primary annotation. [95] In general, most of these steps were performed manually and in sometime, in low-throughput rate in many projects. Therefore, automation tools are required to automate these steps.

Among the three modules, the module of peptides mapping to the genome received most of the automation efforts. Mainly, two tools are available now performing this step, the proteogenomic mapping tool and *Pepline*. [77, 99] While, the proteogenomic mapping tool performs the mapping step only, *PepLine* performs peptide sequence identification, peptide mapping to the genome and evaluating the proteogenomics novelty of the peptide. However, there are two main drawbacks for *PepLine*. First, *PepLine* uses peptide sequence tags (PSTs) to perform spectrum interpretation. Therefore, it cannot work with a platform that uses the database searching method for peptide sequence identification. Secondly, *PepLine* is optimized to handle with specific type of MS/MS data, the quadrupole time-of-flight (QTOF) MS/MS data (*PepLine* is reviewed in details in Chapter 1). [77] Nevertheless, these tools represents the eager for automating these three processes, though they are not enough for performing them in all types of genomes and all types of MS/MS data. Furthermore, none of them includes a visualization module that can visualize the gene annotation, the identified peptides and the novelty of each peptide.

In this chapter, I present the ProteoGenomic Features evaluator (PGFeval), a software tool for peptide's proteogenomic novelty assessment and gene annotation visualization. PGFeval analyzes the peptides obtained from the mass spectrometry-based proteome experiments and mapped to the genome and show graphical annotations indicates the proteogenomic novelty of these peptides e.g. peptides from intronic regions, peptide spanning exon acceptor or peptide spanning exon donor (see below). The input of PGFeval is the peptides' mapping results and the current genome annotation (both in GFF3 format), while the output of PGFeval is the updated genome annotation (in GFF3 format), two reports for genes and peptides (in CSV format) and the visualization of the genome annotation per gene and peptide's proteogenomic novelty (in

PNG, JPEG or GIF image format) (Figure 3.1). PGFeval is the first attempt to automate the peptide’s proteogenomic novelty assessment with integrated visualization module.



Input

1) Peptide files
2) Current annotation (Both in GFF3 format)

1. **Annotation_Updater**
   Combines and updates the current annotation by adding the peptides' lines to the corresponding gene model in the current annotation files.
2. **Gene-Models_Getter**
   Gets gene models and gene models' features (e.g. exons, introns, UTRs and novel peptides) and submits them to the drawing module and evaluation module.
3. **Gene-Models_Drawer**
   Draws gene models and gene models' features (e.g. exons, introns, UTRs and novel peptides) and adds file's header, summary and legend.
4. **Peptides_Evaluator**
   Evaluates the peptides identified from sources other than the reference database and adds them to one of the novel peptide's cluster (Intronic, Exon Acceptor Spanning and Exon Donor Spanning).
5. **Peptides_Evaluation_Drawer**
   Draws the evaluation mark corresponding to the peptide's cluster in each peptides.

Output

1) Updated annotation in GFF3 format
2) Graphical illustration of the new annotation (png, jpg or gif format)
3) Reports (CSV format)
   A) Gene model report
   B) Peptide report

Figure 3.1 The design and architecture of PGFeval

3.3 PGFeval architecture and design

The genomic features can be visualized using tools such as the generic genome browser (Gbrowse) or UCSC genome browser [102, 103], but determination of whether or not the peptide represents a novel genomic feature and the type of novelty, e.g., intronic or exon-boundary spanning, cannot be done with these tools. We consider a peptide novel if it does not exist in the protein database. Therefore, all the peptides identified from the other three databases used in the analysis presented in this thesis are considered novel. However, this does not mean that such a peptide represents a novel proteogenomic feature. The peptide may be aligned to a known coding region, but may not exist in the protein database, due to its incompleteness [72, 74]. Hence, we need an evaluation tool and algorithm to assess the genomic novelty of each novel peptide. Therefore, I developed PGFeval (ProteoGenomic Features Evaluator), an evaluation and visualization tool using perl and the GD library (http://www.libgd.org), which evaluates the genomic novelty of each peptide and draws the whole gene model with graphical annotation that incorporates the genomic novelty of the peptides. PGFeval updates the annotation files by merging the peptides GFF3 files with the current annotation GFF3 file. Then, analyzes the updated annotation file and uses the type, start and end columns to draw the gene and its structural elements, such as the UTRs and exons and peptides (see below). PGFeval works on one gene a time and can be used in high-throughput mode, as described below.

3.3.1 PGFeval architecture

The architecture of PGFeval was designed in a modular fashion, where each module performs certain function. Such modularity allows improvement, replacement or fixing part of the program without the need of modifying the whole program, just modifies the required module. [77] PGFeval consists of five modules (Figure 3.1). Here I'll briefly describe each of them.

1. Annotation Updater Module

The *Annotation Updater Module* in the first module of PGFeval to work in case of submitting the standard required inputs (current genome annotation and peptide files both in GFF3 format). This module reads the peptide file(s) and separate the peptides based on the genes that they are mapped to. This, it creates list of genes and the peptides mapped to each gene. Then it reads the annotation file and separates it into individual genes. Thus, it creates list of all genes and list of genes with identified and mapped peptides. Finally, the *Annotation Updater Module* merges the two lists by appending the peptides of each gene to the end of its annotation and excludes genes without identified and mapped peptides. The output of this module is an updated annotation file(s) that contains the peptides mapping information (Figure 3.1). If this file(s) already exists, the modular design of PGFeval allows starting from the next step directly without using the *Annotation Updater Module*.

2. Gene Model Getter Module

The Gene Model Getter Module encapsulates the functions that get the gene model properties and features. The gene model name, number of cDNAs (alternative splicing isoforms), length of each of them, number of exons, introns, UTRs of each of them…etc are some of the gene model propertied and features that the *Gene Model Getter Module* gets in order to do calculations required for drawing the gene model. The getter functions output is saved in a list to be presented to the next module, the *Gene Model Drawer Module* to draw the gene model (Figure 3.1).

3. Gene Model Drawer Module

The *Gene Model Drawer Module* stats by creating a blank image with width and height calculated from the outputs of the *Gene Model Getter Module*. Next, the *Gene Model Drawer Module* start drawing the gene elements starting with the mRNA in the bottom then the cDNAs staked on the top of each other and finally the peptides. For each cDNA, the *Gene Model Drawer Module* draws its components (3'UTRs, 5'UTRs, exons and introns) with different colors and shapes that ease the differentiation of each of them (Figure 3.2). Then, the *Gene Model Drawer Module* draws the peptides identified from each database in one track with different colors (Figure 3.1).

4. Peptide Novelty Evaluator Module

The *Peptide Novelty Evaluator Module* is the core module of PGFeval functionality. It performs the main task of the program, the novelty evaluation. The *Peptide Novelty Evaluator Module* gets its inputs also from the *Gene Model Getter Module* by receiving the coordinates (start and end columns) of each exon and intron in each cDNA and the mapping coordinates of each peptide mapped to the gene. Then, PGFeval uses a special algorithm (see below) to evaluate the proteogenomic novelty of the peptide and if PGFeval found the peptide novel, it continues to cluster the peptide according to its novelty (Figure 3.1). The *Peptide Novelty Evaluator Module* output is saved in a list to be presented to the next module, The *Peptide Novelty Drawer Module* to add the graphical representation of the peptide novelty to the graphical output and to the final report (see below).

5. Peptide Novelty Drawer Module

The *Peptide Novelty Drawer Module* comes in the end of the gene annotation and peptides novelty analysis chain, to visualize the peptide novelty and draw the gene model summary and legend. The *Peptide Novelty Drawer Module* receives its inputs from the *Peptide Novelty Evaluator Module, which* is the peptide ID, its coordinates (start and end) and its novelty category. Using this information the *Peptide Novelty Evaluator Module*, adds graphical annotations to the gene model visualization exported by the *Gene Model Drawer Module*, through adding circles with different colors for each peptide indicating its novelty cluster (Figure 3.2). Then the *Peptide Novelty Drawer Module* adds the gene model summary and the shape and color legends to the image (Figure 3.2). The gene model summery summarizes the gene model properties such as number of cDNAs, number of peptides, number of novel peptides, number of peptides identified from each database and number of peptides from each novelty cluster. Such numbers are all driven from the gene annotation analysis done by the *Gene Model Getter Module* (see above).

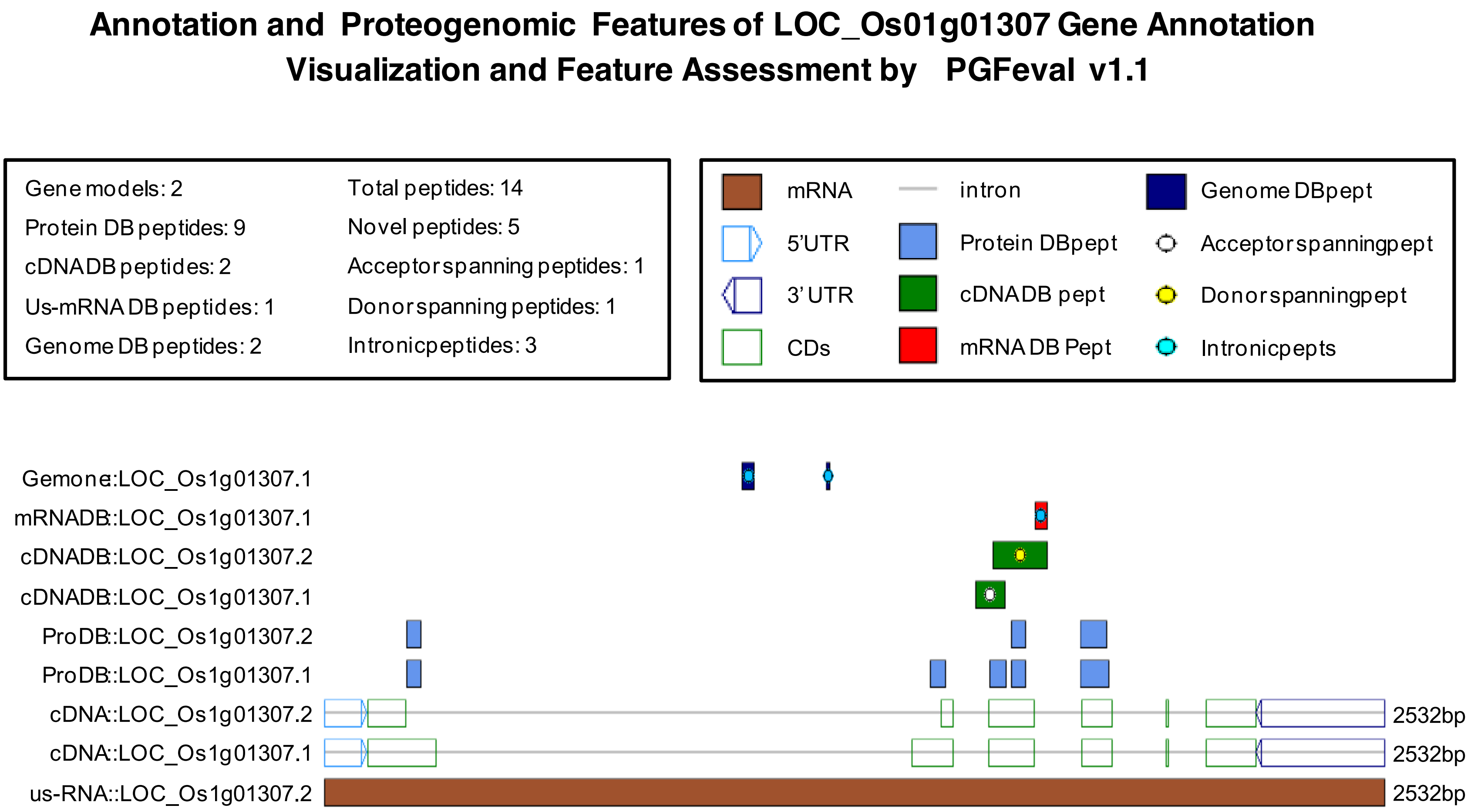


Figure 3.2 Example of the graphical output of PGFeval

3.3.2 Algorithm for peptide novelty assessment implemented in PGFeval

To evaluated the proteogenomic novelty of the peptides, PGFeval uses a special evaluation algorithm that determines if the peptide is novel or not and, if novel, it determine the type of its novelty.

3.3.2.1 Peptide's proteogenomic novelty

Previously, it is mentioned that the peptide is *Novel* if it cannot be driven from the protein database (not existing in any annotated protein). However, the *proteogenomic novelty* of a peptide is whether this peptide is mapped to novel genomic region or from not. Novel peptides mapped to exons, for instance, are not adding new genomic information to our knowledge while, peptides mapped to introns indicate the possible existence of missed exon or even a completely new alternative splicing isoform. Thus, we cluster the peptides according to their proteogenomic novelty into six different categories listed in table 3.1 and illustrated in figure 3.3A.

3.3.2.2 Peptide's proteogenomic-novelty assessment algorithm

The Peptide's proteogenomic-novelty assessment algorithm implemented in PGFeval starts the evaluation by investigating the updated gene annotation resulted from the *Annotation Updater Module* and analyzed by *Gene Model Getter Module* (see above). The algorithm (Figure 3.3b) selects the peptides only for evaluation, by confirming that the *Feature* column value is "peptide" in the GFF3 file. Next, it checks the source of identification of each peptide. Since we do not consider the peptide identified from the protein database novel, the algorithm selects the peptides identified from the nucleotide sequence databases only for evaluation through confirming the identification source using the value of the *Source* column. After creating list of peptides to be evaluated, the algorithm put the other genomic features (e.g. exons, introns, UTRs…etc) in a list to be compared with the peptide list.

The algorithm performs the comparison by comparing the peptide's coordinates (start and end) with the features' coordinated to find if the peptide lays inside any know coding-region such as exons, then it will be a confirmatory peptides that confirms the expression of this region and, therefore, will be added to the *Known* cluster. If the peptide did not match any feature, the algorithm performs the next check by checking if the peptide overlaps the start or end of any feature. If this test returns false, this means the peptide is mapped to an intron and, therefore, will be added to the *Intronic* cluster. While, if the test returns true, the algorithm performs one more test through checking the overlap position. If the peptides overlaps the start of the exon, it will be added to the *Exon Acceptor Spanning* cluster while, if the peptides overlaps the end of the exon, it will be added to the *Exon Donor Spanning* cluster (Figure 3.3B, Table 3.1).

Table 3.1 Novelty terms and categories used in PGFeval

Novelty	Description
Novel Genomic Feature	New genomic information obtained from the proteogenomic analysis performed on the novel peptides.
Novel Peptide	Peptide identified from database other than the protein database (not existing at any annotated protein).
Intronic peptides	Peptide mapped to intronic regions.
Acceptor Spanning Peptides	Peptides spanning the acceptor (start) of the exon
Donor Spanning Peptides	Peptides spanning the donor (end) of the exon.
3'UTR Peptides	Peptide mapped to 3'UTR regions (available in the updated PGFeval version only).
5'UTR Peptides	Peptide mapped to 5'UTR regions (available in the updated PGFeval version only).
Intergenic Peptides	Peptides mapped to intergenic regions.
Genes to be revised	Genes with novel peptides mapped to novel genomic regions.

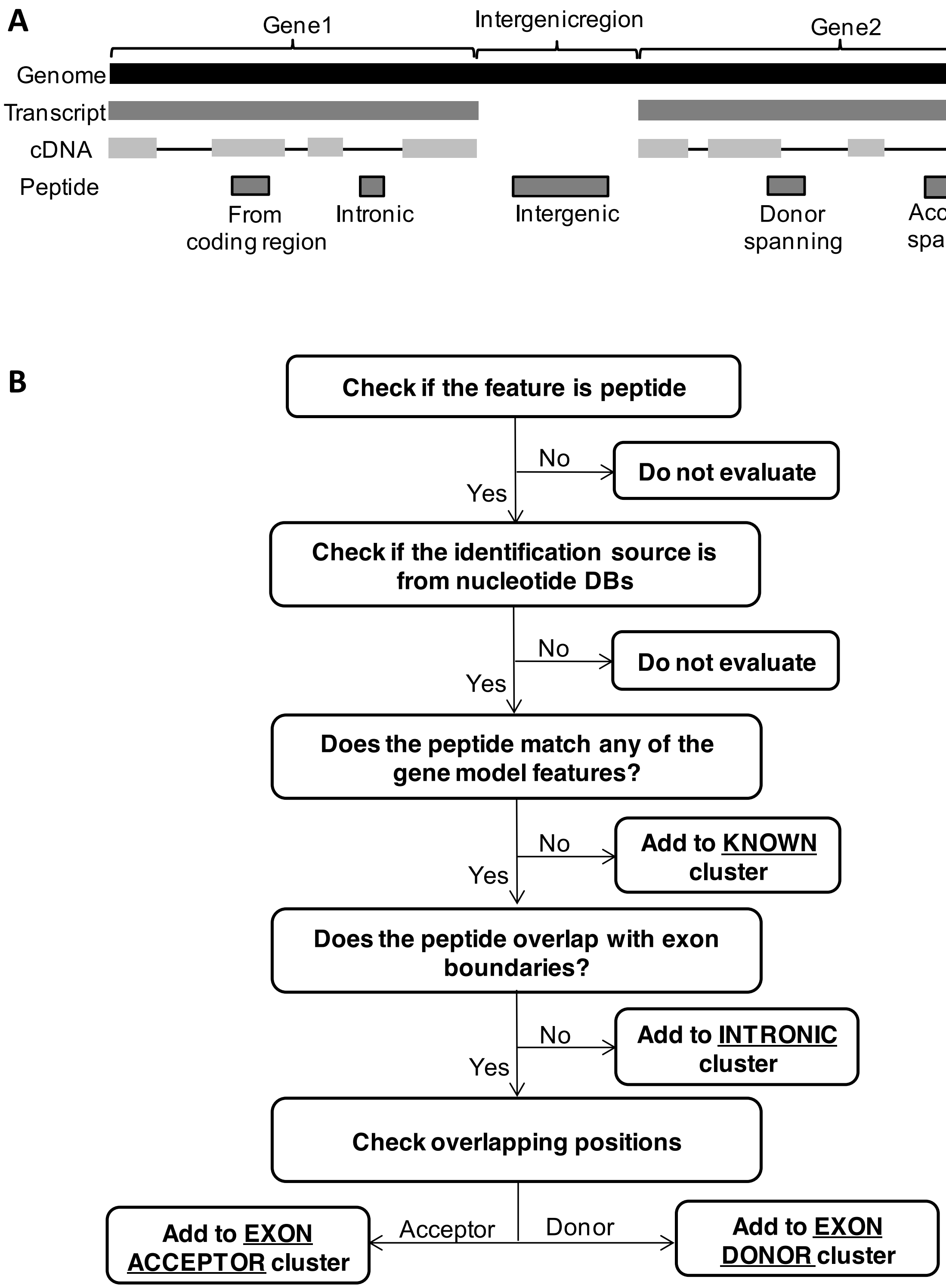

A
Gene1
Intergenicregion
Gene2
Genome
Transcript
cDNA
Peptide
From coding region
Intronic
Intergenic
Donor spanning
Acceptor spanning
B
Check if the feature is peptide
No
Do not evaluate
Yes
Check if the identification source is from nucleotide DBs
No
Do not evaluate
Yes
Does the peptide match any of the gene model features?
No
Add to KNOWN cluster
Yes
Does the peptide overlap with exon boundaries?
No
Add to INTRONIC cluster
Yes
Check overlapping positions
Acceptor
Donor
Add to EXON ACCEPTOR cluster
Add to EXON DONOR cluster

Figure 3.3 Assessment and visualization of a peptide's proteogenomic novelty. A) Schematic illustration of peptide's proteogenomic novelty clusters. B) The assessment algorithm used in evaluating the peptide's proteogenomic novelty in PGFeval.

3.4 PGFeval Output

PGFeval first output (Figure 3.1) is the updated gene annotation GFF3 file results from the *Annotation Updater Module*, which can be in one single file, file per chromosome or file per gene. The second output is graphical visualization of the gene annotation, the identified peptides and the peptide's proteogenomic cluster (Figure 3.2) which can be in PNG, JEPG or GIF image formats. The last output is the analysis reports. PGFeval exports two CSV reports, a genes report and a peptides report, in a master-slave style. The genes report contains one entry per gene summarizing the gene's features such as total peptides, number of novel peptides and novel genomic features, while the peptide report contains one entry per peptide, indicating its gene and assessment result, such as novelty, cluster and identification source. Thus, the two reports can be easily analyzed using any spreadsheet software or imported into any relational database as two tables with one-to-many relationship e.g. the rice proteogenomics database (described in chapters 4 and 5).

3.5 PGFeval updates

3.5.1 Updated peptide's proteogenomic-novelty assessment algorithm

The peptides expressed from the untranslated-region (UTR) are known to have special importance since it plays crucial regulatory and inhibitory roles. [104, 105] Thus, I developed PGFeval to be able to detect the peptide identified from the 3'UTR and 5'UTR. To develop this function, I modified both the peptide's proteogenomic-novelty assessment algorithm and the graphical output. The peptide's proteogenomic-novelty assessment algorithm added two more tests 1) for the peptide of the *Known* cluster, check if the feature that the peptide mapped to is UTR. If this check returns true, it performs the next check to find if this UTR is 3'UTR or 5'UTR and add the peptide to the corresponding cluster (Figure 3.4A). Figure 3.4B shows graphical illustration for the updated peptide's proteogenomic-novelty assessment algorithm.

3.5.2 PGFeval high-throughput mode

The original version of PGFeval was designed to work with one gene at a time. This, however, was inconvenient since, in many cases, one needs to work with hundreds or even thousands of genes in a single analysis. Therefore, I added one more module to PGFeval that added a high-throughput mode to PGFeval to handles high-throughput data (up to thousands of genes). The high-throughput module takes folder of GFF3 files; each file contains one single gene. Then, it passes them one by one to the slandered PGFeval program. The output of the high-throughput mode is the same as the normal mode except that it is multiple outputs gathered in one folder. The updated version of PGFeval was used in the analysis reported in chapter 2 and chapter 5.

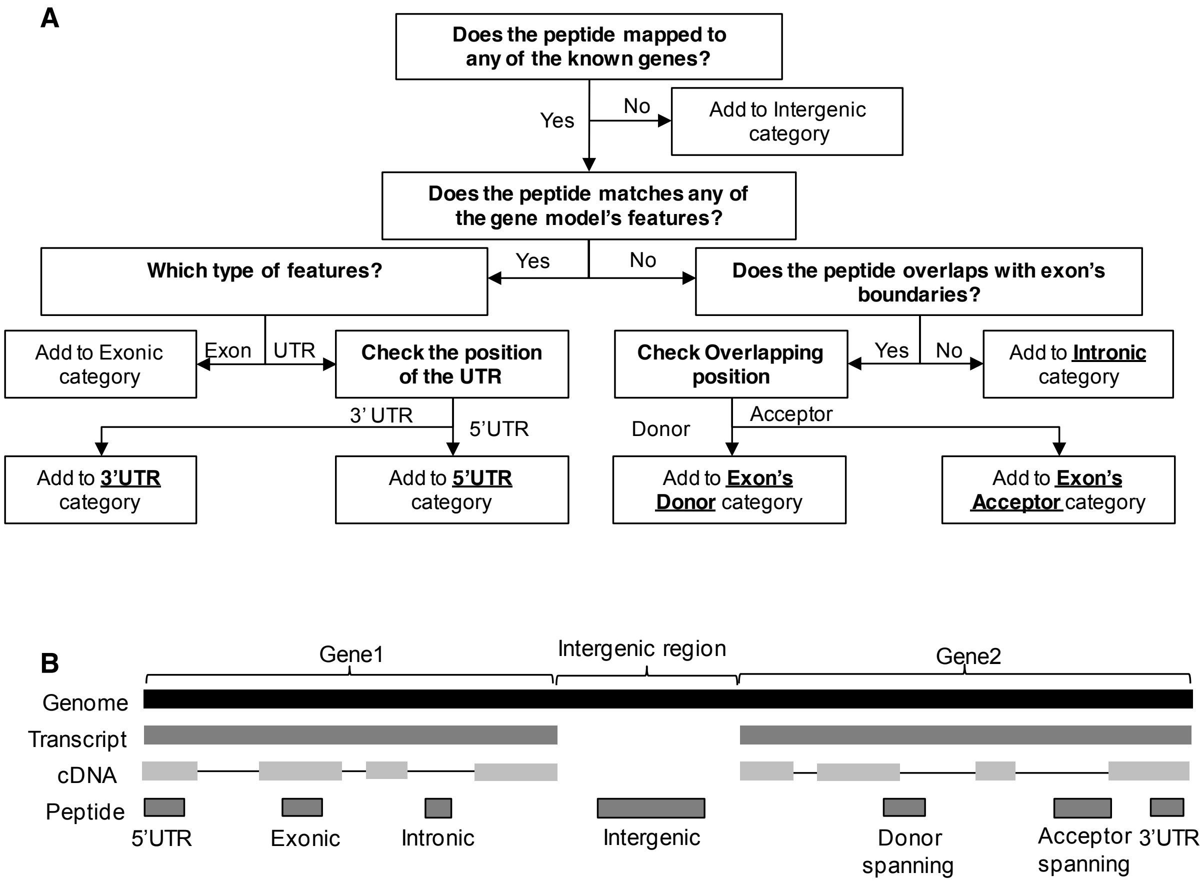

A
Does the peptide mapped to any of the known genes?
Yes
No
Add to Intergenic category
Does the peptide matches any of the gene model's features?
Yes
No
Which type of features?
Does the peptide overlaps with exon's boundaries?
Add to Exonic category
Exon
UTR
Check the position of the UTR
Check Overlapping position
Yes
No
Add to Intronic category
3' UTR
5'UTR
Donor
Acceptor
Add to 3'UTR category
Add to 5'UTR category
Add to Exon's Donor category
Add to Exon's Acceptor category
B
Gene1
Intergenic region
Gene2
Genome
Transcript
cDNA
Peptide
5'UTR
Exonic
Intronic
Intergenic
Donor spanning
Acceptor spanning
3'UTR

Figure 3.4 PGEeval updates A) The updated peptide's proteogenomic novelty assessment algorithm. B) Schematic illustration of peptide novelty categories of the updated version of PGFeval.

3.6 Chapter Conclusion

Proteogenomics, the inclusion of proteome information in the genome annotation process, presents a breakthrough in the advancement of this process. The proteome information adds an experimental confirmation of the genes expression in the translational level. Thus, it confirms the transcription and translation of the predicted genes. Further, it is capable to identify novel non-annotated genes. However, proteogenomics faces several challenges, such as the assessment of the proteogenomic novelty of the identified peptides, which determines if the peptide points novel genomic information or not. In this work, I present PGFeval, the first attempt to automate the proteogenomic-novelty assessment process as well as the visualization of the genome annotation, gene model components, peptides mapped to the genes and the proteogenomic novelty of these peptides. PGFeval was designed in a modular way that allows future improvements and integration with other tools. It was used in several analysis reported in this thesis (Chapters 2, 4, 5) proving its utility in the performed proteogenomic analysis.

Chapter 4

Developing the bioinformatics pipeline (3) The rice proteogenomics database (OryzaPG-DB).

4.1 Chapter Abstract

Proteogenomics aims to utilize experimental proteome information for refinement of genome annotation. Since mass spectrometry-based shotgun proteomics approaches provide large-scale peptide sequencing data with high throughput, a data repository for shotgun proteogenomics would represent a valuable source of gene expression evidence at the translational level for genome re-annotation. Here, I present OryzaPG-DB, a rice proteome database based on shotgun proteogenomics, which incorporates the genomic features of experimental shotgun proteomics data. This version of the database was created from the results of 27 nanoLC-MS/MS runs on a hybrid ion trap-orbitrap mass spectrometer, which offers high accuracy for analyzing tryptic digests from undifferentiated cultured rice cells. Peptides were identified by searching the product ion spectra against the protein, cDNA, transcript and genome databases from Michigan State University, and were mapped to the rice genome. Approximately 3200 genes were covered by these peptides and 40 of them contained novel genomic features. Users can search, download or navigate the database per chromosome, gene, protein, cDNA or transcript and download the updated annotations in standard GFF3 format, with visualization in PNG format. In addition, the database scheme of OryzaPG was designed to be generic and can be reused to host similar proteogenomic information for other species. OryzaPG is the first proteogenomics-based database of the rice proteome, providing peptide-based expression profiles, together with the corresponding genomic origin, including the annotation of novelty for each peptide. The OryzaPG database was constructed and is freely available at http://oryzapg.iab.keio.ac.jp/.

4.2 Introduction

Among high-throughput experimental methods, genome sequencing represents a turning point in the understanding of biological systems. Nevertheless, the biological significance of the sequenced genome cannot be understood unless the protein-coding genes and their products are accurately identified. Thus, genome annotation has become major issue [4, 106, 107]. Genome annotation is the process of gene structure and function determination, and it usually takes place after genome sequencing and before data deposition in a database or databank [87, 107 , 108].

In typical genome annotation work, experimental and computational methods are integrated to analyze the huge volume of sequence data [11, 15, 87, 107]. Thus, genome annotation is highly dependent on the expression evidence, usually transcriptional, provided by experiments and the algorithms implemented in the computational tools [2]. Consequently, the annotation process suffers from several limitations. For instance, most of the sequenced genomes lack rich transcriptional evidence, e.g., a full-length cDNA library. Even when such information is available, evidence of expression at the transcriptional level does not necessarily imply translation into a protein [2, 109]. Therefore, annotation is highly reliant on *de novo* annotations of protein-coding genes performed using gene prediction programs [2, 87, 107].

On the other hand, gene/protein prediction tools have proven their usefulness and utility in the annotation process. However, the prediction accuracy varies from one tool/algorithm to another and from one organism to another, depending on the genome complexity [2, 88, 90, 107]. For instance, in the human and *Arabidopsis* genomes, the prediction accuracy amounted to 50% and ~66%, respectively, indicating the need for better identification and validation methods [88, 89].

Mass spectrometry-based proteomics, as an experimental approach to measure proteins, can provide translation-level expression evidence for the predicted protein-coding genes; this is the so-called proteogenomics approach of using large-scale proteome data in genome annotation refinement [2, 4, 5, 110]. This approach seems the best option for identification and validation of protein-coding genes, or at least a significant portion of them, in an independent and unambiguous way. This can be achieved by detecting the naturally occurring proteins (proteomics) and systematically mapping them back to the genome sequence (genomics) [2, 4, 5, 110]. In addition to validating predicted gene models at the translation level [22, 24], proteogenomics has other useful applications, such as finding new gene models [11], determination of protein start and termination sites [25], finding and verifying splice isoforms at the protein level [13] and verification of hypothetical and putative genes/proteins [25, 31]. The results of proteogenomic studies are usually made freely available via specialized databases such as AgBase [111] or are included in databases developed particularly to host data from specific projects, such as the AtProteome database developed to host the *Arabidopsis* proteogenomics data [7]. Overall, proteogenomics represents a promising approach for application to both completed and newly sequenced genomes.

Rice (*Oryza sativa*) is one of the most important food crops; almost half of the world's population is estimated to rely totally or partially on it. Moreover, rice considered a model organism because of its relatively small genome (12 chromosomes and ~370 Mbp) [112, 113]. The whole genome sequence and annotation have been published and updated several times (5 builds for the genome and 6 builds for the annotation to date) [97, 101, 114]. However, there has been little attempt to include proteome information in the genome-wide annotation, except for the work of Itoh and colleagues, who used rice proteome data, available through the rice proteome database [115], to confirm 834 ORFs [114]. The virtual absence of proteome-based genome annotation for rice is possibly due to the absence of accurate and detailed rice proteome information.

In this chapter, I present OryzaPG-DB, a rice proteome database based on shotgun proteogenomics. Unlike the currently available rice proteome database [115], which provides the 2D-PAGE-based proteome, OryzaPG-DB contains peptides obtained from shotgun-based proteomics with their product ion spectra, as well as updated annotations, side by side with the corresponding protein, cDNA, transcript and genomic sequences and information.

4.3 Construction and content

4.3.1 Generation of a reference dataset by shotgun proteomics

To perform proteogenomics-based genome annotation refinement of rice, we firstly generated a dataset by using a shotgun proteomics approach with an LTQ-orbitrap high-accuracy mass spectrometer for analysis of undifferentiated cultured rice cells (see Materials and Methods). The experimental workflow is shown in Figure 4.1. Three pre-fractionation procedures, SDS-PAGE at the protein level, and strong cation exchange chromatography (7 runs) [116] and isoelectric focusing (7 runs) [43] at the peptide level, were employed prior to nanoLC-MS/MS analysis to extend the proteome coverage. As a result, 156,871 product ion spectra were acquired from 27 LC-MS/MS runs. Undifferentiated cultured rice cells were selected as the first sample to construct our bioinformatics pipeline and data repository system.

Regarding the databases for MASCOT [59] search, the Rice Genome Annotation Project, currently at Michigan State University (MSU), offers rice protein, cDNA, transcript and genome databases (MSU DB) as *de novo* annotation of the rice genome sequence, with further improvements and modifications using full-length cDNA and EST alignment [97] Thus, we decided to use these four databases for peptide/protein identification, since each of them provides a special opportunity to identify novel peptides. The protein database provides all potential protein and peptide sequences. The cDNA database is similar to the protein database, but allows searching six-frame translations of the nucleotide sequence, and therefore, it is possible to identify exon-exon junction peptides and exon-skipping events from the two databases [13, 28]. The transcript database includes introns, so we can identify exon-intron spanning and intronic peptides. Finally, the genome database includes the intergenic regions, offering the potential to find new non-annotated genes. Thus, each of the four databases affords specific search advantages (Figure 4.2).

As a result, we identified 14,955 unique peptides from the database searching against MSU protein database (V6.1). Further, we identified 166 additional unique peptides from three nucleotide sequence databases (cDNA, transcript and genome databases). Then, the obtained peptides were mapped to the annotated MSU genome (V6.1) to get the proteins (cDNAs) and unspliced mRNAs (genes) found in this study. Genes with peptides mapped to novel regions such as intron, exon-intron boundary and non-coding region, are counted as "Genes to be revised", indicating that the annotation of these genes needs revision.

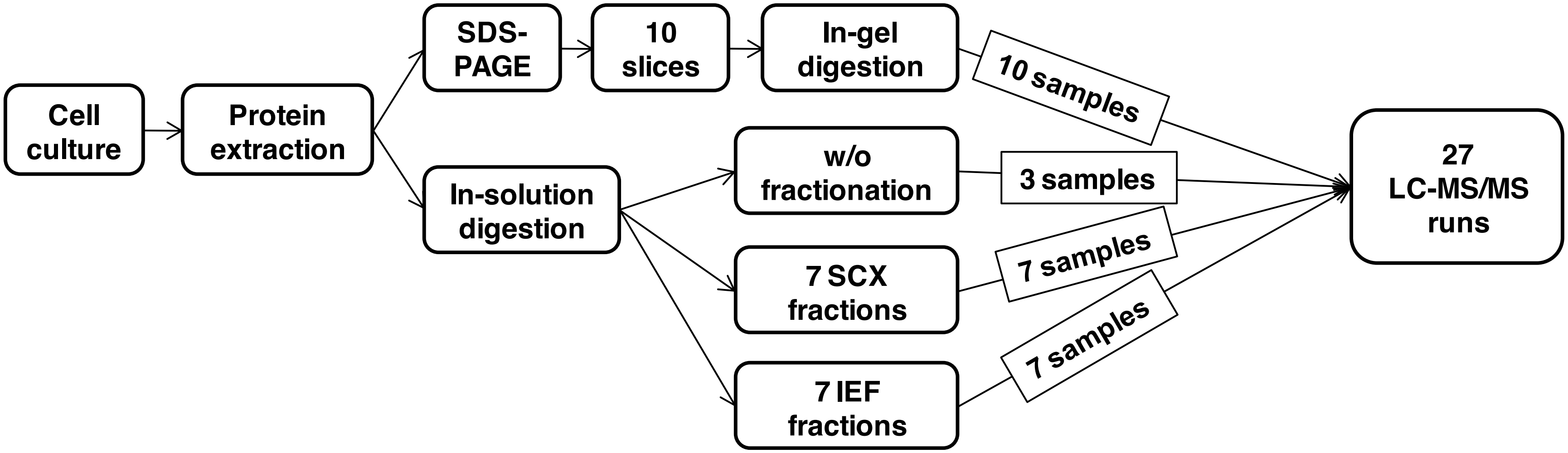


Figure 4.1 The experimental workflow to obtain the rice proteome. Proteins from undifferentiated cultured cells were extracted, digested, and pre-fractionated, and 27 samples were prepared for the subsequent LC-MS/MS analysis.

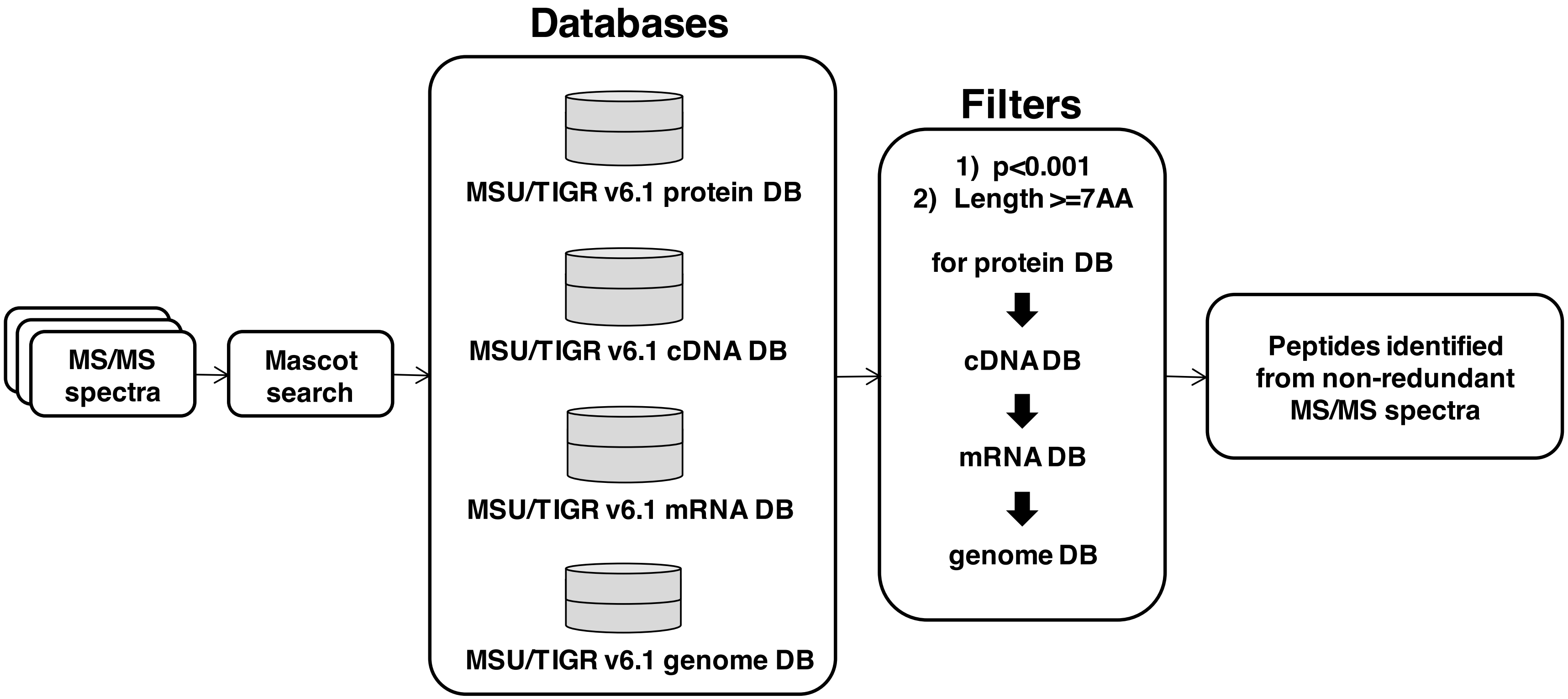


Figure 4.2 Informatics analysis flowchart to create list of peptides/proteins identified from non-redundant product ion spectra.

4.3.2 Proteogenomics analysis to find novel genomic features

Next, we performed proteogenomic data analysis using bioinformatics approaches to map the identified peptides back to the genome and find novel genomic features as follows:

Download the original annotation from MSU genome browser with only the MSU Osa1 Rice Gene Models and MSU Osa1 Rice Loci features selected. The original files can also be obtained from the OryzaPG-DB download page.

Align all peptides identified from the MSU protein, cDNA and transcript databases to their corresponding genomic origin (genomic-unspliced mRNA), using the Basic Local Alignment Search Tool (BLAST) [47]. The alignments were performed using a local version of NCBI BLAST (blast2seq) [49] and perl script.

Extract the alignment results of the peptides identified from the MSU genome database directly from MASCOT output files.

Create perl scripts that read the alignment results and convert them to standard GFF3 format. Each peptide's alignment was converted to a GFF3 line indicating its type, identification source, start, end, parent and OryzaPG-DB peptide ID.

Map the peptides identified from the MSU genome to the genes by comparing the peptide alignment coordinates (start and end) with the gene coordinates. If a peptide's start and end are between or overlapping with a gene's start and end, we map this peptide to that gene and create a GFF3 line similar to the one described above.

Update the original annotation files by appending the peptides' GFF3 lines obtained from our analysis to the end of the corresponding gene. So far, we have created an updated annotation in GFF3 format containing the original annotation and the proteome information. Thus, the "Type" column in the updated GFF3 files includes the type "peptide" beside the original types (3'UTR, 5’UTR, CDS,…etc). However, the identified novel peptides require further analysis to find out whether or not they represent novel genomic features.

The peptide’s preoteogenomic novelty was performed using PGFeval (Chapter 3). The analysis of PGFeval revealed 51 new genomic features in 40 genes. The majority of the novel features consisted of intronic peptides (36), while the exon boundary-spanning peptides consisted of 13 donor-spanning and 2 acceptor-spanning peptides. The remaining novel peptides were mapped to known coding regions.

4.3.3 Generic scheme design for the relational database

We attempt to design a generic and simple database scheme that would be suitable for such proteogenomic data and that could be used in similar shotgun proteogenomics projects. As long as the annotation main unit is the gene, in our design (Figure 4.3), the database main entity is the gene as well. We used the MSU V6.1 locus as our gene ID, joining all other gene components together. All other database entities, such as protein, cDNA, unspliced mRNA and peptide, take the gene ID as a foreign key. Further, protein, cDNA, and un-spliced mRNA have their own IDs and aliases that were used in MSU V6.1. Peptide IDs start with P, C, T or G, which indicates that the identification source of the peptide is the protein, cDNA, transcript or genome database, respectively. The letter is followed by the serial number of the peptide in its sample, *e.g.,* P345 is the peptide number 345 among the peptides identified from the protein database. The peptide ID also joins the peptide with its MASCOT results page and product ion spectral details. In the updated annotation files, the peptides are assigned to their parents using the notation "parent ID:peptide ID" e.g. LOC_Os06g01230.1:P62531.

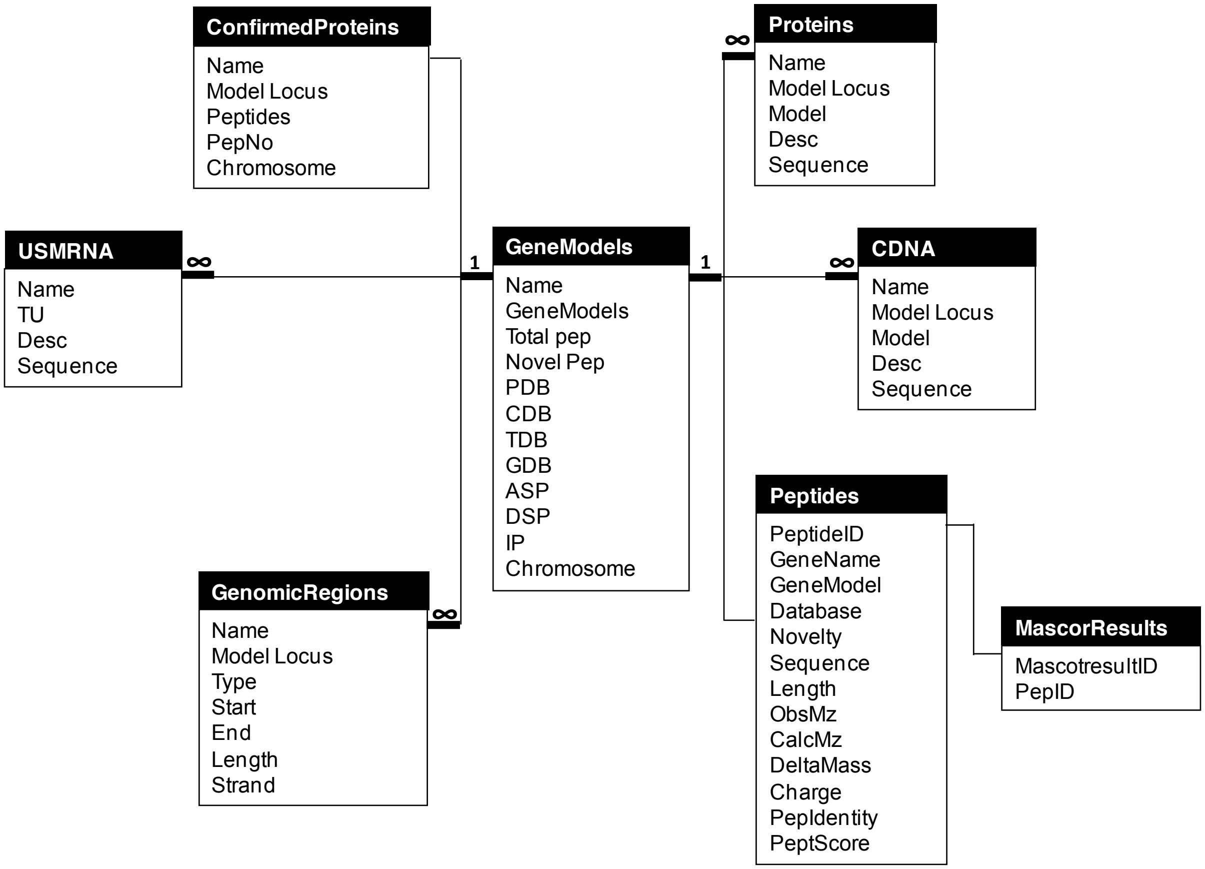


Figure 4.3 OryzaPG database scheme.

4.3.4 Database implementation and web interface development

As mentioned above, PGFeval exports two reports: genes report and peptides report. Both reports are designed in master-slave style. Thus, both were imported directly into the database. The protein, cDNA, and transcript information such as the IDs, aliases, descriptions, lengths and sequences were extracted from the FASTA files and the GFF3 files obtained from the MSU website and MSU genome browser, then converted to tables using perl scripts. Next, HTML files, similar to MASCOT peptide view files, were generated and imported into the OryzaPG-DB server. The data were later imported into a database implemented using the MySQL server. The annotation files (GFF3) and the visualization files (PNG) are stored in the web server directly. The whole system is thus a two-tier web-based system.

The web interface was developed using HTML, Java script and the server side scripting was done using PHP. The database was implemented using the MySQL server. We host the system on Microsoft Internet Information Service (IIS V7.5) on a Dell server running Windows 7 at the Institute for Advanced Biosciences (IAB), Keio University.

4.3.5 OryzaPG-DB Application Programming Interfaces (APIs)

The application programming interface (API) is an interface implemented by the application to allow interaction with the operating system or other programs. An API determines the protocol and parameters required to run certain functions or parts of the program and to return the results of its execution [117]. In OryzaPG-DB, we provide users with several URL APIs for data retrieval. For each entity, we provide users with an API that returns the results per record, per chromosome or for the whole genome. The data are returned in tabular view or in FASTA format with minimum formatting to allow easy processing. The complete list of the available APIs and their parameters can be found on OryzaPG-DB API Guide, available in the OryzaPG-DB website.

4.4 Utility and Discussion

4.4.1 Shotgun proteogenomics

The current rice genome annotation includes 56,797 genes, of which most are either putative (23,348 genes) or hypothetical (8,885 genes) or conserved hypothetical (2,003 genes) [97, http://rice.plantbiology.msu.edu]. Thus, the total number of genes for which experimental expression evidence is lacking represents more than 60% of the total annotated genes. Moreover, the available expression evidence (for 6,311 genes, representing about 10% of the total) is based on transcription, which does not necessarily imply translation to protein [2]. This indicates the need for a novel approach to improve and refine the rice genome annotation.

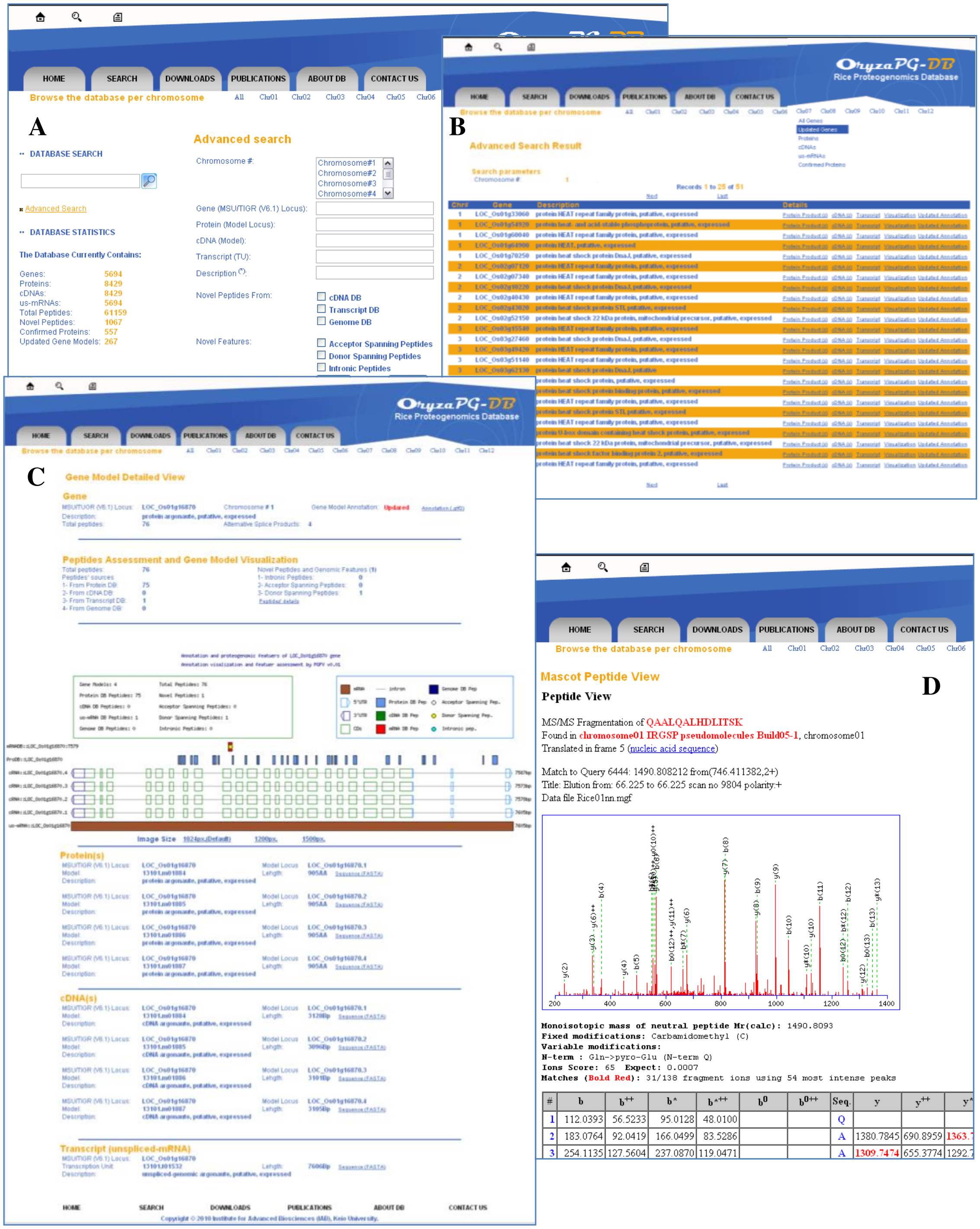
To perform genome annotation refinement of rice by means of a proteogenomics approach, we firstly need accurate and high-throughput proteome information. Thus, we generated the rice MS/MS-based proteome using our highly accurate nanoLC-MS/MS proteomics facility (see Materials and Methods). We started with undifferentiated cultured rice cells to generate data for the construction of our bioinformatics pipeline and data repository system, because a relatively unbiased expression profile of the rice proteome was expected, based on the report that an *Arabidopsis thaliana* proteomics study using cultured cells covered over 70% of the differentiated organ proteome [7]. We plan to generate similar datasets for all vegetative organs throughout the rice life cycle.

The generated data were compared against four databases (protein, cDNA, transcript and genome) for peptide/protein identification and the resultant peptides were filtered using Mascot score and peptide length to select peptides with high identification confidence and high specificity (p value < 0.001 and false-positive rate (FPR) <=1%). Then, the peptides identified from the protein database together with the novel peptides identified uniquely from the other three databases were used to create list of peptides identified from non-redundant product ion spectra.

We utilized these peptides to perform proteogenomic analysis for the rice genes within our sample coverage. Our analysis revealed novel genomic features in 40 genes. In addition, 112 peptides, from the genome database-identified peptides, were mapped to intergenic regions, indicating the possible existence of non-annotated genes.

4.4.2 OryzaPG-DB Utility

OryzaPG-DB provides the scientific community with the first high-throughput MS/MS-based rice proteome information and proteogenomic results, publicly available through a comprehensive web interface (Figure 4.4). The web interface provides several means of data acquisition. Users can browse the database displaying data for genes, updated gene models, proteins, cDNAs and transcripts for the whole genome or per chromosome. In each case, the details of each record will be displayed together with links to related products, genomic origin and identified peptides. In addition, the sequence in FASTA format, the annotation in GFF3 format and visualization in PNG are available for download per displayed record. Further, external links to other databases such as NCBI, MSU and RAP are available. Users can also search the database using keyword search or parameter search, and the search results are displayed per gene. The download page provides users with all the data generated in this project and the data used to perform the proteogenomic analysis. Links to the websites that contains the original datasets and instructions for how to download the original annotations are also provided.



A

Advanced search

DATABASE SEARCH

Advanced Search

DATABASE STATISTICS

The Database Currently Contains:

Genes:	5694
Proteins:	8429
cDNAs:	8429
us-mRNAs:	5694
Total Peptides:	61159
Novel Peptides:	1067
Confirmed Proteins:	557
Updated Gene Models:	267

Chromosome #: Chromosome#1 Chromosome#2 Chromosome#3 Chromosome#4
Gene (MSU/TIGR (V6.1) Locus):
Protein (Model Locus):
cDNA (Model):
Transcript (TU):
Description (*):
Novel Peptides From: □ cDNA DB □ Transcript DB □ Genome DB
Novel Features: □ Acceptor Spanning Peptides □ Donor Spanning Peptides □ Intronic Peptides

B

Advanced Search Result

Search parameters
Chromosome #: 1

Records 1 to 25 of 51

C

Gene Model Detailed View

Gene
MSU/TIGR (V6.1) Locus: LOC_Os01g16870 Chromosome # 1 Gene Model Annotation: Updated Annotation (.gff)
Description: protein argonaute, putative, expressed
Total peptides: 76 Alternative Splice Products: 4

Peptides Assessment and Gene Model Visualization
Total peptides: 76
Peptides' sources
1- From Protein DB: 75
2- From cDNA DB: 0
3- From Transcript DB: 1
4- From Genome DB: 0
Novel Peptides and Genomic Features (1)
1- Intronic Peptides: 0
2- Acceptor Spanning Peptides: 0
3- Donor Spanning Peptides: 1
Peptides' details

Protein(s)

cDNA(s)

Transcript (unspliced-mRNA)

Copyright © 2010 Institute for Advanced Biosciences (IAB), Keio University.

D

Mascot Peptide View

Peptide View

MS/MS Fragmentation of QAALQALHDLITSK
Found in chromosome01 IRGSP pseudomolecules Build05-1, chromosome01
Translated in frame 5 (nucleic acid sequence)

Match to Query 6444: 1490.808212 from(746.411382,2+)
Title: Elution from: 66.225 to 66.225 scan no 9804 polarity:+
Data file Rice01nn.mgf

Monoisotopic mass of neutral peptide Mr(calc): 1490.8093
Fixed modifications: Carbamidomethyl (C)
Variable modifications:
N-term : Gln->pyro-Glu (N-term Q)
Ions Score: 65 **Expect:** 0.0007
Matches (Bold Red): 31/138 fragment ions using 54 most intense peaks

#	b	b++	b*	b*++	b0	b0++	Seq.	y	y++	y*
1	112.0393	56.5233	95.0128	48.0100			Q			
2	183.0764	92.0419	166.0499	83.5286			A	1380.7845	690.8959	1363.7
3	254.1135	127.5604	237.0870	119.0471			A	1309.7474	655.3774	1292.7

Figure 4.4 OryzaPG screenshots. (A) The advanced search form. (B) Example of advanced search results. (C) Gene detailed view including information on the gene, its protein products, cDNAs, transcripts and graphical visualization. D) Peptide view and the product ion spectra. The browse bar at the top of all pages allows the user to navigate through the database displaying data for all chromosomes or per chromosome.

4.5 Materials and Methods

4.5.1 Sample Preparation

1. In-gel digestion

Rice cultured cells (~180 mg) were frozen in liquid nitrogen and then disrupted with a Multi-beads shocker (MB400U; Yasui Kikai), as previously reported [118]. SDS-PAGE was performed using Wako Super Sep Ace precast gel (Osaka, Japan). The gel was stained with negative gel stain MS kit (Wako), and sliced into 10 slices, followed by in-gel digestion as described [42]. These digested samples were desalted using StageTips with C18 Empore disk membranes (3 M) [119] for the subsequent LC-MS/MS analysis.

2. In-solution digestion

The same disrupted cells were suspended in 0.1 M Tris–HCl (pH 9.0), transferred to the homogenizer and homogenized. Then, the homogenate was reduced with dithiothreitol, alkylated with iodoacetamide, and digested with Lys-C, followed by dilution and trypsin digestion as described [120]. These digested samples were desalted using StageTips with C18 Empore disk membranes (3 M). The peptide concentration of the eluates was adjusted to 1.0 mg/ml with 0.1% TFA and 80% acetonitrile. The resultant peptides were pre-fractionated using two methods, strong cation exchange (SCX) (7 fractions) [116] and isoelectric focusing (IEF) (7 fractions) [43], resulting in 14 peptide samples, which were evaluated by LC-MS/MS.

4.5.2 Peptide pre-fractionation

1. SCX pre-fractionation

Pre-fractionation of digested peptides was carried out according to the StageTips fractionation protocol [121], with a strong cation exchange (SCX) disk (SCX-StageTips [42]). Elution with

20-500 mM ammonium acetate solutions containing 15% acetonitrile afforded five fractions. Eluates, including the flow-through fraction, were desalted with C18-StageTips.

1. IEF pre-fractionation

Another portion of the in-solution-digested peptides was pre-fractionated by use of the IEF pre-fractionation method [43] with a ZOOM® IEF Fractionator (Invitrogen) according to the manufacturer's protocol. Digested peptides (890 μg) were loaded onto the ZOOM® IEF Fractionator, and five IEF fractions, enriched in peptides with p*I* values of 3–4.6, 4.6–5.4, 5.4–6.2, 6.2–7, and 7–10, were obtained.

4.5.3 LC-MS/MS analysis

An LTQ-Orbitrap XL (Thermo Fisher Scientific, Bremen, Germany) coupled with a Dionex Ultimate3000 pump (Germering, Germany) and an HTC-PAL autosampler (CTC Analytics AG, Zwingen, Switzerland) was used for nanoLC-MS/MS analyses. A self-pulled needle (150 mm length × 100 μm I.D., 6 μm opening) packed with ReproSil C18 materials (3 μm, Dr. Maisch, Ammerbuch, Germany) was used as an analytical column with "stone-arch" frit [122]. A spray voltage of 2400 V was applied. The injection volume was 5 μL and the flow rate was 500 nL/min. The mobile phases consisted of (A) 0.5% acetic acid and (B) 0.5% acetic acid and 80% acetonitrile. A three-step linear gradient of 5% to 10% B in 5 min, 10% to 40% B in 60 min, 40% to 100% B in 5 min and 100% B for 10 min was employed. The MS scan range was m/z 300–1500. The top ten precursor ions were selected in the MS scan by Orbitrap with R = 60,000, and for subsequent MS/MS scans by ion trap in the automated gain control (AGC) mode where AGC values of 5.00e+05 and 1.00e+04 were set for full MS and MS/MS, respectively. The normalized CID was set to 35.0. A lock mass function was used for the LTQ-Orbitrap XL to obtain constant mass accuracy during gradient analysis [123].

4.5.4 Database search and Sequence alignment

Peptides and proteins were identified by MASCOT v2.2 (Matrix Science, London, U.K.) [59] against the MSU rice protein, cDNA, transcript and genome databases [97] with strict specificity allowing for 1 missed cleavage only. MASCOT identification parameters were Carbamidomethyl (C) as a fixed modification and Acetyl (N-term), Gln->pyro-Glu (N-term Q), Glu->pyro-Glu (N-term E) and Oxidation (M) as partial modifications. The product ion mass tolerance was 0.80 Da, while the precursor ion mass tolerance was 3 ppm. Peptides were rejected if the MASCOT score was below the 99.9% (p value < 0.001) confidence limit based on the "identity" score of each peptide in all database searches. To increase the identification accuracy and peptide specificity, we accepted peptides with at least seven amino acids [93, 100]. Alignment was performed using a local version of NCBI BLAST (blast2seq) Windows version [47, 49] and perl script. We used the default parameters of BLAST. Perl scripts were written for all other data analysis and manipulation steps (see results).

4.6 OryzaPG-DB Availability and Requirements

OryzaPG-DB is freely available at http://oryzapg.iab.keio.ac.jp. In the development of Oryza-PG DB, we followed the usual standards of web applications development and the Java scripts employed are cross-browser scripts. We have confirmed that OryzaPG-DB is fully functional on four web-browsers, Google Chrome, Mozilla Firefox, Microsoft Internet Explorer and Safari, in five operating systems, Windows XP, Vista and 7, Linux Ubuntu and Mac OS (10.5), with no need for any plug-ins or special system requirements.

4.7 OryzaPG-DB Future developments

Plans for further development of OryzaPG-DB are mainly focused on the content and consequently also the interface. We plan to extend the data to include rice root, stem, leaf blade and other organs as soon as we generate those proteomes. In addition, proteogenomic analysis will be available for all genes covered by the new samples. The interface, therefore, will be updated to allow browsing the data by sample, organ, etc., and we will also add advanced search parameters, enabling auto-generation of updated FASTA sequences using experimentally based genome re-annotation. Chapter 5 describes the developments, content expansion and newly added features in the current version of the rice proteogenomics database OryzaPG-DB (v1.1).

4.8 Chapter Conclusions

The rapid growth of available sequenced genomes requires novel approaches to identify genes and their functions, as well as sustainable data repository systems to store the accumulated data and make it publicly available for researchers. Proteogenomics is a novel approach combining MS/MS-based proteomics with genomic information and bioinformatics to enhance genome annotation. In this chapter, we present OryzaPG-DB, the data repository system of the Rice Proteogenomics Project. OryzaPG-DB provides interested rice biologists with the MS/MS-based proteome and the results of proteogenomic analysis, together with all the genomic information within our coverage. The database currently contains the results for cells from undifferentiated culture, and it is planned to be updated periodically with the results of analysis of samples from all vegetative organs of rice. We believe OryzaPG-DB will be an important resource and data-serving tool for rice biologists.

Chapter 5

Developing the bioinformatics pipeline (4) OryzaPG-DB (v1.1): The rice proteogenomics database development, expansion and new features.

5.1 Chapter Abstract

The recently developed rice proteogenomics database (OryzaPG-DB), described in chapter 4, is the first sustainable resource for rice shotgun-based proteogenomics, providing information on peptides identified in rice protein digested peptides measured by means of liquid chromatography-tandem mass spectrometry (LC-MS/MS), and mapping of the peptides to their genomic origins and the genomic novelty of each peptide. The sequences of the peptides, proteins, cDNAs and genes, and the gene annotations are available for download in FASTA and GFF3 formats, respectively. Further, an annotated visualization of the gene models, corresponding peptides and genomic novelty is available for each gene, and MS/MS spectra are available for each peptide. In this chapter, I am discussing the utilization of OryzaPG-DB and describing its recent development, content expansions and newly added features in the current version (OryzaPG-DB v1.1).

5.2 Introduction

The rapid improvement of analytical instruments for biological research made it are capable of producing hundreds of gigabytes of raw data every day resulting in the accumulation of enormous amounts of both raw and analyzed data [95]. Such huge amounts of data contain valuable biological information, which can only be uncovered by performing appropriate computational analysis using proper bioinformatics tools. Since this data is very rich, applying different kinds of computational analysis or using different bioinformatics tools can lead to different discoveries according to the goal of the research. Therefore, such biological data are generally deposited in biological databases, which are collections or libraries of biological data sorted and organized in a way that allows easy access, management and updating in a sustainable format so that interested scientists can have consistent access over a prolonged period [124]. Data sustainability and ease of use are the main benefits of storing biological data in databases. Further, long-term projects, e.g. the human genome project, require extendable and sustainable databases that can store data from different phases of the projects [125, 126]. Almost all kinds of biological data are stored in databases including, for example, sequence data, structural data, interaction data and proteogenomic data [124].

Proteogenomics is an alliance between proteomics and genomics, and aims mainly to use proteomics data and technologies for identifying novel genomic features, such as novel un-annotated genes, and for improving, correcting or confirming the structural and functional genome annotation [2]. However, it is also applicable to deeper and wider goals, such as understanding the mechanisms of environmental adaptation between different species, discovery of biomarkers and identification of the targets of antibodies [3, 32, 33]. In plants, particularly, *Arabidopsis* received most of the proteogenomic efforts [7, 11]. However, proteogenomic analysis was performed for several other plants, such as rice, and several major plant pathogens [9, 14, 35, 127].

Proteogenomic analyses usually rely on high-throughput genomic and proteomic data, such as genome sequencing and liquid chromatography-tandem mass spectrometry (LC-MS/MS) data. The scheme of proteogenomic analysis can differ from project to project depending on the available genomic and proteomic data, the status of the genome annotation of the organism in question (annotated or newly sequenced) and the availability of sufficient informatics tools and computational power to deal with such large amounts of data [95]. However, three steps are always shared among different proteogenomic projects: 1. Mapping proteomic data (peptide sequences derived from the MS/MS analysis) to the genome. 2. Evaluating the genomic novelty of the proteomic data 3. Updating or confirming the current genome annotation by integrating the newly discovered genomic information into the current genome annotation [2, 95]. These considerations distinguish proteogenomics databases from other types of biological databases.

Thus, a proteogenomic database is a biological database that holds genomic and proteomic sequence data, together with the mapping of the proteomic data to the genome and the genome annotation. It can also store other information concerning the organism's genes and proteins, such as mRNA and cDNA sequences and biological or biochemical functions. Therefore, proteogenomic databases require special design and implantation [14].

5.3 The Rice Proteogenomic Database (OryzaPG-DB)

The Rice Proteogenomic Database (OryzaPG-DB), chapter 4, incorporates the genomic features of experimental shotgun proteomics data for rice (*Oryza sativa*) [14]. It was developed using whole proteome data of rice undifferentiated cultured cells, generated from 27 nanoLC-MS/MS runs on a hybrid ion trap-orbitrap mass spectrometer, and the rice genome annotation database at Michigan State University (MSU), which offers rice protein, cDNA, transcript and genome databases and rice genome annotation [97]. The MS/MS spectra are filtered using our previously described method [128] then searched against the above-mentioned databases in the order mentioned, using Mascot (v.2.3) [59] with peptide acceptance criteria >= 99.9%. The lists of identified peptides are merged and filtered to remove all peptides shorter than seven amino acids [93]. Then, the list of accepted peptides is filtered again to remove redundancy. Therefore, the final list contains only accurate and unique peptides (here, unique means a unique combination of sequence and modifications).

The identified peptides were used to perform proteogenomic analysis of the rice genome by mapping the identified peptides to their genomic origins. Mapping was performed through alignment of peptides identified from the protein, cDNA and transcript databases to the un-spliced genomic mRNA of the corresponding genes. Peptides identified from the genome database were aligned to the corresponding chromosome then mapped to the genes using the alignment coordinates. Next, the alignment results were converted to GFF3 format and the peptide’s GFF3 line was appending to the annotation of the corresponding gene and new GFF3 files were created. The new files are submitted to the ProteoGenomic Features Evaluator (PGFeval), chapter 3, software tool to evaluate the genomic novelty of each peptide [14]. PGFeval analysis provided 51 novel genomic features in 40 rice genes. Further, PGFeval exported a visualized annotation file for each gene in PNG image format. Finally, a tailored proteogenomic database was designed and implemented to host all this data, with the capability of expansion to host data from future phases of rice proteogenomics analysis, and to make the data publicly available for interested scientists in the rice biology community.

5.4 OryzaPG-DB design

Since proteogenomics data has a complex structure, as described above, a tailored database design is required for a proteogenomic database. Further, when we designed OryzaPG-DB, we decided to make the design as generic as possible, in order to help other researchers working on the design of databases suitable for proteogenomic data [14]. In the original design (Chapter4, Figure 4.3), the main entity of the database is the gene. Then, information on the corresponding protein(s), LC-MS/MS measured peptide(s), cDNA(s), mRNA(s), mapping information and updated gene annotation are attached to this entity. In addition, several files are attached to each gene, including protein, peptide, cDNA and mRNA sequences, gene annotation in GFF3 format and visualization of the gene annotation (PGFeval output).

5.5 Updates on OryzaPG-DB design and development

The original version of OryzaPG-DB has been publicly available online since 2010 and was officially opened in 2011. Since that time, several developments have been added to the original version. Here, we review the recent updates and newly added features in the current version (OryzaPG-DB v1.1).

5.5.1 OryzaPG-DB updated database design

The above design of OryzaPG-DB was suitable for hosting the rice proteogenomic data available at the database launch. However, we were aware that this design would need to be updated to accommodate new proteomes from other samples/organs or other types of analysis, such as phosphoproteome analysis, as mentioned in the future work section of the original publication on OryzaPG-DB [14]. Thus, the database design of the current version of OryzaPG-DB (v1.1) has been updated to include information about the experimental sample (Figure 5.1).

Adding the ability to include sample information makes the design sufficiently generic to host all proteogenomic data of an organism while retaining information about the source of each peptide and the expression organ of each gene/protein, as well as providing the basis for several other grouping features that can be useful in performing functional analysis, organ-specific analysis and/or inter-organ comparisons. It is now possible, with the updated design, to select peptides identified from a certain sample and/or organ and/or analysis and/or database, with the corresponding genes. For instance, in a proteogenomic study with extensive sampling, such as the *Arabidopsis* proteogenomic study [7], it is important to have an easy way to create lists of genes/peptides identified from various combinations of sample, organ and condition for comparison in order to find organ or life stage specific biomarkers. Such a task is now straightforward with the updated design of OryzaPG-DB v1.1.

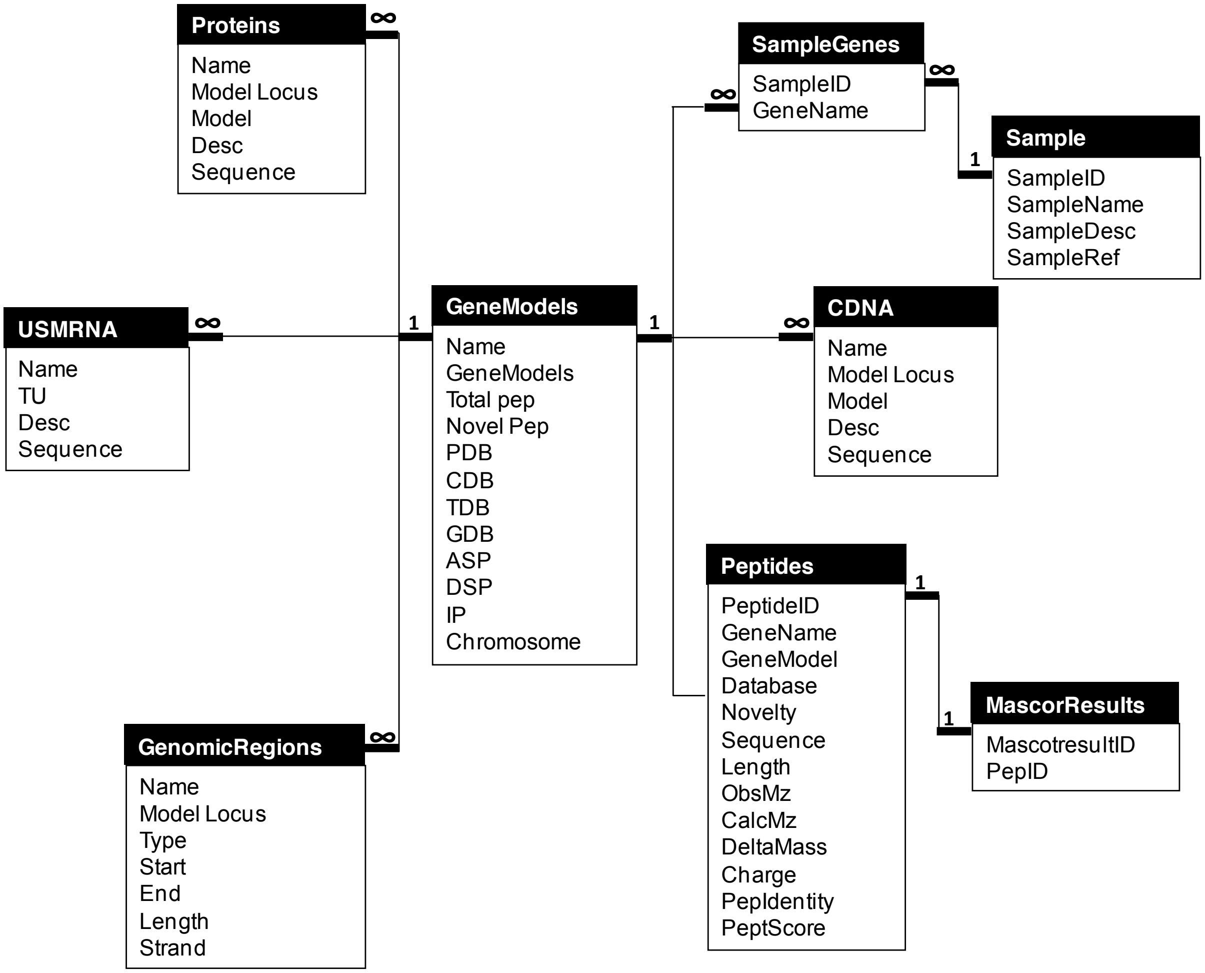


Figure 5.1 Updated design of OryzaPG-DB (v1.1)

5.5.2 OryzaPG-DB new browsing options

The original version of OryzaPG-DB had the “browse per chromosome” feature that allows fetching data of all genes, updated genes, proteins, cDNAs or transcripts of all chromosomes or a single chromosome. However, updating the database design consequently requires updating the “browse per chromosome” options to add sample level options. In the current version, we added the sample level to the “browse per chromosome” menu. This allows fetching data of all genes, genes identified in a particular sample/organ/analysis or genes overlapping samples for all chromosomes or single chromosome (Figure 5.2). With the new “*Save to File*” option (see below), it is possible to save the browsing/search results to a CSV file for download, so that creating lists of genes, proteins, cDNAs, transcripts or peptides per sample, or those overlapping samples for all chromosomes or a single chromosome, has become a single-click task.

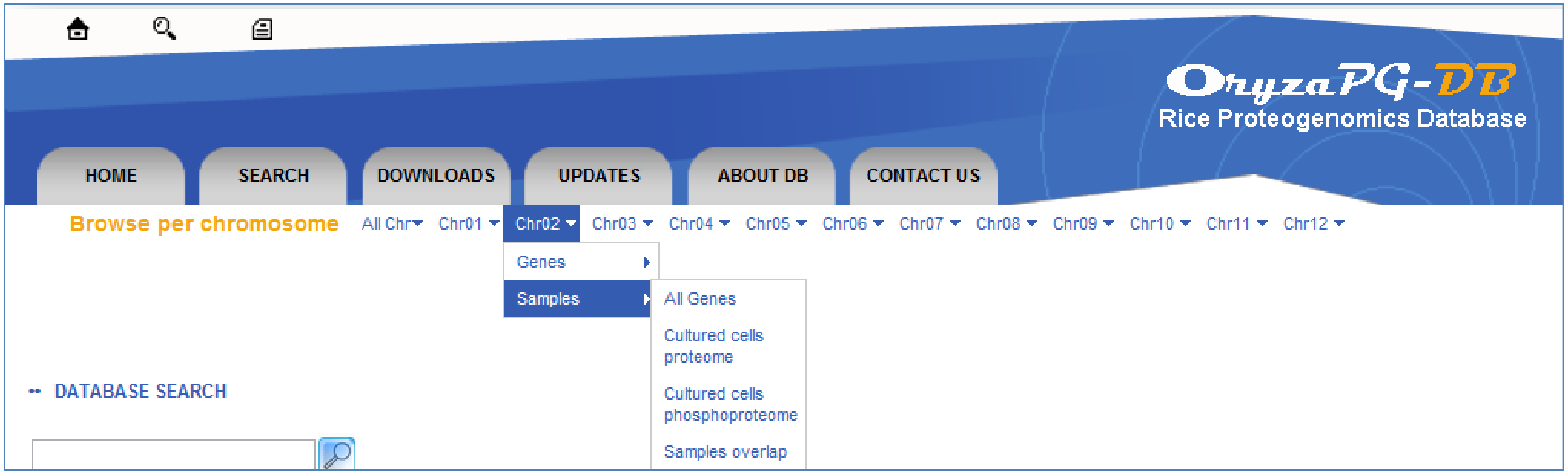


Figure 5.2 Updated browsing options of OryzaPG-DB (v1.1)

5.5.3 OryzaPG-DB new and updated application programming interfaces (APIs)

An application programming interface (API) is an implemented interface that allows other programs or the operating system to interact with the whole program or particular parts or functions of the program. [117]. To increase the usability of the data stored in OryzaPG-DB, we provided six URL APIs to help researchers fetch OryzaPG-DB dynamically with scripts or integrate OryzaPG-DB data into their applications. The URL APIs of OryzaPG-DBv1.1 contains seven APIs (Table 5.1), which are updated versions of the six APIs of the original OryzaPG-DB plus a new API for *samples*. The seven APIs are briefly described.

Genes API: Allows users to retrieve the gene information for a particular gene, all genes in a particular chromosome or all chromosomes.

Updated Genes API: Allows users to retrieve the gene(s) with updated annotations and novel genomic features; per gene, all updated genes in one chromosome or all updated genes.

Proteins API: Allows users to retrieve the protein information for a particular gene, all genes in a particular chromosome or all genes. The result can be shown in tabular view or in FASTA format.

cDNAs API: Allows users to retrieve the cDNA information for a particular gene, all genes in a particular chromosome or all genes. The result can be shown in tabular view or in FASTA format.

Transcripts API: Allows users to retrieve the transcript (unspliced-genomic mRNA) information for a particular gene, all genes in a particular chromosome or all genes. The result can be shown in tabular view or in FASTA format.

Peptides API: Allows users to retrieve peptides information for peptides identified from a particular gene or gene products (protein, cDNA or mRNA), all genes in a particular chromosome or all genes covered by our analysis. The result can be shown in tabular view or in FASTA format. Also, a special parameter can be used to select novel peptides only instead of all peptides.

Samples API (new): Allows users to retrieve the genes identified in a particular sample, all samples or overlapping samples for a particular chromosome or all chromosomes.

For all APIs, a new option was added that allows saving the API execution result to a CSV file by setting the *to_file* parameter to true (Table 5.1).

Table 5.1 OryzaPG-DB v1.1 URL APIs

URL API	Chr. [3]	Gene [3]	FASTA	Novelty	Sample	To file [7]
Genes	Chr=X[1]	Gene=Locus[2]	NA	NA	NA	to_file=1
Updated Genes	Chr=X[1]	Gene=Locus[2]	NA	NA	NA	to_file=1
Proteins	Chr=X[1]	Gene=Locus[2]	fasta=1	NA	NA	to_file=1
cDNAs	Chr=X[1]	Gene=Locus[2]	fasta=1	NA	NA	to_file=1
Transcript	Chr=X[1]	Gene=Locus[2]	fasta=1	NA	NA	to_file=1
Sample[5]	Chr=X[1]	NA	NA	NA	sid =Y[6]	to_file=1
Peptides[4]	Chr=X[1]	Gene=Locus[2]	fasta=1	Novel=1	NA	to_file=1

[1] X: from 1 to 12, if X is out of this range, the API will show data for all genes in the system.

[2] Locus is the MSU V6.1 locus e.g. LOC_Os01g01689.

[3] Gene and Chr (chromosome) parameters cannot be used at the same time.

[4] In the case of peptides, the API cannot work without parameters.

[5] SID and Chr parameters cannot be used at the same time.

[6] See the ABOUT DB page to get the Y value corresponding to each sample.

[7] If the to_file=1, the API result will be saved to a CSV file, otherwise the result will be displayed.

5.6 OryzaPG-DB data expansion

A key feature of biological databases is their expandability, so that the database can expand to host more data related to the original content when such data becomes available. The OryzaPG-DB original and updated database designs both aim to host rice proteogenomics data in an expandable and sustainable way, as described above. In the current version of OryzaPG-DB v1.1, the rice proteogenomic data derived from the proteogenomic analysis of the original 27 nanoLC-MS/MS runs of cultured rice cells were recently expanded to 61 runs in total, demonstrating the sustainability of OryzaPG-DB (Table 5.2). The updated design of the database, by adding sample information, is able to distinguish peptides and genes identified in each sample or those identified in both samples (Figure 5.1). The proteogenomic analysis of the newly added sample covers 845 new genes which were not present in the original OryzaPG-DB coverage, and adds new peptides and/or novel genomic features to 914 of the originally existing genes, expanding the database coverage to 3,973 genes. The numbers of genes with novel peptides and genes with novel genomic features are increased from 119 and 40 to 160 and 62, respectively.

Table 5.2 OryzaPG-DB current contents [1]

	OryzaPG-DB v1.1
Employed dataset(s)	61 LC-MS/MS runs
Confirmatory peptides [2]	18,214
Novel genomic features	98
Genes	3,973
Genes with novel peptides	160
Genes to be updated	62

[1] As of March 28, 2012.

[2] Peptides identified from the protein databases.

5.7 OryzaPG-DB new and updated features

The updated database design and the recent data expansion required the development of several new features that take advantage of the new developments and improve data fetching. In this section, I present a brief description of some of the new features, which are mainly related to database search and content download.

1- Adding *sample* to the advanced search: this option makes use of the new database design, in which one can limit the search to be within the genes/peptides of a certain sample or within the genes/peptides identified across all samples.

2- Adding new peptide novelty categories to the advanced search: the advanced search form in the original version of OryzaPG-DB allows the user to limit the search results to show genes with particular type of peptide novelty, e.g. showing genes with intronic peptides only. Since we have adopted the newer version of PGFeval (Chapter 3) [14], the new form includes two new peptide novelty categories, 3'UTR and, 5'UTR, indicating peptides identified from the 3'UTR and 5'UTR, respectively.

3- Adding *Save to File* option to the database browsing results: the database browsing feature allow fetching data by sample or genes for all chromosomes or a single chromosome (see above). The retrieved data is displayed per gene with links to all details related to this gene such as protein, cDNA, mRNA, peptides and annotation. Consequently, the data is usually huge, so it is displayed in pages of 50 genes per page. In the original version of OryzaPG-DB, there was no way to display or save all retrieved data through database browsing. Therefore, we developed the *Save to File* option that appears in all database browsing results, and allows saving all the data to a downloadable file in CSV file format.

4- Adding *Save Search Results* option to the database searching results: similar to the database browsing feature, the database searching feature can display huge amounts of data and there was no way to display all this data or save it. Therefore, we added the *Save Search Results* option that allows saving all database searching results to a downloadable file in CSV file format in a similar way and format to the *Save to File* option described above.

5.8 Expanded utility and availability of OryzaPG-DB (v1.1)

OryzaPG-DB is the first database that provides a sustainable resource for proteogenomic analysis of an economically important crop that includes genes, gene products (mRNA, cDNA and protein), experimental expression evidence (MS/MS peptide spectra) and mapping of the peptides to their genomic origin. Further, the sequences of each gene and its products and the gene annotation are available in GFF3 format and can be graphically visualized. Such data can be of great value for plant biologists in general and rice biologists in particular. Furthermore, the generic OryzaPG-DB database design provides a template that should be applicable to data from other similar projects/analyses. The new or updated features are discussed in detail above. OryzaPG-DB is freely available online at the servers of the Institute for Advanced Biosciences (IAB), Keio University, Japan at http://oryzapg.iab.keio.ac.jp/.

5.9 The future of the rice proteogenomics database

The current version of OryzaPG-DB, version 1.1, includes several developments and features that were foreshadowed in the future work section of our original article describing OryzaPG-DB [14], such as data expansion, adding sample level information and updating the advanced search parameters, together with features that not mentioned then, but which we thought would improve the utility of the database such as *save to file* and *save search results* options. Future developments are expected to focus mainly on data expansion and proteogenomic analysis of newly added data. In addition, more informatics updates will be included, such as offering downloadable Perl script that will be useful for automation of OryzaPG-DB data acquisition through the available URL APIs.

Conclusions

Proteogenomics is a systems biology approach that utilizes proteomics and genomics to obtain better understanding of the genome structure and function. Peptide spectra obtained using means of high-throughput proteomics, such as mass spectrometry-based shotgun proteomics, and genomic sequences obtained by modern genome sequencing instruments, such as next-generation genome sequencers, are analyzed together using bioinformatics tools in order to improve, correct or confirm the genome structural and functional annotations. This analysis is mainly through searching the peptide spectra against database(s) of genomic sequences to identify the peptide sequence corresponds to each spectrum. Next, the identified peptides are mapped to the genome in order to know its genomic origin. The peptide mapping process can reveal several types of novel genomic information. For instance, peptides mapped to genomic regions that was annotated as "intergenic", point new un-annotated genes. Further, peptides mapped to "known genes" can be mapped to introns, pointing either an incorrect annotation of the gene or a new alternative splicing isoform. Furthermore, peptides mapped to "known genes" and inside the gene mapped to an exon, represent an affordable mean of experimental confirmation for the expression of computationally predicted gens.

However, in large-scale proteogenomic analysis, both the proteomic and genomic data are enormous and, therefore, analyzing them together is one of the major challenges in proteogenomics. In spite of the efforts made by the proteomics and bioinformatics societies to facilitate the peptide identification process in large-scale studies, the growing spectra generation rate of the modern mass spectrometers and the growing size of the sequence databases continuously require developing of new computational methods, algorithms and software tools.

In this thesis, I intensely surveyed the efforts done by the proteomics and bioinformatics societies in the last decade to facilitate the peptide identification process in large-scale studies, either through developing novel computational methods, sophisticated database searching algorithms or software tools (chapter 1). Next, I described the development and applications of a novel bioinformatics pipeline for large-scale proteogenomic studies (chapters 2~5). The developed bioinformatics pipeline has three main components. First, novel computational method facilitates peptides identification from large-scale proteomics and large-sized databases named Mass Spectrum Sequential Subtraction (MSSS) method (chapter 2). Second, software tool for evaluating the proteogenomic novelty of the identified peptide named The ProteoGenomic Features Evaluator (PGFeval) software tool (chapter 3). Third, database for hosting proteogenomic data, hosting the rice proteogenomic data of this work, named The Rice Proteogenomics Database (OryzaPG-DB) database (chapter 4 and 5).

MSSS is a novel bioinformatics method speeds up the peptide-sequence identification from searching large peptide MS/MS spectra datasets, such as MS/MS datasets generated by the modern mass spectrometry instruments, against large nucleotide databases, such as six-frame translation of the genome databases. Searching one protein database and three nucleotide sequence databases, MSSS successfully decreased the number of search queries to 50% and the overall search time to 75%, on average, through sequential removing the identified spectra after each database search and creating new files contain the unidentified spectra only to be used in the next search. Further, MSSS had no effect on the peptide identification capability or the identification false-positive rate (FPR).

The Rice Proteogenomics Database (OryzaPG-DB) is the first rice proteome database based on shotgun proteogenomics. OryzaPG-DB includes the genomic features of experimental shotgun proteomics data of rice together with their product ion spectra, updated annotations for rice genes, side by side with the corresponding protein, cDNA, transcript and genomic sequences. The first version of OryzaPG-DB (v1.0) was created from the results of 27 nanoLC-MS/MS runs, analyzing tryptic digests from undifferentiated cultured rice cells. Later, the recent version (v1.1) includes data resulted from another 34 nanoLC-MS/MS runs of the phosphoproteome of the rice undifferentiated cultured rice cells. Currently, OryzaPG-DB is available online and covers 3,973 genes and 6,291 proteins/cDNAs with 18,214 unique peptides.

The ProteoGenomic Features Evaluator (PGFeval), is a software tool that developed specially to analyze the peptide-mapping data (data of mapping the peptides to the genome) to evaluate the peptide's proteogenomic novelty and visualize gene model structure and features. PGFeval process the peptides mapping data and the current genome annotation to output an updated gene annotation, graphical visualization of the gene model structure and features and two reports describing the proteogenomic novelty of the peptides (peptide report) and the novel genomic features, if any, per gene (genes report).

The developed pipeline was used to analyze rice proteome and phosphoproteome data (61 nanoLC-MS/MS runs) revealing 98 novel genomic feature in 62 rice genes. In addition, MSSS method has several potential applications in cancer biology such as finding cancer-related somatic mutations, which can lead to discovery of new drug targets and/or biomarkers.

References

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

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

3. de Groot A, Dulermo R, Ortet P, Blanchard L, Guerin P, Fernandez B, Vacherie B, Dossat C, Jolivet E, Siguier P *et al*: **Alliance of proteomics and genomics to unravel the specificities of Sahara bacterium Deinococcus deserti**. *PLoS Genetics* 2009, **5**(3):e1000434.

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

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

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

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

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

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

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

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

12. Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, Eng JK, Aderem A, Boyle R, Brunner E, Donohoe S *et al*: **Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry**. *Genome Biology* 2005, **6**(1):R9.

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

14. Helmy M, Tomita M, Ishihama Y: **OryzaPG-DB: Rice Proteome Database based on Shotgun Proteogenomics**. *BMC Plant Biology* 2011, **11**(1):63.

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

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

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

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

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

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

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

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

23. Jaffe JD, Stange-Thomann N, Smith C, DeCaprio D, Fisher S, Butler J, Calvo S, Elkins T, FitzGerald MG, Hafez N *et al*: **The complete genome and proteome of Mycoplasma mobile**. *Genome Research* 2004, **14**(8):1447-1461.

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

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

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

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

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

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

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

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

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

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

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

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

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

37. Gupta N, Benhamida J, Bhargava V, Goodman D, Kain E, Kerman I, Nguyen N, Ollikainen N, Rodriguez J, Wang J *et al*: **Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes**. *Genome Research* 2008, **18**(7):1133-1142.

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

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

40. Helmy M, Sugiyama N, Tomita M, Ishihama Y: **Onco-proteogenomics: a novel approach to identify cancer-specific mutations combining proteomics and transcriptome deep sequencing**. *Genome Biology* 2010, **supp 11**:p17.

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

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

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

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

45. Kapp E, Schütz F: **Overview of Tandem Mass Spectrometry (MS/MS) Database Search Algorithms**. *Current Protocols in Protein Science* 2007, **49**:25.2.1–25.2.19.

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

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

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

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

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

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

52. Zhou C, Chi H, Wang LH, Li Y, Wu YJ, Fu Y, Sun RX, He SM: **Speeding up tandem mass spectrometry-based database searching by longest common prefix**. *BMC Bioinformatics* 2010, **11**(1):577.

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

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

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

56. Turse JE, Marshall MJ, Fredrickson JK, Lipton M, S. , Callister SJ: **An Empirical Strategy for Characterizing Bacterial Proteomes across Species in the Absence of Genomic Sequences**. *PloS one* 2010, **5**(11):e13968.

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

58. Eng JK, McCormack AL, Yates JR: **An approach to correlate MS/MS data to amino acid sequences in a protein database**. *Journal of the American Society for Mass Spectrometry* 1994, **5**:976-989.

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

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

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

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

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

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

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

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

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

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

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

70. Yates JR, Eng JK, Clauser KR, Burlingame AL: **Search of sequence databases with uninterpreted high-energy collision-induced dissociation spectra of peptides**. *Journal of the American Society for Mass Spectrometry* 1996, **7**(11):1089-1098.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

85. Liu J, Carrillo B, Yanofsky C, Beaudrie C, Morales F, Kearney R: **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* 2005, **4**:4441-4444.

86. Helmy M, Sugiyama N, Tomita M, Ishihama Y: **MSSS: Mass spectrum sequential subtraction speeds up searching large peptide MS/MS spectra datasets against large nucleotide databases for proteogenomics**. *Genes to Cells* in press.

87. Koonin E, Galperin M: **Sequence-Evolution-Function: Computational Approaches in Comparative Genomics**: USA, Kluwer Academic Publishers; 2003.

88. Guigo R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E *et al*: **EGASP: the human ENCODE Genome Annotation Assessment Project**. *Genome Biology* 2006, **7 Suppl 1**:S2 1-31.

89. Allen JE, Pertea M, Salzberg SL: **Computational gene prediction using multiple sources of evidence**. *Genome Research* 2004, **14**(1):142-148.

90. Coghlan A, Fiedler TJ, McKay SJ, Flicek P, Harris TW, Blasiar D, Stein LD: **nGASP--the nematode genome annotation assessment project**. *BMC Bioinformatics* 2008, **9**:549.

91. Fu Y, Yang Q, Sun R, Li D, Zeng R, Ling CX, Gao W: **Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry**. *Bioinformatics* 2004, **20**(12):1948-1954.

92. Li D, Fu Y, Sun R, Ling CX, Wei Y, Zhou H, Zeng R, Yang Q, He S, Gao W: **pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry**. *Bioinformatics* 2005, **21**(13):3049-3050.

93. Choudhary JS, Blackstock WP, Creasy DM, Cottrell JS: **Matching peptide mass spectra to EST and genomic DNA databases**. *Trends Biotechnology* 2001, **19**(10 Suppl):S17-22.

94. Kuster B, Mortensen P, Andersen JS, Mann M: **Mass spectrometry allows direct identification of proteins in large genomes**. *Proteomics* 2001, **1**(5):641-650.

95. Helmy M, Tomita M, Ishihama Y: **Peptide Identification by Searching Large-Scale Tandem Mass Spectra against Large Databases: Bioinformatics Methods in Proteogenomics**. *Genes Genomones and Genomics* 2012, **6**(Special Issue 1):76-85.

96. Ning K, Fermin D, Nesvizhskii AI: **Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets**. *Proteomics* 2010, **10**(14):2712-2718.

97. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L *et al*: **The TIGR Rice Genome Annotation Resource: improvements and new features**. *Nucleic Acids Research* 2007, **35**(Database issue):D883-887.

98. Nakagami H, Sugiyama N, Mochida K, Daudi A, Yoshida Y, Toyoda T, Tomita M, Ishihama Y, Shirasu K: **Large-scale comparative phosphoproteomics identifies conserved phosphorylation sites in plants**. *Plant Physiology* 2010, **153**(3):1161-1174.

99. Sanders WS, Wang N, Bridges SM, Malone BM, Dandass YS, McCarthy FM, Nanduri B, Lawrence ML, Burgess SC: **The proteogenomic mapping tool**. *BMC Bioinformatics* 2011, **12**:115.

100. Kim W, Silby MW, Purvine SO, Nicoll JS, Hixson KK, Monroe M, Nicora CD, Lipton MS, Levy SB: **Proteomic detection of non-annotated protein-coding genes in Pseudomonas fluorescens Pf0-1**. *PloS one* 2009, **4**(12):e8455.

101. IRGSP: **The map-based sequence of the rice genome**. *Nature* 2005, **436**(7052):793-800.

102. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ *et al*: **The UCSC Genome Browser Database**. *Nucleic Acids Research* 2003, **31**(1):51-54.

103. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A *et al*: **The generic genome browser: a building block for a model organism system database**. *Genome Research* 2002, **12**(10):1599-1610.

104. Kondo T, Plaza S, Zanet J, Benrabah E, Valenti P, Hashimoto Y, Kobayashi S, Payre F, Kageyama Y: **Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis**. *Science* 2010, **329**(5989):336-339.

105. Bergamaschi C, Jalah R, Kulkarni V, Rosati M, Zhang GM, Alicea C, Zolotukhin AS, Felber BK, Pavlakis GN: **Secretion and biological activity of short signal peptide IL-15 is chaperoned by IL-15 receptor alpha in vivo**. *Journal of Immunology* 2009, **183**(5):3064-3072.

106. Fullwood MJ, Wei CL, Liu ET, Ruan Y: **Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses**. *Genome Research* 2009, **19**(4):521-532.

107. Siezen RJ, Hijum SAFTV: **Genome (re-)annotation and open-source annotation pipelines. Microbial Biotechnology**. *Microbial Biotechnology* 2010, **3**(4):362-369.

108. Reed JL, Famili I, Thiele I, Palsson BO: **Towards multidimensional genome annotation**. *Natuar Reviews Genetics* 2006, **7**(2):130-141.

109. Brent MR: **Steady progress and recent breakthroughs in the accuracy of automated genome annotation**. *Natuar Reviews Genetics* 2008, **9**(1):62-73.

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

111. McCarthy FM, Bridges SM, Wang N, Magee GB, Williams WP, Luthe DS, Burgess SC: **AgBase: a unified resource for functional analysis in agriculture**. *Nucleic Acids Research* 2007, **35**(Database issue):D599-603.

112. Sasaki T: **Current status of and future prospects for genome analysis in rice** Springer-Verleg, Japan; 1999.

113. Matsumoto T, Wu J, Antonio BA, Sasaki T: **Development in rice genome research based on accurate genome sequence**. *International Journal of Plant Genomics* 2008, **2008**:348621.

114. Itoh T, Tanaka T, Barrero RA, Yamasaki C, Fujii Y, Hilton PB, Antonio BA, Aono H, Apweiler R, Bruskiewich R *et al*: **Curated genome annotation of Oryza sativa ssp. japonica and comparative genome analysis with Arabidopsis thaliana**. *Genome Research* 2007, **17**(2):175-183.

115. Komatsu S, Tanaka N: **Rice proteome analysis: a step toward functional analysis of the rice genome**. *Proteomics* 2005, **5**(4):938-949.

116. Wu CC, MacCoss MJ, Howell KE, Yates JR, 3rd: **A method for the comprehensive proteomic analysis of membrane proteins**. *Nature Biotechnology* 2003, **21**(5):532-538.

117. Tulach J: **Practical API Design: Confessions of a Java Framework Architect**. New York: Apress 2008.

118. Sugiyama N, Nakagami H, Mochida K, Daudi A, Tomita M, Shirasu K, Ishihama Y: **Large-scale phosphorylation mapping reveals the extent of tyrosine phosphorylation in Arabidopsis**. *Molecular Systems Biology* 2008, **4**:193.

119. Rappsilber J, Ishihama Y, Mann M: **Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics**. *Analytical Chemistry* 2003, **75**(3):663-670.

120. Saito H, Oda Y, Sato T, Kuromitsu J, Ishihama Y: **Multiplexed two-dimensional liquid chromatography for MALDI and nanoelectrospray ionization mass spectrometry in proteomics**. *Journal of Proteome Research* 2006, **5**(7):1803-1807.

121. Rappsilber J, Mann M, Ishihama Y: **Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips**. *Nature Protocols* 2007, **2**(8):1896-1906.

122. Ishihama Y, Rappsilber J, Andersen JS, Mann M: **Microcolumns with self-assembled particle frits for proteomics**. *Journal of Chromatography* 2002, **979**(1-2):233-239.

123. Olsen JV, de Godoy LM, Li G, Macek B, Mortensen P, Pesch R, Makarov A, Lange O, Horning S, Mann M: **Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap**. *Molecular and Cellular Proteomics* 2005, **4**(12):2010-2021.

124. Birney E, Clamp M: **Biological database design and implementation**. *Briefings in Bioinformatics* 2004, **5**(1):31-38.

125. McGourty C: **Human genome project. Dealing with the data**. *Nature* 1989, **342**(6246):108.

126. Hyman SE: **Genome-sequencing anniversary. The meaning of the Human Genome Project for neuropsychiatric disorders**. *Science* 2011, **331**(6020):1026.

127. Bindschedler LV, McGuffin LJ, Burgis TA, Spanu PD, Cramer R: **Proteogenomics and in silico structural and functional annotation of the barley powdery mildew Blumeria graminis f. sp. hordei**. *Methods* 2011, **54**(4):432-441.

128. Ravichandran A, Sugiyama N, Tomita M, Swarup S, Ishihama Y: **Ser/Thr/Tyr phosphoproteome analysis of pathogenic and non-pathogenic Pseudomonas species**. *Proteomics* 2009, **9**(10):2764-2775.

Appendix

Abbreviations

ABLCP	Algorithm Based on Longest Common Prefix
API	Application Programming Interface
BLAST	Basic Local Alignment Search Tool
CPU	Central Processing Unit
EST	Expressed Sequence Tags
FP	False Positives
FPR	False Positive Rate
GFF3	Generic File Format version 3
GPF	Genomic Peptide Finder
HGPdb	Human Genome Peptide Database
HUPO PPP	Human Proteome Organization Plasma Proteome Project
IEF	Isoelectric Focusing
LC-MS/MS	Liquid Chromatography-Tandem Mass Spectrometry
LSH	Locality Sensitive Hashing
MGF	Mascot Generic Format
MS	Mass Spectrometry
MS IIS	Microsoft Internet Information Service
MSSS	Mass Spectrum Sequential Subtraction
MSU-DB	Michigan State University Database
MW	Molecular Weight
ORF	Open Reading Frames
PGFeval	ProteoGenomic Features evaluator
pI	Peptide Isoelectric Focusing Value
PST	Peptide Sequence Tags
PTM	Posttranslational Modifications

QTOF	Quadrupole Time-Of-Flight
RAP	Rice Annotation Project
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SCX-StageTips	Strong Cation Exchange-StageTips
SNP	Single Nucleotide Polymorphism
TP	True Positives
UTR	Untranslated Regions