Submitted for HPP WG (Chromosome-based Subgroup)
Proposed for Publishing at HUPO HPP Web Page

Ministry of Education and Science of the Russian Federation

National Human Proteome Project Working Group

Russian Proteome Society (RHUPO)

ROADMAP

Proteome of the 18th Human Chromosome: Gene Centric Identification of Transcripts, Proteins and Peptides

(DRAFT, May 31st 2010)

Moscow, 2010

Roadmap for the National Gene Centric Human Proteome Project

- Ministry of Education and Science of the Russian Federation
- Ministry of Health and Social Development of the Russian Federation
- Department of Science, High Technologies and Education of the Government of the Russian Federation
- Russian Academy of Sciences
- Russian Academy of Medical Sciences
- Russian Proteome Society (RHUPO)

www.proteome.ru www.ibmc.msk.ru

(c) ARCHAKOV A.[1], BYKOV V. [2], GOVORUN V. [3], IVANOV V. [3], KHLUNOV A. [4], LISITSA A. [1], MAKAROV A. [5], SAGDEEV R. [6], SKRYABIN K. [7]

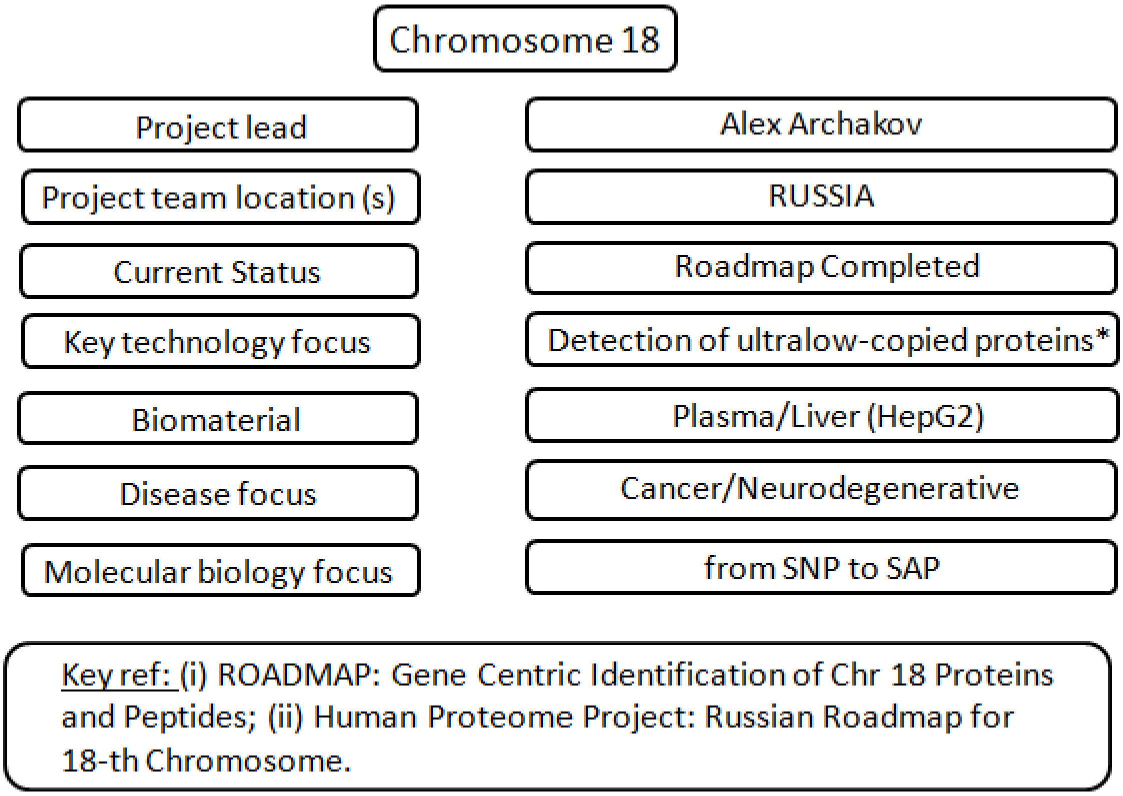
1 – V.N. Orekhovich Institute of Biomedical Chemistry (Moscow)
2 – Joint-Stock Company “NT-MTD” (Zelenograd)
3 – M.M. Shemjakin & Yu.A.Ovchinnikov Institute of Bioorganic Chemistry (Moscow)
4 – Department of Science, High Technologies and Education of Government of Russian Federation
5 – V.A. Engelgard Institute of Molecular Biology (Moscow)
6 – International Tomography Center (Novosibirsk)
7 – Center “Bioengineering” (Moscow)

Abstract

The Human Proteome Project is focused on the inventory of all human proteins and the revelation of interactions occurring between them. In its scale, this project is superior to the Human Genome Project formally finished in 2001. The basic problems that HPP is facing are: impossibility of detecting single molecules in biological material and a situational nature of a proteome, i.e. the dependence of protein composition on the time and type of tissue and cells. Due to the presence of single amino acid polymorphisms (SAP), alternative splicing (AS) and post-translational modifications (PTM), as part of a whole proteome, one might assume the availability of several million of various protein forms.

As of now, the road map of the project is prepared as the Russian national gene-centric initiative. Within the pilot phase there will be revealed the transcripts of the 18th chromosome in the liver cells and 286 proteins in the liver cells and blood plasma and, also, novel technologies will be developed: as such, a technology on the basis of coupling AFM and MS, analytical device with the usage of nanowire sensing element, cloud informational system with the personal supercomputers.

The present stage of experimental realization of the project includes the identification of 220 transcripts encoding proteins of the 18th chromosome in HepG2, 90 proteins of the 18th chromosome identified with 3D-LC-MS/MS Ion Trap method in HepG2 and, also, the revelation of 36 proteins in blood plasma. Using bio-informational methods, the proteins of the 18th chromosome are characterized, including the assessment of medical significance of proteins and revelation of potential protein partners.



Chromosome 18	
Project lead	Alex Archakov
Project team location (s)	RUSSIA
Current Status	Roadmap Completed
Key technology focus	Detection of ultralow-copied proteins*
Biomaterial	Plasma/Liver (HepG2)
Disease focus	Cancer/Neurodegenerative
Molecular biology focus	from SNP to SAP

Key ref: (i) ROADMAP: Gene Centric Identification of Chr 18 Proteins and Peptides; (ii) Human Proteome Project: Russian Roadmap for 18-th Chromosome.

* Down to concentration 10^{-16} M in plasma and ~10 copies per liver (HepG2) cell

Contents

1. Introduction 5
2. Gene Centric Approach 6
3. Implementation Schedule 8
4. Chromosome Selection..9
5. Chromosome 18...11
6. Biomaterial: Liver Tissue and Blood Plasma 12
7. Standardization and Proteotyping..16
8. Analysis of Transcripts and High/Medium-copied Proteins 19
9. Peptidome-based Drug Design 22
10. Detection of Low/Ultralow-copied Proteins...22
11. Interactomics 24
12. Knowledgebase...26
13. Short-term Deliverables (5-year view) 27
14. Long-term Deliverables (15 year view)...28
15. Concluding Remarks 29
References 31

1. Introduction

Akin the Human Genome Project for deciphering genes of the whole genome, the ultimate goal of the Human Proteome Project (HPP) is to make an inventory of all proteins by these genes. However, even the most convinced supporters of proteomics are in doubt that the inventory of all proteins might be made/achieved in the near future. Firstly, the methods available in proteomics do not allow the revelation of single copies of protein molecules in biomaterial. For such cases in genomics, the amplification by polymerase chain reaction (PCR) is used [1]. Secondly, if genome is mostly the same in each cell of the body, the protein composition changes essentially depending on biological fluid, type of the cells and tissues. Thirdly, there is a vast number both of translational and post-translational modifications of protein molecules, i.e. wide diversity of various protein forms are accounted for 1 gene.

Moreover, genome is a constant feature of an organism; it is generally invariable regardless of physiological or pathological conditions and its sequence does not depend on time. Proteome, on the contrary, depends on time, physiological stimulus and, thus, is situational [2]. The proteome composition is highly variable and, depending on time, the content of proteins and their modified versions change widely. The question whether the aforementioned problems are an impassable obstacle to start the work on the Human Proteome Project – is quite in order.

To answer this question it is worthwhile to draw an analogy with the Human Genome Project, roughly completed in 2001 [3]. That was an averaged genome of 6 individuals where few of expressed sites were revealed in its composition apart from regulatory elements [4]. However, big changes have taken place from then onward: a number of individual genomes are readily available, and thousands of them are expected in near future [5]. The individual peculiarities of genomes in diseased and healthy people have been analyzed to reveal splice-variants and single nucleotide polymorphisms (SNP). However, the work on genotyping is far from consummation since complete decoding of a whole genome of an individual is not fully applicable for wide-scale application. Still there exists a problem of proper genome assembling from multitude of sequenced DNA fragments. The function is somehow understood for 3-5% of genetic material available, essentially, by the expressed sites. The properties and assignment of the rest 95% of genome remain unstudied.

Actually, the above intricate situation with “completely decoded” human genome [6] makes it possible to answer the question whether the study of proteome should be started. Obviously, the problem of complete human genome study being yet incomparably easier than the proteome inventory, still remains undetermined in full. However, today one can speak about its resolution thanks to the Human Genome Program that was begun about 20 years ago. Just because of the necessity to create a similar long-term project, the specialists call to start the work on Human Proteome Project, despite clear comprehension that the task won't be settled in the near time.

Start of work on the Human Proteome Project will allow the participants to establish the targets and to set more or less common approaches to meet these targets. Presently, there is no common methodological approach to implement the complete proteome. The most straightforward solution is the so called "gene-centric approach" [7]. Presently the identification of proteins is carried out by matching the mass-spectra of biomaterial to the whole genome. Gene-centric approach targets each participant towards a certain portion of the whole genome, a chromosome, to perform the deep analysis of proteins, originated from this chromosome.

2. Gene Centric Approach

Actually, gene centric approach focused on investigation of proteins relatively to the genes is already in common use. Classical proteomics is based on the genome knowledge, thus, the transfer from the analysis of all genes to the analysis of a single chromosome is just the narrowing of a task compared to genome-based paradigm predominant today. Narrowing of research goal at expense of gene-centricity makes it more realistic by contrast with an alternative approach of tissue-based proteomics [8] comprising the analysis of proteins of separate organs and tissues based on the information on the whole genome. In comparison to gene centric approach, the inventory of a whole proteome in various organs and tissues is more ambitious challenge that seems unreal in the near future.

Application of gene centric approach will allow concentrating the efforts on solution of important technological issues. Firstly, it is necessary to resolve the problem of concentration sensitivity in analytical means of proteomics providing revelation of the low- and ultralow-copied[1] proteins in biomaterial [9]. In the absence of an analog of PCR for proteins, the concentration detection limit (DL) becomes a real challenge [10]. Just a small part of proteins represented in high/medium-copied range is currently available for proteomic inventory (Fig. 1). At the same time, it is reasonable to expect that early biomarkers of diseases will be present in low- and ultralow-copied ranges.

[1] Roadmap operates the ranges of protein concentration in biomaterial, which correspond to the detection limits of currently used analytical methods [10]. These are: high-copied ($\geq 10^{-6}$ M), medium-copied (10^{-7} -- 10^{-10} M), low-copied (10^{-11} -- 10^{-14}) and ultralow-copied ($< 10^{-14}$ M). "Copied" is used instead of "abundance" to emphasize the roadmap's technological focus on detection of single molecules.

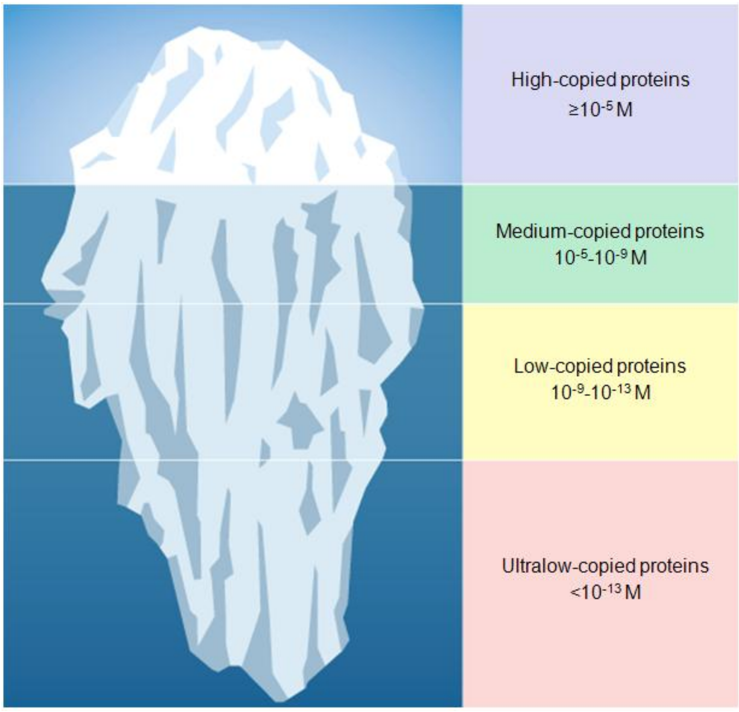


Fig. 1. Just a small “tip” of the proteome “iceberg” is accessible for currently available analytical methods of proteomics. The majority of proteins, including disease-specific biomarkers, are most likely situated in the “bottom” part of ultralow concentrations.

This roadmap is based on the concept of utilizing the nanotechnology to achieve the ultr-low copied proteins, buried at the bottom of the proteome iceberg. Nanotechnologies actively manifest themselves in genomics providing the possibility of working with single molecules of nucleic acids [11]. If similar approach could be applicable for proteins then the problem of concentration sensitivity in proteomics would be solved as well [10]. Maybe it would be possible to settle, at once, the related problem coming from situational proteome variability. By the opportunity to analyze the cells and bio-fluids with a several tens of protein copies being detected, the proteome dynamics will be expressed as the dependency of copies against time, and, most likely, it will also involve the emergence of new proteins and/or their modifications.

Gene centric approach enables to perform the systematic large-scale comparison between the transcriptome and proteome. With regard to the diversity of SAP and AS, the analysis of transcripts of just a single chromosome seems to be more realistic than investigation of translation products on a wide genome scale. Gene centric correspondence of transcripts to proteins will allow identifying the correlation between the level of expression at mRNA and protein level.

Restricted by the products of a single chromosome, one may attack the problem of defining the interactions and interrelations of proteins, i.e. to study an interactome. It means that there would be made an analysis of the interactions of proteins of the chosen chromosome with all the rest proteins coded by the human genome.

Thus, the advantages of gene centric approach are obvious. Firstly, it makes possible the usage of traditional proteomic methods based on the results of analysis of complete genome and, secondly, it is the concentration of efforts of separate research groups on a clearly designated task of studying gene products of a single chromosome.

3. Implementation Schedule

Roadmap envisages the implementation of the HPP will take 8-10 years for a single chromosome. Implementation includes the pilot stage (3-4 years) and master stage (5-7 years). During the pilot stage, at least, one protein for each gene of the target human chromosome would be identified with rough estimate on the level of its expression and predominant modifications. Pilot stage will deliver knowledge on the distribution of proteins across selected tissues at the sensitivity level up to tens of molecules per cell [12]. The pilot stage of the proposed roadmap is comprised of genomic and proteomics activities. The former include:

- establishing the criteria for chromosome selection;
- deep sequencing of the selected chromosome, refinement of ORFs;
- data-mining for the determined transcripts for the genes of the chosen chromosome;
- experimental design for identification of transcripts in hepatocytes and cell line HepG2.

Proteomic activities of the pilot stage cover the milestones listed below:

- data-mining for proteins expressed by selected chromosome, justifying liver tissue and blood plasma as a biomaterial of choice
- standardization of transcriptomic and proteomic technologies, standardization of plasma proteome variability by studying samples from healthy volunteers;
- identification of high/medium-copied proteins by MS/MS, identification of low-copied proteins using MRM technology, identification of peptidome;
- proteotyping and proteogenomic profiling of proteins from the chosen chromosome in the liver tissue, cell line HepG2 and in blood plasma;
- definition of an interactome complement of the selected chromosome;
- creation of gene centric knowledgebase on proteins of selected chromosome.

The master stage of roadmap implementation is aimed to essentially increase the sensitivity of proteomic technologies and, through this, to increase the scope of information on the protein molecules encoded by the selected chromosome. Two milestones are ahead of the master phase:

- creation of hybrid technology based on coupling of atomic force microscopy and mass-spectrometry for identification of low- and ultralow-copied proteins;

- identification of proteins in concentration range from 10^{-13} to 10^{-18} M in blood plasma and at a level ~ 10 copies per liver cell.

4. Chromosome Selection

Due to gene centric concept, chromosome selection becomes an important prerequisite of the national HPP. From the general consideration, the following selection criteria were suggested to rank the chromosomes:

- number of protein-coding genes;
- abundance of genes that, according to the available literary data, are relevant to the diseases;
- lack of immunoglobulin-coding genes.

According to the immunogenetics resource data (IMGT) [13], total of 430 immunoglobulin-coding genes are scattered over the genome; about one-half of these genes are pseudogenes. The genes of immunoglobulins are non-uniformly distributed over chromosomes. Most of them are located at the chromosomes #2, #14 and #22. The chromosomes #3-7, 10-13, 17, 19-21 do not at all contain immunoglobulins; about 1-2 genes are found in chromosomes ##1, 8, 9 and #18 where in the chromosomes ##1, 9 and 18 the immunoglobulins are presented as pseudogenes.

On the average there are 921±443 protein-coding genes per 1 chromosome in a human genome. The chromosomes ##13, 18 and 21 containing by 353, 286 and 253 genes, respectively, differ in considerably less quantity of genes against the average. These particular chromosomes have become the top priority subjects of discussion on the HUPO HPP workshops. In particular, chromosome #13 was selected by the Korean scientists and chromosome #21 was proposed by the American and Canadian scientists (see Fig. 2). In view of preceding criterion, it should be noted that the genes of immunoglobulins are not present in the chromosomes #13 and #21, whereas in the chromosome #18 the only immunoglobulin IG-lambda is coded by pseudogene.

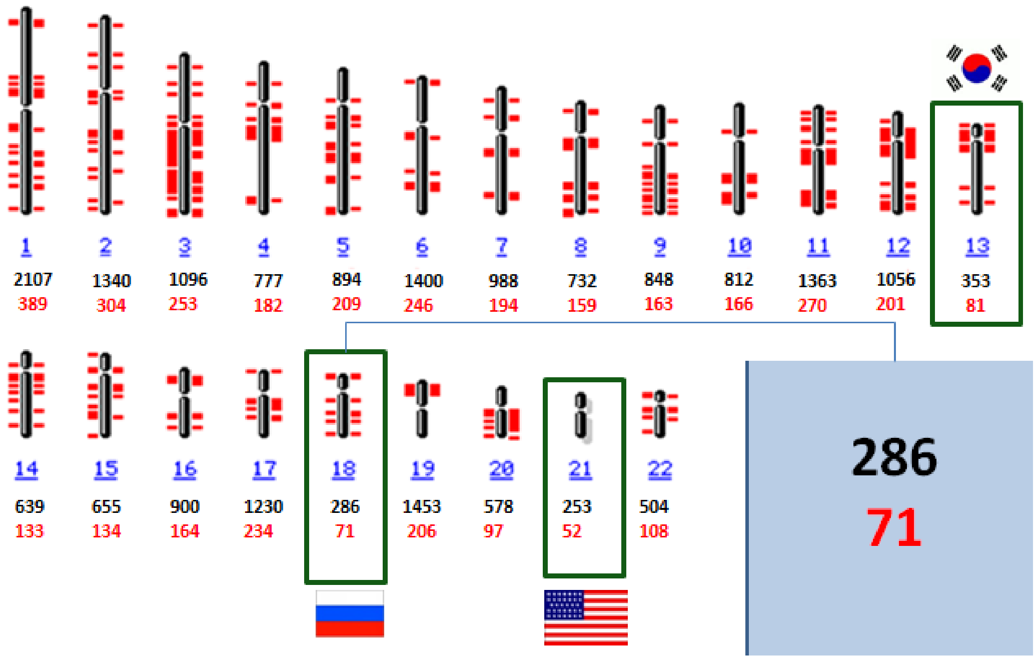


Fig. 2. The chromosomes are human genome are depicted. The numbers of autosomes are given in blue color beneath each chromosome, the number of protein-coding genes (in black) and the number of disease-related genes (in red).

Considering the work scope, chromosome #21 containing only 253 genes seems the most suitable target. However, this chromosome comes short of medical relevance that can be assessed as a proportion of disease-associated genes in the total quantity of genes in a chromosome. Association of genes and their respective products with the diseases was determined based on GeneCards data [14]. Table 1 presents the ratio of disease-associated genes in some chromosomes. In the average about 20% of protein-coding genes proved to be linked to diseases in one way or another. Chromosome #18 has the lead in the rating list, whereas chromosomes #13 and #21 are behind the 18th chromosome upon the present criterion about 2% and 4%, respectively.

Table 1. Total quantity of protein-coding genes (PCG) and the percentage (%) of disease-associated genes is given. The data for 10 chromosomes with maximum percentage values are shown. According to GeneCards data as at February 15, 2010.

Chr #	**18**	**4**	**5**	**3**	**13**	**2**	**8**	**22**	**14**	**21**
Number of PCG	286	777	894	1096	353	1340	732	504	639	253
%	24.8	23.4	23.4	23.1	22.9	22.7	21.7	21.4	20.8	20.6

In spite of considerable (more than fivefold, compare chromosome #2 and chromosome #21 in the table) differences in the quantity of genes being contained in human chromosomes, the percentage of disease-associated genes does not essentially differ (~3-4%). This observation is also true for other chromosomes not shown in the Table 1. A peer value of chromosomes relatively to the purposes of gene centric HPP urged Russian Working Group towards 18-th chromosome to manage the roadmap.

5. Chromosome 18

From a medical point of view, key feature of the 18th chromosome is high frequency of trisomy described as Edwards syndrome [15]. Children with trisomy of the 18th chromosome are born with a periodicity about once every 6000 and die predominantly aged up to 12 months because of numerous developmental diseases [16]. It may be assumed that overexpression of 18-chr-coded proteins lead to severe disturbances in the organism.

The genetic studies have indicated that the 18th chromosome is associated with a wide range of diseases (Fig. 3). Thus, an important chromosome mutation is a translocation between the 14th and 18th chromosomes leading to follicular lymphoma, with an overexpression of bc12 oncogene [17]. Relation to the oncological diseases is actually typical for the chosen chromosome. What stand out in Fig. 3 are the diseases like retinoblastoma, colorectal cancer, pancreatic cancer, synovial sarcoma, leukemia and squamous-cellular carcinoma. Another socially important focus of the 18th chromosome appears in relation to the diseases of the central neural system, including amyloid neuropathy, Parkinson disease and Schizophrenia.

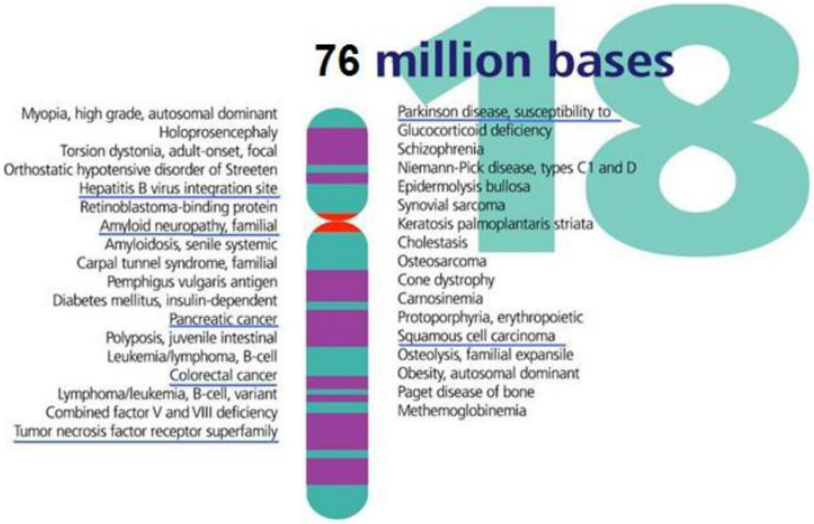

76 million bases
18
Myopia, high grade, autosomal dominant
Holoprosencephaly
Torsion dystonia, adult-onset, focal
Orthostatic hypotensive disorder of Streeten
Hepatitis B virus integration site
Retinoblastoma-binding protein
Amyloid neuropathy, familial
Amyloidosis, senile systemic
Carpal tunnel syndrome, familial
Pemphigus vulgaris antigen
Diabetes mellitus, insulin-dependent
Pancreatic cancer
Polyposis, juvenile intestinal
Leukemia/lymphoma, B-cell
Colorectal cancer
Lymphoma/leukemia, B-cell, variant
Combined factor V and VIII deficiency
Tumor necrosis factor receptor superfamily
Parkinson disease, susceptibility to
Glucocorticoid deficiency
Schizophrenia
Niemann-Pick disease, types C1 and D
Epidermolysis bullosa
Synovial sarcoma
Keratosis palmoplantaris striata
Cholestasis
Osteosarcoma
Cone dystrophy
Carnosinemia
Protoporphyria, erythropoietic
Squamous cell carcinoma
Osteolysis, familial expansile
Obesity, autosomal dominant
Paget disease of bone
Methemoglobinemia

Fig. 3. Diseases associated with the genetic aberrations in chromosome 18. Illustration is taken from the web-site of the Human Genome Project [18].

Chromosome 18 contains 76 million of base pairs [19]. In the average, for each gene of this chromosome there are by 3.1 variants of alternative splicing. The total number of genes is estimated at 480 (RefSeq, Feb 2010); however, as has already been stated, only 286 of them (ProteinAtlas, Feb 2010) code the potentially expressed proteins whereas the rest are either pseudogenes or code the RNA molecules. According to the UniProt [20], translation was confirmed by presence of respective cDNAs for 89% of genes of chromosome 18. Out of 286 proteins of the 18th chromosome, 104 (proteins) were revealed with immunohistochemical methods in tissue samples in the course of the ProteinAtlas project [21]. With proteomic methods 226 proteins of the chromosome under study were identified: mass spectra of these proteins are available in PRIDE database [22]. Out of all proteins coded by the 18th chromosome, 90 proteins are simultaneously revealed by immunohistochemical and mass-spectrometric methods, whereas 136 proteins are present exclusively in PRIDE, and 14 – exclusively in ProteinAtlas. The data on identification for 46 proteins is currently lacking.

6. Biomaterial: Liver Tissue and Blood Plasma

Although gene centric approach is generally tissue-independent, the selection of biomaterial is necessary to define the general guidelines for coordinated implementation of the roadmap. Unlike genome, the human proteome is tissue-specific; therefore, from the very beginning it is necessary to decide on the biomaterial.

Tissue/organ-based dimension of the roadmap provides the links between gene centric HPP and pre-existing HUPO initiatives. Among HUPO initiatives liver proteome project (HLPP, [23]) and plasma proteome project (HPPP, [24]) are most profoundly developed. The experience of these initiatives if inherited by the gene centric HPP establishes a steady springboard for the pilot phase. Other tissues should complement the master phase in case some chromosome-specific gene products still will be missing in plasma and liver even as the ultralow-copied instances.

Because of difficult access to a biomaterial of the liver, the hepatocellular carcinoma–derived cell lines are applied in proteomics. Some of them possess the features of hepatocytes, in particular, HepG2 line is a nonaggressive cell line that does not cause a tumor in nude mice. This line is used as a model of hepatocytes for investigation of their cytophysiological properties [25]. Proteomic analysis of HepG2 has revealed its substantial similarity with the primary culture of the hepatocytes [26]. Therefore, HepG2 line may serve as a reference biomaterial for investigation of the liver proteome.

The works on identification of the liver proteome are extensively carried on in Russia for the last 5 years. The technological approaches to identification of membrane proteins of microsomal liver fraction are elaborated [27]. Microheterogeneity of the cytochromes P450 in the liver tissue was thoroughly investigated using SDS-PAGE [28, 29]. Xenobiotic-induced profiles of differentially expressed proteins were determined in a series of 2DE-based experiments [30]. The sufficient level of activity in the area of proteome investigation indicates, that there exists established collaboration between clinical and research centers, which has to be an integral part of the roadmap.

Any tissue, including the liver tissue, is formed by the cells of several types. Proteome of a tissue sample is a complex mixture of proteomes of cell types that complicates the interpretation of the results of the proteins' identification. The present-day trends in proteomics of the human tissues call for the usage of the laser capture microdissection (LCM) for excision of homogeneous cell population. Application of this method in proteomics is at the start point because of the need for a large protein amounts for proteomic analysis as against the analysis of nucleic acid. In recent years, the improvement of microdissection methods makes it possible to pick up 2x10000 cells in a sample (1-4 mkg of protein) [31] that is compatible with requirements for protein identification by LC-MS/MS method. LCM was used for the separation of hepatocytes in proteomic investigations of the liver [32, 33], although the wide spread application of LCM in this area is severely hampered by the laboriousness of preparative protocol. Within the liver-oriented focus of the roadmap it is anticipated to enforce the LCM approaches.

On a protein level, liver is closely linked to blood plasma. There is high correlation between liver proteome and blood plasma proteome [34]. PRIDE reported that practically all proteins of the 18th chromosome revealed in the liver were also indentified in blood plasma. Just 3 of 18-chr-coded proteins occur exclusively in the liver, being absent in blood plasma and other organs. Vice versa, 216 proteins of 18-th chromosome were identified in blood plasma, of which 110 were also present in the liver tissue.

Blood is a kind of a collector containing all proteins of a human. Plasma proteome, in fact, includes proteomes of other tissues as a subset. Blood plasma is a pioneer area of focus in the human proteomics [35]. The main objective in plasma investigation was the application of the results at medical diagnostics. Blood plasma to the utmost answers the purposes imposed upon the target of clinical molecular research. It is the most accessible for minimally invasive selection by the human tissue. Compared to more available biological fluids (saliva, urine), plasma, as a part of blood, is homeostatic to a greater extent, i.e. it is characterized by the generally invariable proportions of high-copied proteins.

At a pilot stage of the Human Plasma Proteome Project (HPPP) initiative, standard samples of the human blood plasma were analyzed with different proteomic methods in dozen laboratories [36]. The main problem in the proteomic analysis of blood plasma is vast dynamic range of protein concentration (~10 fold difference of concentrations). Most of the methods used for the analysis of unfractioned plasma make possible the revelation of only few major proteins, whereas no consideration has been given to the low-copied proteins significant for medical diagnostics [37].

Fig. 4 demonstrates the diagram reflecting current visions on the structure of blood plasma proteome. It is seen that nearly 60% of blood plasma-identified proteins are functioning in cells, while "classical" plasma proteins come to only 13% and immunoglobulins account for 8%. Of cellular proteins, 53% are localized in cytoplasm, 9% - in the core nucleus and 19% - in the membrane, whereas the secreted proteins' fraction comprises less than 14%. These data show that plasma proteome contains the proteins relevant to the molecular processes in cells and, hence, these proteins can be selected as the candidate disease biomarkers.

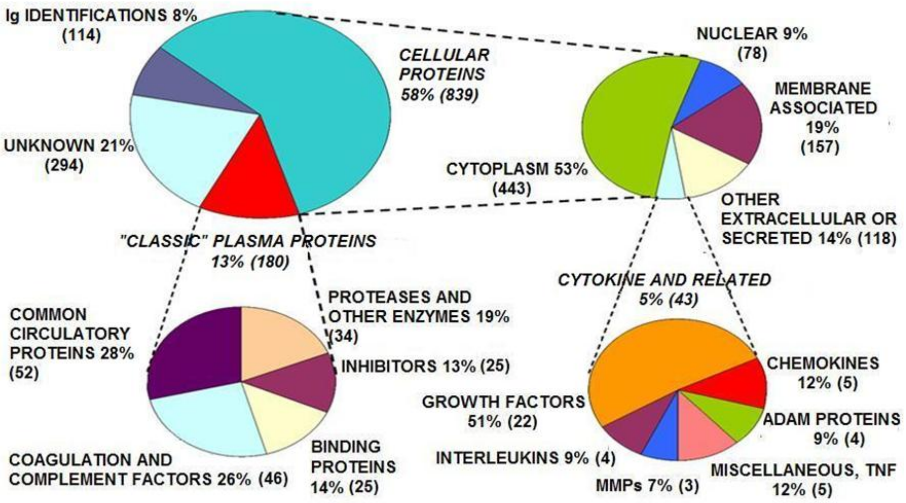


Fig. 4. "No plasma – no money" - S.Hanash, HUPO HPP Meeting, 2009. Grouping of identified blood plasma proteins according to [38], [39].

Despite technological complications, there is no doubt that blood plasma is a biomaterial of choice for the gene centric HPP. One should not be puzzled by such transformation of gene centric paradigm to the case of specific type of biomaterial. It should be emphasized that plasma proteome might contain subproteomes of various organs and tissues which presently could not be detected because of low sensitivity of the current proteomic technologies.

To establish the starting point for plasma proteome analysis we referred to the list of 3020 proteins, which came up as a result of a pilot phase of HPP [24] (1175 protein identifications from this list were independently confirmed by 4 analytical methods [40]). In HPPP, protein was interpreted as a nonredundant product of one gene without regard to modifications and processing.

Gene centric approach requires the distribution of the identified proteins in pursuant to a chromosome they are encoded by. Plasma-identified proteins are uniformly distributed through the chromosomes, *i.e.* the more genes present in a chromosome, the more proteins coded by these genes has been detected in blood plasma (Fig. 5).

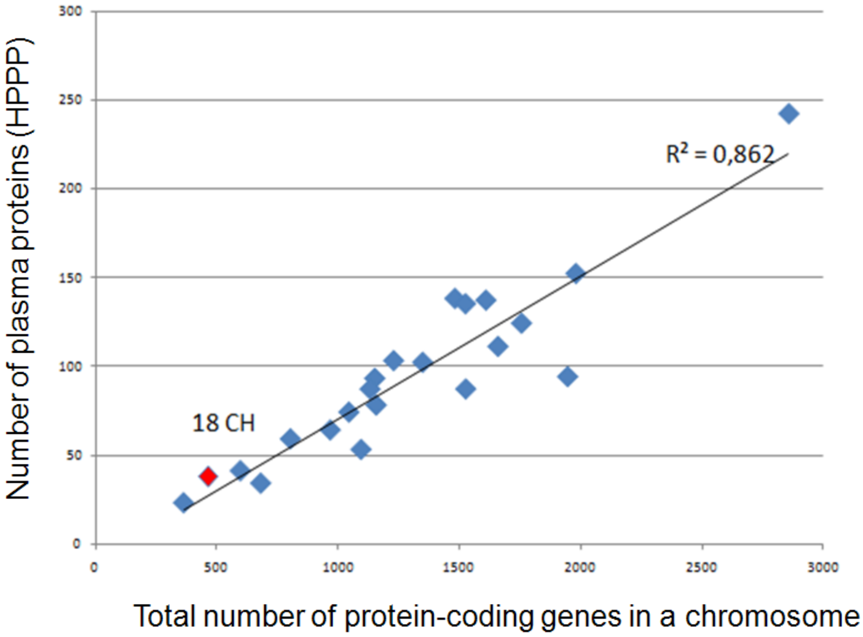


Fig. 5. The relationship between the quantity of plasma-identified proteins and the total number of protein-coding genes in a chromosome. The point corresponding to the 18th chromosome is indicated as red diamond. Plasma proteins are taken from [41].

Fig. 5 presents the mapping on the human chromosomes the information on proteins identified in blood plasma. High correlation (R^2 ~0.9) between the number of protein-coding genes in a chromosome and the number of plasma-identified proteins is observed demonstrating that each chromosome is coding the blood plasma proteins proportionally to the number of its genes. It means that plasma in reference to the chromosomes may also be characterized as “collector”.

All the above considered, it must be speculated that plasma as a biomaterial could play the same crucial part in HPP as the blood cells did in HGP. For this purpose, proteomics is in need of new technologies for identification of low- and ultralow-copied proteins. As to high- and medium-copied proteins, the issues of standardization of high-throughput experiment become topical.

7. Standardization and Proteotyping

The methods used in proteomics for identification of high/medium-copied proteins are presently based on the liquid chromatography coupled with the tandem mass spectrometry (LC-MS/MS). The question of inter-lab reproducibility of this method is still an open issue as the results of identification are highly dependent on over 46 subtle technical parameters [42]. The diversity of results observed in the analysis of similar samples made in different laboratories is avoidable with the method of standardization. For this purpose, independent investigations were carried out in different scientific centers: the same standard samples both of artificially obtained mixtures of recombinant proteins [43] and the proteins' mixture of yeast lysate [44] were characterized with LC-MS/MS method. The differences in a set of proteins of one test sample identified by different researchers were based not only on the reasons involved with technical procedure but, to a large extent, on the algorithms of mass-spectral data analysis [44]. The recent publications related to the subject make clear that for the goals of HPP the state-of-the-art mass-spectrometric platforms should routinely undergo tests using the standard samples.

One of the suitable standards for multicentral studies is the complex proteomic standard developed by Agilent Inc. (USA) in collaboration with the leading laboratories in the field of proteomics. This set specially designed for testing LC-MS/MS systems contains the extract of proteins of standard bacteria strain *Pyrococcus furiosus* (Pfu). The quantity of proteins contained in this reagent exceeds 1500 that keeps in complexity with the samples coming from a human organism. Moreover, the advantage of Pfu standard is a low similarity degree of protein sequences between a microorganism and a human. Pfu reagent is proposed as a standard for comparison and quality control of LC-MS/MS technology set up in the laboratories-participants of the Russian HPP roadmap.

Standard peptides are used for the quantitative analysis of proteins with mass spectrometry that is carried out through multiple reaction monitoring (MRM). The artificially produced peptide standards are mixed with bio-sample or probed separately to adjust MS regimens [45]. The usage of standards in analysis essentially reduces the variability of the results. Thus, in multi-site research of the same samples a high degree of reliability of the results of MRM-analysis was shown [46]. Therefore, for the usage of MRM-analysis within the project one must adhere to standard protocols. In course of roadmap implementation the artificially created standardized protein mixtures are proposed to perform cross-laboratory evaluation of quantitative assignments.

Characteristic property of biological studies, in particular, the analysis of proteome, is the fact that a measurement error is due not only to the instrumentation and technological aspects but also to the individual variability of an investigated object. Revelation of such variability is the essence of proteotyping problem that is recently given increasingly much attention [47]. The proposed national HPP roadmap contributes to the determination of individual variability of proteomes of the blood plasma and liver.

Revelation of statistical variation range of the level of blood plasma proteins is an essential provision for definition of the disease biomarker [48]. Determination of plasma proteome variability will require the analysis of individual differences in healthy people. It requires recruiting the volunteers whose health status corresponds to the existing standards that is affirmed by the health authorities. Within the roadmap scope there will be made a statistical analysis of concentration variability and the reference intervals for plasma proteins coded by the 18th chromosome will be estimated.

Determination the reference values of the protein expression levels in liver is also worthwhile, however the peculiar format of respective experiments is assigned to the master stage of the roadmap. Cholecystectomy and hemangioma resection are most frequent surgery cases to obtain the conditionally normal liver samples.

Thus, standardization in the HPP should be aimed, on the one hand, at provision of reference operational modes of high-throughput equipment and, on the other hand, at the definition of permissible standards of inter-individual differences of a proteome.

Proteotyping as another issue related to the standardization of the proteome research. In this case the standards has to be applied to the definition of protein structure, as a product of single amino acid polymorphisms (SAPs, [49]), alternative splicing (AS, [50]) and post-translational modifications. In response to the problem of protein structural microheterogeneity two approaches are under consideration. The first “top-down” approach consists in precise measurement of the mass of whole proteins [49],[51], herein the shift of measured mass against the theoretical mass is related to the pattern of single-aminoacid polymorphisms and post-translational modifications. High accuracy that is required for the registration of such indicative shifts in mass spectra is provided by the up-to-date devices like OrbiTrap or FT-ICR. However, on registration of the whole protein molecules with high molecular mass the sensitivity level is not high: about 10^{-5} – 10^{-6} M [52]. Low sensitivity gives no way of applying the “top-down” approach for pursuing the announced roadmap goals.

The sensitivity of proteotyping could be increased using the “bottom-up” approach using conventional mass-spectrometric analysis of proteolytic peptides. Herewith, the sensitivity increases by 4-5 orders and reaches the level 10^{-8} – 10^{-9} M or even 10^{-12} M in MRM mode.

During the pilot phase of HPP roadmap it is planned to use an iterative algorithm of identification of proteotyping peptides [53]. According to this algorithm, first the identification of proteins versus complete genome database is made. Further, the variants of sequences containing SAP, AS and putative PTMs for identified proteins are included into the database. On the database extension the repeated identification is made, which yields information on the presence of microheterogeneic variants of proteins in a sample. The identification of proteotypic peptide containing SAPs, AS or PTMs is followed in MRM mode (see below).

Based on information available in current databases it is possible to evaluate roughly the amount of microheterogeneic variants of the proteins relevant to the 18th chromosome. For the genes of the 18th chromosome there are 650`220 variants of nucleotide polymorphims (according to National Center for Biotechnology Information [54]) of which 2386 are nonsynonymic (nsSNP) and can be potentially observed as SAPs.

For the proteins of the 18th chromosome UniProt database contains information on 796 single aminoacid polymorhisms. Overwhelmingly, they are putative SAPs the availability of which one would expect based on the data on nsSNPs in relevant genes. On average, 4 SAPs would be expected per protein. Among all protein-coding genes of the 18th chromosome the greatest quantity – 168 SAPs fall on the gene from Niemann-Pick type C (NPC) family. There are 168 sites of single aminoacid polymorphism which gives rise to 2^{168}~$3.74x10^{50}$ of combinatorial variants of primary structure in population. The functions of NPC are linked to intracellular cholesterol transport; the defects of this gene are associated with the development of sphingomyelinosis (Niemann-Pick disease). It should be noted that in our experiments the NPC gene had been identified through the analysis of proteome of a cell line HepG2.

The data on AS and PTMs have also been estimated using information available in UniProt. It appeared that for 120 proteins of the 18th chromosome at least 1 splice isoform is known. According to UniProt, alternative splicing of this chromosome delivers in the average 2.81 protein products per gene (that is not different from the average statistical data on the whole genome) and the greatest variety of the isoforms is observed for NFATCI gene (9 isoforms). Post-translational modifications were deposited in UniProt for approx. 60% of proteins coded by the 18th chromosome. As usual the most frequently observable is phosphorylation (112 proteins), glycosylation (53 proteins) and acetylation (36 proteins). UniProt database reports, on average, 4 modifications for a protein, herewith, among proteins with the known function the most modifications (23, including 21 sites of phosphorylation of serine and threonine and 2 sites of acetylation of lysine and methionine) is known for the regulatory protein with identifier Q9Y212 involved in meningioma pathogenesis.

Of the above averages one might give minimal estimation of a quantity of protein isoforms originating from the 18th chromosome. Assuming that for each of 286 protein-coding genes there are, on average, 4 single aminoacid polymorphisms, 2.81 variants of alternative splicing and by 4 post-translational modifications (ignoring combinatorial variants) we obtain N=286x(4+2.81+4)~3046 of structurally different isoforms. That is, by the most conservative estimate, the scope of work that should be done during implementation of the 18-th chromosome roadmap.

8. Analysis of Transcripts and High/Medium-copied Proteins

To illustrate the possibilities and limitations of existing approaches readily available for implementing HPP, we performed an analysis of asynchronous cell line of hepatocytes HepG2, samples of the liver tissue of healthy donors and samples of blood plasma. The analysis of cells and tissue was carried out at the transcriptome level with the application of full-genome arrays from Agilent Inc. At proteome level the biomaterial was analyzed using multi-dimensional chromatography coupled with MS/MS protein identification.

By means of whole genome transcriptional microarrays, the expression of 10`865 genes was revealed in cell line (347 genes of this amount referred to the 18th chromosome). In Ensembl database [55] only 179 proteins conformed to the identifiers of these genes, whereas the others constitute the temporary experimental titles of fragments of unknown genes.

Inspecting GeneOntology revealed, that the 18-th chromosome assigned transcripts take part in 30 various biochemical processed. Respective genes mostly contribute to the regulation of transcription (24 proteins) and encode transport proteins (19 proteins). Localization of the products of transcribed genes is considerably diverse: membrane proteins make up about 1/3 of the expressed transcripts (53), the nuclear ones are presented by 43 genes and 37 proteins are arranged within the cytoplasm. According to UniProt assignments, only 22 of the obtained proteins relate to a progression of a disease. Part of these diseases is associated with nsSNPs which are able to be translated as SAPs (see above).

A correlation between transcriptome of cell line and transcriptome of the liver tissue in healthy people, obtained during cholecystectomy, revealed that more than 73% of active genes agree. So, cell line is an adequate model representing basic physiological processes of the liver tissue. If the transcripts of the 18th chromosome are considered exclusively, then the level of concordance is slightly below – about 62%.

Among the advantages of cell line are, certainly, simplicity in generation of a material in sufficient quantities and high degree of experimental standardization. On the other hand, these cells are cancerous with functional features being partially lost in cultivation process. An evident disadvantage in the work with liver samples is a complexity in obtaining a reasonable quantity of samples, a need for separation of homogenous population of a tissue cells, and an individual distinctions between samples.

The complementary proteomics analysis of aforementioned biosamples was performed using 3D separation method [27]. The proteins were separated on a reversed-phase column to 5 fractions, each of which then was hydrolysed and analyzed with MudPIT technology [56]. We have identified over 4100 proteins of HepG2, 91 of which belong to the 18th chromosome. Coupling these identifications to the “Natural Variants” section of respective UniProt records we found, that 17 proteins carry disease-associated mutations, including cardiomyopathy, cataract, cancer and previously mentioned Niemann-Pick disease.

Some of identified proteins were not revealed on a previous stage of a whole genome transcripts investigation. Transcripts of 17 genes were not indicated above the microarray intensity threshold, nevertheless, their corresponding protein products were identified in a proteome with significant score. Such discordance between transcriptome and proteome data is observed quite often [57].

To confirm this observation, we examined differences in protein abundance using label-free spectrum counting approach [58] with tandem mass spectrometry (MS/MS). Figure 6 shows protein-abundance *vs* corresponding mRNA-expression intensities. Protein-abundances (number of peptides participated in protein identification) were not correlated with their mRNA absolute fluorescent intensities ($R^2 = 0.02$). This may indicate a low stability of mRNA compared to the proteins' molecules as well as technical reasons should not be ruled out: dependence of the signal level on the quantity and specificity of on-chip probes for a certain gene [59].

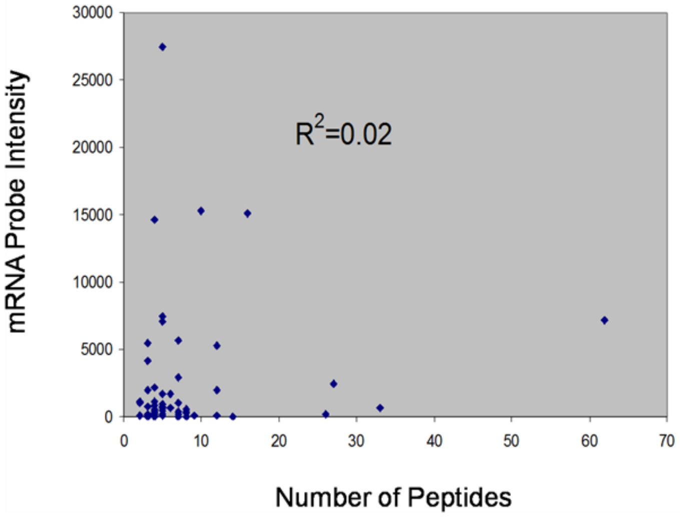


Fig 6. Scatter plot of peptide counter versus corresponding mRNA expression level for the transcripts/proteins of 18-th chromosome.

Using MudPIT platform, the analysis of protein composition of blood plasma was also carried out. Among 2000 of identified protein products we revealed 66 proteins that are encoded by genes of the 18th chromosome.

The analysis of the 18th chromosome's transcripts and proteins encoded by them had shown that sufficient intersection between mRNA and protein data enables the realization of gene centric approach. The discrepancies, which are observed in Fig. 6, might also root themselves in the metrological incorrectness of comparing the microarray signal intensity with the peptide abundance index. The roadmap has to solve such type of incorrectness by expressing content as a number of copies of biomolecules. The rationale behind counting individual molecules instead of measuring their concentrations rests upon the concept of reverse Avogadro number [10]. Today we already witness this concept gaining strength in the technology of single-molecular DNA sequencing [60]. Within the roadmap the reverse Avogadro number is utilized to bridge the conventional proteomics to the nanoproteomics [9],[61].

Proteogenomic profiling. Correlation of transcriptome with proteome makes it possible to perform the feed-back formatting of a genome. Reverse genomics (proteogenomic typing) is the task of analyzing the splice variants at the protein level to decipher the true boundaries of coding frames along the genome and to elucidate the unannotated genes [62]. For proteogenomic typing important is the information on not only the whole proteins but also on their low-molecular fragments – peptides.

9. Peptidome-based Drug Design

The localization and identification of possible “peptide” gene products (proteins with a mass less 10 kDa) and, also, the peptide families being the products of native and pathological degradation of the proteins by proteolytic enzymes is, in practice, a “white spot” at the gene annotation. A great many of the degradation products might possess physiological activity.

Of special interest is the identification of peptides being the products of protein degradation and revelation of their specific biological and marker activity. The roadmap envisages the peptidome as additional information that is assembled according to the gene centric principle on a chromosome-based scaffold. The biological activity of the peptides is considered as a proteomic-driven entry point for elaboration of the drug prototypes [63].

For systemic characterization of a peptidome it is necessary to elaborate the standard methods of peptide extraction and desorption from the major tissue and plasma proteins. The methods of extraction should provide the reproducibility of the results of structural-functional analysis using the methods of high accuracy mass spectrometry. Although any peptides detected in serum or blood plasma are of interest in such experiments, there is the possibility of their selective identification and identification of pre-proteins encoded by the 18th chromosome. The methods of selective mass-spectrometric identification allow the targeted searching of the given peptide fragments at a low-abundance range of concentrations.

10. Detection of Low/Ultralow-copied Proteins

The roadmap was built upon the axiom, that the main problem in the analysis of biological material is the poor sensitivity of the current proteomic technologies. Even most sophisticated affinity-powered protein identification methods can hardly attain the level below 10^{-12} M. Consequently most of the proteomic analyses of biological material allow detecting just a few hundreds of high/medium abundance proteins while the low abundance proteins remain inaccessible for the investigation (Fig. 1). The problem of identification of low abundance proteins is, obviously, the basic technical challenge which should be overcome within the HPP.

Multiple reaction monitoring (MRM, [45]) is a mass-spectrometric method with the high reliability of proteins and peptides detection. MRM provides high sensitivity and wide dynamic range of protein concentrations not provided by other MS-based methods. According to the published data, the sensitivity this method is in the range 10^{-12}-10^{-14} M for yeast [12] (which makes few hundred copies of the proteins per cell).

To use the MRM method it is necessary to define a set of proteotypic peptides, which will undergo monitoring. Physical and chemical properties of proteotypic peptide should enable efficient ionization and show a characteristic mass-spectrum with well pronounced peaks. To identify a certain protein it is necessary that proteotypic peptide would not only be well ionized, but also it should be unique, i.e. not found in the proteins encoded by other genes.

One may expect that application of MRM in combination with the up-to-date methods of protein fractionation will make it possible to reach the limit of detection of low-abundance proteins, *i.e.* to reach the concentration level 10^{-15} M. Further increase of sensitivity may be based on application of the method of irreversible protein fishing from biological material on the surface of chemically activated biochip. Registration of proteins fished on the surface will be carried out with the aid of molecular detector, for example, atomic force microscope and, at the same time, mass-spectrometric identification of detected molecules will be performed [64].

Irreversible fishing enables the concentration of low/ultralow-copied proteins on a biochip with chemically activated groups through which a covalent protein immobilization occurs. Theoretical estimates show that the protein concentration on the surface of the chip using irreversible fishing can increase to 8 orders [65]. This means that fished low-copied proteins will be turned into the category of medium-copied and ultralow-copied proteins - to the category of low-copied proteins.

The usage of atomic force microscope as a molecular detector allows counting the number of protein molecules immobilized on the biochip after fishing [66],[67]. As the AFM sensitivity is at the level of single molecules, the process of molecules’ registration is no longer a limiting step in proteome analysis technology [65]. Combining AFM with mass-spectrometry will allow not only the visualization and counting the protein molecules, but their identification as well.

Despite the fact that fishing in combination with molecular detection is a prospective line in technological development of proteomics, there is a crucial limitation of atomic force microscopy (AFM) - low speed of surface scanning. Most rapid commercial models operate at the 1-10 Hz scan rate, which takes approx. 3 hours to visualize the small area of 400 um^2. The progress in this direction is connected to the increase of the scan rate. One of the achievements indicates that the scan rate can be increased 1000 fold, up to 1 kHz [68]. That reduces the scanning time of a typical area of 400 um^2 down to 10 sec. In addition, the achieved scanning rate enables to observe the motion of the molecules in real time. It means that in addition to the size and shape, AFM will possess an option to identify the protein specifically by the dynamic changes of its conformation. In future, the real-time AFM will be able to capture the situational changes of the proteome by monitoring the formation and decay of functional complexes addressing the intractome layer.

11. Interactomics

Investigation of human interactome provides new level of understanding of molecular mechanisms and machines in normal and pathological cases giving great impact for treatment of diseases. Some vision of the chromosome-centered approach for untangling the interactome is provided below.

On the pilot phase of roadmap implementation the database mining will be performed to elucidate interactions of proteins from chromosome 18. The data on protein-protein interactions will be fetched from APID2NET resource [69], which unites information from DIP, MINT, BioGRID, IntAct, HRPD and BIND. The information on metabolic protein interactions will be accessed through KEGG [70]. The protein networks will be ranked according to their relevance to the disease focuses of the roadmap (see Section "Chromosome 18"). The selected batches of interacting proteins will be examined through existing collections of mass-spectra to determine the reliable identifications. From such inventory we would expect a list of proteotypic peptides for the proteins residing aside chromosome 18, but interacting within the 18-chr-related pathways and complexes. Such proteotypic peptides will be included to the MRM procedure.

Genes of chromosome 18th have variability of Interpro domains. Totally 231 proteins are assigned to the Interpro domains, of which 33 Zinc-finger C2H2-type domains and 51 Cadherin and Cadherin-like domains, 11 proteins having protease inhibitor, 14 serpin domains and 136 domain of other types. Such data gives evidence that many of chromosome 18 proteins have similar domain architecture. This information could be used for filtering of known and predicted interactions, for example, proteins with the same GO annotation or/and with the same domains, expressed in the same tissues could share similar partners.

KEGG database contains 60 proteins of 18th chromosome associated with signaling and metabolic pathways. It was observed, that not more than 3-4 proteins per pathway showing that chromosome 18th proteins are randomly scattered over various cellular processes. Fig. 7 shows an example for the fragment of TGF signaling pathway. There are four proteins coded by 18-th chromosome genes involved to this pathway, namely Smad4, Smad2, Smad7 and ROCK. Smad7 inhibits Smad2 and Smad4 and Smad2 assemble the transcription complex, which activates pathways of cell growth, migration, apoptosis pathways [71],[72]. In medical applications this pathway is mainly considered in relation to the cancer of such origin as lung, breast and prostate [73].

According to PRIDE the protein identifications exists for 82 proteins of the depicted pathway. These proteins were identified in 5-10 experiments with blood plasma, however, in most of the experiments only a single proteotypic peptide is reported and no mass-spectra details are available. In addition to the products of 18-chromosome the TGF pathway indicates the partners from other chromosome, of which, for example, SMURF2 (Q9HAU4, chromosome#17) and SMURF1 (Q9HCE7, chromosome #7) were also found in PRIDE. However, for some proteins of TGF pathway (*e.g.* Q96S42 or Q6GTN3) there is no MS evidence, therefore such proteins have to be identified in the course of interactomics-related roadmap activities.

The master stage of the roadmap comprises the experimental design to decipher the new interactions involving the proteins translated form 18-th chromosome. First the *in silico* elucidation of putative protein-protein interactions will be performed using data-mining and predictive algorithms like sequence local similarity of known interacting partners [74] and comparative genomics of protein phylogenetic profiles [75],[76].

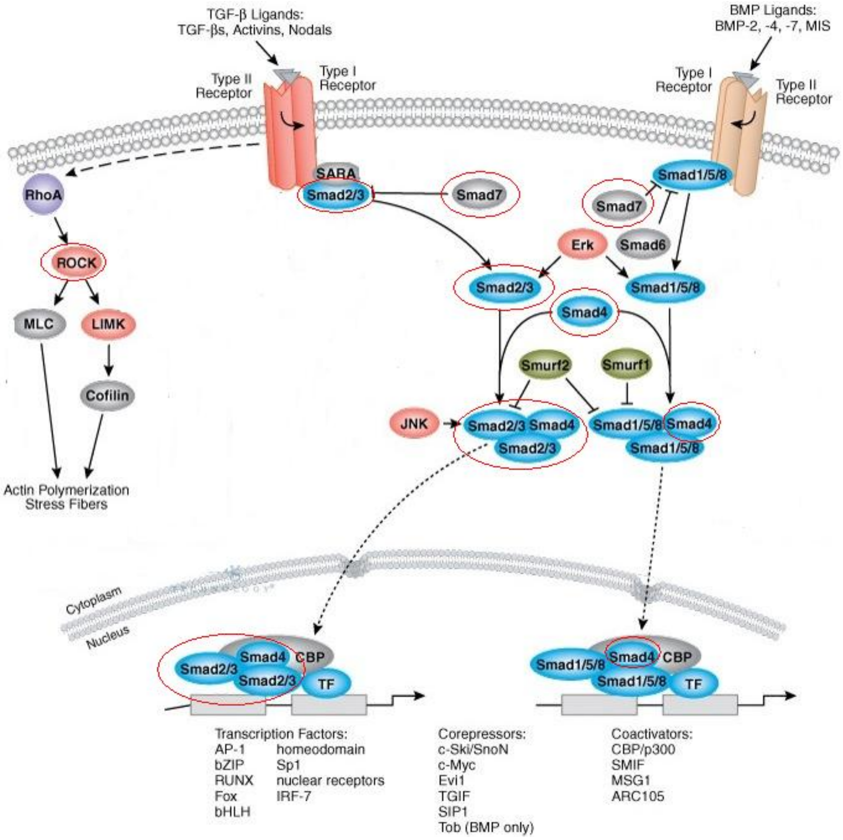


Fig. 7. Fragment of TGF-beta signaling pathway involving proteins of chromosome 18 origin (shown by red ovals).

Combination of *in silico* and experimental methods gives opportunity for validation known interaction in different experimental conditions and for validation *in silico* predicted novel protein-protein interactions for building reliable protein-protein networks of 18th chromosome proteins.

Plasmon resonance biosensor method in combination with mass-spectrometry identification of proteins provides high quality measurements, the protocol is divided into three main steps [77],[78]: (1) extraction of target protein in complex with protein partners on chip of optical biosensor after injection of cell lysate (or its chromatography fraction) – molecular fishing, (2) elution of extracted proteins from optical chip; (3) mass-spectrometry analysis and identification of proteins. Assembly of experimental facts about protein interactions and identifications, their relation to the different experimental conditions and pathological processes will comprise the information to be shared through the knowledgebase interface.

12. Knowledgebase

Databases were used as a means of informational support of Human Genome Project [79]. Great complexity of the proteome compared to genome surprises one to provide the software not only for the data storage but for its analytical processing as well.

Within the roadmap it is offered to create the gene-centric knowledgebase on the proteins of the 18th chromosome. In case the approach will be successful it could be applied to other chromosomes as well.

As shown in Fig. 8 the knowledgebase functions as mediator between roadmap activities and public repositories, necessary for implementing the molecular biology work. Users of the knowledgebase interrogate public resources through the interface, which accumulates the so-called strategic knowledge [80]. Accumulated knowledge is spread over the HPP consortia to foster the intelligent solutions within a changing environment [81].

At the top of generalization the knowledgebase computes the indices, which enable to monitor the status of roadmap implementation. The first index estimates the medical relevance of the implemented activities. Such index has been already discussed while selecting the chromosome. Using the knowledgebase, the information extracted through data-/text-mining is matched against the newly identified proteins and peptides. The index value also depends upon the increase of sequence coverage of disease-associated proteins, upon the number of proteotypic SAP-containing peptides with putative diagnostic utility, etc.

Second index indicates the overall progress roughly as a percentage of protein-coded genes assigned to the appropriate MS data. It depends on the accuracy of collected mass-spectra and also on the abundance of protein identifications across different independent experiments.

Third block of the knowledgebase deals with the technology focus of the roadmap. The relevant activities comprise the estimation of the proportion of the ultralow-copied proteins of the targeted chromosome.

The implementation of knowledgebase will be performed on the basics of the text-mining algorithms, many of which are actively utilized in biomedical research. Information management technologies based on explicit analysis of MeSH terms [82],[83] and protein bibliography mapping [84] are expected to comprise the core of the knowledgebase. With a goal to deliver the individualized content knowledgebases would create demand for the personal supercomputing [85], as previously Human Genome Project promoted the cluster computing.

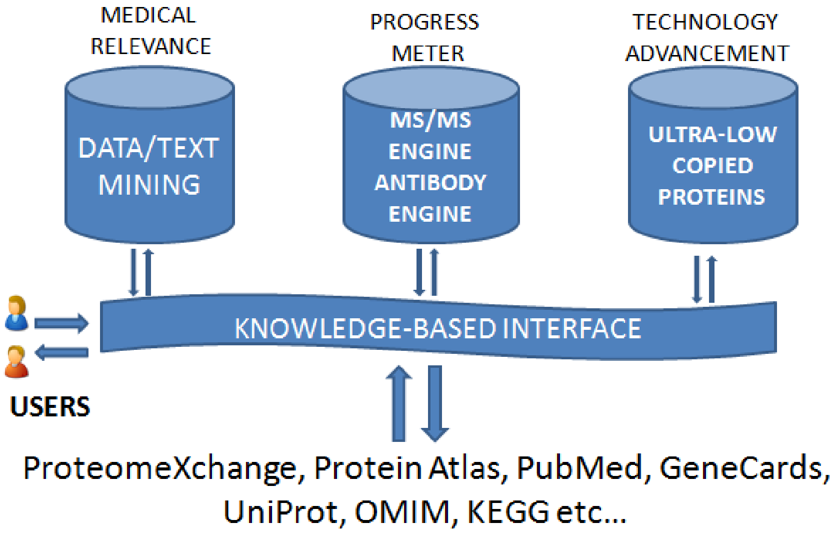

MEDICAL RELEVANCE
DATA/TEXT MINING
PROGRESS METER
MS/MS ENGINE ANTIBODY ENGINE
TECHNOLOGY ADVANCEMENT
ULTRA-LOW COPIED PROTEINS
KNOWLEDGE-BASED INTERFACE
USERS
ProteomeXchange, Protein Atlas, PubMed, GeneCards, UniProt, OMIM, KEGG etc...

Fig. 8. Knowledgebase concept for Human Proteome Project.

13. Short-term Deliverables (5-year view)

The pilot stage if roadmap implementation, focused upon the inventory of high/medium-copied proteins and peptides of the 18th chromosome, will deliver in 3-years:

(i) at least one protein product for each gene of 18-th chromosome with at least 1 proteotypic peptide;

(ii) list of frequently observed modifications, splice variants and single-aminoacid polymorphisms;

(iii) ranking of proteins assigned to 18-th chromosome according to their medical relevance (biomarkers and/or drug targets) and frequency of identification in plasma and liver cells (HepG2 cells);

(iv) reference ranges of concentrations of high/medium abundant proteins in norm;

(v) standardization protocols for proteomic procedures;

(vi) ranking of protein interactions according to their medical relevance and accessibility to mass-spectrometry identifications;

(vii) experimental protocol for chromosome-centered interactomics.

14. Long-term Deliverables (15 year view)

Having resolved the problem of detection of low/ultralow-copied proteins there will be created a fundamentally new method for registration of biomolecules with the usage of nanotechnologies. Based on a hybrid technology, there will be produced industrial complexes combining the capacities of mass spectrometry and atomic force microscopy. One may objectively predict that the sensitivity will reach the level 10^{-18} M in blood plasma and 10-100 copies of protein molecule for 1 cell of the liver tissue. Application of hybrid complexes for the scientific purposes will allow the revelation of the early biomarkers directly associated with the development of pathological processes.

At the phase of practical application, there will emerge the prototypes of molecular sensors based on the nanowires [86]. These devices will cause sharp cheapening and large scale distribution of the multiparameteric medical tests. In their reliability level the devices will be applicable for point-of-care diagnostics. Technological achievements of the master phase of the roadmap will make it possible to proceed to monitoring diagnostics of the changes of molecular systems of an organism which will provide the scientific basis for personalized medicine.

The schema in Fig. 9 provides the overview of major opportunities build into the structure of the roadmap. A starting point is pointed by two already elaborated laboratory prototypes: atomic-force microscope coupled with MS and nanowire sensor. The roadmap vector evolves from the laboratory prototypes towards the clinical application through discovery of new biomarkers. Further in this direction the point-of-care devices are foreseen, which creates the technological environment for the personalized and decentralized medicine.

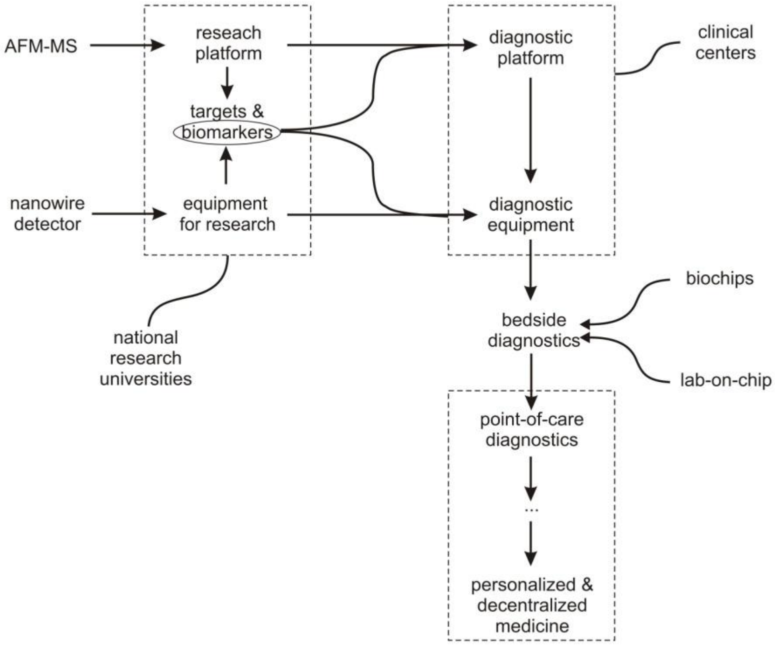


Fig. 9. Long-term expectations from the gene centric Human Proteome Project.

15. Concluding Remarks

The scope of the HPP requires the distribution of work between participants. In the same way that it has been done for the “Human genome” project, it is suggested to distribute the work based on the gene centric approach [7], i.e. to identify protein products of genes according to their distribution over chromosomes.

Gene centric approach, unlike the formerly used tissue/disease-based approaches, enables the organization of proteome study as a roadmap. The distinction of the roadmap from research plan is that according to the roadmap the scientific problems are resolved taken in conjunction with the advances in analytical technologies that, in turn, create basics for the development of translational medicine.

Roadmap, elaborated by the Russian scientists is focused on the assembling the protein parts list coded by the genes of the 18th chromosome. This is comparatively tiny chromosome, which approx. 80% of coded proteins already identified.

The present roadmap envisages the pilot and the master phases. During 3-year pilot phase it is planned to identify the proteins of the 18th chromosome in three types of biological material: blood plasma, cell culture HepG2 and liver tissue. The goal of pilot phase is to identify, at least, one protein for each gene, to determine the level of its expression and predominant modifications. The results of implementation of the pilot phase will be the data on individual variability of proteome of blood plasma and liver tissue.

The master phase of roadmap will be realized in 5 years; it is involved with experimental revelation of modifications for all proteins of the 18th chromosome. Modifications include single aminoacid polymorphisms, the products of alternative splicing and post-translational modifications.

The roadmap puts forth several tasks that span over both phases of roadmap implementation:

- genome/transcriptome analysis using single-molecular DNA/RNA readers [60] to get more specific gene-coding sequence of the 18th chromosome and to elucidate the alternatively spliced transcripts;
- detection of medium- and low-copied proteins by utilizing multi-dimensional separation with MRM technology; proteotyping and proteogenomic profiling; deciphering the chromosome-centered portion of interactome;
- advanced technology for detection of ultralow-copied proteins by integration of molecular detectors with mass-spectrometry [64],[78].

The advance of analytical technologies is the basic element of the Russian Roadmap. To our opinion, the top target in this direction is the usage of nanotechnological approaches. As a deliverable of participating in HPP, Russia is planning to elaborate new equipment for proteomic studies combining the opportunities of mass-spectrometry and atomic-force microscopy. With such AFM-MS complexes one may expect to attain the sensitivity at the level 10^{-18} M in blood plasma and 10-100 copies of protein molecules for 1 cell of liver tissue.

Roadmap meets the translational medicine by incorporating the plan for construction of molecular sensors based on nanoscale semiconductor elements. In foreseeable future these devices will provide affordable means for complex, multiparametric molecular analyses. Appropriate reliability of such devices opens up a way for the point-of-care and even home environment diagnostics. Roadmap master phase establishes a benchmark for the personalized medicine through creating prerequisite hardware for fostering individual prophylaxis and treatment of socially significant diseases.

References

[1] R.K. Saiki, D.H. Gelfand, S. Stoffel, S.J. Scharf, R. Higuchi, G.T. Horn, K.B. Mullis, and H.A. Erlich, "Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase.," *Science (New York, N.Y.)*, vol. 239, 1988, pp. 487-91.

[2] M. Frenkel-Morgenstern, A.A. Cohen, N. Geva-Zatorsky, E. Eden, J. Prilusky, I. Issaeva, A. Sigal, C. Cohen-Saidon, Y. Liron, L. Cohen, T. Danon, N. Perzov, and U. Alon, "Dynamic Proteomics: a database for dynamics and localizations of endogenous fluorescently-tagged proteins in living human cells.," *Nucleic acids research*, vol. 38, 2010, pp. D508-12.

[3] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al., "Initial sequencing and analysis of the human genome.," *Nature*, vol. 409, 2001, pp. 860-921.

[4] I. Consortium, "Finishing the euchromatic sequence of the human genome.," *Nature*, vol. 431, 2004, pp. 931-45.

[5] K.U. Mir, "Sequencing genomes: from individuals to populations.," *Briefings in functional genomics & proteomics*, vol. 8, 2009, pp. 367-78.

[6] M. Snyder, J. Du, and M. Gerstein, "Personal genome sequencing: current approaches and challenges.," *Genes & development*, vol. 24, 2010, pp. 423-31.

[7] HUPO, "A gene-centric human proteome project: HUPO--the Human Proteome organization.," *Molecular & cellular proteomics : MCP*, vol. 9, 2010, pp. 427-9.

[8] L.A. Liotta and E.F. Petricoin, "Beyond the genome to tissue proteomics.," *Breast cancer research : BCR*, vol. 2, 2000, pp. 13-4.

[9] Y.D. Ivanov, V.M. Govorun, V.A. Bykov, and A.I. Archakov, "Nanotechnologies in proteomics.," *Proteomics*, vol. 6, 2006, pp. 1399-414.

[10] A.I. Archakov, Y.D. Ivanov, A.V. Lisitsa, and V.G. Zgoda, "AFM fishing nanotechnology is the way to reverse the Avogadro number in proteomics.," *Proteomics*, vol. 7, 2007, pp. 4-9.

[11] J. Hahm and C.M. Lieber, "Direct Ultrasensitive Electrical Detection of DNA and DNA Sequence Variations Using Nanowire Nanosensors," *Nano Letters*, vol. 4, 2004, pp. 51-54.

[12] P. Picotti, B. Bodenmiller, L.N. Mueller, B. Domon, and R. Aebersold, "Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics.," *Cell*, vol. 138, 2009, pp. 795-806.

[13] J. Robinson, M.J. Waller, S.C. Fail, H. McWilliam, R. Lopez, P. Parham, and S.G. Marsh, "The IMGT/HLA database.," *Nucleic acids research*, vol. 37, 2009, pp. D1013-7.

[14] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet, "GeneCards: integrating information about genes, proteins and diseases.," *Trends in genetics : TIG*, vol. 13, 1997, p. 163.

[15] J.H. EDWARDS, D.G. HARNDEN, A.H. CAMERON, V.M. CROSSE, and O.H. WOLFF, "A new trisomic syndrome.," *Lancet*, vol. 1, 1960, pp. 787-90.

[16] B. Goc, Z. Walencka, A. Włoch, E. Wojciechowska, D. Wiecek-Włodarska, J. Krzystolik-Ładzińska, K. Bober, and J. Swietliński, "Trisomy 18 in neonates: prenatal diagnosis, clinical features, therapeutic dilemmas and outcome.," *Journal of applied genetics*, vol. 47, 2006, pp. 165-70.

[17] K. Gu, W.C. Chan, and R.C. Hawley, "Practical detection of t(14;18)(IgH/BCL2) in follicular lymphoma.," *Archives of pathology & laboratory medicine*, vol. 132, 2008, pp. 1355-61.

[18] "Human Genome Project Information http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml."

[19] C. Nusbaum, M.C. Zody, M.L. Borowsky, M. Kamal, C.D. Kodira, T.D. Taylor, C.A. Whittaker, J.L. Chang, C.A. Cuomo, K. Dewar, M.G. FitzGerald, X. Yang, A. Abouelleil, N.R. Allen, S. Anderson, T. Bloom, B. Bugalter, J. Butler, A. Cook, D. DeCaprio, R. Engels, M. Garber, A. Gnirke, N. Hafez, J.L. Hall, C.H. Norman, T. Itoh, D.B. Jaffe, Y. Kuroki, J. Lehoczky, A. Lui, P. Macdonald, E. Mauceli, T.S. Mikkelsen, J.W. Naylor, R. Nicol, C. Nguyen, H. Noguchi, S.B. O'Leary, K. O'Neill, B. Piqani, C.L. Smith, J.A. Talamas, K. Topham, Y. Totoki, A. Toyoda, H.M. Wain, S.K. Young, Q. Zeng, A.R. Zimmer, A. Fujiyama, M. Hattori, B.W. Birren, Y. Sakaki, and E.S. Lander, "DNA sequence and analysis of human chromosome 18.," *Nature*, vol. 437, 2005, pp. 551-5.

[20] "UniProt http://www.uniprot.org/."

[21] "Human Protein Atlas http://www.proteinatlas.org/."

[22] "PRoteomics IDEntifications database (PRIDE) http://www.ebi.ac.uk/pride/."

[23] F. He, "Human liver proteome project: plan, progress, and perspectives.," *Molecular & cellular proteomics : MCP*, vol. 4, 2005, pp. 1841-8.

[24] G.S. Omenn, D.J. States, M. Adamski, T.W. Blackwell, R. Menon, H. Hermjakob, R. Apweiler, B.B. Haab, R.J. Simpson, J.S. Eddes, E.A. Kapp, R.L. Moritz, D.W. Chan, A.J. Rai, A. Admon, R. Aebersold, J. Eng, W.S. Hancock, S.A. Hefta, H. Meyer, Y. Paik, J.

Yoo, P. Ping, J. Pounds, J. Adkins, X. Qian, R. Wang, V. Wasinger, C.Y. Wu, X. Zhao, R. Zeng, A. Archakov, A. Tsugita, I. Beer, A. Pandey, M. Pisano, P. Andrews, H. Tammen, D.W. Speicher, and S.M. Hanash, "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*, vol. 5, 2005, pp. 3226-45.

[25] C.J. Mee, H.J. Harris, M.J. Farquhar, G. Wilson, G. Reynolds, C. Davis, S.C. van IJzendoorn, P. Balfe, and J.A. McKeating, "Polarization restricts hepatitis C virus entry into HepG2 hepatoma cells.," *Journal of virology*, vol. 83, 2009, pp. 6211-21.

[26] A. Slany, V.J. Haudek, H. Zwickl, N.C. Gundacker, M. Grusch, T.S. Weiss, K. Seir, C. Rodgarkia-Dara, C. Hellerbrand, and C. Gerner, "Cell characterization by proteome profiling applied to primary hepatocytes and hepatocyte cell lines Hep-G2 and Hep-3B.," *Journal of proteome research*, vol. 9, 2010, pp. 6-21.

[27] V.G. Zgoda, S.A. Moshkovskii, E.A. Ponomarenko, T.V. Andreewski, A.T. Kopylov, O.V. Tikhonova, S.A. Melnik, A.V. Lisitsa, and A.I. Archakov, "Proteomics of mouse liver microsomes: performance of different protein separation workflows for LC-MS/MS.," *Proteomics*, vol. 9, 2009, pp. 4102-5.

[28] N.A. Petushkova, I.P. Kanaeva, A.V. Lisitsa, G.F. Sheremetyeva, V.G. Zgoda, N.F. Samenkova, I.I. Karuzina, and A.I. Archakov, "Characterization of human liver cytochromes P450 by combining the biochemical and proteomic approaches.," *Toxicology in vitro : an international journal published in association with BIBRA*, vol. 20, 2006, pp. 966-74.

[29] A.V. Lisitsa, N.A. Petushkova, H. Thiele, S.A. Moshkovskii, V.G. Zgoda, I.I. Karuzina, A.L. Chernobrovkin, O.G. Skipenko, and A.I. Archakov, "Application of slicing of one-dimensional gels with subsequent slice-by-slice mass spectrometry for the proteomic profiling of human liver cytochromes P450.," *Journal of proteome research*, vol. 9, 2010, pp. 95-103.

[30] V. Zgoda, O. Tikhonova, A. Viglinskaya, M. Serebriakova, A. Lisitsa, and A. Archakov, "Proteomic profiles of induced hepatotoxicity at the subcellular level.," *Proteomics*, vol. 6, 2006, pp. 4662-70.

[31] L. Zang, D. Palmer Toy, W.S. Hancock, D.C. Sgroi, and B.L. Karger, "Proteomic analysis of ductal carcinoma of the breast using laser capture microdissection, LC-MS, and 16O/18O isotopic labeling.," *Journal of proteome research*, vol. 3, pp. 604-12.

[32] D.J. Johann, J. Rodriguez-Canales, S. Mukherjee, D.A. Prieto, J.C. Hanson, M. Emmert-Buck, and J. Blonder, "Approaching solid tumor heterogeneity on a cellular basis by tissue proteomics using laser capture microdissection and biological mass spectrometry.," *Journal of proteome research*, vol. 8, 2009, pp. 2310-8.

[33] G. Marko-Varga, M. Berglund, J. Malmström, H. Lindberg, and T.E. Fehniger, "Targeting hepatocytes from liver tissue by laser capture microdissection and proteomics expression profiling.," *Electrophoresis*, vol. 24, 2003, pp. 3800-5.

[34] L. Beretta, "Comparative analysis of the liver and plasma proteomes as a novel and powerful strategy for hepatocellular carcinoma biomarker discovery.," *Cancer letters*, vol. 286, 2009, pp. 134-9.

[35] L. Anderson and N.G. Anderson, "High resolution two-dimensional electrophoresis of human plasma proteins.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, 1977, pp. 5421-5.

[36] A.J. Rai, C.A. Gelfand, B.C. Haywood, D.J. Warunek, J. Yi, M.D. Schuchard, R.J. Mehigh, S.L. Cockrill, G.B. Scott, H. Tammen, P. Schulz-Knappe, D.W. Speicher, F. Vitzthum, B.B. Haab, G. Siest, and D.W. Chan, "HUPO Plasma Proteome Project specimen collection and handling: towards the standardization of parameters for plasma proteome samples.," *Proteomics*, vol. 5, 2005, pp. 3262-77.

[37] M.A. Karpova, S.A. Moshkovskii, I.Y. Toropygin, and A.I. Archakov, "Cancer-specific MALDI-TOF profiles of blood serum and plasma: biological meaning and perspectives.," *Journal of proteomics*, vol. 73, 2010, pp. 537-51.

[38] Y. Shen, J. Kim, E.F. Strittmatter, J.M. Jacobs, D.G. Camp, R. Fang, N. Tolié, R.J. Moore, and R.D. Smith, "Characterization of the human blood plasma proteome.," *Proteomics*, vol. 5, 2005, pp. 4034-45.

[39] "Proteomics National Center for Research Resource http://www.pnl.gov/biology/programs/msd/ncrr.stm."

[40] N.L. Anderson, M. Polanski, R. Pieper, T. Gatlin, R.S. Tirumalai, T.P. Conrads, T.D. Veenstra, J.N. Adkins, J.G. Pounds, R. Fagan, and A. Lobley, "The human plasma proteome: a nonredundant list developed by combination of four separate sources.," *Molecular & cellular proteomics : MCP*, vol. 3, 2004, pp. 311-26.

[41] "HUPO Plasma Proteome Project http://www.bioinformatics.med.umich.edu/hupo/ppp."

[42] P.A. Rudnick, K.R. Clauser, L.E. Kilpatrick, D.V. Tchekhovskoi, P. Neta, N. Blonder, D.D. Billheimer, R.K. Blackman, D.M. Bunk, H.L. Cardasis, A.L. Ham, J.D. Jaffe, C.R. Kinsinger, M. Mesri, T.A. Neubert, B. Schilling, D.L. Tabb, T.J. Tegeler, L. Vega-Montoto, A.M. Variyath, M. Wang, P. Wang, J.R. Whiteaker, L.J. Zimmerman, S.A. Carr, S.J. Fisher, B.W. Gibson, A.G. Paulovich, F.E. Regnier, H. Rodriguez, C. Spiegelman, P. Tempst, D.C. Liebler, and S.E. Stein, "Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses.," *Molecular & cellular proteomics : MCP*, vol. 9, 2010, pp. 225-41.

[43] A.W. Bell, E.W. Deutsch, C.E. Au, R.E. Kearney, R. Beavis, S. Sechi, T. Nilsson, and J.J. Bergeron, "A HUPO test sample study reveals common problems in mass spectrometry-based proteomics.," *Nature methods*, vol. 6, 2009, pp. 423-30.

[44] D.L. Tabb, L. Vega-Montoto, P.A. Rudnick, A.M. Variyath, A.L. Ham, D.M. Bunk, L.E. Kilpatrick, D.D. Billheimer, R.K. Blackman, H.L. Cardasis, S.A. Carr, K.R. Clauser, J.D. Jaffe, K.A. Kowalski, T.A. Neubert, F.E. Regnier, B. Schilling, T.J. Tegeler, M. Wang, P. Wang, J.R. Whiteaker, L.J. Zimmerman, S.J. Fisher, B.W. Gibson, C.R. Kinsinger, M. Mesri, H. Rodriguez, S.E. Stein, P. Tempst, A.G. Paulovich, D.C. Liebler, and C. Spiegelman, "Repeatability and reproducibility in proteomic identifications by liquid

chromatography-tandem mass spectrometry.," *Journal of proteome research*, vol. 9, 2010, pp. 761-76.

[45] M.A. Kuzyk, D. Smith, J. Yang, T.J. Cross, A.M. Jackson, D.B. Hardie, N.L. Anderson, and C.H. Borchers, "Multiple reaction monitoring-based, multiplexed, absolute quantitation of 45 proteins in human plasma.," *Molecular & cellular proteomics : MCP*, vol. 8, 2009, pp. 1860-77.

[46] T.A. Addona, S.E. Abbatiello, B. Schilling, S.J. Skates, D.R. Mani, D.M. Bunk, C.H. Spiegelman, L.J. Zimmerman, A.L. Ham, H. Keshishian, S.C. Hall, S. Allen, R.K. Blackman, C.H. Borchers, C. Buck, H.L. Cardasis, M.P. Cusack, N.G. Dodder, B.W. Gibson, J.M. Held, T. Hiltke, A. Jackson, E.B. Johansen, C.R. Kinsinger, J. Li, M. Mesri, T.A. Neubert, R.K. Niles, T.C. Pulsipher, D. Ransohoff, H. Rodriguez, P.A. Rudnick, D. Smith, D.L. Tabb, T.J. Tegeler, A.M. Variyath, L.J. Vega-Montoto, A. Wahlander, S. Waldemarson, M. Wang, J.R. Whiteaker, L. Zhao, N.L. Anderson, S.J. Fisher, D.C. Liebler, A.G. Paulovich, F.E. Regnier, P. Tempst, and S.A. Carr, "Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma.," *Nature biotechnology*, vol. 27, 2009, pp. 633-41.

[47] M.J. Roth, B.A. Parks, J.T. Ferguson, M.T. Boyne, and N.L. Kelleher, ""Proteotyping": population proteomics of human leukocytes using top down mass spectrometry.," *Analytical chemistry*, vol. 80, 2008, pp. 2857-66.

[48] T.H. Corzett, I.K. Fodor, M.W. Choi, V.L. Walsworth, K.W. Turteltaub, S.L. McCutchen-Maloney, and B.A. Chromy, "Statistical analysis of variation in the human plasma proteome.," *Journal of biomedicine & biotechnology*, vol. 2010, 2010, p. 258494.

[49] D. Nedelkov, "Population proteomics: investigation of protein diversity in human populations.," *Proteomics*, vol. 8, 2008, pp. 779-86.

[50] J. Godovac-Zimmermann, O. Kleiner, L.R. Brown, and A.K. Drukier, "Perspectives in spicing up proteomics with splicing.," *Proteomics*, vol. 5, 2005, pp. 699-709.

[51] B.A. Parks, L. Jiang, P.M. Thomas, C.D. Wenger, M.J. Roth, M.T. Boyne, P.V. Burke, K.E. Kwast, and N.L. Kelleher, "Top-down proteomics on a chromatographic time scale using linear ion trap fourier transform hybrid mass spectrometers.," *Analytical chemistry*, vol. 79, 2007, pp. 7984-91.

[52] N.M. Karabacak, L. Li, A. Tiwari, L.J. Hayward, P. Hong, M.L. Easterling, and J.N. Agar, "Sensitive and specific identification of wild type and variant proteins from 8 to 669 kDa using top-down mass spectrometry.," *Molecular & cellular proteomics : MCP*, vol. 8, 2009, pp. 846-56.

[53] G. Alves, A.Y. Ogurtsov, and Y. Yu, "RAId_DbS: mass-spectrometry based peptide identification web server with knowledge integration.," *BMC genomics*, vol. 9, 2008, p. 505.

[54] "National Center for Biotechnology Information http://www.ncbi.nlm.nih.gov/."

[55] "Ensembl www.ensembl.org."

[56] C.M. Delahunty and J.R. Yates, "MudPIT: multidimensional protein identification technology.," *BioTechniques*, vol. 43, 2007, p. 563, 565, 567 passim.

[57] G. Chen, T.G. Gharib, C. Huang, J.M. Taylor, D.E. Misek, S.L. Kardia, T.J. Giordano, M.D. Iannettoni, M.B. Orringer, S.M. Hanash, and D.G. Beer, "Discordant protein and mRNA expression in lung adenocarcinomas.," *Molecular & cellular proteomics : MCP*, vol. 1, 2002, pp. 304-13.

[58] A.T. Kopylov, V.G. Zgoda, and A.I. Archakov, "[Label-free quantitative analysis of proteins using mass-spectrometry]," *Biomedit͡sinskai͡a khimii͡a*, vol. 55, pp. 125-39.

[59] R. Kakuhata, M. Watanabe, T. Yamamoto, E. Obana, N. Yamazaki, M. Kataoka, T. Ooie, Y. Baba, T. Hori, and Y. Shinohara, "Importance of probe location for quantitative comparison of signal intensities among genes in microarray analysis.," *Journal of biochemical and biophysical methods*, vol. 70, 2008, pp. 926-31.

[60] M. Tsutsui, M. Taniguchi, K. Yokota, and T. Kawai, "Identifying single nucleotides by tunnelling current.," *Nature nanotechnology*, vol. 5, 2010, pp. 286-90.

[61] A. Archakov, "Introducing Nanoproteomics, a new section in PROTEOMICS," *PROTEOMICS*, vol. 7, 2007, pp. 4409-4412.

[62] J. Armengaud, "Proteogenomics and systems biology: quest for the ultimate missing parts.," *Expert review of proteomics*, vol. 7, 2010, pp. 65-77.

[63] V.T. Ivanov and O.N. Yatskin, "Peptidomics: a logical sequel to proteomics.," *Expert review of proteomics*, vol. 2, 2005, pp. 463-73.

[64] A.L. Kaysheva, Y.D. Ivanov, V.G. Zgoda, P.A. Frantsuzov, T.O. Pleshakova, N.V. Krokhin, V.S. Ziborov, and A.I. Archakov, "Visualization and identification of hepatitis C viral particles by atomic force microscopy combined with MS/MS analysis," *Biochemistry (Moscow) Supplement Series B: Biomedical Chemistry*, vol. 4, 2010, pp. 15-24.

[65] A. Archakov, Y. Ivanov, A. Lisitsa, and V. Zgoda, "Biospecific irreversible fishing coupled with atomic force microscopy for detection of extremely low-abundant proteins.," *Proteomics*, vol. 9, 2009, pp. 1326-43.

[66] V.Y. Kuznetsov, Y.D. Ivanov, V.A. Bykov, S.A. Saunin, I.A. Fedorov, S.V. Lemeshko, H.B. Hoa, and A.I. Archakov, "Atomic force microscopy detection of molecular complexes in multiprotein P450cam containing monooxygenase system.," *Proteomics*, vol. 2, 2002, pp. 1699-705.

[67] V.Y. Kuznetsov, Y.D. Ivanov, and A.I. Archakov, "Atomic force microscopy revelation of molecular complexes in the multiprotein cytochrome P450 2B4-containing system.," *Proteomics*, vol. 4, 2004, pp. 2390-6.

[68] M. Shibata, H. Yamashita, T. Uchihashi, H. Kandori, and T. Ando, "High-speed atomic force microscopy shows dynamic molecular processes in photoactivated bacteriorhodopsin.," *Nature nanotechnology*, vol. 5, 2010, pp. 208-12.

[69] J. Hernandez-Toro, C. Prieto, and J. De Las Rivas, "APID2NET: unified interactome graphic analyzer.," *Bioinformatics (Oxford, England)*, vol. 23, 2007, pp. 2495-7.

[70] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, "KEGG for representation and analysis of molecular networks involving diseases and drugs.," *Nucleic acids research*, vol. 38, 2010, pp. D355-60.

[71] A. Villanueva, C. García, A.B. Paules, M. Vicente, M. Megías, G. Reyes, P. de Villalonga, N. Agell, F. Lluís, O. Bachs, and G. Capellá, "Disruption of the antiproliferative TGF-beta signaling pathways in human pancreatic cancer cells.," *Oncogene*, vol. 17, 1998, pp. 1969-78.

[72] Y. Shi and J. Massagué, "Mechanisms of TGF-beta signaling from cell membrane to the nucleus.," *Cell*, vol. 113, 2003, pp. 685-700.

[73] G. Yang and X. Yang, "Smad4-mediated TGF-beta signaling in tumorigenesis.," *International journal of biological sciences*, vol. 6, 2010, pp. 1-8.

[74] B. Sobolev, D. Filimonov, A. Lagunin, A. Zakharov, O. Koborova, A. Kel, and V. Poroikov, "Functional classification of proteins based on projection of amino acid sequences: application for prediction of protein kinase substrates.," *BMC bioinformatics*, vol. 11, 2010, p. 313.

[75] M. Pellegrini, E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, 1999, pp. 4285-8.

[76] M.A. Piatnitskiĭ, A.V. Lisitsa, and A.I. Archakov, "[Prediction of functionally related proteins by comparative genomics in silico]," *Biomedit͡sinskai͡a khimii͡a*, vol. 55, pp. 230-46.

[77] O. Buneeva, O. Gnedenko, V. Zgoda, A. Kopylov, V. Glover, A. Ivanov, A. Medvedev, and A. Archakov, "Isatin-binding proteins of rat and mouse brain: proteomic identification and optical biosensor validation.," *Proteomics*, vol. 10, 2010, pp. 23-37.

[78] D. Nedelkov, "Integration of SPR biosensors with mass spectrometry (SPR-MS).," *Methods in molecular biology (Clifton, N.J.)*, vol. 627, 2010, pp. 261-8.

[79] B.J. Strasser, "Genetics. GenBank--Natural history in the 21st Century?," *Science (New York, N.Y.)*, vol. 322, 2008, pp. 537-8.

[80] O. Kuchar, J. Reyes-Spindola, and M. Benaroch, "Cognitive Bioinformatics: Computational Cognitive Model for Dynamic Problem Solving," *Proceedings of the Third IEEE International Conference on Cognitive Informatics*, 2004, pp. 84-92.

[81] D. Hull, S.R. Pettifer, and D.B. Kell, "Defrosting the digital library: bibliographic tools for the next generation web.," *PLoS computational biology*, vol. 4, 2008, p. e1000204.

[82] "Collexis http://collexis.com/."

[83] "Neosemantic Framework (ProContent) http://www.neosemanticsoft.com/."

[84] Z. Hu, I. Mani, V. Hermoso, H. Liu, and C.H. Wu, "iProLINK: an integrated protein resource for literature mining.," *Computational biology and chemistry*, vol. 28, 2004, pp. 409-16.

[85] R. Hussong, B. Gregorius, A. Tholey, and A. Hildebrandt, "Highly accelerated feature detection in proteomics data sets using modern graphics processing units.," *Bioinformatics (Oxford, England)*, vol. 25, 2009, pp. 1937-43.

[86] M. Lee, D. Lee, S. Jung, K. Lee, Y.S. Park, and W. Seong, "Measurements of serum C-reactive protein levels in patients with gastric cancer and quantification using silicon nanowire arrays.," *Nanomedicine : nanotechnology, biology, and medicine*, vol. 6, 2010, pp. 78-83.