Applying Capability Modelling in the Genomics Diagnosis Domain: Lessons Learned

Francisco Valverde[1] and Maria Jose Villanueva[1],

[1] Universitat Politècnica de València, Camino de Vera S/N
46022, Valencia
{fvalverde, mvillanueva}@pros.upv.es

Abstract. Because of the evolution of sequencing technologies, Bioinformatics Workflow Management Systems (BWMS) are a popular software for geneticists to describe workflows for analysing genomic data. Although these systems improve development productivity, they are far from being widely accepted by this community. The lack of rigorous conceptual modelling-practices explains the complexity to adapt this genetic analysis software to context changes. In order to face this adaptation issue, we propose using the capability notion as a modelling primitive for providing a sound conceptual background. This paper analyses, from a capability-driven perspective, how daily practices in a bioinformatics SME could be represented as capabilities. From this real scenario, we state current capabilities and explain how they can be supported using current BWMS. As a lessons learned, we discuss how the introductions of capability-driven development could improve their daily work

Keywords: Capability-driven development, Workflow systems, Conceptual modelling

1 Introduction

Current DNA analyses require a complex computer procedure to perform quality results. As some authors claim [1-2], although a lot of commercial and open-source analysis software is available, it is difficult to find a solution that supports all geneticist's requirements. Therefore geneticists have no option but to develop their own custom software. The most common development approach is to assemble and reuse open-source software components (mapping algorithms, data processing utilities, visualizers, file parsers, etc.) using a scripting-oriented language. Due to this practice, there is a huge amount of isolated scripts, frameworks and command-line tools that perform the same functionality with slight differences.

An interesting approach for improving this scenario is the use of bioinformatics workflow management systems (BWMS) to describe their tailored software as workflows made up of software components that manipulate genetic data. Current BWMS are, to some extent, end-user oriented because they provide some guidance for creating experiments, such as visual notations or wizards. However, when they try to use them, BMWS lack of suitable domain abstractions [3]. In addition, changes usually require addressing underlying technological issues. For that reasons, they find BMWS still very complex and they are reluctant to learn programming [4][5] to tailor them to their requirements.

From an Information Systems perspective, we have detected a serious problem that affects the expected quality of those tools: the lack of sound Conceptual Modeling practices. In previous works, we have presented a holistic Conceptual Schema for the Human Genome (CSHG) [6], some applications for genetic analysis based on this CSHG [7], and the first attempts to provide an ontological background based on UFO [8].

In this paper we explore how conceptual modelling practices can improve the design and development of genetic analyses. In particular, our goal is to explore how the notion of capability can help to better understand and better design those genetic analyses. Capability-driven development (CDD) [9] is a novel approach for addressing evolving scenarios like the one we present in this work. From the business perspective, a capability is defined as the ability and capacity that enables an enterprise to achieve a business goal in a certain operational context. From the technical perspective, capability delivery requires dynamic utilization of resources and services in dynamically changing contexts.

The main contribution of this work is to analyse CDD in a domain such challenging and evolving as Genomics. We want to evaluate that applying the capability notion, functional semantics associated to BWMS are more precise, design complexity is reduced and productivity is increased. Our first step is to identify which capabilities geneticists demand from BWMS: following the principles of Action Research [10], we observed geneticists from IMEGEN (Instituto de Medicina Genomica) [11], a Spanish bioinformatics SME, when using their current genetic analysis software to carry out disease diagnoses. From the lessons learned in this real use case, a set of capabilities are identified and modelled using three widely known open-source BWMS.

The rest of the paper is structured as follows: Section 2 describes several approaches that address bioinformatics software development using BWMS. Section 3 explains the IMEGEN geneticists' scenario for genetic disease diagnosis. Section 4 specifies a concrete set of four capabilities detected in the IMEGEN use case. Section 5 evaluates how these four different capabilities are supported by three popular BWMS. Section 6 discusses how CDD can improve the current context, and section 7 states the conclusions and the future work.

2 Related Work

Several authors have also noticed geneticists' difficulties when they use software tools to accomplish their work. On the one hand, Lacroix and Menager [12] carried out an evaluation of BWMS and noticed several points of improvement regarding extensibility, functionality, usability, understandability, scalability, and efficiency. Also, Barker and Hemert [4] performed a comparison between business and scientific workflow technologies and conducted a survey about different BWMS. Both identified several issues about usability, sustainability and tool adoption, however, both evaluations were performed in 2005 and 2007 respectively, and conclusions may not be longer accurate because tools have evolved since then. Looking critically at these works, there is a lack of a rigorous, clear conceptual modelling background.

Cohen, et al. [12] analysed different real workflows from the micro-array data analysis domain created by scientists using current BWMS. Authors discuss why scientists are not yet prepared to use BWMS, and they describe several technological challenges that current BWMS should address regarding reusability, adaptability and usability. Additionally, McPhillips et al. [5] also believe that environments must provide to geneticists some assistance for the design and implementation of workflows. They enumerated several properties that should be satisfied, such as, clarity, well-formedness, predictability, recordability, reportability, reusability, scientific data modelling and automatic optimization. Again, we realized that the point of view of the analysis is basically solution space-oriented (product-oriented), instead of problem space-oriented (conceptual model-driven).

Trying to overcome this lack of conceptual modelling-based approaches, our work contributes to genetic analysis design by: i) proposing the use of the capability notion as basic modelling unit; ii) identifying several capabilities directly associated with geneticists' requirements; and iii) assessing how BWMS support those capabilities. Specifically, we describe four dimensions or properties that are not considered by the aforementioned approaches: the goal KPIs, the context, the capacity and the ability. With this, we introduce a modelling perspective with a more suitable abstract representation of the domain.

3 Illustrative scenario: The genetic disease diagnosis process

IMEGEN [11] is a SME specialized in the diagnosis of hundreds of genetic diseases and several diagnosis techniques. A generic scenario from their genetic diagnosis process is made up of the following steps:

1. Sequencing Phase: The patient's DNA sample is introduced in the sequencing machine with a set of reactives and, as a result, a set of small sequences is obtained.
2. Sequence Treatment Phase: Geneticists build and check the correctness of the DNA sequence from the set of pieces obtained by sequencing machines. To do this, they must import the set of sequences into a software tool.
3. Alignment Phase: The complete sequence is compared against a reference sequence in order to obtain the differences between them. To do this, they must import the reference sequence and the consensus sequence obtained from the previous phase into a software tool.

4. Knowledge Phase: Each difference is characterized as a genetic variation. If a variation has been previously described, additional information related to a genetic disease should be retrieved. To do this, they must search each variation in different online databases.
5. Report Phase: All genetic variations and their associated information are gathered in a report. To do this, they use a spreadsheet to represent the information gathered.

IMEGEN's geneticists adopted a software solution to semi-automate this manual procedure. But with the called Next Generation Sequencing (NGS) technologies [15] this solution was out to date; although conceptually the process remained the same.

As a solution to this handicap, current BWMS aim to provide geneticists, with a platform to create their own workflows, and evolve them accordingly. The use of these environments benefits geneticists because: 1) Their interests, such as financial and temporal requirements, depend only on themselves instead of IT companies; 2) They are able to create exactly what they expect from a software product; and 3) They control the created product in case of any change is required.

However, we detected a common problem: the lack of modelling practices. For instance, the need to compare a sample sequence against a reference sequence is always the same problem, even if the evolution of technology changes continuously. This change is represented in several BWMS with the introduction of a new alignment algorithm as a new modelling primitive. Mixing the stable, conceptual part of the problem, with the more volatile, technology-oriented concrete solution, is being a serious problem shared among different BWMS.

The benefits of using BWMS and models to describe the semantics of a process are clear in this sort of situations, in which neither semantics nor the problem model has changed. Thus, evolution is described as an update model projection into concrete software technologies, such as toolkits, algorithms and information systems, which would be changed according to the situation. The contribution of this work is to use the CDD methodology to describe such processes to be implemented by a BWMS.

4 Capabilities in a genomic analysis scenario

Using the experience gained after the study of the IMEGEN scenario, we identified a list of mandatory capabilities to be supported in their work. According to the metamodel for capability specification introduced in [9], capabilities are described in terms of five main properties:

- Goal: Desired state of affairs that needs to be attained.
- Goal KPI: Key performance indicator (KPI) for monitoring the achievement of a goal.

- Context: The context encompasses the information characterizing the situation in which a business capability should be provided.
- Capacity: Availability of resources, e.g. money, time, personnel, tools, for delivering the capability
- Ability: Level of available competence, were competence is understood as talent, intelligence and disposition, of a subject or enterprise to accomplish a goal.

We use these five properties for describing and characterizing each detected capability. Additionally, for each capability an example is shown to illustrate how the capability is deployed in real practice. Next, the capabilities are described:

4.1 Access to genomic public data sources (C1)

The genetics domain is such a young field that the best way to achieve new genetic discoveries is sharing as much genetic knowledge as possible. As a result, geneticists feed from data stored in different public repositories spread around the web. To use this data, geneticists specify queries to create datasets from several public repositories. As an example, we can consider this query: "Specify a query to retrieve a reference sequence from a gene, whose identifier is "BRCA1" from the NCBI repository"

- **Goal**: For a giving genetic disease, retrieve all the information publicly available on the repositories list.
- **Goal KPI**: Increase the % of detected variations with relevant information for diagnosis.
- **Context**: Public databases are accessible using different sources: restful APIs, relational databases and HTML pages. It is common the inclusion of relevant new databases.
- **Capacity**: a private database management system for storing indexes about previously detected variations and internal data for analysis.
- **Ability**: knowledge about data sources provides relevant and trustful information from the genomic diagnosis point of view.

4.2 Integration with external services (C2):

Bioinformatics community spends a lot of resources in the development of algorithms, toolkits, and web services specialized in the automation of different genetic tasks. Geneticists design their experiments by combining some of these services into a workflow. Hence, geneticists must be provided with a mechanism to design the integration and configuration of external services. Example: *Integrate a custom algorithm that translates a DNA sequence into a RNA sequence*.

- **Goal**: Include into the genetic analysis external services that provide data processing functionalities.
- **Goal KPI**: Increase the number of services currently integrated into the design environment.
- **Context**: External services are available as web services (using SOAP or Rest interfaces) and as command-line utilities. New and updated services are published regularly according to novel research.
- **Capacity**: Computer server for the deployment and integration of external services.
- **Ability**: Knowledge about the genetic tasks to be performed and the functional requirements.

4.3 Conceptual description and management of new genetic data (C3):

One of the main problems of the genetic domain is the lack of widely accepted conceptual models to express genetic data. Geneticists work with different formats that are basically raw text files following a tabular structure. It is common that new formats arise with new sequencing technologies. Hence, geneticists require a mechanism for managing new genetic data representations and extract the relevant information. Example: *Use an entity "Gene" and use its attribute "Gene Name" as a parameter of a genetic task instead of column 3 of the file Gene.fasta*.

- **Goal:** Support the management of the common data formats from sequencing technologies
- **Goal KPI:** Increase the number of supported data formats
- **Context:** Each sequencing technology could potentially provide a specific data format as there is no standard. Data is stored as flat text files that must be firstly parsed to be used by BWMS.
- **Capacity:** Text based processing tools or scripts to transform data files into structured data, i.e inside a database.
- **Ability:** Software programmers with the skills to implement translator modules and text parsers.

4.4 Reporting (C4)

During the execution of their experiments, geneticists need to store all the data transformations and results that are obtained in every step. This information is essential to check the correct execution of the experiment and also, in order to reconstruct the experiment under the same conditions. Hence, geneticists must be able to specify the creation of result reports. Example: *Specify a report that contains the date and the number of Genes obtained by a given query performed through the NCBI database*.

- **Goal:** Create a report with the relevant data from a genetic disease diagnosis.
- **Goal KPI:** Increase the number of results correctly described in the report.
- **Context:** Report information changes according to the disease to be diagnosed and the customer profile.
- **Capacity:** Reporting software for generating doc or pdf files from the results of the genetic analysis.
- **Ability:** Software programmers with the skills to implement the report templates and the import mechanisms

5 Evaluation of capabilities supported by BWMS

From a conceptual modelling perspective, the next step is to show that the modelling of these four capabilities can improve the understanding and the subsequent implementation of genetic analysis. This modelling exercise also provides a useful framework to conceptually compare which BWMS is more suitable for deploying the capability.

Following this reasoning line, we have deployed the presented capabilities into three of the most reference open-source BWMS. Commercial tools have been discarded in this analysis, because they cannot be properly evaluated using trial or evaluation versions. We claim that the discussion reported in this section could be replicated without any significant constraint using other tools that provide a similar functionality. For the evaluation, we have used the genetic disease diagnosis process described in section 3, which was specified as a “main” capability. Using that scenario, we have configured the specific BWMS and deployed a workflow that supports each capability. Table 1 summarizes the support degree of each capability in three levels as: 1) “supported (S)”; 2) “partially supported (PS)”; or 3) “not supported (N)”. Next, the results are detailed for each tool:

Table 1. Capabilities supported by the analyzed BWMS

	Taverna	Galaxy	eBioFlow
C1: Access to Public Data Sources	PS	PS	PS
C2: Integration with External Services	S	PS	N
C3: Conceptual description of genetic data	PS	N	PS
C4: Reporting	PS	PS	N

5.1 Taverna

Taverna is an open-source environment for the design, edition and execution of scientific workflows created by the MyGrid Team [15]. Its main objective is to aid with the definition of in-silico experiments through the integration of web services specially focused on the biological domain. Taverna integrates functionality through myExperiment, a social network to share scientific workflows, and the Biocatalogue [17], a curated catalogue of web services for the life sciences.

Taverna supports C2, because different services specifications, such as WSDL, BeanShell or Biomoby, are supported, and local tools can be integrated using a command line application. It partially supports C1 because it provides a set of services for retrieving data from different biological data sources, but it is only possible to use predefined queries. Additionally, the expressivity of the workflow language only supports the definition of specific software tools, such as "run-MiraProgram" or "runBlastn2seq" instead of high-level bioinformatic tasks, such as "assembly task" or "local alignment task".

It also partially supports C3, because conceptual descriptions of data can be defined using the MOBY-S ontology [18]. However, this ontology is difficult to search, as it contains duplicate entries, poor descriptions and entities not related with the biological domain. Regarding C4, it is partially supported because it is possible to display data and to store results in text files, but custom reports cannot be specified.

5.2 Galaxy

Galaxy is an open-source web-based environment for the execution of biological services created by the Galaxy Team [19]. Its main purpose is to aid geneticists with their data intensive biological research through the definition of web interfaces for biological data retrieval and services execution. With this purpose, it provides different interfaces that access to UCSC [20] and Ensembl [21] databases and toolkits.

Galaxy partially supports C1, as users can only retrieve data through predefined queries provided by the environment' interfaces, for example the web search interface provided by the USCS database. Also C2 is partially supported, because there is a wide array of popular bioinformatics services provided by default. Additionally, Galaxy supports the integration of new services using programming scripts, but this approach is not suitable for geneticists that lack of programming knowledge. Regarding C4, Galaxy outputs data using a mechanism based on HTML templates. However, the usability and expressivity provided with this solution is not enough to consider this capability fully-supported.

5.3 eBioFlow

eBioFlow is an open-source workflow management system to design and execute biological workflows developed in the academic environment as a proof of concept of a series of PhD dissertations [22]. Its main purpose is to improve other BWMS proposals focusing on the support of data provenance, the usability of the user interface and the execution of workflows step by step.

It partially supports C1, as it provides a set of services that execute a specific query against a specific data source, but it is not possible to retrieve specific data properties. Also, C3 is partially supported, as there is a service for creating data structures according to a taxonomy of entities related with the biological and the bioinformatics domain. However, the specification of each data structure is ambiguous, relationships among structures cannot be detected and there is a confusing mixture between biological properties and computer properties. Neither C2 nor C4 are supported by this BWMS.

6 Discussion

Using a conceptual modelling approach based on the notion of capability, we have simplified the deployment of the aforementioned process. One main improvement is that using capabilities specification, workflows are easy to understand to geneticists. They state that capabilities specification are a nice and organized documentation of their process.

Regarding the analysis, the main conclusion is that Taverna is the most complete environment, as it takes into account the aforementioned capabilities and partially supports all of them. However, specific domain knowledge and entities must be introduced to improve the end-user satisfaction. On the contrary, Galaxy provides less capabilities and its main advantage is that data retrieval and local data uploading is easier. Galaxy is specifically designed for the bioinformatics domain, it is more geneticists friendly and has a simpler workflow notation than Taverna. Its main disadvantage is that the workflow language is not as expressive as the provided by Taverna. Finally, eBioFlow provides a good workflow language in terms of expressivity and a user-friendly interface. This tool also includes some interesting features such as different workflow design perspectives, automatic creation of workflow entities from data types or partial workflow execution. But, it lacks of the advanced bioinformatics functionality that Taverna and Galaxy provide.

From this preliminary analysis, we validated that the previous issues (before applying the CDD methodology) were a consequence of modelling workflows using software-oriented concepts and the lack of documented specification. Current BWMS provide software components oriented to the solution domain, i.e, programming frameworks and command-line tools to implement specific bioinformatics functionality. In order to overcome this situation, the application of

CDD is useful to improve the process understanding by end-users. Reasoning in terms of the well-supported concepts of goal, goal KPIs, context, capacity and ability, geneticists can design the expected capabilities instead of designing just workflows. This idea has been discussed with IMEGEN geneticists in order to come up with a better evolution approach. Specifically, we have found a set of improvements regarding the analysed capabilities:

- **Integration with public data sources**: Current environments only support the execution of predefined queries against a single data source at a time. We have proposed the definition of concept models, a modelling construct proposed by CDD, specifically designed to gather all the concepts regarding the information to be retrieved. Hence, geneticists will construct queries using this conceptual model instead of programming for specific APIs or database schemas. Applying a model-to-code transformation approach is also feasible to transform a query against this model to several specific SQL queries that extract data from each biological data source. As a drawback, it will be necessary the inclusion of a mechanism that combines the query responses and expresses the resulting datasets according to the conceptual model.
- **Data import from genetic flat files**: Some environments manage genetic data by means of text files, which can have different formats. For extracting information, users have to indicate which fields inside that file contain the data. Other BWMS provide a conceptualization of genetic data using taxonomies or biological ontologies. However, usually, these conceptualizations are ambiguous or little intuitive, and geneticists are unable to select the correct entity. We have proposed that every piece of data available in the BWMS must be managed as a domain entity. This entity, with a set of properties, represents precisely and unambiguously a concept of the biological or the bioinformatics domain. Following the approach of the previous points, an entity will be instantiated when data is retrieved in the BWMS and it will be available to become an input or output of a task of the workflow.
- **Reporting**: Report generation is constrained to the execution of software components that generate a specific report. We have proposed the definition of a model-oriented mechanism that uses the conceptual model to create personalized reports, where all the domain data available in the environment can be selected to be reflected in the report.

While several issues have been identified in current BWMS, the main problem is the steep learning curve. For geneticists, it is a highly time-consuming task to fully understand and master all the features provided by a BWMS. As a solution, we believe that geneticists should use a business environment where all the technological and low-level details are hidden. This approach addresses the current situation because CDD offers [9] “an approach to business and IT development … to produce solutions that are fit for changing business contexts, while taking the advantage of emerging technology solutions”.

The application of CDD is not a generic approach to address any bioinformatics task, but it is useful in highly evolving processes like the diagnosis of genetic diseases. Using a CDD environment, geneticists and software analysts will be able to define their diagnosis process in an agile way as a set of bioinformatics processes and services. Additionally, process models and patterns defined as a capability can guide the generation of a workflow specification for a BWMS.

7 Conclusions and Future Work

This work discusses the software issues that geneticists must face daily to accomplish their genetic analyses, and the benefits of using the CDD methodology. In order to address the unresolved features, it is discussed how CDD can be useful. The work shows how capabilities can help to put conceptual modelling in practice, in order to come up with a solution that improves current practices. Hence, we encourage the definition of capabilities specifically oriented to geneticists, whose benefits are: 1) they can focus on the experiments specification; 2) they can abstract technological details and programming issues as domain issues; 3) they can avoid the necessity to learn different BWMS specific management and features; and 4) they have flexibility to change and to evolve the processes that represent their research. As a result of this preliminary study, we state that applying a CDD approach IMEGEN genetic analyses are improved. The next steps are to go further in the specification of additional capabilities and to address the deployment of the presented capabilities.

Acknowledgments Thanks to the Instituto de Medicina Genómica (IMEGEN, http://www.imegen.es) for its support and the provision of information. This work has been partially supported by the EU-FP7 funded project no: 611351 CaaS - Capability as a Service in Digital Enterprises.

References

[1] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A.: The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20, 1297-1303 (2010)

[2] Bartlett, J.C., Toms, E.G.: Developing a protocol for bioinformatics analysis: An integrated information behavior and task analysis approach. Journal of the American Society for Information Science and Technology 56, 469-482 (2005)

[3] Ludascher, B., Weske, M., McPhillips, T., Bowers, S.: Scientific Workflows: Business as Usual? In: Dayal, U., Eder, J., Koehler, J., Reijers, H. (eds.) Business Process Management, vol. 5701, pp. 31-47. Springer Berlin / Heidelberg (2009)

[4] Barker, A., van Hemert, J.: Scientific Workflow: A Survey and Research Directions. In: Wyrzykowski, R., Dongarra, J., Karczewski, K., Wasniewski, J. (eds.) Parallel Processing and Applied Mathematics, vol. 4967, pp. 746-753. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)

[5] McPhillips, T., Bowers, S., Zinn, D., Lud\"a, s.B.: Scientific workflow design for mere mortals. Future Gener. Comput. Syst. 25, 541-551 (2009)

[6] Pastor, O. Conceptual Modeling Meets the Human Genome. ER, 2008, 1-11

[7] Villanueva, M., Valverde, F., Levín, A., Pastor Lopez, O.: Diagen: A Model-Driven Framework for Integrating Bioinformatic Tools. In: Nurcan, S. (ed.) IS Olympics: Information Systems in a Diverse World, vol. 107, pp. 49-63. Springer Berlin Heidelberg (2012).

[8] Ferrandis, A. M. M.; Pastor, O. Ló. & Guizzardi, G. Applying the Principles of an Ontology-Based Approach to a Conceptual Schema of Human Genome. ER, 2013, 471-478

[9] Zdravkovic, J., Stirna, J., Henkel, M., Grabis, J.: Modeling Business Capabilities and Context Dependent Delivery by Cloud Services. In: Salinesi, C., Norrie, M., Pastor, Ó. (eds.) Advanced Information Systems Engineering, vol. 7908, pp. 369-383. Springer Berlin Heidel

[10] Wieringa, R. & Morali, A. Technical Action Research as a Validation Method in Information Systems Design Science. DESRIST, 2012, 220-238

[11] IMEGEN: Instituto de Medicina Genómica. “http://www.imegen.es”(2015)

[12] Lacroix, Z., Ménager, H.: Evaluating workflow management systems for bioinformatics. (2005)

[13] Cohen-Boulakia, S., Leser, U.: Search, adapt, and reuse: the future of scientific workflows. SIGMOD Rec. 40, 6-16 (2011)

[14] Shendure, J., Ji, H.: Next-generation DNA sequencing. Nat Biotech 26, 1135-1145 (2008)

[15] Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M.R., Li, P., Oinn, T.: Taverna: A tool for building and running workflows of services. Nucleic Acids Research 34 (Web Server Issue), W729-W732 (2006)

[16] De Roure, D., Goble, C., Bhagat, J., Cruickshank, D., Goderis, A., Michaelides, D., Newman, D.: myExperiment: Defining the Social Virtual Research Environment. pp. 182 -189 %&. (Year)

[17] Bhagat, J., Tanoh, F., Nzuobontane, E., Laurent, T., Orlowski, J., Roos, M., Wolstencroft, K., Aleksejevs, S., Stevens, R., Pettifer, S., Lopez, R., Goble, C.A.: BioCatalogue: a universal catalogue of web services for the life sciences. Nucleic Acids Research 38, W689-W694 (2010)

[18] Wilkinson, M.D., Links, M.: BioMOBY: An open source biological web services proposal. Briefings in Bioinformatics 3, 331-341 (2002)

[19] Goecks, J., Nekrutenko, A., Taylor, J., Team, T.G.: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology 11, R86 (2010)

[20] Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D., Kent, W.J.: The UCSC Genome Browser Database. Nucleic Acids Research 31, 51-54 (2003)

[21] Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., Clamp, M.: The Ensembl genome database project. Nucleic Acids Research 30, 38-41 (2002)

[22] Wassink, I., Ooms, M., Neerincx, P., van der Veer, G., Rauwerda, H., Leunissen, J., Breit, T., Nijholt, A., van der Vet, P.: e-BioFlow: Improving Practical Use of Workflow Systems in Bioinformatics. In: Khuri, S., Lhotska, L., Pisanti, N. (eds.) Information Technology in Bio- and Medical Informatics, ITBAM 2010, vol. 6266, pp. 1-15. Springer Berlin / Heidelberg (2010)