Bioscientific data processing and modeling

Kok, J.N.; Lamprecht, A.-L.; Verbeek, F.J.; Wilkinson, M.D.; Margaria, T.; Steffen, B.

Bioscientific Data Processing and Modeling

Joost Kok[1], Anna-Lena Lamprecht[2], Fons J. Verbeek[1], and Mark D. Wilkinson[3]

[1] Leiden Institute of Advanced Computer Science, Leiden University, 2300 RA Leiden, The Netherlands
{joost,fverbeek}@liacs.nl
[2] Chair for Service and Software Engineering, University of Potsdam, 14482 Potsdam, Germany
lamprecht@cs.uni-potsdam.de
[3] Centro de Biotecnología y Genómica de Plantas, Parque Científico y Tecnológico de la U.P.M., Campus de Montegancedo, 28223 Pozuelo de Alarcón (Madrid), Spain
mark.wilkinson@upm.es

With more than 200 different types of "-omic" data [1] spanning from sub-molecular, through molecular, cell, cell-systems, tissues, organs, phenotypes, gene-environment interactions, and ending at ecology and organism communities, the problem and complexity of bioscientific data processing has never been greater. Often data are generated in high-throughput studies with the aim to have a sufficient volume to find patterns and detect rare events. For these high-throughput approaches new methods have to be developed in order to assure integrity of the volume of data that is produced. At the same time efforts to integrate these widely-varying data types are underway in research fields such as systems biology. Systems-level research requires yet additional methodologies to pipeline, process, query, and interpret data, and such pipelines are, themselves, objects of scientific value if they can be re-used or re-purposed by other researchers.

This ISoLA 2012 special track focuses at the various topics concerned with the discovery and preservation of knowledge in the biosciences. The track comprises four papers, of which three are concerned with algorithms for image analysis, and one with a new workflow management methodology. The following gives a brief overview of these two thematic areas and of all the papers in the track.

Algorithms for Image Analysis

Although imaging and bioinformatics are research fields in their own right, there exists a quite substantial overlap between these two areas. On the interface of these two fields we find typically with image analysis as well as with the study of interoperability of image information to other bio-molecular information resources. In the life-sciences image analysis spans quite an application area ranging from molecular biology to interpretation of areal imagery for ecology. Then there is medical imaging focussed on patients and health care. Here we focus on the imaging from the molecules to (small) organisms; the imaging device is the microscope and the field is pre-clinical research.

T. Margaria and B. Steffen (Eds.): ISoLA 2012, Part II, LNCS 7610, pp. 7–11, 2012.

In microscopy imaging, at least, three issues are important, the first being obtaining and organizing the images, then analyzing the images and reducing the scene to numbers so that patterns can be found and analyzed, next, the information in the image needs to be represented properly. The analysis of images requires images to be acquired in large volumes so that patterns are statistically meaningful. Moreover, large volumes are required to detect rare events. A trend in life sciences research is, therefore, to approach problem with a high-throughput workflow. This puts demands in the acquisition phase, that need be largely automated but also on the processing phase. The latter requires algorithms that are robust and reproducible; here we present two examples on different levels of resolution, one on the organismal/tissue level [2,3] and one at the cellular level [4,5]; application of high-throughput to cellular systems is also referred to as cytomics. The specific algorithms that are presented here are designed and evaluated with the specific requirements for high-throughput analysis in mind.

Further processing of the features extracted from the images requires frameworks for pattern recognition specific to the data at hand [4,6,2]. However, we need to be able to integrate images as well as the resulting analysis in systems that include a reference model. Such systems are now being made on the level of the model system: e.g. mouse [7], the zebrafish [8,9], but also on the level of the organ. The brain is a good example for that, the rodent brain is used as a model for the human brain and specific reference systems for integration are being developed for the rodent brain [7]. The integration requires intelligent use of reference systems on the semantic level [4,10]. Therefore well maintained ontologies will be extremely important to maintain and disclose the large amounts of data that are currently produced. Ultimately, resources for genomic and molecular research will be integrated with image based resources. The challenge for the scientific community is to do this right.

The first paper of this ISoLA track, **Using multiobjective optimization and energy minimization to design an isoform-selective ligand of the 14-3-3 protein** (Hernando Sanchez-Faddeev, Michael T.M. Emmerich, Fons J. Verbeek, Andrew H. Henry, Simon Grimshaw, Herman P. Spaink, Herman W. van Vlijmen and Andreas Bender) [11], presents an approach for de novo design of protein ligands based on evolutionary multiobjective optimization. It shows that multiobjective optimization with evolutionary algorithms can be successfully employed in selective ligand design.

The paper **Segmentation for High-throughput Image Analysis: Watershed Masked Clustering** (by Kuan Yan and Fons J. Verbeek) [12] is concerned with high-throughput analysis of images of cells. It describes a new segmentation algorithm for high-throughput imaging, which is in particular suitable for image analyses in the fields of cytomics and high-throughput screening. The algorithm has been used with good results in a number of studies and is reported to perform better than previous algorithms for this task.

In **Efficient and Robust Shape Retrieval from Deformable Templates** (Alexander E. Nezhinski and Fons J. Verbeek) [13] an algorithmic framework for the automated detection of shapes in images through deformable templates is

presented. For demonstration purposes, it is applied to a biological case study, namely to high-throughput screening images of zebrafish larvae, and the algorithm is reported to be particularly accurate and robust.

Workflow Management

In recent years, numerous software systems have been developed for specifically supporting the management of scientific workflows (see, e.g., [14,15] for surveys). Research in this comparatively new field is currently going into many different directions. At the previous ISoLA in 2010, we focused on workflow management for scientific applications in the scope of a symposium track on "Tools in scientific workflow composition" [16], which comprised papers on subjects such as tools and frameworks for workflow composition, semantically aware workflow development, and automatic workflow composition, as well as some case studies, examples, and experiences.

Particularly interesting and challenging in the field of scientific workflow management is currently the research concerned with the use of semantics-based methods for automating workflow composition (see, e.g., [17,18]). Some examples of concrete systems which have lately been applied for semantics-based, (semi-) automatic workflow composition in the bioinformatics domain are the Bio-jETI framework [19,20,21] that makes use of workflow synthesis techniques to translate abstract, high-level workflow specifications into concrete, executable workflow instances, the jORCA [22,23] system that automatically creates pipelines of web services given the desired input and output data types, the SADI and SHARE frameworks [24,25] that facilitate on-the-fly service discovery and execution based on OWL-annotated data, and the Wings (Workflow INstance Generation and Selection) [26] extension for the Pegasus [27] grid workflow system that provides functionality for (semi-) automatic workflow creation based on semantic representations and planning techniques. Some of these systems have also been presented in the scope of the ISoLA 2010 track.

In this context, and as a continuation of the ISoLA 2010 paper on semantics-guided workflow construction in the Taverna workbench [28], the fourth paper of our track addresses the problem of workflow sharing and re-purposing in bioinformatics: In **OWL-DL domain models as abstract workflows** (Ian Wood, Ben Vandervalk, Luke McCarthy and Mark D. Wilkinson) [29], the authors discuss the growing popularity of formal analytical workflows, and the associated difficulty in re-using these workflows due to their rigidity. To overcome these issues, they present an original approach where a domain-concept, modeled in OWL-DL and based on the SADI and SHARE frameworks, can be used dynamically as a workflow template, which is then concretized into a Web Service workflow at run-time. Moreover, the semantics inherent in these domain-models can act as a form of workflow annotation. The authors propose that, over time, these abstract workflows may be easier to share and repurpose than conventional "concrete" workflows. The paper demonstrates the approach by automatically reproducing a published comparative genomics analysis through creating an OWL-DL representation of the biological phenomenon being studied.

References

1. McDonald, D., Clemente, J., Kuczynski, J., Rideout, J., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., Knight, R., Caporaso, J.: The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. GigaScience 1(1), 7 (2012)
2. Stoop, E., Schipper, T., Rosendahl Huber, S., Nezhinsky, A., Verbeek, F., Gurcha, S., Besra, G., Vandenbroucke-Grauls, C., Bitter, W., van der Sar, A.: Zebrafish embryo screen for mycobacterial genes involved in the initiation of granuloma formation reveals a newly identified ESX-1 component. Disease Model Mechanisms, 526–536 (2011)
3. Nezhinsky, A.E., Verbeek, F.J.: Pattern Recognition for High Throughput Zebrafish Imaging Using Genetic Algorithm Optimization. In: Dijkstra, T.M.H., Tsivtsivadze, E., Marchiori, E., Heskes, T. (eds.) PRIB 2010. LNCS, vol. 6282, pp. 301–312. Springer, Heidelberg (2010)
4. Larios, E., Zhang, Y., Yan, K., Di, Z., LeDévédec, S., Groffen, F., Verbeek, F.J.: Automation in Cytomics: A Modern RDBMS Based Platform for Image Analysis and Management in High-Throughput Screening Experiments. In: He, J., Liu, X., Krupinski, E.A., Xu, G. (eds.) HIS 2012. LNCS, vol. 7231, pp. 76–87. Springer, Heidelberg (2012)
5. LeDévédec, S., Yan, K., de Bont, H., Ghotra, V., Truong, H., Danen, E., Verbeek, F., van de Water, B.: A Systems Microscopy Approach to Understand Cancer Cell Migration and Metastasis. Journal Cellular and Molecular Life Sciences 67(19), 3219–3240 (2010)
6. Yan, K., Larios, E., LeDevedec, S., van de Water, B., Verbeek, F.J.: Automation in Cytomics: Systematic Solution for Image Analysis and Management in High Throughput Sequences. In: Proceedings IEEE Conf. Engineering and Technology (CET 2011), vol. 7 (2011)
7. Hawrylycz, M., Baldock, R.A., Burger, A., Hashikawa, T., Johnson, G.A., Martone, M., Ng, L., Lau, C., Larsen, S.D., Nissanov, J., Puelles, L., Ruffins, S., Verbeek, F., Zaslavsky, I., Boline, J.: Digital Atlasing and Standardization in the Mouse Brain. PLoS Comput. Biol. 7(2), e1001065+ (2011)
8. Belmamoune, M., Potikanond, D., Verbeek, F.: Mining and analysing spatio-temporal patterns of gene expression in an integrative database framework. Journal of Integrative Bioinformatics 7(3)(128), 1–10 (2010)
9. Verbeek, F., Boon, P., Sloetjes, H., van der Velde, R., de Vos, N.: Visualization of complex data sets over Internet: 2D and 3D visualization of the 3D digital atlas of zebrafish development. In: Proc. SPIE 4672, Internet Imaging III, pp. 20–29 (2002)
10. Slob, J., Kallergi, A., Verbeek, F.J.: Observations on Semantic Annotation of Microscope Images for Life Sciences. In: Marshall, M.S., Burger, A., Romano, P., Paschke, A., Splendiani, A. (eds.) SWAT4LS. CEUR Workshop Proceedings, vol. 559, CEUR-WS.org (2009)
11. Sanchez-Faddeev, H., Emmerich, M.T., Verbeek, F.J., Henry, A.H., Grimshaw, S., Spaink, H.P., van Vlijmen, H.W., Bender, A.: Using Multiobjective Optimization and Energy Minimization to Design an Isoform-Selective Ligand of the 14-3-3 Protein. In: Margaria, T., Steffen, B. (eds.) ISoLA 2012, Part II. LNCS, vol. 7610, pp. 12–24. Springer, Heidelberg (2012)
12. Yan, K., Verbeek, F.J.: Segmentation for High-throughput Image Analysis: Watershed Masked Clustering. In: Margaria, T., Steffen, B. (eds.) ISoLA 2012, Part II. LNCS, vol. 7610, pp. 25–41. Springer, Heidelberg (2012)

13. Nezhinsky, A.E., Verbeek, F.J.: Efficient and Robust Shape Retrieval from Deformable Templates. In: Margaria, T., Steffen, B. (eds.) ISoLA 2012, Part II. LNCS, vol. 7610, pp. 42–55. Springer, Heidelberg (2012)
14. Taylor, I.: Workflows for E-Science: Scientific Workflows for Grids. Springer (2007)
15. Wikipedia: Bioinformatics workflow management systems — Wikipedia, The Free Encyclopedia (2012) (Online; last accessed June 25, 2012)
16. Kok, J.N., Lamprecht, A.-L., Wilkinson, M.D.: Tools in Scientific Workflow Composition. In: Margaria, T., Steffen, B. (eds.) ISoLA 2010, Part I. LNCS, vol. 6415, pp. 258–260. Springer, Heidelberg (2010)
17. Chen, L., Shadbolt, N.R., Goble, C.A., Tao, F., Cox, S.J., Puleston, C., Smart, P.R.: Towards a Knowledge-Based Approach to Semantic Service Composition. In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 319–334. Springer, Heidelberg (2003)
18. Lord, P., Bechhofer, S., Wilkinson, M.D., Schiltz, G., Gessler, D., Hull, D., Goble, C.A., Stein, L.: Applying Semantic Web Services to Bioinformatics: Experiences Gained, Lessons Learnt. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 350–364. Springer, Heidelberg (2004)
19. Lamprecht, A.L., Margaria, T., Steffen, B.: Bio-jETI: a framework for semantics-based service composition. BMC Bioinformatics 10(suppl. 10), S8 (2009)
20. Lamprecht, A.L., Naujokat, S., Margaria, T., Steffen, B.: Semantics-based composition of EMBOSS services. Biomedical Semantics 2(suppl. 1), S5 (2011)
21. Lamprecht, A.L., Naujokat, S., Steffen, B., Margaria, T.: Constraint-Guided Workflow Composition Based on the EDAM Ontology. In: Burger, A., Marshall, M.S., Romano, P., Paschke, A., Splendiani, A. (eds.) Proceedings of the 3rd Workshop on Semantic Web Applications and Tools for Life Sciences (SWAT4LS 2010), vol. 698, CEUR Workshop Proceedings (December 2010)
22. Martín-Requena, V., Ríos, J., García, M., Ramírez, S., Trelles, O.: jORCA: easily integrating bioinformatics Web Services. Bioinformatics 26(4), 553–559 (2010)
23. Karlsson, J., Martín-Requena, V., Ríos, J., Trelles, O.: Workflow Composition and Enactment Using jORCA. In: Margaria, T., Steffen, B. (eds.) ISoLA 2010, Part I. LNCS, vol. 6415, pp. 328–339. Springer, Heidelberg (2010)
24. Wilkinson, M.D., Vandervalk, B., McCarthy, L.: SADI Semantic Web Services - 'cause you can't always GET what you want! In: Proceedings of the IEEE Services Computing Conference, APSCC 2009, December 7-11, pp. 13–18. IEEE Asia-Pacific, Singapore (2009)
25. Vandervalk, B.P., McCarthy, E.L., Wilkinson, M.D.: SHARE: A Semantic Web Query Engine for Bioinformatics. In: Gómez-Pérez, A., Yu, Y., Ding, Y. (eds.) ASWC 2009. LNCS, vol. 5926, pp. 367–369. Springer, Heidelberg (2009)
26. Gil, Y., Ratnakar, V., Deelman, E., Mehta, G., Kim, J.: Wings for Pegasus: creating large-scale scientific applications using semantic representations of computational workflows. In: Proceedings of the 19th National Conference on Innovative Applications of Artificial Intelligence, vol. 2, pp. 1767–1774. AAAI Press (2007)
27. Deelman, E., Singh, G., Hui Su, M., Blythe, J., Gil, A., Kesselman, C., Mehta, G., Vahi, K., Berriman, G.B., Good, J., Laity, A., Jacob, J.C., Katz, D.S.: Pegasus: a framework for mapping complex scientific workflows onto distributed systems. Scientific Programming Journal 13, 219–237 (2005)
28. Withers, D., Kawas, E., McCarthy, L., Vandervalk, B., Wilkinson, M.: Semantically-Guided Workflow Construction in Taverna: The SADI and BioMoby Plug-Ins. In: Margaria, T., Steffen, B. (eds.) ISoLA 2010, Part I. LNCS, vol. 6415, pp. 301–312. Springer, Heidelberg (2010)
29. Wood, I., Vandervalk, B., McCarthy, L., Wilkinson, M.D.: OWL-DL Domain Models as Abstract Workflows. In: Margaria, T., Steffen, B. (eds.) ISoLA 2012, Part II. LNCS, vol. 7610, pp. 56–66. Springer, Heidelberg (2012)