Chapter 19
Urinary Protein Biomarker Database: A Useful Tool for Biomarker Discovery

Chen Shao

Abstract An open-access biomarker database offers a convenient tool for researchers to acquire existing knowledge about proteins and diseases by simply querying its Web site. Biologists can use the biomarker database to assess the confidence and disease specificity of their own research results by cross-study comparison, and bioinformaticians can use it to discover new relationships between diseases and proteins by reanalyzing data via new strategies. This chapter introduces the urinary protein biomarker database, a manually curated database that aim to collect all studies of urinary protein biomarkers from published literature. In the current stage, this database includes very few disease-specific biomarker candidates that have been reported by multiple studies, reflecting current status in the field of urinary biomarker discovery. We believe that this situation will be improved with the development of technologies and accumulation of data, and a more complete and precise biomarker database will play more important role in future studies.

Keywords Database · Urinary biomarkers

19.1 Rationale for a Urinary Biomarker Database

Urine is an ideal source of biomarkers. In comparison to plasma, urine has some unique advantages that make it a suitable source for both physiological research and disease biomarker discovery. Firstly, urine can be collected continuously and noninvasively. Secondly, the urinary proteome directly reflects the condition of the urinary system. Thirdly, since the urinary proteome contains a number of plasma proteins, some changes of the plasma proteome can also be found in urine.

C. Shao (✉)
National Key Laboratory of Medical Molecular Biology, Department of Pathophysiology,
Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences,
5 Dong Dan San Tiao, Beijing, China
e-mail: scshaochen@126.com

Y. Gao (ed.), *Urine Proteomics in Kidney Disease Biomarker Discovery*,
Advances in Experimental Medicine and Biology 845,
DOI 10.1007/978-94-017-9523-4_19

Therefore, urine is not only a good source for the study of urological diseases, but can also reflect the status of the entire body [4]. There have been a considerable amount of urinary biomarker studies based on different experimental platforms for a variety of diseases, such as bladder [2] and prostate cancer [12], renal disorders [3], and cardiovascular diseases [9].

The development of proteomic technologies offers possibility of identification and quantification of all peptide/proteins in biological samples simultaneously. Currently, a single proteomic assay can identify dozens or hundreds of peptide/proteins that express differently between normal and disease conditions. However, due to limitation in patient samples and experimental resource, usually only a small proportion of these ‘differentially expressed proteins’ were selected for consequent validation experiments, while the rest of them were absent in any further analysis or even not reported in the published articles. The relative low throughput in the validation phase decreases the efficiency of whole biomarker discovery workflow and results in a waste of the abundant information achieved in the discovery phase. Researchers need to pay attention on this ‘ignored information’ if they want to make fully use of the high-throughput proteomic technologies. Collecting this information into a database is the first step for in-depth data analysis.

On the other hand, a considerable amount of the ‘differentially expressed proteins’ identified in a proteomic analysis does not reflect the real difference between normal and disease conditions, but are caused by some relatively random factors, such as experimental errors and variation among urine samples. Enlarging the analyzed sample size is an idea solution to eliminate these influence factors, but it costs too much experimental resources since urinary proteome has been reported to vary even among healthy individuals [8] and is affected by a number of physiological factors [7, 10]. This chapter suggests that cross-study comparison may be a much easier way to partially solve this problem. The concept is that if the same trend of abundance change is observed for a protein in more than one distinct study, the chance that this observation is caused by random factors would be significantly decreased. The comparison can be highly simplified by collecting results of existing urinary biomarker studies into an open-access database.

Additionally, building a biomarker database can facilitate the assessment of disease specificity for biomarker candidates. Only biomarkers with rigorous disease specificity can be used to distinguish diseases with similar signs and symptoms and consequently guide the choice of drugs and treatments. By querying a biomarker database, biomarker candidates that are related to multiple diseases can be easily picked out, so that researchers can better focus on biomarker candidates with high disease specificity and save their efforts from those that have low potential to distinguish different diseases.

In summary, a urinary biomarker database offers a convenient tool for cross-study comparison. It can help biologists to assess the confidence and disease specificity of their own research results and allows bioinformaticians to discover new relationships between diseases and proteins by reanalyzing existing data via new strategies. A list of existing databases of normal urinary proteome or urinary biomarkers is shown in Table 19.1.

Table 19.1 Urinary proteome databases

	URL	Content
HKUPP database	http://www.hkupp.org/	Proteome of normal kidney and urine
Urinary exosomes protein database	http://dir.nhlbi.nih.gov/papers/lkem/exosome/index.htm	304 proteins identified from exosomes in normal human urine [11]
MAPU urine dataset	http://www.mapuproteome.com	1,543 normal human urinary proteins identified by Adachi et al. [1]
Clinical urine proteomic database	http://alexkentsis.net/urineproteomics/	Urinary proteins that were annotated to be associated with diseases by machine learning and text mining methods [6]
Urinary peptide biomarker database	http://mosaiques-diagnostics.de/diapatpcms/mosaiquescms/front_content.php?idcat=257	CE-MS results of naturally occurring human urinary peptides in different pathophysiological conditions [14]
Urinary protein biomarker database	http://122.70.220.102/biomarker/	A literature-curated database of protein biomarker or biomarker candidates in human and animal urine [13]

19.2 Establishment of the Urinary Protein Biomarker Database

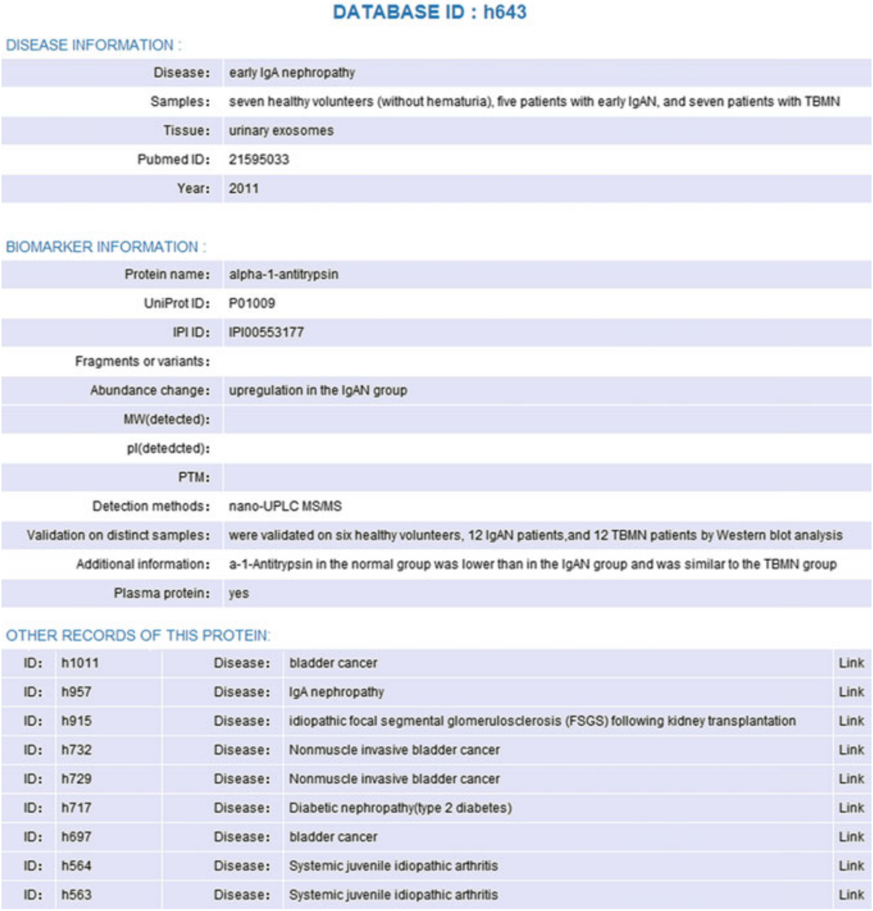
In 2011, we published the Urinary Protein Biomarker Database (UPB database), a manually curated database compiling results of urinary protein biomarker studies from published literature of proteomic studies as well as small-scale experiments such as ELISA and Western blot. Manually, curation ensures the minimum mistakes appear in the process of database establishment.

All of the protein/peptides that were reported to have abundance change under disease conditions are considered as biomarker candidates and collected in this database. No extra filtration of biomarker confidence is used since the aim of building this database was to preserve the original result of literature and make the database as comprehensive as possible. Users can use information such as detection method and sample size displayed in the database to help assess data confidence themselves. Table 19.2 shows data statistics of this database.

Particularly, a very small proportion of records in this database are 'negative records,' in which changes of protein abundance were reported as not statistically significant under disease conditions. These records are included because the same proteins were identified as biomarker candidates for the same diseases by other studies. Including negative data in the biomarker database is important, since it may help researchers to assess data confidence better by analyzing conflict results for the same biomarker candidate.

Table 19.2 Data statistics of the UPB database after an update in July 2013

	Human dataset	Animal dataset			
		Rat	Mouse	Others	Total
Articles	348	49	19	8	76
Records	993	299	62	18	379
Diseases	119	29	13	7	49
Biomarkers	819	253	62	18	333
Proteins	458	161	62	16	239



DATABASE ID : h643

DISEASE INFORMATION :

Disease:	early IgA nephropathy
Samples:	seven healthy volunteers (without hematuria), five patients with early IgAN, and seven patients with TBMN
Tissue:	urinary exosomes
Pubmed ID:	21595033
Year:	2011

BIOMARKER INFORMATION :

Protein name:	alpha-1-antitrypsin
UniProt ID:	P01009
IPI ID:	IPI00553177
Fragments or variants:	
Abundance change:	upregulation in the IgAN group
MW(detected):	
pI(detedcted):	
PTM:	
Detection methods:	nano-UPLC MS/MS
Validation on distinct samples:	were validated on six healthy volunteers, 12 IgAN patients,and 12 TBMN patients by Western blot analysis
Additional information:	a-1-Antitrypsin in the normal group was lower than in the IgAN group and was similar to the TBMN group
Plasma protein:	yes

OTHER RECORDS OF THIS PROTEIN:

ID:	h1011	Disease:	bladder cancer	Link
ID:	h957	Disease:	IgA nephropathy	Link
ID:	h915	Disease:	idiopathic focal segmental glomerulosclerosis (FSGS) following kidney transplantation	Link
ID:	h732	Disease:	Nonmuscle invasive bladder cancer	Link
ID:	h729	Disease:	Nonmuscle invasive bladder cancer	Link
ID:	h717	Disease:	Diabetic nephropathy(type 2 diabetes)	Link
ID:	h697	Disease:	bladder cancer	Link
ID:	h564	Disease:	Systemic juvenile idiopathic arthritis	Link
ID:	h563	Disease:	Systemic juvenile idiopathic arthritis	Link

Fig. 19.1 A Webpage displaying records in the UPB database

For each biomarker candidate in this database, the following information was collected (if available): definition and sample size of the disease and control groups, experimental procedures and instrument types, protein information such as fold change, fragment, variant and post-translational modification (PTM), experimental molecular weight and pI. In particular, proteins were queried in the plasma proteome list [5] to infer its origin (Fig. 19.1).

This database is open access to nonprofit researchers in the community. The Web site now allows users to browse and download the complete database. Users are strongly welcome to submit their own data to this database.

19.3 Analyzing Data in the UPB Database

Analyzing data in the UPB database reveals some important aspects in urinary biomarker discovery and database construction. In this database, biomarker candidates identified by different proteomic methods overlapped poorly with each other. Approximately, half of the records were identified only by proteomic methods and reported in only one study. Besides false positive results generated from technic errors or limited sample size, this phenomenon might be caused by several other reasons. The first one is that authors do not always report the whole protein list that is identified to have significant abundance change in disease conditions. Sometimes, they only report proteins that are thought to have higher potential to act as biomarkers or those have not been reported by other studies. Lack of original experimental data is an instinct problem in the construction of a literature-based database. Secondly, since different proteomic strategies vary a lot in the methods of sample preparation, separation, identification, and quantification, proteins or peptides with particular property (i.e., hydrophobicity) may be preferred in one strategy but cannot be identified in another one. Thirdly, although some studies are for the same disease, their results may not be comparable due to different criteria for the selection of samples to disease and control groups. So, detailed description of patients in each group should be included in biomarker database. The poor overlap rate among different proteomic studies for the same disease in this database makes researchers difficult to perform some in-depth bioinformatical or statistical methods, such as meta-analysis.

Studies based on animal models also overlapped poorly with those based on human samples. However, considering that the inter-organism overlap rate is not lower than the overlap rate among different proteomic studies of human samples, no clear conclusion can be made to the question that how well these animal models mimic real human diseases.

In the current stage, the UPB database includes very few disease-specific biomarker candidates that have been reported by multiple studies. Whereas a large proportion of biomarker candidates in this database are considered to have relatively low potential for clinical usage due to lack of disease specificity or further validation to prove their confidence. This reflects current status in the field of urinary biomarker discovery. We believe that this situation will be improved with the development of technologies and accumulation of data, and a more complete and precise biomarker database will play more important role in future studies.

19.4 Potential Usage of the Biomarker Database

A biomarker database facilitates researchers acquire existing knowledge about proteins and diseases. By simply querying the biomarker database, researchers can find answers to some important questions to help them assess biomarker candidates they identified. For example, they may want to know whether these proteins have been reported to be biomarkers or biomarker candidates for the same disease by other groups, and if so, whether the fold changes they observed agree or conflict with previous studies. By querying a biomarker database of animal models, they can also easily find out that whether orthologous of these proteins have been studied by animal models but still lack of validation in human samples. In addition, it is also important to know whether these proteins have been previously identified as biomarkers or biomarker candidates for other diseases, which would indicate their poor disease specificity.

A biomarker database is also a useful bioinformatics tool to study the pathophysiology of diseases with the hypothesis that diseases sharing biomarkers may share the same injury sites or pathophysiological processes. For a ‘new’ disease where the pathogenesis or injury sites are not clear (for example, a new drug with unknown toxicity), if the fold changes of urinary proteins caused by this disease are known, researchers can query the protein list in the database to link the disease to other diseases that cause similar fold changes in these proteins. The injury site, pathophysiological process, and severity of the ‘new’ disease can then be inferred by its relationship to the other diseases.

Disease–protein network is plotted to display relationships between diseases and proteins deposited in a biomarker database. Moreover, a disease–disease network can be plotted by linking diseases sharing the same proteins as biomarkers or biomarker candidates, while a protein–protein network can be plotted by linking proteins that were found to be related to the same disease. Researchers can possibly dig out novel relationships among those diseases and proteins by analyzing structures or topological characters of these networks. In the previously published article [13], we built a weighted disease–disease network in which the weight of each link was defined as the number of biomarker candidates shared by two diseases. This network was then clustered into seven densely connected subnetworks solely based on its topological structure. Most diseases in the same subnetwork are known to share similar injury sites or pathophysiological processes, indicating that the result of clustering was very rational biologically. This example suggests that network analysis of the biomarker database offers a new angle of view for the similarity among diseases, and therefore, it may be helpful to study the pathophysiology of diseases.

19.5 Future Work

Existing biomarker database only includes basic information of diseases and proteins. Extending more information about proteins and diseases to the current database can make the database more convenient for users. As listed in Fig. 19.2, the UPB database can be improved in several aspects.

19.5.1 Disease Information

Besides urinary protein/peptide biomarkers, nonprotein and non-urinary biomarkers are also essential to disease diagnosis. More biomarker data can be collected from datasets of genomic and metabolic studies as well as studies of other kinds of biological samples. Some open-access disease databases, such as the Online Mendelian Inheritance in Man (OMIM) database and the Kyoto Encyclopedia of Genes and Genomes (KEGG) disease database, also provide useful information, such as biomarkers, pathways, drugs, and drug targets.

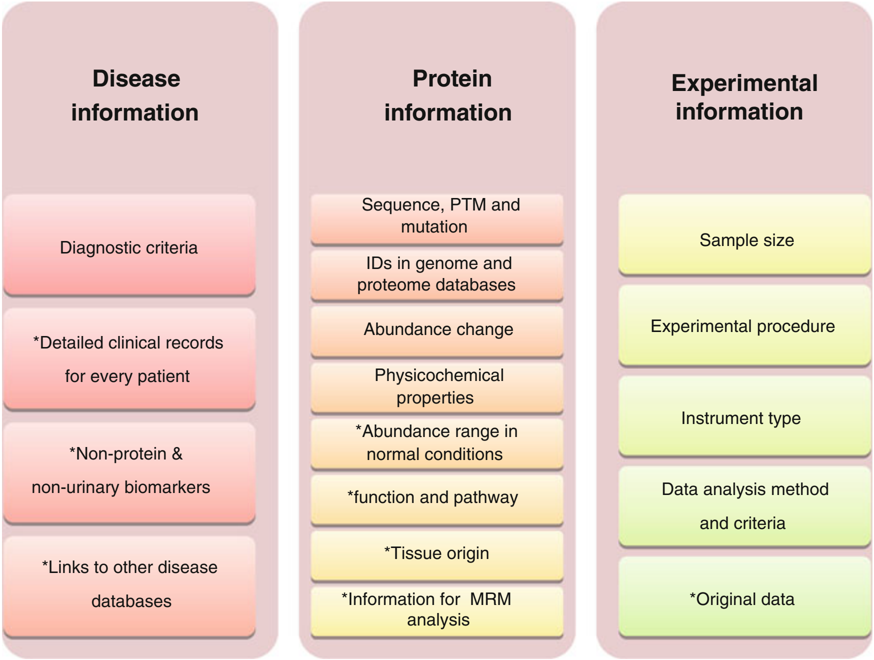

Disease information
Diagnostic criteria
*Detailed clinical records for every patient
*Non-protein & non-urinary biomarkers
*Links to other disease databases
Protein information
Sequence, PTM and mutation
IDs in genome and proteome databases
Abundance change
Physicochemical properties
*Abundance range in normal conditions
*function and pathway
*Tissue origin
*Information for MRM analysis
Experimental information
Sample size
Experimental procedure
Instrument type
Data analysis method and criteria
*Original data

Fig. 19.2 Design for the future urinary biomarker database. Contents marked with *asterisks* have not been included in the current UPB database

19.5.2 Protein Information

Variation of urinary proteome appears in different physiological conditions and among different individuals, affecting the result of urinary biomarker study. By including abundance ranges of proteins in normal conditions, the database can offer a good reference to help researchers determining changes of protein abundance that are caused by diseases. The database can also be improved by adding more protein functional information. The functional information could be acquired from open-access bioinformatics resources, e.g., the Gene Ontology database, KEGG pathway database, protein–protein interaction databases, and the Web site of Protein Atlas Project for protein tissue expression profiles.

In addition, targeted proteomic approaches such as multiple reaction monitoring (MRM) can validate biomarker candidates in the high throughput and high-accuracy manner. To facilitate researchers whom may want to validate biomarker candidates in the database via MRM assay, list of proteotypic peptides and their MS/MS spectra for each protein can be included in the database.

19.5.3 Experimental Information

Achieving original experimental data allows database builders or bioinformaticians to reanalyze results from different studies in the same criteria, so that they can get much better control of data confidence and achieve more precise result in the cross-study comparison. However, it is hard to acquire original data in a literature-based database. Fortunately, more and more researchers would like to upload the original files of proteomic experiments to an online repository in the recent years, this situation will be improved in the near future.

References

1. Adachi J, Kumar C, Zhang Y, Olsen JV, Mann M (2006) The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins. Genome Biol 7:R80
2. Bhatt J, Cowan N, Protheroe A, Crew J (2012) Recent advances in urinary bladder cancer detection. Expert Rev Anticancer Ther 12:929–939
3. Bonomini M, Sirolli V, Magni F, Urbani A (2012) Proteomics and nephrology. J Nephrol 25:865–871
4. Decramer S, Gonzalez de Peredo A, Breuil B, Mischak H, Monsarrat B, Bascands JL, Schanstra JP (2008) Urine in clinical proteomics. Mol Cell Proteomics 7:1850–1862
5. Farrah T, Deutsch EW, Omenn GS, Campbell DS, Sun Z, Bletz JA, Mallick P, Katz JE, Malmstrom J, Ossola R, Watts JD, Lin B, Zhang H, Moritz RL, Aebersold R (2011) A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. Mol Cell Proteomics 10(9):M110–006353

6. Kentsis A, Monigatti F, Dorff K, Campagne F, Bachur R, Steen H (2009) Urine proteomics for profiling of human disease using high accuracy mass spectrometry. Proteomics Clin Appl 3:1052–1061
7. Khan A, Packer NH (2006) Simple urinary sample preparation for proteomic analysis. J Proteome Res 5:2824–2838
8. Liu X, Shao C, Wei L, Duan J, Wu S, Li X, Li M, Sun W (2012) An individual urinary proteome analysis in normal human beings to define the minimal sample number to represent the normal urinary proteome. Proteome Sci 10:70
9. Napoli C, Zullo A, Picascia A, Infante T, Mancini, FP (2013) Recent advances in proteomic technologies applied to cardiovascular disease. J Cell Biochem 114:7–20
10. Oh J, Pyo JH, Jo EH, Hwang SI, Kang SC, Jung JH, Park EK, Kim SY, Choi JY, Lim J (2004) Establishment of a near-standard two-dimensional human urine proteomic map. Proteomics 4:3485–3497
11. Pisitkun T, Shen RF, Knepper MA (2004) Identification and proteomic profiling of exosomes in human urine. Proc Natl Acad Sci USA 101:13368–13373
12. Sardana G, Diamandis EP (2012) Biomarkers for the diagnosis of new and recurrent prostate cancer. Biomark Med 6:587–596
13. Shao C, Li M, Li X, Wei L, Zhu L, Yang F, Jia L, Mu Y, Wang J, Guo Z, Zhang D, Yin J, Wang Z, Sun W, Zhang Z, Gao Y (2011) A tool for biomarker discovery in the urinary proteome: a manually curated human and animal urine protein biomarker database. Mol Cell Proteomics 10:M111–010975
14. Siwy J, Mullen W, Golovko I, Franke J, Zurbig P (2011) Human urinary peptide database for multiple disease biomarker discovery. Proteomics Clin Appl 5:367–374