Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

Experiences in Designing a Distributed Service-Oriented Platform for In-Silico Biomarker Discovery and Validation

Ioan Lucian Muntean*, Bogdan Rotaru*, Ramona Suharoschi† and Cristina Iuga‡
*Computer Science Department
Technical Univesity of Cluj-Napoca, Romania
†Food Science and Technology Department
University of Agricultural Sciences and Veterinary Medicine, Cluj-Napoca, Romania
‡Drug Analysis Department
University of Medicine and Pharmacy, Cluj-Napoca, Romania

***Abstract*—In order to design efficient biomarkers for personalized medicine, one needs correlated insights from different multi-omics perspectives, such as proteomics, genomics, transcriptomics etc. In this paper, we address the challenge of integrating multi-omics data repositories for in-silico biomarker discovery and validation. For that, we have designed a software platform based on the genomic and proteomic data overlay.**

Currently, this platform provides an integrated way of querying bioinformatics major data sources such as GenBank, UniProt, and ArrayExpress. Our approach integrates the client-side API of these resources with the support of web services and cloud (Google App Engine), as enabling technologies. As such, the extension of the platform with further omics data sources becomes straightforward. The platform was evaluated on the Google Cloud. The first results are promising, showing that the queries can easily be formulated and executed against these public community resources, the overhead introduced by our services being rather small.

I. Motivation

Many modern techniques for the early detection and diagnosis of diseases such as the cancer are based on biomarkers. A biomarker is a molecule used as an indicator for biological states. Further uses of biomarkers are in the personalized medicine, where the prevention and therapeutic interventions are guided by the response of the patient to specific biomarkers.

The effectiveness of biomarker discovery and validation (BMVD) can be greatly improved by analyzing and correlating experimental data and findings from multiple 'omics' areas such as proteomics (currently the most widely used for BMVD), genomics, transcriptomics etc. In this way, the specificity, sensitivity, and reliability of BMs could be significantly increased in-silico, before proceeding with the clinical trials. The major challenges of such a multi-omics data overlay approach are:

- processing large amount of experimental data
- biological understanding of the different 'omics' data;
- definition of an appropriate experimental deisgn in order to accomplish statistical requirements (significance, sensitivity, accuracy, robustness) for BMDV into clinical trials;
- increased complexity of the integration due to significant differences in experimental data/data storage formats, data collection and access policies, scientific review of collected data, interfaces for querying data etc.

In this paper, we address the challenge of integrating different 'omics' data repositories for in-silico BMVD. For tackling the integration challenge, we propose the use of the cloud and web service technologies. We facilitate the integration of heterogeneous open web resources such as NCBI (GenBank [1]), EMBL-EBI (ArrayExpress [2]), and UniProt [3]. Our development and integration efforts are bundled into the emerging novel software platform BioGenProtOMICS. Our approach integrates the client-side API of these resources with the support of web services and cloud (Google App Engine), as enabling technologies. As such, the extension of the platform with further 'omics' data sources becomes straightforward. The platform was evaluated on the Google Cloud. The first results are promising, showing that the queries can easily be formulated and executed against these public community resources, the overhead introduced by the our services being rather small.

The remaining of this paper is structured as follows: Section II positions our work among other approaches from bioinformatics. The *in-silico* biomarker discovery and validation process we follow in our work is introduced in Sec. III. In Sec. IV are briefly introduced the main integration challenges and our choice of technologies. Section V presents software platform we propose for searching for biomarkers. The evaluation results of our work are the focus of Sec. VI. Section VII concludes the paper and gives an outlook on possible future extensions.

II. Related Work

Scientists have to interrogate many of these databases or web sites for each gene in their candidate gene list; they must learn and remember how to navigate several web sites, each

978-1-4673-4358-9/12/$31.00 ©2012 IEEE

of which accepts dissimilar sets of gene identifiers (Entrez Gene, Ensembl, Refseq, UniGene, and other), thus doing the navigation difficult and time-consuming. Despite that these resources are highly informative separately, the databases with open content have more capabilities, if provided in an integrated, centralized context indexed in a robust manner.

Even though at this time there is no single resource/tool that entirely describes all the information that a researcher might want to find out about a specific gene, a few integrative approaches towards this goal have been developed and include BioGPS [4], Ondex [5], NIF [6], GeneCards [7], and several databases for cataloging web resources, such as PathGuide [8]. The integrated and exhaustive use of biological information remains unsolved due to the large number of available databases and to their fragmentation. ID/naming problems and differences in biological data present additional hurdle towards wide and complex biological data integration.

Other approaches with focus on integration and relevant to our work are GabiPD [9], neXTProt [10], GEMS [11], TCGA [12], and caBIG[13]. The first one is focused only on the integration of databases produced within the GABI projects and with NCBI UniGene, having applicability primarily to plants. It is based on Java and Perl, the search uses the concept of GreenCards for aggregating the search results. neXTProt is an integrative search tool residing around the UniProt resources, with focus on the integration and cross-databases search of protein data. It lacks though of support for many genomics and microarray databases. Perhaps the closest one to our approach is the GEMS system. As BioGenProtOMICS, it discovers biomarkers from microarray gene expression data. Moreover, it is focused on the automated cancer diagnosis. The range of data sources available with GEMS is much smaller than the targeted one for BioGenProtOMICS, the two-layer architecture of GEMS being potentially less scalable with respect to integration and to the high-throughput discovery of biomarkers. TCGA and caBIG are large infrastructure projects for cancer research using bioinformatics. Our software platform follows their recommendations (e.g. RDF database format and SPARQL [12] querying language), to enable the inter-operation with their resources and applications. We envisage a potential use of our BioGenProtOMICS software as an application on caBIG.

III. *In-Silico* Biomarker Discovery and Validation

In conformity with Biomarkers Definitions Working Group a 'biological marker (biomarker-BM) is a characteristic that is objectively measured and evaluated as an indicator of normal biological process, pathogenic processes or pharmacological responses to a therapeutic intervention [14]. Biological markers could be divided into diagnostic, prognostic, predictive and therapeutic response markers, and are represented by different genes expression, altered or mutated genes, miRNA, transcription factors, RNA, proteins, lipids, carbohydrates, small metabolites molecules and modified expression of those molecules that can be correlated with a biological aspects or a clinical outcome [15]. The biomarker discovery through mining a wide range of repositories is inherently and has a high degree of parallel and distributed processing. The biomarker discovery and validation flow could be based on experimental data and laboratory process or a preliminary computational (in-silico) process of potential candidates' biomarker discovery and validation (BMDV). The technology employed in BMDV process could be exhaustive, based on high-throughput technology or classical, robust molecular technology, with a high variety of data types. Our study is based on a computational (in-silico) discovery methodology followed by in-silico validation. Linking expressional data gained through genomic and proteomic study to biological pathways of interest underlying with a comprehensive understanding of system biology.

The BMDV approach in this study is based on the following steps: 1) medical problem; 1a) gene expression datasets; 2)generate a set of candidate genes and/or proteins; 3) identify differential expression genes and/or proteins; 3a) ranking of genes and/or proteins (using statistical tools for filtering significantly differentially expressed data; 3b) removal of non-significant data; 4)data overlay (overlap- potential biomarker candidates); 5) statistics filtering; 6)List of meaningful genes and/or proteins; 7) biological knowledge extraction (link discovery); 8) filtering service (based on PubMed literature data); 9) in-silico dry BM validation by cross-validation method; 9a) split data (training set and testing set). A more detailed description of the BMDV process we follow can be found in [16].

IV. The Integration Realm

A. Omics Data Sources

GenBank is built and distributed by the National Center for Biotechnology Information (NCBI). GenBank data is available to scientists through FTP or through a wide range of retrieval and analysis web services. GenBank records, consisting of both sequences and annotations, uniquely identified by the accession number. This accession number is employed in the matching of GenBank records with data from other providers. The programmatic access to GenBank is done by means of web service calls, Entrez suite etc.

Microarray data is shared in the end-consumer communities through the ArrayExpress repository established at the European Bioinformatics Institute in 2002. This is a public repository for microarray data that supports the MIAME requirements and stores well-annotated raw and normalized data. The data can be retrieved by accession number or queried by various parameters such as species, author and array platform. The ArrayExpress query interface provides the ability to query for Experiments, Protocols and Array designs by their various attributes, such as species, authors or array platforms. There are a few ways to access the biological data with ArrayExpress: REST-style queries, JSON web service format etc.

The Universal Protein Resource (UniProt) provides a comprehensive central resource of protein sequences and functional annotation. UniProt provides the scientific community with a single, centralized, authoritative resource for protein

sequences and functional information. UniProt queries can be performed using REST services, UniProtJAPI (Java API) etc.

B. Choice of Technology

Google App Engine [17] (GAE) is a cloud-enabled engine for applications. App Engine applications are built with the GAE software development kit, are easy to maintain, and, when deployed on the Google Cloud, they scale as the traffic and data storage needs grow. Whereas GAE applications are similarly organized and run as the web applications, there are also limitations of the former ones such as the lack of multi-threading support. The enabling technology chosen for integration is web services. These are commonly used for shaping and reducing the interface of software modules or even of entire applications that run remotely. All the 'omics' data sources introduced above provide REST services for querying their data. As such, we take this technology as our primer choice.

From the integration perspective, the web services offer a powerful way of programmatically accessing remote software. The drawback integration solutions based on web services have is their intrusive character, though.

V. The BioGenProtOMICS Platform

In order to support the *in-silico* discovery and validation of biomarkers, we have designed a software platform that supports the BMDV process introduced in Sec. III and handles the integration of 'omics' data sources (see Sec. IV).

A. Architecture Overview

BioGenProtOMICS has a multi-layer distributed architecture, as depicted in Fig. 1. The responsibilities of its layers are:

- Presentation focuses on taking the user query and on organizing the results retrieved from the different data sources;
- Service Layer represents the interface of the Processing Layer responsible for executing the main operations of the BMDV (matching results from different sources, filtering data, or enhanced searching using Gene Ontology (GO));
- Processing Layer composes queries in a generic format, common to the integrated data sources, mines retrieved data etc.;
- Integration Layer provides a common API for interrogating the external data sources, implements client side querying functionality corresponding to the public services of the data sources;
- Remote Layer handles the placement of computing tasks created by the platform. Target computing environments are grids and clouds.

So far, a first prototypical implementation of the platform was implemented. It consists of a user interface for specifying user queries, a query interface common to the external data sources, implementations for ArrayExpress, GenBank, and UniProt being developed as web services, and basic statistics

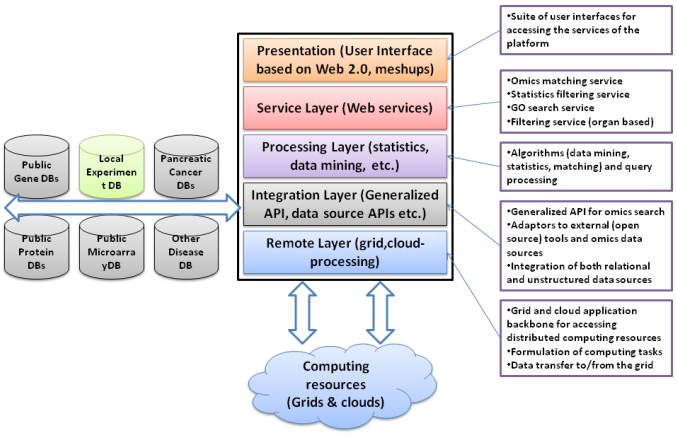

Fig. 1. The multi-layer architecture of the BioGenProtOMICS platform.

operation on the retrieved data. The processing of the query results takes place for now within the web services, thus, in the environment where these are deployed (web application server, Google Cloud). As such, the backbone of the platform available, making possible incremental additions of new search capabilities and further data sources. An ontology-based search engine is under development at the moment of writing. It uses RDF representations for enhancing the simultaneous search in multiple 'omics' data sources.

B. Integration of 'Omics' Data Sources

The focus of the initial work of the platform was on the prototypical development of the layers and on coping with the technical challenges of the integration.

The client-side capabilities of ArrayExpress, UniProt, and GenBank public services are encapsulated in web services implementing the same query interface. The results returned by our services have a common structure, allowing for common/similar handling of the data, regardless its source. This is done in two steps: First, results from the public services are bound to Java classes using JAXB. Second, the resulting data types are transformed in the common type using the Data Transfer Object (DTO) pattern. As such, the DTO objects get decoupled from their initial types, allowing for their generic processing.

Figure 2 shows two operations available in the UniProt Service of our platform. The results of the data binding is represented on the right side of the diagram. The ArrayExpress and GenBank services have a similar organization to the one in Fig. 2. This makes possible the retrieval of data sets corresponding to a single or to a list of identifiers, and, in subsequent calls, of refined biological data corresponding to the additional filters specified in the search query.

The generalized programming interface for executing the queries against the 'omics' data sources is implemented in the External Data Retrieval Service (part of the Integration Layer), which further delegates the call with appropriate parameters to one or more specialized services (see Fig. 3). Similarly, the formulation of processing tasks, their submission, and

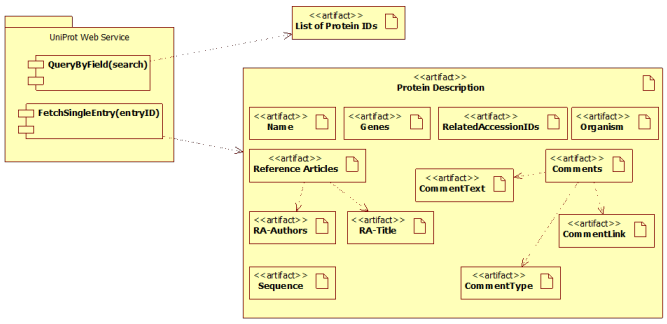

Fig. 2. Basic operations and resulting bounded protein data structures of the UniProt Service.

the retrieval of data from the remote computing environment are available to the platform through the Remote Processing Service. Details specific to the respective grid (here Globus Toolkit-based) or cloud (here GAE) are encapsulated separately. In order to integrate additional resources, one has to

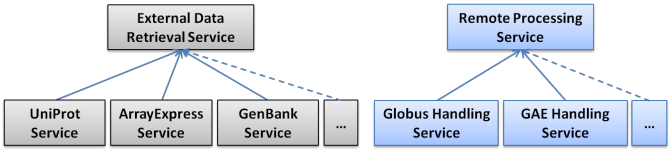

Fig. 3. Internal organization of the Integration Layer.

provide specific adaptations of these two generic services. One major advantage of our approach to the integration is the limitation of the propagation of the changes introduced in the platform by new adaptations.

In its current status, the BioGenProtOmics platform allows the scientists to automatically retrieve data of interest correlated from multiple sources, to filter it, and to further concentrate the exploration of bioinformatics data sources. Although these functionalities significantly simplify the search for biomarkers, a full BMDV process is not possible with the platform yet.

VI. Results

A. Experimental Setup and Queries

The results presented in this section have been obtained on the evaluation environment depicted in Fig. 4. Two important components of the retrieval of query results have been evaluated: the extraction of result data from the external 'omics' sources using the platform services and the basic processing of this as DTOs (see Sec. V). This setup is relevant for targeted deployment scenarios of the entire platform.

Choice of queries: For prostate cancer, the clinical conduct is dependent predominantly on pathological exam (tumor grade, stage), tumor localization (local, invasion: seminal vesicle invasion, lymph node invasion). The differential diagnostic is refer to malignant (prostate cancer or prostate adenocarcinoma) vs. non-malignant tumors (benign prostatic hyperplasia). We expect a good prognostic in the case of an

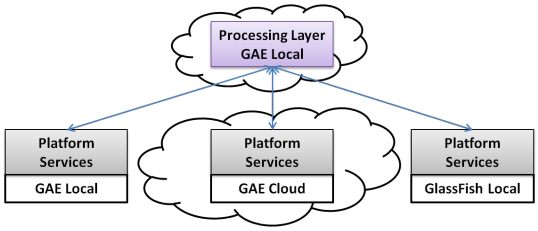

Fig. 4. Evaluation environment: platform services deployed locally (on both GlassFish and GAE) or on the Google cloud environment, and processing layer deployed on local GAE.

early diagnostic that is taken at an early stage - precancerous lesions (high-grade intraepithelial neoplasia and atypical small acinar proliferation) vs. low-grade intraepithelial neoplasia (benign lesions), aggressive prostate cancer vs. indolent prostate cancer. We could estimate the prognostic related with the indication of outcome regardless of the specific treatment the patient receives (radical prostatectomy) or to prediction that indicate the likelihood of response to a specific therapy (biochemical recurrence).

All of these could be taken into consideration as key-words (see Table I) and for each of this condition we could expect to have significantly different findings into databases that could help the clinical decision. As such, the set of queries from Table I has been performed on all the integrated data sources.

TABLE I
EVALUATION QUERIES AND THE CORRESPONDING NUMBER OF ENTITIES RETRIEVED BY THE PLATFORM FROM THE 'OMICS' DATA SOURCES UNIPROT (UP), ARRAYEXPRESS (AE), AND GENBANK (GB), RESP.

Id	Query	#UP	#AE	#GB
q1	prostate cancer	990	476	102253
q2	normal prostate	359	176	13017
q3	prostate adenocarcinoma	180	399	12375
q4	aggressive prostate cancer	6	21	2624
q5	radical prostatectomy	2	25	13160
q6	prostate specific antigen	146	10	24038
q7	biochemical recurrence	2	12	4082
q8	seminal vesicle invasion	7	4	914
q9	lymph node invasion	11	22	13403

B. Evaluation of the Integration Approach

The results of retrieving the protein data from UniProt, corresponding to the evaluation queries, are presented in Fig. 5. Fetch Time represents the time needed for a data retrieval service to pull data from an external data source. It is a relevant indicator for the responsiveness of the integration solution provided in the platform.

Normalizing the Fetch Time with the number of entities retrieved by each query each of the deployments GlassFish, GAE Local, and GAE Cloud, we got an average Fetch Time/Entity of 26 ms, 18 ms, and 6 ms resp. The cloud deployment led to the fastest execution of our services. From the local deployments, GAE Local execution was generally faster then GF Local. Exceptions are q1 and q2, for which the GF Local execution was faster than GAE Local. Out of all evaluation queries, these two retrieve the largest number

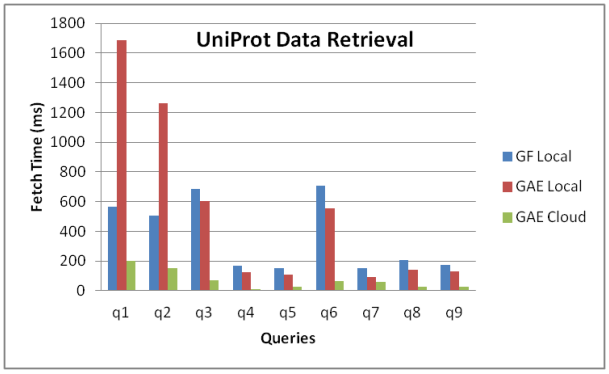

Fig. 5. Transfer time of the results corresponding to the queries from Table I executed against UniProt.

of resulting entities. The threshold for choosing between the local deployments appears to be around 200 retrieved entities, which is close to the results obtained for the execution of q3.

Measurements of experiment data retrieval from ArrayExpress, corresponding to the evaluation queries, are presented in Fig. 6. The average Fetch Time/Entity obtained was 14 ms (GF

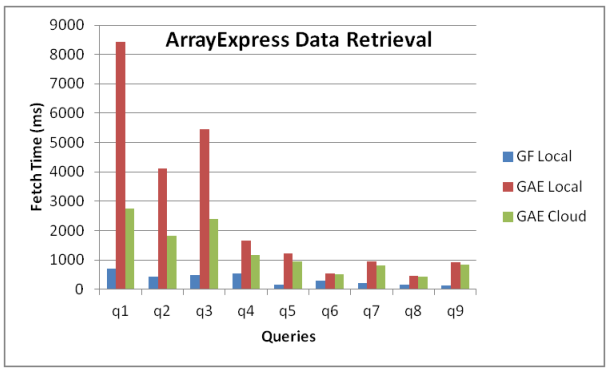

Fig. 6. Transfer time of the results corresponding to the queries from Table I executed against ArrayExpress.

Local), 52 ms (GAE Local), and 42 ms (GAE Cloud), resp. The local deployment using GlassFish is significantly faster than the GAE-based ones. Results obtained on GAE Cloud are still better than GAE Local, as in the case of UniProt.

C. Discussion

The execution of the queries from Table I against PubMed through GenBank with our GenBank service exhibits a behavior similar to the ones in Fig. 5. The Fetch Time is much higher in the case of GenBank, situation that is firstly explained by the numerous data entries retrieved (see Table I). Out of the three deployment models for the services of our platform, we believe that GAE Cloud is the most promising one. GF local is also an option to consider, especially for handling large resulting data sets.

Experience 1. The choice of integration strategy using web services has a big extension potential. It allowed us to fully control the input queries, the timing of the invocation of the retrieval services from external sources, the processing of the resulting data. This allows us to efficiently interact with the 'omics' data sources.

Experience 2. A downside of our integration approach is represented by the additional effort that is invested in the homogenization of both the search queries and the retrieved data. The different data sources have different search syntax and semantics that needs to be properly understood and employed, as a prerequisite for the correct functioning of the BioGenProtOMICS platform. Furthermore, the format of the data retrieved depends on its data source. In order to allow for common handling of the multi-omics data in our platform, we need to transform it. Nevertheless, common data types and search semantics/syntax brings great advantage in the formulation of the BMDV process tasks.

Experience 3. The design of the platform with the cloud deployment model in mind makes the resulting software prototype very flexible. Various parts of it (either layer-wise or services within the same layer) could be run on cloud resources, benefiting from the inherent advantages of clouds (in our case, elasticity in the first place). Nevertheless, before choosing the deployment target of any of the parts of the platform, special attention must be paid to the amount of data and processing the respective part needs to handle. As seen in Fig. 6, the use of a local application server might more efficient in some cases.

The proposed integration solution represents a solid base for retrieving 'omics' data necessary for performing the steps 2)–9) from the BMDV process (see Sec. III). The entities delivered by the data retrieval services of the platform can now be statistically processed, filtered etc.

VII. Conclusions

The paper proposes an integration approach of multi-omics data sources, based on web services and cloud, with the aim of performing *in-silico* biomarker discovery and validation. This work prepared the integration backbone of the novel software platform BioGenProtOMICS. The implemented services provide the query results in a homogeneous way to the processing tasks required by the BMDV process. Further integration of 'omics' data sources could be handled similarly. In order to enable fully featured BMDVs, our future work needs to focus on the enhancing of the integrated search with semantics support on the one hand. On the other hand, we need to further investigate cloud technologies that will allow a more flexible deployment of the platform layers and services. Of potential interest for future research in this direction are Amazon EC2 (commercial) and StratusLab cloud [18] (open source) technologies.

Acknowledgements

This work was partly supported by the FP7 financing contract IGE RI-261560.

References

[1] D. Benson, I. Karsch-Mizrachi, D. Lipman, J. Ostell, and E. Sayers, "Genbank," *Nucleic Acids Res*, vol. 40, pp. 48–53, 2012.

[2] H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, R. Mani, T. Rayner, A. Sharma, E. William, U. Sarkans, and A. Brazma, "ArrayExpress–a public database of microarray experiments and gene expression profiles," *Nucleic Acids Res*, vol. 35, no. Database issue, Jan. 2007. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/17132828

[3] C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek, "The Universal Protein Resource (UniProt): an expanding universe of protein information," *Nucleic Acids Research*, vol. 34, no. suppl 1, pp. D187–D191, 2006. [Online]. Available: http://nar.oxfordjournals.org/content/34/suppl_1/D187.abstract

[4] C. Wu, C. Orozco, J. Boyer, M. Leglise, J. Goodale, S. Batalov, C. Hodge, J. Haase, J. Janes, J. Huss, and A. Su, "BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources," *Genome Biology*, vol. 10, no. 11, pp. R130+, 2009. [Online]. Available: http://dx.doi.org/10.1186/gb-2009-10-11-r130

[5] J. Kohler, J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Regg, C. Rawlings, P. Verrier, and S. Philippi, "Graph-based analysis and visualization of experimental results with ondex," *Bioinformatics*, vol. 22, no. 11, pp. 1383–1390, 2006. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/22/11/1383.abstract

[6] A. Gupta, W. Bug, L. Marenco, X. Qian, C. Condit, A. Rangarajan, H. M. Muller, P. L. Miller, B. Sanders, J. S. Grethe, V. Astakhov, G. M. Shepherd, P. W. Sternberg, and M. E. Martone, "Federated access to heterogeneous information resources in the neuroscience information framework (nif)," *Neuroinformatics*, vol. 6, no. 3, pp. 205–207, September 2008.

[7] M. Safran, I. Dalah, J. Alexander, N. Rosen, T. Iny Stein, M. Shmoish, N. Nativ, I. Bahir, T. Doniger, H. Krug, A. Sirota-Madi, T. Olender, Y. Golan, G. Stelzer, A. Harel, and D. Lancet, "Genecards version 3: the human gene integrator," *Database*, vol. 2010, 2010. [Online]. Available: http://database.oxfordjournals.org/content/2010/baq020.abstract

[8] G. D. Bader, M. P. Cary, and C. S, "Pathguide: a pathway resource list," in *Nucleic Acids Research, 34(suppl 1):D504506*, 2006.

[9] D. M. R. o PachA (3)n, A. Nagel, J. Neigenfind, R. Wagner, R. Basekow, E. Weber, B. Mueller-Roeber, S. Diehl, and B. Kersten, "Gabipd: the gabi primary database - a plant integrative omics database," *Nucleic Acids Research*, vol. 37, no. Database-Issue, pp. 954–959, 2009.

[10] L. Lane, "nextprot, a new knowledgebase on human proteins," Available from Nature Precedings ¡http://dx.doi.org/10.1038/npre.2010.5104.1¿, 2010.

[11] A. Statnikov, I. Tsamardinos, Y. Dosbayev, and C. F. Aliferis, "GEMS: a system for automated cancer diagnosis and biomarker discovery from microarray gene expression data," *Int J Med Inform*, vol. 74, no. 7-8, pp. 491–503, Aug. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&\#38;db=PubMed\&\#38;dopt=Citation\&\#38;list_uids=15967710\%20

[12] H. F. Deus, D. F. Veiga, P. R. Freire, J. N. Weinstein, G. B. Mills, and J. S. Almeida, "Exposing the cancer genome atlas as a SPARQL endpoint," *Journal of Biomedical Informatics*, vol. 43, no. 6, pp. 998–1008, 2010.

[13] G. A. Komatsoulis, *Collaboration in Cancer Research Community: Cancer Biomedical Informatics Grid (caBIG)*. John Wiley & Sons, Inc., 2011, pp. 261–280. [Online]. Available: http://dx.doi.org/10.1002/9781118026038.ch17

[14] A. J. Atkinson, W. A. Colburn, V. G. Degruttola, D. L. Demets, G. J. Downing, D. F. Hoth, J. A. Oates, C. C. Peck, R. T. Schooley, B. A. Spilker, J. Woodcock, and S. L. Zeger, "Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework*," *Clin Pharmacol Ther*, vol. 69, no. 3, pp. 89–95, Mar. 2001. [Online]. Available: http://dx.doi.org/10.1067/mcp.2001.113989

[15] J. H. Phan, R. A. Moffitt, T. H. Stokes, J. Liu, A. N. Young, S. Nie, and M. D. Wang, "Convergence of biomarkers, bioinformatics and nanotechnology for individualized cancer treatment." *Trends in biotechnology*, vol. 27, pp. 350–8, 2009 Jun 2009.

[16] R. Suharoschi, C. Iuga, N. Crisan, D. Pamfil, O. Balacescu, and I. Muntean, "A proposed curation protocol for discovery cancer potential biomarker candidates," *Agricultura, Agricultural Practice and Science*, vol. 81-82, accepted 2012.

[17] D. Sanderson, *Programming Google App Engine: Build and Run Scalable Web Apps on Google's Infrastructure*, 1st ed. O'Reilly Media, Inc., 2009.

[18] C. Loomis, M. Airaj, M.-E. Bégin, E. Floros, S. Kenny, and D. O'Callaghan, *StratusLab Cloud Distribution*. 12 Back Chapman St, Newcastle upon Tyne, UK: Cambridge Scholars, Jan. 2012, pp. 260–282. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00676252/