RESEARCH ARTICLE

BiomarkerDigger: A versatile disease proteome database and analysis platform for the identification of plasma cancer biomarkers

Seul-Ki Jeong[1], Min-Seok Kwon[1], Eun-Young Lee[1], Hyoung-Joo Lee[1], Sang Yun Cho[1], Hoguen Kim[2], Jong Shin Yoo[3], Gilbert S. Omenn[4], Ruedi Aebersold[5,6], Sam Hanash[7] and Young-Ki Paik[1]

[1] Department of Biochemistry, Yonsei Proteome Research Center and Biomedical Proteome Research Center, Sudaemoon-ku, Seoul, Korea
[2] Department of Pathology, Yonsei University College of Medicine, Yonsei University, Sudaemoon-ku, Seoul, Korea
[3] Korea Basic Science Institute, Daejeon, Korea
[4] Departments of Internal Medicine and Human Genetics and Center for Computational Medicine and Biology, University of Michigan, Ann Arbor, MI, USA
[5] Institute for Systems Biology, Seattle, WA, USA
[6] Institute of Molecular Systems Biology, ETH Zurich; and Faculty of Sciences, University of Zurich, Switzerland
[7] Fred Hutchinson Cancer Research Center, Seattle, WA, USA

We have developed a proteome database (DB), BiomarkerDigger (http://biomarkerdigger.org) that automates data analysis, searching, and metadata-gathering function. The metadata-gathering function searches proteome DBs for protein–protein interaction, Gene Ontology, protein domain, Online Mendelian Inheritance in Man, and tissue expression profile information and integrates it into protein data sets that are accessed through a search function in BiomarkerDigger. This DB also facilitates cross-proteome comparisons by classifying proteins based on their annotation. BiomarkerDigger highlights relationships between a given protein in a proteomic data set and any known biomarkers or biomarker candidates. The newly developed BiomarkerDigger system is useful for multi-level synthesis, comparison, and analyses of data sets obtained from currently available web sources. We demonstrate the application of this resource to the identification of a serological biomarker for hepatocellular carcinoma by comparison of plasma and tissue proteomic data sets from healthy volunteers and cancer patients.

Received: July 13, 2008
Revised: March 16, 2009
Accepted: April 20, 2009

Keywords:
BiomarkerDigger / Database / HUPO / Plasma proteome

1 Introduction

The rapid growth and expansion of the proteomics field has resulted in exponential growth in protein data sets that are accessible in various formats. Clinical proteomics studies typically involve collection and processing of clinical samples, protein/peptide separations, and identification of clinically relevant proteins using MS methods [1]. There are no standard methods for the collection, comparison, or presentation of proteome data, which makes cross-study comparisons difficult. The analysis of data would greatly benefit from the availability of software capable of coordinated cross-study analysis.

Numerous systematic databases (DBs) have been developed for depositing, retrieving, and mining data sets. Freely accessible DBs include PRIDE [2], PeptideAtlas [3], UniPep [4], the Global Proteome Machine (GPM) [5], PhosphoPep [6], Proteome Commons and its Tranche file-sharing system (www.tranche.proteomecommons.org), Human Protein Reference Database (HPRD) [7], the Human Protein

Correspondence: Professor Young-Ki Paik, Department of Biochemistry, College of Life Science and Biotechnology, Yonsei Proteome Research Center, Yonsei University, 134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Korea
E-mail: paikyk@yonsei.ac.kr
Fax: +82-2-393-6589

Abbreviations: DBs, databases; **GO,** Gene Ontology; **HCC,** hepatocellular carcinoma; **HPPP,** HUPO Plasma Proteome Project; **HPRD,** Human Protein Reference Database; **IPI,** International Protein Index; **KO,** Kyoto Encyclopedia of Genes and Genomes Ortholog; **MRM,** multiple reaction monitoring; **PPI,** protein–protein interaction; **XML,** eXtensible Markup Language

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 1. List of databases for collection of protein information

Database name and URL	Used information	Release number/date
GO http://geneontology.org	GO terms and their hierarchies	15-Jan-2008
HPRD http://www.hprd.org	Expressed tissues and references PPI	Release 7 (09-Jan-2008)
IPI http://www.ac.uk/IPI	Cross-references information	Release 3.39 (07-Feb-2008)
KEGG http://www.genome.jp/kegg	Cross-references information Pathway	Release 45 (01-Jan-2008)
OMIM^{TM} http://www.ncbi.nlm.nih.gov/omim/	Human genes and genetic disorders	07-Mar-2008[a)]
UniProtKB-Swiss-Prot/TrEmbL http://www.uniprot.org	Cross-reference information Sequence, Mw Domain Description Expressed tissues and references GO PPI	Release 12.8 (05-Feb-2008)

a) OMIM^{TM} McKusick, VA. 1998 [54].

Atlas [8], MitoP2 [9], PeroxisomeDB [10], LMPD [11], and ChromDB [12]. PRIDE, PeptideAtlas, Tranch, and GPM are well-known proteome repositories for the storage of proteins and peptide identifications from proteomics experiments (see Mead *et al.* [13] for details). The remaining DBs offer similar but more specialized functions. PhosphoPep is a phosphoproteome resource, which is a part of the PeptideAtlas project. PhosphoPep provides a search function to detect the sites of phosphorylation on individual proteins, utilities for the use of the phosphopeptide data for targeted proteomics experiments such as multiple reaction monitoring (MRM) and is searchable by spectral matching. MitoP2 and PeroxisomeDB are designed to provide comprehensive proteome expression information specific for the mitochondria and the peroxisomes, respectively. LMPD and ChromDB manage information related to lipid-associated proteins and chromatin-associated proteins, respectively. The UniPep DB provides useful information on both N-linked glycosides that have been experimentally verified and predicted N-glycopeptides in the human proteome. HPRD is an object DB that integrates a wealth of information relevant to the function of human proteins in healthy and disease state. The Human Protein Atlas contains expression and localization data for proteins in a large variety of healthy human tissues, cancer cells, and cell lines including displayable immunohistochemistry images. Finally, Michigan Molecular Interactions [14] offers a data integration function through which various DBs can be merged based on accession numbers and related cross-reference identifier information, and visualized with Cytoscape (http://mimi.ncibi.org).

To maximize the potential values of proteomic data sets compiled from different experiments, 'search' and 'cross-comparison' are essential parts of any proteome DB [13]. The above-cited resources are not designed to support analysis functions relevant to biomarkers including cross-study comparison of proteins, prediction of molecular function and pathway analysis, classification of proteins according to the functions and expression patterns. BiomarkerDigger presented here is designed to encompass both the DB and web-based tools for users to deposit their data sets and to search stored data sets (DB function). BiomarkerDigger functions as a web-tool by supporting a comparative cross-analysis of proteome to derive disease-related biological information. Here we demonstrate the utility of this resource to the identification of a serological biomarker for hepatocellular carcinoma (HCC).

2 Materials and methods

2.1 Collection of information

To collect diverse types of information from various DBs, we created java-embedded 'idMapper' and 'dataExtractor' modules. 'idMapper' was created to identify UniprotKB accession codes of the deposited proteins using cross-reference information obtained from International Protein Index (IPI) [15], UniProtKB-Swiss-Prot/TrEmBL [16], and Kyoto Encyclopedia of Genes and Genomes Ortholog (KO) [17]. This 'idMapper' is useful for the comparative analyses of different proteome data sets because it can convert the protein lists generated by a given experiment or study (*e.g.* plasma) to the list of UniProtKB accession codes. The converting function was enhanced by linking idMapper to the Protein Identifier Cross-Referencing service [18]. The 'dataExtractor' facilitates the extraction of protein information collected from different DBs using 'idMapper' and integrates it into BiomarkerDigger. The information that is collected includes amino acid sequence, molecular weight,

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

domain structure, functional descriptions, tissue expression pattern, Gene Ontology (GO) annotations, interacting partners, and association with genetic disorders. Pathway information was extracted from KO, which has a four-tiered hierarchical structure. Table 1 lists the DBs from which information was collected. This information was stored as meta-data in BiomarkerDigger.

2.2 Compilation of cancer-related protein information

Information about known protein biomarker candidates was compiled from a recent comprehensive review of human biomarkers [19]. Oncogene product information was obtained from the Cancer Gene Census provided by the Sanger Center [20]. The list of cancer-related proteins covers 1261 proteins that have been reported to be differentially expressed by human cancers, and provides valuable information about each protein including utility as a clinical marker, concentration in normal plasma, and availability of antibodies [19]. The Cancer Gene Census contains a list of those genes that cause cancer when mutated. The proteins in this list are categorized by mutation type, tumor shape, and cellular origin (*e.g.* somatic or germline cell) [20]. Information regarding cancer, cytogenetics and clinical entities in oncology, and cancer-prone disease was obtained from both 'Atlas of Genetics and Cytogenetics in Oncology and Haematology' [21] and CancerGenes [22]. Information as to candidate biomarker that is an organ-specific was from OncoDB.HCC [23] and Prostate Gene Database [24]. OncoDB.HCC [23] contains information as to those genes and their loci that showed a significant expression change in HCC specimen (tissue/patients plasma), whereas Prostate Gene Database [24] provides the gene list that is related to prostate disease and cancer. From this compiled information, a list of 2006 cancer-related proteins was obtained.

2.3 Peptide data processing

To gain useful information from tryptic peptides, we installed 'peptideCutter' and 'peptideMapper'; the former provides peptide sequence and expected mono-isotopic molecular weights of trypsin-digested proteins and the latter generates a list and count of the proteins that contain the peptide products. These two modules were used to prepare a list of peptides that appear in other proteins and their frequency of occurrence (Fig. 1).

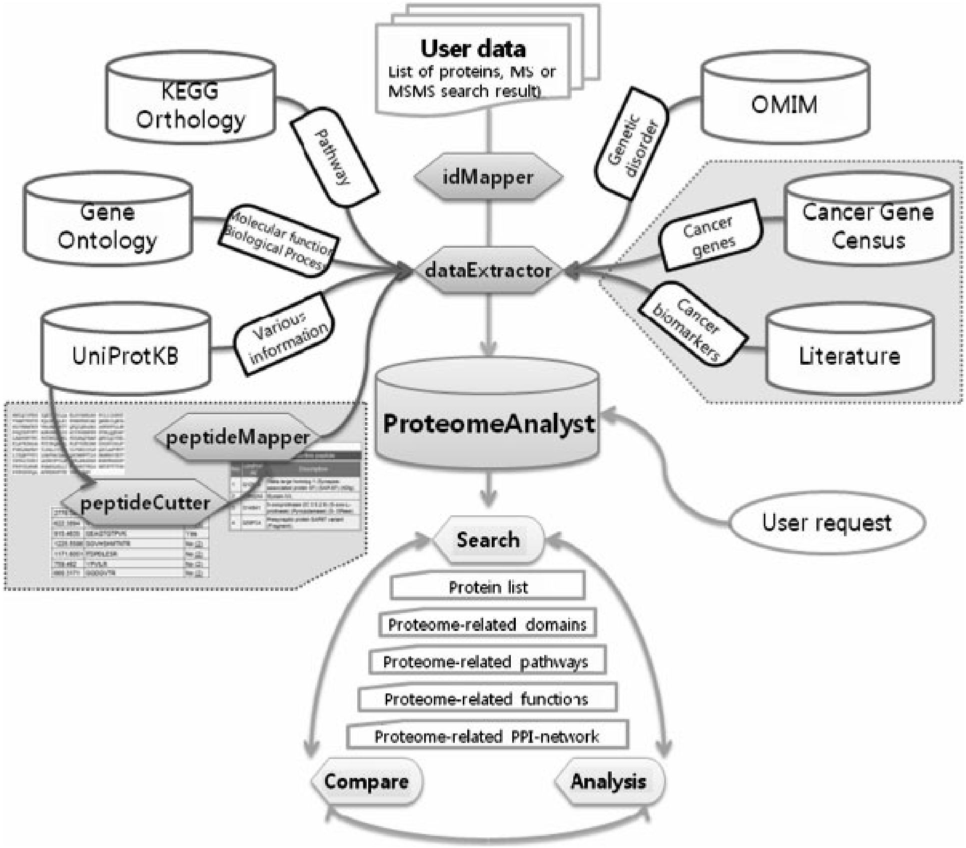


Figure 1. Outline of Biomarker-Digger structure.

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

2.4 Preparation of HCC clinical specimen samples and peptide analysis

Human plasma and liver tissues from healthy individuals (control) and HCC patients were obtained with informed consent from College of Medicine, Yonsei University (Seoul, Korea). Access to human tissues was governed by the guidelines of the Yonsei Medical Center Institutional Review Board. Trypsin-digested sample sets (plasma and tissue) were separated by 2-D LC (PF2D) [25]. Fractionated peptides were analyzed using a nano-LC-MS/MS system and identified using SEQUEST (Bio Works Software Version 3.2, Thermo Scientific) [25]. To obtain a more accurate protein list, MASCOT (MASCOT release 2.1, Matrix Science) was used to filter out MS/MS spectra those proteins that were commonly identified through these two search engines were selected. The search parameter used for MASCOT search engine was the same as that in SEQUEST with the significant threshold $p<0.05$.

2.5 Web-interface

The web-interface implemented using Apache (2.2.6, www.apache.org) and PHP (5.2.4, www.php.net) enables user-deposited information to be processed and provides query results from MySQL DB in an HTML format (Fig. 2). The pie chart was generated using Google chart (http://code.google.com/apis/chart). The 'network' was viewed using network viewer operated by a JAVA applet. This DB was constructed with an entity-relationship model (Fig. 3) using MySQL (5.0.48, www.mysql.com), which is linked to Apache and PHP, enabling query results to be received and reported through the Web.

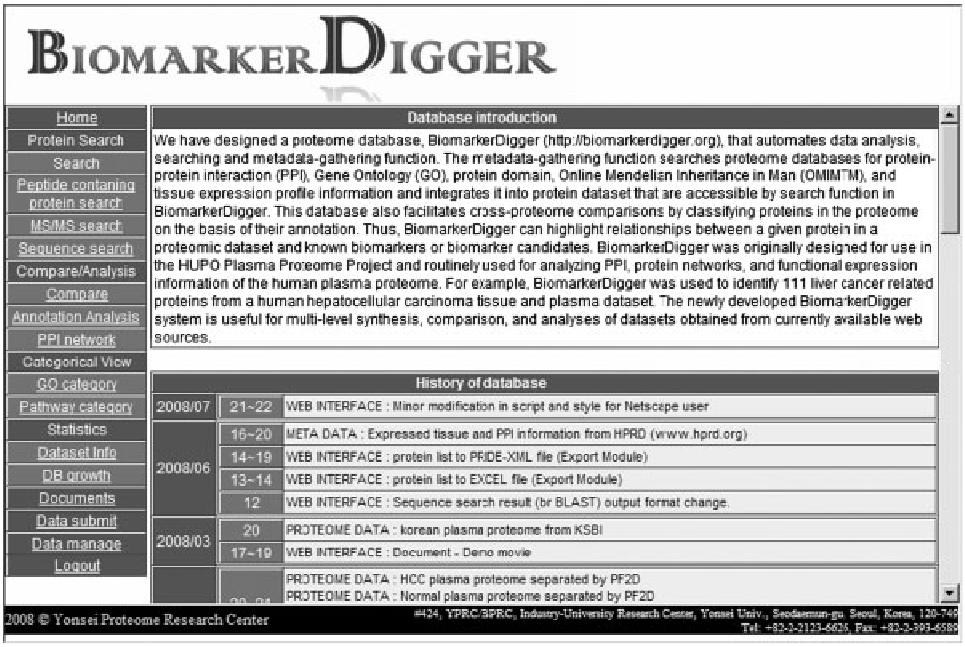
2.6 Network analysis

For the network analysis of protein–protein interaction (PPI), we used mostly the available information as to well-known interaction partner proteins that had been obtained from both HPRD [7] and UniProtKB [16]. The network is composed of vertices and edges which function as connection between two different vertexes. In the PPI network, vertices represent protein, whereas edges stand for interactions between two proteins. The topology of the network was measured by both the degree centrality and closeness centrality in which the former represents the number of interacting partners with one protein while the latter indicates the number of steps that should be taken from one protein to the rest of proteins within a network as obtained by the following

$$Cc(v) = \frac{\sum_{t \in V,\, t \neq v} D(v,t)}{N} \tag{1}$$

N is the number of proteins in network, V the all proteins in network, t, v the protein in network, $D(v,t)$ the minimum number of proteins to connect protein 'v' and 't' within a network.

Modularity is obtained by using the edge-betweeness centrality [26].



Database introduction

We have designed a proteome database, BiomarkerDigger (http://biomarkerdigger.org), that automates data analysis, searching and metadata-gathering function. The metadata-gathering function searches proteome databases for protein-protein interaction (PPI), Gene Ontology (GO), protein domain, Online Mendelian Inheritance in Man (OMIM™), and tissue expression profile information and integrates it into protein dataset that are accessible by search function in BiomarkerDigger. This database also facilitates cross-proteome comparisons by classifying proteins in the proteome on the basis of their annotation. Thus, BiomarkerDigger can highlight relationships between a given protein in a proteomic dataset and known biomarkers or biomarker candidates. BiomarkerDigger was originally designed for use in the HUPO Plasma Proteome Project and routinely used for analyzing PPI, protein networks, and functional expression information of the human plasma proteome. For example, BiomarkerDigger was used to identify 111 liver cancer related proteins from a human hepatocellular carcinoma tissue and plasma dataset. The newly developed BiomarkerDigger system is useful for multi-level synthesis, comparison, and analyses of datasets obtained from currently available web sources.

History of database

2008/07	21~22	WEB INTERFACE : Minor modification in script and style for Netscape user
2008/06	16~20	META DATA : Expressed tissue and PPI information from HPRD (www.hprd.org)
	14~19	WEB INTERFACE : protein list to PRIDE-XML file (Export Module)
	13~14	WEB INTERFACE : protein list to EXCEL file (Export Module)
	12	WEB INTERFACE : Sequence search result (br BLAST) output format change.
2008/03	20	PROTEOME DATA : korean plasma proteome from KSBI
	17~19	WEB INTERFACE : Document - Demo movie
		PROTEOME DATA : HCC plasma proteome separated by PF2D PROTEOME DATA : Normal plasma proteome separated by PF2D

2008 © Yonsei Proteome Research Center

#424, YPRC/SPRC, Industry-University Research Center, Yonsei Univ., Seodaemun-gu, Seoul, Korea, 120-749
Tel: +82-2-2123-6626, Fax: +82-2-393-6589

Figure 2. Summary of BiomarkerDigger web-page.

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

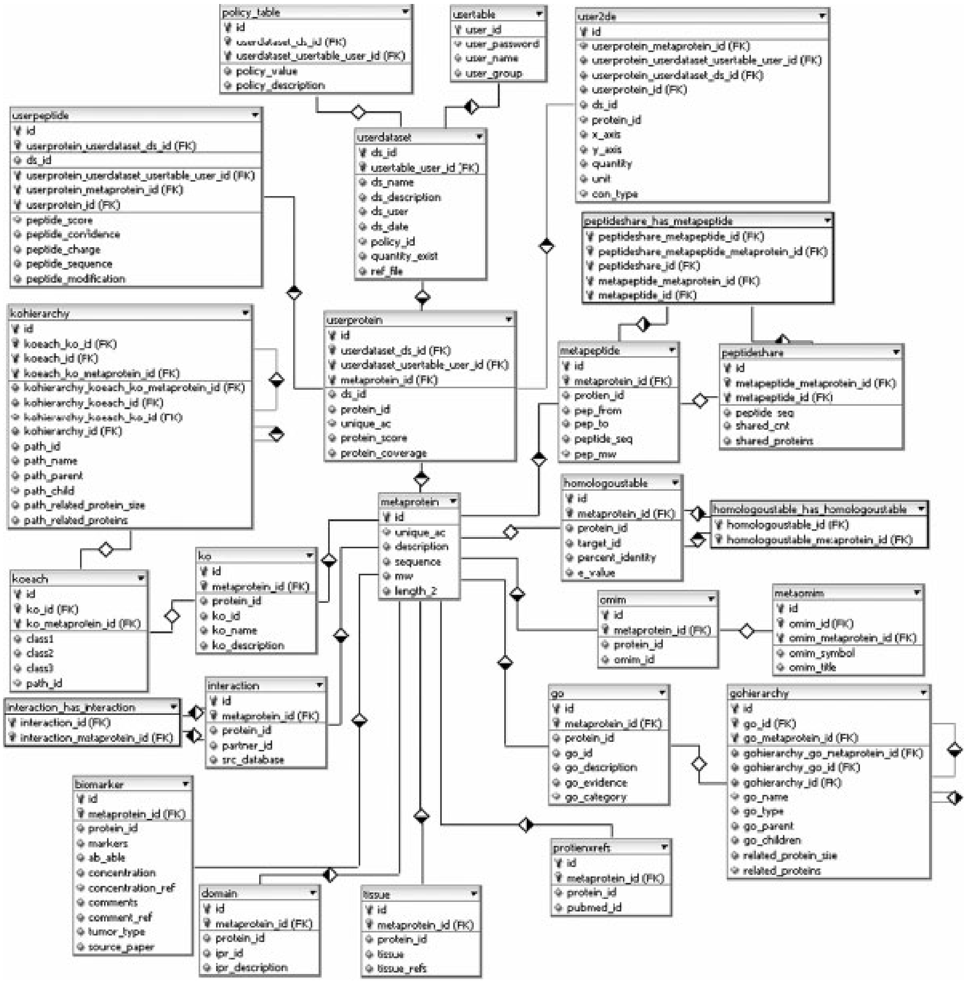


Figure 3. Entity-relationship diagram of BiomarkerDigger. Boxes represent entities or tables and lines show relationships between the entities. Symbol '-◈-' represents 1: n relation; symbol '◇' shows 1:1 relation.

3 Results and discussion

3.1 Overview of BiomarkerDigger

BiomarkerDigger incorporates three major functions (Fig. 1). First, it has an integrated DB function that compiles proteome data and displays search query results as requested by the end user. Second, it has an automated information collection feature that retrieves additional data stored in various DBs. Finally, it has a comparative analysis function that mines the data stored in BiomarkerDigger. With these three functions, users can store, analyze, and compare proteomic data, with respect to a specific disease of interest (Fig. 4).

3.2 Data submission

Using the 'Submit' menu, users can submit their data set(s) to 'BiomarkerDigger' for analysis using functionalities in this system. Users may submit protein lists that contain accession codes and those eXtensible Markup Language (XML) files obtained from any type of search engines (*e.g.* MASCOT, SEQUEST and ProteinProphet [27]). PRIDE XML files may also be submitted. Data submission can be done using a protein list, without additional information that shall be incorporated into the DB by BiomarkerDigger (Fig. 4). Files submitted from different sources may contain various archived protein sequence information that can be used for protein identification, and therefore may have different accession codes and descriptions for the same protein. This would be expected to cause redundancies that could complicate cross-comparisons of proteins from different experimental data sets. To overcome this limitation, we developed 'idMapper' to reduce redundancies and show uniqueness by converting the accession codes of the proteins in the list into the UniProtKB primary accession codes.

3.3 Data search and query options

Information deposited in BiomarkerDigger can be searched using multiple query methods *via* web-interface. The query options consist of protein description, molecular weight,

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

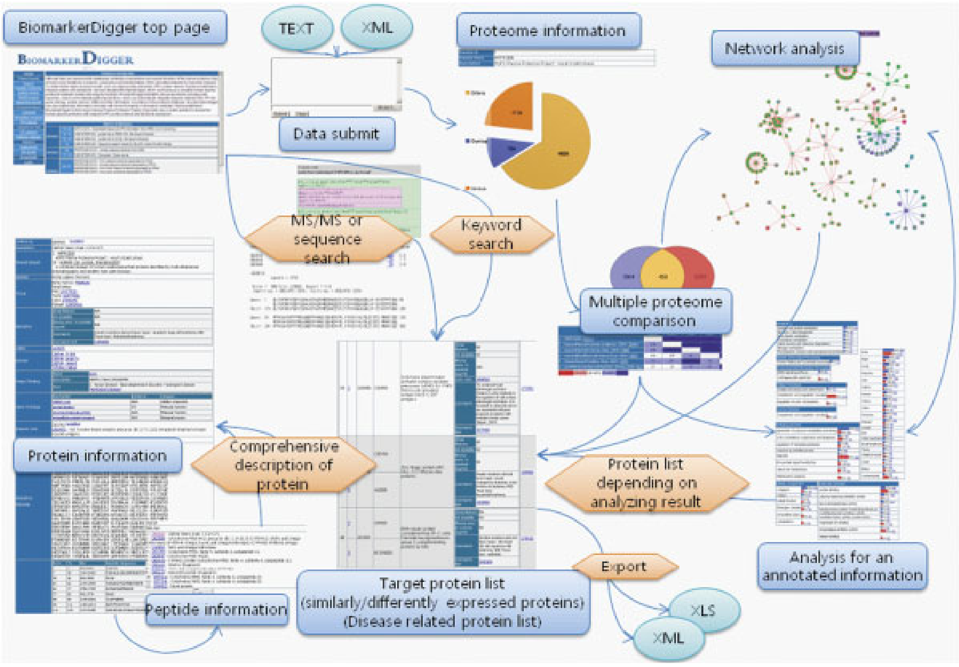

BiomarkerDigger top page
BIOMARKERDIGGER
TEXT
XML
Data submit
Proteome information
Network analysis
MS/MS or sequence search
Keyword search
Multiple proteome comparison
Protein information
Comprehensive description of protein
Protein list depending on analyzing result
Analysis for an annotated information
Export
Peptide information
Target protein list (similarly/differently expressed proteins) (Disease related protein list)
XLS
XML

Figure 4. The BiomarkerDigger user interface flow diagram.

sequence length, protein domains, annotated GO term, gene symbol, and accession codes for the most common public DBs (NCBI, UniProt, and IPI). Search queries can be based on the relationship to a specific metabolic pathway, genetic disorder, cancer gene, biomarkers, or biomarker candidates. It is also possible to search for data sets stored in BiomarkerDigger using MS/MS data or protein sequences through X!tandem [28] or BLAST [29], which are embedded in BiomarkerDigger. BiomarkerDigger also provides a search function to identify other proteins in a shared metabolic pathway, classified by KO term. These diverse search functions expand the utility of BiomarkerDigger and allow it to identify any protein (s) or protein family that has been functionally categorized (Fig. 4).

3.4 Data export

Information stored or processed in BiomarkerDigger can be extracted in Excel and XML file formats when generating search result reports using the export function (Fig. 4). Since XML files comply with the standard file format set by the HUPO Proteomics Standard Initiatives [30], researchers can directly submit their own data set as an XML file to PRIDE or other public DBs that follow HUPO PSI guidelines (http://www.psidev.info/). Thus, the BiomarkerDigger DB can process, store, search, and report various proteomic experimental results in multiple file formats.

3.5 Data gathering and mining

Interpretation of proteomic experimental data for the purpose of biomarker discovery or comparative proteomic analysis requires the gathering of all the necessary information associated with the candidate proteins and proteomes (Table 1, Fig. 1). The protein lists identified by search engines on the basis of MS or MS/MS peak lists contain DB accession codes and descriptions, but additional information is needed for the interpretation of data. Information related to pathways, functional annotation, and interacting partners can be acquired by accessing relevant DBs. BiomarkerDigger uses an automated integrated mapping function that interconnects various types of DB accession codes to facilitate the presentation of gathered information. BiomarkerDigger can also generate peptide information for submitted proteins using ‘peptideCutter’ and ‘peptideMapper’. These modules use the protein information files to generate tryptic peptide lists, protein sequence reports, predicted molecular weight, and a list of human protein that contains a given peptide. The information collected for specific proteins can also be viewed as a report file through web-interface, which provides both a comprehensive description and associated analytical data extracted from the DB. The report files also contain information regarding protein function including tissue expression patterns, related pathways, GO annotation, interacting proteins, functional domains, and disease associations. In particular, cancer-related information is searchable and includes cancer type, gene products of oncogenes, association with genetic disorders, and occurrences as biomarkers or biomarker candidates. Thus, the peptide and associated information compiled in the report provided by BiomarkerDigger can be used to relate an individual protein to a specific disease. To evaluate biomarker candidates, one must perform the relevant experiment under well-controlled conditions using high-throughput, sensitive, and quantita-

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

tive assays [31, 32]. For example, to carry out MRM, it is necessary to identify biomarker-specific peptides (*i.e.* proteotypic peptides) in a given disease protein and obtain related information [33, 34]. In this regard, BiomarkerDigger can also provide with those parameters that are essential for MRM analysis: protein identity, sequence length, and molecular weight.

3.6 Comparison of the interpreted experimental data

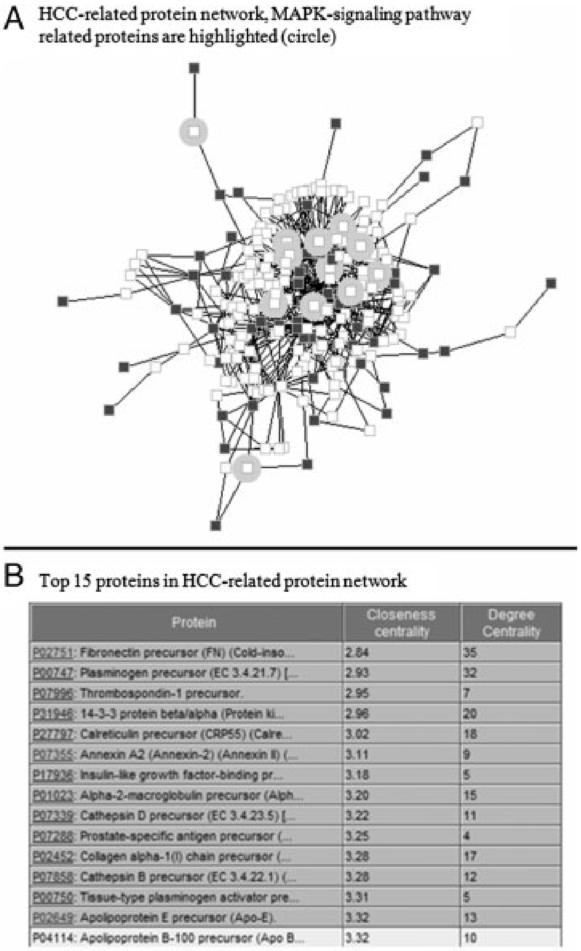
The comparative analysis tool of BiomarkerDigger web-interface uses a simple logical operation to facilitate the comparative analysis of data in different formats. However, assessment of similarities and differences between two or more experiments can only be done by the comparison of protein lists. This is not applied to other data that have been collected using non-standard methods (*e.g.* a different scoring system based on different MS analysis methods). Results from comparisons of protein lists can be viewed as a compiled protein list, Venn diagram, or heat map, which can be used for further comparisons or analyses. It is also possible to search and export comparison results as previously described (Fig. 4).

3.7 Analysis and comparison of annotated information

BiomarkerDigger facilitates a direct comparison of similarity and identifies annotated information between two data sets through a simple interface supported by web. Annotated proteomic information can be viewed in two modes: within a single data set and between two different data sets. If a single data set is selected, BiomarkerDigger displays a histogram showing the distribution of annotations within the selected data set. If two data sets are selected, differences in the annotation distribution are presented as a graph that displays fold differences. It also generates compilations consisting of GO terms, KO pathways, protein domains, and tissue expression patterns. GO hierarchy terms provide complementary functional information when sorted by biological process, molecular function, or cellular localization. The presence of homology domains is also an important indicator of protein function. Finally, when KO information is available, it is possible to carry out a proteome-wide comparative analysis of an involvement of any metabolic pathway.

3.8 Analysis of PPI networks

BiomarkerDigger uses the existing PPI data to create PPI networks that can be analyzed with respect to modularity and topology (Fig. 5). Network functions allow partitioning into several modules depending on the intensity of the interaction between individual elements within the network. Although proteins present in one module may exert strong interaction with others within the same module, proteins present in other modules may exert less interaction [26, 35]. Users can divide a network into smaller modules through which they may be able to obtain protein function and metabolic pathway information limited to one of the modules in the network. The degree distribution conveys information content as a function of the relatedness of one protein to another [36]. Closeness centrality provides a measure as to how important a given protein is with respect to the overall protein network [36]. When two data sets are compared, BiomarkerDigger can identify both commonalities and differences in topology and modularity. From the analysis of modularity and topology, it is also possible to predict the functional distribution, functional relatedness, importance of each protein with respect to the network,


A
HCC-related protein network, MAPK-signaling pathway related proteins are highlighted (circle)
B
Top 15 proteins in HCC-related protein network

Protein	Closeness centrality	Degree Centrality
P02751: Fibronectin precursor (FN) (Cold-inso...	2.84	35
P00747: Plasminogen precursor (EC 3.4.21.7) [...	2.93	32
P07996: Thrombospondin-1 precursor.	2.95	7
P31946: 14-3-3 protein beta/alpha (Protein ki...	2.96	20
P27797: Calreticulin precursor (CRP55) (Calre...	3.02	18
P07355: Annexin A2 (Annexin-2) (Annexin II) (...	3.11	9
P17936: Insulin-like growth factor-binding pr...	3.18	5
P01023: Alpha-2-macroglobulin precursor (Alph...	3.20	15
P07339: Cathepsin D precursor (EC 3.4.23.5) [...	3.22	11
P07288: Prostate-specific antigen precursor (...	3.25	4
P02452: Collagen alpha-1(I) chain precursor (...	3.28	17
P07858: Cathepsin B precursor (EC 3.4.22.1) (...	3.28	12
P00750: Tissue-type plasminogen activator pre...	3.31	5
P02649: Apolipoprotein E precursor (Apo-E).	3.32	13
P04114: Apolipoprotein B-100 precursor (Apo B...	3.32	10

Figure 5. PPI network in HCC. (A) A PPI network constructed using the list of 111 HCC-related proteins; black square boxes represent cancer-related proteins. (B) The top fifteen proteins extracted from the protein list based on closeness centrality. The degree centrality for the two largest sub-networks is shown. Known cancer-associated proteins are indicated in dark grey.

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 2. BiomarkerDigger database composition

	No. of data set	No. of entry	Disease-related proteins	Biomarker/biomarker candidate related	Data source
Plasma proteome[a)]	4	5031	2483	679	PRIDE, literature [37, 38, 39, 41]
Highly abundant protein associated protein[b)]	1	203	146	52	Literature [40]
Glycoprotein[c)]	2	746	552	174	Literature [42, 43]
CSF[d)]	1	2141	1118	262	Literature [44]
PF2D-HCC[e)]	2	320	267	99	BiomarkerDigger(YPRC) [25]
PF2D-normal[e)]	2	423	350	128	BiomarkerDigger(YPRC) [25]
Narrow 2D-normal Korean plasma[f)]	6	64	48	25	BiomarkerDigger(YPRC)
Korean plasma[f)]	1	248	156	74	BiomarkerDigger(KSBI)
Total[g)]	19	7223	3536	870	

a) Anderson *et al.* [37]; Shen *et al.* [38]; Rose *et al.* [39]; Omenn *et al.* [41].
b) Zhou *et al.* [40].
c) Yang *et al.* [42]; Zhang *et al.* [43].
d) Pan *et al.* [44].
e) Lee *et al.* [25].
f) Unpublished data.
g) Not including duplicate entries.

communication between proteins, and differences between two data sets. A PPI-based network can serve as a model for assessment of the overall phenomenon under investigation.

3.9 Contents of the current version

BiomarkerDigger contains a plasma protein data set derived from recently published reports [37–40] and a protein list from the HUPO Plasma Proteome Project (HPPP_3020) [41]. Glycoproteins listed in HPPP [42] and other sources [43] were used as a glycoprotein reference DB. Human cerebrospinal fluid proteome results obtained from multidimensional chromatography and MS/MS [44] were used for a cerebrospinal fluid reference. Table 2 outlines the entry data sets and their sources. Information as to other biofluids would also be available in the future. The results obtained from various reference data sets and a comparative analysis of proteins differentially expressed in HCC patients can be used to search for HCC biomarker candidates.

3.10 Application examples

BiomarkerDigger is useful for single and multiple comparative analyses and is a versatile tool for analyzing proteome data sets that have been merged or split from the existing proteome data sets on the basis of biological processes, cellular components, tissue distribution, biological pathway relationships, or other user-defined criteria. The latter feature enables targeted analysis. Using this unique feature of BiomarkerDigger, multiple comparative analyses of clinical sample sets (*e.g.* HCC tumor tissue *versus* non-tumor tissue or HCC plasma *versus* healthy plasma), and various selection criteria, we identified a total of 111 HCC-related proteins. BiomarkerDigger can also be used to create a PPI network for these proteins that showed a central clustering pattern around well-characterized cancer-related proteins.

3.10.1 Comparative analysis between different proteomes

As an attempt to screen for serologic biomarker candidates, we performed a comparative analysis of healthy and HCC plasma proteomes using in-house data sets from healthy plasma samples and HPPP DB [41]. As shown in Fig. 6A, we were able to identify 13 HCC-related plasma proteins that were not previously detected in the plasma protein data set generated by our in-house parallel analysis of healthy plasma. Of the 13 HCC-related plasma proteins, three were not found in the HPPP data set [41] or the healthy plasma data set generated by our laboratory (Supporting Information Table S1, list 1). Focusing on those glycoproteins that might be detected in plasma proteins [42, 43] it was revealed that 64% of HCC plasma proteins (32 of 50) and 41% of healthy plasma proteins (33 of 61) were glycoproteins (Fig. 6B). Of these 13 HCC-related plasma proteins, 8 were predicted to be glycoproteins

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

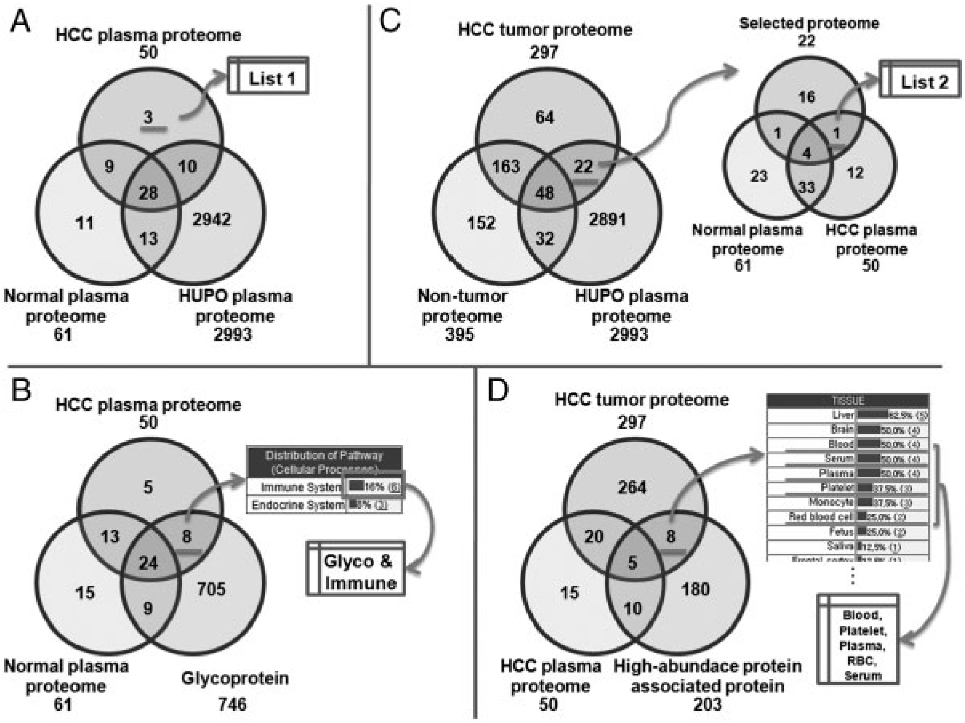


Figure 6. Comparison of multiple datasets. (A) Protein lists from in-house HCC and healthy plasma data sets and the HUPO plasma proteome database [41]. (B) Glycoprotein list compiled from the in-house HCC and healthy plasma proteome data sets and other plasma protein data sets [42, 43]. (C) Tissue proteome distribution patterns (left) and reclassification of particular groups of proteins present in plasma. (D) Glycoprotein list (right) extracted from the list of proteins present in HCC plasma and tumor tissues (left), and proteins associated with high-abundance proteins [40].

(Supporting Information Table S1, Glyco). A comparative analysis of the tissue proteome showed that 86 proteins were found only in HCC tumor tissues (Fig. 6C) and of these, 70 were listed in the HPPP DB where 80 non-tumor-tissue proteins were also found. A subset of 22 of the 86 HCC tumor-tissue-only proteins were found in common in the HPPP DB and the tumor-tissue protein list. As an example, a comparison of plasma proteins from the HPPP DB to those from the healthy plasma protein list produced in our laboratory showed that only IGKV1-5 protein was detected in the HCC-related plasma protein list but not in the healthy plasma protein list (Supporting Information Table S1, list 2). This suggests that these proteins are involved in hepatocarcinogenesis.

We usually deplete HCC and healthy plasma of the six or seven most abundant plasma proteins before either 2-DE or 2-D LC-MS/MS using a series of affinity columns, each one depleting a single high-abundance protein [25, 44–46]. To determine whether any of the HCC-related proteins is associated with high-abundance proteins removed during the depletion process, a comparative analysis was made among HCC tumor tissues (non-depleted), the HCC plasma proteome (depleted), and the high-abundance protein associated proteins [40, 45, 46]. As shown in Fig. 6D, eight tumor-tissue proteins that had not been previously found in plasma were identified, confirming their association with high-abundance proteins as previously reported [40]. Based on UniProtKB and HPRD annotations of tissue samples, six of these eight proteins were also found to be expressed in blood, platelets, plasma, red blood cells, and serum (Table 3).

Table 3. HCC tissue proteins that appear to be associated with high-abundance plasma proteins but are not detected in HCC

Accession code	Description	Tissue expression (annotated[a)])
P02671	Fibrinogen α chain precursor [contains: fibrinopeptid...	Platelet, plasma, blood, liver
P01024	Complement C3 precursor [contains: complement C3 β cha...	Platelet, serum, plasma, liver
P69905	Hemoglobin subunit α (hemoglobin α chain) (α-...	Platelet, red blood cell, blood
P68871	Hemoglobin subunit β (hemoglobin β chain) (β-glo...	Red blood cell, serum, blood
P02751	Fibronectin precursor (FN) (cold-insoluble globulin) (CIG).	Plasma, serum, liver
P02649	Apolipoprotein E precursor (Apo-E).	Serum, plasma, blood

a) Annotated information from UniProtKB comment and expression section of HPRD.

3.10.2 Selection criteria for HCC-related biomarkers

We used BiomarkerDigger to perform a comparative analysis of various protein data sets, including HCC tissue proteins, healthy tissue proteins, plasma proteins of healthy and HCC subjects, the HPPP data set and known

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

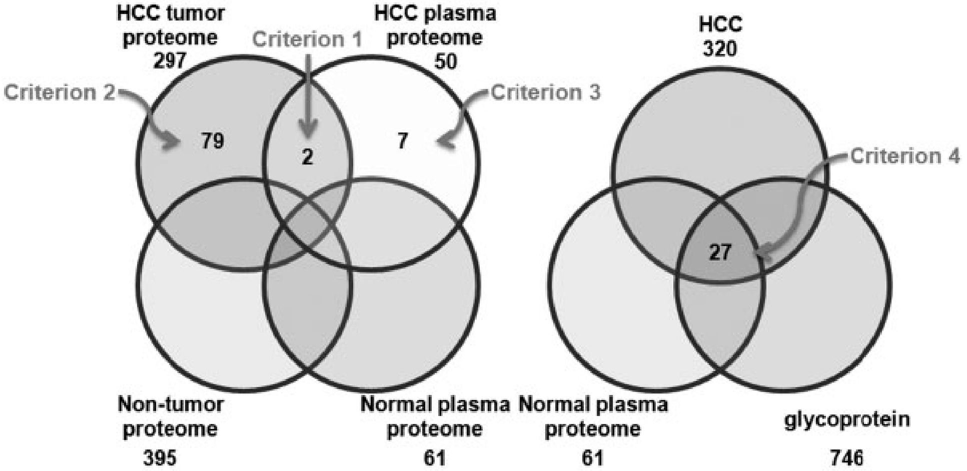


Figure 7. The criteria for selecting HCC-related serum proteins from HCC liver tissue, HCC plasma, healthy liver tissue, and healthy plasma proteome data sets.

glycoproteins [42, 43]. A total of 111 non-redundant HCC-related proteins were identified by performing a homology search using the BLAST search program (>75% similarity, *e*-value <0.001) (Supporting Information Table S1) [29]. Potential HCC biomarker candidates belonging to several categories were thus identified (Fig. 7).

Selection Criterion 1: Proteins detected in HCC plasma/tissues but not in healthy plasma/tissue samples. Two proteins that were likely preferentially expressed during hepatocarcinogenesis were identified as potential biomarker candidates (Supporting Information Table S1). Proteins that appeared in either healthy tissues or plasma, regardless of HCC status, were eliminated.

Selection Criterion 2: Proteins found in HCC liver tumor tissues, but not in healthy tissue/plasma or HCC plasma samples. Proteins detected in HCC tissues but not in healthy tissues were classified as HCC-specific proteins. The subset not detected in either healthy or HCC plasma was not considered to be released into plasma or other biofluids. A total of 79 proteins belonged to this category; annotation information related to serological terms was available for four of these proteins, which therefore would be considered as HCC biomarker candidates potentially detectable in blood or plasma.

Selection Criterion 3: Proteins detected in HCC plasma, but not detected in healthy plasma or healthy/HCC tissue samples. Seven proteins were in this category that may represent non-specific markers.

Selection Criterion 4: Glycoproteins identified in HCC tissues/plasma and healthy plasma. Proteins identified in HCC tissue/plasma could be newly expressed in HCC, differentially expressed only in HCC possibly due to a change in function, or constitutively present regardless of disease state. Of the proteins whose function might have changed in association with HCC, we focused on glycoproteins because glycosylation is the most probable type of structural change that could result in alteration in the protein function [47]. We found a total of 27 such proteins, 15 of which were known to be associated with cancer [19, 22, 23].

3.10.3 PPI network analysis of HCC-related proteins

Using BiomarkerDigger, we created a PPI network for the 111 HCC-related biomarker candidates (Fig. 5). In this HCC-related protein network, most of the proteins (top 14 proteins), which were centrally positioned are related with cancer. Fibronectin (closeness centrality 2.84, degree centrality 35) has been implicated in extracellular matrix-receptor interaction, small cell lung cancer, increased in renal cell cancer, and differentially expressed in HCC, whereas fibronectin is known to interact with the tumor necrosis factor ligand superfamily, Hepatocyte growth factor, and other cancer-related proteins (*e.g.* Coagulation factor XIII, ADAMTS-4, Matrix metalloproteinase-9) [17, 48–51]. 14-3-3 β (closeness centrality 2.96, degree centrality 20) has been implicated in cell growth, cell death, and lung cancer, whereas 14-3-3 β/α is known to interact with the B-Raf and C-Raf proto-oncogenes, casein kinase, and other cancer-related proteins (*e.g.* insulin-like growth factor 1 receptor precursor, tumor necrosis factor, α-induced protein 3) [17, 52, 53]. Thus, BiomarkerDigger readily identified several well-known key proteins involved in cancer-related protein networks. This approach provides a suitable framework for further studies of interacting proteins, in terms of both target binding and involvement in common pathways. It also provides useful related information, such as the number of the network protein, the identity of key network proteins, and the relationship of the network proteins to cancer.

4 Concluding remarks

In conclusion, BiomarkerDigger is a versatile DB that provides safe data storage, data extraction, user-friendly search functions, and flexible reporting of results in a various formats. It also facilitates automated collection and display of protein information, enabling a comparative

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

proteomic analysis of the experimental results. Using BiomarkerDigger, it is possible to identify distribution patterns and compare specific proteins based on GO terms, KO pathways, protein homology domains, and tissue expression patterns. This assessment can also be extended to a PPI network analysis, which yields information about regulatory modularity and topology of proteins in a protein network model.

BiomarkerDigger can be used to perform a comprehensive comparative proteome analysis of healthy and HCC plasma and tissue samples in order to readily identify HCC-related proteins based on a wide variety of selection criterion. Thus, BiomarkerDigger is a flexible search tool that offers a diversity of functions to identify potential biomarkers for specific diseases using automated mining of available DB and literature resources. Work is in progress to assess the expression level of HCC biomarker candidates in clinical samples. The PPI network mapping function of BiomarkerDigger is also currently being expanded to contain an analysis tool for the prediction of function, domain, and pathway-based interactions between proteins.

All source codes and sql files (for DB schema, meta-data) can be downloaded at http://www.biomarkerdigger.org/.

This study was supported by a grant from the Korea Health 21 R&D project (Ministry of Health, Welfare and Family Affairs, Republic of Korea, A030003)and PALL company (to YKP).

The authors have declared no conflict of interest.

5 References

[1] Paik, Y. K., Kim, H., Lee, E. Y., Kwon, M. S. *et al.*, Overview and introduction to clinical proteomics. *Methods Mol. Biol.* 2008, *428*, 1–31.

[2] Jones, P., Côté, R. G., Martens, L., Quinn, A. F. *et al.*, PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.* 2006, *34*, D659–D663.

[3] Deutsch, E. W., Eng, J. K., Zhang, H., King, N. L. *et al.*, Human plasma peptide atlas. *Proteomics* 2005, *5*, 3497–3500.

[4] Zhang, H., Loriaux, P., Eng, J., Campbell, D. *et al.*, UniPep – a database for human N-linked glycosites: a resource for biomarker discovery. *Genome Biol.* 2006, *7*, R73.

[5] Craig, R., Cortens, J. P., Beavis, R. C., Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* 2005, *3*, 1234–1242.

[6] Bodenmiller, B., Malmstrom, J., Gerrits, B., Campbell, D. *et al.*, PhosphoPep – a phosphoproteome resource for systems biology research in Drosophila Kc167 cells. *Mol. Syst. Biol.* 2007, *3*, 139.

[7] Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z. *et al.*, Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* 2003, *13*, 2363–2371.

[8] Uhlen, M., Ponten, F., Antibody-based proteomics for human tissue profiling. *Mol. Cell Proteomics* 2005, *4*, 384–393.

[9] Prokisch, H., Andreoli, C., Ahting, U., Heiss, K. *et al.*, MitoP2: the mitochondrial proteome database – now including mouse data. *Nucleic Acids Res.* 2006, *34*, D705–D711.

[10] Schlüter, A., Fourcade, S., Domènech-Estévez, E., Gabaldón, T. *et al.*, PeroxisomeDB: a database for the peroxisomal proteome, functional genomics and disease. *Nucleic Acids Res.* 2007, *35*, D815–D822.

[11] Cotter, D., Maer, A., Guda, C., Saunders, B. *et al.*, LMPD: LIPID MAPS proteome database. *Nucleic Acids Res.* 2006, *34*, D507–D510.

[12] Gendler, K., Paulsen, T., Napoli, C., ChromDB: the chromatin database. *Nucleic Acids Res.* 2007, *36*, D298–D302.

[13] Mead, J. A., Bianco, L., Bessant, C., Recent developments in public proteomic MS repositories and pipelines. *Proteomics* 2009, *9*, 861–881.

[14] Jayapandian, M., Chapman, A., Tarcea, V. G., Yu, C. *et al.*, Michigan Molecular Interactions (MiMI): putting the jigsaw puzzle together. *Nucleic Acids Res.* 2007, *35*, D566–D571.

[15] Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y. *et al.*, The International Protein Index: an integrated database for proteomics experiments. *Proteomics* 2004, *4*, 1985–1988.

[16] The UniProt Consortium, The universal protein resource (UniProt) 2009. *Nucleic Acids Res.* 2009, *37*, D169–D174.

[17] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F. *et al.*, From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* 2006, *34*, D354–D357.

[18] Côté, R. G., Jones, P., Martens, L., Kerrien, S. *et al.*, The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics* 2008, *8*, 401.

[19] Polanski, M., Anderson, L. N., A list of candidate cancer biomarkers for targeted proteomics. *Biomarker Insights* 2006, *2*, 1–48.

[20] Futreal, P. A., Coin, L., Marshall, M., Down, T. *et al.*, A census of human cancer genes. *Nat. Rev. Cancer* 2004, *4*, 177–183.

[21] Huret, J. L., Dessen, P., Bernheim, A., Atlas of genetics and cytogenetics in oncology and haematology, year 2003. *Nucleic Acids Res.* 2003, *31*, 272–274.

[22] Higgins, M. E., Claremont, M., Major, J. E., Sander, C. *et al.*, Cancer genes: a gene selection resource for cancer genome projects. *Nucleic Acids Res.* 2007, *35*, D721–D726.

[23] Su, W. H., Chao, C. C., Yeh, S. H., Chen, D. S. *et al.*, OncoDB.HCC: an integrated oncogenomic database of hepatocellular carcinoma revealed aberrant cancer target genes and loci. *Nucleic Acids Res.* 2007, *35*, D727–D731.

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

[24] Li, L. C., Zhao, H., Shiina, H., Kane, C. J. *et al.*, PGDB: a curated and integrated database of genes related to the prostate. *Nucleic Acids Res.* 2003, *31*, 291–293.

[25] Lee, H. J., Kang, M. J., Lee, E. Y., Cho, S. Y. *et al.*, Application of a peptide-based PF2D platform for quantitative proteomics in disease biomarker discovery. *Proteomics* 2008, *8*, 3371–3381.

[26] Yoon, J., Blumer, A., Lee, K., An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality. *Bioinformatics.* 2006, *22*, 3106–3108.

[27] Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R. *et al.*, A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 2004, *75*, 4646–4658.

[28] Craig, R., Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004, *20*, 1466–1467.

[29] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J. *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997, *25*, 3389–3402.

[30] Orchard, S., Taylor, C. F., Jones, P., Montechi-Palazzo, L. *et al.*, Entering the implementation era: a report on the HUPO-PSI Fall workshop 25-27 September 2006, Washington DC, USA. *Proteomics* 2007, *7*, 337–339.

[31] Gortzak-Uzan, L., Ignatchenko, A., Evangelou, A. I., Agochiya, M. *et al.*, A proteome resource of ovarian cancer ascites: integrated proteomic and bioinformatic analyses to identify putative biomarkers. *J. Proteome Res.* 2007, *7*, 339–351.

[32] Rifai, N., Gillette, M. A., Carr, S. A., Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.* 2006, *24*, 971–983.

[33] Keshishian, H., Addona, T., Burgess, M., Kuhn, E. *et al.*, Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution. *Mol. Cell Proteomics* 2007, *6*, 2212–2229.

[34] Anderson, L., Hunter, C. L., Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol. Cell Proteomics* 2006, *5*, 573–588.

[35] Hartwell, L. H., Hopfield, J. J., Leibler, S., Murray, A. W. *et al.*, From molecular to modular cell biology. *Nature* 1999, *402*, C47–52.

[36] Wouter, D. N., Andrej, M., Vladimir, B., *Exploratory Social Network Analysis with Pajek*, Cambridge University Press, New York 2005.

[37] Anderson, N. L., Polanski, M., Pieper, R., Gatlin, T. *et al.*, The human plasma proteome: a nonredundant list developed by combination of four separate sources. *Mol. Cell Proteomics* 2004, *3*, 311–326.

[38] Shen, Y., Jacobs, J. M., Camp, D. G., Fang, R. *et al.*, Ultra-high-efficiency strong cation exchange LC/RPLC/MS/MS for high dynamic range characterization of the human plasma proteome. *Anal. chem.* 2004, *76*, 1134–1144.

[39] Rose, K., Bougueleret, L., Baussant, T., Böhm, G. *et al.*, Industrial-scale proteomics: from liters of plasma to chemically synthesized proteins. *Proteomics* 2005, *4*, 2125–2150.

[40] Zhou, M., Lucas, D. A., Chan, K. C., Issaq, H. J. *et al.*, An investigation into the human serum ''interactome''. *Electrophoresis* 2005, *25*, 1289–1298.

[41] Omenn, G. S., States, D. J., Adamski, M., Blackwell, T. W. *et al.*, Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* 2005, *5*, 3226–3245.

[42] Yang, Z., Hancock, W. S., Chew, T. R., Bonilla, L. *et al.*, A study of glycoproteins in human serum and plasma reference standards (HUPO) using multilectin affinity chromatography coupled with RPLC-MS/MS. *Proteomics* 2005, *5*, 3353–3366.

[43] Zhang, H., Liu, A. Y., Loriaux, P., Wollscheid, B. *et al.*, Mass spectrometric detection of tissue proteins in plasma. *Mol. Cell Proteomics* 2007, *6*, 64–71.

[44] Pan, S., Zhu, D., Quinn, J. F., Peskind, E. R. *et al.*, A combined dataset of human cerebrospinal fluid proteins identified by multi-dimensional chromatography and tandem mass spectrometry. *Proteomics* 2007, *7*, 469–473.

[45] Cho, S. Y., Lee, E. Y., Lee, J. S., Kim, H. Y. *et al.*, Efficient prefractionation of low-abundance proteins in human plasma and construction of a two-dimensional map. *Proteomics* 2005, *5*, 3386–3396.

[46] Cho, S. Y., Lee, E. Y., Kim, H. Y., Kang, M. J. *et al.*, Protein profiling of human plasma samples by two-dimensional electrophoresis. *Methods Mol. Biol.* 2008, *428*, 57–75.

[47] Turnbull, J. E., Field, R. A., Emerging glycomics technologies. *Nat. Chem. Biol.* 2007, *3*, 74–77.

[48] Hegele, A., Heidenreich, A., Kropf, J., von Knobloch, R. *et al.*, Plasma levels of cellular fibronectin in patients with localized and metastatic renal cell carcinoma. *Tumour Biol.* 2004, *25*, 111–116.

[49] Honda, M., Kaneko, S., Kawai, H., Shirota, Y. *et al.*, Differential gene expression between chronic hepatitis B and C hepatic lesion. *Gastroenterology* 2001, *120*, 955–966.

[50] Delpuech, O., Trabut, J. B., Carnot, F., Feuillard, J. *et al.*, Identification, using cDNA macroarray analysis, of distinct gene expression profiles associated with pathological and virological features of hepatocellular carcinoma. *Oncogene* 2002, *21*, 2926–2937.

[51] Li, C., Hong, Y., Tan, Y. X., Zhou, H. *et al.*, Accurate qualitative and quantitative proteomic analysis of clinical hepatocellular carcinoma using laser capture microdissection coupled with isotope-coded affinity tag and two-dimensional liquid chromatography mass spectrometry. *Mol. Cell Proteomics* 2004, *3*, 399–409.

[52] Xiao, T., Ying, W., Li, L., Hu, Z. *et al.*, An approach to studying lung cancer-related proteins in human blood. *Mol. Cell Proteomics* 2006, *4*, 1480–1486.

[53] Eggstein, S., Manthey, G., Hirsch, T., Baas, F. *et al.*, Raf-1 kinase, epidermal growth factor receptors, and mutant Ras proteins in colonic carcinomas. *Dig. Dis. Sci.* 1996, *41*, 1069–1075.

[54] McKusick, V.A., *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders*, 12th Edn), Johns Hopkins University Press, Baltimore 1998.

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim