JCBi JOURNAL OF CLINICAL BIOINFORMATICS

REVIEW Open Access

Bioinformatic-driven search for metabolic biomarkers in disease

Christian Baumgartner[1*], Melanie Osl[1], Michael Netzer[1], Daniela Baumgartner[2]

Abstract

The search and validation of novel disease biomarkers requires the complementary power of professional study planning and execution, modern profiling technologies and related bioinformatics tools for data analysis and interpretation. Biomarkers have considerable impact on the care of patients and are urgently needed for advancing diagnostics, prognostics and treatment of disease. This survey article highlights emerging bioinformatics methods for biomarker discovery in clinical metabolomics, focusing on the problem of data preprocessing and consolidation, the data-driven search, verification, prioritization and biological interpretation of putative metabolic candidate biomarkers in disease. In particular, data mining tools suitable for the application to omic data gathered from most frequently-used type of experimental designs, such as case-control or longitudinal biomarker cohort studies, are reviewed and case examples of selected discovery steps are delineated in more detail. This review demonstrates that clinical bioinformatics has evolved into an essential element of biomarker discovery, translating new innovations and successes in profiling technologies and bioinformatics to clinical application.

Biomarkers, profiling technologies and bioinformatics

By definition, biomarkers are "objectively measured indicators of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention, and ... are intended to substitute for a clinical endpoint (predict benefit or harm) based on epidemiological, therapeutic, pathophysiological or other scientific evidence (Biomarkers Definitions Working Group, 2001)" and have a variety of functions [1]. From the clinical perspective, biomarkers have a substantial impact on the care of patients who are suspected to have disease, or those who have or have no apparent disease. According to this categorization, biomarkers can be classified into diagnostic, prognostic and screening biomarkers. The latter are of high interest because of their ability to predict future events, but currently there are few accepted biomarkers for disease screening [2-4].

Advances in omic profiling technologies allow the systemic analysis and characterization of alterations in genes, RNA, proteins and metabolites, and offer the possibility of discovering novel biomarkers and pathways activated in disease or associated with disease conditions [5-7]. The proteome, as an example, is highly dynamic due to the diversity and regulative structure of posttranslational modifications, and gives an in-depth insight into disease; this is because protein biomarkers reflect the state of a cell or cellular subsystem determined by expression of a set of common genes. Many interesting proteins related to human disease, however, are low-abundance molecules and can be analyzed by modern mass-spectrometry (MS) -based proteomics instrumentations, even if these technologies are somewhat limited due to their moderate sensitivity and the dynamic range necessary for high-throughput analysis [8]. In metabolomics, metabolite profiling platforms, using tandem mass spectrometry (MS/MS) coupled with liquid chromatography (LC), allow the analysis of low-molecular weight analytes in biological mixtures such as blood, urine or tissue with high sensitivity and structural specificity, but still preclude the analysis of large numbers of samples [9,10]. More recently, whole spectrum analysis of the human breath in liver disease or cancer using ion-molecule reaction (IMR) or proton transfer reaction (PTR) mass spectrometry represents a further layer of potential applications in the field of biomarker discovery, as a breath sample can be obtained non-invasively

* Correspondence: christian.baumgartner@umit.at
[1]Research Group for Clinical Bioinformatics, Institute of Electrical, Electronic and Bioengineering, University for Health Sciences, Medical Informatics and Technology (UMIT), Hall in Tirol, Austria
Full list of author information is available at the end of the article

and its constituents directly reflect concentrations in the blood [11,12].

In general, the search, verification, biological and biochemical interpretation and independent validation of disease biomarkers require new innovations in high-throughput technologies, biostatistics and bioinformatics, and thus make necessary the interdisciplinary expertise and teamwork of clinicians, biologists, analytical- and biochemists, and bioinformaticians to carry out all steps of a biomarker cohort study with professional planning, implementation, and control. Generally in human biomarker discovery studies, a variety of experimental designs are used. These include case-control or more complex cohort study designs such as crossover or serial sampling designs. Retrospective case-control studies is the type of epidemiological study most frequently used to identify biomarkers, by comparing patients who have a specific medical condition (cases) with individuals who do not have this condition but have other similar phenotypic and patient specific characteristics (controls). In contrast, longitudinal cohort studies allow patients to serve as their own biological control, which reduces the interindividual variability observed in multiple cohort studies as well as the technology platform-based variability due to a moderate signal-to-noise ratio [13].

Bioinformatics plays a key role in the biomarker discovery process, bridging the gap between initial discovery phases such as experimental design, clinical study execution, and bioanalytics, including sample preparation, separation and high-throughput profiling and independent validation of identified candidate biomarkers. Figure. 1 shows the typical workflow of a biomarker discovery process in clinical metabolomics.

In this survey article, we review and discuss emerging bioinformatic approaches for metabolomic biomarker discovery in human disease, delineating how data mining concepts are being selected and applied to the problem of identifying, prioritizing, interpreting and validating clinically useful metabolic biomarkers.

Quality controlled collection and integration of biomedical data

Central to biomedical research is a Good Clinical Practice (GCP) compliant data collection of patient-related records, which accommodates the quality controlled collection and tracking of samples and additional study material. This practice necessitates a carefully executed, standardized integration of generated omic/epigenetic data and clinical information including biochemistry, pathology and follow-up. If required, it also must be made complete with data from public repositories such as Enzyme, KEGG, Gene Ontology, NCBI Taxonomy, SwissProt or TrEMBL and literature (e.g PubMed) using appropriate data warehouse solutions. In the past few years in particular, the bioinformatics community has made great progress in developing data warehouse applications in a biomedical context for improved management and integration of the large volumes of data generated by various disciplines in life sciences.

A data warehouse is a central collection or repository that continuously and permanently stores all of the relevant data and information for analysis. Coupled with intelligent search, data mining and discovery tools, it enables the collection and processing of these data to turn them into new biomedical knowledge [14,15]. Technically, we need to distinguish between the back room and front room entities, as these two parts are usually separated physically and logically. While the back room holds and manages the data, the front room usually enables data accession and data mining. In comprehensive biomarker cohort studies, a data warehouse is an essential bioinformatic tool for standardized collection and integration of biomedical data, as well as meta-analysis of clinical, omic and literature data under the constraints of well-phenotyped patients' cohorts to discover and establish new biomarkers for early diagnosis and treatment.

Fundamental statistic concepts, data mining methods and meta-analysis

Once a biomarker cohort study has been set up, and sample collection, preparation, separation and MS analysis have been carried out, an extensive technical review of generated data is essential to ensure a high degree of consistency, completeness and reproducibility in the data.

Data preprocessing, as a preliminary data mining practice performed on the raw data, is necessary to transform data into a format that will be more easily and effectively processed for the purpose of targeted analyses. There are a number of methods used for data preprocessing, including data transformation (e.g. logarithmic scaling of data) and normalization, e.g. using z-transformation, data sampling or outlier detection. In particular, the problem of detecting and cleaning datasets from outliers is a crucial task in data preprocessing. Thus, a careful handling of outliers is warranted to avoid manipulation and distortion of statistical results, which complicates a useful interpretation of biological findings. Traditional statistical approaches propose observations as outliers that are deemed unlikely with respect to mean and standard deviation, assuming normal data distribution. A common model uses the interquartile ranges and defines an outlier as observation outside the interquartile range IQR = Q_3 - Q_1, where Q_1 and Q_3 are the first and third quartiles. However, alternative data mining methods try to overcome concepts based on the assumption that data is normally

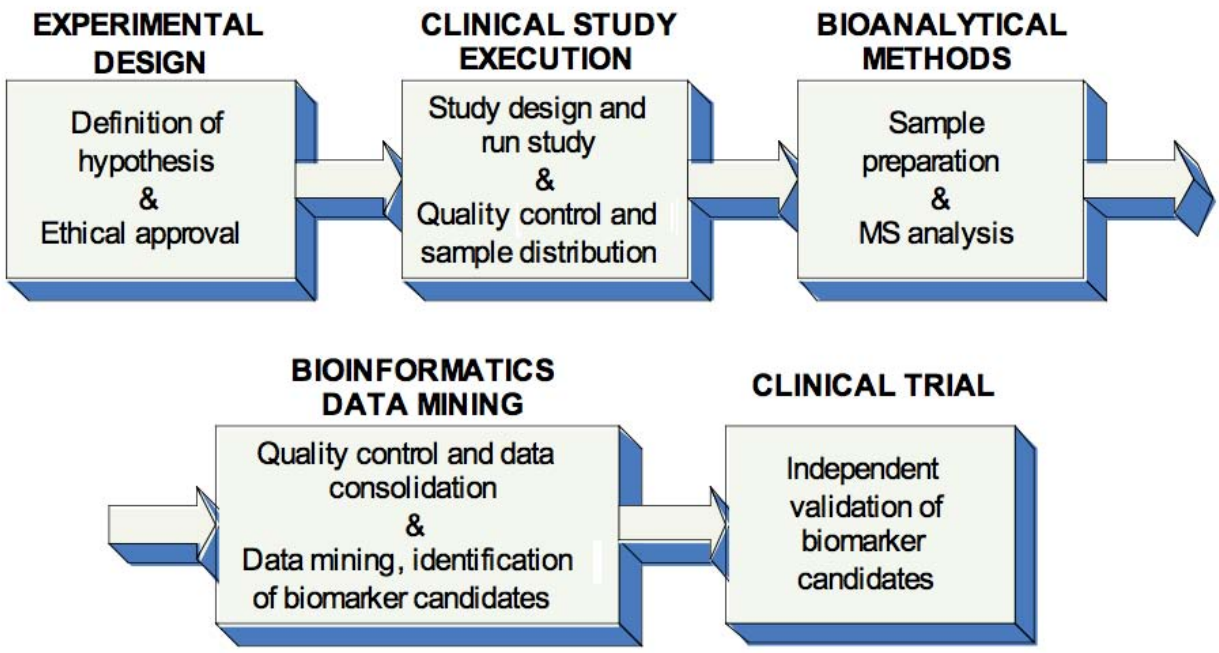

Figure 1 Biomarker discovery process in human disease using an MS-based metabolite profiling platform.

distributed, by using distance-based approaches or defining the outlier problem via a local neighborhood of data points in a given data space, such as the local outlier factor (LOF) or the algorithm LOCI, using a local correlation integral for detecting outliers [16-18]. These methods show high value in treating the problem of outlier detection, especially in multiple biomarker search problems.

In recent years, various powerful data mining and statistical bioinformatics methods have been propagated for identifying, prioritizing and classifying robust and generalizable biomarkers with high discriminatory ability [19-27]. Principal data mining tasks in biomarker discovery, such as the identification of biomarker candidates in experimental data (feature selection) and classification, are "supervised" because study cohorts are well phenotyped in carefully designed and controlled clinical trials. Therefore, data vectors are determined by a set of tuples, $T = \{(c_j, a) \mid c_j \in C, a \in A\}$, where c_j is a class label from the collection C of pre-classified cohorts (normal, diseased, various stages of disease, treated, at rest, during stress, etc.), and $A = \{a \mid a_1, \ldots, a_n\}$ is the set of concentrations of low-molecular weight biomolecules such as nucleotides, amino and organic acids, lipids, sugars, etc., if molecules are predefined and quantified, or simple m/z values from generated raw mass spectra. In this area, basic data mining concepts for the search of biomarker candidates constitute filter- and wrapper-based feature selection algorithms, and more advanced paradigms like embedded or ensemble methods [27-31]. However, if class membership is (partly) unknown, semi- or unsupervised techniques (cluster analysis) are helpful tools for biomarker search and interpretation. Note that many unsupervised feature selection methods treat this task as a search problem. Since the data space is exponential in the number of examined features, the use of heuristic search procedures are necessary where the search is combined with a feature utility estimator to evaluate and assess the relative merit of selected subsets of features. Supervised clustering, for example, opens a new research field in biomarker discovery to be employed when class labels of all data are known, with the objective of finding class pure clusters. Table 1 gives a survey of widely-used supervised feature selection techniques, useful for the identification of candidate biomarkers in data sets gathered from well-phenotyped cohort studies, considering both basic types of paired and unpaired test hypotheses [32-40].

Recently, combined biomarkers constructed by mathematical expressions such as quotients or products have been utilized to significantly enhance their predictive value, as demonstrated in newborn screening [41,42]. For example, a simple model for screening for phenylananine hydroxylase deficiency (PKU), a common congenital error of metabolism, was proposed by the ratio Phe/Tyr (Phe is phenylananine and Thy is tyrosine), to describe the irreversible reaction A→B of a reactant A

Table 1 Commonly used supervised data mining methods for the search and prioritization of biomarker candidates in independent and dependent samples

Independent samples	Method	Basic principle and key features of the method	Reference
	Unpaired null hypothesis testing (Two-sample t-test*, Mann-Whitney-U test°)	- univariate filter method - P value serves as evaluation measure for the discriminatory ability of variables - is an accepted statistical measure - appropriate for two class problems only - P value is sample size dependent	Lehmann, *Springer Verlag*, 2005 [32]
	Principal component analysis (PCA)#	- unsupervised projection method - PCA calculates linear combinations of variables based on the variance of the original data space - appropriate for multiple class problems - visualizable loading and score plots (scores can be labeled according to class membership) - no ranking and prioritization of features possible	Jolliffe, *Springer Verlag*, 2005 [33], Ringnér, *Nat Biotechnol*, 2008 [34]
	Information gain (IG)	- univariate filter method - IG calculates how well a given feature separates data by pursuing reduction of entropy - appropriate for multiple class problems - quick and effective ranking of features - IG scores permit prioritization of features	Hall and Holmes, *IEEE Trans Knowl Data Eng*, 2003 [28]
	ReliefF (RF)	- multivariate filter method - RF score relies on the concept that values of a significant feature are correlated with the feature values of an instance of the same class, and uncorrelated with the feature values of an instance of the other class - appropriate for multiple class problems - RF scores permit prioritization of features	Robnik-Sikonja & Kononenko, *Mach Learn*, 2003 [35] Hall and Holmes, *IEEE Trans Knowl Data Eng*, 2003 [28]
	Associative voting (AV)	- multivariate filter method - AV uses a rule-based evaluation criterion by a special form of association rules; considers interaction among features - appropriate for two class problems only - AV scores permit prioritization of features - restriction of the rule search space necessary	Osl et al., *Bioinformatics*, 2008 [36]
	Unpaired Biomarker Identifier (uBI)	- univariate filter method - statistical evaluation score by combining a discriminance measure with a biological effect term - appropriate for two class problems only - quick and effective ranking of features - uBI scores permit prioritization of features - uBI scores closely related to pBI scores	Baumgartner et al., *Bioinformatics*, 2010 [13]
	Guilt-by-association feature selection (GBA-FS)	- multivariate subset selection method - GBA-FS uses a hierarchical clustering with correlation as distance measure; the most relevant features of each cluster are assessed by their discriminatory power, as measured for example by two-sample t-test - accounts for redundancy between features - appropriate for two class problems only	Shin et al., *J Biomed Inform*, 2007 [37]
	Support vector machine-recursive feature elimination (SVM-REF)	- embedded selection method - SVM-REF uses optimized weights of SVM classifier to rank features - appropriate for two class problems only	Guyon et al., *Mach Learn*, 2002 [38]
	Random forest models (RFM)	- embedded selection method - RFM uses bagging and random subspace methods to construct a collection of decision trees aiming at identifying a complete set of significant features - appropriate for multiple class problems	Enot et al., *PNAS*, 2006 [39]
	Aggregating feature selection (AFS)	- ensemble selection method - aggregating multiple feature selection results to a consensus ranking, e.g. using the concept of weighted voting or by counting the most frequently selected features to derive the consensus feature subset - appropriate for multiple class problems	Saeys et al., *Lecture Notes in Artificial Intelligence*, 2008 [30]

Table 1 Commonly used supervised data mining methods for the search and prioritization of biomarker candidates in independent and dependent samples *(Continued)*

	Stacked feature ranking (SFR)	- ensemble selection method - stacked learning architecture to construct a consensus feature ranking by combining multiple feature selection methods - appropriate for multiple class problems - feature selection by optimizing the discriminatory ability (AUC)	Netzer et al., *Bioinformatics*, 2009 [31]
	Wrapper approach	- evaluating the merit of a feature subset by accuracy estimates using a classifier - produces subsets of very few features that are dominated by stronger and uncorrelated attributes - increased computational runtime; necessitates heuristic search methods like forward selection, backward elimination, or more sophisticated methods such as genetic algorithms	Hall and Holmes, *IEEE Trans Knowl Data Eng*, 2003 [28]
Dependent samples	Paired null hypothesis testing (Paired t-test*, Wilcoxon signed-rank test°)	- univariate filter method - P value serves as evaluation measure for the discriminatory ability of variables - is an accepted statistical measure - appropriate for two class problems only - P value is sample size dependent - two dependent samples	Lehmann, *Springer Verlag*, 2005 [32]
	Repeated measure analysis	- univariate and multivariate approaches - mixed model analysis (GLMM, General Linear Mixed Model) - time series (multiple time points) analysis	Crowder & Hand, *Analysis of repeated measures*, 1990 [40]
	Paired Biomarker Identifier (pBI)	- univariate filter method - pBI uses a statistical evaluation score by combining a discriminance measure with a biological effect term - appropriate for two class problems only - pBI scores permit prioritization of features - pBI scores closely related to uBI scores	Baumgartner et al., *Bioinformatics*, 2010 [13]

* data normal distributed, ° data non-normal distributed. # PCA is an unsupervised method also used for data containing class information. All algorithms are run on continuous data as data generated in metabolomics are usually of metric nature. Data can represent absolute metabolite concentrations (given as intensity counts or more specific in μmol/L if internal standards are available) or simple m/z values from raw or preprocessed mass spectra.

into a product B, caused by an impaired enzyme activity [43]. In this manner, models of single and combined predictors, as built upon *a priori* knowledge of abnormal pathways like those shown above, exhibit high potential to develop screening models with high discriminatory ability. Ultimately, the process of identifying clinically relevant biomarkers is an ambitious data-mining task, bringing together various computational concepts of feature ranking, subset selection and feature construction by attribute combination.

The identification of a set of relevant, but not redundant, predictors is important for building prognostic and diagnostic models. Ding and Peng, for example, presented a minimum redundancy feature selection approach on microarray data, demonstrating significantly better classification accuracy on selected minimized redundant gene sets than those obtained through standard feature ranking methods [44]. Most commonly, individual features are ranked in terms of a quality criterion, out of which the top k features are selected. However, most feature-ranking methods do not sufficiently account for interactions and correlations between the features, and therefore redundancy is likely to be encountered in the selected features. Recently, Osl et al., presented a new algorithm, termed Redundancy Demoting (RD), that takes an arbitrary feature ranking as input, and improves the predictive value of a selected feature subset by identifying and demoting redundant features in a postprocessing modality [45]. The authors define redundant features as those that are correlated with other features, but are not relevant in the sense that they do not improve the discriminatory ability of a selected feature set. This means that although correlated biomarkers exhibit potential reactions and interactions among biomolecules in a biological pathway, they do not provide a substantial increase in predictive value if they are redundant. On the other hand, if they are not redundant, they may be good candidates to further enhance the predictive value of selected multiple biomarkers.

For building predictive models on biological data, a wide spectrum of machine learning methods is available: These include discriminant analysis methods like linear discriminant analysis or logistic regression analysis,

decision trees, the k-nearest neighbor classifier (k-NN), an instance-based learning algorithm, the Bayes classifier, a probabilistic method based on applying the Bayes' theorem, support vector machines, a method that uses a kernel technique to apply linear classification techniques to nonlinear classification problems or artificial neural networks [46-53]. A more detailed review of these methods, however, is beyond the scope of this article.

As an advanced and more sophisticated layer of data analysis, meta-analysis is used with the objective of improving single experiment results and identifying common clinical and biological relevant patterns [54,55]. Meta-analysis of data may contain different steps: (i) scoring disease-relevance of candidate biomarkers by integrated analysis of the different clinical and experimental data (which may arise from multiple clinical studies), (ii) building statistical models on preselected candidates, derived by coupling methods such as feature selection and logistic regression analysis that result in the highest discriminatory ability with respect to the targeted patient cohorts or populations, (iii) performing correlation analysis to analyze 'omics' data under constraints defined by the patient data, (vi) examining various performance characteristics of biomarker candidates e.g. through decision-analytic outcome modeling. Receiver-operating-characteristics (ROC) analyses of related discriminatory models with specific sensitivities and specificities are used as input parameters for decision models, calculating expected epidemiologic and economic consequences for individuals and public health of the evolving health-care technologies under assessment.

Generalizability and validation of biomarkers

Objective measures to assess the predictive value and generalizable power of selected candidate biomarkers are sensitivity, specificity, the product of sensitivity and specificity, or the area under the ROC curve (AUC). These measures are useful and valid only if they are determined on independent samples (e.g. cases versus controls). In serial sampling studies, alternative measures are needed to assess the predictive value of biomarkers in a similar manner. Very recently, a new objective measure for expressing the discriminatory ability (DA) in dependent samples was developed by our group [13]. The discriminance measure DA is defined as the percent change of analyte levels in a cohort in one direction versus baseline, and acts as a feature analogously to the product of sensitivity and specificity when addressing an unpaired test problem. Thus, a DA value of 0.5 in paired testing corresponds exactly to a product or AUC of 0.5 in unpaired testing, demonstrating no discrimination, while a DA of 0.75 or 1.00 indicates good or perfect discrimination.

Using both related discrimination measures, i.e. the product of sensitivity and specificity, and DA, a clinically useful prioritization of biomarkers - for example, into classes of weak, moderate and strong predictors - is possible independently of the study design (e.g. case-control versus serial sampling study). Very recently, Lewis et al. and Baumgartner et al. published a prospective longitudinal biomarker cohort study that was carried out to identify, categorize, and profile kinetic patterns of early metabolic biomarkers of planned (PMI) and spontaneous (SMI) myocardial infarction [56,13]. Figure. 2 depicts a kinetic map of selected circulating metabolites from a human model of PMI that faithfully reproduces SMI [57]. Promising metabolites were selected and prioritized into classes of different predictive value by using the so-called pBI scoring model, developed for longitudinal biomarker cohort studies where each patient serves as his/her own control [13]. In the given example, each circulating metabolite is able to be categorized at each time point of analysis in order to qualitatively and quantitatively assess the dynamic expression pattern of metabolic biomarkers after myocardial injury. Using this approach, a set of promising putative biomarker candidates could be identified as early as 10 minutes after the event.

In general, identified biomarker candidates need to be validated using larger sample sets, covering a broad

Metabolites	t_{10}	t_{60}	t_{120}	t_{240}
Hypoxanthine	160	198	5	-10
Malonic Acid	99	84	-39	-76
Kynurenine	-67	-59	58	75
Trimethylamine-N-Oxide	-67	-63	-5	11
Xanthine	63	83	34	-58
Alanine	-51	-27	-5	53
Threonine	-38	-30	-26	-52
Homoserine	-37	-37	-23	-75
Lactic Acid	-30	-11	30	89

-	+	
< -73	> 73	strong predictors
< -44	> 44	moderate predictors
< -21	> 21	weak predictors

Figure 2 Kinetic map of metabolites on PMI data at 10, 60, 120, and 240 minutes after myocardial injury, using the pBI scoring model for prioritization of selected metabolites into groups of weak, moderate and strong predictors. Values indicate absolute pBI scores. The thresholds for prioritization are denoted below in the list of analytes. Red color increments indicate decreasing levels, blue increasing levels. In this study, a series of metabolites in pathways associated with myocardial infarction could be identified, some of which change as early as 10 minutes after injury, a time frame where no currently available clinical biomarkers are present [13,56].

cross section of patients or populations. However, if no independent cohort for validation is available, especially if further samples are costly, hazardous or impossible to collect, cross validation is an accepted statistical strategy to assess generalizability on a single derivation cohort at this discovery stage. Usually, stratified 10-fold cross-validation is applied, which is the statistical practice of partitioning a sample of data into ten subsets, where each subset is used for testing and the remainder for training, yielding an averaged overall error estimate. For very small samples, leave-one-out cross validation using one observation for testing and n-1 observations for training is proposed to generalize findings. Alternatively, bootstrapping or permutation modalities can be used as powerful approaches for statistical validation [58-60].

As an example, Figure. 3 shows the predictive value of multiple metabolites in newborn screening data on a single derivation cohort *with* and *without* stratified 10-fold cross validation. The data set contains concentrations of 43 analytes, i.e. amino acids and acyl-carnitines, separated into 63 cases (medium-chain acyl-CoA dehydrogenase deficiency, MCADD) and 1241 healthy controls [61]. This result clearly demonstrates the strong disagreement in discriminatory ability between non- and cross-validated analyte subsets, and confirms the necessity of this computational modality for pre-selecting robust and generalizable candidate biomarkers, eliminating the potential bottleneck of taking too many candidates to the validation phase. Meta-analysis is a next logical step to further strengthen such results. However, after these crucial discovery steps, prospective trials are ultimately needed to validate the clinical benefit of assessing expression patterns of selected biomarker candidates before they can go into clinical routine.

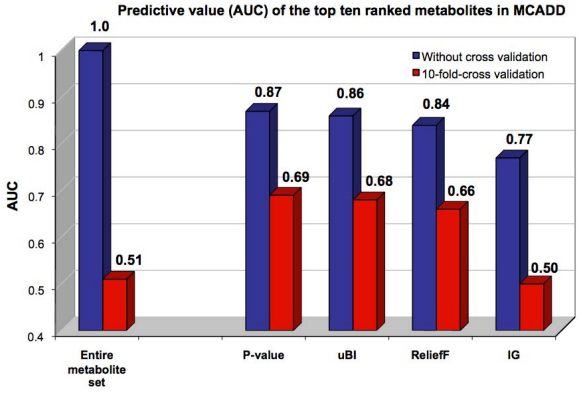

Figure 3 AUC analysis on the entire metabolite set (bars in the left), and on a set of the top ten ranked metabolites using four common feature selection methods, i.e. two sample t-test (P-value), the unpaired Biomarker Identifier (uBI), ReliefF, and Information gain (IG) on MCADD data (bars in the right). Red bars represent the predictive value expressed by the AUC of selected analyte sets, determined on a single derivation cohort *with* cross validation and blue bars *without* cross-validation. Interestingly, using the entire metabolite set (43 analytes) for distinguishing between the two groups, the discriminatory ability dropped from AUC = 1.0 (without cross validation) to AUC = 0.51 after 10-fold cross validation, thus indicating no discrimination between the cohorts. On the selected subset, the AUC dropped by 15% to 25% after cross validation, demonstrating weak predictive value and thus low generalizability of the selected subset in this experiment.

Analysis after biomarker identification

One challenging research area in bioinformatics is the biological and biochemical interpretation of identified putative marker candidates by means of mining the most likely pathways. In metabolomics, various explorer tools such as cPath, Pathway Hunter Tool (public) or Ingenuity Pathway Analysis and MetaCore (commercial) are available to visualize, map and reconstruct a spectrum of possible pathways between relevant metabolites identified by feature selection [62,63]. Most tools extract metabolic information from metabolic network databases like KEGG and provide algorithms which allow (i) querying of thousands of endogenous analytes from those databases, (ii) displaying biochemical pathways with their involved metabolite and enzymes, and (iii) reconstructing and visualizing the most likely pathways related to the identified key metabolites [24,64,65]. These tools also provide an interactive analysis of biochemical pathways and entities such as metabolites, enzymes or reactions and allow a quick and direct functional annotation of experimental findings. As an example, Figure. 4 shows the most likely pathway in the KEGG database, addressing altered concentration levels of arginine (Arg) and ornithine (Orn), respectively, in patients afflicted with severe metabolic syndrome and cardiovascular disease (MS+) versus healthy controls. Both candidate metabolites, which are closely associated with the D-Arg & D-Orn metabolism in the urea cycle, were identified by feature selection from targeted MS profiling data [24,66,67].

Direct hyperlinks to databases such as OMIM, Swiss-Prot or Prosite reveal supplementary information about these entities that can help researchers learn more about the underlying biochemical and biological mechanisms. It is obvious that emerging bioinformatics tools for exploring metabolic pathways and networks, thus allowing for mapping expression profiles of genes or proteins simultaneously onto these pathways, are of high importance for the biological interpretation of biomarkers from a systems biology viewpoint [68-70]. Such tools thus contribute to a better understanding of how genes, proteins and metabolites act and interact in such networks, and consequently how human diseases manifest themselves.

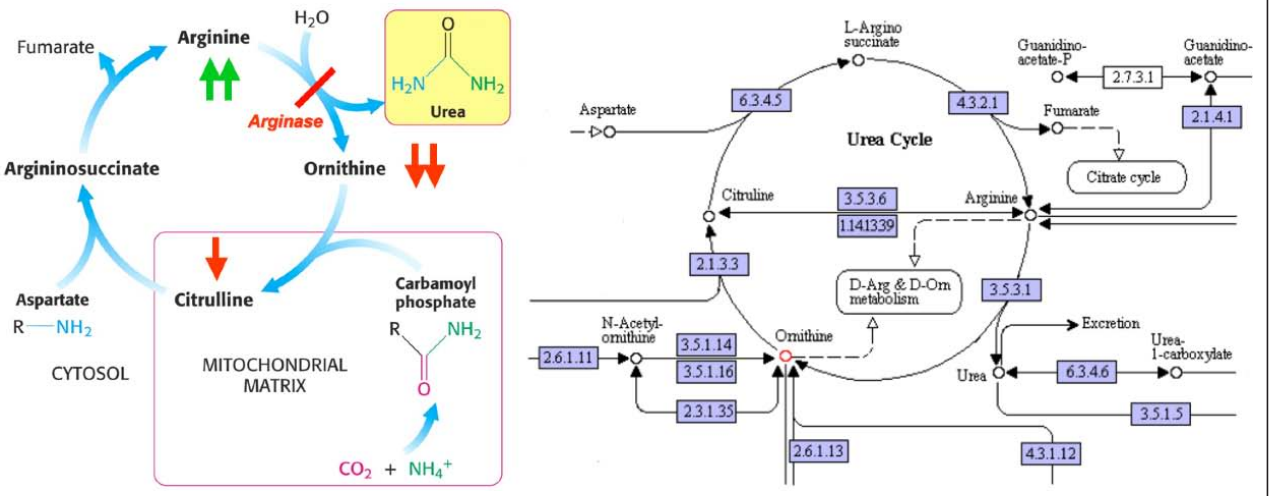

Figure 4 The high and low concentration levels of arginine (Arg) and ornithine (Orn), respectively, in patients afflicted with severe metabolic syndrome and cardiovascular disease (MS+) versus healthy controls, implied an impacted enzyme arginase in the urea cycle (left figure). The urea cycle and associated pathways from the KEGG database are depicted in the right figure. Findings could be confirmed by literature [66,67].

Conclusions and final remarks

In this article we have discussed the complementary power of modern profiling technologies and bioinformatics for metabolomic biomarker discovery in human disease. The discovery and interpretation of new biomarkers, however, depends on a comprehensive view of genomics, transcriptomics, proteomics and metabolomics [71]. In particular, proteomics and metabolomics offer excellent insights into disease, because function, structure or turnover of proteins, typically regulated via post-translational modifications, as well as metabolites, which act as end products of cellular processes, define the phenotypic heterogeneity of disease [72-74]. Therefore, great interest in the discovery of new biomarkers originates from their wide range of clinical applications, fundamental impact on pharmaceutical industry, and the current public health burden. Biomarkers, once qualified for clinical use, can aid in diagnosis and prediction of life-threatening events, confirm drug's pharmacological or biological action mechanisms, or serve as early and objective indicators of treatment efficiency in patients [75-78]. Theranostics, an emerging field in personalized medicine, utilizes molecular biomarkers to select patients for treatments that are expected to benefit them and are unlikely to produce side effects, and provides an early indication of treatment efficacy in individual patients. Therefore, theranostic tests, which lead to rapid and more accurate diagnosis and allow for a more efficient use of drugs, and thus improved patient management, are increasingly used in cancer, cardiovascular and infectious diseases, or prediction of drug toxicity [79,80].

In summary, clinical bioinformatics has evolved into an essential tool in translational research, transforming fundamental bioinformatics research to clinical application by exploiting novel profiling technologies, biological databases, data mining and biostatistics methods for speeding up biomarker and drug discovery. These useful innovations will ultimately improve individualized clinical management of patient health and will also reduce costs of drug development.

Abbreviations (in alphabetical order)

AFS: aggregating feature selection; Arg: arginine; AUC: area under the ROC curve; AV: associative voting; DA: discriminatory ability; GBA-FS: guild-by-association feature selection; GC: gas chromatography; GCP: good clinical practice; IG: information gain; IMR: ion-molecule reaction; IQR: interquartile range; KEGG: Kyoto Encyclopedia of Genes and Genomes; k-NN: k-nearest neighbor classifier; LC: liquid chromatography; LOCI: local correlation integral; LOF: local outlier factor; MCADD: medium-chain acyl-CoA dehydrogenase deficiency; MS: mass spectrometry; MS+: metabolic syndrome + cardiovascular disease; OMIM: Online Mendelian Inheritance in Man; Orn: ornithine; pBI: paired biomarker identifier; PCA: principal component analysis; Phe: phenylanine; PMI: planned myocardial infarction; PKU: phenylananine hydroxylase deficiency; PTR: proton transfer reaction; Q_1: first quartile; Q_3: third quartile; RD: redundancy demoting; RF: relief; RFM: random forest model; ROC: receiver operating characteristics; SFR: stacked feature ranking; SMI: spontaneous myocardial infarction; SVM-REF: support vector machine-recursive feature elimination; Thy: tyrosine; uBI: unpaired biomarker identifier.

Acknowledgements

The authors gratefully acknowledge support from the Austrian Genome Research Program GEN-AU and its "Bioinformatics Integration Network (BIN III)" project.

Author details

[1]Research Group for Clinical Bioinformatics, Institute of Electrical, Electronic and Bioengineering, University for Health Sciences, Medical Informatics and Technology (UMIT), Hall in Tirol, Austria. [2]Clinical Division of Pediatric Cardiology, Department of Pediatrics, Innsbruck Medical University, Austria.

Authors' contributions
CB and DB conceptualized and wrote the manuscript. MO and MN prepared table and figures and commented on the paper. All authors have read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 12 June 2010 Accepted: 20 January 2011
Published: 20 January 2011

References

1. Biomarkers Definitions Working Group: **Biomarkers and surrogate endpoints: preferred definitions and conceptual framework.** *Clin Pharmacol Ther* 2001, **69**:89-95.
2. Gerszten RE, Wang TJ: **The search for new cardiovascular biomarkers.** *Nature* 2008, **451**:949-952.
3. Lintula S, Hotakainen K: **Developing biomarkers for improved diagnosis and treatment outcome monitoring of bladder cancer.** *Expert Opin Biol Ther* 2010, **10**:1169-1180.
4. Melander O, Newton-Cheh C, Almgren P, Hedblad B, Berglund G, Engström G, Persson M, Smith JG, Magnusson M, Christensson A, Struck J, Morgenthaler NG, Bergmann A, Pencina MJ, Wang TJ: **Novel and conventional biomarkers for prediction of incident cardiovascular events in the community.** *JAMA* 2009, **302**:49-57.
5. Ackermann BL, Hale JE, Duffin KL: **The role of mass spectrometry in biomarker discovery and measurement.** *Curr Drug Metab* 2006, **7**:525-539.
6. Hood BL, Stewart NA, Conrads TP: **Development of high-throughput mass spectrometry-based approaches for cancer biomarker discovery and implementation.** *Clin Lab Med* 2009, **29**:115-138.
7. Kulasingam V, Pavlou MP, Diamandis EP: **Integrating high-throughput technologies in the quest for effective biomarkers for ovarian cancer.** *Nat Rev Cancer* 2010, **10**:371-378.
8. Rifai N, Gerszten RE: **Biomarker discovery and validation.** *Clin Chem* 2006, **52**:1635-1637.
9. Dettmer K, Aronov PA, Hammock BD: **Mass spectrometry-based metabolomics.** *Mass Spectrom Rev* 2007, **26**:51-78.
10. Lenz EM, Wilson ID: **Analytical strategies in metabonomics.** *J Proteome Res* 2007, **6**:443-458.
11. Millonig G, Praun S, Netzer M, Baumgartner C, Dornauer A, Mueller S, Villinger J, Vogel W: **Non-invasive diagnosis of liver diseases by breath analysis using an optimized ion-molecule reaction-mass spectrometry approach: a pilot study.** *Biomarkers* 2010, **15**:297-306.
12. Bajtarevic A, Ager C, Pienz M, Klieber M, Schwarz K, Ligor M, Ligor T, Filipiak W, Denz H, Fiegl M, Hilbe W, Weiss W, Lukas P, Jamnig H, Hackl M, Haidenberger A, Buszewski B, Miekisch W, Schubert J, Amann A: **Noninvasive detection of lung cancer by analysis of exhaled breath.** *BMC Cancer* 2009, **9**:348.
13. Baumgartner C, Lewis GD, Netzer M, Pfeifer B, Gerszten RE: **A new data mining approach for profiling and categorizing kinetic patterns of metabolic biomarkers after myocardial injury.** *Bioinformatics* 2010, **26**:1745-1751.
14. Pfeifer B, Aschaber J, Baumgartner C, Dreiseitl RS Modre, Schreier G, Tilg B: **A data warehouse for prostate cancer biomarker discovery.** *Proceedings of the 8th International Conference on Bioinformatics & Computational Biology, BIOCOMP 2007* Las Vegas, NV: CSREA Press; 2007, 323-327.
15. Lee TJ, Pouliot Y, Wagner V, Gupta P, Stringer-Calvert DWJ, Tenenbaum JD, Karp PD: **BioWarehouse: a bioinformatics database warehouse toolkit.** *BMC Bioinformatics* 2006, **7**:170.
16. Knorr EM, Ng RT, Tucakov V: **Distance-based outliers: algorithms and applications.** *VLDB Journal* 2000, **8**:237-253.
17. Breunig MM, Kriegel H, Ng RT, Sander J: **LOF: Identifying density-based local outliers.** *Proceedings of the ACM SIGMOD International Conference on Management of Data* 2000, 93-104.
18. Papadimitriou S, Kitagawa H, Gibbons PB, Faloutsos C: **LOCI: Fast outlier detection using the local correlation integral.** *Proceedings of the 19th International Conference on Data Engineering* 2003, 315-326.
19. Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armañanzas R, Santafé G, Pérez A, Robles V: **Machine learning in bioinformatics.** *Brief Bioinform* 2006, **7**:86-112.
20. Bhaskar H, Hoyle DC, Singh S: **Machine learning in bioinformatics: a brief survey and recommendations for practitioners.** *Comput Biol Med* 2006, **36**:1104-1125.
21. Shulaev V: **Metabolomics technology and bioinformatics.** *Brief Bioinform* 2006, **7**:128-139.
22. Tarca AL, Carey VJ, Chen XW, Romero R, Drăghici S: **Machine learning and its applications to biology.** *PLoS Comput Biol* 2007, **3**:e116.
23. Barla A, Jurman G, Riccadonna S, Merler S, Chierici M, Furlanello C: **Machine learning methods for predictive proteomics.** *Brief Bioinform* 2008, **9**:119-128.
24. Baumgartner C, Graber A: **Data mining and knowledge discovery in metabolomics.** In *Successes and new directions in data mining*. Edited by: Masseglia F, Poncelet P, Teisseire M Hershey. New York: Information Science Reference; 2008:141-166.
25. Lee JK, Williams PD, Cheon S: **Data mining in genomics.** *Clin Lab Med* 2008, **28**:145-166, viii.
26. Inza I, Calvo B, Armañanzas R, Bengoetxea E, Larrañaga P, Lozano JA: **Machine learning: an indispensable tool in bioinformatics.** *Methods Mol Biol* 2010, **593**:25-48.
27. Datta S, Pihur V: **Feature selection and machine learning with mass spectrometry data.** *Methods Mol Biol* 2010, **593**:205-229.
28. Hall MA, Holmes G: **Benchmarking attribute selection techniques for discrete class data mining.** *IEEE T Knowl Data En* 2003, **15**:1437-1447.
29. Saeys Y, Inza I, Larrañaga P: **A review of feature selection techniques in bioinformatics.** *Bioinformatics* 2007, **23**:2507-2517.
30. Saeys Y, Abeel T, Van de Peer Y: **Robust feature selection using ensemble feature selection techniques.** *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. Part II* Berlin, Heidelberg: Springer-Verlag; 2008, 313-325, Vol. 5212 of Lecture Notes in Artificial Intelligence.
31. Netzer M, Millonig G, Osl M, Pfeifer B, Praun S, Villinger J, Vogel W, Baumgartner C: **A new ensemble-based algorithm for identifying breath gas marker candidates in liver disease using ion molecule reaction mass spectrometry.** *Bioinformatics* 2009, **25**:941-947.
32. Lehmann EL, Romano JP: *Testing statistical hypotheses*. 3 edition. New York, NY: Springer Verlag; 2005.
33. Jolliffe IT: *Principal component analysis* New York, NY: Springer Verlag; 2002.
34. Ringnér M: **What is principal component analysis?** *Nat Biotechnol* 2008, **26**:303-304.
35. Robnik-Sikonja M, Kononenko I: **Theoretical and empirical analysis of relieff and rrelief.** *Mach Learning* 2003, **53**:23-69.
36. Osl M, Dreiseitl S, Pfeifer B, Weinberger K, Klocker H, Bartsch G, Schäfer G, Tilg B, Graber A, Baumgartner C: **A new rule-based data mining algorithm for identifying metabolic markers in prostate cancer using tandem mass spectrometry.** *Bioinformatics* 2008, **24**:2908-2914.
37. Shin H, Sheu B, Joseph M, Markey MK: **Guilt-by-association feature selection: identifying biomarkers from proteomic profiles.** *J Biomed Inform* 2008, **41**:124-136.
38. Guyon I, Weston J, Barnhill S, Vapnik V: **Gene Selection for Cancer Classification using Support Vector Machines.** *Mach Learn* 2002, **46**:389-422.
39. Enot DP, Beckmann M, Overy D, Draper J: **Predicting interpretability of metabolome models based on behavior, putative identity, and biological relevance of explanatory signals.** *Proc Natl Acad Sci USA* 2006, **103**:14865-14870.
40. Crowder MJ, Hand DJ: *Analysis of repeated measures* Boca Raton, FL: Chapman &Hall/CRC Press; 1990.
41. Baumgartner C, Böhm C, Baumgartner D: **Modelling of classification rules on metabolic patterns including machine learning and expert knowledge.** *J Biomed Inform* 2005, **38**:89-98.
42. Ho S, Lukacs Z, Hoffmann GF, Lindner M, Wetter T: **Feature construction can improve diagnostic criteria for high-dimensional metabolic data in newborn screening for medium-chain acyl-CoA dehydrogenase deficiency.** *Clin Chem* 2007, **53**:1330-1337.
43. Chace DH, Sherwin JE, Hillman SL, Lorey F, Cunningham GC: **Use of phenylalanine-totyrosine ratio determined by tandem mass spectrometry to improve newborn screening for phenylketonuria of early discharge specimens collected in the first 24 hours.** *Clin Chem* 1998, **44**:2405-2409.
44. Ding C, Peng HC: **Minimum Redundancy Feature Selection from Microarray Gene Expression Data.** *Proceedings of the Second IEEE*

Computational Systems Bioinformatics Conference Stanford, CA; 2003, 523-528.
45. Osl M, Dreiseitl S, Netzer M, Cerqueira F, Pfeifer B, Baumgartner C: **Demoting redundant features to improve the discriminatory ability in cancer data.** *J Biomed Inform* 2009, **42**:721-725.
46. Hosmer DW, Lemeshow S: *Applied logistic regression.* 2 edition. New York, NY: Wiley; 2000.
47. Quinlan JR: *C4.5: Programs for Machine Learning* San Francisco, CA: Morgan Kaufmann; 1993.
48. Mitchell TM: *Machine learning* Boston, MA: McGraw-Hill; 1997.
49. Gelman A, Carlin JB, Stern HS, Rubin DB: *Bayesian data analysis.* 2 edition. Boca Raton, FL: Chapman & Hall/CRC Press; 2004.
50. Shawe-Taylor J, Cristianini N: *Kernel methods for pattern analysis* Cambridge, UK: Cambridge University Press; 2004.
51. Yang ZR: **Biological applications of support vector machines.** *Brief Bioinform* 2004, **5**:328-338.
52. Raudys S: *Statistical and neural classifiers* London: Springer-Verlag; 2001.
53. Harper PR: **A review and comparison of classification algorithms for medical decision making.** *Health Policy* 2005, **71**:315-331.
54. Bagos PG, Nikolopoulos GK: **A method for meta-analysis of case-control genetic association studies using logistic regression.** *Stat Appl Genet Mol Biol* 2007, **6**:Article17.
55. Ramasamy A, Mondry A, Holmes CC, Altman DG: **Key issues in conducting a meta-analysis of gene expression microarray datasets.** *PLoS Med* 2008, **5**:e184.
56. Lewis GD, Wei R, Liu E, Yang E, Shi X, Martinovic M, Farrell L, Asnani A, Cyrille M, Ramanathan A, Shaham O, Berriz G, Lowry PA, Palacios I, Tasan M, Roth FP, Min J, Baumgartner C, Keshishian H, Addona T, Mootha VK, Rosenzweig A, Carr SA, Fifer MA, Sabatine MS, Gerszten RE: **Metabolite profiling of blood from individuals undergoing planned myocardial infarction reveals early markers of myocardial injury.** *J Clin Invest* 2008, **118**:3503-3512.
57. Lakkis NM, Nagueh SF, Kleiman NS, Killip D, He ZX, Verani MS, Roberts R, Spencer WH: **Echocardiography-guided ethanol septal reduction for hypertrophic obstructive cardiomyopathy.** *Circulation* 1998, **98**:1750-1755.
58. Witten IH, Frank E: *Data Mining - Practical Machine Learning Tools and Techniques with JAVA Implementations* San Francisco, CA: Morgan Kaufmann Publishers; 2000.
59. Kohavi R: **A study of cross-validation and bootstrap for accuracy estimation and model selection.** *Proceedings of the 17th International Joint Conference on Artificial Intelligence* 1995, **2**:1137-1143.
60. Xiao Y, Hua J, Dougherty ER: **Quantification of the impact of feature selection on the variance of cross-validation error estimation.** *EURASIP J Bioinform Syst Biol* 2007, 16354.
61. Baumgartner C, Böhm C, Baumgartner D, Marini G, Weinberger K, Olgemöller B, Liebl B, Roscher AA: **Supervised machine learning techniques for the classification of metabolic disorders in newborns.** *Bioinformatics* 2004, **20**:2985-2996.
62. Cerami EG, Bader GD, Gross BE, Sander C: **cPath: open source software for collecting, storing, and querying biological pathways.** *BMC Bioinformatics* 2006, **7**:497.
63. Rahman SA, Advani P, Schunk R, Schrader R, Schomburg D: **Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC).** *Bioinformatics* 2005, **21**:1189-1193.
64. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 1999, **27**:29-34.
65. Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P, Goto S, Kanehisa M: **KEGG Atlas mapping for global analysis of metabolic pathways.** *Nucleic Acids Res* 2008, **36(Web Server issue)**:W423-426.
66. Jorda A, Cabo J, Grisolia S: **Changes in the levels of urea cycle enzymes and in metabolites thereof in diabetes.** *Enzyme* 1981, **26**:240-244.
67. Wu G, Meininger CJ: **Impaired arginine metabolism and NO synthesis in coronary endothelial cells of the spontaneously diabetic BB rat.** *Am J Physiol* 1995, **269**:H1312-1318.
68. Ng A, Bursteinas B, Gao Q, Mollison E, Zvelebil M: **Resources for integrative systems biology: from data through databases to networks and dynamic system models.** *Brief Bioinform* 2006, **7**:318-330.
69. Schuster S, von Kamp A, Pachkov M: **Understanding the roadmap of metabolism by pathway analysis.** *Methods Mol Biol* 2007, **358**:199-226.
70. Ekins S, Nikolsky Y, Bugrim A, Kirillov E, Nikolskaya T: **Pathway mapping tools for analysis of high content data.** *Methods Mol Biol* 2007, **356**:319-350.
71. Kussmann M, Raymond F, Affolter M: **OMICS-driven biomarker discovery in nutrition and health.** *J Biotechnol* 2006, **124**:758-787.
72. Street JM, Dear JW: **The application of mass-spectrometry-based protein biomarker discovery to theragnostics.** *Br J Clin Pharmacol* 2010, **69**:367-378.
73. Kohn EC, Azad N, Annunziata C, Dhamoon AS, Whiteley G: **Proteomics as a tool for biomarker discovery.** *Dis Markers* 2007, **23**:411-417.
74. Baumgartner C, Rejtar T, Kullolli M, Akella LM, Karger BL: **SeMoP: A New Computational Strategy for the Unrestricted Search for Modified Peptides Using LC-MS/MS Data.** *J Proteome Res* 2008, **7**:4199-4208.
75. Beger RD: **Cambridge Healthtech Institute's 7th Annual, identifying and validating metabolic markers for drug development and clinical studies.** *Expert Rev Mol Diagn* 2007, **7**:113-115.
76. Hong H, Goodsaid F, Shi L, Tong W: **Molecular biomarkers: a US FDA effort.** *Biomark Med* 2010, **4**:215-225.
77. Ganesalingam J, Bowser R: **The application of biomarkers in clinical trials for motor neuron disease.** *Biomark Med* 2010, **4**:281-297.
78. Rosenson RS: **New technologies personalize diagnostics and therapeutics.** *Curr Atheroscler Rep* 2010, **12**:184-186.
79. Jain KK: **Personalised medicine for cancer: from drug development into clinical practice.** *Expert Opin Pharmaco* 2005, **6**:1463-1476.
80. Netto GJ, Epstein JI: **Theranostic and prognostic biomarkers: genomic applications in urological malignancies.** *Pathology* 2010, **42**:384-94.

doi:10.1186/2043-9113-1-2
Cite this article as: Baumgartner *et al.*: Bioinformatic-driven search for metabolic biomarkers in disease. *Journal of Clinical Bioinformatics* 2011 **1**:2.