Industrial Engineering Journal
ISSN: 0970-2555
Volume : 51, Issue 04, April : 2022

The human body fluid proteome: quantitative profiling and computational prediction

Mr. Gopal Behera[1], Hemanta Paikray[2]*

[1] Assistant Professor Dept. Of Computer Science and Engineering, NIT , BBSR*
[2] Assistant Professor,Dept. Of Computer Science and Engineering, NIT , BBSR
gopalbehera@thenalanda.com hemantpaikray@thenalanda.com*

Abstract

Thanks to advances in high-throughput biotechnologies, recent proteome studies of body fluids have led to the discovery of many new disease biomarkers and therapeutic drugs. Meanwhile, enormous progress has been made in discovering the proteome of body fluids, resulting in the discovery of more than 15,000 different proteins in the main fluids of the human body. However, current proteomics techniques still share the challenges of effectively addressing the wide range of protein modifications in these fluids. To this end, computational work using statistical and machine learning methods has shown early success in identifying biomarker proteins in certain human diseases. In this article, we first summarized the experimental progress using a combination of conventional and high-throughput techniques and the main findings, and focused on current research on 16 types of body fluid proteins. Next, emerging computational work on protein prediction based on support vector machine, alignment algorithm and protein-protein interaction network was also mapped, followed by a discussion on the algorithm and applications. Finally, we also discuss other critical issues related to these topics and conclude the review by proposing future perspectives specifically for the clinical disease discovery application of biomarkers.

Key words: body-fluid proteome; protein prediction; clinical application; biomarker discovery

Introduction

Human body fluids are biological fluids that are either excreted or secreted from the bodies of living people [1]. They include, but not limited to, plasma/serum, saliva, urine, cerebrospinal fluid, seminal fluid, amniotic fluid, tear fluid, bronchoalveolar lavage

fluid, milk, synovial fluid, nipple aspirate fluid, cervicovaginal fluid, pleural effusion, sputum, exhaled breath condensate and pancreatic juice. It has been widely accepted that human body fluids contain disease-associated proteins that are secreted or leaked from pathological tissues across the body and are often

easily obtainable through noninvasive procedures [2]. To date, over 15 000 different proteins have been identified in major human body fluids.

For decades, proteomic applications have spanned across different fields in biomedical and biochemistry research [3] and considered body fluids as the easy and attractive targets to pro-file [4]. Since the first research on serum globulin separation in 1937 [5], numerous reports on human body-fluid proteomes have been documented. Especially after the use of two-dimensional gel electrophoresis (2-DE) [6], several instrumental milestones appear. For example, in 1970, Freeman and Smith resolved 60 protein components in plasma using conventional gel filtration [7], which clearly demonstrated the complex composition of plasma and the feasibility of profiling blood proteins using those techniques. However, despite its popularity, 2D electrophoresis was known to have limits in terms of its very low efficiency in the analysis of hydrophobic proteins and high sensitivity to the dynamic range and quantitative distribution [8]. Such drawbacks necessitate high-power analytical techniques. In the early 1900s, Sir J.J. Thomson developed mass spectrometry (MS) technique when he obtained mass spectra of small gaseous ions [9]. Since then, MS has become the method of choice for analyz-ing complex protein samples [10]. Since early 1980s, MS moved from an analytical technique applicable only to small volatile compounds to applications on large biomolecules [11] and has been widely used for comprehensive profiling of human body- fluids proteomes. For examples, the Human Plasma Proteome Project initiated by international Human Proteome Organization has involved a collaboration of many laboratories using MS technology and compiled a core dataset of 3521 distinct proteins in human plasma [12]. In addition, the Sys-BodyFluid database published in 2009 contained 11 kinds of body-fluid proteomes and over 10 000 proteins [13]. Many recent studies concentrated on the discovery of protein biomarkers in pathologic conditions such as cancers, metabolic disease and brain disease [14]. In this regard, it is well-known that the major bottleneck challenges in biomarker discovery lie in the quantitative analysis of highly specific proteins [15]. For example, diseases such as Sjögren's syndrome, bacterial and viral infectious diseases, and oral cancer all cause alterations of salivary protein expression [16]. Similarly, urine drains from the urinary tract and is there-fore particularly enriched in proteins deriving from the kidney, bladder and prostate [17]. Many hereditary glomerular disease proteins have been identified in urine, such as podocin, alpha- actinin-4, CD2-

associated protein, myosin-9, myosin 1E, integrin alpha 3 and cubilin, for which the quantitative measure is key to the applications [18].

The ability of MS to identify and to increasingly precisely quantify thousands of proteins from complex fluids has a broad impact on biomedical research [19]. However, regardless the evolving technology, protein identification is still considered as a challenging topic simply because a large amount of proteins are subject to a variety of modifications in body fluids, making the proteome composition highly complex. To facilitate such research, a few computational pipelines have been developed to characterize molecular features of various types of secreted pro-teins and provide new predictions using statistical and machine-learning methodologies. In 2008, Cui *et al.* [20] firstly proposed a machine-learning strategy to predict if a protein is likely to enter into bloodstream using support vector machine (SVM) clas- sifier. Soon after that, several related studies were reported to identify secreted proteins associated with different body fluids, including blood [21], urine [22,23], saliva [24,25] and others [26]. In addition to protein identification, those predictors can be used to identify potential biomarkers for specific human diseases based on the context-dependent genomics data [20]. Figure 1 shows the major event nodes related to body-fluid proteome research. In e following sections, we first review the major tech- niques and discoveries in protein identification and then focused on the computational work in this field in terms of the methodologies and applications. The discussion will be centered around critical issues related to future application in human fluid proteomics research.

Major methodological strategies for body-fluids protein profiling

Modern proteomic tools have provided different technical frameworks for handling proteome complexity in human body fluids [27]. Several previous works have addressed important issues related to the standardization of sample collection, separation and processing [28,29]. As a summary, Figure 2 shows the currently used analytical workflows, including technologies used to fractionate and analyze proteome in either qualitative or quantitative manner [30].

The qualitative separation was mainly through 1-DE, 2-DE and chromatography. Although 2-DE is low-cost, reproducible and visual, questions remain concerning its ability of handling protein co-migration [31] and limitations in protein analysis for high- or low-molecular weight proteins as well as those of proteins with extreme isoelectric point (p*I*) values [32]. In contrast, multiple liquid chromatography (LC) techniques and their continuous improvements in separation components are providing further advances and enabling increasingly effective large-scale proteomics [33].

A number of isotope-labeling approaches are available for quantitative proteomic analysis [34], including 2D difference gel electrophoresis [35], isotope-coded affinity tag [36], stable isotope labeling by amino acids in cell culture [37], isobaric tags for relative and absolute quantification [38]. Although in general isotopic labeling technology is deemed successful, it has some technical limitations due to the high costs of the labeling reagents, computational difficulties and the error-prone nature [39]. The ion intensity-based label-free quantitative approach has gradually gained more popularity and provides an alterna- tive powerful tool to resolve and identify thousands of proteins from a complex biological sample [40]. It is rapid and sensitive and can increase the protein dynamic range by 3- to 4-fold compared with 2-DE [41]. Similarly, protein chip has also been employed as a simple-to-use technology that offers the capa- bility of differentiating proteins and quantifying the abundance [42,43].

MS has become an indispensable analytical tool in quanti- tative protein analysis. Particularly, both matrix-assisted laser desorption ionization–time of flight (MALDI–TOF) MS and tan- dem MS (MS/MS) can provide excellent mass accuracy, high resolution, high sensitivity and direct analysis from complex mixtures [44].

Proteomic analysis on 16 types of human body fluids

In this section, we review proteomic research on 16 major types of body fluids since 2001 and summarize the major discovery of body-fluid proteins on Figure 3 shows the distribution of the 16 types of body fluids in human body.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 51, Issue 04, April : 2022

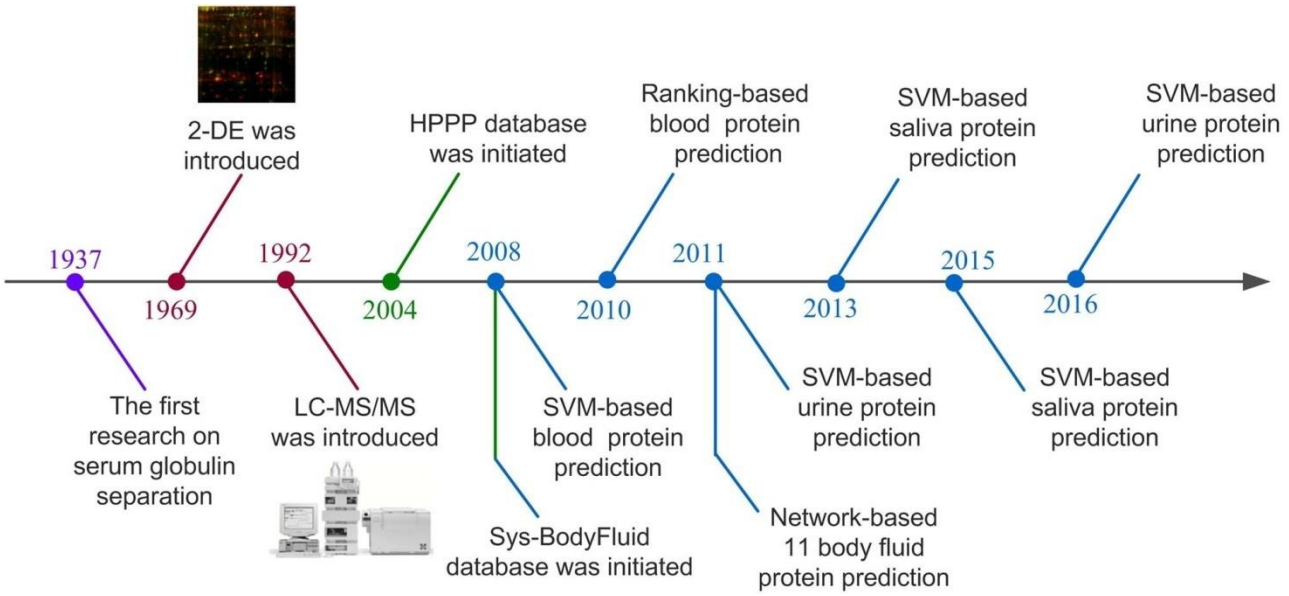

Figure 1. Major events related to proteomics technology development and body-fluid proteome research.

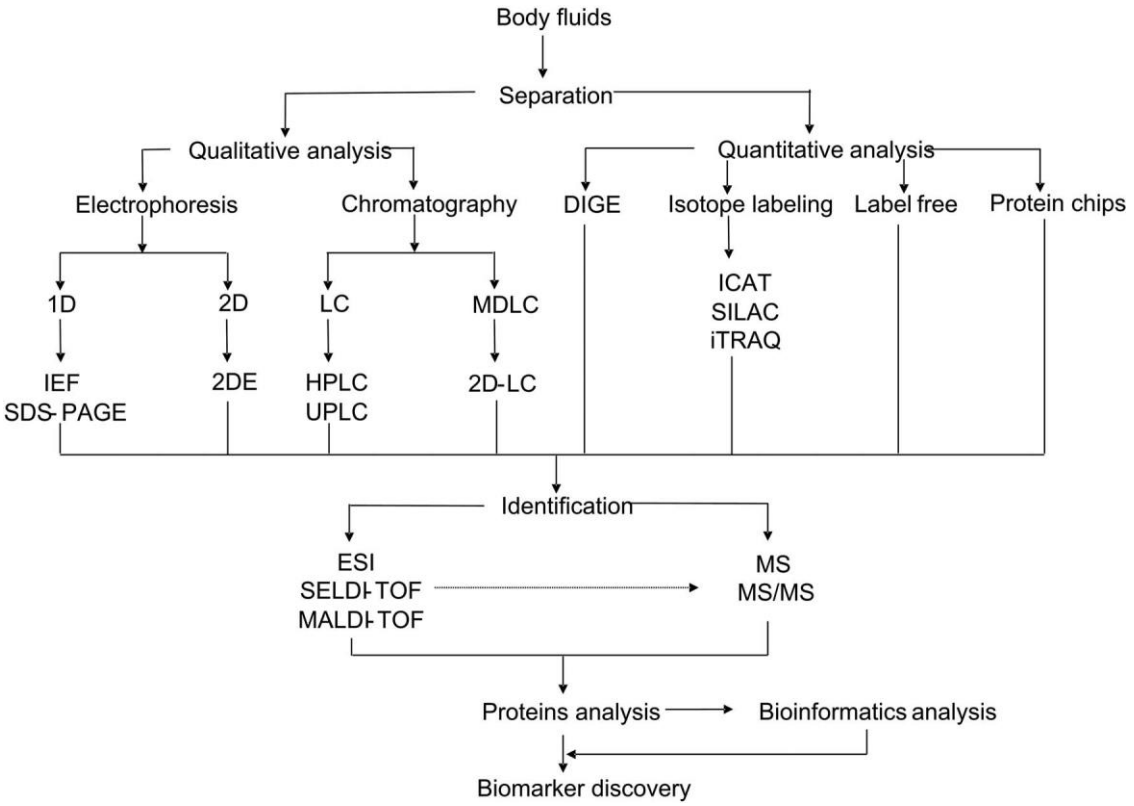

Figure 2. Overview of different strategies used for human body fluids analysis aiming biomarker discovery.

Plasma/serum

Blood plasma is believed to have the most complex human- derived proteome [45] and has attracted high volume of research attentions [45–111]. Owing to the importance of plasma proteins, several large-scale proteomic efforts have been carried out on human plasma proteins [112]. To date, over 12 000 different plasma proteins have been identified with high confidence, which provides the largest set of circulating proteins as themost commonly-used pool for finding potential biomarkers for clinical diagnosis. In the meantime, great challenges remainbecause of the complex modification of proteins in blood.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 51, Issue 04, April : 2022

Saliva

Saliva mainly comes from parotid, submandibular, sublingual and several minor glands, and is a dilute aqueous solutionconsisting of electrolytes, minerals, buffers and proteins [113]. The collection of saliva is simple, noninvasive and cheap and can be easily repeated [28]. The saliva proteome research has led to the identification of more than 4000 different protein species[4,105,107,113–142]. In the context of clinical proteomics, it hasgained increasing potential for disease diagnosis using saliva proteins, especially in oral cancers [105, 130,137] and periodontaldiseases [129].

Urine

Urine is a complex fluid comprised of proteins from the differentsources, including the filtration of the blood within the glomeru-lus and secretion form the kidneys and the urogenital tract[143]. Urine has the advantage to be obtained in large volume. In 1979, the Anderson's group published the first studies by 2-DE on normal urine [144], in which they identified only the major components. Up to today, more than 8000 proteins have been identified in human urine [17,18,107,143, 145–165]. Early success has made toward the development of candidate biomarker in urine for various urogenital diseases, including acute kidney injury, bladder cancer and diabetic nephropathy [164].

Cerebrospinal fluid

Cerebrospinal fluid (CSF) is in continuum with the extra-cellular fluid of the central nervous system (CNS) and is produced by the choroid plexus that surrounds the brain [14]. Several proteomic studies were conducted to identify the proteome of human cerebrospinal fluid, and over 6000 proteins have been identified [107,150,166–179]. CSF is a promising source for studying protein biomarkers of diseases in the CNS [170] and provides an acces-sible liquid pool in the brain [173].

Seminal fluid

Seminal fluid is the liquid component of sperm [180]. In the case of studying human seminal plasma, the main aim would be the discovery of new biomarkers for prostate and testis cancers[181]. In addition, it also sheds new light into the fundamental aspects of the human sperm and points to new potential pro- teins involved in male infertility [182].

Amniotic fluid

Amniotic fluid (AF) contains cells of fetal origin and a wide range of fetal proteins, and is formed from fetal urine and secretions [183]. Proteomic profiles of amniotic fluid have been generated by several groups using different methods since 1997 [184]. As animportant source of biomarkers for fetal pathologies, amniotic fluid has been widely studied for diagnosis of many pregnancy- related pathologies and genetic diseases [15,185–189], includ- ing fetal abnormalities [15], gestational age-dependent changes[189] and so on.

Tear fluid

Tear fluid (TF) is a complex mixture of secretions produced by the lacrimal gland, goblet cells, cornea and vascular sources [190]. Many methods have been used to map tear protein profiles,including different MS technologies [191], such as MALDI–TOF [192] and LC/MS [193]. TF is becoming an increasingly important source for finding biomarkers for eye-related diseases, such as Graves' ophthalmopathy [194].

Bronchoalveolar lavage fluid

Bronchoalveolar lavage fluid (BALF) is a clinical body fluid used in sampling of the soluble protein contents of the airway lumen [195]. One of the earliest attempts to map the protein compo- nents of normal human BALF has identified 49 proteins [196]. Since then, more than 1000 proteins have been identified [195– 205]. BALF also has the great advantage of easy collection and lung-disease indication therefore has been widely studied in ventilator-associated pneumonia [201], lung cancer [198,204], lung adenocarcinoma [197] and chronic obstructive pulmonarydisease (COPD) [206].Milk

Human milk contains many bioactive proteins that serve asthe first source of nutrition for mammalian infants [207]. Over 1700 proteins have been identified in human milk [208]. Among them, milk fat globule membrane [209] and human colostrum [210] have become important targets for proteomics research. In an effort to explore the benefits that human milk can provide, numerous proteomic studies investigated the proteins in milk whey [211,212], which comprises 40.0% of the total milk proteins and has strong implication in growth/maintenance and immu- nity support.

Synovial fluid

Synovial fluid is a serum filtrate located in the joints that con- tains proteins from surrounding tissues, articular cartilage, syn- ovial membrane and bone [213]. Many research on synovial- fluid proteome focused on the rheumatoid arthritis [213,214] and osteoarthritis [2, 214–220]. To date, only less than 1000 proteins can be identified in synovial fluid.

Nipple aspirate fluid

Nipple aspirate fluid (NAF) is a fluid secreted by the epithelial cells of the mammary ductal and lobular system, and it containsa set of specific breast tissue proteins [221, 222]. Therefore, NAF proteome is a valuable source of breast cancer biomarkers.For decades, the literature on NAF and breast secretions has expanded considerably and more than 2000 proteins have beenidentified [221–228].

Cervical-vaginal fluid

Cervical-vaginal fluid (CVF) consists of water, electrolytes, low- molecular-weight organic compounds, cells and a wide range of proteins and proteolytic enzymes [229]. Up to today, about 600 proteins were identified in CVF by seven research groups [189,229–234]. CVF could play a critical role in spontaneous preterm birth by detecting biomarkers and potential molecular networks.

Pleural effusion

Pleural effusion (PE) is the excess fluid in the pleural space, which exists in lung cancer patients and also forms due to many benign ailments [235]. To date, about 1300 proteins have been detected in PE [236–240] and a number of potential biomarkers were evaluated, such as lung surfactant protein A, cystatin-C, vascular endothelial growth factor and so on.

Sputum

Sputum is a readily accessible biological fluid, and its compo- sition may change by different disease [241]. Sputum contains biomarkers of inflammation in common chronic airway dis-eases, such as asthma and COPD [242].

Exhaled breath condensate

Exhaled breath condensate (EBC) is a biological fluid consisting of aerosol droplets and water vapor, and can be obtained by freezing exhaled air under conditions of spontaneous breathing [243]. EBC composition reflects the physiological state of the lung and consequently, and, in principle, can be used toidentify and monitor several pathologies, including asthma, COPD, bronchiectasis, cystic fibrosis, acute respiratory distress syndrome, infectious and neoplastic lung diseases [243]. Approximately 220 proteins were identified in EBC, which is considerably lower than those identified in other body fluids [243–249].

Pancreatic juice

Pancreatic juice is often used for pancreatic cancer detection [250]. Only a few studies have been published on the identifi- cation of pancreatic juice proteins. Over 740 unique proteins were identified including known pancreatic cancer tumor markers and proteins over expressed in pancreatic cancers [250–253].

Clearly, apart from the applausive progresses made in the field of human body-fluid proteomics, there are significant discrepancies between different proteomic discoveries, which is mainly caused by biased sample selection and preparation, technical difference of proteomic profiling, and distinct rules toward result interpretation. Nevertheless, the accumulation of publically-available proteomics data has shown great potential in facilitating various quantitative analysis in a broad array of biomedical applications.

Computational predictions onbody-fluid proteome

In the last decade, the large-scale proteomics studies have encounter challenges in large dynamic range of the protein abundance [22] and high experimental costs (both in material and time) [254]. As alternative strategies, several computational methods for protein prediction based on statistics and machine learning have been developed and demonstrated promising performance [20–26].

Overview of learning-based prediction models

Intuitively, the discovery of proteins in different body fluids can be formulated into a classification problem, where pub- lished experimental data can be used for training a classifierto infer undiscovered instances. In fact, different learning-basedapproaches have been documented in the literature, including the following: (i) SVM-based classification: In 2008, Cui *et al.* [20] firstly proposed a computational method for prediction ifa secreted protein was likely to enter into bloodstream based on a SVM classifier. Since then, similar other works include a classifier that used physiochemical properties and amino acid composition features to infer whether a protein can be excreted into urine [22,23], and a computational model for identification of origins of detected proteins in urine; classifiers for identifying human salivary proteins and applications in head and neck cancer biomarker discovery [24,25]; (ii) ranking-based prediction: Liu *et al.* [21] presented a computational framework for blood- secretory protein prediction using manifold ranking algorithm, which ranks all the candidate proteins according to the possi- bility of being blood-

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 51, Issue 04, April : 2022

secreted. (iii) Network-based prediction: Hu *et al.* [26] has developed a novel approach that employed protein–protein interaction (PPI) network to predict human secreted proteins related to different body fluids.

In general, all these data-driven predictions require the col- lection of known body-fluid proteins for training and validation of the model [20], as well as molecular features as instance descriptors, as shown in Figure 4. Each approach introduces a

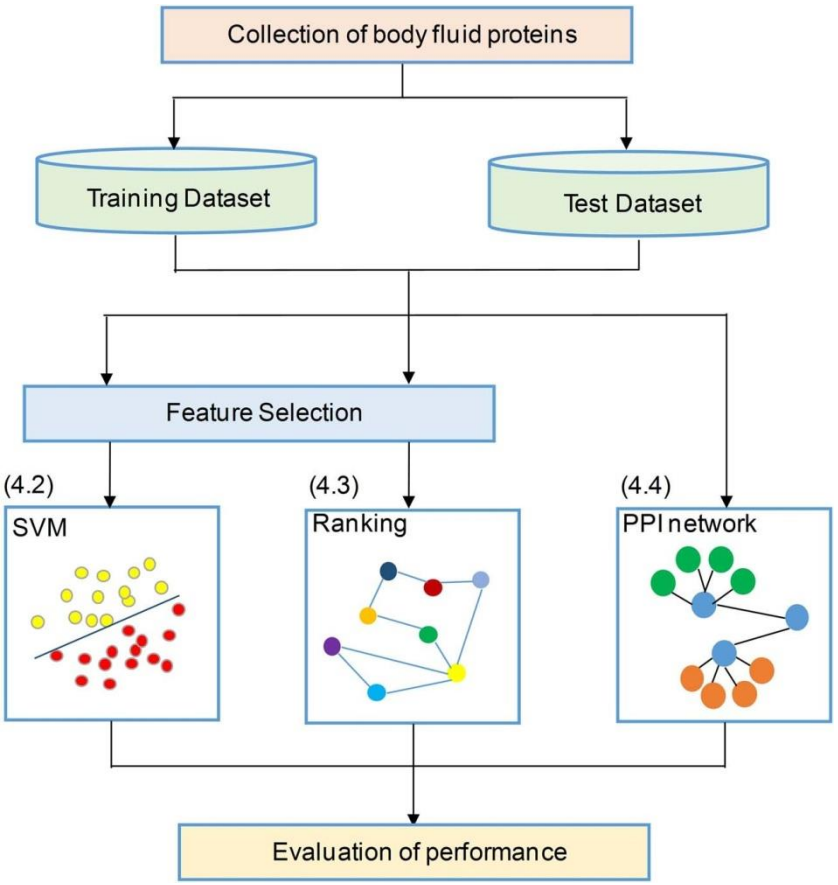

Figure 4. Summary workflow of the statistical and machine learning process for prediction of body-fluid proteomes.

unique set of analytical or computational challenges. In the nextsection, we will focus on several key issues on each topic.

SVM-based secreted protein prediction

Among all the learning-based methods, SVM has become the most popular and powerful classifier for body-fluid proteins because of its easy use and its compelling performance. Note that SVM emphasizes the idea of maximizing the margin ordegree of separation in the two-class or multiple-class clas-sification in the training process [23]. To ensure a successful application in body-fluid proteomic study, the following steps are key to train a reliable SVM classifier.

Data collection

It is essential to collect human body-fluid proteins that are experimentally detected by multiple proteomic studies from the public databases or literatures. For examples, Sys-BodyFluid database [13] contains over 10 000 proteins from 11 kinds of bodyfluids. Plasma proteome database contains information on 10 546 proteins detected in serum/plasma [112]. In addition, some body-fluid protein datasets can be collected from publishedliterature. In [20], the authors collected a total of 1620 human proteins that are annotated as secretory proteins from the Swis-sprot and SPD database [255]. Approximately 305 of those pro-teins match at least two peptides and hence are considered as secreted proteins into blood—a common practice for protein identification based on MS data. To ensure the good quality, this study only used 305 proteins that has met two criteria (both secreted and serum/plasma detected), as the positive dataset and did not include proteins that leak into the blood as a result of cell damage (e.g. cardiac myoglobin released into plasma aftera heart attack).

For binary classification through SVM, a negative dataset of non-body-fluid proteins is always required. Since such data are often not well-defined, it often requires a reliable way to generate the negative datasets, e.g. through a random selection of representative from all non-secreted related protein families, as defined in Pfam protein families [23]. Specifically, the negative data generation includes the following steps: (i) obtaining all human proteins and Pfam families from the UniProt database, (ii) mapping the known fluid proteins (positive data) to Pfam family,
(iii) excluding the families which include known fluid proteins and (iv) randomly selecting representatives from each remaining family to construct the negative data with comparable size. Cui *et al*. used a large test set containing 98 secretory proteins and 6601 non-secretory proteins of human along with other additional data to evaluate the models [20].

Feature selection

Numerous protein features have been used to train the clas- sification model that can predict human body-fluid proteins, which can be categorized into four types of property [24]: (i) sequence properties, (ii) structural properties, (iii) domains and motifs properties and (iv) physicochemical properties,

Since not all the initial features are related to a specific application, it is often useful to remove features that are noisy or irrelevant when predicting a specific group of body- fluid proteins [267]. A simple *t*-test is often used to determine the significance of a feature in terms of distinguishing two classes. Based on the derived *P*-value, a *q*-value is calculated to control the false discovery rate [268], where
is used as the threshold for removing non-contributing features. Furthermore, a classic feature-selection method known as recursive feature elimination [269] based on SVM is employed to remove features with weak classification

where TP, FP, TN and FN mean the number of true positives, false positives, true negatives and false negatives, respectively. *N* is the total number of proteins for prediction in a given test set [23]. For instance, the sensitivity relates to the classifier's ability to correctly identify the positive examples, that is, body-fluids proteins, while the specificity relates to the ability to correctly identify the negative examples, that is, non-body-fluids proteins. Note that MCC [272] is generally used as a balance measure of the quality of two-class classifications when the classes are of very different sizes. Additionally, the area under curve (AUC) is the average value of sensitivity for all possible values of specificity [273]. Last, the performance can be assessed using *k*-fold cross-validation [274] to identify the optimized model, e.g. the one achieves the highest AUC of the recall-precision curve precision. For example, in the study of blood-secreted protein [20], the SVM classifier achieved ~90% sensitivity and ~98% specificity on the test set containing 98 secretory proteins and 6601 non-secretory proteins of human with AUC as 0.96. Several additional datasets were used to further assess the performance in that study [20].

Ranking-based models

Different from SVM-based binary classifier that often requires a clean negative dataset of non-body-fluid for training, ranking- based algorithms can be employed to rank all the candidate proteins according to the possibility of being in body fluids [21]. For example, the manifold ranking algorithm [275], initially proposed to rank data points along their underlying manifold by analyzing their relationship in Euclidean space [276], has been used for to identify proteins in blood [21]. Specifically, a manifold ranking algorithm uses two datasets, a true sample set (as positive set) and an unknown sample set (as background set). According to the relevance of the unknown sample set with the true samples, the individual members of the unknown sample set can be ranked [21]. An intuitive description of this algorithm is as follows: a weighted graph is first formed, where each node represents one sample and an edge with weight score represents the similarity between the two nodes in the feature space; all the nodes then propagate their scores to the nearby points via the weighted graph; the propagation process is repeated until a global stable state is reached (which means convergence), and all the nodes except the true sample will have their own scores according to which they will be ranked.

Specifically, Liu *et al*. performed the analysis following the steps shown in Figure 5 [21]. A total of 11 394 proteins was used to

Figure 5. The workflow of the ranking-based models.

training the model, where 253 high-confidence secreted proteins were used as positive data the rest are background data. As a result, 3681 proteins were identified as human plasma proteins. Novel blood proteins were ranked based on their relevance to the core set of experimentally validated blood proteins. The higher the ranked proteins, the more likely to be body-fluid proteins. The AUC to evaluate the prediction performance is 66.3% in Liu's study [21]. Although ranking method provides an alternative solution for single class classification, it is not always advantageous over binary SVM when one can generate pseudo negative examples.

Network-based models

Considering interacted proteins may be secreted into the same body fluid to perform their functions [26], the PPI information was used in the prediction. The PPI community has been char- acterized by a wide and open distribution of proteomic data through the collection of PPI and pathway information [277]. For example, the human PPI networks were retrieved from STRING, a database dedicated to both physical and functional interactions of human proteins [26].

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 51, Issue 04, April : 2022

PPI networks can be intuitively modeled as a static graph $G = V, E$, where V is the set of nodes (proteins), and E is the set of edges (PPI) [278]. The weight of undirected edge between each pair of nodes represents the interaction confidence score in the PPI network.

This network method for body-fluid proteome prediction requires only a true sample set and the rest of the procedure is follows:

i. Define the relationship f between the protein set (the true sample set) and the body fluid. $f = 1$ means this protein can be secreted into the certain body fluid, otherwise $f = 0$;
ii. Denote the interaction confidence score w between the query proteins with the protein set in the PPI network ($w = 0$ means no interaction);
iii. Formulate the likelihood score s as the sum of the interaction confidence scores of the query protein with its interacting proteins that can be secreted into a certain body fluid j;
iv. The most likely body fluid F, where the protein is secreted should be the one with the maximum score.

Jackknife test cross-validation methods are used to examine a predictor for its effectiveness in practical application. For the jth order prediction, the accuracy φ_j obtained by the jackknife test can be formulated as

$$\varphi_j = \frac{N_j}{M} \quad j = 1, 2, \ldots, m \qquad (7)$$

where N_j represents the number of the secreted proteins, whose jth order predicted body fluid is one of the true body fluids, and M represents the total number of proteins in the PPI network.

Given a query protein, the higher the likelihood score, the more likely they are to be secreted into a certain body fluid [26]. In [26], a breakdown of the 529 human secreted proteins from 11 different types of body fluids based on the literature search were used in the training dataset according to the, and 57 blood-secreted proteins were used to test this method. The model achieved 96% accuracy based on validation.

As shown in Table 2, all those methods have shown promis- ing prediction power in the identification of body-fluid proteins. Particularly, the average performance (AUC or accuracy) of inde- pendent test across all these computational methodologies is 87.0%, while the average accuracy of SVM-based methods is 90.0%.

Discussions and future perspective

As mentioned earlier, a useful repertoire of proteomics tech- nologies is currently available for disease diagnosis and clinical-related applications. Our article reviews a large collection of different approaches involved in the proteomic data analysis of human body fluids, both experimentally and computation- ally. Current successes of the wet experimental technologies for protein characterization have been obvious. So far, there are over 15 000 different proteins discovered in major human body fluids. As discussed earlier, the largest sample dataset includes over 12 000 plasma and serum proteins; on the contrary, the smallest set is on EBC, including approximate 220 proteins. Further development of those technologies, especially with MS, will likely reduce sample requirement, increase the throughput and more effectively uncover various types of protein alterations such as post-translational modifications [279].

In the meantime, a great variety of computational tools has been developed to assist the analysis of body-fluid proteome and has shown promising performance, especially in novel protein discovery. Since the first predictor was proposed in 2008 to anno- tate the body fluids where human protein can be secreted into blood stream [20], it is anticipated that such methods will benefit the relevant experimental researches and stimulate a series of follow-up investigations into this emerging and challenging area. As reviewed previously, machine learning-based prediction through, e.g. SVM, has proved to be highly effective in terms of identifying novel secreted proteins and disease biomarkers. Sim-ilarly, both ranking and PPI network methods have made promis- ing progress in body-fluid proteomics research. Although SVM- based prediction has achieved decent performance, it still has room for improvement through possibly increasing the size and quality of the positive training set and including more relevant features. This, however, may raise concern of computation com- plexity in the ranking algorithm. In general, a few key aspects to ensure a good performance of those computational predictions include a proper data collection of high quality experimentally- detected proteins to train the model, a comprehensive collection of molecular features underlying possible mechanisms of the secretion and effective techniques for feature and model selec- tion.

Note that in general when learning larger dataset with high-dimensional features, new challenges arise. Conventional machine-learning techniques were somewhat limited in processing high-dimensional data [280]. New approach based on deep learning will likely lead to more successes in the near future because it requires very little engineering by hand and can easily take advantage of the increasingly-accumulated data available in the field [281]. As an example, the deep neural network-based model introduced in [282] can be another promising method that facilitates the understanding of body-fluid proteome and accelerate biomarker discovery in human disease.

A reliable prediction of human body-fluid proteins allows for effective targeted search for biomarker in body fluids. When

further combined with other information such as disease-associated transcriptomic data, as reviewed above, such framework provides an upstream tool that is highly useful for finding candidate biomarkers associated with human diseases or physiological phenotypes. Often a combination of several proteins can form a signature panel for non-invasive test in clinical practice for diseases diagnosis. Ongoing effort in identifying and designing such effective panels for disease detection represents other major research topics in this field, which is beyond the scope of this review. All in all, a highly innovative and integrative approach leveraging the strength of both experimental profiling and computational prediction should be further pursued along the current research line to accelerate the process toward successful clinical applications.

References

1. Wu CC, Duan JC, Liu T, *et al.* Contributions of immunoaffin-ity chromatography to deep proteome profiling of human biofluids. *J Chromatogr B Anal Technol Biomed Life Sci* 2016;**1021**:57–68.
2. Peffers MJ, Mcdermott B, Clegg PD, *et al.* Comprehensive protein profiling of synovial fluid in osteoarthritis follow- ing protein equalization. *Osteoarthr Cartil* 2015;**23**:1204–13.
3. Tanaka Y, Akiyama H, Kuroda T, *et al.* A novel approach and protocol for discovering extremely low-abundance proteins in serum. *Proteomics* 2006;**6**:4845–55.
4. Hu S, Wang JH, Meijer J, *et al.* Salivary proteomic and genomic biomarkers for primary sjögren's syndrome. *Arthritis Rheum* 2007;**56**:3588–600.
5. Tiselius A. Electrophoresis of serum globulin: electrophoretic analysis of normal and immune sera. *Biochem J* 1937;**31**:313–7.
6. Margolis J, Kenrick KG. Two-dimensional resolution of plasma proteins by combination of polyacrylamide disc and gradient gel electrophoresis. *Nature* 1969;**221**:1056–7.
7. Freeman T, Smith J. Human serum protein fractionation bygel filtration. *Biochem J* 1970;**118**:869–73.
8. Rabilloud T, Chevallet M, Luche S, *et al.* Two-dimensional gel electrophoresis in proteomics: past, present and future. *J Proteome* 2010;**73**:2064–77.
9. Thomson JJ. Rays of positive electricity and their applica- tion to chemical analyses. *Nature* 1914;**92**:549–50.
10. Burlingame AL, Boyd RK, Gaskell SJ. Mass spectrometry. *Anal Chem* 1976;**60**:268–303.
11. Roepstorff P. Mass spectrometry in protein studies from genome to function. *Curr Opin Biotechnol* 1997;**8**:6–13.
12. Omenn GS. The human proteome organization plasma pro- teome project pilot phase: reference specimens, technology platform comparisons, and standardized data submissions and analyses. *Proteomics* 2004;**4**:1235–40.
13. Li SJ, Peng M, Li H, *et al.* Sys-BodyFluid: a systematical database for human body fluid proteome research. *Nucleic Acids Res* 2009;**37**:D907–12.
14. Ogata Y, Charlesworth MC, Muddiman DC. Evaluation of protein depletion methods for the analysis of total-, phospho- and glycoproteins in lumbar cerebrospinal fluid. *J Proteome Res* 2005;**4**:837–45.
15. Cho CK, Shan SJ, Winsor EJ, *et al.* Proteomics analysis of human amniotic fluid. *Mol Cell Proteomics* 2007;**6**:1406–15.
16. Zeng Z, Hincapie M, Pitteri SJ, *et al.* A proteomics platform combining depletion, multi-lectin affinity chromatography (M-LAC), and isoelectric focusing to study the breast cancer proteome. *Anal Chem* 2011;**83**:4845–54.
17. Marimuthu A, O'Meally RN, Chaerkady R, *et al.* A compre- hensive map of the human urinary proteome. *J Proteome Res* 2011;**10**:2734–43.
18. Hogan MC, Johnson KL, Zenka RM, *et al.* Subfractiona- tion, characterization, and in-depth proteomic analysis of glomerular membrane vesicles in human urine. *Kidney Int* 2014;**85**:1225–37.