The process chain for peptidomic biomarker discovery

Michael Schrader[a,b,*] and Hartmut Selle[b]
[a]*Department of Biotechnology and Bioinformatics, University of Applied Sciences, Weihenstephan, D-85350 Freising, Germany*
[b]*BioVisioN AG, Feodor-Lynen-Str. 5, D-30625 Hannover, Germany*

Abstract. Over the last few years the interest in diagnostic markers for specific diseases has increased continuously. It is expected that they not only improve a patient's medical treatment but also contribute to accelerating the process of drug development. This demand for new biomarkers is caused by a lack of specific and sensitive diagnosis in many diseases. Moreover, diseases usually occur in different types or stages which may need different diagnostic and therapeutic measures. Their differentiation has to be considered in clinical studies as well. Therefore, it is important to translate a macroscopic pathological or physiological finding into a microscopic view of molecular processes and vice versa, though it is a difficult and tedious task. Peptides play a central role in many physiological processes and are of importance in several areas of drug research. Exploration of endogenous peptides in biologically relevant sources may directly lead to new drug substances, serve as key information on a new target and can as well result in relevant biomarker candidates. A comprehensive analysis of peptides and small proteins of a biological system corresponding to the respective genomic information (peptidomics® methods) was a missing link in proteomics. A new peptidomic technology platform addressing peptides was recently presented, developed by adaptation of the striving proteomic technologies. Here, concepts of using peptidomics technologies for biomarker discovery are presented and illustrated with examples. It is discussed how the biological hypothesis and sample quality determine the result of the study. A detailed study design, appropriate choice and application of technology as well as thorough data interpretation can lead to significant results which have to be interpreted in the context of the underlying disease. The identified biomarker candidates will be characterised in validation studies before use. This approach for discovery of peptide biomarkes has potential for improving clinical studies.

Glossary

- *Peptide:* Oligo- and polypeptides with a molecular mass below 20 kDa
- *Endogenous peptide:* A peptide generated *in vivo* within a biological system or subsystem
- *Peptidome:* All endogenous peptides present in a biological system or subsystem at a given time
- *Peptidomics analysis:* Comprehensive analysis of peptides present in a biological system or subsystem

1. Biomarker discovery

1.1. Medical and pharmaceutical need for proteomic biomarkers

There is a constant need for new diagnostics and biomarkers for the correct and early diagnosis of diseases. A biomarker by definition is a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes or pharmacologic responses to a therapeutic intervention [8]. Therefore, molecular analytes which correlate with disease states are searched for to develop new assay systems [13,23,60]. The corresponding research so far has delivered few new molecules. Along with the decryption of the human genome as the blueprint of human life, the knowledge about the molecular nature of the substances produced in our

[*] Corresponding author. E-mail: michael.schrader@fh-weihenstephan.de.

ISSN 0278-0240/06/$17.00

body evolving from the genome becomes more and more important. The risk factors for a disease might be defined genetically, but the development of a diseased state mostly does not involve changes in the genome itself. Changes occur usually downstream in regulatory processes like gene expression, protein synthesis and processing [34]. Typical approaches thus include the analysis of RNA, proteins and nowadays also of peptides. The measurement of RNA quantities using for example multi-analyte chips is one possibility, but RNA levels not necessarily depict protein expression [3,20, 23]. The subsequent processing of proteins is a widely occurring principle that results in metabolic variation in living systems. With a systemic view about proteins and peptides as the main gene products, there is a great potential to better understand the complex regulatory systems of the human body and to identify new molecules that can be used as biomarkers or diagnostics. With the integration of modern proteomics [3,43] and peptidomics [45] technologies into biomedical research, it is anticipated that the number of new molecular biomarker entities will substantially increase. The requirements for biomarkers in clinical development of new drugs and toxicology are somewhat different compared to biomarkers for diagnostic use, but the same set of new technologies has the potential to be applied in these fields, too [28,60].

The new possibilities of mass spectrometry enormously improved the sensitivity and selectivity of proteomics and peptidomics enabling a comprehensive analysis of proteins and peptides [2,6]. Over recent years 2D-gel electrophoresis in combination with mass spectrometry has become the main proteomics research tool [3]. This methodology addresses the analysis of proteins in a molecular mass range between about 10 and 200 kDa. However, smaller proteins and native peptides are not yet covered by the standard proteomics methodologies.

1.2. Peptidomics – why peptides as biomarkers?

The term peptidomics was first introduced in February 2000 at the ABRF conference "From Singular to Global Analyses of Biological Systems". It was coined as a short version of "peptide proteomics" and describes the comprehensive qualitative and quantitative analysis of all peptides and small protein of a given biological system. We and a second group independently published different peptidomics concepts in parallel for the analysis of peptides in body fluids [45] and in endocrine research [11]. The number of peptidomics papers has grown steadily ever since.

Peptides are oligomers or polymers of amino acids linked by peptide bonds. There is no official clear-cut definition distinguishing between peptides and proteins [24]. We refer to the term 'peptide' for oligo- and polypeptides in the range from dipeptides to molecules of about 20 kDa. This choice was mainly based on the differences between the physico-chemical properties of peptides and proteins. A further reason for this definition of peptides is their physiological discrimination in the human organism by the cut-off of the kidney filtration of blood plasma components. Hemofiltration is used to replace the kidney function in patients with end-stage renal disease using the same cut-off by removing compounds with molecular masses below 20 kDa whereas plasma proteins are retained [46].

Peptides are involved in almost all physiological areas and are tightly regulated by proteolytic control [49]. Quite frequently one precursor is cleaved into several biologically active peptides and occasionally this processing varies in different types of cells or tissues [15, 22]. These complex processing patterns of peptides can be used for diagnostic purposes [36,40]. Moreover, further processing that could also be of diagnostic use occurs in the extra-cellular compartments of the body [37]. Insulin illustrates such an example: The two insulin chains are released from one precursor molecule, the proinsulin, by enzymatic removal of the C-peptide, a segment connecting the two insulin chains within proinsulin. The insulin molecule is used as a therapeutic compound but also for the diagnosis of diabetes. In addition, the C-peptide as side product from the processing is used as a diagnostic marker. In diabetes diagnosis it is preferred compared to insulin since its biological half-life is longer and it is easier to measure.

Several peptides are already commercially available as diagnostic markers or are under validation in clinical research: Insulin and C-peptide are used in diabetes and beta-amyloid peptide has been established as a marker for Alzheimer's disease. Relevant peptides can be processed protein fragments as well as products of degeneration and metabolism of proteins. Some regulatory messengers, such as hormones and cytokines and their pre-processed or processed forms are useful as diagnostics. Procalcitonin is the current gold standard for diagnosis of sepsis [58]. Another very recent example of a peptide biomarker is the diagnosis of cardiovascular diseases by means of brain natriuretic peptide (BNP) that is cleared very fast by the body. An assay for an

Table 1
Examples of important human peptides which are used as diagnostics or biomarkers

Peptide	Molecular mass [kDa]	Clinical use
Gastrin	2.1	Ulcus, diarrhoe
C-peptide	3.1	Diabetes
Beta-amyloid peptide	4.8	Alzheimer's disease
Insulin	5.8	Diabetes
Osteocalcin	5.8	Osteoporosis
NT-proBNP	8.6	Cardiovascular diseases
Procalcitonin (PCT)	13	Sepsis

N-terminal propiece of 8.5 kDa (NT-proBNP) [26,33] is now used as a marker for congestive heart failure with good success.

In addition to a use in diagnostics, such molecular parameters are needed to monitor the efficacy of a therapy in case of a lack of accessible appropriate clinical parameters. Such biomarkers are for example needed for degenerative diseases with slow progression. A well-known example of such a biomarker is the beta-amyloid peptide in the cerebrospinal fluid of patients suffering from Alzheimer's disease, indicating the status of amyloid and plaque formation [7]. Table 1 lists prominent examples for diagnostic products derived from peptide discovery. The original peptides discovered have a molecular mass well below 15 kDa. Several further peptide-derived compounds are currently under clinical development. All these examples demonstrate the use of peptides in different disease areas. With better molecular parameters at hand, it is expected to accelerate drug development in such diseases.

2. Process chain

2.1. Background and hypothesis

Biomarker development and application require a multidisciplinary approach. Teams with very different skills and competences are needed for success [13]. The complex nature of the scientific work can be reduced by establishing defined separate modules that are assembled to a process chain (Fig. 1). This breakdown of complexity allows smaller teams to focus all necessary skills on their respective core competence. Nevertheless, it has to be taken into account that the weakest module of the process chain will determine the overall quality of its output. Furthermore, failures in the early phases of that chain can not be compensated downstream by sophisticated technology or data mining. Very often soft skills such as communication are critical and have to be addressed by a competent project management.

To start a pepdidomics-based discovery project aiming at new biomarkers possible hypotheses about the involvement of peptides have to be defined and reviewed; an assumption about changes in the concentration of peptide analytes has to be made. Such changes occur very likely in diseases where proteolytic enzymes, peptide hormones or protein processing are involved. This includes degenerative diseases like Alzheimer's disease and osteoporosis or diabetes and obesity that comprise hormonal dysregulation.

As a starting point, a hypothesis is intended to model the progression of the disease. Biomarker discovery includes the analysis of a dynamic disease process and the characteristics of such a process have to be considered in the study design. This systematic approach has been underestimated in the past and the paradigm for global "omics" analyses has to be changed into a view for dynamic systems that is comparable to the systematics that is well established in non-equilibrium physical chemistry [52]. For the time being, it should be helpful to apply at least simple models of the progression of disease, for example:

- Acute change that results in two almost discrete states
- Continuous change of the disease progression
- Stepwise change of disease progression
- Occurrence of transition states between health and disease

As typically a plethora of peptide analytes is analysed, only a restricted number of well-defined sample groups should be included in a study. For the selection of groups of patients it has to be taken into account, whether the disease reaches a defined endpoint or if it can progress into further states which may be even not specific for that disease, i.e. for all patients. Confounding diseases might be included and sometimes it is impossible to define a group of "healthy" controls. The preceding state or starting point of the investigated

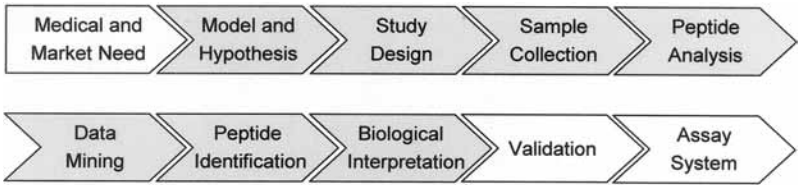


Fig. 1. The process chain of peptide biomarker discovery comprises several consecutive steps. Each step is represented by a box. Grey boxes correspond to topics reviewed in this article.

disease might not be healthy but another disease or physiological status which is difficult to differentially diagnose against the target status. Thus, different states of disease may be compared that have to be defined precisely before experimental work starts. Samples are ordered into groups representing a well-defined status. The more groups are used the more detailed information can be obtained. But it also complicates the data mining process, as it is illustrated by our study on Alzheimer's disease, where the Alzheimer group was compared to two different control groups: subjects without cognitive impairment and patients suffering from another type of dementia [48].

With an appropriate model in mind it is possible to decide on the most suitable groups and time points to recruit samples. A wrong decision at this point can already substantially derogate the success of the whole project, obliterating the subsequent work in the process chain. The process of model-based biomarker discovery is a more complicated approach than typically described in the literature. However, it is possible to rely on knowledge generated within decades of research in different disciplines [56].

Designing a good hypothesis needs a lot of experience and expertise, thus clinical experts usually have to take the lead. The hypothesis in general has to be adjusted to the clinical context of a disease or treatment. In case of an commercial output, market requirements have to be taken into account as well. Interdisciplinary communication has to solve this task, which often is impaired by a lack of comprehension of the clinical and technology experts. Therefore, an experienced team, where the necessary professional skills are combined appropriately, is an invaluable basis for projects on biomarker research.

2.2. Study design and samples

Many proteomic studies have been published during the last years describing the use of very limited numbers of clinical samples. This only partially led to progress in diseases of difficult aetiology such as Alzheimer's disease [9,43] and cancer [1,38]. However, many other studies did not give significant results and it has to be accepted that the same rules in principle have to be applied in biomarker discovery as they are used in clinical studies [32]. This means that it should be estimated within the study design phase what chance of success a project has under certain assumptions. With this knowledge at hand, it is possible to decide whether a study should take place or if parameters have to be adjusted before the start of sample analysis. The selection and appropriateness of criteria and parameters has to be done at such an early stage and not after completion of a lot of analytical work.

The number of samples is the most critical parameter in many biomarker discovery projects since it is difficult to collect sufficient numbers of suitable clinical samples with a good overall quality. A mandatory criterion is a very good and reliable clinical diagnosis by an experienced physician based on the available gold standard. Such samples are often difficult to obtain, very expensive and limited in amount and volume. For larger sample sets usually several clinical centres are included which increases the variance of sample quality and the diagnosis. Centre effects may be the sources of systematic errors. Though a multi-centre based project may result in a biomarker that addresses a wider population, it should be restricted to later discovery phases. In early phases samples should be as homogenous as possible. Biomarker candidates from the early phase shall be verified or differentiated with further sample sets, including several sources. Nonetheless, all samples have to fulfil the same criteria and the need of such samples led to the establishment of so-called "biobanks" that are built up by companies or clinical centres.

The sample size depends on several parameters and has to be estimated on the basis of the underlying hypothesis and the biological as well as the technological variance. In the characterisation of diagnostic tests a terminology was developed to quantify discriminatory accuracy [50] which can be adopted to characterise discovery projects. The samples size n, i.e. the number of samples necessary for finding differences, can be depicted as a function of several factors which have to be considered for estimation:

$n = f$ (ROC$_{\mathrm{tar}}$, ROC$_{\mathrm{dia}}$, S$_{\mathrm{pep}}$, V$_{\mathrm{bio}}$, V$_{\mathrm{tec}}$, Δc); with

ROC$_{\mathrm{tar}}$: given value for the integral of the receiver operating characteristic curve, combining the expected sensitivity and specificity for a final diagnostic assay

ROC$_{\mathrm{dia}}$: value for the integral of the ROC curve concerning a given hypothesis which is available in the chosen clinical setting

S$_{\mathrm{pep}}$: sensitivity of the peptidomic analysis

V$_{\mathrm{bio}}$: biological variance and sample heterogeneity

V$_{\mathrm{tec}}$: technological variance of the applied methods

Δc: Expected difference (dynamics) in marker concentrations

The sample size has to be increased in non-linear fashion for increasing discriminatory power of a marker, reflected by a high ROC value, with the uncertainness of the clinical diagnosis and increasing biological variance of the samples as well as technological variance. The sensitivity of the peptidomic analysis has a twofold input. With increasing sensitivity more potentially interesting molecules can be semi-quantitatively analysed. The increasing number of analytes increases the risk of false positive results that again can be overcome by more samples. The expected effect in change of peptide concentrations correlates in negative fashion with the sample size.

A typical sample size for simple biomarker studies often is below $n = 10$ per group which means that practically no statistical analysis is possible in the case of multi-variate data. This number can only serve as a starting point to check sample quality and analytical instrumentation. To perform a reliable statistical interpretation that delivers robust marker candidates, sample numbers usually exceed a few dozens and can easily be above 100. Since clinical samples are usually expensive and collection is very time consuming, experiments with suitable animal model systems may be considered for the first steps of the discovery process – wherever appropriate [44]. In this case the study can be designed in an optimal way and sample availability is not restricted to samples that are usually taken in clinical settings.

Many early proteomic discovery projects started with a simple assumption of comparing healthy and diseased states using too few samples of different and often not precisely defined clinical stages, e.g. in cancer research. This led to observations where the corresponding proteins or peptides were linked to such unspecific effects as acute phase reaction, cell lysis or other general clinical symptoms and biochemical processes [59]. The analysis of the acute phase response in human blood plasma was in fact one of the early quantitative proteomic studies demonstrating a potential diagnostic application [17].

At this stage it shall be considered to collect sufficient material not only for experiments on marker discovery but also (1) for method development in pilot studies in order to avoid problems in later phases of the process chain and (2) to allow identification of the molecular nature of interesting hits. For the data mining process clinical data like a pathological and physiological characterisation of the patients by experts is mandatory. Data must be documented in a standardised format as spreadsheet or database, providing a uniform input for the data mining process.

Special attention has to be paid to the procedure of sample collection. The biochemical synthesis of peptides in higher organisms is usually performed via protein precursors. Peptides are therefore indirect gene products occurring after a series of processing steps. Specific enzymatic proteolysis is the most important post-translational modification for peptides and in particular for peptide hormones such as insulin. Enzymatic activities therefore have to be carefully stopped immediately after sample collection. Different measures or combinations thereof are applied, such as a substantial change of pH, the reduction of certain metal ion concentrations (e.g. by EDTA complexation) and/or addition of specific inhibitors. If applicable, a size dependent separation of molecules larger than 20 kDa reduces the concentration of enzymes which typically reside in the range of 25–30 kDa. Furthermore, rigorous cooling to at least 4°C and freezing for long-term storage should be applied immediately after sample collection. This is of special importance in studies that focus on endogenous peptides. The information content of a peptide hormone often is that specific that the removal of only one amino acid residue from the peptide chain sufficiently alters the physico-chemical properties and changes its biological function.

In conjunction with the problems of interacting with different clinical partners which are usually lacking sufficient time for research and are no experts in protein chemistry, it is very critical to condense and communicate the necessary information. Detailed standard operating procedures (SOP) help to establish a standardised procedure for sample collection and handling, thereby reducing sample variability. However, the practical implementation in the clinic should be discussed in advance and controlled during sample collection; a customized training of the responsible clinical staff may be helpful for complex procedures. Samples that do not fulfil stringent criteria should be excluded before analysis as such samples will not improve but rather impair the overall results. In a study recently performed for identifying new biomarkers for Alzheimer's disease, more than 300 samples were included to gain enough information to properly reduce the number of potential peptide biomarker candidates [48].

In conclusion, well selected, collected and documented clinical samples are a key factor of successful biomarker discovery. It is worthwhile to pay specific care to the samples as these otherwise may deliver serious and even irreversible errors in the analysis. A sufficient number of individual samples according to study design is a prerequisite to achieve a valid result.

2.3. Sample analysis

Peptides in general usually occur in small quantities within complex biological matrices. This is attributed to the complexity of biological sources, the small concentration of the single components and the overwhelming amounts of a few house-keeping proteins. Peptide research is thus substantially driven by innovations in analytical chemistry [45,49]. Due to the physico-chemical properties of peptides, a separation by chromatographic methods is favourable [31]. The best detection instrumentation is mass spectrometry (MS) since it is sensitive, extremely accurate and allows quantification. Especially the impressive boost in mass spectrometry during the last 20 years [2,42] improved the analysis of peptides by several orders of magnitude in terms of sensitivity, specificity and speed.

Before these methods can be applied, a stringent sample preparation is necessary. Biological samples usually have a much higher content of larger proteins as compared to peptides. In blood plasma, albumin is dominating the whole protein content with about 40 to 50 g/L. This corresponds to almost 1 mmol/L in molecular concentration. However, typical concentrations of peptides in blood vary between about micromolar and picomolar concentrations for peptide hormones. Moreover, blood plasma and cerebrospinal fluid as other biological samples consist of a very complex mixture of peptides [21,41]. A very effective sample preparation is thus needed to remove three to nine magnitudes in concentration of proteins still allowing a reproducible analysis of peptides. Single depletion steps like immunoprecipitation, solid-phase extraction or ultrafiltration typically allow a factor of hundred to thousand in enrichment of peptides which still is not sufficient. Several methods have to be combined and carefully optimised. For the sake of reproducibility every step of the sample preparation has to be examined in terms of relative quantification. This can be monitored by the addition of internal standards and in a later stage by means of common abundantly occurring components in the sample that are used in a similar way as internal standards. Though not as important as reproducibility, the recovery of the analytes should be ascertained, being not too low.

The result typically is a modular process chain that allows a robust sample preparation and subsequent analysis [12,47]. A rigorous standardisation and automation of all parts of the analysis such as sample preparation, chromatography and mass spectrometry is the basis for a semi-quantitative interpretation of the data. All instrumentation and disposables have to be evaluated carefully, otherwise substantial differences can be delivered by variances of the technology alone as it was shown for surface-enhanced laser desorption/ionisation (SELDI) based analysis as an example [6]. A high sample throughput facilitates the analysis of high sample numbers and replicate measurement, e.g. for statistical reasons.

A simple technological approach is the use of SELDI analysis which typically focuses on peptides [1,9,38]. Here sample preparation takes place without subsequent chromatographic fractionation but by means of specific mass-spectrometric targets with different surface affinity characteristics. This allows an easier data mining process as each sample is represented in a single mass spectrum. However, this leads to a drastic decrease of the number of signals due to superimposition of signals as a result of the omission of a preceding chromatographic fractionation and as a result of the low resolution of the used mass spectrometer. Mass-spectrometric peaks are not fully resolved and it is thus difficult to distinguish between different molecules. Furthermore, the lacking separation leads to substantial suppression effects if one or a few analytes dominate a sample. The outstanding advantage of this system is its simple use by almost any group interested in biomarker discovery without needing a lot of specialised skill and experience.

The above mentioned problems can be overcome by coupling of a chromatographic fractionation process with an mass-spectrometric analysis [45,47]. Such approach requires a specific and sophisticated technology platform to secure the quality of the process chain which so far has been accomplished only by few groups. As an alternative chromatography and mass spectrometry may be coupled online [25,57]. Again, a high reproducibility has to be achieved which includes extensive validation of the analytical steps. Every step can be monitored by use of added internal standards or abundant components within the sample itself. Finally, the processes have to be carried out according to standard operating procedures (SOPs), automated and checked for quality by experienced staff using a set of defined parameters.

2.4. *Data mining*

The interpretation of large data sets is still a developing field in biology and the application of strict procedures for data processing and statistical analysis is often overcome with rather simple rules. It is not sufficient to just state that a twofold increase in intensity is a significant result [54] but the significance of signal differences has to be verified. The significance of changes in the data sets depends on about the same set of parameters as the sample size ($\mathrm{ROC}_{\mathrm{dia}}$, $\mathrm{S}_{\mathrm{pep}}$, $\mathrm{V}_{\mathrm{bio}}$, $\mathrm{V}_{\mathrm{tec}}$, $\Delta\mathrm{c}$, n) including also the number of samples itself. The quality of the data set usually increases with sample size, but less than linear. Critical is that the technological variance has to be less than the biological variance in this context. Often this is not known prior to sample analysis and therefore, data mining includes an iterative analytical process. In every step it has to be verified that the results are robust.

In the case of peptidomic biomarker discovery analysing several thousand analytes [21,29,48] is included in the data mining process. Samples are analysed in several independent analytical sets of samples each delivering a data set for a separate independent statistical analysis. A specifically designed software has been developed for automated analysis of mass spectrometric data as well as a data mining process for detecting the most robust markers. For each analytical set non-parametric statistics is performed for any mass-spectrometric signal: absolute and relative differences, non-parametric U tests and ROC curves (receiver operator characteristic) in order to discriminate between the patient groups. Clinical data are included and correlated to the experimental outcome; e.g. the occurrence of a pathological status or clinical symptom is related to the mass-spectrometric signal intensity and correlational analysis is performed. The resulting list of candidates is sorted based on a statistical parameter, e.g. the ROC values, and the lists from all analytical sets are combined. Selection criteria like a minimal ROC value or a p-value are defined before analysis and all signals that meet these criteria in all analytical sets are considered as marker candidates [48].

2.5. *Peptide identification*

After successful extraction and quantification of peptides, the next important aspect is sequence identification. This enables the interpretation of the biological context, supports the validation and facilitates patent protection of the results of biomarker discovery.

In several early proteomic studies using SELDI mass spectrometry of clinical samples (e.g. [1,38]), a sophisticated data mining process was also applied leading to promising results. However, the choice of hypothesis is debatable as cancer patients were compared with healthy controls but not with other oncologic diseases. This approach did not take sufficiently into account the multifactorial nature of cancer which is assumed to develop in several stages before the disease can be diagnosed. Subsequent studies did not find the same set of markers [16]. One specific problem of this SELDI-based analysis is the difficulty in identifying the molecular nature of the marker candidates, a limitation that has not yet been fully overcome for this technology. Though it has been discussed whether this knowledge is a prerequisite in biomarker discovery [14, 16] we strongly believe that in the current learning phase of applying proteomics and peptidomics technology, the identification of the candidates is essential to verify the whole process chain for biomarker discovery. In addition, the knowledge of the identity (sequence) of biomarker candidates is a very likely prerequisite for obtaining intellectual property. Although these pioneering SELDI-based experiments did not fulfil the initial expectations they gave a remarkable impetus to the area of biomarker discovery in general. Below, approaches allowing the sequence identification of biomarker candidates by mass spectrometric analysis will be introduced. The SELDI approach may be supplemented by these technologies that have proven to deliver the required information.

Presently, most sequence information is generated by mass-spectrometric methods followed by database comparison owing to the high-throughput nature of the technique compared with Edman sequencing or amino acid analysis [2]. The identification of peptides from complex biological sources is still a challenge that could be overcome by further development of mass spectrometric instrumentation [10]. Proteins are usually digested with trypsin into several peptides and identified by comparison of the such cleavage pattern with databases. For peptides, the number of possible specific trypsin cleavage sites is typically too small for an identification via this approach. As a prerequisite, the original peptide has to be purified to a high extent. With the modern mass spectrometry instrumentation, the identification of the amino acid sequence of peptides and small proteins is favourably achieved with a "top-down" approach which involves high-resolution mass spectrometric measurement and fragmentation of intact ionised molecules in the gas phase [27].

The absence of practical experience in top-down identification of endogenous peptides from complex mixtures has recently been overcome by the development of optimised experimental protocols [5,35]. After a reduction of complexity of biological samples by liquid chromatography and mass-spectrometric fragmentation, the resulting characteristic data of multiply charged fragment ions can be identified by using specific software for automated database searches. The existing databases, whether they cover proteins, genomes or ESTs are of great help for acceleration and automation of identification processes, although there is still a lack of specific databases and software to search for post-translational modifications. A specific workflow combining initial liquid chromatography followed by offline mass spectrometry was established for the separation and identification of peptide components from complex mixtures. A Q-TOF spectrometer is used for a very powerful purification step and subsequent mass analysis of the fragments. The quadrupolar ion optics very effectively separates the ions of interest which can be analysed subsequently in the TOF (time of flight) part after collision-induced dissociation of these peptides. This approach was for example successfully applied for the identification of endogenous peptides in cerebrospinal fluid [21,35,51] and murine brain tissue [35,53]. This top-down methodology is generally applicable for peptides up to 9 kDa from body fluids, tissues or other biological sample sources and also allows for the identification of post-translational modifications. If the molecular mass range of the proteins of interest clearly exceeds 10 kDa, the application of either classical peptide mapping [2] or the application of more sophisticated FT-ICR-MS instrumentation [27] is favourable.

2.6. Biological interpretation

At this point of the process chain the original project's hypothesis has to be reviewed in order to verify, disprove or modify the initial assumptions, in order to change the study design or consider a termination of the project. A modification or disproval might request a modified or even a new hypothesis for the molecular nature of the disease. Another type or preparation of sample may be required which may result in a new design for the project.

In the case of peptide biomarker candidates first the corresponding protein precursors and genes are examined. These are surveyed for relevance in the disease context by database and literature retrieval. Moreover, it should be evaluated if further peptides from the same precursor or gene are also potential candidates or not. With mass spectrometric technology, this is rather easy as, based on the precursor sequence, the potential molecular masses can directly be calculated and possible candidates can be screened by use of automated software tools [30]. The processing sites of the peptide candidates are another source of information. These may lead to the proteolytic enzymes that generated the peptides which could contain further important information and can be compared to the disease model.

During the process of evaluation, it should be kept in mind that not all detected differences must causally be linked to the disease, although biomarkers with an obvious interrelation to the underlying disease are the easiest to interpret and the most interesting for further studies [34]. The process chain is completed with a ranking summarising all important information gathered from the analysis, data mining and biological interpretation modules. This list or matrix should then be used to select the best candidates for further development or to design a new discovery project.

2.7. Further steps

A very promising feature of peptidomic analysis is the availability of several methods for quantitative analysis of selected components. Immunoassay based measurement with radioimmunoassay (RIA) or enzyme-linked immunosorbent assay (ELISA) are well established. Moreover, mass spectrometry can also be used to determine the amounts of specific peptides with the modern instruments capable of both improving the purification and performing the quantification [25,45]. Sample preparation can be performed by affinity [19, 55] or chromatographic purification, with the analysis conducted either online or offline. The combination of affinity preconcentration with rather low specificity combined with subsequent mass spectrometry seems very promising [4]. Electrospray MS [18,25] and MALDI-TOF-MS [19,55] are both suitable detection methods for quantitative measurements. The quantification can be performed by comparison with either external [55] or internal [18,19] references. The subsequent steps concerning validation and initiation of the development of an assay system are reviewed in two papers from experts from the diagnostics industry [23, 60].

3. Example of a peptidomics application: Differential peptide display in cerebrospinal fluid

A body fluid of specific interest in neurodegenerative diseases is cerebrospinal fluid (CSF), as CSF is a known source for neuropeptides and peptide biomarkers in neurodegenerative diseases such as Alzheimer's disease [7]. By using different combinations of liquid chromatography and mass spectrometry many peptide components in CSF were identified [39,51]. CSF does not contain such an overwhelming content of high abundant proteins like blood plasma. Many of the high abundant peptides are generated by specific proteolytic processing of neurospecific protein precursors like secretory proteins [51]. Applying a high-resolution peptidomic display, a specific peptide pattern from CSF was shown [21].

This work from our group was the first example of biomarker discovery in CSF using the differential peptide display® (DPD) technology for a systematic screen of all CSF peptides in relation to a clinical condition: CSF from two patients suffering from a primary CNS lymphoma (PCNSL) was analysed and the peptide pattern was compared with that obtained from samples from three subjects undergoing routine myelography. More than 6,000 signals were detected with each signal's position characterised by its relative molecular mass and its elution time during the preceding chromatography. This precision allowed the comparison of the two peptidomes: Several differences were found and one, that was detected with higher intensity in the PCNSL samples, was identified as a 24 amino-acid fragment of serum albumin. This identification of the peptide that correlates to the mass-spectrometric signal from the DPD analysis allows the evaluation of the pathophysiologic relevance of that difference in the peptide pattern. The increase of that particular albumin fragment can be explained by barrier disruption that typically occurs in tumour patients. The complete albumin molecule is an established marker for the integrity of the brain barriers with the ratio of albumin concentration in CSF and blood correlating with the extent of the barrier disruption. The smaller peptide might even easier traverse the blood-CSF barrier indicating a less severe barrier disruption at an earlier time point. Further CSF analyses should include relevant clinical data like the albumin quotient of CSF and blood as well as correlational analyses for establishing relations between candidate peptides and those clinical parameters [21,29].

Neurodegenerative diseases pose a growing medical need. Especially Alzheimer's disease (AD) is expected to drastically increase in the next decades. For the diagnosis of AD peptide biomarkers in CSF, that originate from the processing of amyloid beta protein by secretases, are already established [7]. A few peptidomic studies were undertaken to identify further biomarkers for this severe disease. Using SELDI-MS several potential peptide candidates were identified [9]. However, these studies involved rather small sample numbers resulting in data that still have to be validated. Furthermore, a lack of sufficient quantities of CSF from relevant patients hampers the precise identification of the peptides of interest. Thus, as proposed above, such studies should include several hundred samples of CSF and a sufficient data mining strategy to address biological and technological variability and diagnostic uncertainty [48]. The patient groups shall be defined prior to sample collection, e.g. by balancing parameters like age between the groups or including only comparable subgroups concerning genetic background or external factors, e.g. lifestyle. A late consideration during the data mining process can only compensate partially – if at all – the negligence of early selection and balancing of subjects. The example of AD also illustrates the importance of documentation of clinical data since a matching of samples from AD and control groups concerning the age of subjects has to be considered for sample selection and statistical evaluation of the data.

4. Conclusions and outlook

A systematic project design is necessary to discover biomarkers by applying new peptidomic technology. The available technologies are sensitive, highly automated and of robust quality allowing reproducible analyses of semi-quantitative data. Several thousand peptide signals are commonly detected in a broad dynamic range of concentration and differential comparisons by specific software tools for data mining pinpoint biologically relevant biomarkers. The scalability of the analysis allows identification (sequencing) of biomarker candidates and their subsequent biological evaluation.

The process chain is robust and reliable enough for application in industrial or clinical settings to discover reliable biomarker candidates and perform initial validation steps to reduce false positive candidates. Moreover, the possibility to search in protein and DNA databases has made the identification much easier with high success rates. The underlying analytical methods are still in rapid development. Nevertheless, the (sub-) discipline of peptidomics is still a rather new and growing field of research with this new molecular terra incognita, which is very likely a wealthy source of new information suitable for biomarker and drug discovery as well as the monitoring of clinical studies.

Acknowledgements

The contribution of Peter Schulz-Knappe, Hans-Dieter Zucht, Rüdiger Hess, Michael Jürgens, Harald Tammen, Jens Lamerz, Gabriele Heine, Kerstin Kupke and Birgit Kemper of BioVisioN in setting up the process chain is gratefully acknowledged. The open atmosphere within discussions during the collaboration with Roche Diagnostics as industrial partner in the phase of optimising the process chain was of inspiring help. We would like to thank Andreas Dessauer, Gerda Huber, Johann Karl and Rudolf Reiter and all the other contributors in that collaboration. Finally, we appreciate very much the input of Martin Schuhmann who introduced CSF to us as an object of peptide analysis.

This work was supported in part by grants from the German Ministry for Education and Research, BMBF (FKZ 0312815), and the Lower Saxony Ministry for Economy and Technology (203.19-32329-5-354).

BioVisioN®, Peptidomics® and Differential Peptide Display® are registered trademarks of BioVisioN AG, Hannover, Germany.

References

[1] B.L. Adam, Y. Qu, J.W. Davis et al., *Cancer Res* **62** (2002), 3609–3614.

[2] R. Aebersold and M. Mann, *Nature* **422** (2003), 198–207.

[3] N.L. Anderson and N.G. Anderson, *Electrophoresis* **19** (1998), 1853–1861.

[4] N.L. Anderson, N.G. Anderson, L.R. Haines, D.B. Hardie, R.W. Olafson and T.W. Pearson, *J Proteome Res* **3** (2004), 235–244.

[5] G. Baggerman, P. Verleyen, E. Clynen, J. Huybrechts, A. De Loof and L. Schoofs, *J Chromatogr B Analyt Technol Biomed Life Sci* **803** (2004), 3–16.

[6] R. Bischoff and T.M. Luider, *J Chromatogr B Analyt Technol Biomed Life Sci* **803** (2004), 27–40.

[7] K. Blennow and H. Hampel, *Lancet Neurol* **2** (2003), 605–613.

[8] G.A. Carr, *Clin Pharmacol Ther* **69** (2001), 89–95.

[9] O. Carrette, I. Demalte, A. Scherl et al., *Proteomics* **3** (2003), 1486–1494.

[10] P. Chaurand, S.A. Schwartz and R.M. Caprioli, *Anal Chem* **76** (2004), 87A–93A.

[11] E. Clynen, G. Baggerman, D. Veelaert et al., *Eur J Biochem* **268** (2001), 1929–1939.

[12] E. Clynen, A. De Loof and L. Schoofs, *Gen Comp Endocrinol* **132** (2003), 1–9.

[13] W.A. Colburn, *J Clin Pharmacol* **43** (2003), 329–341.

[14] K. Cottingham, *Anal Chem* **75** (2003), 472A–476A.

[15] R.E. Dalbey and D.S. Sigman, *The enzymes. Vol.22: Co-and posttranslational proteolysis of proteins,* Academic Press, San Diego, 2002.

[16] E.P. Diamandis, *Clin Chem* **49** (2003), 1272–1275.

[17] N.S: Doherty, B.H. Littman, K. Reilly, A.C. Swindell, J.M. Buss and N.L. Anderson, *Electrophoresis* **19** (1998), 355–363.

[18] C. Fierens, L.M. Thienpont, D. Stockl, E. Willekens and A.P. De Leenheer, *J Chromatogr* **A896** (2000), 275–278.

[19] J. Gobom, K.O. Kraeuter, R. Persson, H. Steen, P. Roepstorff and R. Ekman, *Anal Chem* **72** (2000), 3320–3326.

[20] S.P. Gygi, Y. Rochon, B.R. Franza and R. Aebersold, *R. Mol Cell Biol* **19** (1999), 1720–1730.

[21] G. Heine, H.-D. Zucht, M.U. Schuhmann et al., *J Chromatogr B Analyt Technol Biomed Life Sci* **782** (2002), 353–361.

[22] V.Y.H. Hook, *Proteolytic and Cellular Mechanism in Prohormone and Proprotein Processing,* R.G. Landes Company: Austin, Texas, 1998.

[23] S.E. Ilyin, S.M. Belkowski and C.R. Plata-Salaman, *Trends Biotechnol* **22** (2004), 411–416.

[24] IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN), *Eur J Biochem* **138** (1984), 9–37.

[25] H. John, M. Walden, S. Schafer, S. Genz and W.G. Forssmann, *Anal Bioanal Chem* **378** (2004), 883–897.

[26] J. Karl, A. Borgya, A Gallusser at al., *Scand J Clin Lab Invest Suppl* **230** (1999), 177–181.

[27] N.L. Kelleher, *Anal Chem* **76** (2004), 197A–203A.

[28] S. Kennedy, *Biomarkers* **7** (2002), 269–290.

[29] J. Lamerz, H. Selle, L. Scapozza et al., Correlation-Associated Peptide Networks of Human Cerebrospinal Fluid, *Proteomics* **5** (2005), 2789–2798.

[30] J. Lamerz, R. Crameri, L. Scapozza, T. Mohring, H. Selle and H.-D. Zucht, *Peptide Sequence Prediction Supported by Correlation-Associated Networks in Human Cerebrospinal Fluid,* submitted for publication, 2005.

[31] N. Lundell, *J Chromatogr* **639** (1993), 97–115.

[32] D. Machin, M.J. Campbell, P.M. Fayers and A.P.Y. Pinol, *Sample Size Tables for Clinical Studies,* second edition, Blackwell Science, London, 1997.

[33] P.A. McCullough, T. Omland and A.S. Maisel, *Rev Cardiovasc Med* **4** (2003), 72–80.

[34] G.L. Miklos and R. Maleszka, *Proteomics* **1** (2001), 30–41.

[35] T. Mohring, M. Kellmann, M. Jurgens and M. Schrader, *J Mass Spectrom* **40** (2005), 214–226.

[36] C.L. Nilsson, A. Brinkmalm, L. Minthon, K. Blennow and R. Ekman, *Peptides* **22** (2001), 2105–2112.

[37] C.A. Owen and E.J. Campbell, *J Lab Clin Med* **134** (1999), 341–351.

[38] E.F. Petricoin, A.M. Ardekani, B.A. Hitt et al., *Lancet* **359** (2002), 572–577.

[39] M. Ramstrom and J. Bergquist, *FEBS Lett* **567** (2004), 92–95.

[40] J.F. Rehfeld and J.P. Goetze, *Curr Mol Med* **3** (2003), 25–38.

[41] R. Richter, P. Schulz-Knappe, M. Schrader et al., *J Chromatogr B Biomed Sci Appl* **726** (1999), 25–35.

[42] P. Roepstorff, *Curr Opin Biotechnol* **8** (1997), 6–13.

[43] C. Rohlff, *Electrophoresis* **21** (2000), 1227–1234.

[44] D.G. Rudmann and S.K. Durham, *Toxicol Pathol* **27** (1999), 111–114.

[45] M. Schrader and P. Schulz-Knappe, *Trends Biotechnol* **19** (2001), S55–S60.

[46] P. Schulz-Knappe, M. Raida, M. Meyer, E.A. Quellhorst and W.G. Forssmann, *Eur J Med Res* **1** (1996), 223–236.

[47] P. Schulz-Knappe, H.D. Zucht, G. Heine, M. Jurgens, R. Hess and M. Schrader, *Comb Chem High Throughput Screen* **4** (2001), 207–217.

[48] H. Selle, J. Lamerz, K. Buerger et al., *Identification of Novel Biomarker Candidates by Differential Peptidomics Analysis of Cerebrospinal Fluid in Alzheimer's Disease,* Combin. Chem. & High Throughput Screen, 2005, accepted for publication.

[49] N. Sewald and H.-D. Jakubke, *Peptides: Chemistry and Biology,* 1st Reprint 2003 ed., Wiley-VCH Verlag GmbH: Weinheim, 2002.

[50] D.E. Shapiro, *Stat Methods Med Res* **8** (1999), 113–134.

[51] M. Stark, O. Danielsson, W.J. Griffiths, H. Jornvall and J. Johansson, *J Chromatogr B Biomed Sci Appl* **754** (2001), 357–367.

[52] R. Strohman, *Nat Biotechnol* **21** (2003), 477–479.

[53] M. Svensson, K. Skold, P. Svenningsson and P.E. Andren, *J Proteome Res* **2** (2003), 213–219.

[54] C. Tilstone, *Nature* **424** (2003), 610–612.

[55] K.A. Tubbs, D. Nedelkov and R.W. Nelson, *Anal Biochem* **289** (2001), 26–35.

[56] P. Vicini, M.R. Gastonguay and D.M. Foster, *Crit Rev Biomed Eng* **30** (2002), 379–418.

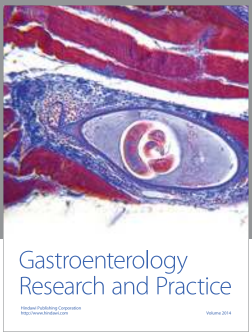
[57] W. Wang, H. Zhou, H. Lin et al., *Anal Chem* **75** (2003), 4818–4826.

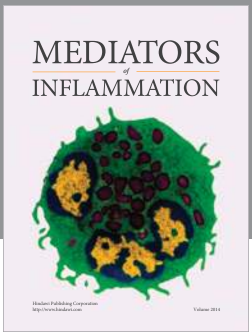
[58] J. Whicher, J. Bienvenu and G. Monneret, *Ann Clin Biochem* **38** (2001), 483–493.

[59] R. Zhang, L. Barker, D. Pinchev et al., *Proteomics* **4** (2004), 244–256.

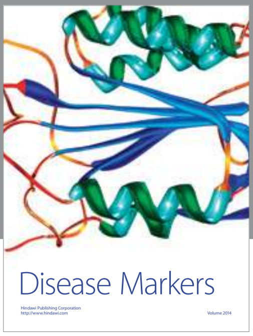
[60] J.W. Zolg and H. Langen, *Mol Cell Proteomics* **3** (2004), 345–354.

The Scientific World Journal

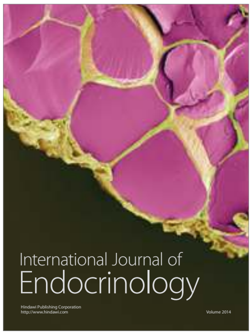

Gastroenterology Research and Practice


MEDIATORS of INFLAMMATION

Journal of Diabetes Research


Disease Markers

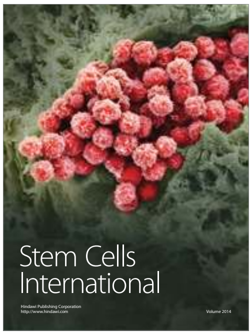
Journal of Immunology Research

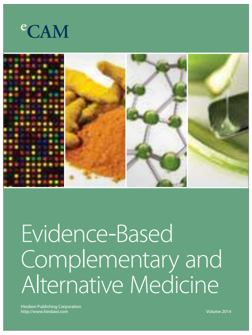

International Journal of Endocrinology

PPAR Research

BioMed Research International

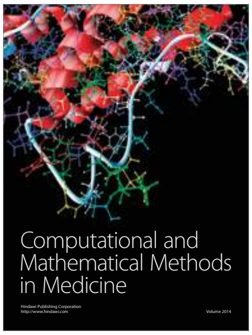
Journal of Ophthalmology

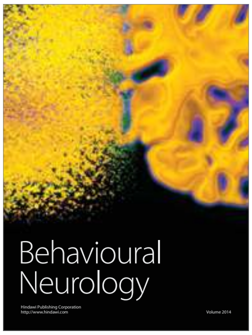

Stem Cells International


eCAM
Evidence-Based Complementary and Alternative Medicine

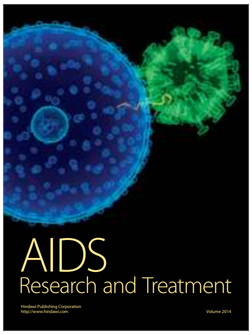
Journal of Obesity

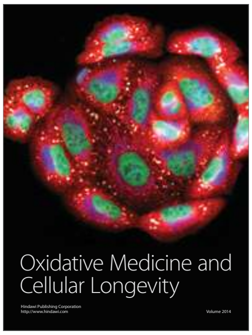
Journal of Oncology


Computational and Mathematical Methods in Medicine


Behavioural Neurology

Parkinson's Disease


AIDS Research and Treatment


Oxidative Medicine and Cellular Longevity