Mining Interesting Clinico-Genomic Associations: The HealthObs Approach

George Potamias[1], Lefteris Koumakis[1], Alexandros Kanterakis[1], Vassilis Moustakis[1,3], Dimitrsi Kafetzopoulos[2], and Manolis Tsiknakis[1].
1 Institute of Computer Science (ICS)
2 Institute of Molecular Biology & Biotechnology (IMBB)
Foundation for Research & Technology – Hellas (FORTH)[1]
Vassilika Vouton, P.O. box 1385, 71110 Heraklion, Crete, Greece
3 Technical University of Crete, Department of Production Engineering & Management, 73100 Chania, Crete, Greece
{potamias,koumakis,kantale,moustaki,tsiknaki}@ics.forth.gr,
kafetzo@imbb.forth.gr

Abstract. HealthObs is an integrated (Java-based) environment targeting the seamless integration and intelligent processing of distributed and heterogeneous clinical and genomic data. Via the appropriate customization of standard medical and genomic data-models HealthObs achieves the semantic homogenization of remote clinical and gene-expression records, and their uniform XML-based representation. The system utilizes data-mining techniques (association rules mining) that operate on top of query-specific XML documents. Application of HealthObs on a real world breast-cancer clinico-genomic study demonstrates the utility and efficiency of the approach.

1 Introduction

As the number of electronic clinical records and respective data repositories increases, the seamless integration of the respective data repositories coupled with knowledge discovery operations offer the potential for the automated discovery of valuable clinical knowledge. Furthermore, the completion of the human genome drives us to the post-genomics era. In this environment the newly raised scientific and technological challenges push for trans-disciplinary team science and translational research. As it is noted by J. Grimson: '... *Patient empowerment fuelled by the Internet coupled with post genomics will ultimately lead to a health system which focuses more on promoting wellness rather than on treating illness ... Such a system must be centred on the patient (citizen) and their health status and management. The existence of a longitudinal Electronic Health Care Record is fundamental to bringing about this paradigm shift in the healthcare system*' [11].

The vision is to compact major diseases, such as cancer, on an *individualized* diagnostic, prognostic and treatment manner. This requires not only an understanding of the genetic basis of the disease – acquired, for example, from patient's gene expression profiling studies [4, 13, 21], but also the correlation of this data with knowledge normally processed in the clinical setting. Coupling the knowledge gained from genomics and from clinical practice is of crucial importance and presents a major challenge for on-going and future clinico-genomic trials [15]. Such *evidential* knowledge will enhance health care professionals' decision-making capabilities, in an attempt to meet the raising evidence-based medicine demand.

Recently, and in the context of three research projects – PrognoChio (http://www.ics.forth.gr/~analyti/PrognoChip/isl_site/index.html, [6]), INFOBIOMED (www.infobiomed.net, [10]), and ACGT (http://www.eu-acgt.org, [15]), we have designed and implemented an integated clinico-genomics environment [7]. The environment is enhanced by a *Mediation* infrastructure through which linkage and integration of patients' clinical and genomic (e.g., nicroarray gene-expression) data is achieved [2]. The clinical information systems being utilized are components of an integrated clinical systems' infrastructure built in the region of Crete, Greece [16]. These systems are: (a) *Onco-Surgery* information system – manages information related to patient identification and demographic information, medical history, patient risk factors, family history of malignancy, clinical examinations and findings, results of laboratory exams (mammography, ultrasound, hematological etc.), pre-surgical and post-surgical therapies, as well as therapy effectiveness and follow-up; and (b) *Histo-Pathology* information system – manages information related to patients samples' histopathologic evaluation and TNM staging (tumor size, lymph node involvement, and metastatic spread). Engaged CISs comply with relevant medical information and data models, such as: SNOMED CT® (http://www.snomed.org/), ICD (http://www.cdc.gov/nchs/ icd9.htm), and LOINC® (http://www.regenstrief.org/loinc/). Data and information exchange between the two CIS is based on the HL7 (Health Level 7) messaging standard (http://www.hl7.org). The experimental study presented in this paper (section 4) deploys the two CIS to store and manage patients' clinico-histopathology information and data drawn from an anonymized public domain clinico-genomic study [13]. In this respect we are not confronted with ethical, legal and security issues (even if the whole infrastructure provides high-level security services).

With the help of the *Mediator*, the biomedical investigator can form clinico-genomic queries through the web-based graphical user interface of the Mediator and translates them into an equivalent set of local sub-queries, which are executed directly against the constituent databases (i.e., clinical and genomic/microarray information systems). Then, results are combined for presentation to the user and/or transmission to further analysis.

Access to distributed and heterogeneous data sources and collection of respective data items are not end in itself. What is desirable is the exploitation of data, hence the possibility for exporting useful and comprehensible conclusions. In this context we have designed and developed an integrated clinico-genomic knowledge discovery scenario enabled by a multi-strategy data-mining approach. The scenario is realized by the smooth integration of three data-mining techniques: clustering, association rules mining and feature selection [3,14]. In this scenario, clustering is performed on gene-expression datasets in order to induce indicative clusters of genes, called *Metagenes*. Clustering is performed with *discr_k-means* – a revision of the k-means algorithm that primarily identifies clusters of co-regulated *binary*-valued genes. To overcome the error-prone variance of gene-expression levels, gene-expression values are *discretized* (following a data pre-processing discretization step) into two nominal values: 'low' and 'high'. Putting it in more molecular biology terms, the 'low' and 'high' gene-expression levels correspond to 'DOWN'- and 'UP'-regulated status of the genes, respectively. The discr_k-means algorithm resembles similar approaches presented in [17] and [20]; for a more detailed description of the discr_k-means algorithm, please refer to [3,14]. After convergence of discr_k-means, each metagene is linked with respective (patient) samples and obeys a special characteristic: all of its genes exhibit a 'strong' gene-expression profile for all of its linked samples, i.e., exhibit solely 'high' (or, solely 'low') expression levels. Then, the quest is forwarded towards the identification of associations between metagenes and specific clinico-histopathological profiles.

In this paper we focus on the presentation and utilization of an association rule mining system, called HealthObs [5, 8, 12], to discover interesting and indicative association between patients' clinical and genomic gene-expression profiles, i.e., the metagenes. In the next section we present the architectural specifics of HealthObs. Section 3 presents the basic HealthObs operations and functionality. In section 4 we present results of using HealthObs on a real world breast-cancer study. In the last section we conclude and present hints for further research and development.

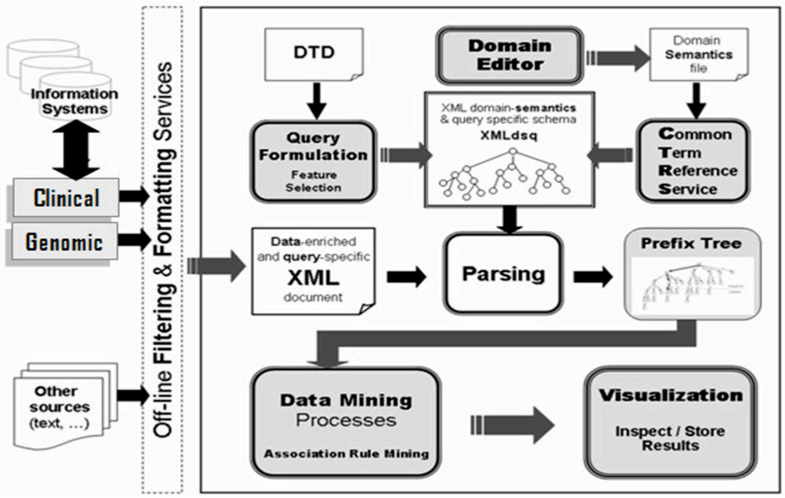

Fig. 1. Architecture and Components of HealthObs

2 Architectural and Operational Set-Up of HealthObs

HealthObs is an integrated system that offers: (i) semantic homogenization of respective distributed and heterogeneous clinical and genomic data sources, (ii) uniform representation of the respective data-items as realized by standard clinical and microarray data-models, and (iii) intelligent processing of XML-formatted documents enabled by the discovery of interesting clinical associations implemented by the customization of association rules mining (ARM) techniques [9,18]. HealthObs offers a population-oriented view on the distributed patients' clinico-genomic information as recorded in respective clinical and genomic/microarray information systems.

Association rules mining [9,18] is among the most advanced and interesting methods for finding interesting patterns and indicative trends in data. The definition of an ARM problem has as follows: Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. An association rule is an implication of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \varnothing$. The rule $X \Rightarrow Y$ has confidence c in the transaction set D if c% of transactions in D that contain X also contains Y. The rule $\mathrm{X} \Rightarrow \mathrm{Y}$ has support s in the transaction set D if s% of transactions in D contains $X \cup \mathrm{Y}$. Given a set of transactions D, the ARM problem is to discover the associations that have support and confidence values higher that the user specified *minimum support*, and *minimum confidence* levels, respectively.

An outline of the reference architecture underlying HealthObs is shown in Figure 1 (above) where, the basic operational modules of the system are also shown. Central to the architecture is a single data-enriched XML file which contains information and data from distributed and heterogeneous clinical and genomic information systems. Accumulation of data and their XML formatting are performed off-line. To this end, the Mediator infrastructure [2] is utilized in order to mediate and query federated clinical and genomic information sources and recall the relevant query-specific data items. For each query, and with the aid of custom made filtering and formatting operations, the respective query-specific XML file is created. HealthObs initiates and base its operations on such data-enriched XML files.

3 Basic Operations and Functionality of HealthObs

Query formulation supports the representation of the inquiry presented to the system. For instance, a user may decide to investigate and assess the confidence of associations between a focused number of clinical and genomic features. For example, between histological/biochemical-tests, such as `'ER'` (Estrogen Receptor) status and prognostic features, such as patients' `'METASTASIS'` status, on one hand, and genomic features, e,g., 'UP'/'DOWN'-regulated status of specific genes, on the other. In Figure 2, the system's feature-selection/focusing interface is shown. The features to be selected correspond to the instance elements being present in the data-enriched XML file to process.

In Figure 2 note the unique characteristic of HealthObs that relate to the specification of the desired form of 'focused' association rules to induce: rules to induce: (a) if the user only check-tick (✓) a feature, then this feature may or may-not be present in the rule, i.e., not-obligatory feature, (b) if the user not only checks a feature but post an 'IF' (✓IF, e.g., `ER`, `MG33g9c19`) or a 'THEN' (✓Then, e.g., `METASTASIS)` tick on it then, the presence of the feature in the association rules is obligatory in the 'IF' or, 'THEN' part of the rule, respectively.

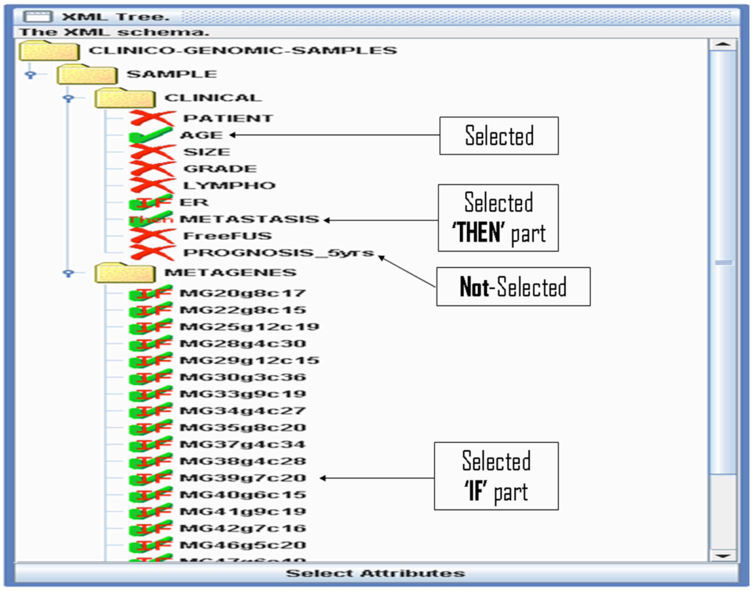

Fig. 2. Feature focusing and query formulation in HealthObs

3.1 Semantic Homogenization

Upon presentation of the inquiry and selection of the respective query features, HealthObs activates the Common Term Reference Service (CTRS) component. CTRS support the placement of the query in context of domain's semantics, e.g., involved medical and genomic nomenclatures and data-models. The SNOMED/CT, ICD and LOINC medical ontologies and nomenclatures (see section 1), as well as the 'Minimum Information About Microarray Experiments' (MIAME, http://www.mged.org/Workgroups/MIAME/miame.html) microarray/gene-expression data-models are utilized. CTRS incorporates (user's) specifications for the semantics of the domain (e.g., valid reference-ranges for lab findings, enabled by the transformation of numerical values to qualitative equivalents or, assignment of continuous gene-expression values to qualitative ones, e.g., into 'high'/'low' expression levels). Activation of the CTRS component results to the creation of an intermediate XML domain semantics and query specific schema ('XMLdsq' tree in Figure 1). XMLdsq is a restriction of the given DTD grammar and helps to: (i) focus the inquiry on the user selected features, and (ii) semantically homogenize the content of the data-enriched XML file. For the editing of the domain-semantics file, and its customization to different domains, we have also developed a special tool, the 'Domain Editor', made operational within the HealthObs environment (for detail see [8]).

3.2 The Prefix-Tree Structure

The recalled query-specific clinical data are kept in the corresponding data-rich XML-documents, and the implemented ARM operations are performed exclusively on top of these documents. The implemented ARM operations rely on the principles of the Apriori algorithm [18]. Adaptation of Apriori-like functionality on top of XML structures is based on a specially devised XML parser enabled by object-oriented search operations. Following RDF/XML techniques, the parser reads/scans the XML document in order to identify composite/atomic observations and homogenize their content (with CCTR service).

In the core of the ARM process is the identification of all frequent itemsets. Usually this is achieved by multiple-scans of the data (in our case, of the XML-document). Thus information-space search operations should be efficient. To enhance on efficiency we rely on a prefix-tree – a special tree-like data structure, that passes the data only once, the *prefix-tree* [1,19]. A prefix-tree structure makes no distinction between internal and leaf nodes. In this structure, nodes do not contain sets, but only information about sets (e.g. counters). Each edge in the tree is labelled with an item, and each node contains the information for the set of items labelling the edges of its path to the root. Prefix-trees store both frequent sets and candidate sets in the same tree.

4 HealthObs in Practice

The utility of HealthObs system was assessed byn applying it on a real-world breast-cancer study [13]. This study profiles the expression of ~24800 genes on 78 breast-cancer patients. The aim was to reveal (potentially) interesting and indicative *individualized* (i.e., target-population oriented) *clinico-genomic profiles*.

Characterization and classification of a disease, and prediction of respective patients' clinical outcome could be performed with reference either to solely clinico-histopathological patient profiles (CHPPs or, clinical phenotypes) -the clinical classification of the disease or, to solely genomic (i.e., microarray gene-expression) patient profiles (GEPPs or, genomic phenotypes) – provided that specific and reliable gene-markers are available. If this presents the decision-making track in the course of a clinico-genomic research trial, the most challenging task is the *knowledge discovery* track which works in a more-or-less inverse way. That is, starting from observable clinico-histopathological disease states the quest targets the identification of respective molecular signatures or, gene-markers able to discriminate between the different disease states.

Based on the central-dogma of molecular biology, CHPPs could be fully 'shaped' and *causally determinable* by respective GEPPs. In this setting, the quest is forwarded towards the following target: "*which clinico-histopathology phenotypes relate and how with which gene-expression phenotypes*?" Such a discovery-driven scenario falls into the *individualized* medicine context -GEPPs may be utilised to 'screen' respective CHPPs, to refine the clinical decision-making process, and finally identify specific patients groups (i.e., cohorts) as more suitable for specific clinical follow-up procedures. The whole endeavour aims to the identification of *abductive* and *inductive* inferential 'rules'.

As we have already mentioned (see section 1) we have designed and developed a *Mediation* infrastructure to recall patients' clinical and gene-expression data from respective clinical and microarray information systems [2]. With the utilization of a clustering operation – realized by the customization of k-means clustering technique on categorical data [3,14], we induce indicative clusters of genes, called *Metagenes* that meet a special characteristic: all of its genes exhibit a 'strong' gene-expression profile for all of its linked samples, i.e., exhibit solely 'high'/'UP'-regulated or, solely 'low'/'DOWN'-regulated expression levels. For example, with 'MG39g7c20=DOWN' we denote a cluster with id=39 ('MG39'), which includes 7 genes ('g7') and covers 20 cases ('c20'), and for all 20 cases all the respective genes exhibit a DOWN value (i.e., are down-regulated or, exhibit 'low' expression levels). A total of 22 such metagenes were induced when the following genes' pre-filtering was applied: p-value ≤ 0.01 and a 2-fold difference in at least 5 samples (similar filtering was applied in the original reference study [13]).

HealthObs was called to induce associations between the induced metagene-values and respective clinical feature-values for the available set of 78 patient samples. The target clinical feature was set to 'METASTASIS' (in the reference breast-cancer study metastasis is considered as 'YES' ('good') or 'NO' ('bad') if it occurred in less than five years or not, respectively. A total of 32 association rules were induced (22 concluding to 'METASTASIS=NO', and 10 to 'METASTASIS=YES') when the following parameters were applied: min-sup = 13% (i.e., at least 10 samples), and min-conf = 60%. By visual inspection of the rules (offered by HealthObs's graphical interface), we were able to identify some interesting associations with potential clinical decision-making value.

For example, one of the rules is:

ER=pos ⇒ METASTASIS=NO

Confidence=**63**%, Support: **63**% = **49** cases

Another, related with the above, rule that was induced is:

ER=pos & MG39g7c20=DOWN ⇒ METASTASIS=NO

Confidence=**100**%, Support: **13**% = **10** cases

ER (Estrogen Receptor) factor possesses a distinct prognostic value for breast-cancer patients. In a 'positive' ('pos') ER state the prognosis is considered as good (i.e., no metastasis). The first rule, above, validates partially this, i.e., it is true in 63% of the cases. With the inclusion of gene-expression information and knowledge the evidence of a good prognosis could be improved. This is what the second rule states and suggests: with the knowledge that all genes in metagene MG39g7c20 are in 'DOWN'-regulated state then, the good prognosis is definite (i.e., 100% confident). Note that the second rule covers just 10 cases, less than the 49 cases (~ 9%) covered by the first rule. This could be considered as an approach to *individualization* of prognosis in the context of a *molecular medicine* environment. There are other association rules induced by HealthObs that cover other cases. For example the rule below engages two metagenes, is also 100% confident, and covers another sub-population of 7 cases:

ER=pos & MG9g6c22=UP & MG38g4c28=UP ⇒ METASTASIS=NO

Confidence=**100**%, Support: ~**9**% = **7** cases

Of course the findings are valid for the specific case-study that refers to a limited set of samples. Further evaluation and validation of results depends on the initiation of specific and targeted clinico-genomic trials that acquire adequate numbers of (statistically stratified) patients' samples. The running times for the discr-k_means and the ARM component of HealthObs were 13 and 3 seconds, respectively; the figures are indicative for the efficiency of the whole approach and of the respective clustering and ARM/HealthObs implementations.

5 Conclusions

We have presented a methodology for mining distributed and heterogeneous clinical and genomic data sources implemented within the context of the HealthObs environment. HealthObs represents an integrated platform with inter-operating software components that offers: (i) semantic homogenization of heterogeneous data resources, (ii) operationalization of ARM operations on-top of XML-formatted clinico-genomic data items, and (iii) flexible query-formulation and mining operations. Preliminary results on applying HealthObs on a real-world clinico-genomic (breast-cancer) study demonstrate the utility of the approach.

Our future research and development plans include: (a) design and development of appropriate human computer interfaces, accompanied with user-profiling capabilities for the personalized delivery of the results, (b) experimentation with other clinico-genomic domains and assessment of the clinical/genomic validity of the results, (c) incorporation of other data-mining operations (e.g., rule discovery), and (c) implementation of 'active-query' capabilities where, discovered clinico-genomic associations from pre-selected records are tested for potential differences and deviations so that, specific alarms could be broadcasted to the interested clinical researcher.

References

1. A. Amir, R. Feldman R., and R. Kashi, A new and versatile method for association generation, *Information Systems* 2, 333-347, (1997).
2. A. Analyti, H. Kondylakis, D. Manakanatas, M. Kalaitzakis, D. Plexousakis, and G. Potamias, Integrating Clinical and Genomic Information through the PrognoChip Mediator, *Lecture Notes in Bioinformatics* **4345**, 250-261, (2006).
3. D. Kanterakis, G. Potamias, Supporting Clinico-Genomic Knowledge Discovery: A Multistrategy Data Mining Process, *Lecture Notes in Computer Science* **3955**, pp. 520-524 (2006).
4. F. Cardoso, Microarray technology and its effect on breast cancer (re)classification and prediction of outcome, *Breast Cancer Res.*, **5**, 303-304, (2003).
5. G. Potamias and V. Moustakis, Knowledge Discovery from Distributed Clinical Data Sources: The Era for Internet-Based Epidemiology, in: 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, (2001), pp. 3638- 3641 vol.4.
6. G. Potamias, A. Analyti, D. Kafetzopoulos, M. Kafousi, T. Margaritis, D. Plexousakis, P. Poirazi, M. Reczko, Y. Tollis, E. Sanidas, E. Stathopoulos, M. Tsiknakis, and S. Vassilaros, Breast Cancer and Biomedical Informatics: The PrognoChip Project, in: Proceedings of the 17th IMACS World Congress Scientific Computation, Applied Mathematics and Simulation, Paper T3-I-68-1066.

7. G. Potamias, D. Kafetzopoulos, and M. Tsiknakis, Integrated Clinico-Genomics Environment: Design and Operational Specification, *Journal for Quality of Life Research* (JQLR), **2**(1), pp. 145-150 (2004).
8. G. Potamias, L. Koumakis, and V. Moustakis, Mining XML Clinical Data: the HealthObs System. *Ingénierie des Systèmes d'Information* **10**(1), pp. 59-79 (2005).
9. H. Mannila, H. Toivonen, and A.I. Verkamo, Efficient algorithms for discovering association rules, in: KDD-94: AAAI Workshop on Knowledge Discovery in Databases, (2001), pp. 181-192, 1994.
10. H.P. Eich, G. de la Calle, C. Diaz, S. Boyer, A.S. Pena, B.G. Loos, P. Ghazal, and I. Bernstein, Practical Approaches to the Development of Biomedical Informatics: the INFOBIOMED Network of Excellence. *Stud Health Technol Inform.*, **116**, pp. 39-44, (2005).
11. J. Grimson, Delivering the electronic healthcare record of the 21st century, *International Journal of Medical Informatics* **64**, pp. 111-127 (2001).
12. L. Koumakis, HealthObs: Health Observatory. An integrated system of data mining and knowledge discovery over distributed and heterogeneous clinical sources, Department of Computer Science, University of Crete MSc thesis (in Greek), 2004.
13. L.J. van 't Veer, et al., Gene expression profiling predicts clinical outcome of breast cancer, *Nature* **415**, pp. 530-536 (2002).
14. M. May, G. Potamias, and S. Rüping, Grid-based Knowledge Discovery in Clinico Genomic Data, *Lecture Notes in Bioinformatics* **4345**, pp. 219-230 (2006).
15. M. Tsiknakis, D. Kafetzopoulos, G. Potamias, A. Analyti, K. Marias, and A. Manganas, Building a European biomedical grid on cancer: the ACGT Integrated Project, *Stud Health Technol Inform.*, **120**, pp. 247-258, (2006).
16. M. Tsiknakis, D. Katehakis and, S. Orphanoudakis, An open, component-based information infrastructure for integrated health information networks, *International Journal of Medical Informatics* **68**(1-3), pp. 3-26 (2002).
17. O.M. San, V. Huynh, and Y. Nakamori, An alternative extension of the k-means algorithm for clustering categorical data, *Int. J. Appl. Math. Comput. Sci.* 14(2), pp. 241-247 (2004).
18. R. Agrawal, H. Manilla, R. Srikant, H. Toivonen, and I. A. Verkamo, Fast discovery of association rules, in: Advances in Knowledge Discovery and Data Mining, (AAAI/MIT Press, 1995), pp. 307-328.
19. R.J. Jr. Bayardo, Efficiently mining long patterns from databases, *SIGMOD Record* **27**(2), pp. 85-93 (1998).
20. S. Gupta, S. Rao, and V. Bhatnagar, K-means Clustering Algorithm for Categorical Attributes, *Lecture Notes in Computer Science* **1676**, pp. 203-208 (1999).
21. S.K. Gruvberger, M.Ringnér, P. Eden, A. Borg, M. Ferno, C. Peterson, and P.S Meltzer, Expression profiling to predict outcome in breast cancer: the influence of sample selection, *Breast Cancer Res.* **5**(1), pp. 23-26 (2003).