Pacific Symposium on Biocomputing 13:580-591(2008)

ENABLING INTEGRATIVE GENOMIC ANALYSIS OF HIGH-IMPACT HUMAN DISEASES THROUGH TEXT MINING

JOEL DUDLEY AND ATUL J. BUTTE

Stanford Medical Informatics, Departments of Medicine and Pediatrics
Stanford University School of Medicine
Stanford, CA 94305-5479, USA

Our limited ability to perform large-scale translational discovery and analysis of disease characterizations from public genomic data repositories remains a major bottleneck in efforts to translate genomics experiments to medicine. Through comprehensive, integrative genomic analysis of all available human disease characterizations we gain crucial insight into the molecular phenomena underlying pathogenesis as well as intra- and inter-disease differentiation. Such knowledge is crucial in the development of improved clinical diagnostics and the identification of molecular targets for novel therapeutics. In this study we build on our previous work to realize the next important step in large-scale translational discovery and analysis, which is to automatically identify those genomic experiments in which a disease state is compared to a normal control state. We present an automated text mining method that employs Natural Language Processing (NLP) techniques to automatically identify disease-related experiments in the NCBI Gene Expression Omnibus (GEO) that include measurements for both disease and normal control states. In this manner, we find that 62% of disease-related experiments contain sample subsets that can be automatically identified as normal controls. Furthermore, we calculate that the identified experiments characterize diseases that contribute to 30% of all human disease-related mortality in the United States. This work demonstrates that we now have the necessary tools and methods to initiate large-scale translational bioinformatics inquiry across the broad spectrum of high-impact human disease.

1. Introduction

1.1. *The Role of Text Mining in Translational Bioinformatics*

As the pace at which genomic data is generated continues to accelerate, propelled by technological advances and declining per-experiment costs, our ability to utilize these data to address long-standing problems in clinical medicine continues to lag behind[1]. It is only through the correction of this disparity that we can overcome one of the major obstacles in translating fundamental discoveries from genomic experiments into the world of medicine for the benefit of public health and society[2, 3].

Owing to its capabilities as a high-bandwidth molecular quantification and diagnostic platform, the RNA expression detection microarray has emerged as a premier tool for characterizing human disease[4-6] and developing novel diagnostics[7, 8]. Fortunately, the data generated by microarray experiments is routinely warehoused in a number of public repositories, providing opportunities to address an unprecedented depth and breadth of data for translational research. These repositories include the NCBI Gene Expression Omnibus (GEO)[9], ArrayExpress at EBI[10], and the Stanford Microarray Database[11]. GEO is the largest among these repositories, offering 157,850 samples (microarrays) from 6,062 experiments as of this writing. Given GEO's exponential growth, it is unlikely to lose this position of predominance for the foreseeable future. In light of these characteristics, it is clear that GEO stands as a model public genomic data repository against which novel bioinformatics methods for large-scale translational discovery may be rigorously designed, evaluated and applied.

We recently described a method for the automated discovery of disease-related experiments within GEO using Medical Subject Heading (MeSH) annotations derived from associated PUBMED identifiers[12]. This represented an important first step in enabling large-scale translational discovery by providing an automated means through which an entire body of publicly available genomic data can be mined comprehensively for human disease characterizations. It also demonstrated the utility of applying text mining methods in translational research, as well as their potential role in realizing a fully automated pipeline for translational bioinformatics discovery and analysis of the human "diseasome". The ultimate goal of such an effort is to comprehensively analyze the whole of disease-related experiments for the purpose of developing novel therapeutics and improved clinical protocols and diagnostics. If such a pipeline were realized, we would be able to ask an entirely new class of questions about the nature of human disease, e.g., "Which genes are significantly differentially expressed across all known autoimmune diseases?"

In order to uncover the many putative links between gene expression and human disease, we must first be able to compare the global gene expression of a disease state with that of a comparable disease-free, or normal control state. Given the sheer volume of experiments available in repositories like GEO, there is a need to develop automated tools and techniques to enable the identification of such states on a large-scale.

1.2. *Objective and Approach*

In this study we seek to develop a robust text mining method to automatically identify disease-related GEO experiments that contain samples for both disease and normal control states. To accomplish this, we utilize an upper-level representation of an experiment in GEO known as a GEO DataSet (GDS), in which samples are organized into biologically informative collections known as subsets. These subsets are defined by GEO curators who group samples from a particular experiment according to the experimental axis under examination (e.g.

disease state or *agent*). Each subset is annotated with a brief, free-text description used to further elucidate the nature of the subset (e.g. *disease-free* or *placebo*). The pertinent attributes and relationships of the GEO GDS are illustrated in Figure 1. The definition of GDS has not kept pace with the addition of experiments (GSE), and as of this writing there are 1,936 GDS defined in GEO representing 32% of the total GSE.

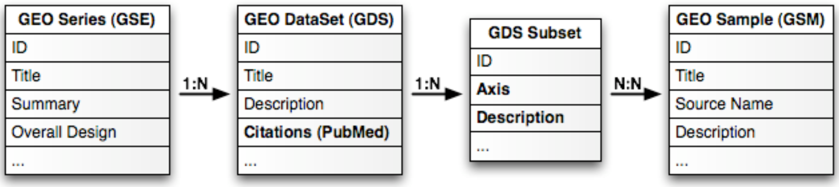
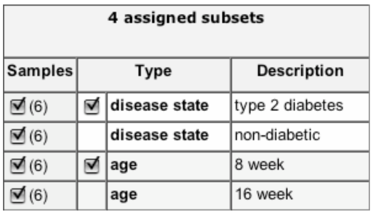


Figure 1. The relationship between GEO Samples, GEO DataSets, GDS Subsets, and GEO Samples is illustrated. The attributes utilized by the proposed method are shown in bold. The label over the arrows indicates the cardinality of the relationship.

We propose that these subset text attributes can be evaluated to determine if a particular subset is representative of either a disease state or a normal control. While the vocabulary used to denote the experimental axes for a subset is principally controlled, currently comprised of twenty-four distinct terms, their utilization within a GDS and their application to sample collections is left completely to curator discretion. Furthermore we find that the content of the descriptions associated with each subset is free-text, constrained by no declared or discernable convention or controlled vocabulary. An example of these subset annotations is shown in figure 2. It is not possible to elucidate control subsets from the experimental axis annotation alone, as these annotations aim to classify the experimental variable being measured (e.g. *cell type* or *development stage*), rather than to describe the context of measurement instances. Thus we are faced with the difficult problem of elucidating the context of each subset based on the free-text descriptions associated with each subset.

Fortunately, simple frequency analysis reveals that a small number of terms commonly used to describe a normal control state are found in the associated subset descriptions for disease-related GDS in high frequency. As shown in Figure 3, the distribution of subset description phrases follows a Zipf-like distribution, with the common used control terms *control*, *normal*, and *wild type* representing the most frequently used phrases by-experiment and by-samples across all disease-related GDS subset descriptions. Thus, it is reasonable to suggest that the problem of large-scale



4 assigned subsets			
Samples		**Type**	**Description**
☑ (6)	☑	**disease state**	type 2 diabetes
☑ (6)		**disease state**	non-diabetic
☑ (6)	☑	**age**	8 week
☑ (6)		**age**	16 week

Figure 2. Example GDS subset designations for GDS402 taken from the GEO website.

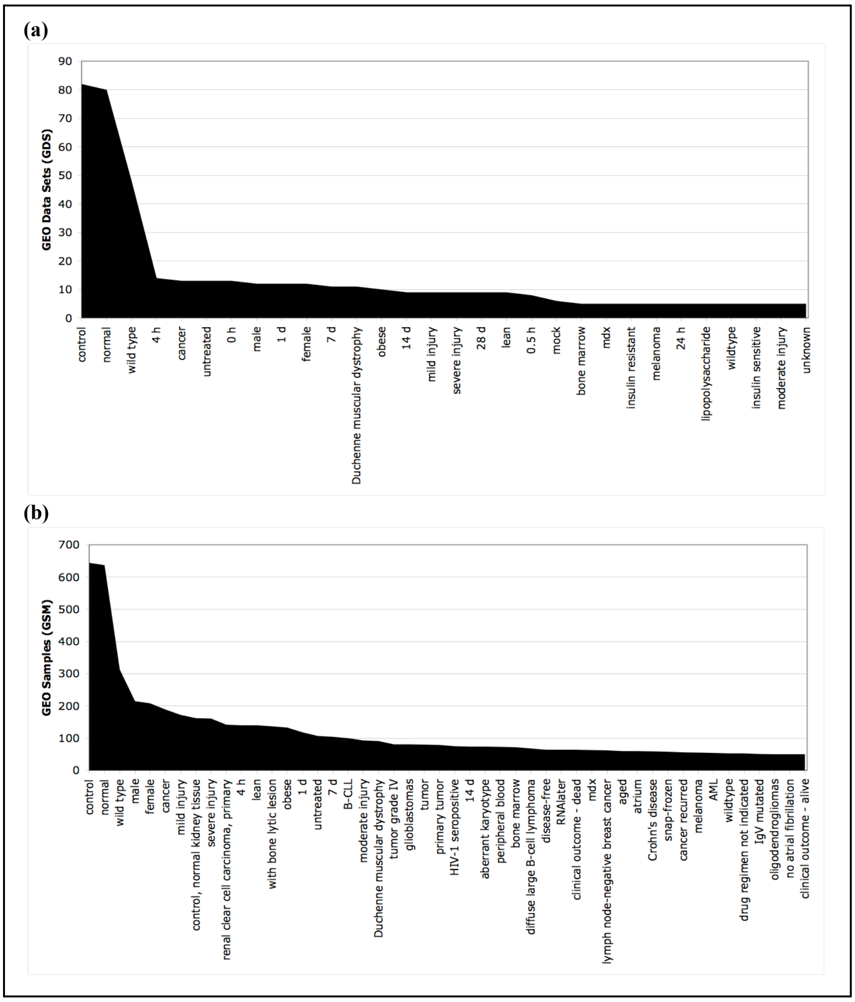


Figure 3. Distribution of GDS subset annotation phrases for all disease-related GDS. The distributions are filtered to terms annotating > 5 GDS and > 50 GSM for display purposes. The distribution shows that the **(a)** majority of disease-related GDS contain subsets annotated with a small set of common control phrases, **(b)** representing a major proportion of samples.

normal control detection within GEO is tractable by the fact that a simple pattern matching approach using three common normal control phrases will identify controls in a majority of experiments representing a majority of samples. However this technique alone is insufficient as many control subsets for unique disease characterizations are found in the “long-tail” of the frequency distributions. In some cases common control terms are found within the subset description, but they do not represent a disease-free state (e.g. *skin cancer control*). In other cases a control subset is annotated using a disease negation scheme (e.g. *diabetes-free*). In such cases the application of a simple pattern matching technique would result in either a false positive or a false negative respectively.

To manage such complex cases we make use of the Unified Medical Language System (UMLS) Metathesaurus[13] to identify terms representing a human disease. With disease terms identified, it is possible to infer control subsets that are implied rather than explicit, for example the negation of a disease term implies a normal control, and avoids incorrectly identifying control subsets that are annotated in a contradictory manner (e.g. *normal skin cancer*).

1.3. *Evaluating the Impact of Translational Text Mining*

The impact of any exercise in *translational* text mining cannot be fully assessed without a clear quantitative evaluation of the clinical impact and overall benefit to human health. For it is through such clinical imperatives that translational bioinformatics is distinguished. It is tempting to measure the clinical impact of the proposed method by way of the total number of unique diseases for which a disease vs. normal control state was identified, however not every human disease carries the same clinical impact.

Therefore in addition to traditional performance measures, we propose to measure translational impact along the axis of human disease-related mortality. In this context, impact is based on the coverage of disease characterizations over the total disease-related human mortality, quantified by the number of deaths for which a disease is responsible. This impact measure is intuitive, because it is reasonable to assume that the diseases causing the greatest number of deaths are the diseases that have the greatest impact on clinical practice.

2. Methods

2.1. *Identifying Disease-Related Experiments*

Similar to our previously described method[12], the disease-related experiments were identified using a MeSH-based mapping approach. We used a February 15th, 2007 snapshot of the Gene Expression Omnibus (GEO)[9] which was parsed into a normalized structure and stored in a relational database.

For the 1,231 GEO DataSets (GDS) experiments associated with a PUBMED identifier, we downloaded the corresponding MEDLINE record and extracted the MeSH using the BioRuby toolkit (http://www.bioruby.org). The extracted MeSH terms were stored in a relational database along with the associated GDS identifier, resulting in 20,654 distinct mappings. These mappings were joined with the UMLS (2007AA release) Concept Names and

Sources (MRCONSO) and Semantic Types (MRSTY) tables to identify GDS associated with MeSH terms having any of the semantic types among *Injury or Poisoning* (T037), *Pathologic Function* (T046), *Disease or Syndrome* (T047), *Mental or Behavioral Dysfunction* (T048), *Experimental Model of Disease* (T050), or *Neoplastic Process* (T191) as disease-related GDS.

2.2. *Control Subset Detection*

For each disease-related GDS we obtained data for the associated subsets using the aforementioned relational snapshot of GEO. The subsets of each disease-related GDS were enumerated and their descriptions evaluated to elucidate control subsets. As previously mentioned, a sizeable proportion of disease-related GDS (41%) have subsets annotated with the common control terms *control*, *normal* and *wild type* or some slight variation thereof. These common control terms were assembled into a set, and any subset with a description annotation comprised of a single term from this set was identified as a normal control subset. Subset descriptions were also transformed into stemmed, word case, spacing and hyphenation variants using porter stemming and regular expressions to detect control term variants (e.g. *controlled* becomes *control*, *wild-type* becomes *wild type*), which represented an additional 14% of disease-related GDS. If any such variant of a common control term was matched in a subset annotation, then the subset was identified as a normal control.

Curiously, a small proportion of disease-related GDS (3%) did not have any subsets defined. It is not clear as to why this was the case. It could be that these GDS are incompletely curated, and subset definitions will be applied in later releases of GEO. Consequently these GDS were removed from consideration.

Subset descriptions not containing common control terms were evaluated using more sophisticated techniques to account for negation and lexical variation.

2.3. *Handling Negation*

We find that GDS subsets are frequently annotated using a negation scheme in which a subset representative of a disease state will be annotated with a UMLS disease concept and the control will be expressed as the negation of that disease concept (e.g. *diabetic* vs. *non-diabetic*). Therefore the identification of control subsets was expanded to include subsets that are annotated using this disease-negation pattern.

The detection of negations in natural languages is non-trivial[14], however there are several properties of GDS subset labels that increase the tractability of the problem. GDS subset descriptions are typically terse (average of 10.7

characters per description), and therefore the word distance between the negation signal and the concept is negligible. This aids negation detection by minimizing a common source of error in tokenizing negation parsers[15], and eliminates the need to engage more complex Natural Language Processing (NLP) approaches, such as parse tree based negation classification[16], to link negation symbols to disjoint disease concepts. Given these properties we chose to identify negation-based control subsets using a modified version of the NegEx algorithm[17]. The NegEx algorithm is a regular-expression based algorithm for the detection of the explicit negation of terms indexed by UMLS. NegEx has been shown to have 78% sensitivity and 84.5 % positive predictive value when detecting negations in medical discharge summaries[17]. It is expected that NegEx will perform better in the detection of negation-based control subsets, as complex syntactic structures, which are not present in terse subset labels, were a major source of error in detecting negations in verbose discharge summaries. Additionally, we constrained the NegEx algorithm to detect negation for UMLS-mapped terms exhibiting any of the five aforementioned disease-related semantic types rather than the broader fourteen semantic type categories used by the unmodified algorithm.

We found that in some cases, a subset description will exhibit the negation of a valid disease term, but does in fact lead to a false positive since the negation is also a valid disease state (i.e. *non-Hodgkins Lymphoma)*. To correct for this case, we first query UMLS to ensure that description does not represent a disease state.

2.4. *Handling Lexical Variations*

In some cases the description for a control subset was expressed in a manner that is lexically inconsistent with the terms used to describe the disease state. For example, GDS887 defines the following subset labels for the *disease state* axis: *type 1 diabetes*, *type 2 diabetes*, and *non-diabetic*. In order to automatically link the subset labeled *non-diabetic* as the negated control of the subset labeled *type 1 diabetes*, we must derive that these lexically incompatible labels are in fact semantically related.

Lexical variations were automatically reconciled using the Normalized Word Index table (MRXNW_ENG) in UMLS. The Normalized Word Index contains tokenized, uninflected forms of UMLS terms, derived either algorithmically or through the SPECIALIST lexicon. Using this table we find that the terms *type 1 diabetes* and *diabetic* share a common association with at least one Concept Unique Identifier (CUI) (C0011854). Therefore we can infer that the subset labeled *non-diabetic* is in fact a valid negated control of the subset labeled *type 1 diabetes*.

2.5. *Performance Evaluation*

To evaluate performance we used an expert human reviewer as a "gold standard", and divided control subsets into two distinct groups. The first group, Group A, represents control subsets identified using common control terms, and the second group, Group B, represents control subsets that did not contain common control terms, and therefore were evaluated using the negation-based approach. We randomly sampled positively and negatively identified control subsets from both groups and calculated True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) counts after each subset from the random samples was evaluated by the expert human evaluator, who positively or negatively identified control subsets. From these counts we calculated sensitivity = TP/TP+FN, specificity = TN/TN+FP, Positive Predictive Value (PPV) = TP/TP+FP, Negative Predictive Value (NPV) = TN/TN+FN, and F1 score = 2·(PPV·(TP/TP+FN))/PPV+(TP/TP+FN). These values were also computed across both groups to provide an overall evaluation of performance for the proposed method.

2.6. *Evaluating Clinical Impact from Mortality Data*

U.S. mortality data from 1999 to 2004 were obtained from the Centers for Disease Control and Prevention (CDC) using the Wide-ranging Online Data for Epidemiologic Research (WONDER) system (http://wonder.cdc.gov). Causes of Death were specified using International Classification of Disease (ICD) codes (10th edition). These codes were mapped to their corresponding MeSH using the MRCONSO table in UMLS. We acknowledge that many ICD10 codes have no direct mapping to MeSH in UMLS, with only ~15% of ICD10 directly linked to MeSH terms. Computational translation between UMLS source vocabularies is an active area of research, with several promising approaches emerging[19, 20]. However it is beyond the scope of this paper to participate in this budding area of research. Therefore we only map ICD10 codes to MeSH terms when they are directly related under the same concept identifier (CUI) in UMLS to provide a minimum estimate of impact. The number of deaths mapped to disease-related GDS in this manner was used to calculate the total disease-related mortality impact.

3. Results

In mapping GDS to MeSH terms, we find that 1,231 (78%) of the 1,588 GDS in our GEO database snapshot were associated with a PUBMED identifier. From the resulting 20,654 GDS to MeSH mappings, we find that 513 GDS are associated with MeSH terms having at least one of the six semantic types considered to be disease-related (T037, T046, T047, T048, T050 or T191).

In detecting common normal control phrases in subset annotations, we find that control subsets are identified in 56% of disease-related GDS. Using the negation and lexical variation compensation techniques, we are able to identify control subsets in an additional 33 GDS, resulting in the automated identification of control subsets in a total of 62% of disease-related GDS. This results in a set of 13,840 samples spanning 141 unique disease-related concepts.

We manually inspected the 38% of disease-related GDS for which normal control subsets could not be identified, and found that they fell into a handful of general categories. A number of GDS experiments were designed to characterize or differentiate among disease subtypes (e.g. expression profiling across different cancer cell lines), and therefore contain no true control subsets. Others annotated subsets using proprietary identifiers for cell lines and animal strains. The latter accounts for a major source of sensitivity dampening in evaluating control subsets. Detailed performance metrics are illustrated in Table 1.

Table 1. Performance Evaluation of Control Detection.

	Sensitivity	Specificity	PPV	NPV	F1
Group A (common control terms) (n=100)	0.979	1.000	1.000	0.980	0.989
Group B (negation-based controls) (n=100)	0.428	0.983	0.937	0.750	0.588
Combined (Group A+Group B) (n=200)	0.750	0.911	0.984	0.840	0.851

We were successful in mapping 2,019 ICD10 codes to MeSH terms, covering 18% of the ICD10 codes represented in the mortality data, and 42% of the total mortality. Using MeSH headings, we were able to map 42% the disease-related GDS with normal controls to ICD10 codes. These ICD10 codes mapped to 77 unique ICD10 codes in the mortality data representing 4,219,703 combined deaths over 5 years, or 30% of the total human disease-related mortality in the United States in the same period. Note that this is a minimum estimation given the limited mapping between ICD10 and MeSH in UMLS.

4. Discussion

Given the current pace of growth experienced by international genomic data repositories, it may be only six years before researchers have access to more than a million microarray samples. Yet, even with less than half that amount available today, it has not been possible to link any significant portion of these genomic measurements to the broad molecular characteristics underlying the broad spectrum of human disease. Here we describe a method that enables the creation of such links, and lays the groundwork for the development of a robust translational bioinformatics pipeline that can be applied to both current and forthcoming volumes of public genomic data.

Through this method we find that we can automatically identify normal control subsets in GDS representing 141 unique disease states and conditions. While cancers make up a significant proportion of the associated diseases, afflictions such as Alzheimer's disease, heart disease, diabetes and other diseases having a major impact on human mortality are also represented.

The techniques developed for the identification of negated control subsets and the reconciliation of lexical variations will become increasingly important as GEO continues its exponential growth. Even if the percentage of disease-related GDS experiments containing non-obvious control subset designations remains the same (17%) or even slightly less, these techniques could enable the automated translational analysis of thousands of disease-related microarray samples.

We have now proven that it is not only possible, but also completely tractable to apply these methods to our current public data collections in an attempt to characterize the broad spectrum of high-impact human disease. Despite the fact that we were only able to identify control subsets in 20% of the total GDS found in GEO, and ultimately only 6% of the total experiments contained within GEO, we were able to associate these GDS experiments with diseases contributing to 30% of the total human mortality in the United States.

The next critical step is to develop a means by which those experiments without associated PUBMED identifiers can be automatically evaluated to identify additional disease-related experiments. In addition, these techniques must be further generalized so that they can be applied to additional public repositories containing data from microarrays and other genome-scale measures.

We acknowledge that while this study provides a successful proof of concept and demonstration of utility, it does not provide a finished product. Therefore the method will not be made available as a public resource, however it will enable the creation of more biologically relevant downstream resources.

Conclusion

Using GEO as a model public data repository, we have developed text mining techniques that enable completely new types and scales of translational research.

As these techniques are applied to new and expanding public data repositories, by means of translational bioinformatics, we will be given the opportunity to discover the fundamental molecular principals and dynamics that underlie the whole of high-impact human disease. It is from this vantage that we will begin to realize the novel diagnostics and therapeutics long-promised in this post-genomic era.

Acknowledgments

The authors would like to thank Alex Morgan for providing critical feedback on an early draft of the manuscript, and Alex Skrenchuck for HPC support. The work was supported by grants from the Lucile Packard Foundation for Children's Health, National Library of Medicine (K22 LM008261), National Institute of General Medical Sciences (R01 GM079719), National Human Genome Research Institute (P50 HG003389), Howard Hughes Medical Institute, and the Pharmaceutical Research and Manufacturers of America Foundation.

References

1. C. A. Ball, G. Sherlock and A. Brazma, Funding high-throughput data sharing. *Nature biotechnology* **22**, 1179-83 (2004)
2. E. A. Zerhouni, Translational and clinical science--time for a new vision. *N Engl J Med* **353**, 1621-3 (2005)
3. M. Chee, R. Yang, E. Hubbell, A. Berno, X. Huang, D. Stern, J. Winkler, D. Lockhart, M. Morris and S. Fodor, Accessing genetic information with high-density DNA arrays. *Science* **274**, 610-4 (1996)
4. S. Calvo, M. Jain, X. Xie, S. A. Sheth, B. Chang, O. A. Goldberger, A. Spinazzola, M. Zeviani, S. A. Carr and V. K. Mootha, Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat Genet* **38**, 576-82 (2006)
5. K. Mirnics and J. Pevsner, Progress in the use of microarray technology to study the neurobiology of disease. *Nat Neurosci* **7**, 434-9 (2004)
6. E. E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. Guhathakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, P. Y. Lum, A. Leonardson, R. Thieringer, J. M. Metzger, L. Yang, J. Castle, H. Zhu, S. F. Kash, T. A. Drake, A. Sachs and A. J. Lusis, An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* **37**, 710-7 (2005)
7. A. M. Glas, A. Floore, L. J. Delahaye, A. T. Witteveen, R. C. Pover, N. Bakx, J. S. Lahti-Domenici, T. J. Bruinsma, M. O. Warmoes, R. Bernards, L. F. Wessels and L. J. Van't Veer, Converting a breast cancer microarray

signature into a high-throughput diagnostic test. *BMC Genomics* **7**, 278 (2006)

8. G. J. Gordon, R. V. Jensen, L. L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker and R. Bueno, Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res* **62**, 4963-7 (2002)
9. T. Barrett, T. O. Suzek, D. B. Troup, S. E. Wilhite, W. C. Ngau, P. Ledoux, D. Rudnev, A. E. Lash, W. Fujibuchi and R. Edgar, NCBI GEO: mining millions of expression profiles--database and tools. *Nucleic Acids Res* **33**, D562-6 (2005)
10. A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, A. Oezcimen, P. Rocca-Serra and S. A. Sansone, ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* **31**, 68-71 (2003)
11. G. Sherlock, T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J. C. Matese, S. S. Dwight, M. Kaloper, S. Weng, H. Jin, C. A. Ball, M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein and J. M. Cherry, The Stanford Microarray Database. *Nucleic Acids Res* **29**, 152-5 (2001)
12. A. J. Butte and R. Chen, Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. *AMIA Annu Symp Proc* 106-10 (2006)
13. D. A. Lindberg, B. L. Humphreys and A. T. McCray, The Unified Medical Language System. *Methods of information in medicine* **32**, 281-91 (1993)
14. R. M. April and M. E. Caroline, The ambiguity of negation in natural language queries to information retrieval systems. *J. Am. Soc. Inf. Sci.* **49**, 686-692 (1998)
15. P. G. Mutalik, A. Deshpande and P. M. Nadkarni, Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. *J Am Med Inform Assoc* **8**, 598-609 (2001)
16. Y. Huang and H. J. Lowe, A novel hybrid approach to automated negation detection in clinical radiology reports. *J Am Med Inform Assoc* **14**, 304-11 (2007)
17. W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper and B. G. Buchanan, A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics* **34**, 301-10 (2001)
18. O. Bodenreider, S. J. Nelson, W. T. Hole and H. F. Chang, Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. *AMIA Annu Symp Proc* 815-9 (1998)
19. C. Patel and J. Cimino, Mining Cross-Terminology Links in the UMLS. *AMIA Annu Symp Proc* 624-8 (2006)