Hindawi
BioMed Research International
Volume 2020, Article ID 8740468, 9 pages
https://doi.org/10.1155/2020/8740468

*Research Article*

# Development of an MS Workflow Based on Combining Database Search Engines for Accurate Protein Identification and Its Validation to Identify the Serum Proteomic Profile in Female Stress Urinary Incontinence

**Sara El Jadid**,[1,2] **Taoufik Bensellak**,[2] **Raja Touahni**,[1] **and Ahmed Moussa**[2]

[1]*Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco*
[2]*National School of Applied Sciences, Abdelmalek Essaadi University, Tangier, Morocco*

Correspondence should be addressed to Sara El Jadid; eljadidsara@gmail.com

Received 29 May 2020; Accepted 30 July 2020; Published 18 August 2020

Academic Editor: Hely M. Häggman

A critical stage of shotgun proteomics is database search, a process which attempts to match the experimental spectra to the theoretical one. Given the considerable time and effort spent in analysis, it is self-evident for a researcher to aspire for rigorous computational analysis and a more confident and accurate peptide/protein identification. Mass spectrometry (MS) has been applied across several clinical disciplines. The pathophysiology of Stress Urinary Incontinence (SUI), caused by a damaged pelvic floor, has become a boundless disease altering the quality of life worldwide. Although some studies pointed markers that can be bioindicators for SUI, these findings raise the issue of sensitivity and specificity. Therefore, it is critical to have a sensitive and specific analytical approach to identify markers that have been associated with protective and deleterious associations in disease. Here, we describe our designed and developed workflow for protein identification from tandem mass spectrometry that uses multiple search engines. We apply our workflow to an existing study addressing the pathophysiology of SUI. We demonstrate how using the combined approach together with high-performance computing techniques can surmount the challenges of complex analyses and extended computing time. We also compare the relative performance of each combination. Our results suggest that a combination of MS-GF+ and COMET represents the best sensitivity-specificity trade-off, outperforming all other tested combinations. The approach was also sensitive and accurately identified a set of protein that was shown to be markers for categories of diseases associated with the pathophysiology of SUI. This workflow was developed to encourage proteomic researchers to adopt MS-based techniques for accurate analysis and to promote MS as a routine tool to the clinical cohorts.

## 1. Introduction

Stress Urinary Incontinence (SUI) is caused by weakened pelvic floor muscles or a weakened urethral sphincter leading to urine leaks whenever there is sudden physical pressure applied to the abdomen or bladder. This type of urinary incontinence causes sudden spurts of leaked urine when someone coughs, laughs, or sneezes, or with straining and exertion. In women, physical changes can also contribute to stress incontinence, including pregnancy, vaginal deliveries, and menopause. One of the major contributors to stress incontinence is estrogen deficiency, which is thought to increase the chances of leakage by lowering muscle pressure around the urethra. Age, parity, and family history are other risk factors that have been previously noted [1, 2]. Other risk factors influencing the occurrence of SUI include age, delivery mode, concomitant diabetes mellitus, hereditary factors, ethnicity, neurological illnesses, obesity, Parkinson's disease, parity, recurrent urinary tract infections, and pregnancy [3].

SUI has become a widespread disease that affects profoundly the quality of life worldwide [4]. The etiology of SUI is still not completely known, although several studies have suggested that serum and tissue proteins can be bioindicators for SUI. Some of these markers, however, were found to be nonspecific or did not affect the pathophysiology, while others showed no association with SUI [5, 6].

While earlier studies have identified the serum proteomic profile in patients with SUI using database search engines, these findings raise a number of questions, including the issue of sensitivity and specificity.

The present research tries to revisit and implement a workflow for identifying proteins from tandem mass spectrometry data to complement findings from an existing study addressing the pathophysiology of SUI [7]. Mostly, a workflow of shotgun proteomics is a whole process of mass spectrometry data analysis aiming to address a biological question. Over recent years, shotgun proteomics has made tremendous progress and has become the most comprehensive and versatile tool for studying proteins in a large scale [8]. This technique is aimed at identifying proteins in complex mixtures using HPLC in combination with MS/MS.

A shotgun proteomics workflow starts with the extraction of the proteins to be studied, from a tissue or cell, followed by the digestion step using a digestive enzyme, commonly trypsin [9], which generates a group of peptides [10]. These peptides are afterwards inspected by liquid chromatography coupled to mass spectrometry, where they are separated by C18 chromatography in a first step. Secondly, they are electrosprayed into the mass spectrometer where ions are sorted according to their mass/charge ratio. After acquisition and signal processing, we visualize the fragmentation spectra. The data are then analyzed to quantify and identify the specific proteins [11]. The final step of the workflow is the functional analysis where the appropriate proteins are placed on the context of the biological question of interest. One of the most important steps in the process is the identification of proteins. Typically, there are two most common strategies for peptide/protein identification: database search engines, which attempt to compare and match the experimental spectra to identified spectra belonging to peptide libraries, and de novo search engines, which attempt to predict amino acid sequence from its tandem mass spectrum without the assistance of database.

In the aforementioned study of SUI, Marianne Koch et al. identified the serum proteomic profile in patients with SUI. Database search algorithms, X!Tandem [12] and Mascot [13], were used to perform peptide identification. The strategy of database search is a crucial element of shotgun proteomics study [14]. After the acquisition of the experimental spectra, database search algorithms are used to define the best sequence match to the spectrum.

Various algorithms have been developed to carry out searches of MS/MS data and estimate the probability of a match, but they differ in terms of sensitivity and specificity. Sensitivity measures how accurate analysis can detect the smallest number of targeted proteins, while specificity measures the analysis accuracy in singling out target proteins from other (noncontributing) proteins that may be present in the sample [15]. This process can lead to significant false-positive results. The overall estimation of false positives in database search engine results is given as the false discovery rate (FDR)—a measure of the false peptide spectral matches (PSMs) out of the accepted ones [16]. Although various strategies exist to estimate FDR, the target-decoy (TD) database search remains the most commonly used one in shotgun proteomics. In this strategy, the database search is performed on both the true, known as the target and a null database, known as the decoy. The null model is mandatory to estimate the FDR. The decoy database is generated from the target database using different methods, either by randomization, permutation, or reversal. The TD approach assumes that the number of false PSMs in decoy search and false PSMs in target search will be equal above a given threshold score. For the TD approach, the database search can be carried out in two different ways: concatenated or separated. The concatenated search is performed by combining both the target and decoy database together. While the separated search implies searching both the target and decoy database separately.

It is popularly assumed that the accuracy and dynamic range of the analysis increase as the number of PSMs is maximized [17]. Thus, since search engines identify different subsets of PSMs, the idea of combining the capacity of various search engines seems natural to gain a better and more accurate result. Furthermore, in the case where the tandem mass spectra have a consistent fragmentation and a good signal to noise ratio, the identification of the correct sequence match remains fairly straightforward. However, multiple database search algorithms can be combined to perform the analysis when the tandem mass spectra have poor quality or an irregular fragmentation [18]. The results of different search engines can be widely divergent. The disagreement between algorithms is due to the flexibility allowed by some programs to use different types of tandem mass spectrometry data and/or modification patterns.

Handling this large number of spectra, while still minimizing computation time and memory expenses, has become a crucial issue in proteomics research. Fortunately, the recent emergence of cluster computing and high-performance computing (HPC) provides an opportunity to reduce the computation time high-throughput analyses partly by using parallelization to apply more processor power.

The central tenant of our work is to combine the results of various search engines to profit from their varying selectivity and to assess the likely advance made in the previous study. High-performance computing and clustering will be used to manage the complexity of running several search engines and reduce computational time.

In the following sections, we describe our deployed analysis pipeline for protein identification using multiple search engines together with HPC. We also discuss the unique contributions of search engines by comparing their single performance on the SUI dataset. We provide a comparison of the various search engine combinations and highlight those with better sensitivity-specificity tradeoffs. Finally, we present and discuss the accuracy of the identified set of proteins.

The workflow is made available through the following GitHub link: https://github.com/taoufik-elpho/Serum-Analysis-Elpho.

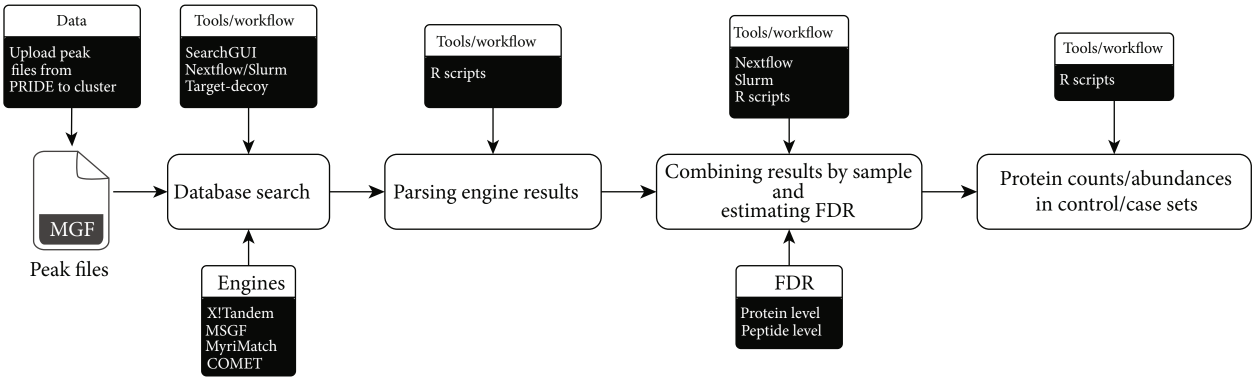

Figure 1: Workflow of the developed approach. The deployed pipeline for identifying proteins using combined database search engines and HPC from data uploading to result generation.

## 2. Material and Methods

The data used in this follow-up study are those of the previous study (PXD008553). The samples are blood serum samples collected from 19 SUI patients and 19 controls. Only 32 samples were publicly available in PRIDE from all the 38 samples. However, four samples were not case-control matched; therefore, a total of 24 samples (12 samples of patients with SUI and 12 controls) were available for the analysis.

The main modifications considered during sample preparation are serum albumin depletion, digestion using a combination of trypsin and Gluc-C, and peptide separation using nano-HPLC. The identification was performed by searching the human Swiss-Prot database on April 15, 2019, and the search was carried out including a decoy database, with an FDR cut-off set to 0.1%. The software used in this analysis is the following: X!Tandem version Vengeance 2015.12.15.2, COMET [19] version 2018.01.rev.3, MS-GF+ version 2018.04.09 [20], and MyriMatch version 2.2.10165 [21] as search engines; PeptideShaker for identification result interpretation [22]; and SearchGUI version 3.3.10 for running and configuring the searches [23].

Figure 1 summarizes the main steps of our deployed workflow. Pipeline scripting was performed using Nextflow, a tool enabling scalable and reproducible computational workflows through software containers. It also simplifies the deployment of complex distributed pipelines [24]. The searches were executed in parallel on a Linux cluster running Slurm for job scheduling and task management [25]. Then, SearchGUI was used to configure the search parameters. The main search configurations were the selection of both Glu-C and trypsin as digestion enzymes; determining carbamidomethyl on Cys being a fixed modification; definition of phosphorylation on Ser, Thr, and Tyr; and oxidation on Met as variable modifications. When the search results are generated, in-house scripts developed in R were used to combine the results of the different search engines and validate peptides/protein by readjusting PTM localization scores and redesigning the protein inference. Once these previous tasks were completed, we started downloading and processing the results. Considering the formats outputted by MyriMatch and MS-GF+ are dissimilar than X!Tandem and COMET output, scripts for processing and parsing results were written using R for this purpose and for executing decoy-based FDR estimation. Error rate calculations were performed on both protein and peptide levels. First, FDR was estimated for each engine per sample; then, a new FDR (1%) was set up taking into account hits found in all engines.

A multitool result requires overcoming some conspicuous obstacles. For merging search engine results, in addition to the issue of different output file formats generated by each engine, the problem of PSM quality matching between engines needs to be clarified.

The quality of a PSM is expressed by a different score parameter for each search engine, which can make matching between a set of PSMs corresponding to the same spectrum difficult [26]. Once the problem of different output file formats is overcome by converting the original search engine output to a common format, we used a decoy match approach by adding decoy proteins in the database search step to surmount the PSM quality matching issue.

The main outcome measures were the proteins detected by sample through the combination of search engine results, the proteins detected in SUI samples and not detected in controls, and the proteins detected in control samples and not in SUI. For the statistical analysis, only proteins present at least 6 times in the same group were used. Those proteins were analyzed and matched against the DisGeNET—a platform of the largest publicly available collections of genes and variants associated with human diseases—(https://www.disgenet.org/) and the KEGG (Kyoto Encyclopedia of Genes and Genomes; https://www.genome.jp/tools/kaas/).

Search engine unique contributions were also evaluated. This task was accomplished by comparing the number of accepted proteins by FDR for each search engine per sample. An upset of control vs. case sample by the search engine was also performed.

## 3. Results and Discussion

While comparing serum samples of SUI patients and controls, we identified 13 induced proteins (abundantly found only in SUI samples) and 26 depleted proteins (abundant only in control samples) as illustrated in Tables 1 and 2.

Table 1: List of proteins detected in SUI samples and not in controls.

| UniProt accession | Protein | Associated gene |
|---|---|---|
| Q96KM6 | Zinc finger protein 512B | ZNF512B |
| Q9HDB5 | Neurexin-3-beta | NRXN3 |
| Q7RTY9 | Serine protease 41 | PRSS41 |
| M0R2J8 | Doublecortin domain-containing protein 1 | DCDC1 |
| O15444 | C-C motif chemokine 25 | CCL25 |
| P04437 | T cell receptor alpha variable 29/delta variable 5 | TRAV29DV5 |
| Q9NQB0 | Transcription factor 7-like 2 | TCF7L2 |
| A8MU46 | Smoothelin-like protein 1 | SMTNL1 |
| P49768 | Presenilin-1 | PSEN1 |
| P20340 | Ras-related protein Rab-6A | RAB6A |
| P16871 | Interleukin-7 receptor subunit alpha | IL7R |
| O43613 | Orexin receptor type 1 | HCRTR1 |
| P28566 | 5-Hydroxytryptamine receptor 1E | HTR1E |

Table 2: List of proteins detected in control samples and not in SUI.

| UniProt accession | Protein | Associated gene |
|---|---|---|
| Q9NZW5 | MAGUK p55 subfamily member 6 | MPP6 |
| Q8WYA0 | Intraflagellar transport protein 81 homolog | IFT81 |
| P04350 | Tubulin beta-4A chain | TUBB4A |
| O94763 | Unconventional prefoldin RPB5 interactor 1 | URI1 |
| Q9H4A4 | Aminopeptidase B | RNPEP |
| Q9UH65 | Switch-associated protein 70 | SWAP70 |
| P31371 | Fibroblast growth factor 9 | FGF9 |
| Q99717 | Mothers against decapentaplegic homolog 5 | SMAD5 |
| O95168 | NADH dehydrogenase [ubiquinone] 1 beta Subcomplex subunit 4 | NDUFB4 |
| Q9ULT8 | E3 ubiquitin-protein ligase HECTD1 | HECTD1 |
| Q96Q35 | Flagellum-associated coiled-coil domain-containing protein 1 | FLACC1 |
| Q9BYX4 | Interferon-induced helicase C domain-containing protein 1 | IFIH1 |
| A3KMH1 | von Willebrand factor A domain-containing protein 8 | VWA8 |
| P14618 | Pyruvate kinase PKM | PKM |
| Q9P253 | Vacuolar protein sorting-associated protein 18 homolog | VPS18 |
| Q9P0U3 | Sentrin-specific protease 1 | SENP1 |
| Q9H201 | Epsin-3 | EPN3 |
| A0JP26 | POTE ankyrin domain family member B3 | POTEB3 |
| Q99999 | Galactosylceramide sulfotransferase | GAL3ST1 |
| Q70CQ1 | Ubiquitin carboxyl-terminal hydrolase 49 | USP49 |
| Q9Y6X6 | Unconventional myosin-XVI | MYO16 |
| Q8NGF1 | Olfactory receptor 52R1 | OR52R1 |
| Q8WTV1 | THAP domain-containing protein 3 | THAP3 |
| Q8N126 | Cell adhesion molecule 3 | CADM3 |
| Q9UEG4 | Zinc finger protein 629 | ZNF629 |
| A6NDP7 | Myeloid-associated differentiation marker-like protein 2 | MYADML2 |

The four most abundant proteins identified in SUI patients were Q9NQB0, Q93074, A8MU46, and P28566. All four proteins were found involved in prenatal exposure delayed effects and body weight change [27–29]. Smoothelin-like protein 1 plays a role in smoothing muscle fibers and mediating vascular adaptation to exercise and regulating contraction and relaxation of skeletal. It is furthermore implicated in fetal growth retardation disease and memory disorders [30].

Transcription factor 7-like 2 is implicated in blood glucose homeostasis. Genetic variants of its gene have been found to be associated with an increased risk of type 2 diabetes [31].

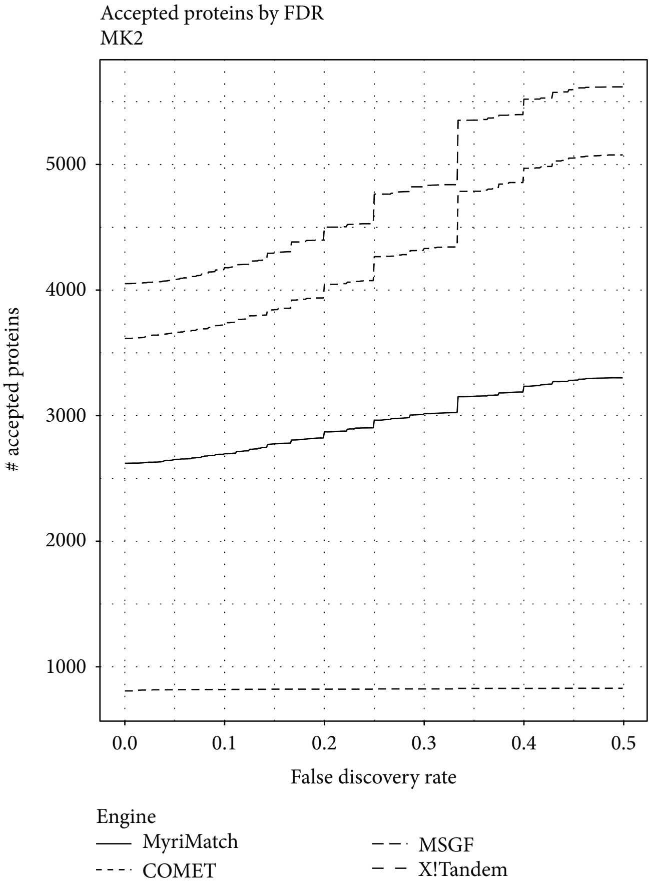

Figure 2: Comparison of each single-engine performance. False discovery rate vs. accepted proteins by each search engine.

A mediator protein was found in SUI samples, the MED12 protein. This protein is crucial for activating CDK8 kinase which plays a role in modulating mediator-polymerase II interactions to regulate transcription initiation and reinitiation rates [32]. Phenotypes linked to Med12 are Lujan-Fryns syndrome, Ohdo syndrome, and X-linked Opitz-Kaveggia syndrome known as FG syndrome.

A G-protein-coupled receptor was also found in significantly higher abundance in SUI patients. 5HT1E (G-protein-coupled receptor for serotonin) is also known to act as a receptor for different psychoactive substances and alkaloids [33]. This protein is shown to be involved in mental disorders and substance-related disorders [34].

Five proteins were found fairly abundant in control samples. They were shown to be involved in two main class diseases, neurodegenerative diseases and female urogenital diseases and pregnancy complications.

A member of the fibroblast growth factor family, P31371, is involved in a variety of biological processes, such as cell growth, tissue repair, and embryonic development. It is also thought to have a role in brain tissue regeneration [35]. This protein was associated with prenatal injuries and fetal death [36, 37].

An enzyme was singled out in control samples; it is a member of the protein family ubiquinone oxidoreductase subunit NDUFB4. O95168 was shown to be implicated in the mitochondrial membrane respiratory chain complex I assembly, mitochondrial electron transport, and response to oxidative stress [38, 39]. Links with nerve degeneration and nervous system diseases were determined for this protein [40, 41].

A further enzyme was found abundantly in controls belonging to the peptidase C19 family. It is known to catalyze various reactions, such as peptide and isopeptide bonds formed by the C-terminal Gly of ubiquitin [42]. It was also revealed that Q70CQ1 performs deubiquitinating of histone H2B at “Lys-120” and acts as a regulator of pre-mRNA splicing [43]. Implications in urogenital abnormalities, female infertility, and female genital diseases were demonstrated for USP49 [44–46].

THAP3 is an element of the THAP1/THAP3-HCFC1-OGT complex. The protein is implicated in the regulation of the transcriptional activity of the cell cycle-specific gene PRM1 [47]. THAP3 participates also in some molecular functions such as DNA binding and metal ion binding [48]. Besides its implication in urogenital diseases and pregnancy complications, THAP3 is shown to be a biomarker for Polycystic ovary syndrome [49].

The last protein found in high abundance in controls is myeloid-associated differentiation marker-like protein 2. MYADML2 is an integral component of the membrane and is predicted to localize to the cytoplasm. It is found to be involved in congenital, hereditary, and neonatal diseases and abnormalities and uterine diseases [50, 51].

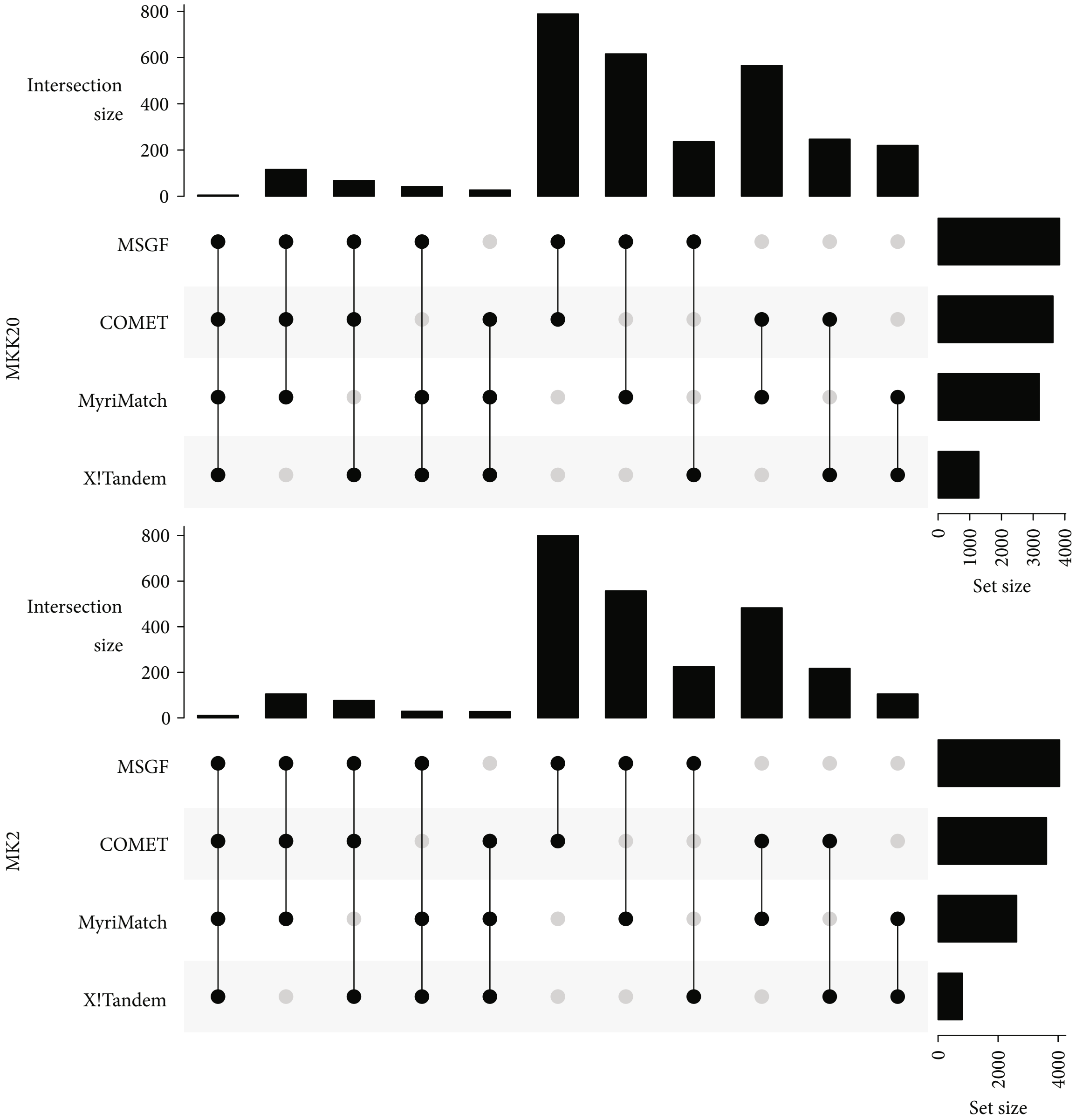

Figure 3: Comparison of the different combinations of search engines. Assessment of all the possible combinations of the four used engines through case control (MK2-MKK20). The set size represents the number of identified proteins by every single engine. While the intersection size represents the number of identified proteins by each combination performed.

Search engine unique contribution was also assessed. We evaluated the unique performances of accepted proteins by FDR for each of the four search engines. Figure 2 shows that MS-GF+ performance was outstanding by identifying the larger number of accepted proteins followed by COMET who competed well. It should be mentioned that the weak performance of X!Tandem could be blamed to its scoring function.

Figure 3 compares the number of identified proteins by each engine per group, as well as the intersection of result search engines through different combinations. The total number of detected proteins in controls is slightly higher than those detected in SUI samples, MS-GF+ reached 4000 proteins in control against 3000 proteins in SUI. This difference is due to the biological variance of the sample. MS-GF+ outperformed MyriMatch, COMET, and X!Tandem for both groups.

One can also notice that the combination of MS-GF+ and COMET behaves better than others. Rationally, such a combination is expected to produce results more correct than those of other combinations. It is logical to assume that the dissimilarity of the scoring function used by each of the two engines leads to a better separation between correct and incorrect identifications. MS-GF+ uses a robust probabilistic model while COMET uses a descriptive model. It is popularly assumed that algorithms based on the descriptive approach show a better sensitivity, whereas algorithms based on the

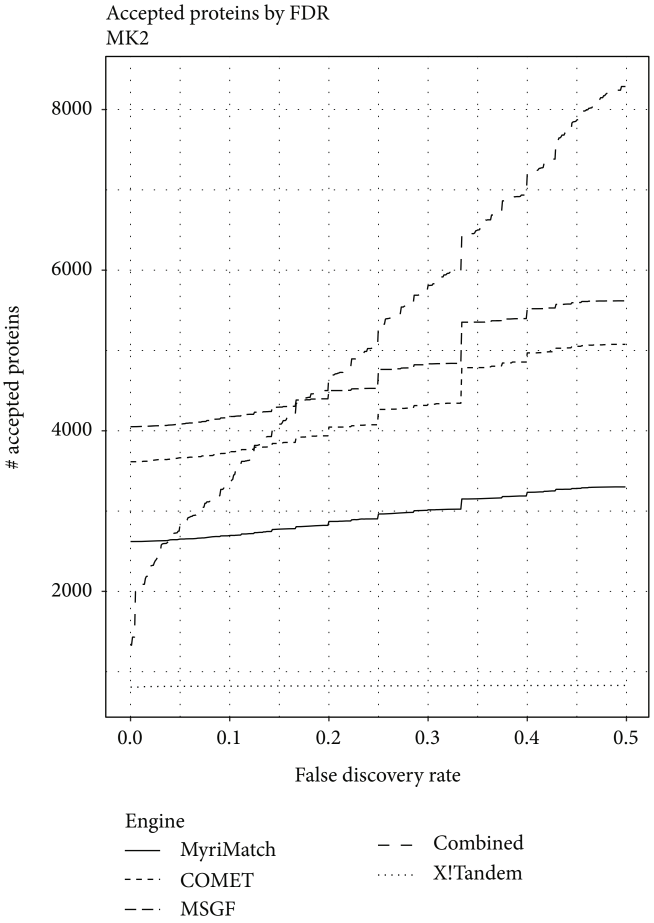

Figure 4: Comparison of the combined method with every single engine. FDR vs. accepted proteins by the combined approach and by each search engine. Within the usable range of FDR values (<0.05), our method shows an accurate selectivity comparing to single-engine results. Passing the usable range of FDR (higher FDR values), for each single engine, almost all the hits are already accepted which explains the slow increase. While the combined approach will continue increasing for the simple fact of summing the reported hits from all engines, the number of accepted proteins for the combined approach is found highly correlated to the FDR.

probabilistic approach show a better specificity. This is the probable reason that the MS-GF+ and COMET combination outran the other combination.

Finally, we compared the unique contributions of the four search engines (MS-GF+, COMET, MyriMatch, and X!Tandem) with the combined approach that represents the combination of the four search engine results. As illustrated in Figure 4, the combined approach shows an accurate selectivity within the usable range of FDR (from 0 to almost 1%). Passing this range, the slow increase of other curves is explained by the fact that all the hits are already reported. The combined approach still increases to the simple fact of summing the identified hits from each engine.

The combined approach combined four search engine results, which implies four different scoring functions. Accordingly, the approach benefited from the selectivity and sensitivity of each search engine.

## 4. Conclusion

In this work, we have utilized an existing study of SUI to assess and validate our developed workflow for protein identification using a combined approach of database searching together with HPC.

Thirteen proteins were found exclusively in SUI samples and are known to be associated with prenatal exposure delayed effects and body weight change. The twenty-six proteins found exclusively in control samples belonged to two main classes of female urogenital diseases related to risk factors influencing SUI. Our analysis is agnostic to whether these proteins are causal in the development process of SUI or simply the result of SUI. Further studies of the identified proteins are required to answer this question.

We have designed and implemented a workflow for protein identification by combining multiple search engine results. We have opted for high-performance computing and cloud computing for managing the multiple searches and reducing the computational time expense.

By examining different search engine combinations, we showed that MS-GF+ and COMET led to a dramatic improvement in protein identification accuracy. Thus, the selection of search engines should also consider the complementarity of their scoring model.

In processing the same dataset using our developed approach, we were able to identify a more accurate set of proteins shown to be involved in diseases associated with risk factors affecting the pathology. In fact, the KEGG and enrichment analyses showed the top four most significantly induced proteins were involved mainly in prenatal exposure delayed effects and body changes, and the five most significantly depleted proteins were shown to be implicated mostly in two main class diseases: neurodegenerative diseases and female urogenital diseases and pregnancy complications. Given that the aforementioned risk factors influencing SUI generally include neurological illnesses, obesity, Parkinson's disease, parity, recurrent urinary tract infections, and pregnancy, thus, our results show that the developed approach succeeded to be more accurate during the identification process. In general, combining search engines serves to benefit from the strength of each engine and thus complement the peptide/protein identification.

We have also demonstrated that the combined approach improves the specificity and sensitivity of the analysis by increasing the confidence of identified proteins.

## Data Availability

The datasets generated and analyzed during the current study as well as the whole developed workflow are fully reproducible and available in the GitHub repository, https://github.com/taoufik-elpho/Serum-Analysis-Elpho.

## Conflicts of Interest

The authors have declared no conflict of interest.

## References

[1] X. Fritel, V. Ringa, N. Varnoux, A. Fauconnier, S. Piault, and G. Bréart, "Mode of delivery and severe stress incontinence. A cross-sectional study among 2,625 perimenopausal women," *BJOG: An International Journal of Obstetrics and Gynaecology*, vol. 112, no. 12, pp. 1646–1651, 2005.

[2] L. L. Subak, H. E. Richter, and S. Hunskaar, "Obesity and urinary incontinence: epidemiology and clinical research update," *The Journal of Urology*, vol. 182, 6 Supplement, pp. S2–S7, 2009.

[3] I. Milsom and M. Gyhagen, "The prevalence of urinary incontinence," *Climacteric*, vol. 22, no. 3, pp. 217–222, 2019.

[4] A. D. Garely and N. Noor, "Diagnosis and surgical treatment of stress urinary incontinence," *Obstetrics and Gynecology*, vol. 124, no. 5, pp. 1011–1027, 2014.

[5] R. Cartwright, A. C. Kirby, K. A. O. Tikkinen et al., "Systematic review and metaanalysis of genetic association studies of urinary symptoms and prolapse in women," *American Journal of Obstetrics and Gynecology*, vol. 212, no. 199, pp. 124–191, 2015.

[6] P. Skorupski, K. Jankiewicz, P. Miotla, M. Marczak, B. Kulik-Rechberger, and T. Rechberger, "The polymorphisms of the *MMP-1* and the *MMP-3* genes and the risk of pelvic organ prolapse," *International Urogynecology Journal and Pelvic Floor Dysfunction*, vol. 24, no. 6, pp. 1033–1038, 2013.

[7] M. Koch, W. Umek, E. Hanzal et al., "Serum proteomic pattern in female stress urinary incontinence," *Electrophoresis*, vol. 39, no. 8, pp. 1071–1078, 2018.

[8] T. Nilsson, M. Mann, R. Aebersold, J. R. Yates 3rd, A. Bairoch, and J. J. Bergeron, "Mass spectrometry in high-throughput proteomics: ready for the big time," *Nature Methods*, vol. 7, no. 9, pp. 681–685, 2010.

[9] J. R. Wiśniewski and M. Mann, "Consecutive proteolytic digestion in an enzyme reactor increases depth of proteomic and phosphoproteomic analysis," *Analytical Chemistry*, vol. 84, no. 6, pp. 2631–2637, 2012.

[10] Y. H. Lee, H. T. Tan, and M. C. Chung, "Subcellular fractionation methods and strategies for proteomics," *Proteomics*, vol. 10, no. 22, pp. 3935–3956, 2010.

[11] M. Y. Hein, K. Sharma, J. Cox, and M. Mann, *Handbook of Systems Biology*, Elsevier BV, Amsterdam, 2013.

[12] R. Craig and R. C. Beavis, "TANDEM: matching proteins with tandem mass spectra," *Bioinformatics*, vol. 20, no. 9, pp. 1466-1467, 2004.

[13] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis*, vol. 20, no. 18, pp. 3551–3567, 1999.

[14] W. Zhang and X. Zhao, "Method for rapid protein identification in a large database," *BioMed Research International*, vol. 2013, no. 7, Article ID 414069, 2013.

[15] A. J. Saah and D. R. Hoover, ""Sensitivity" and "specificity" reconsidered: the meaning of these terms in analytical and diagnostic settings," *Annals of Internal Medicine*, vol. 126, no. 1, pp. 91–94, 1997.

[16] A. I. Nesvizhskii, "A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics," *Journal of Proteomics*, vol. 73, no. 11, pp. 2092–2123, 2010.

[17] D. Shteynberg, A. I. Nesvizhskii, R. L. Moritz, and E. W. Deutsch, "Combining results of multiple search engines in Proteomics," *Molecular & Cellular Proteomics*, vol. 12, no. 9, pp. 2383–2393, 2013.

[18] M. Bern, D. Goldberg, W. H. McDonald, and J. R. Yates 3rd., "Automatic quality assessment of peptide tandem mass spectra," *Bioinformatics*, vol. 20, Supplement 1, pp. i49–i54, 2004.

[19] J. K. Eng, T. A. Jahan, and M. R. Hoopmann, "Comet: an open-source MS/MS sequence database search tool," *Proteomics*, vol. 13, no. 1, pp. 22–24, 2013.

[20] S. Kim and P. Pevzner, "MS-GF+ makes progress towards a universal database search tool for proteomics," *Nature Communications*, vol. 5, no. 1, 2014.

[21] D. L. Tabb, C. G. Fernando, and M. C. Chambers, "MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis," *Journal of Proteome Research*, vol. 6, no. 2, pp. 654–661, 2007.

[22] M. Vaudel, J. M. Burkhart, R. P. Zahedi et al., "PeptideShaker enables reanalysis of MS-derived proteomics data sets," *Nature Biotechnology*, vol. 33, no. 1, pp. 22–24, 2015.

[23] H. Barsnes and M. Vaudel, "A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines," *Journal of Proteome Research*, vol. 17, no. 7, pp. 2552–2555, 2018.

[24] D. P. Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, "Nextflow enables reproducible computational workflows," *Nature Biotechnology*, vol. 35, no. 4, pp. 316–319, 2017.

[25] A. B. Yoo, M. A. Jette, and M. Grondona, "JSSPP," *Lecture Notes in Computer Science*, vol. 2862, 2003.

[26] A. Lin, J. J. Howbert, and W. S. Noble, "Combining high-resolution and exact calibration to boost statistical power: a well-calibrated score function for high-resolution MS2 data," *Journal of Proteome Research*, vol. 17, no. 11, pp. 3644–3656, 2018.

[27] M. Puttabyatappa, J. D. Martin, V. Andriessen et al., "Developmental programming: changes in mediators of insulin sensitivity in prenatal bisphenol A-treated female sheep," *Reproductive Toxicology*, vol. 85, pp. 110–122, 2019.

[28] Z. Drobna, A. Talarovicova, H. E. Schrader, T. R. Fennell, R. W. Snyder, and E. F. Rissman, "Bisphenol F has different effects on preadipocytes differentiation and weight gain in adult mice as compared with Bisphenol A and S," *Toxicology*, vol. 420, no. 420, pp. 66–72, 2019.

[29] M. C. Kang, N. Kang, S. Y. Kim et al., "Popular edible seaweed, *Gelidium amansii* prevents against diet-induced obesity," *Food and Chemical Toxicology*, vol. 90, pp. 181–187, 2016.

[30] C. Wang, R. Niu, Y. Zhu et al., "Changes in memory and synaptic plasticity induced in male rats after maternal exposure to bisphenol A," *Toxicology*, vol. 322, no. 322, pp. 51–60, 2014.

[31] A. F. Hansen, A. Simić, B. O. Åsvold et al., "Trace elements in early phase type 2 diabetes mellitus-a population-based study. The HUNT study in Norway," *Journal of Trace Elements in Medicine and Biology*, vol. 40, pp. 46–53, 2017.

[32] H. J. Baek, Y. K. Kang, and R. G. Roeder, "Human mediator enhances basal transcription by facilitating recruitment of transcription factor IIB during preinitiation complex assembly," *The Journal of Biological Chemistry*, vol. 281, no. 22, pp. 15172–15181, 2006.

[33] J. M. Zgombick, L. E. Schechter, M. Macchi, P. R. Hartig, T. A. Branchek, and R. L. Weinshank, "Human gene S31 encodes the pharmacologically defined serotonin 5-hydroxytryptamine1E receptor," *Molecular Pharmacology*, vol. 42, no. 2, pp. 180–185, 1992.

[34] Y. S. Kim, M. Yang, W. K. Mat et al., "*GABRB2* haplotype association with heroin dependence in Chinese population," *PLoS One*, vol. 10, no. 11, article e0142049, 2015.

[35] X. Zhang, O. A. Ibrahimi, S. K. Olsen, H. Umemori, M. Mohammadi, and D. M. Ornitz, "Receptor specificity of the fibroblast growth factor family. The complete mammalian FGF family," *The Journal of Biological Chemistry*, vol. 281, no. 23, pp. 15694–15700, 2006.

[36] A. R. Hindman, X. M. Mo, H. L. Helber et al., "Varying susceptibility of the female mammary gland to *in utero* windows of BPA exposure," *Endocrinology*, vol. 158, no. 10, pp. 3435–3447, 2017.

[37] K. M. Gaworecki, R. W. Chapman, M. G. Neely, A. R. D'Amico, and L. Bain, "Arsenic exposure to killifish during embryogenesis alters muscle development," *The Journal of Toxicological Sciences*, vol. 125, no. 2, pp. 522–531, 2012.

[38] D. A. Stroud, E. E. Surgenor, L. E. Formosa et al., "Accessory subunits are integral for assembly and function of human mitochondrial complex I," *Nature*, vol. 538, no. 7623, pp. 123–126, 2016.

[39] J. Murray, S. W. Taylor, B. Zhang, S. S. Ghosh, and R. A. Capaldi, "Oxidative damage to mitochondrial complex I due to peroxynitrite," *The Journal of Biological Chemistry*, vol. 278, no. 39, pp. 37223–37230, 2003.

[40] I. A. Adedara, O. Owoeye, I. O. Awogbindin, B. O. Ajayi, J. B. Rocha, and E. O. Farombi, "Diphenyl diselenide abrogates brain oxidative injury and neurobehavioural deficits associated with pesticide chlorpyrifos exposure in rats," *Chemico-Biological Interactions*, vol. 296, pp. 105–116, 2018.

[41] P. T. Theunissen, J. F. Robinson, J. L. A. Pennings et al., "Transcriptomic concentration-response evaluation of valproic acid, cyproconazole, and hexaconazole in the neural embryonic stem cell test (ESTn)," *Toxicological Sciences*, vol. 125, no. 2, pp. 430–438, 2012.

[42] V. Quesada, A. Díaz-Perales, A. Gutiérrez-Fernández, C. Garabaya, S. Cal, and C. López-Otín, "Cloning and enzymatic analysis of 22 novel human ubiquitin-specific proteases," *Biochemical and Biophysical Research Communications*, vol. 314, no. 1, pp. 54–62, 2004.

[43] Z. Zhang, A. Jones, H. Y. Joo et al., "USP49 deubiquitinates histone H2B and regulates cotranscriptional pre-mRNA splicing," *Genes & Development*, vol. 27, no. 14, pp. 1581–1595, 2013.

[44] A. Colciago, L. Casati, O. Mornati et al., "Chronic treatment with polychlorinated biphenyls (PCB) during pregnancy and lactation in the rat part 2: effects on reproductive parameters, on sex behavior, on memory retention and on hypothalamic expression of aromatase and 5alpha-reductases in the offspring," *Toxicology and Applied Pharmacology*, vol. 239, no. 1, pp. 46–54, 2009.

[45] M. M. Milesi, J. Varayoud, J. G. Ramos, and E. H. Luque, "Uterine ER$\alpha$ epigenetic modifications are induced by the endocrine disruptor endosulfan in female rats with impaired fertility," *Molecular and Cellular Endocrinology*, vol. 454, pp. 1–11, 2017.

[46] S. M. Dickerson, S. L. Cunningham, H. B. Patisaul, M. J. Woller, and A. C. Gore, "Endocrine disruption of brain sexual differentiation by developmental PCB exposure," *Endocrinology*, vol. 152, no. 2, pp. 581–594, 2011.

[47] T. Clouaire, M. Roussigne, V. Ecochard, C. Mathe, F. Amalric, and J. P. Girard, "The THAP domain of THAP1 is a large C2CH module with zinc-dependent sequence-specific DNA-binding activity," *Proceedings of the National Academy of Sciences*, vol. 102, no. 19, pp. 6907–6912, 2005.

[48] R. Mazars, A. Gonzalez-de-Peredo, C. Cayrol et al., "The THAP-zinc finger protein THAP1 associates with coactivator HCF-1 and O-GlcNAc transferase: a link between dyt6 and dyt3 dystonias," *Journal of Biochemistry*, vol. 285, pp. 13364–13371, 2010.

[49] R. F. Savaris, J. M. Groll, S. L. Young et al., "Progesterone resistance in PCOS endometrium: a microarray analysis in clomiphene citrate-treated and artificial menstrual cycles," *The Journal of Clinical Endocrinology and Metabolism*, vol. 96, no. 6, pp. 1737–1746, 2011.

[50] Y. Dang, F. Wang, and C. Liu, "Real-time PCR array to study the effects of chemicals on the growth hormone/insulin-like growth factors (GH/IGFs) axis of zebrafish embryos/larvae," *Chemosphere*, vol. 207, pp. 365–376, 2018.

[51] R. G. Berger and W. G. Foster, "Bisphenol-A exposure during the period of blastocyst implantation alters uterine morphology and perturbs measures of estrogen and progesterone receptor expression in mice," *Reproductive Toxicology*, vol. 30, no. 3, pp. 393–400, 2010.