Bio-Search Computing: Integration and global ranking of bioinformatics search results

Marco Masseroli*, Giorgio Ghisalberti, Stefano Ceri

Dipartimento di Elettronica e Informazione, Politecnico di Milano,
Piazza Leonardo da Vinci 32, 20133 Milano, Italy

Summary

In the Life Sciences, numerous questions can be addressed only by comprehensively searching different types of data that are inherently ordered, or are associated with ranked confidence values. We previously proposed Search Computing to support the integration of the results of search engines with other data and computational resources. This paper presents how well known bioinformatics resources can be described as search services in the search computing framework and integrated analyses over such services can be carried out. An initial set of bioinformatics services has been described and registered in the search computing framework and a bioinformatics search computing (Bio-SeCo) application using these services has been created. This current prototype application, the available services that it uses, the queries that are supported, the kind of interaction that is therefore made available to the users, and the future scenarios are here described and discussed.

1 Introduction

In the Life Sciences, questions are often complex and simultaneously regard several different functional and structural aspects of an organism and its biomolecular entities (e.g. the genes expressed in certain conditions, their mutations and their involvement in pathological phenotypes or diseases, the proteins with their protein domains and 3D structure, their participation in different biochemical pathways and biological processes, etc.). An example is the following: "Which drugs treat diseases that are likely to be associated with a given genetic mutation?" Such questions can be addressed only by exploring, comprehensively searching and globally evaluating the numerous available data and their relationships, which are of different types and often inherently ordered or associated with ranked confidence values.

Access to these data is being increasingly provided by web services, which often offer specific *search services*, i.e. bioinformatics services that provide results (often ranked) of user defined searches within data repositories. These services provide users with rapid and selective access to biomedical data from potentially huge repositories. However, individual search tools are often ineffective for use in applications in which the answer to a request involves combining results from more than one search engine. In particular, available search services [1] typically provide *vertical* search capabilities, in that they are focused on a single topic. They seek individual items that meet the criteria specified in a request, whereas in practice information relevant to a biomedical requirement may be spread over several resources. Furthermore, it is often essential to combine multiple vertical search services to perform multi-topic searches, where the different individual topic searches either refine or augment previous results. For example, if the user is interested in knowing which genes both encode proteins with high sequence similarity to a given protein and are significantly expressed in the same given biological condition or tissue, current practice typically involves the integration of results from three different searches (for similar proteins, protein encoding

* To whom correspondence should be addressed. E-mail: marco.masseroli@polimi.it

genes and gene expressions), where the individual search results are themselves likely to be ranked by some criteria [2]. Such an integration task, taking account of the rankings, is termed a *multi-topic search* and may be carried out manually or by a custom program, by composing services in workflows through the use of flexible service oriented architectures (SOA).

The definition of workflows to infer new knowledge from existing datasets by using available services is an often performed activity in bioinformatics [3]. Notable examples of workflow systems supporting such activity include Taverna [4], Wings/Pegasus [5], [6], Galaxy [7], Triana [8] and Kepler [9]. Taverna, the most known and used in bioinformatics, has been used to support experimental investigation into a variety of research areas. It is a language and computational model designed to support the automation of complex, service-based and data-intensive scientific processes. Yet, Taverna and the other available workflow systems and data integration platforms are not able to deal with rankings and scores of both intermediate results and global combinations; thus, they do not provide support for multi-topic search.

Search Computing [10], [11] has been proposed to support the integration of search engine results from different areas with other data and computational resources. The innovative contribution of this new infrastructure, that sets it apart from previous works, is its support for combining different data sources and dealing also with ranked partial results, taking them into account in order to provide a global ranking of the integrated partial ranked data.

This paper complements a previous study [2] of the envisaged relevance of search computing to the Life Sciences (in particular to information integration and support for ordered data in the Life Sciences) by illustrating and discussing the application of search computing in a bioinformatics use case. Besides the demonstration of the previously described principle through the implementation of the here discussed bioinformatics application, it also provides a view to identifying the extent to which the existing platform for multi-topic search provides useful facilities for representing and integrating bioinformatics search services to be used in biomedical applications.

2 Search Computing

Search Computing (http://www.search-computing.eu/) is a new approach that provides the abstractions, methods, tools and computing systems required to express multi-topic queries, also over ranked data sources, and to build their answers [11]. Figure 1 represents the overall conceptual architecture of the search computing system.

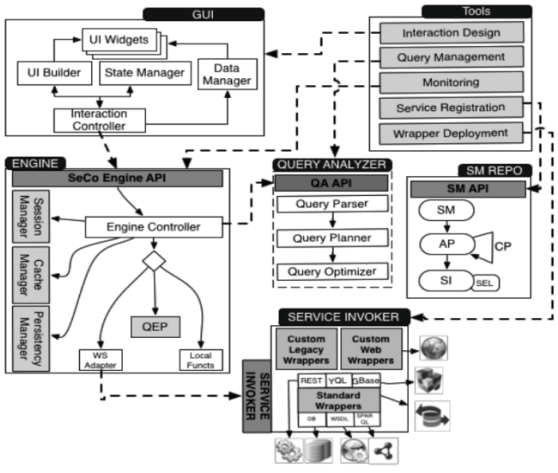

Figure 1: Search computing architecture

The search computing system architecture includes a variety of tools that cover service development and publishing, a query execution environment, as well as application registration and query tuning functionalities. A service registration environment eases the creation of wrappers to adapt existing services to the system architecture. A repository stores the definitions of wrappers and wrapped data sources, which are used for the deployment of search-based applications in generic areas.

Search computing can be used to describe well known bioinformatics and biomedical resources as search services, and carry out integrated analyses over such services [2]. In particular, this makes explicit how different ranked data (e.g. from sequence comparisons, gene expression results, functional annotation analyses, or several other different bioinformatics aspects) can be integrated in a way that takes account of the rankings of the different types of data and analyses. In so doing, ranking is innovatively used as a first class citizen for data integration in the Life Sciences. Thus, search computing and its information exploration paradigm based on semantic resource framework (a multi-level description of data sources, including search services) provide a platform for expressing requests over multiple search services, such that the results of the integrated requests take account of the rankings of individual search results. Figure 2 describes some of the several different types of biomedical data for which search services are available. It also shows the relationships between such data types provided by the numerous bioinformatics services available and how they constitute a semantic resource framework, which can be leveraged for complex search computing. Thus, by using available web services for searching individual bioinformatics data types and taking advantage of the relationship data they supply and the attributes they define for providing a ranking, search computing techniques can be innovatively applied to efficiently explore available data and search for globally ranked answers to complex biomedical questions.

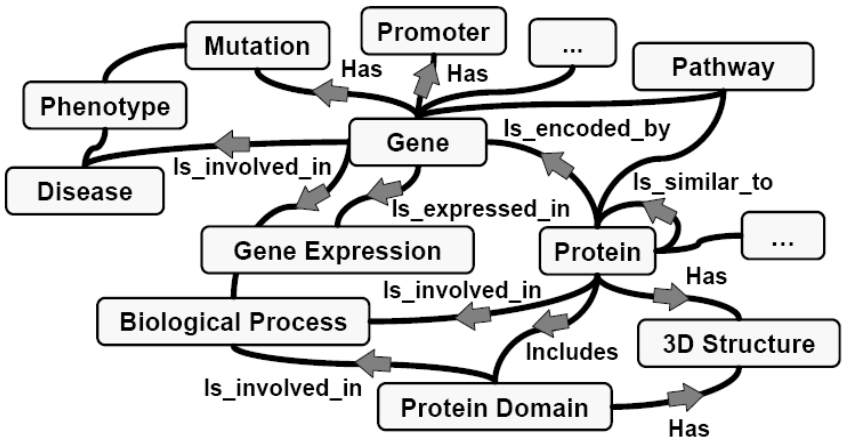

Figure 2: Biomedical Semantic Resource Framework

3 Multi-topic Complex Searches: Implementation and Results

We have described and registered an initial set of bioinformatics services in the developed search computing framework. A bioinformatics search computing (Bio-SeCo) application using these services has been created and made publicly available at http://www.search-computing.org/UIDemoBio/. It can answer the following paradigmatic complex case study question: "Which genes encode proteins in different organisms with high sequence similarity to a given protein X and are significantly (over or under) co-expressed in the same given biological tissue Y?" This is just one of the many questions that are supported by suitably composing services shown in Figure 2, by means of a scientist-friendly interface. Indeed, such

type of interesting and complex queries cannot be automatically answered by any other currently available system.

The above multi-topic question is decomposed into the following three single biomedical topic sub-queries, each focused upon one of the nodes of Figure 2: "Which proteins in different organisms have the highest sequence similarity to a given protein X?"; "Which genes encode which proteins?"; and "Which genes are (over or under) co-expressed in the same given tissue Y?". Each of these sub-queries is mapped to an available search service, i.e. a sequence similarity search program such as BLAST, in one of its implementations (e.g. WU-BLAST [12] - http://www.ebi.ac.uk/blast2/), a query service in a database of genomic and proteomic data as our GFINDer (http://www.bioinformatics.polimi.it/GFINDer/) GPDW [13], and a search engine over a repository of gene expression data such as ArrayExpress Gene Expression Atlas [14] (http://www.ebi.ac.uk/gxa/), respectively. Arcs of Figure 2 allow connecting services one to another, thereby building the single query resolving the original problem.

3.1 Search service modelling for Bio-SeCo semantic resource framework

According to the search computing framework [11], each of the abovementioned nodes is modelled with *access patterns*, which describe the service and its input (I), output (O) and ranked output (R) attributes for the specific data accesses available; arcs are modelled by *connection patterns* explaining how service attributes are used for joining access patterns. In the proposed example we use the following three access patterns and two access patterns:

WU-BLAST(SequenceAlignmentProgram^I, SearchedDB^I, QuerySequenceID^I, Email^I, FoundSequenceDB^O, FoundSequenceID^O, FoundSequenceSymbol^O, FoundSequenceName^O, BestAlignment.Expectation^R)

GPDW_Protein2Gene(ProteinDB^I, ProteinID^I, ProteinDB^O, ProteinID^O, GeneDB^O, GeneID^O, GeneSymbol^O, Organism^O)

ArrayExpress(GeneSymbol^I, Organism^I, ExpressionRegulation^I, Condition^I, View^I, GeneSymbol^O, Organism^O, ExperimentalFactor^O, FactorValue^O, ExpressionRegulation^O, StudyNumber^R, BestStudy.Pvalue^R)

ExistsCodingGene(WU-BLAST, GPDW_Protein2Gene):
[(WU-BLAST.FoundSequenceDB = GPDW_Protein2Gene.ProteinDB
AND WU-BLAST.FoundSequenceID = GPDW_Protein2Gene.ProteinID)]

ExistsExpressedGene(GPDW_Protein2Gene, ArrayExpress):
[(GPDW_Protein2Gene.GeneSymbol = ArrayExpress.GeneSymbol
AND GPDW_Protein2Gene.Organism = ArrayExpress.Organism)]

3.2 Search query submission

After selecting the registered services to be used, the specific question to be answer can be specified by setting the query specific input values. For the paradigmatic multi-topic case study question above considered, they are the *protein X* (indicated through its ID), the *type of co-expression* (*over* or *under*) and the *biological tissue Y*. Service specific input values, other than those provided at query execution time by the connected services as specified by the defined connection patterns, can be specified at service registration time as service default input values (e.g. the sequence alignment program to use, or the sequence database to search, for the WU-BLAST service). In the prototypical demonstrative Bio-SeCo application that we created (http://www.search-computing.org/UIDemoBio/), through a simple graphical interface the user can set query specific input values and launch the search query.

3.3 Query execution and retrieved results

The submitted multi-topic query is then executed by the search computing platform by calling the involved services with the user-defined input query values. In the case of our paradigmatic question, through the *WU-BLAST* access pattern, the search computing platform calls one of the BLAST programs (e.g. *SequenceAlignmentProgram* = "BLASTP"), available in the WU-BLAST web service registered in the platform, to search, in one of the protein sequence databases available to WU-BLAST (e.g. *SearchedDB* = "uniprotkb_swissprot"), for protein sequences highly similar to the sequence of the user specified protein X ID (e.g. *QuerySequenceID* = "uniprot:P24593"). Ranked results can be obtained in increasing order of probability that the alignment found is not better than an alignment found by chance. The IDs of the protein sequences found with the best alignment (i.e. with the lowest probability) are retained. Then, based on the *ExistsCodingGene* connection pattern, they are passed as input to the *GPDW_Protein2Gene* query service registered in the platform. This query service is automatically invoked by the search computing platform, according to the automatically defined query plan and its registered access pattern and service interface, in order to query the GPDW for genes encoding the proteins retrieved by the WU-BLAST web service. The symbol and organism name of the obtained genes are retrieved. Then, based on the *ExistsExpressedGene* connection pattern and the *ArrayExpress* accession pattern, they and the user specified expression type (e.g. *ExpressionRegulation* = "up") and biological tissue Y (e.g. *Condition* = "kidney") search constraints are sent as input to the ArrayExpress Gene Expression Atlas search engine, registered in the platform. ArrayExpress ordered search results include those genes, among the input ones, that in the ArrayExpress Archive are reported in decreasing order of probability to be significantly over co-expressed in the kidney.

3.4 Integration and global ranking of retrieved partial results

Partial search results provided by the individual services considered are composed according to the semantic resource network and taking into account their partial rankings, when available. Global ranking of integrated results is performed by composing partial rankings according to a predefined weighted combination function. This function can be arbitrary complex and describe the relationships between the partial rankings to be combined. Yet, if the partial rankings are expressed in the same units and range values, the simple weighted product of the partial rankings can be used as global ranking. This has been done also for the considered paradigmatic query, since its partial rankings to be combined are sequence alignment expectation values (from the WU-BLAST *BestAlignment.Expectation* attribute) and differential gene expression *p*-values (from the ArrayExpress *BestStudy.Pvalue* attribute), which are both dimensionless values expressed in the 0.0-1.0 value range.

Results provided by our Bio-SeCo application for the example input values given above are shown in Table 1. They represent the ordered list of the top genes that encode proteins with the highest sequence similarity to the *UniProt P24593* protein (*human Insulin-like growth factor-binding protein 5*) and are significantly *over expressed* in the *kidney* tissue or within any of its parts. According to the partial ranked results provided on April 29th, 2011 by the WU-BLAST, GPDW and ArrayExpress services registered in our search computing platform, they constitute the global ranked answer to the considered multi-topic paradigmatic case study question that the Bio-SeCo application can automatically compute by integrating the retrieved partial ranked results.

Table 1: Global ranked results provided by search computing to the case study question for the user input QuerySequenceID = "*uniprot:P24593*", ExpressionRegulation = "*up*" and Condition = "*kidney*". Expectation: BLAST expectation value; Diff. Expr.: gene differential expression type; for all table items, ArrayExpress Experimental Factor = "Organism part" (not shown)

WU-BLAST			GPDW_Protein2Gene		ArrayExpress				Global Rank
Similar Protein ID (UniProt)	Similar Protein Name	Expectation	Gene Symbol	Organism	Factor Value	Diff. Expr.	Study Number	*P*value	
Q07079	Insulin-like growth factor-binding protein 5	$1.80E^{-137}$	Igfbp5	Mus musculus	kidney	UP	10	$1.00E^{-11}$	**$1.80E^{-148}$**
P24594	Insulin-like growth factor-binding protein 5	$3.80E^{-137}$	Igfbp5	Rattus norvegicus	kidney	UP	3	$1.00E^{-11}$	**$3.80E^{-148}$**
P24593	Insulin-like growth factor-binding protein 5	$9.40E^{-141}$	IGFBP5	Homo sapiens	kidney cortex	UP	1	$8.54E^{-08}$	**$8.03E^{-148}$**
P24593	Insulin-like growth factor-binding protein 5	$9.40E^{-141}$	IGFBP5	Homo sapiens	kidney medulla	UP	1	$2.78E^{-05}$	**$2.61E^{-145}$**
P15473	Insulin-like growth factor-binding protein 3	$1.50E^{-53}$	Igfbp3	Rattus norvegicus	kidney	UP	2	$1.00E^{-11}$	**$1.50E^{-64}$**
P47878	Insulin-like growth factor-binding protein 3	$1.10E^{-52}$	Igfbp3	Mus musculus	kidney	UP	11	$1.00E^{-11}$	**$1.10E^{-63}$**
P47879	Insulin-like growth factor-binding protein 4	$3.50E^{-41}$	Igfbp4	Mus musculus	kidney	UP	11	$1.00E^{-11}$	**$3.50E^{-52}$**
P24592	Insulin-like growth factor-binding protein 6	$1.40E^{-34}$	IGFBP6	Homo sapiens	kidney	UP	1	$1.00E^{-11}$	**$1.40E^{-45}$**
P21743	Insulin-like growth factor-binding protein 1	$1.60E^{-33}$	Igfbp1	Rattus norvegicus	kidney	UP	3	$1.00E^{-11}$	**$1.60E^{-44}$**
P47876	Insulin-like growth factor-binding protein 1	$2.70E^{-33}$	Igfbp1	Mus musculus	kidney	UP	8	$1.00E^{-11}$	**$2.70E^{-44}$**
P22692	Insulin-like growth factor-binding protein 4	$3.90E^{-39}$	IGFBP4	Homo sapiens	kidney cortex	UP	1	$1.82E^{-04}$	**$7.10E^{-43}$**
P08833	Insulin-like growth factor-binding protein 1	$9.70E^{-32}$	IGFBP1	Homo sapiens	kidney	UP	2	$1.00E^{-11}$	**$9.70E^{-43}$**
P12843	Insulin-like growth factor-binding protein 2	$1.90E^{-30}$	Igfbp2	Rattus norvegicus	kidney	UP	2	$1.00E^{-11}$	**$1.90E^{-41}$**
P18065	Insulin-like growth factor-binding protein 2	$5.90E^{-29}$	IGFBP2	Homo sapiens	kidney medulla	UP	1	0.016	**$9.44E^{-31}$**
P18065	Insulin-like growth factor-binding protein 2	$5.90E^{-29}$	IGFBP2	Homo sapiens	kidney cortex	UP	1	0.017	**$1.00E^{-30}$**
P35572	Insulin-like growth factor-binding protein 6	$1.20E^{-16}$	Igfbp6	Rattus norvegicus	kidney	UP	1	$1.00E^{-11}$	**$1.20E^{-27}$**
Q9R118	Serine protease HTRA1	0.0087	Htra1	Mus musculus	kidney	UP	9	$1.00E^{-11}$	**$8.70E^{-14}$**
Q61581	Insulin-like growth factor-binding protein 7	0.021	Igfbp7	Mus musculus	kidney	UP	10	$1.00E^{-11}$	**$2.10E^{-13}$**
Q92563	Testican-2	0.0032	SPOCK2	Homo sapiens	kidney cortex	UP	1	$3.23E^{-06}$	**$1.03E^{-08}$**
Q92743	Serine protease HTRA1	0.011	HTRA1	Homo sapiens	kidney cortex	UP	1	0.032	**$3.52E^{-04}$**
Q9NQ30	Endothelial cell-specific molecule 1	0.043	ESM1	Homo sapiens	kidney medulla	UP	1	0.011	**$4.73E^{-04}$**
Q9NZV1	Cysteine-rich motor neuron 1 protein	0.19	CRIM1	Homo sapiens	kidney medulla	UP	1	0.019	**$3.61E^{-03}$**

As expected, the resulting genes correctly include the human IGFBP5 gene, which encodes the input protein; in fact, this gene is known to be over expressed in kidney [15]. They also

comprise some other genes that are members of the same *insulin-like growth factor binding protein* (IGFBP) family in human and other organisms (actually only in human, mouse and rat since ArrayExpress provides gene expression data in kidney only for these organisms). It is worth noticing that not all gene members of the IGFBP family (in human, mouse and rat) are present in the computed result list; this correctly reflects the data provided by the considered bioinformatics services. In fact, IGFBP3 and IGFBP7 in human, Igfbp2 and Igfbp6 in mouse and Igfbp4 in rat do not result over expressed in kidney from the ArrayExpress data. Similarly, from the WU-BLAST data, the protein codified by the rat Igfbp7 gene is not among the amino acid sequences, contained in the searched UniProtKB/Swiss-Prot database, that are highly similar to the input protein sequence (actually, it is present only in the UniProtKB/TrEMBL database).

Towards the end of the resulting ordered list, a few other genes that codify proteins less similar in sequence to the input protein and that are less significantly over expressed in kidney are also included. Interestingly, the human IGFBP5 gene that encodes the input protein is not the first of the list; in fact, its two Igfbp5 ortholog genes in mouse and rat, although with some differences in sequence from the input protein, result to be more significantly over expressed in kidney according to the ArrayExpress data.

4 Discussion and Future Work

The created Bio-SeCo application here described is a demonstrator of the capabilities of the search computing technology to be effectively applied to efficiently search for globally ranked answers to complex biomedical questions. It fully enables the user to run the considered example type of multi-topic biomedical query and obtain the computed global search results, which depend on the data (and their quality) provided by the individual bioinformatics search services considered. Yet, it is just a first prototype that will soon be improved. In particular, our near future work will address application flexibility in all aspects of user interaction, i.e. in query definition and expansion, ranking composition function setting, and result visualization and browsing.

The fixed multi-topic case study question considered is just a simple paradigmatic example of how the search computing approach works and of the type of questions that can be addressed with search computing. Other interesting biomedical questions, with even more relevant bioinformatics predictions, can be similarly created and answered. In the next Bio-SeCo release, we will provide functionalities for flexible query definition by enabling the user to build its own search through a graphical interface. According to the semantic resource network of search services registered in the search computing framework, it will be possible to define an initial global search and then expand or refine it by adding partial searches in new topic areas, remove previously included search services, modify input and filter parameters of considered search services, or roll back to previous search results. For instance, for the considered example query, interesting expansions could regard the search for the involvement in a specific biological process of the genes found significantly expressed in the same given biological tissue, or the search for specific domains within the proteins found similar in sequence to a given protein.

Weight coefficients of the ranking composition function will be definable and interactively modifiable by the user at query time in order to allow customizing global ranking calculation also according to the specific partial results retrieved by each individual search service composed.

Bio-SeCo result visualization interface, usable by all Web browsers, will also be improved in order to allow interactively browsing and expanding individual search results and highlighting global combinations of results with particular relevance.

Finally, several additional bioinformatics services will be registered in the search computing framework; this will enable bio-scientists to build multi-topic queries over a variety of services, thereby increasing their ability to easily express and efficiently solve complex biomedical questions.

5 Conclusions

This paper has presented and discussed the new Search Computing approach and a prototypical demonstrative Bio-Search Computing application. The developed application demonstrates how available bioinformatics resources can be described as search services in the search computing framework, and that the framework then supports scientists in formulating and efficiently executing multi-topic biomedical queries. Such queries are numerous in the Life Sciences and only can be addressed by comprehensively evaluating different types of data, which are often inherently ordered or associated with ranked confidence values. By providing direct support for ranking as a first class citizen in data integration, search computing provides distinctive data integration features, not supported in other available data integration or workflow platforms. In so doing, search computing can support exploratory search and curiosity driven browsing of biomedical data, thus enabling ambitious data driven biomedical knowledge discovery and verification.

Acknowledgements

This research is part of the "Search Computing" (SeCo) project (2008-2013), funded by the European Research Council (ERC), under the 2008 call for "IDEAS Advanced Grants".

References

[1] C. A. Goble, K. Belhajjame, F. Tanoh, J. Bhagat, K. Wolstencroft, R. Stevens, S. Pettifer, E. Nzuobontane, H. McWilliam, T. Laurent and R. Lopez. BioCatalogue: a curated Web Service registry for the Life Science community. *ISMB/ECCB 2009*. Technology Track: TT40, 2009.

[2] M. Masseroli, N. W. Paton and I. Spasić. Chapter 15: Search Computing and the Life Sciences. In: S. Ceri and M. Brambilla, editors. *Search Computing - Challenges and Directions*. LNCS, vol. 5950, pp. 291-306. Springer, Heidelberg, D., 2010.

[3] E. Deelman, D. Gannon, M. Shields and I. Taylor. Workflows and e-Science: An overview of workflow system features and capabilities. *Future Gener. Comput. Syst*. 25(5): 528-540, 2009.

[4] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. Pocock, P. Li and T. Oinn. Taverna: a tool for building and running workflows of services, *Nucleic Acids Res*. 34(Web Server issue): 729-732, 2006.

[5] Y. Gil, P. Gonzalez-Calero, J. Kim, J. Moody and V. Ratnakar. A semantic framework for automatic generation of computational workflows using distributed data and component catalogs. *J. Exp. Theor. Artif. Intel*. to appear, 2011.

[6] E. Deelman, G. Singh, M. H. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, G. B. Berriman, J. Good, A. C. Laity, J. C. Jacob and D. S. Katz. Pegasus: A

framework for mapping complex scientific workflows onto distributed systems. *Scientific Programming* 13(3): 219-237, 2005.

[7] A. Nekrutenko. Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the Life Sciences. *Genome Biol*. 11(8): R86, 2010.

[8] D. Churches, G. Gombas, A. Harrison, J. Maassen, C. Robinson, M. Shields, I. Taylor and I. Wang. Programming scientific and distributed workflow with Triana services. *Concurr. Comput*. 18(10): 1021-1037, 2006.

[9] B. Ludäscher, I. Altintas and C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao and Y. Zhao. Scientific workflow management and the Kepler system. *Concurr. Comput*. 18(Workflow in Grid Systems): 1039-1065, 2005.

[10] D. Braga, S. Ceri, F. Daniel and D. Martinenghi. Mashing up search services. *IEEE Internet Comput*. 12(5): 16-23, 2008.

[11] S. Ceri and M. Brambilla, *Search Computing – Challenges and Directions*. LNCS, vol. 5950. Springer, Heidelberg, D. 2010.

[12] R. Lopez, V. Silventoinen, S. Robinson, A. Kibria, and W. Gish. WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res*. 31(13): 3795-3798, 2003.

[13] M. Masseroli, S. Ceri and A. Campi. Integration and mining of genomic annotations: Experiences and perspectives in GFINDer data warehousing. In: N. W. Paton, P. Missier and C. Hedeler (eds.). *Data Integration in the Life Sciences. 6th International Workshop, DILS 2009*. LNCS (LNBI), vol. 5647, pp. 88-95. Springer, Heidelberg, D. 2009.

[14] H. Parkinson, U. Sarkans, M. Shojatalab, N. Abeygunawardena, S. Contrino, R. Coulson, A. Farne, G.G. Lara, E. Holloway, M. Kapushesky, P. Lilja, G. Mukherjee, A. Oezcimen, T. Rayner, P. Rocca-Serra, A. Sharma, S. Sansone and A. Brazma. ArrayExpress - A public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*. 33(Database issue): D553-D555, 2005.

[15] M. R. Schneider, E. Wolf, A. Hoeflich and H. Lahm. IGF-binding protein-5: Flexible player in the IGF system and effector on its own. *J. Endocrinol*. 172(3): 423-440, 2002.