頭頸部腫瘍

Chemotherapy for head and neck cancer

五月女　隆
Takashi SAOTOME
癌研究会癌研有明病院化学療法科

◎頭頸部扁平上皮癌の化学療法はシスプラチンと5-FU持続静注が主体となるため，外来化学療法の機会は多くない．近年タキサン系薬剤の導入に伴い，従来のシスプラチン・5-FU併用療法にドセタキセルを加える試みがなされ，術前化学療法としての検討で良好な成績が得られている．ドセタキセルは本来外来で投与可能な薬剤であり，シスプラチンについては大量補液のための連日の通院あるいは代替薬としてのカルボプラチンやネダプラチンの使用，5-FUに関してはインフュージョンポンプの使用あるいは代替薬としての経口薬(UFT，S-1)の使用など，入院治療から外来治療に切り替えようとする試みがなされているが，抗腫瘍効果や生存期間への寄与が証明されないかぎり薬剤の軽率な切り替えは慎むべきである．本稿では将来の外来化学療法への移行を見据え，頭頸部扁平上皮癌の化学療法を組み込んだ標準的治療戦略の現状について述べる．

Key word　頭頸部腫瘍，扁平上皮癌，シスプラチン，5-FU，ドセタキセル

頭頸部腫瘍にはおもなものとして鼻腔・副鼻腔～口腔～咽喉頭～食道に発生する扁平上皮癌と甲状腺癌がある．比較的まれなものとして唾液腺癌(耳下腺・顎下腺・舌下腺)があり，ほとんどが腺癌である．甲状腺癌は手術が中心となるため，化学療法は姑息的な役割に限られる．唾液腺癌や甲状腺癌に対し化学療法を行う機会は多くなく，またいまのところ標準化学療法が存在しないため，本稿では頭頸部扁平上皮癌に対する化学療法について述べる．

薬剤の選択

欧米のデータであるが，頭頸部扁平上皮癌に対する単剤の第Ⅱ相試験レベルの有効率は，メソトレキセート31%，ブレオマイシン21%，シスプラチン28%，カルボプラチン22%，5-FU 15%，イホスファミド23%，パクリタキセル15～40%，ドセタキセル30～33%，ビノレルビン18%，ゲムシタビン13%，トポテカン14%となっている[1]．頭頸部扁平上皮癌に対する化学療法は消化器癌と同様，まずフッ化ピリミジン系薬剤が中心となり発

サイドメモ　頭頸部癌に対する分子標的療法

最近，標準的薬物療法の一部として組み込まれつつある分子標的薬剤は，ゲフィチニブ，イマチニブなどの経口薬のみでなく，トラスツズマブ，リツキシマブなどの注射剤もその投与の簡便さや軽度の毒性から外来で用いられることが多い．頭頸部癌では腫瘍増殖のメカニズムにおいて上皮成長因子受容体(epidermal growth factor receptor：EGFR)が乳癌，大腸癌などと同様に重要な役割を果たしていることが判明した．EGFRを標的とした薬剤がいくつか開発され，頭頸部癌にも応用すべく臨床試験が組まれた．EGFRに対する抗体であるcetuximab(わが国未発売)は単剤での試験，進行・再発例に対するCDDPとの併用，放射線治療との併用について検討され，有効性が証明された．その結果を受け，アメリカFDA(food and drug administration)では頭頸部癌に対する放射線治療との同時併用，標準療法後セカンドラインとしての使用を認めた．EGFRチロシンキナーゼの阻害物質のゲフィチニブやerlotinib(わが国未発売)，血管新生阻害薬であるベバシズマブの効果についても報告があいついでいる．

perspectives Journal of proteome research

Personalized Medicine and Proteomics: Lessons from Non-Small Cell Lung Cancer

György Marko-Varga,[†,‡] Atsushi Ogiwara,[†,§,∥] Toshihide Nishimura,[§,∥] Takeshi Kawamura,[§,∥] Kiyonaga Fujii,[§,∥] Takao Kawakami,[§,∥] Yutaka Kyono,[∥] Hsiao-kun Tu,[∥] Hisae Anyoji,[∥] Mitsuhiro Kanazawa,[∥] Shingo Akimoto,[∥] Takashi Hirano,[⊥] Masahiro Tsuboi,[⊥] Kazuto Nishio,[§] Shuji Hada,[#] Haiyi Jiang,[×] Masahiro Fukuoka,[△] Kouichiro Nakata,[◆] Yutaka Nishiwaki,[+] Hideo Kunito,[§] Ian S. Peers,[◇] Chris G. Harbron,[◇] Marie C. South,[∞] Tim Higenbottam,[▽,£] Fredrik Nyberg,[*,▽,☆] Shoji Kudoh,[ƒ] and Harubumi Kato[⊥]

Respiratory Biological Sciences, AstraZeneca R&D Lund, SE-221 87 Lund, Sweden, Clinical Proteome Center, Tokyo Medical University, Shinjuku Sumitomo Building 17F, 2-6-1 Nishishinjuku, Shinjuku, Tokyo 163-0217, Japan, Medical ProteoScope Company, Limited, Shinjuku Sumitomo Building 17F, 2-6-1 Nishishinjuku, Shinjuku, Tokyo 163-0217, Japan, Department of Surgery, Tokyo Medical University, 6-7-1 Nishishinjuku, Shinjuku, Tokyo 160-0023, Japan, Clinical Division, Research & Development, AstraZeneca K.K., Umeda Sky Building Tower East, 1-88, 1-chome, Ohyodo-naka, Kita-ku, Osaka 531-0076, Japan, Clinical Science Department, Research & Development, AstraZeneca K.K., Umeda Sky Building Tower East, 1-88, 1-chome, Ohyodo-naka, Kita-ku, Osaka 531-0076, Japan, Department of Medical Oncology, Kinki University School of Medicine, 377-2,Ohno-higashi, Osakasayama-city 589-8511, Osaka, Japan, Department of Respiratory Diseases, Toho University School of Medicine, 6-11-1, Omori-nishi, Ota-ku, Tokyo 143-8541, Japan, Dept. of Thoracic Oncology, National Cancer Centre Hospital East, 6-5-1, Kashiwanoha, Kashiwa-city, Chiba 277-8577, Japan, Statistical Sciences, AstraZeneca R&D Alderley Park, Cheshire, UK, Cancer & Infection Statistics, AstraZeneca R&D Alderley Park, Cheshire, UK, Medicine & Science, AstraZeneca R&D Charnwood, Loughborough LE11 5RH, Leicestershire, UK, Sheffield University, Sheffield, UK, Epidemiology, AstraZeneca R&D Mölndal, SE-431 83 Mölndal, Sweden, Institute of Environmental Medicine, Karolinska Institute, Box 210, SE-171 77 Stockholm, Sweden, and 4th Department of Internal Medicine, Nippon Medical School, 1-1-5, Sendagi, Bunkyo-ku, Tokyo 113-8603, Japan

Received January 26, 2007

Personalized medicine allows the selection of treatments best suited to an individual patient and disease phenotype. To implement personalized medicine, effective tests predictive of response to treatment or susceptibility to adverse events are needed, and to develop a personalized medicine test, both high quality samples and reliable data are required. We review key features of state-of-the-art proteomic profiling and introduce further analytic developments to build a proteomic toolkit for use in personalized medicine approaches. The combination of novel analytical approaches in proteomic data generation, alignment and comparison permit translation of identified biomarkers into practical assays. We further propose an expanded statistical analysis to understand the sources of variability between individuals in terms of both protein expression and clinical variables and utilize this understanding in a predictive test.

Keywords: personalized medicine • gefitinib • therapy • interstitial lung disease • non-small cell lung cancer • biomarkers • predictive test • mass spectrometry • statistical analysis • proteomics

* To whom correspondence should be addressed. Epidemiology, AstraZeneca R&D Mölndal, SE-431 83 Mölndal, Sweden; Tel, +46 31 706 5203; Fax, +46 31 776 3828; E-mail, Fredrik.Nyberg@astrazeneca.com.

† György Marko-Varga and Atsushi Ogiwara made equal contributions to this manuscript.

‡ Respiratory Biological Sciences, AstraZeneca R&D Lund.

§ Clinical Proteome Center, Tokyo Medical University.

∥ Medical ProteoScope Co., Ltd.

⊥ Department of Surgery, Tokyo Medical University.

Clinical Division, Research & Development, AstraZeneca K.K.

× Clinical Science Department, Research & Development, AstraZeneca K.K.

△ Department of Medical Oncology, Kinki University School of Medicine.

◆ Department of Respiratory Diseases, Toho University School of Medicine.

+ Dept. of Thoracic Oncology, National Cancer Centre Hospital East.

◇ Statistical Sciences, AstraZeneca R&D Alderley Park.

∞ Cancer & Infection Statistics, AstraZeneca R&D Alderley Park.

▽ Medicine & Science, AstraZeneca R&D Charnwood.

£ Medical School, Sheffield University.

☆ Epidemiology, AstraZeneca R&D Mölndal.

ƒ Institute of Environmental Medicine, Karolinska Institute.

‡ Fourth Department of Internal Medicine, Nippon Medical School.

10.1021/pr070046s CCC: $37.00

Published on Web 07/17/2007

Introduction

A personalized medicine approach uses appropriate biomarkers to select treatments best suited for an individual patient and disease phenotype. A multiple biomarker approach (e.g., proteomics) has the advantage over conventional single biomarkers of combining many different pieces of information. Here, we review the key features of state-of-the-art proteomic profiling and introduce recent analytic developments to build a proteomic toolkit for use in personalized medicine, and we describe how these may be applied in a viable method for exploiting predictive proteomic fingerprints in the clinic. The potential of our proteomics toolkit hopefully brings us one step closer to a practical personalized medicine.

Cancer therapy is moving toward individually selected treatments, chosen not only according to tumor cell type but also based on the patient's predicted responsiveness to different classes of therapy or susceptibility to therapeutic adverse events. This emerging personalized medicine approach draws on both genotype and phenotype information, including protein expression. To implement personalized medicine, we need to develop effective biomarker tests predictive of response to treatment or susceptibility to adverse events. The benefits of personalized medicine are exemplified by considering interstitial lung disease (ILD) among non-small cell lung cancer (NSCLC) patients, which is associated with various kinds of chemotherapy treatment. A personalized medicine approach, using a simple blood test to predict those NSCLC patients at risk of developing ILD, would clearly be of great value.

We review current thinking and present some novel developments in a number of areas that have to be integrated to develop and then practically apply such tests in a clinical setting:

- The large scale collection of reliable and high quality phenotypic and clinical data and blood samples.
- Protein analysis in blood.
- Data acquisition, handling, combining and analysis.
- Interpretation and utilization of results in a clinical setting.

Clinical Background

A Motivating Example: Gefitinib (IRESSA) Treatment of NSCLC. The concepts of proteomics-based personalized medicine discussed in this article are very generally applicable. A motivating example that we will refer to in order to illustrate the potential benefits of personalized medicine is ongoing work in attempting to develop a simple blood test to address the potential occurrence of ILD in seriously ill NSCLC patients, the target group for the NSCLC treatment gefitinib.

Gefitinib is a "small molecule" inhibitor of the enzyme tyrosine kinase of the epidermal growth factor receptor (EGFR) family, such as erbB1. It is an approved therapy for advanced NSCLC in many countries and offers important clinical benefits (tumor shrinkage and improvement in disease-related symptoms) for "end-stage" patients. The large phase III ISEL (IRESSA Survival Evaluation in Lung Cancer) trial demonstrated some improvement in survival with gefitinib which failed to reach statistical significance compared with placebo in the overall population and in patients with adenocarcinoma.[1] However, in preplanned subgroup analyses, a significant increase in survival was shown with gefitinib in patients of Asian ethnicity and in patients who had never smoked.[1]

Analysis of the biomarker data from a subset of patients in the ISEL study suggested that patients with pretreated advanced NSCLC who have tumors with a high EGFR gene copy number (detected by fluorescent in situ hybridization [FISH]) have a higher likelihood of increased survival when treated with gefitinib compared with placebo.[2] Increased HER2 gene copy number has also been seen in tumors from patients who are responsive to gefitinib.[3] Somatic-activating mutations of EGFR in tumor tissue have also been associated with increased gefitinib responsiveness in patients with NSCLC.[4-7] Such mutations are more commonly found in tumor samples from patients of Asian origin and non-smokers.[8]

Following the ISEL subgroup analyses, and the biomarker evidence, it has become important to clarify which patients are more suitable for treatment with gefitinib. Analyses for both somatic-activating mutations and gene copy number require tumor tissue, which is not always available from the time of diagnosis; therefore, a blood test may represent a more versatile option and be of great value to clinicians.

With respect to tolerability, the search for a blood test that might include both genetic and proteomic biomarkers to define patients at risk of adverse effects from a drug, for example interstitial lung disease with gefitinib, is a focus of research.

Interstitial Lung Disease as a Complication in NSCLC Patients. ILD is a disease that afflicts the parenchyma or alveolar region of the lungs.[9] The alveolar septa (the walls of the alveoli) become thickened with fibrotic tissue. Associated with drug use, it can present precipitously with acute diffuse alveolar damage (DAD). The lungs show so-called "ground glass" shadowing on chest radiology, and patients complain of severe breathlessness. There are no effective treatments but patients can be supported by oxygen supplementation, corticosteroid therapy, or assisted ventilation. The process of alveolar damage is however fatal in some patients. ILD is a comorbidity in patients with NSCLC.[10-16] Both diseases are associated with cigarette smoking,[17-20] and ILD is also considered to be associated with various kinds of lung cancer chemotherapy.[21-26]

In the ISEL study of gefitinib in NSCLC mentioned above, ILD-type events occurred in 1% of both placebo and gefitinib-treated patients.[1] Most ILD-type events occurred in patients of Asian origin, where placebo and treated patients had similar prevalences of respectively 4% and 3%. The rate observed in the gefitinib-treated arm was in line with earlier safety data from Japan and a large gefitinib post-marketing surveillance study in Japan (3322 patients), where the reported rate of ILD-type events was 5.8%.[27]

A simple blood test to predict the potential occurrence of ILD in seriously ill NSCLC patients before initiating treatments would clearly be of great value. This article describes the personalized medicine approach, which could be used to provide such a test. Consequently, the proteomics objectives of the preliminary phase of the study we describe were to verify the protein expression alterations in blood plasma from case patients (who developed ILD) and control patients (without ILD) treated by gefitinib, using a liquid chromatography–mass spectrometry/mass spectrometry (LC–MS/MS) proteomics platform.

Data and Sample Collection

To develop a personalized medicine test, it is essential to have access to an adequately sized collection of high quality tissue samples on which to perform proteomics analysis, with corresponding reliable diagnostic and clinical data.

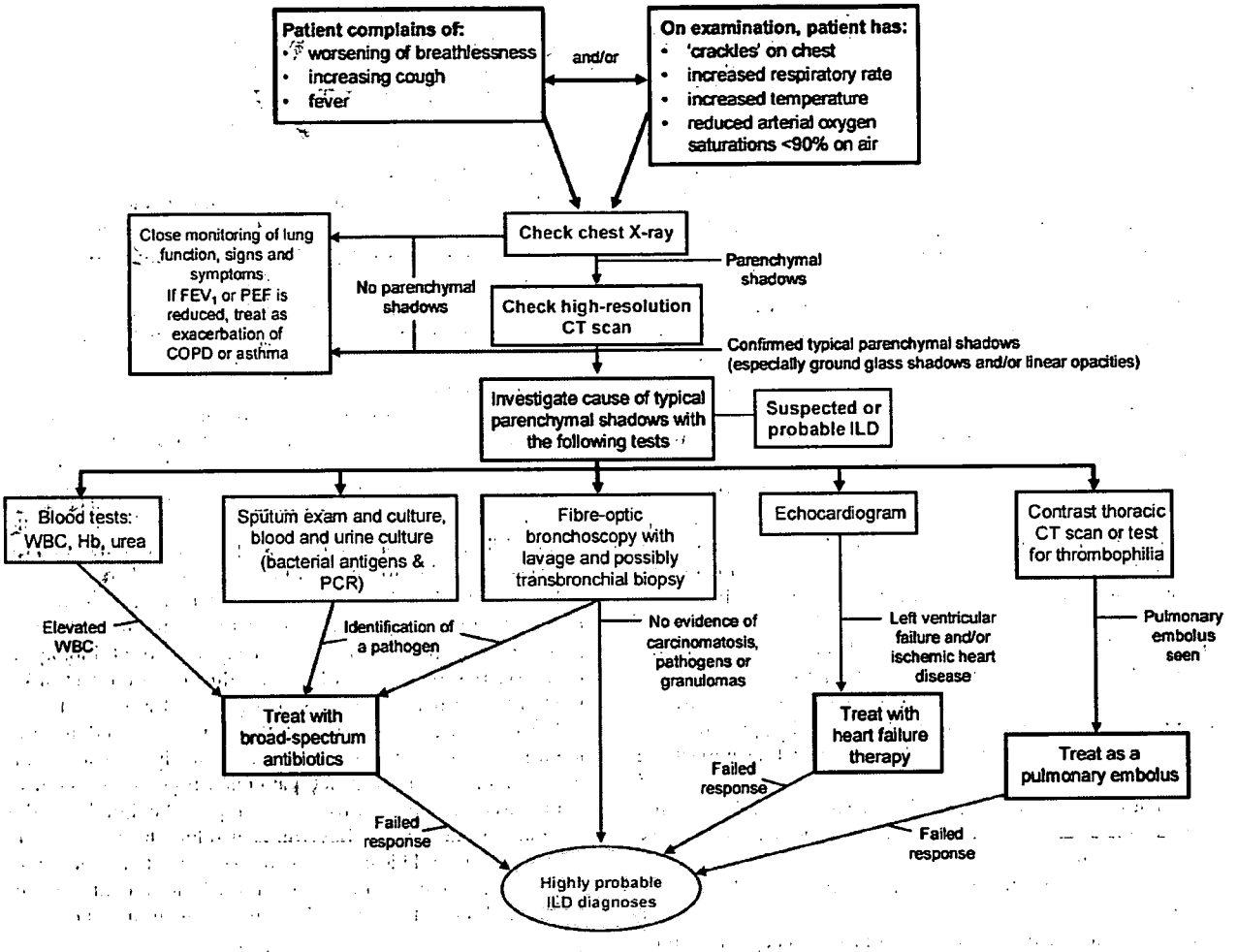


Figure 1. Algorithm for diagnosis of interstitial lung disease (ILD) in non-small cell lung cancer (NSCLC) patients.

As an example, in our work with gefitinib, samples were taken after obtaining informed consent from a nested case-control study, i.e., a case-control study performed within a prospective pharmacoepidemiological cohort of several thousand patients with advanced or recurring NSCLC who had received at least one prior chemotherapy regimen, and who were to be treated with gefitinib or chemotherapy. The main objective of this study was to measure the relative risk of ILD in Japanese patients with NSCLC using gefitinib compared with conventional therapy, with the associated aims of determining the incidence rate of ILD in late stage NSCLC patients and the principal risk factors for this complication.

Central to both the case-control study and the proteomics analysis was the use of internationally agreed criteria for the diagnosis of ILD and an algorithm of diagnostic tests to exclude alternative diseases.[28] Principal investigators in the study were asked to assess all patients for possible ILD using the diagnostic algorithm (Figure 1). Two case review boards of experts from oncology, radiology, and pulmonary medicine were set up to independently establish a consistent final diagnosis of ILD. In addition, extensive standard clinical and demographic risk factor data were collected on all registered cases and controls.

This degree of rigor in establishing accurate phenotypic diagnosis is critical to develop a robust and reliable personalized medicine test, as inaccuracies at this stage will affect all subsequent data analyses. The availability of clinical and risk factor data, and a rigorous epidemiological study design setting for the collection of proteomics samples is also of great value to fine-tune the statistical analysis.

Is Proteomics Ready for Personalized Medicine Applications?

The Human Proteome Map in Plasma. The impetus to develop personalized medicine based on blood samples has encouraged proteomic profiling that identifies individual proteins and multiple "fingerprint" protein patterns. A remaining limitation has been the lack of integration of the technology of protein separation with bioinformatics and statistical methods. Extensive national and international[29,30] collaborations are being implemented to address some of these needs. An important component in this development is the Human Proteome Organization (HUPO; www.HUPO.org), a scientific consortium that supports various programmes to map the proteins expressed in various human tissues, disease states, etc.[31–33] One of these is the Plasma Proteome initiative started in 2002, aiming to annotate and catalog the many thousands of proteins and peptides[34–37] of the human plasma proteome. Recently results from the pilot phase with 35 collaborating laboratories from 13 countries[38–42] and multiple analytical

groups were made publicly available on the Internet (www.bioinformatics.med.umich.edu/hupo/ppp; www.ebi.ac.uk/pride). The combined efforts have generated 15 710 different MS/MS datasets that were linked to the International Protein Index (IPI) protein IDs, and an integration algorithm applied to multiple matches of peptide sequences yielded 9504 IPI proteins identified with one or more peptides[40] and characterized by Gene Ontology, InterPro, Novartis Atlas, and OMIM. Such advances provide an important platform for transforming proteomics from a technology to a useful biomarker tool applicable to personalized medicine.

Protein Analysis in Blood—The Methods. With respect to automated studies, multidimensional chromatography is the main technology used for protein analysis in blood. It is coupled to mass spectrometry either by electrospray ionization (ESI) for analysis in solution or matrix assisted laser desorption/ionization (MALDI) in solid phase applications.[39,41,43-47] Alternatively, ion-trap mass spectrometers are gaining recognition for high-throughput sequencing.[46,48-53] Linking a Fourier transform ion cyclotrone resonance (FTICR) unit to the linear trap can increase the resolution profoundly,[36,54-56] one of several novel principles for strengthening the assignment of protein annotations with the most commonly used protein search engines.[36,47,54-61] For protein annotation, the recent development of a human protein reference database complements these technologies.[61] Studies of protein expression in a variety of biological compartments ranging from sub-cellular to whole organisms have been undertaken with these analytic approaches.[62-70] Some key findings from the HUPO initiatives that impact on methodology include:

- For studies using blood samples, plasma rather than serum is preferred, with ethylenediaminetetraacetic acid (EDTA) as an anticoagulant.[40]
- The abundant proteins in plasma should be depleted prior to analysis.[40]
- Acceptance of protein annotation, i.e., accepted protein identities[39,40] should use standard criteria. These include having two identified peptide sequences from each protein, both with a statistical significance score high enough to ensure a correct sequence confirmation when compared with the corresponding gene sequence entity.[39]

Despite the advances in methodology, important hurdles to using proteomics in a personalized medicine context remain.

Protein Expression Analysis in Blood—Some Important Hurdles. Although protein profiling technology is highly automated and interfaced with database search engines to relate peptide sequences to protein identities and function,[39,40] there are many practical reasons why determining the relative abundance of proteins relevant for prediction purposes is difficult:

- About 90% of proteins are believed to be present only in low copy numbers, i.e., at medium and low abundance levels.[49]
- There can be variation both in the quantity and form of protein expression within normal physiological function.
- Between 300 000 and 3 million human protein species exist as direct gene products or post-translational modifications.[44]
- The relative abundance of the post-translational modifications occurring within the cell is called a Cell-Protein-Index Number (CPIN).[29,30] As an example, if one considers that there are 30 types of phosphorylation variants of a single phosphoprotein, and a hundred possible fold forms of glycosylation of a single glycoprotein, the theoretical CPIN varies considerably depending on the sample complexity.
- The dynamic range of protein expression within cells, between levels of most and least abundant proteins, is in the order of 10^8–10^{10}.[34-36]
- In a typical clinical proteomics study the total cellular protein material in a sample seldom exceeds 10–20 milligrams. Therefore, the least abundant proteins would be present at starting levels not exceeding picograms.
- Recent studies use technology that can identify several thousand proteins in plasma samples,[29] but this still probably only represents a small fraction of the intermediate and processed protein forms. This is due to the current limitation of mass spectrometry not being able to ionize all amino acid sequences and protein modifications with equal efficiency. In most situations, a limited region of the full length protein is sequence annotated.
- The detection of differences in protein expression between groups of interest (e.g., cases and controls) takes place against a background of high variation between individuals within a group, within individuals over time and possible analytic run-to-run variation. Any method used to address this hurdle (which will involve "alignment" for spectral methods) directly impacts the ability to find good protein biomarkers.

Beyond the hurdles above, the fundamental challenge of protein biomarkers is to link the relative abundance of single markers or a fingerprint to clinically important biological processes based on some direct or indirect cause-effect link[29] related to normal or aberrant biological pathways.[47,49] In the following sections, we present the approach used for the identification of protein biomarkers potentially associated with development of ILD in NSCLC patients within the case-control study used as our motivating example. We build on the foundations described above and introduce further analytic developments and ideas relating to proteomic data generation, assaying and alignment to build a proteomics toolkit that can be applied today for personalized medicine approaches.

A State of the Art Clinical Biomarker Analysis System

In the previous section, we described several challenges in proteomic analysis. Here we describe a system and analysis approaches that we have successfully implemented to address some of these issues.

The Components of the Analysis System. The analysis system (Figure 2) uses liquid chromatography-based high-resolution separation of peptides with an interface to tandem MS/MS, a technology which has been attracting great attention as the "shotgun" method of proteome analysis.[44,68-70] With this technology, after depletion of albumin and immunoglobulin G (IgG), all extracted plasma proteins are digested into their specific peptide components by proteolytic enzyme treatment.

The generated peptides are subjected to capillary reverse-phase submicro- to micro-flow liquid chromatography (capillary RP μLC), separated by retention times due to their physicochemical properties, and then detected and sequenced by a linear ion-trap tandem mass spectrometer[71] (LTQ, Thermo Fisher Scientific, San Jose, CA) interfaced with a spray needle tip for ESI of peptides.[70] A two-dimensional quadrupole ion trap mass spectrometer[71] is used, operated in a data-dependent acquisition mode with operational m/z range limits set at 450–2000 (Figure 3, graphs A and B). Automatic switching to MS/MS acquisition mode is made in 1-second scanning cycles, controlled by the XCalibur software. The actual differences between annotated peptide fragment peaks shown in Figure 3, graph C, correspond to the amino acid residue mass, i.e.,

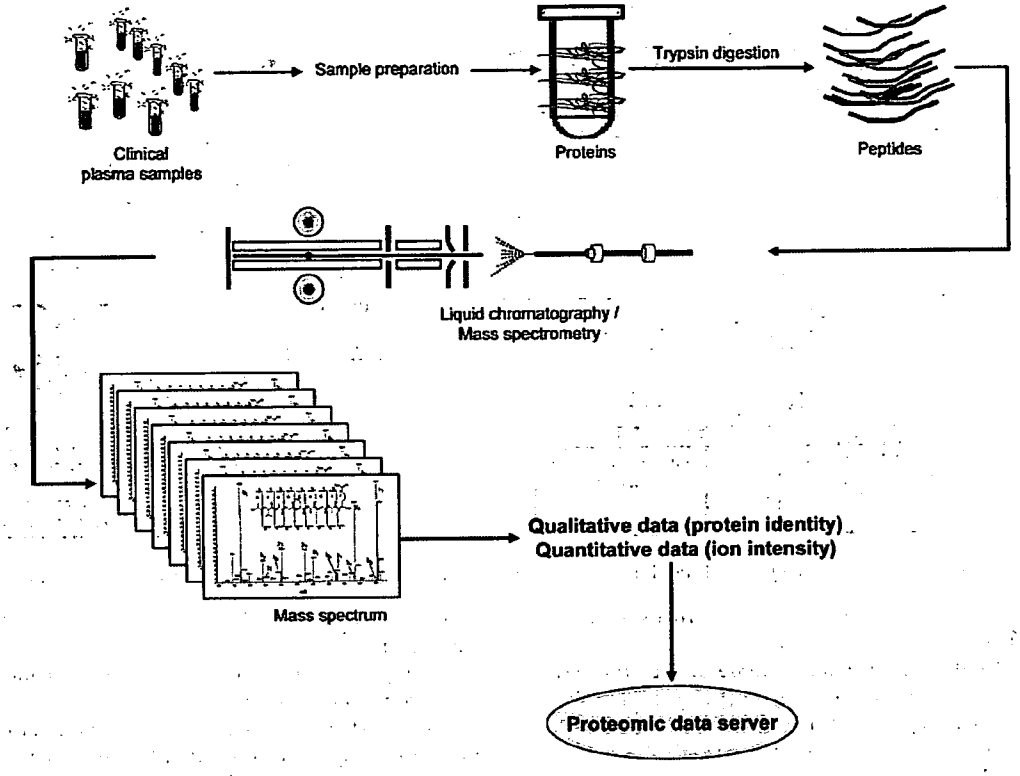


Figure 2. Schematic illustration of the clinical proteomics screening process.

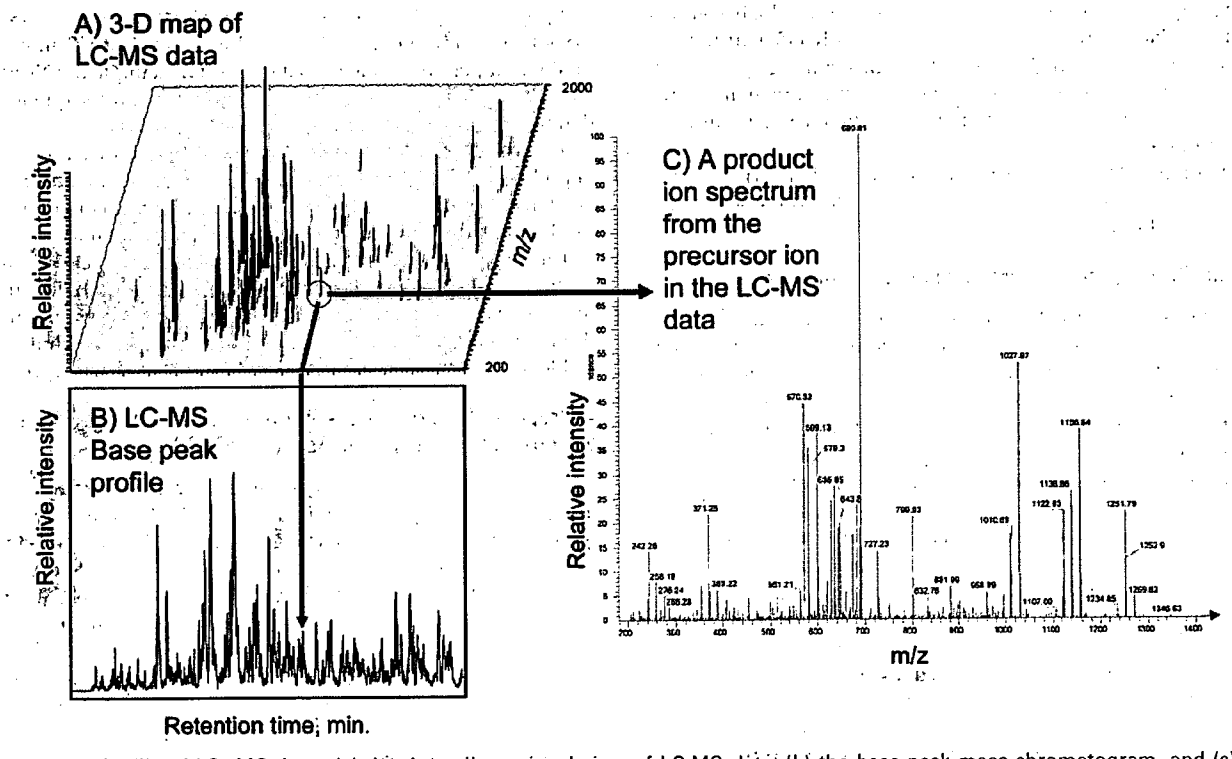


Figure 3. Profile of LC–MS data: (a) the three-dimensional view of LC-MS data; (b) the base-peak mass chromatogram, and (c) a product ion spectrum measured for a precursor ion in data-dependent acquisition mode (with MS acquisition operational *m/z* range set at 450–2000).

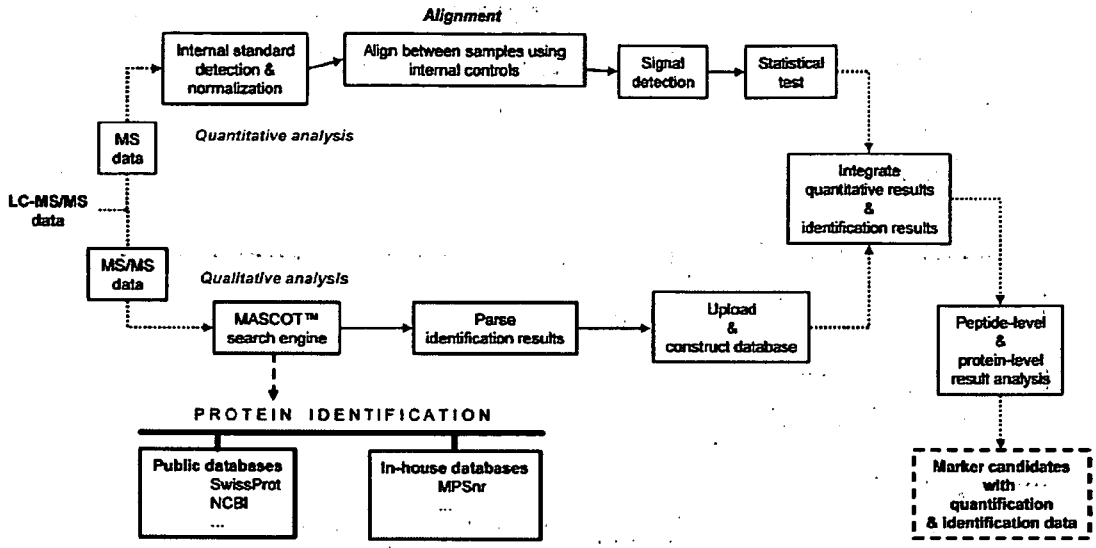


Figure 4. Overview of the data acquisition and database mining process developed within the gefitinib biomarker study.

identify the correct amino acid sequence. Internal standards are used for alignment of retention-times.

How the Methodology Overcomes Some of the Hurdles. The system described above addresses some of the hurdles noted previously. The digestion of all extracted plasma proteins into peptides will reduce the complexity by combining high-resolution nanoflow chromatographic fractionation with the separation power of modern mass spectrometry, performing automated and unattended shotgun sequencing in plasma.[35] Peptides are also more soluble and easier to handle than intact proteins. In addition, the two-dimensional quadrupole ion trap mass spectrometer[71] operates with a high-volume quadrupole electric field that makes it highly efficient to trap ions. The result is high sensitivity, high scanning speed, and better quantification over a wide dynamic range in comparison with the conventional three-dimensional ion-trap instruments.

Finding signals against a background of high variation is a further challenge, and the next section describes some approaches for addressing these.

Initial Data Handling, Processing, and Analysis

Proteomic data analysis process can be considered as consisting of two components (Figure 4). *Quantitative analysis* is used to discover significant differences in peptide signal intensities by comparing two (or more) sample groups. This process uses data collected from an entire MS run to quantify the amount of peptide ions by their respective ion signal intensity. *Qualitative analysis* is used to identify the amino acid sequence of each peptide ion, from the respective product ion spectra. To maximize their value, the results from the two component analyses should be considered in combination.

A typical quantitative analysis may consist of several steps:

1. Normalization: To account for differences in the original sample concentrations. Typically, the total signal intensity is scaled to a constant value for each analyzed sample.

2. Alignment: Correcting for nonlinear fluctuation in retention time between different samples. A variety of methodologies are available for aligning LC–MS data sets. We have found the i-OPAL algorithm (Patent # WO 2004/090526 AI), which is based on the single linkage clustering algorithm[72] and which makes use of internal standard signals, to perform well. Other alignment algorithms include xcms.[73]

3. Peak picking or signal detection: Identifying individual peptide ions within the data.

4. Identify discriminating peptides: A number of methods can be used, often in combination. A common approach is to apply a Student's *t*-test and select peptides which are significant, i.e., with a *p*-value less than the chosen cutoff, and which also show a fold-change or intensity ratio greater than another criterion. Further developments of this aspect are discussed in the Principled Statistical Analysis section.

A popular choice for qualitative analysis is the MASCOT MS/MS ion search program.[74] This may be run against a number of different peptide sequence databases, for example the NCBI Nr, Refseq, Gene Ontology, HUGO, and Swiss-Prot sequence databases. The results of the quantitative analysis can then be combined with the qualitative analysis so that, for example, a peptide must be both discriminating and have annotation—i.e., have achieved a high MASCOT score showing confidence in identification—to be considered a candidate biomarker.

The approaches we have discussed above are focused on finding potentially discriminating proteins of clinical utility. In the following section, we describe the next stage in our thinking, namely how we could rapidly deploy in the clinic a viable method for exploiting a predictive proteomic fingerprint.

A Proposal for Proteomics in the Clinical Setting: Mass Spectrometric Biomarker Assays - MSBA

Although today's technology allows for high-throughput analyses of many proteins rather than a single protein,[30] the details of how such multiplexing assays will be adapted for clinical use have not been well clarified. The Mass Spectrometric Biomarker Assay (MSBA) platform described here was conceived as one example of a rapid and seamless method to progress from identification of a diagnostic more directly to a clinically useful test. MSBA requires only a minute sample amount (5–20 μL) to obtain a read-out from a handful of quantified protein biomarkers (typically 3–35) and automatically analyzes proteins using liquid-phase separation and tandem mass spectrometry with simultaneous quantitation and identification.

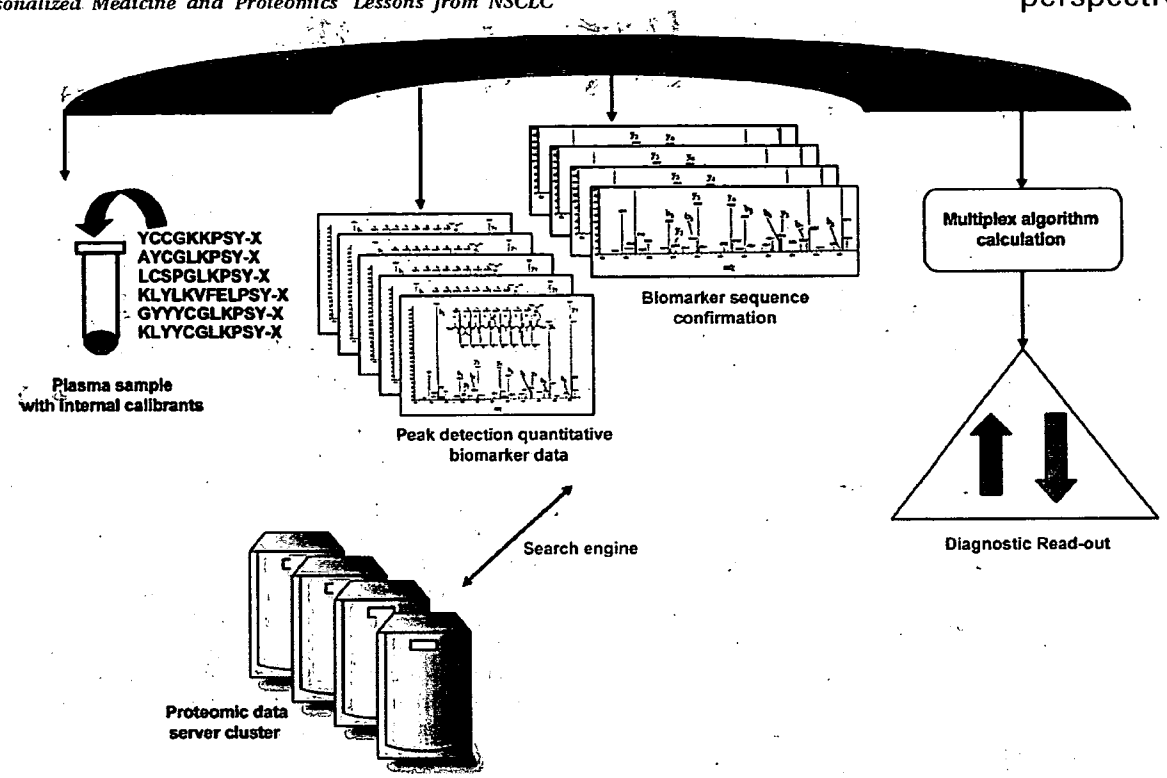


Figure 5. Entire flow of the operational components of Mass Spectrometric Biomarker Assays (MSBA).

The MSBA builds on a pre-defined Multiplex Biomarker list, which is stored within the MSBA database. Each marker entity has the values of masses and the relative retention time index with tolerance parameters. In running a patient sample, the predefined biomarker list is scanned to pick up patient sample signals that match with one of the predefined biomarker signals by satisfying the tolerance criteria (in general ±1 for m/z value and ±2% for relative retention time index). The selected candidate signals are further confirmed using the product ion spectrum. That is, the product ion spectrum is represented as a vector by binning (grouping) the m/z ratio values. Using the cosine correlation between the sample vectors and the reference vectors, we can confirm whether the selected candidate signals are truly assigned as target biomarkers. (A standard threshold value of the cosine correlation is 0.8.)

The process steps within the MSBA cycle are outlined in Figure 5. The calculation of the final multiplex biomarker assay read-out from all of the individual markers can be performed by a variety of applications, as discussed in more detail in the Principled Statistical Modeling Approach section. Figures 6A and B illustrate one approach, calculating a distance score which indicates to what extent a measured sample is distant from the case or control template in terms of predefined multiplex biomarkers.

$$S_{\text{case or control}} = \sqrt{\left[\frac{1}{n(n-2)}\right]\left[n\sum_i y_i^2 - \left(\sum_i y_i\right)^2 - \frac{\left[n\sum_i x_i y_i - \left(\sum_i x_i\right)\left(\sum_i y_i\right)\right]^2}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2}\right]}$$

If the ratio of S_{case} and S_{control} exceeds an MSBA threshold parameter, then the test sample is predicted to be a patient susceptible to develop ILD (ILD case); if not, the test sample is predicted to be a non-susceptible patient (control). We are currently evaluating the MSBA approach in practice.

A Principled Statistical Modeling Approach

We have described an analytical approach based on proteomic data, with various novel developments. However, additional insight is needed to further improve model discrimination and to broaden the focus from the proteomic data to the ultimate goal of prediction using combinations of data. Statistical analysis can be used to provide further refinement by combining information from the full clinical and laboratory datasets.

An advantage of a multiple biomarker approach (e.g., proteomics) compared with standard single biomarker development is the capability to combine information from many different entities. An example is illustrated in Figure 7A. Considering each biomarker alone fails to separate the two groups of subjects, as there is considerable overlap for both biomarkers. Use of two biomarkers in combination completely separates the two groups.

We can also use clinical variables to advantage in the analysis of the peptide patterns. For example, the efficacy of gefitinib appears to be greater in non-smokers, women, patients of Asian origin, and patients with adenocarcinomas.[8] Figure 7B illustrates how, instead of two protein biomarkers, the combination of clinical data (e.g., age) and a proteomic biomarker is able to separate two groups.

On this basis, we propose using a principled statistical analysis approach to first explore and understand the data and

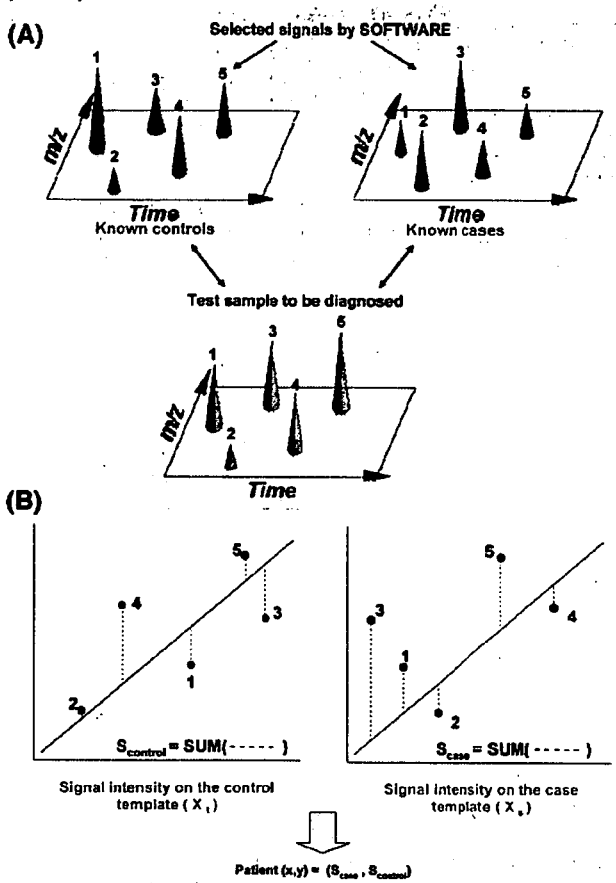


Figure 6. (A) Peptide signal comparison that MSBA (Mass Spectrometric Biomarker Assays) performs of the generated ions from the sample. The comparison is made both with the pattern of the controls and with the pattern of the case group for the corresponding signals. (B) Illustration of the regression model application of the MSBA where control templates and case templates are compared to that of the sample template generated in the analysis process.

then to model it and understand the quality of any models produced. A first step is to perform exploratory data analysis (EDA), for example using principal components analysis (PCA), to understand the major sources of data variation and the covariation between clinical parameters and protein intensity measures. The next step is univariate modeling for each protein marker individually, for example using analysis of covariance (ANCOVA), and an assessment of the effect of clinical parameters across the whole set of protein biomarkers using, for example, the False Discovery Rate as a tool.[75] This provides an understanding of key clinical variables and sources of variation within the data.

The next step is to perform multivariate predictive modeling using the proteins and clinical variables identified as being potentially important. There are a number of mathematical methods described in the literature for performing supervised classification, for example Support Vector Machines,[76] Random Forests,[77] PAM,[78] all of which have been successfully applied to high dimensional genomics data.[79] It remains an important unanswered question which modeling approach, or combination of modeling approaches, will generate the most predictive and robust models for data generated using this technology within a prospective study of this design.

Finally, to confirm that we have a practical prediction, the predictive power of a model must be assessed on a different set of patients from that used to generate the model. There are a number of approaches for external validation given a limited size dataset, for example the sequential approach of building a model based upon currently available data and testing on data from new patients when they become available, or withholding an arbitrary selection of subjects from the modeling as a test set and testing the model on these subjects. Internal validation approaches such as cross-validation or related bootstrapping methods may also be useful to assess the model selection *procedure*, but tend to overestimate the performance of a specific predictive model in subsequent external validation.[80,81] The key properties to consider when selecting an assessment method are to ensure that it will provide both precise and unbiased information regarding the prediction error rate of the potential model to be tested for clinical use. As well as assessing an overall predictive rate, it is also useful to separately assess the predictive rate for both the cases and controls and to consider the relative costs of making these false predictions within a clinical setting. Finally, the prevalence of the condition in question (here ILD) is also a critical factor in estimating what proportion of people predicted to be at risk are truly at risk, and this should also be borne in mind when evaluating a model for potential clinical use. The recently published FDA concept paper on drug-diagnostic co-

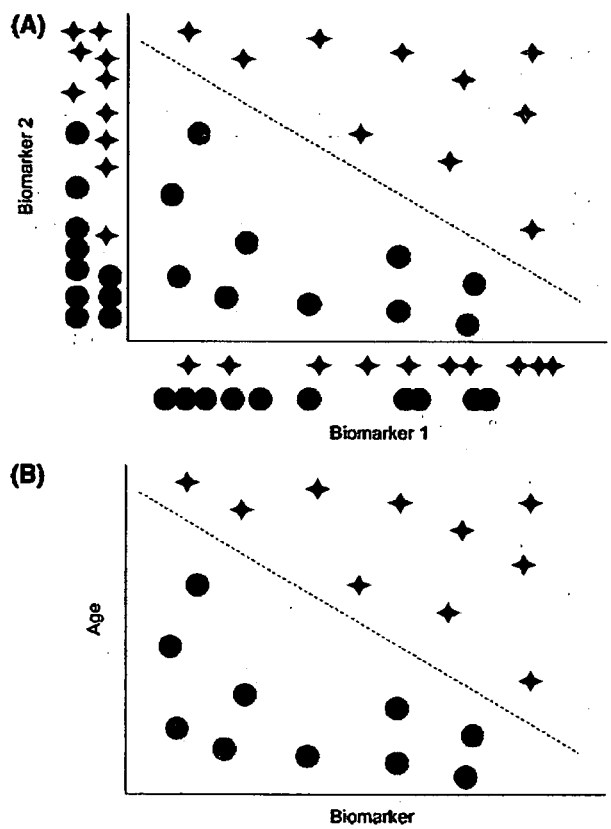


Figure 7. (A) Hypothetical example of the combined disease-linkage effect of two protein biomarkers. (Stars signify affected case individuals, circles non-affected control individuals). (B) Hypothetical example of the combined disease-linkage effect of a biomarker and a clinical variable. (Stars signify affected case individuals, circles non-affected control individuals).

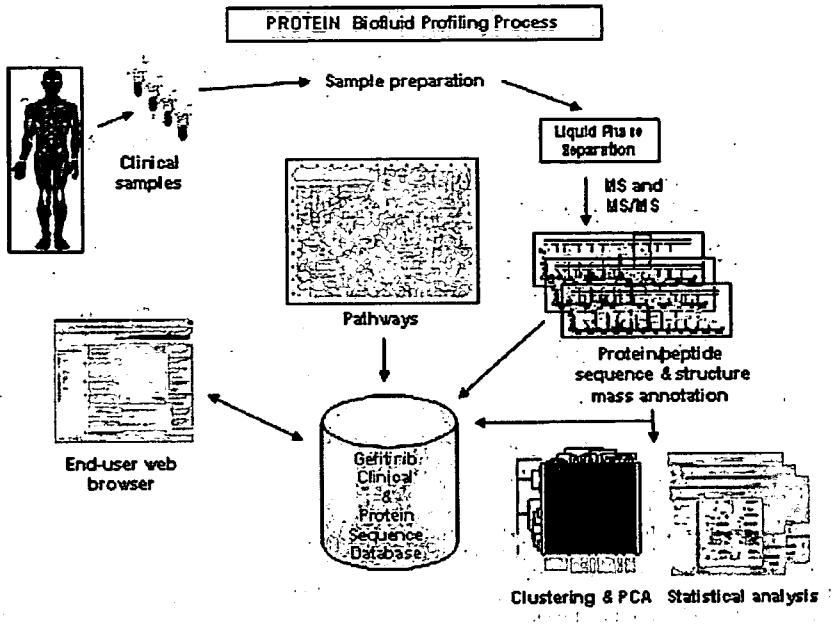


Figure 8. Illustration of the bioinformatics and data processing structure within which MSBA (Mass Spectrometric Biomarker Assays) data are captured, modified and analyzed.

development discusses many of the issues around validating predictive biomarkers.[82]

Finally, it is preferable to be able to assign a biological rationale to the biomarkers. Confidence in the reliability of a biomarker is greatly enhanced if we can correctly understand how it relates to the mechanism and progression of the disease of interest. Figure 8 illustrates a bioinformatics and data processing structure that we have developed to allow us to both conduct interactive exploratory and statistical analyses, and also investigate the disease and pathway linkage of discovered biomarker proteins through direct access to reference databases.

Future Perspectives

Within this paper we have discussed many of the issues that need to considered in developing a personalized medicine approach. A key starting point is that rigorous steps are taken to ensure accurate diagnosis and the careful gathering of both clinical and proteomic data to facilitate the search for peptide patterns.

There are many challenges in performing protein analysis in blood, but mass spectrometry equipment and methods can now be used to generate peptide data with high sensitivity, high scanning speed, and improved quantification. Data handling and processing techniques for steps such as peak alignment and the subsequent methodologies for statistical modeling and analysis are now far enough developed to generate high quality data and robustly analyze these data with confidence.

We have provided details of the MSBA method that can be used to easily translate protein intensities into a practical multiplex assay which can be exploited in the clinic without the need to develop anti-bodies for ELISA. We have also described how an expanded statistical analysis can be used to allow for the individual variance of protein expression to enable us to focus on the proteomic patterns that are actually related to ILD. Finally, we have emphasized the importance of validating the predictive power of a biomarker tool in a way that reflects the real-life setting of intended clinical use.

Hopefully, this combination of developments over a range of different areas brings us one step closer to a practical personalized medicine.

IRESSA is a trademark of the AstraZeneca group of companies.

Acknowledgments. We thank all involved in the Iressa study which provided the inspiration for this overview of personalized medicine approaches, including: the external Epidemiology Advisory Board (Kenneth J. Rothman, Jonathan M. Samet, Toshiro Takezaki, Kotaro Ozasa, Masahiko Ando) for their advice and scientific review of study design, conduct, and analysis; Professor Nestor Müller for his expert input into radiological aspects of ILD diagnosis; all Case Review Board members individually (M. Suga, T. Johkoh, M. Takahashi, Y. Ohno, S. Nagai, Y. Taguchi, Y. Inoue, T. Yana, M. Kusumoto, H. Arakawa, A. Yoshimura, M. Nishio, Y. Ohe, K. Yoshimura, H. Takahashi, Y. Sugiyama, M. Ebina) for their valuable work in blindly reviewing ILD diagnoses, as well as pre-study CT scans for pre-existing comorbidities, the Japan Thoracic Radiology Group, Shiga, Japan for their support of CRB work; and all Hospitals, Clinical Investigators; study monitors, nurses, data managers, other support staff, and the participating patients for providing and collecting the data in the study.

References

(1) Thatcher, N.; Chang, A.; Parikh, P.; Pereira, J. R.; Ciuleanu, T.; von. Pawel, J.; et al. Gefitinib plus best supportive care in previously treated patients with refractory advanced non-small-cell lung cancer: results from a randomised, placebo-controlled, multicentre study (Iressa Survival Evaluation in Lung Cancer). *Lancet* 2005, *366*, 1527–1537.

(2) Hirsch, F. R.; Varella-Garcia, M.; McCoy, J.; West, H.; Xavier, A. C.; Gumerlock, P.; et al. Increased epidermal growth factor receptor gene copy number detected by fluorescence in situ hybridization associates with increased sensitivity to gefitinib in patients with bronchioloalveolar carcinoma subtypes: a Southwest Oncology Group study. *J. Clin. Oncol.* 2005, *23*, 6838–6845.

(3) Cappuzzo, F.; Varella-Garcia, M.; Shigematsu, H.; Domenichini, I.; Bartolini, S.; Ceresoli, G. L.; et al. Increased HER2 gene copy number is associated with response to gefitinib therapy in epidermal growth factor receptor-positive non-small-cell lung cancer patients. *J. Clin. Oncol.* 2005, *23*, 5007−5018.
(4) Araki, J.; Okamoto, I.; Suto, R.; Ichikawa, Y.; Sasaki, J. Efficacy of the tyrosine kinase inhibitor gefitinib in a patient with metastatic small cell lung cancer. *Lung Cancer* 2005, *48*, 141−144.
(5) Kim, K. S.; Jeong, J. Y.; Kim, Y. C.; Na, K. J.; Kim, Y. H.; Ahn, S. J.; et al. Predictors of the response to gefitinib in refractory non-small cell lung cancer. *Clin. Cancer Res.* 2005, *11*, 2244−2251.
(6) Lynch, T. J.; Bell, D. W.; Sordella, R.; Gurubhagavatula, S.; Okimoto, R. A.; Brannigan, B. W.; et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* 2004, *350*, 2129−2139.
(7) Paez, J. G.; Jänne, P. A.; Lee, J. C.; Tracy, S.; Greulich, H.; Gabriel, S.; et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* 2004, *304*, 1497−1500.
(8) Shigematsu, H.; Lin, L.; Takahashi, T.; Nomura, M.; Suzuki, M.; Wistuba II; et al. Clinical and biological features associated with epidermal growth factor receptor gene mutations in lung cancers. *J. Natl. Cancer Inst.* 2005, *97*, 339−346.
(9) American Thoracic Society: American Thoracic Society/European Respiratory Society International Multidisciplinary Consensus Classification of the Idiopathic Interstitial Pneumonias. This joint statement of the American Thoracic Society (ATS), and the European Respiratory Society (ERS) was adopted by the ATS Board of Directors, June 2001 and by The ERS Executive Committee, June 2001. *Am. J. Respir. Crit. Care Med.* 2002, *165*, 277−304.
(10) Raghu, G.; Nyberg, F.; Morgan, G. The epidemiology of interstitial lung disease and its association with lung cancer. *Br. J. Cancer* 2004, *91* (Suppl. 2), S3−S10.
(11) Asada, K.; Mukai, J.; Ougushi, F. Characteristics and management of lung cancer in patients with idiopathic pneumonia. *Jap. J. Thor. Dis.* 1992, *51*, 214−219.
(12) Hubbard, R.; Venn, A.; Lewis, S.; Britton, J. Lung cancer and cryptogenic fibrosing alveolitis. A population-based cohort study. *Am. J. Respir. Crit. Care Med.* 2000, *161*, 5−8.
(13) Matsushita, H.; Tanaka, S.; Saiki, Y.; Hara, M.; Nakata, K.; Tanimura, S.; et al. Lung cancer associated with usual interstitial pneumonia. *Pathol. Int.* 1995, *45*, 925−932.
(14) Ogura, T.; Kondo, A.; Sato, A.; Ando, M.; Tamura, M. Incidence and clinical features of lung cancer in patients with idiopathic interstitial pneumonia. *Nihon Kyobu Shikkan Gakkai Zasshi* 1997, *35*, 294−299.
(15) Takeuchi, E.; Yamaguchi, T.; Mori, M.; Tanaka, S.; Nakagawa, M.; Yokota, S.; et al. Characteristics and management of patients with lung cancer and idiopathic interstitial pneumonia. *Nihon Kyobu Shikkan Gakkai Zasshi* 1996, *34*, 653−658.
(16) Turner-Warwick, M.; Lebowitz, M.; Burrows, B.; Johnson, A. Cryptogenic fibrosing alveolitis and lung cancer. *Thorax* 1980, *35*, 496−499.
(17) Baumgartner, K. B.; Samet, J. M.; Stidley, C. A.; Colby, T. V.; Waldron, J. A. Cigarette smoking: a risk factor for idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* 1997, *155*, 242−248.
(18) Britton, J.; Hubbard, R. Recent advances in the aetiology of cryptogenic fibrosing alveolitis. *Histopathology* 2000, *37*, 387−392.
(19) Iwai, K.; Mori, T.; Yamada, N.; Yamaguchi, M.; Hosoda, Y. Idiopathic pulmonary fibrosis. Epidemiologic approaches to occupational exposure. *Am. J. Respir. Crit. Care Med.* 1994, *150*, 670−675.
(20) Nagai, S.; Hoshino, Y.; Hayashi, M.; Ito, I. Smoking-related interstitial lung diseases. *Curr. Opin. Pulm. Med.* 2000, *6*, 415−419.
(21) Lilly. Gemcitabine prescribing information. http://pi.lilly.com/gemzar.pdf, 2003.
(22) Kunitoh, H.; Watanabe, K.; Onoshi, T.; Furuse, K.; Niitani, H.; Taguchi, T. Phase II trial of docetaxel in previously untreated advanced non-small-cell lung cancer: a Japanese cooperative study. *J. Clin. Oncol.* 1996, *14*, 1649−1655.
(23) Merad, M.; Le Cesne, A.; Baldeyrou, P.; Mesurolle, B.; Le Chevalier, T. Docetaxel and interstitial pulmonary injury. *Ann. Oncol.* 1997, *8*, 191−194.
(24) Wang, G.-S.; Yan, K.-Y.; Perng, R.-P. Life-threatening hypersensitivity pneumonitis induced by docetaxel (taxotere). *Br. J. Cancer* 2001, *85*, 1247−1250.
(25) Erasmus, J. J.; McAdams, H. P.; Rossi, S. E. Drug-induced lung injury. *Semin. Roentgenol.* 2002, *37*, 72−81.
(26) Aviram, G.; Yu, E.; Tai, P.; Lefcoe, M. S. Computed tomography to assess pulmonary injury associated with concurrent chemoradiotherapy for inoperable non-small cell lung cancer. *Can. Assoc. Radiol. J.* 2001, *52*, 385−391.
(27) Yoshida, S. The results of gefitinib prospective investigation. *Med. Drug J.* 2005, *41*, 772−789.
(28) Mueller, N. L.; White, D. A.; Jiang, H.; Gemma, A. Diagnosis and management of drug-associated interstitial lung disease. *Br. J. Cancer* 2004, *91*, S24−S30.
(29) Marko-Varga, G.; Fehniger, T. E. Proteomics and disease—the challenges for technology and discovery. *J. Proteome Res.* 2004, *3*, 167−178.
(30) Marko-Varga, G.; Lindberg, H.; Lofdahl, C. G.; Jonsson, P. H. L.; Dahlback, M.; Lindquist, E.; et al. Discovery of biomarker candidates within disease by protein profiling: principles and concepts. *J. Proteome Res.* 2005, *4*, 1200−1212.
(31) Omenn, G. S. The Human Proteome Organization Plasma Proteome Project pilot phase: reference specimens, technology platform comparisons, and standardized data submissions and analyses. *Proteomics* 2004, *4*, 1235−1240.
(32) Omenn, G. S. Advancement of biomarker discovery and validation through the HUPO plasma proteome project. *Dis. Markers* 2004, *20*, 131−134.
(33) Orchard, S.; Hermjakob, H.; Binz, P. A.; Hoogland, C.; Taylor, C. F.; Zhu, W.; et al. Further steps towards data standardisation: the Proteomic Standards Initiative HUPO 3(rd) annual congress, Beijing 25−27(th) October, 2004. *Proteomics* 2005, *5*, 337−339.
(34) Anderson, N. G.; Matheson, A.; Anderson, N. L. Back to the future: the human protein index (HPI) and the agenda for post-proteomic biology. *Proteomics* 2001, *1*, 3−12.
(35) Anderson, N. L.; Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* 2002, *1*, 845−867.
(36) Jacobs, J. M.; Adkins, J. N.; Qian, W. J.; Liu, T.; Shen, Y.; Camp, D. G.; et al. Utilizing human blood plasma for proteomic biomarker discovery. *J. Proteome Res.* 2005, *4*, 1073−1085.
(37) Anderson, N. G.; Anderson, L. The Human Protein Index. *Clin. Chem.* 1982, *28*, 739−748.
(38) Haab, B. B.; Geierstanger, B. H.; Michailidis, G.; Vitzthum, F.; Forrester, S.; Okon, R.; et al. Immunoassay and antibody microarray analysis of the HUPO Plasma Proteome Project reference specimens: systematic variation between sample types and calibration of mass spectrometry data. *Proteomics* 2005, *5*, 3278−3291.
(39) Martens, L.; Hermjakob, H.; Jones, P.; Adamski, M.; Taylor, C.; States, D.; et al. PRIDE: the proteomics identifications database. *Proteomics* 2005, *5*, 3537−3545.
(40) Omenn, G. S.; States, D. J.; Adamski, M.; Blackwell, T. W.; Menon, R.; Hermjakob, H.; et al. Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* 2005, *5*, 3226−3245.
(41) Patterson, S. D. Data analysis—the Achilles heel of proteomics. *Nat. Biotechnol.* 2003, *21*, 221−222.
(42) Rahbar, A. M.; Fenselau, C. Integration of Jacobson's pellicle method into proteomic strategies for plasma membrane proteins. *J. Proteome Res.* 2004, *3*, 1267−1277.
(43) Ho, Y.; Gruhler, A., Heilbut, A.; Bader, G. D.; Moore, L.; Adams, S. L.; et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 2002, *415*, 180−183.
(44) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* 2003, *422*, 198−207.
(45) Anderson, N. L.; Polanski, M.; Pieper, R.; Gatlin, T.; Tirumalai, R. S.; Conrads, T. P.; et al. The human plasma proteome: a nonredundant list developed by combination of four separate sources. *Mol. Cell. Proteomics* 2004, *3*, 311−326.
(46) Olsen, J. V.; Mann, M. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc. Natl. Acad. Sci. U.S.A.* 2004, *101*, 13417−13422.
(47) Sadygov, R. G.; Liu, H.; Yates, J. R. Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. *Anal. Chem.* 2004, *76*, 1664−1671.
(48) Fujii, K.; Nakano, T.; Kanazawa, M.; Akimoto, S.; Hirano, T.; Kato, H.; et al. Clinical-scale high-throughput human plasma proteome analysis: lung adenocarcinoma. *Proteomics* 2005, *5*, 1150−1159.

(49) Campbell, J. M.; Collings, B. A.; Douglas, D. J. A new linear ion trap time-of-flight system with tandem mass spectrometry capabilities. *Rapid Commun. Mass Spectrom.* 1998, *12*, 1463–1474.
(50) Cha, B. C.; Blades, M.; Douglas, D. J. An interface with a linear quadropole ion guide for an electrospray-ion trap mass spectrometer system. *Anal. Chem.* 2000, *72*, 5647–5654.
(51) Hager, J. W. Product ion spectral simplification using time-delayed fragment ion capture with tandem linear ion traps. *Rapid Commun. Mass Spectrom.* 2003, *17*, 1389–1398.
(52) Syka, J. E.; Marto, J. A.; Bai, D. L.; Horning, S.; Senko, M. W.; Schwartz, J. C.; et al. Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications. *J. Proteome Res.* 2004, *3*, 621–626.
(53) Shen, Y.; Zhao, R.; Belov, M. E.; Conrads, T. P.; Anderson, G. A.; Tang, K.; et al. Packed capillary reversed-phase liquid chromatography with high-performance electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry for proteomics. *Anal. Chem.* 2001, *73*, 1766–1775.
(54) Wu, S. L.; Kim, J.; Hancock, W. S.; Karger, B. Extended Range Proteomic Analysis (ERPA): a new and sensitive LC-MS platform for high sequence coverage of complex proteins with extensive post-translational modifications-comprehensive analysis of beta-casein and epidermal growth factor receptor (EGFR). *J. Proteome Res.* 2005, *4*, 1155–1170.
(55) Olsen, J. V.; de Godoy, L. M.; Li, G.; Macek, B.; Mortensen, P.; Pesch, R.; et al. Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* 2005, *4*, 2010–2021.
(56) Yates, J. R.; Cociorva, D.; Liao, L.; Zabrouskov, V. Performance of a linear ion trap-Orbitrap hybrid for peptide analysis. *Anal. Chem.* 2006, *78*, 493–500.
(57) Anderson, D. C.; Li, W.; Payan, D. G.; Noble, W. S. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J. Proteome Res.* 2003, *2*, 137–146.
(58) Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A.; Clauser, K.; Nesvizhskil, A. The need for guidelines in publication of peptide and protein identification data: working group on publication guidelines for peptide and protein identification data. *Mol. Cell. Proteomics* 2004, *3*, 531–533.
(59) Fenyo, D.; Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* 2003, *75*, 768–774.
(60) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 2003, *75*, 4646–4658.
(61) Peri, S.; Navarro, J. D.; Kristiansen, T. Z.; Amanchy, R.; Surendranath, V.; Muthusamy, B.; et al. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.* 2004, *32*, D497–D501.
(62) Kratchmarova, I.; Blagoev, B.; Haack-Sorensen, M.; Kassem, M.; Mann, M. Mechanism of divergent growth factor effects in mesenchymal stem cell differentiation. *Science* 2005, *308*, 1472–1477.
(63) Dreger, M.; Bengtsson, L.; Schöneberg, T.; Otto, H.; Hucho, F. Nuclear envelope proteomics: novel integral membrane proteins of the inner nuclear membrane. *Proc. Natl. Acad. Sci. U.S.A.* 2001, *98*, 11943–11948.
(64) Giot, L.; Bader, J. S.; Brouwer, C.; Chaudhuri, A.; Kuang, B.; Li, Y.; et al. A protein interaction map of Drosophilia melanogaster. *Science* 2003, *302*, 1727–1736.
(65) Johnson, J. R.; Florens, L.; Carucci, D. J.; Yates, J. R., III. Proteomics in malaria. *J. Proteome Res.* 2004, *3*, 296–306.
(66) Hirsch, J.; Hansen, K. C.; Burlingame, A. L.; Matthay, M. A. Proteomics: current techniques and potential applications to lung disease. *Am. J. Physiol. Lung Cell Mol. Physiol.* 2004, *287*, L1–L23.
(67) Malmström, J.; Larsen, K.; Hansson, L.; Löfdahl, C.-G.; Norregard-Jensen, O.; Marko-Varga, G.; et al. Proteoglycan and proteome profiling of central human pulmonary fibrotic tissue utilizing minaturized sample preparation: A feasibility study. *Proteomics* 2002, *2*, 394–404.
(68) Malmström, J.; Larsen, K.; Malmström, L.; Tufvesson, E.; Parker, K.; Marchese, J.; et al. Proteome annotations and identifications of the human pulmonary fibroblast. *J. Proteome Res.* 2004, *3*, 525–537.
(69) Oh, P.; Li, Y.; Yu, J.; Durr, E.; Krasinska, K. M.; Carver, L. A.; et al. Subtractive proteomic mapping of the endothelial surface in lung and solid tumours for tissue-specific therapy. *Nature* 2004, *429*, 629–635.
(70) Fujii, K.; Nakano, T.; Kawamura, T.; Usui, F.; Bando, Y.; Wang, R.; et al. Multidimensional protein profiling technology and its application to human plasma proteome. *J. Proteome Res.* 2004, *3*, 712–718.
(71) Schwartz, J. C.; Senko, M. W.; Syka, J. E. A two-dimensional quadrupole ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.* 2002, *13*, 659–669.
(72) Sneath, P. H. A.; Sokai, R. R. *Numerical Taxonomy, The principles and practice of numerical classification*; W. H. Freeman and Co.: San Francisco, 1973.
(73) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* 2006, *78*, 779–787.
(74) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence-databases using mass spectrometry data. *Electrophoresis* 1999, *20*, 3551–3567.
(75) Storey, J. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B* 2002, *64*, 479.
(76) Vapnik, V. *Statistical Learning Theory*; Wiley: Chichester, UK, 1998.
(77) Breiman, L. Random forests. *Mach. Learn.* 2001, *45*, 5–32.
(78) Tibshirani, R.; Hastie, T.; Narasimhan, B.; Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 2002, *99*, 6567–6572.
(79) Lee, J. W.; Lee, J. B.; Park, M.; Song, S. H. An extensive comparison of recent classification tools applied to microarray data. *Comp. Stat. Data Anal.* 2005, *48*, 869–885.
(80) Steyerberg, E. W.; Harrell, F. E., Jr.; Borsboom, G. J.; Eijkemans, M. J.; Vergouwe, Y.; Habbema, J. D. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J. Clin. Epidemiol.* 2001, *54*, 774–781.
(81) Bleeker, S. E.; Moll, H. A.; Steyerberg, E. W.; Donders, A. R.; Derksen-Lubsen, G.; Grobbee, D. E.; et al. External validation is necessary in prediction research: a clinical example. *J. Clin. Epidemiol.* 2003, *56*, 826–832.
(82) Food and Drug Administration (FDA): Drug-diagnostic co-development concept paper. Draft—not for implementation. http://www.fda.gov/cder/genomics/pharmacoconceptfn.pdf, 2005.

PR070046S

Protein expression associated with early intrahepatic recurrence of hepatocellular carcinoma after curative surgery

Hideki Yokoo,[1,2] Tadashi Kondo,[1,6] Tetsuya Okano,[1] Kazuaki Nakanishi,[2,3] Michiie Sakamoto,[3,4] Tomoo Kosuge,[5] Satoru Todo[2] and Setsuo Hirohashi[1]

[1]Proteome Bioinformatics Project, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku Tokyo 104-0045; [2]Department of General Surgery, Hokkaido University Graduate School of Medicine, Sapporo Kita-ku, Sapporo 060-0815; [3]Pathology Division, National Cancer Center Research Institute, Tokyo; [4]Department of Pathology, Keio University School of Medicine, 35 Shinanomachi, Shinjuku-ku Tokyo 160-8582; [5]Hepatobiliary and Pancreatic Surgery Division, National Cancer Center Hospital, 5-1-1 Tsukiji, Chuo-ku Tokyo, Japan 104-0045

(Received November 20, 2006/Revised January 9, 2007/Accepted January 11, 2007/Online publication March 25, 2007)

The poor prognosis of patients with hepatocellular carcinoma (HCC) is attributed to intrahepatic recurrence. To understand the molecular background of early intrahepatic recurrence, we conducted a global protein expression study. We compared the protein expression profiles of the primary HCC tissues of 12 patients who showed intrahepatic recurrence within 6 months post surgery with those of 15 patients who had no recurrence 2 years post surgery. Two-dimensional difference gel electrophoresis identified 23 protein spots, the intensity of which was highly associated with early intrahepatic recurrence. To validate the prediction performance of the identified proteins, we examined additional HCC tissues from 13 HCC patients; six with early intrahepatic recurrence and seven without recurrence. We found that all but one of the 13 patients were grouped according to their recurrence status based on the intensity of the 23 protein spots. Mass spectrometry identified 23 proteins corresponding to the spots. Although 13 of 23 have been previously reported to be correlated with HCC, their association with early intrahepatic recurrence had not been established. The identified proteins are involved in signal transduction pathways, glucose metabolism, cytoskeletal structure, cell adhesion, or function as antioxidants and chaperones. The identified proteins may be candidates for prognostic markers and contribute to the improvement of existing therapeutic strategies. (*Cancer Sci* 2007; 98: 665–673)

Despite recent progress in early diagnostic modalities and therapeutic management, the prognosis of hepatocellular carcinoma (HCC) patients remains poor, mainly due to intrahepatic recurrence.[1] The incidence of intrahepatic recurrence in the liver remnant ranges from 50% to 100% of HCC patients who undergo curative resection as a result of either intrahepatic metastasis from the primary tumor or multicentric recurrence[2–6] and the median survival after recurrence is only 11 months.[7] Although various molecular alterations have been correlated with early recurrence of HCC[8–13] and studies on such events furthered our understanding of HCC biology, the underlying mechanisms remain obscure. Studies focusing on individual candidate genes may be insufficient for precisely understanding the background of the malignant behavior of tumor cells, because the features of cancer cells are defined by multiple genetic alterations in a functionally coordinated manner. With this notion in mind, global mRNA expression analyses have been conducted to identify the gene network associated with early intrahepatic recurrence.[14–17] Such molecular pathogenesis studies may lead to the development of gene-based biomarkers and novel therapeutic strategies. However, mRNA expression reflects the protein expression of only a small proportion of genes, probably because of post-translational control of protein quantity.[18–21] To complement gene expression studies, global protein expression profiling, namely a proteomic approach, has been carried out for the study of HCC.[22–28] Conceptually, proteomics reflects the molecular background of cancer phenotypes more directly, as the final product of genetic and epigenetic events is the proteins contained in the cells.

In this study we investigated the protein expression of HCC tissues from patients with early intrahepatic recurrence and from patients without recurrence. We used two-dimensional fluorescence difference gel electrophoresis (2D-DIGE) to generate quantitative protein expression profiles, and data-mining methods and mass spectrometry to determine the proteins associated with early intrahepatic recurrence.

Materials and Methods

Patients. HCC patients who had undergone surgery at the National Cancer Center Hospital from January 1998 to April 2000 were evaluated for inclusion in this study. We assessed 40 patients who showed intrahepatic recurrence within 6 months or no recurrence within 2 years post surgery. The absence of intrahepatic metastasis in the residual liver was monitored by ultrasonography (US), computer tomography (CT) scans and angiography preoperatively, and was further confirmed by intraoperative US. CT scans were also carried out to rule out distant metastasis. Surgical margins were assessed post surgery by experienced pathologists. The patients were monitored for HCC recurrence at least once every 4 months after surgery using US and CT scans. To identify the protein spots associated with early intrahepatic recurrence, we validated the prediction performance of the selected spots using a newly enrolled sample set. The clinico-pathological characteristics of the patient groups are summarized in Table 1. There was no significant difference of virus type between the early recurrence group and non-recurrence group in both the training and test set. The number of primary tumor, venous invasion, and tumor-node-metastasis (TNM) staging were significantly different between the early recurrence and non-recurrence groups in the training set (Table 1). The majority of cases were infected with hepatitis C virus (Table 1). This project was approved by the institutional review board for the use of human subjects of the National Cancer Center.

Protein extraction, fluorescence labeling and separation by two-dimensional polyacrylamide gel electrophoresis. The tissue specimens obtained from the surgically resected tumors were stored in vapor nitrogen until use. The frozen tissues were homogenized in urea lysis buffer (7 M urea, 2 M thiourea, 3% CHAPS, 1% Triton X-100) and incubated on ice for 30 min. After centrifugation at 15 000 *g* for 30 min, the supernatant was

[6]To whom correspondence should be addressed. E-mail: takondo@gan2.res.ncc.go.jp

doi: 10.1111/j.1349-7006.2007.00441.x

Table 1. Clinical information of the patients

	Training set			Test set		
	Early recurrence group (n = 12)	Non-recurrence group (n = 15)	P-value	Early recurrence group (n = 6)	Non-recurrence group (n = 7)	P-value
Age (mean ± SD)	63.2 ± 5.5	67.3 ± 3.6	0.06	60.5 ± 7.2	61 ± 9.1	0.91
Gender (%)						
Male	7 (58.3)	12 (80)	0.06	5 (83.3)	7 (100)	0.46
Female	5 (41.7)	3 (20)		1 (16.7)	0 (0)	
Viral infection (%)						
HBV(+)	1 (8.3)	3 (20)	0.43	3 (50)	2 (28.6)	0.1
HCV(+)	7 (58.4)	10 (66.7)		1 (16.7)	4 (57.1)	
HBV(+)HCV(+)	1 (8.3)	0 (0)		0 (0)	1 (14.3)	
HBV(–)HCV(–)	3 (25)	2 (13.3)		2 (33.3)	0 (0)	
Child-Pugh classification (%)						
A	12 (100)	15 (100)		6 (100)	7 (100)	
B	0 (0)	0 (0)	–	0 (0)	0 (0)	–
C	0 (0)	0 (0)		0 (0)	0 (0)	
ICG-15 (mean ± SD)	19.7 ± 10.5	15.4 ± 9	0.26	22.1 ± 15.3	18.5 ± 9.8	0.63
AFP (ng/mL)						
> 400	4 (33.3)	1 (6.7)	0.13	0 (0)	2 (28.6)	0.46
≤ 400	8 (66.7)	14 (93.3)		6 (100)	5 (71.4)	
Tumor size (mean ± SD)	6.6 ± 5.2	4.1 ± 1.5	0.14	9.3 ± 7.3	4.5 ± 1.3	0.18
Primary lesion (%)						
Single	2 (16.7)	14 (93.3)	< 0.001	1 (16.7)	7 (100)	0.004
Multiple	10 (63.3)	1 (6.7)		5 (83.3)	0 (0)	
Gross type (%)						
Type 1	3 (25)	7 (46.7)		0 (0)	4 (57.1)	
Type 2	4 (33.3)	6 (40)	0.2	2 (33.3)	3 (42.9)	0.06
Type 3	5 (41.7)	2 (13.3)		4 (66.7)	0 (0)	
Tumor differentiation (%)						
Well	0 (0)	3 (20)		0 (0)	2 (28.6)	
Moderately	7 (58.3)	10 (66.7)	0.18	4 (66.7)	4 (57.1)	0.62
Poorly	5 (41.7)	2 (13.3)		2 (33.3)	1 (14.3)	
Venous invasion (%)						
Presence	12 (100)	6 (40)	0.001	5 (83.3)	4 (57.1)	0.31
Absence	0 (0)	9 (60)		1 (16.7)	3 (42.9)	
UICC pTNM stage (%)						
Stage I	0 (0)	0 (0)	< 0.001	0 (0)	0 (0)	0.03
Stage II	0 (0)	12 (80)		0 (0)	4 (57.1)	
Stage IIIA	10 (83.3)	3 (20)		3 (50)	3 (42.9)	
Stage IVA	2 (16.7)	0 (0)		3 (50)	0 (0)	
Background liver (%)						
Normal	0 (0)	1 (6.7)		1 (16.7)	0 (0)	
Chronic hepatitis	9 (75)	12 (80)	0.49	4 (66.6)	6 (85.7)	0.79
Cirrhosis	3 (25)	2 (13.3)		1 (16.7)	1 (14.3)	

AFP, alpha-fetoprotein; HBV, hepatitis B virus; HCV, hepatitis C virus; ICG, indocyanine green; pTNM, pathological TNM; UICC, Union Internationale Centre le Cancer

recovered. Fluorescence labeling of proteins was carried out as described previously (Fig. 1A).[22] In brief, a portion of all samples was mixed together to make an internal control sample, aliquoted and stored at −80°C until use. Fifty micrograms of both the internal control sample and each experimental sample were labeled with 200 nmol of 1-(5-carboxypentyl)-1′-propylindocarbocyamine halide (Cy3) and 1-(5-carboxypentyl)-1′-methylindocarbocyamine halide (Cy5) (GE Healthcare Bio-sciences, Little Chalfont; Buckinghamshire, UK), respectively. After the labeling reaction was terminated with 0.2 mM lysine, equal amounts of the Cy3- or Cy5-labeled samples were mixed and the sample volume was made up to 420 μL with urea lysis buffer containing 65 mM dithiothreitol (DTT) and 1% ampholine (Amersham Biosciences, Upsala, Sweden).

Two-dimensional gel electrophoresis was carried out as in our previous report (Fig. 1A).[22] In brief, the first separation was achieved using IPG DryStrip gels (24 cm long with a pI range between 3.0 and 10.0, Amersham Biosciences) and IPGphor (GE Healthcare Bio-sciences), and the second-dimension separation was achieved using Ettan Dalt II and 9–15% polyacrylamide gradient gels. Identical samples were run in triplicate gels. After electrophoresis, gels were scanned with 2920 2D-Master Imager (GE Healthcare Bio-sciences). The Cy5-image of each sample was normalized by the Cy3-image of the internal standard sample using the DeCyder software (GE Healthcare Bio-sciences).

We analyzed the acquired proteome data, which consisted of the intensity signal ratios of approximately 1400 protein spots from 40 HCC samples, using the Impressionist software (Gene Data, Basel, Switzerland). We assessed the reproducibility of the acquired protein profiles by running an identical sample three times and examining the similarities between the protein expression profiles with scatter plot analysis. The intensity values obtained for each of the protein spots ranged within a twofold difference between the experiments for more than 90% of the

doi: 10.1111/j.1349-7006.2007.00441.x

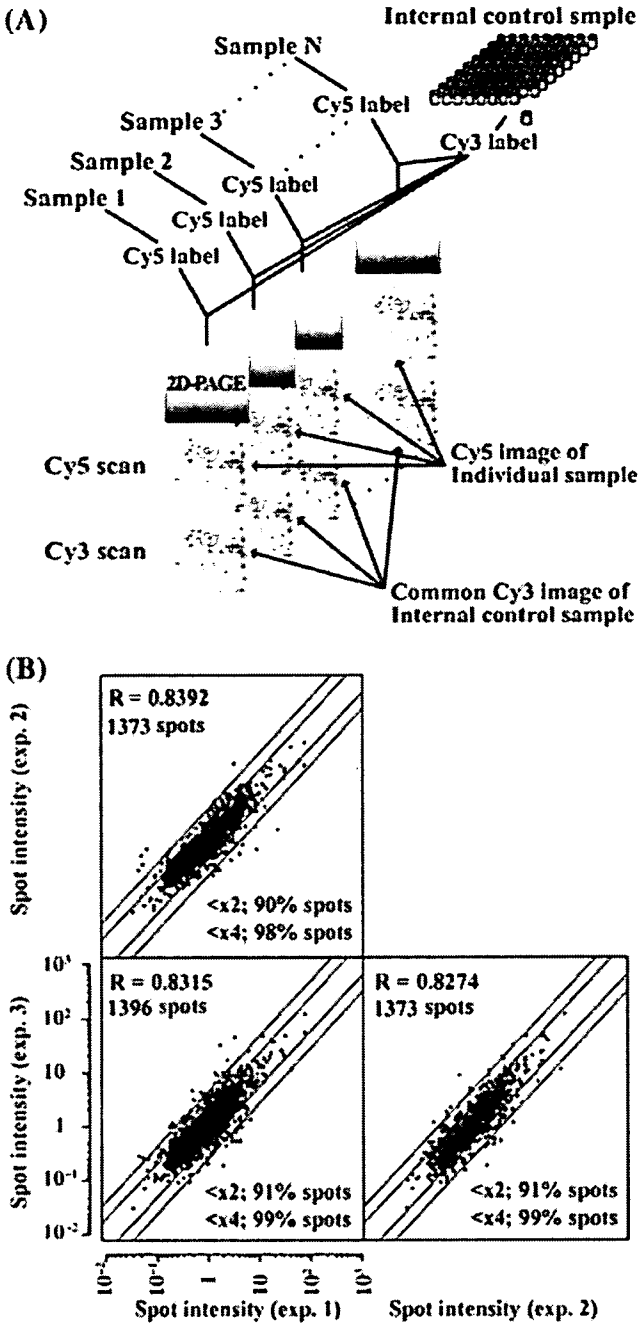


Fig. 1. (A) Fluorescence labeling and separation by two-dimensional polyacrylamide gel electrophoresis (2D-PAGE). A protein sample from a cell line and the internal control sample were labeled with Cy5 and Cy3, respectively. The Cy5-labeled sample was mixed with the Cy3-labeled internal control sample and coseparated in the gel. The Cy5 images such as the one shown here represent the protein expression profile of individual samples while the Cy3 image represents that of the common internal control sample. (B) Scatter plot analysis of three protein expression profiles obtained from the same sample. The intensity values obtained for each of the protein spots ranged within a twofold difference between the experiments for more than 90% of the spots.

spots, while the correlation coefficient for the overall protein expression was at least 0.8274 (Fig. 1B). These results indicated that the protein expression profiles obtained by 2D-DIGE were highly reproducible both in a quantitative and in a qualitative way.

Mass spectrometric identification of proteins. Protein identification was achieved as reported previously.[22] In brief, a 500-μg protein sample was separated by two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) and stained with SYPRO Ruby (Molecular Probes, Eugene, OR, USA). SpotPicker (GE Healthcare Bio-sciences) collected the spots of interest. The tryptic peptides generated by in-gel digestion[22] were subjected to matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) using a Q-Star Pulser-i equipped with an orthogonal injection/MALDI ion source (Applied Biosystems, Framingham, CA, USA). Mass spectra were processed with the Analyst QS program (Applied Biosystems) and Mascot program (Matrix Science, Boston, MA, USA) using the Swiss-Prot database for protein identification.

Results

We created a training set of 27 samples, consisting of 12 samples from patients who had early intrahepatic recurrence within 6 months and 15 samples from patients who had no recurrence within 2 years post surgery. We first examined the similarity between the 2D profiles of the samples in the training set. The correlation efficiency of all paired samples was calculated and the results were summarized in a correlation matrix (Fig. 2A). We found that the protein expression patterns of samples within each group (the non-recurrence and the early recurrence group) were very similar to each other, while they were different to those of the other group, suggesting the presence of a certain proteomic background corresponding to the events leading to the recurrence. This observation allowed us to construct an early recurrence prediction model using spot intensity. We used a support-vector-machine algorithm to build a class prediction model. A support-vector-machine creates the hyper-plane in a multidimensional space to best-distinguish two sample groups.[29] We identified the minimal number of protein spots that can be used to classify the samples with the minimal misclassification error rate. First, we constructed a classifier using a certain number of spots and calculated the misclassification error rate by a leave-one-out cross validation. Then, we ranked the protein spots according to their contribution to the classification model using support vector machine weight. The classification model was then constructed again after the lower-ranked protein spots were deleted, and its misclassification rate was calculated. This process was repeated and the misclassification error rates were plotted as a function of the number of protein spots (Fig. 2B). Although the classifier that included all protein spots had a discriminatory power of 100% accuracy (0% misclassification error rate), a classifier with a smaller spot number (23 spots) also showed 0% misclassification error rate and was considered as the optimal for the classification model to distinguish the samples according to their recurrence status. These 23 protein spots are the minimal protein set to distinguish these two groups based on the support-vector machine algorithm, and the reduction of the number of spots resulted in the increased misclassification error rate.

The classification performance of the 23 selected protein spots was validated using different classification methods. With principal component analysis, which visualizes the correlation of samples in a three-dimensional space created by the intensity of the 23 protein spots, the 27 samples were clearly separated according to their recurrence status (Fig. 2C). A hierarchical clustering study based on similarities in the expression profiles of the 23 protein spots categorized the early intrahepatic recurrence and non-recurrence groups into distinct branches (Fig. 2D). As the separation was ambiguous when all spots were used (data not shown), these 23 protein spots characterized the proteome of the tumors with or without recurrence.

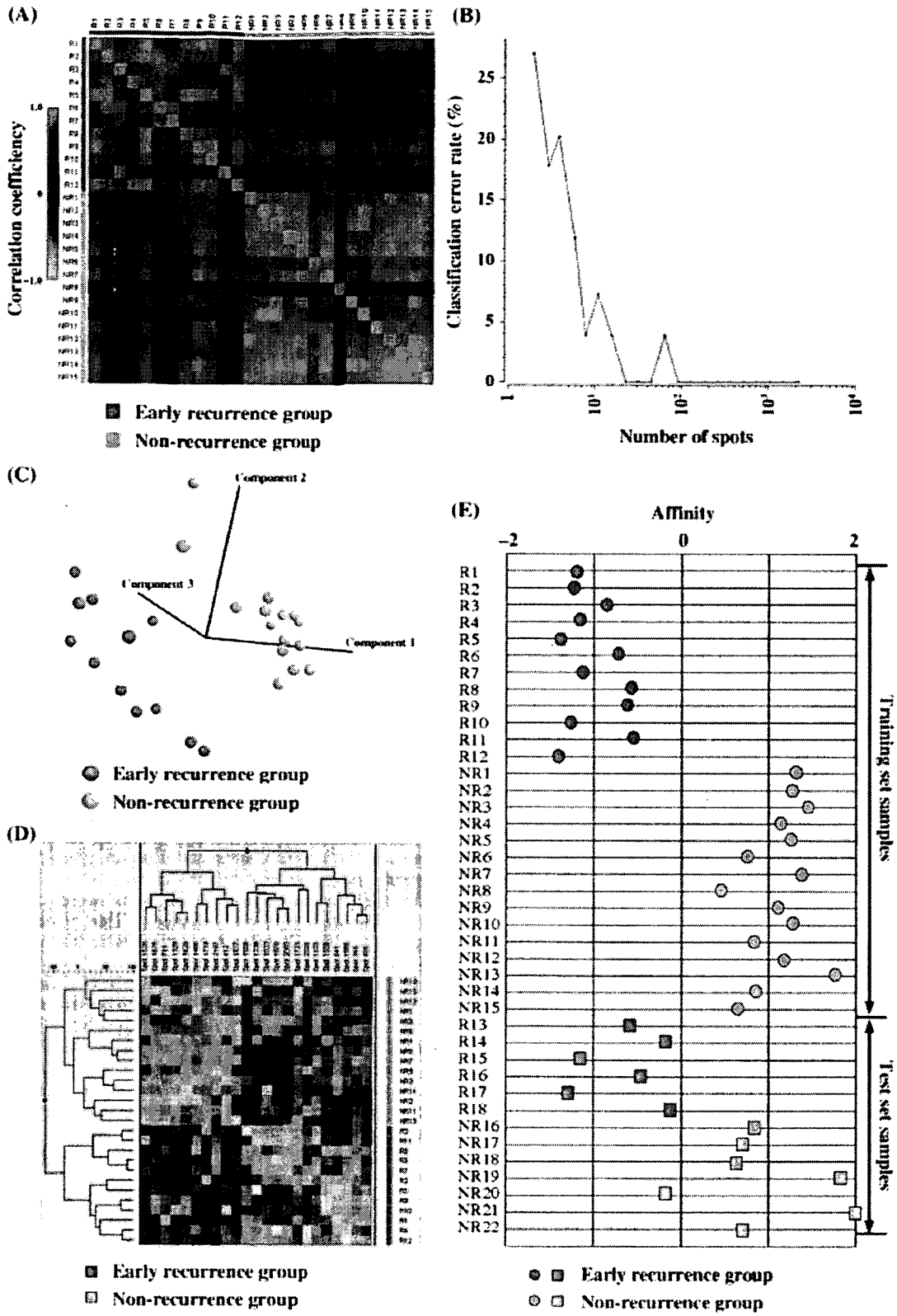


Fig. 2. Protein spots associated with early intrahepatic recurrence. (A) Correlation matrix showing the similarity of the overall feature of proteome. (B) A leave-one-out cross-validation error rate was plotted as a function of spot number, showing that the minimal number of spots to make the error rate 0% was 23. Principal component analysis (C) and hierarchical clustering analysis (D) clearly separated the samples according to their recurrence status using the 23 selected protein spots. (E) All but one of the samples in the training and test set were correctly classified on the basis of the expression pattern of the 23 protein spots. The patients were assigned as those that had early recurrence within 6 months (R1-R18) and those that did not have recurrence for 2 years post surgery (NR1-NR22).

doi: 10.1111/j.1349-7006.2007.00441.x

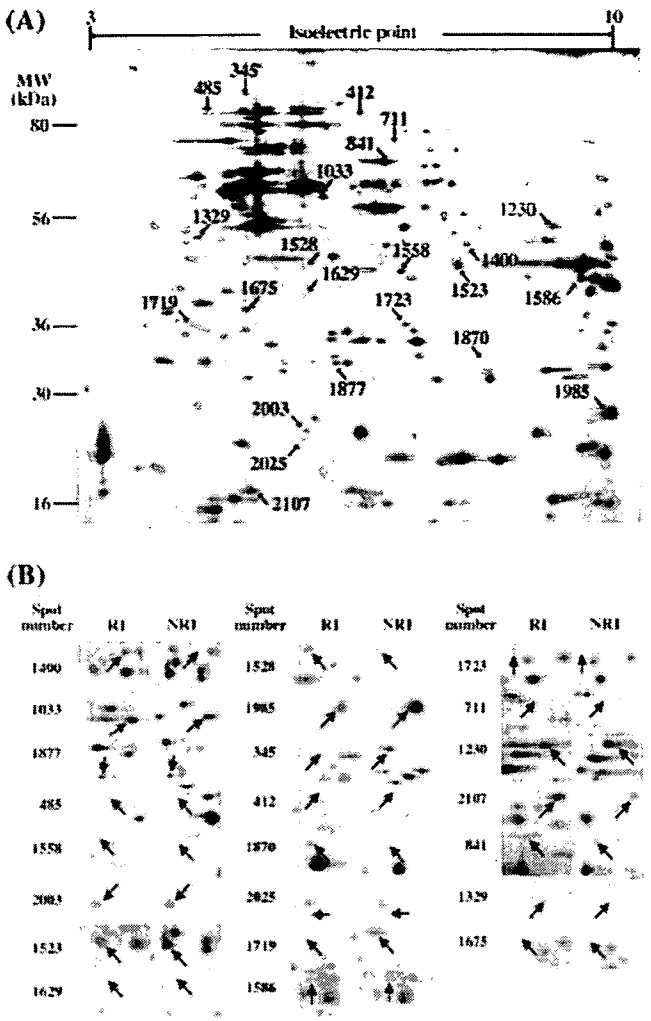


Fig. 3. (A) Localization of the 23 selected spots on the 2D image. (B) The enlarged images are shown to compare the typical image of a sample from the early recurrence cases (R1) and one from the non-recurrence cases (NR1).

The predictive performance of the selected set of spots was validated using a test set, which consisted of early intrahepatic recurrence ($n = 6$) and non-recurrence cases ($n = 7$). We determined the affinity value of each sample by calculating the minimal distance between each sample and the hyper-plane (Fig. 2E). The distance was defined so that the samples classified in the recurrence group had a positive distance value while the samples categorized in the non-recurrence group had a negative value. Figure 2(E) shows that all samples except NR20 were correctly classified into the groups. The expression pattern of the 23 proteins in sample NR20 was more similar to that of the recurrence group, although there were no clinico-pathological features unique to this case.

The portal vein is the main route for intrahepatic metastasis, and portal invasion is correlated with early recurrence.[30] We therefore examined the proteins associated with portal invasion and the number of primary lesions. The training sample set was divided to portal invasion positive ($n = 18$) and negative ($n = 9$) cases (Table 1). We did not find a protein spot set with a misclassification rate of less than 13%. Using the selected protein spots, principal component analysis and hierarchical clustering analysis did not separate the samples according to their portal invasion status (data not shown). The number of primary lesions (single or multiple) has been also associated with early recurrence.[31] We separated the samples in the training set into two groups: cases with single primary tumors ($n = 16$) and multiple primary tumors ($n = 11$) (Table 1). We did not identify any spots based on which the HCC samples could be separated according to the number of primary tumors (data not shown).

The location of the 23 protein spots on the 2D image is demonstrated in Figure 3(A). Figure 3(B) shows the 2D differential expression images of the 23 spots. We used a standardized spot intensity, which was calculated by dividing the intensity of the Cy5 signal by that of the Cy3 signal for each protein spot in the same gel, and the apparent spot intensity in the Cy5 image did not necessarily reflect the normalized intensity (Fig. 3B). Thus, we demonstrate the standardized spot intensity across the samples used to compare the expression level among the samples (Fig. 4). The mean ratios of the identified proteins in the recurrence and non-recurrence groups are summarized in Table 2.

The mass spectrum of tryptic peptides in spot 1400 is shown in Figure 5A as an example of mass spectrometric protein identification. Eleven peptides were matched to those of phosphoserine aminotransferase with an Analyst QS score of 945. One ion

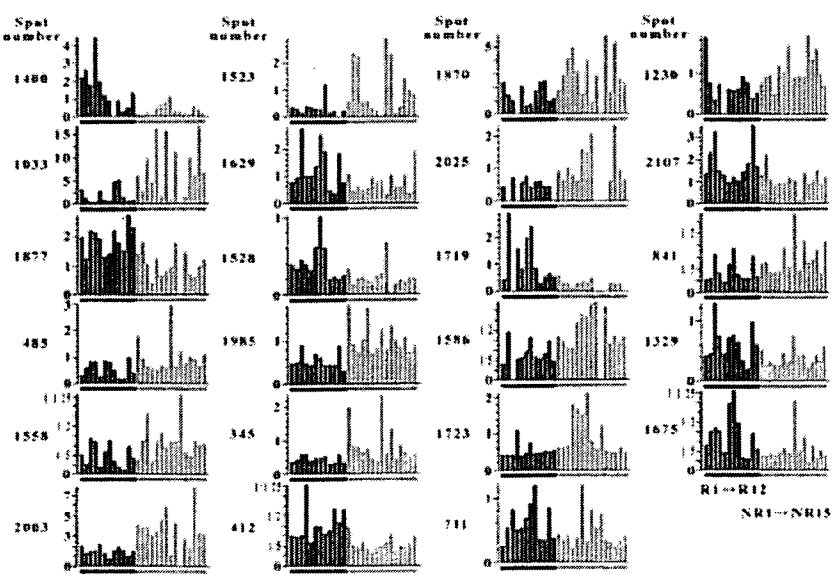


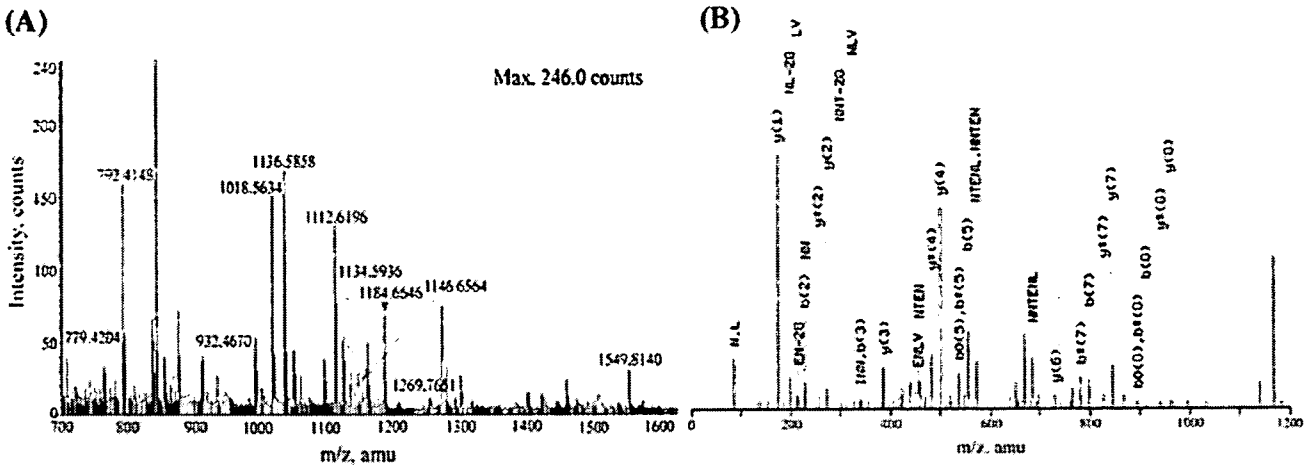
Fig. 4. Quantitative differential display of the 23 selected spots. The standardized spot intensities of all hepatocellular carcinoma (HCC) samples in the training set are displayed. Assignment of patients corresponds to Figure 2.

Table 2. List of the identified 23 proteins

Ranking[‡]	Spot no.[†]	Identification[§]	Acct.no.[¶]	Theoretic[††]		Observed[††]		Fold diff[§§]	MS score[¶¶]	Coverage	MS/MS score[¶¶] (%)[†††]	Function	Known connection to HCC	References
				Mr	pI	Mr	pI	R /NR						
1	1400	Phosphoserine aminotransferase	Q9Y617	40.4	7.7	38.4	7.9	3.8	642	37.6	41	Serine biosynthesis	–	–
2	1033	IFP53/Tryptophanyl-Trna synthetase	P23381	53.2	5.8	48.8	5.8	0.27	212	13.4	66	DNA/RNA synthesis	–	–
3	1877	Peroxiredoxin 3	P30048	27.7	7.7	25.7	5.9	1.82	171	21.9	51	Antioxidant	yes	40
4	485	GRP94 (Tumor rejection antigen 1)	P14625	92.5	4.8	93.5	4.9	0.59	1574	35.1	–	Chaperone	yes	41
5	1558	Annexin A1 (Phospholipase A2 inhibitory protein)	P04083	38.6	6.6	34.1	6.8	0.55	382	14.7	–	Signal transduction	yes	42
6	2003	Superoxide dismutase (Cu-Zn)	P00441	15.8	5.7	21.4	5.8	0.44	306	22.7	31	Antioxidant	yes	45
7	1523	Annexin A2	P07355	38.5	7.6	34.7	7.7	0.42	548	18.9	–	Signal transduction	–	–
8	1629	Annexin A4	P09525	35.8	5.9	31.8	5.8	1.6	647	25.7	–	Signal transduction	–	–
9	1528	L-lactate dehydrogenase B chain	P07195	36.5	5.7	34.4	5.8	1.61	427	10.8	–	Glucose metabolism	yes	44
10	1985	Cyclophilin B (Peptidyl-proiyl cis-trans isomerase B)	P23284	22.7	9.3	22.3	10	0.53	726	43.8	–	Signal ransduction	yes	47
11	345	ORP150	O9Y4L1	111.3	5.2	115	5.2	0.59	671	23.9	–	Chaperone	–	–
12	412	Vinculin	P12003	124.4	5.4	99.8	5.9	1.82	597	49.2	–	Cytoskeletal structure	yes	13,14
13	1870	Flavin reductase	P30043	21.9	7.3	25.8	8.1	0.52	106	14.1	48	Antioxidant	–	–
14	2025	Stathmin	P16949	17.2	5.8	20.9	5.8	0.56	269	34.9	79	Signal transduction	yes	–
15	1719	Protein kinase C inhibitor protein-1 (14-3-3 zeta/delta)	P63104	27.8	4.7	29.2	4.6	2.7	1306	57.1	–	Signal transduction	–	–
16	1586	Heterogeneous nuclear ribonucleoprotein A2/B1	P22626	37.4	9	33.4	9.5	0.55	597	20.1	–	Pre-mRNA processing	–	–
17	1723	26S proteasome non-ATPase regulatory subunit 8	P48556	30	6.9	29.1	6.9	0.64	104	8.9	38	Protein degradation	–	–
18	711	Glutamine amidotransferase	P49915	76.7	6.4	74.6	6.7	1.31	945	51.2	53	DNA and RNA synthesis	–	–
19	1230	Phosphoglycerate kinase 1	P00558	44.6	8.3	40.6	9.1	0.69	234	34.9	–	Glucose metabolism	–	–
20	2107	TCP-1 chaperonin cofactor A	O75347	12.7	5.3	18.1	5.3	1.57	72	26.9	44	Chaperonine	yes	47
21	841	TCP-1-gamma	P49368	60.4	6.1	65.9	6.6	0.64	783	27.8	–	Chaperonine	yes	48
22	1329	Multidrug resistance-associ ated protein MGr1-Ag	P08865	32.9	4.8	39	4.7	1.51	405	15.3	74	Cell adhesion	yes	49
23	1675	Nuclear chloride ion channel 27	O00299	26.9	5.1	30.5	5.2	1.64	1041	44	–	Ion channel	yes	38

[‡]Ranking orders are determined by support vector machine weight. [†]Spot numbers refer to those in Figure 3 and 4. [§]Identifications are according to Swiss-Prot and NCBI database. [¶]Accetion numbers are derived from Swiss-Prot database. [††]Theoretical molecular weight (Mr) and isoelectric point (pI) are determined by Swiss-Prot and the ExPASy databas (http://au.expasy.org). [††]Theoretical molecular weight (Mr) and isoelectric point (pI) are determined by the location of the spots on the 2D image. [§§]Fold differences indicate the ratio between the mean intensity of the recurrence and non-recurrence sample groups. [¶¶]MS scores are calculated by MasCot program. [†††]Coverages indicate the percentage of the length of the identified peptide against the full length protein. [¶¶]MS/MS scores are calculated by MasCot program. The blank (–) indicate that MS/MS analysis was not successfully achieved.

doi: 10.1111/j.1349-7006.2007.00441.x

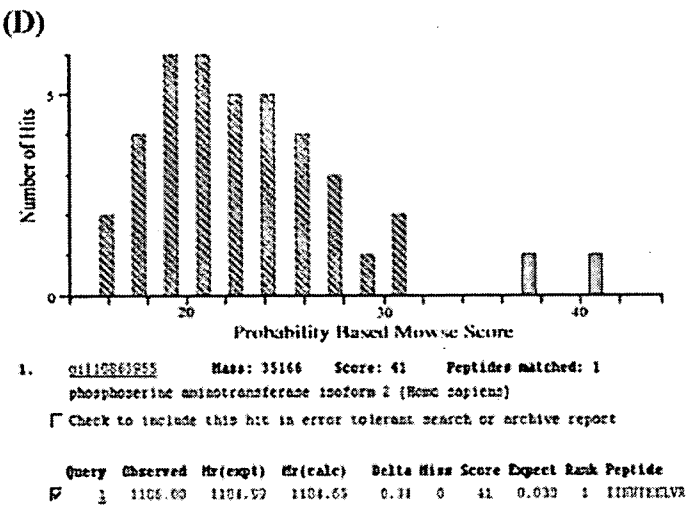


(C) Monotopic mass of neutral peptide Mr (calc): 1184.65

Ions Score: 41 Expect: 0.038

Matches (Bold): 44/104 fragment ions using 52 most intense peaks

#	Immon.	b	b*	b0	Seq.	y	y*	y0	#
1	86.10	114.09			I				10
2	86.10	227.18			I	1072.57	1055.55	1054.56	9
3	87.06	341.22	324.19		N	959.49	942.46	941.48	8
4	87.06	455.26	438.23		N	845.45	828.42	827.44	7
5	74.06	556.31	539.28	538.30	T	731.40	714.38	713.39	6
6	102.05	685.35	668.32	667.34	E	630.36	613.33	612.35	5
7	87.06	799.39	782.37	781.38	N	501.31	484.29		4
8	86.10	912.48	895.45	894.47	L	387.27	370.24		3
9	72.08	1011.55	994.52	993.54	V	274.19	257.16		2
10	129.11				R	175.12	158.09		1



(E)

MDAPRQVVNF GPGPAKLPHS VLLEIQKELL DYKGVGISVL EMSHRSSDFA
KEINNTENLV RELLAVPDNY KVIFLQGGGC GQFSAVPLNL IGLKAGRCAD
YVVTGAWSAK AAEEAKKFGT INIVHPKLGS YTKIPDPSTW NLNPDASYVY
YCANETVHGV EFDFIPDVKG AVLVCDMSSN FLSKPVDVSK FGVIFAGAQK
NVGSAGVTVV IVRDDLLGFA LRECPSVLEY KVQAGNSSLY NTPPCFSIYV
MGLVLEWIKN NGGAAAMEKL SSIKSQTIYE IIDNSQGFYV CPVEPQNRSK
MNIPFRIGNA KGDDALEKRF LDKALELNML SLKGHRSVGG IRASLYNAVT
IEDVQKLAAF MKKFLEMHQL

Fig. 5. Mass spectrometric protein identification. (A) Mass spectrogram of spot 1400. Peptide ions with m/z-values were used for protein identification. The ion peak shown with an arrow was subjected to MS/MS study. (B) MS/MS spectrogram of the peptide with m/z 1184.65. (C, D) Database search by the Mascot program using the MS/MS data. (E) Sequence of phosphoserine aminotransferase. The sequences used for the protein identification are underlined.

peak (m/z 1184.6646) was further processed for tandem mass spectrometry analysis, resulting in identification of the same protein (Fig. 5B–D). The peptide sequence used for the identification is shown in Figure 5(E). The other identified proteins are summarized in Table 2.

Discussion

We identified a set of 23 proteins whose expression pattern is significantly associated with early intrahepatic recurrence of hepatocellular carcinoma. Based on the expression profile of

this set and using hierarchical clustering and principal component analysis, the HCC tumor samples were accurately classified into two groups according to their recurrence status, while they were ambiguously classified when all protein spots were used as the basis for their classification. As the prediction value of these proteins was validated by external validation samples, these proteins may be good candidates for tumor markers to predict early recurrence of HCC. The identification of patients at high risk for recurrence would be useful for further developing therapeutic strategies to possibly include prophylactic liver transplantation, an idea that would require the study of a larger number of HCC samples to be confirmed.

The proteins we identified in this study are involved in a wide range of biological processes, suggesting that various factors determine the malignant behavior of HCC cells. Phosphoserine aminotransferase is involved in the L-serine biosynthesis pathway[32] and enzymic imbalance of serine metabolism in rat hepatoma has been reported previously.[33] As phosphoserine aminotransferase is highly expressed in tissues with a high rate of cell turnover[34] its up-regulation in HCC tissue in cases showing early recurrence may reflect the proliferation potential of the tumor cells. We also identified proteins involved in DNA/RNA synthesis, such as IFP53 and glutamine amidotransferase, premRNA processing, such as hnRNP A2/B1, chaperones such as GRP94, ORP150 and peptidyl-prolyl *cis-trans* isomerase B, chaperonins, such as TCP-1-chaperonin cofactor A and TCP-1 gamma, and proteins in the proteolytic pathway, such as the 26S proteasome subunit. As these proteins play a key role in fundamental biological processes, from protein synthesis to degradation, they may be responsible for the overall proteomic differences observed between cases with and without recurrence (Fig. 2A). We identified three antioxidant proteins, namely peroxiredocin 3, superoxide dismutase and flavin reductase. Oxidative stress has been considered as a modulator of cancer-relevant signaling pathways leading to the accumulation of oncogenic mutations in HCC.[35] Our findings may either reflect the defensive response to the oxidative stress or the fact that tumor cells were clonally selected by the oxidative stress. Indeed, the proteins involved in the signal transduction pathways of the annexin families (annexin A1, 2, and 4) and protein kinase C (14-3-3 zeta/delta), and stathmin, a proliferation regulator, showed different expression levels between the early recurrence and non-recurrence groups. Stathmin depolymerizes microtubules and plays a key role in signal transduction and cellular proliferation. Stathmin has been implicated in carcinogenesis and cancer progression in many types of malignant tumors, and was recently linked with carcinogenesis in HCC.[36] Two enzymes involved in glucose metabolism, phosphoglycerate kinase 1 and L-lactate dehydrogenase, were associated with early recurrence. Tumor formation is generally linked with increased activity of glycolytic enzymes, and a previous proteomic study on liver cancer reported the aberrant regulation of several glycolytic enzymes.[27] In lung cancer patients, a proteomic study revealed that higher expression of PGK1 was also associated with reduced survival[37] suggesting the importance of aberrant glycolysis in cancer progression. Nuclear chloride ion channel 27 was increased during the course of carcinogenesis in HCC[38] and was also up-regulated in the early recurrence group in our study. We observed up-regulation of the resistance-associated protein MGr1-Ag in the early recurrence group. Recently, MGr1-Ag turned out to be a ligand for PCK3145, an antimetastatic synthetic peptide for hormone-refractory prostate cancer[39] possibly presenting a novel molecular therapy target against early recurrence HCC. The early recurrence and non-recurrence groups in the training set had the different number of primary tumor, venous invasion and TNM staging in the training set (Table 1). Therefore, these proteins might be involved in the mechanisms of these clinical features in a certain coordinated manner. However, as these clinical parameters were not significantly different between the two groups in the test set and these 23 proteins could distinguish them, the developed classifier may have more prediction value than these clinical parameters.

We may need an extensive validation study using a large-scale sample set from the multiple hospitals to apply the present results for clinical applications because various clinical backgrounds should influence the prediction performance of the biomarkers. 2D-DIGE is a powerful tool to discover the biomarker candidate, because it can provide quantitative protein expression data in a reproducible way across multiple samples in a relatively short time. However, it may be difficult to use 2D-DIGE as a routine clinical examination tool in local hospitals because of its technical complexity and expensive initial and running costs. A more popular, convenient and less expensive examination tool, such as ELISA should be considered for screening purposes. Array technology has already enabled us to monitor the expression level of multiple proteins across a large number of samples in a less labor-intense, high-throughput and less expensive way, and the development of the specific antibody against the identified protein variants should be the key step to applying the proteomics results in the clinical setting.

Literature validation showed that many of the identified proteins were associated with HCC (Table 2).[38,40–49] However, with the exception of vinculin[14,15] the genes reported to be associated with early recurrence of HCC in transcriptomic studies were not identified in this study, a fact possibly due to the current detection limitations of 2D-DIGE using the CyDye DIGE Fluor minimal dye. To improve the sensitivity of our proteomic studies, we are now considering the use of prefractionation methods, narrow range IPG gels, more sensitive fluorescent dye such as the CyDye DIGE Fluor saturation dye, and larger format 2D gels.[50] We found that the large format 2D gel with CyDye DIGE Fluor saturation dye could increase the number of spots up to 5000.[51] The extended version of 2D-DIGE linking to the modern mass spectrometry should be considered as a key modality of cancer proteomics. In addition, the combination of multiple proteomic modalities such as the above, mass spectrometry and array technology will act in a complementary way and result in a comprehensive view of liver cancer proteomics.

Acknowledgments

This work was supported by a grant from the Ministry of Health, Labor and Welfare and by the Program for Promotion of Fundamental Studies in Health Sciences of the National Institute of Biomedical Innovation of Japan.

References

1 Nagao T, Inoue S, Yoshimi F *et al.* Postoperative recurrence of hepatocellular carcinoma. *Ann Surg* 1990; 211: 28–33.
2 Iwo SC, Chiu JH, Chau GY, Loong CC, Lui WY. Risk factors linked to tumor recurrence of human hepatocellular carcinoma after hepatic resection. *Hepatology* 1992; 16: 1367–71.
3 Harada T, Shigemura F, Kodama S, Higuchi T, Ikeda S, Okazaki M. Hepatic resection is not enough for hepatocellular carcinoma: A follow-up study of 92 patients. *J Clin Gastroenterol* 1992; 14: 245–50.
4 Nagasue N, Uchida M, Makino Y *et al.* Incidence and factors associated with intrahepatic recurrence following resection of hepatocellular carcinoma. *Gastroenterology* 1993; 105: 488–94.
5 Yamamoto J, Kosuge T, Takayama T *et al.* Recurrence of hepatocellular carcinoma after surgery. *Br J Surg* 1996; 83: 1219–22.
6 Cha C, Fong Y, Jarnagin WR, Blumgart LH, DeMatteo RP. Predictors and

doi: 10.1111/j.1349-7006.2007.00441.x