Towards an Ontology for Bioinformatics Research Process

Pietro Crovari[*1][0000−0002−6436−4431], Sara Pidò[*1][0000−0003−1425−1719], and Franca Garzotto[1][0000−0003−4905−7166]

Department of Electronics, Information and Bioengineering, Politecnico di Milano, via Ponzio 34/5, Milan, Italy, 20133 name.surname@polimi.it

Abstract. Next-generation sequencing techniques made possible enormous steps in the sequencing of genomic material. These advancements were not supported by similar progress in the development of tools for extracting knowledge from these data: today interfaces require high Computer Science expertise, being not suitable for most researchers with a biological or clinical background. As a consequence, bioinformatics research is limited by the cognitive barriers introduced by these tools. To overcome this problem, we want to provide an ontology of the research process that can be used as a reference during the development of new tools. To do that, we run a user study to elicit the key conceptual steps of the bioinformatics research process and modelled them using hierarchical task analysis. Then, we show how the resulting process ontology can be exploited to design interfaces that are not only focused on the data involved in the process but also keep into account both the research workflow and the researchers' exploration requirements. Finally, we discuss the implication of this approach on new more usable data intensive bioinformatics tools that keep in consideration both the research workflow and the researchers' requirements. Our work has profound implications on the design of new, accessible bioinformatics tools that can enhance genomic research.

Keywords: Bioinformatics Tertiary Analysis · Hierarchical Task Analysis · Bioinformatics Research Ontology · Conversational Agent · User-Centered Design.

1 Introduction

Due to the large amount of genomic data that has been generated in recent years, storing and processing biological information has created new challenges. In this context, bioinformatics and computational biology have tried to overcome such challenges [9]. Bioinformatics is "the application of computational tools to organize, analyze, understand, visualize, and store information associated with biological macromolecules" [16].

[*] These authors contributed equally to this work

Typically, bioinformatics is subdivided into primary, secondary, and tertiary analysis. The primary data analysis consists of identifying and evaluating raw data, focusing on generating readable sequencing reads (base calling), and ranking the consistency of the basis. The outputs are usually FASTQ files (Illumina). They are inputs of the secondary analysis, that consists of aligning the reads against the human genome and variant calling [22]. Finally, the tertiary analysis is considered the most challenging step since it allows to study the sequencing results. More in detail, it focuses on understanding the raw data using statistical algorithms, machine learning, and data mining ones [18].

Today, there exists many different bioinformatics tools that allow to perform these three types of analysis. The vast majority of these tools, though, are designed keeping in consideration exclusively operational and functional requirements arising from data retrieval and analysis, neglecting the ones that are more typical of Interaction Design and are related to usability and the need to fit with the researcher's profile and the open ended nature of the their goals during the data exploration processes. As a result, most current tools are difficult to use require a significant expertise in Computer Science to enable users exploit data correctly and in the most effective way.

Today, researchers must spend a tremendous cognitive effort in the interaction for using the platform, being distracted from the goal of the interaction, that is the biological interpretation of the data processed.

For these reasons, bioinformatics research should tackle the challenge of creating tools that result usable for the final users [5]. It is therefore important to start thinking about these tools from the researcher's perspective, adopting a user centered perspective since the very early stages of tool design and development. To do that, it is necessary to have a deep understanding of the bioinformatics research process. Only in this way we can have a clear idea of the operational pipeline that software tools aim to support once they have been developed, to have a clear idea of the user requirements and design tools interfaces consistently. To the best of our knowledge, though, no studies focused on the elicitation of such a process.

The work reported in this paper describes how we performed this activity, focusing on the tertiary analysis process and resulting on a conceptual model of it. To do that, we run a user study to gather information on the process, to then conceptualize the model using the hierarchical task analysis framework. Our work brings two major contributions:

1. an ontology, in the form of a hierarchical task tree [27], representing a typical bioinformatics tertiary process, and
2. a concrete example of how the elicited ontology can be exploited to design a usable tool for genomic data retrieval and extraction.

2 State of the Art

The first tools used to perform bioinformatics analysis were mainly based on command lines. Due to the progress in bioinformatics research, many tools have been developed in order to help bioinformaticians during the analysis.

Particularly, it is possible to identify two main types of interfaces used for bioinformatics analysis: the traditional Graphical User Interfaces (GUI) and, recently, the conversational ones have started to be developed. Among all the available GUIs, some worth mentioning are Galaxy [2], OrangeBioLab [8], UCSC Xena [11], Globus Genomics [17] or GenePattern [25]. Galaxy [2] is a scientific workflow, data integration, and data analysis platform focused on the bioinformatics secondary analysis. It provides a quite simple graphical user interface. OrangeBioLab [8] is a data visualization tool that allows to analyse the uploaded data using data mining, machine learning, predictive modeling, feature scoring. UCSC Xena [11] or Globus Genomics [17] are two visual programming interfaces to analyze genomic data. GenePattern [25] is a powerful scientific workflow system that provides access to hundreds of genomic analysis tools to perform bioinformatics analysis through both a web interface or a command line one.

While it is a lot of years that graphical interfaces have been used, conversational interfaces have recently becoming more and more employed, to help bioinformaticians with a simpler and more user-friendly interface. To make some examples, we can mention Ava [14] and Iris [10], two chatbots developed to help data scientists to compose data analytic pipelines. Both are able to build the workflow through the dialogue and transforms it into an executable Jupyter Python notebook. The main difference among them is that Ava is based on a Controlled Natural Language Interface, i.e., the user follows precise directions to construct the workflow, while Iris leaves more freedom to the user.

In the bioinformatics domain, conversational agents have started to be used to retrieve biological data through natural language processing. Some examples are Maggie [21], BioGraphBot [19] and Ok, DNA! [7]. Particularly, Maggie is a conversational interface to extract data from BioCatalog without any support to the user. BioGraphBot is instead a chatbot that allows to translate queries in natural language to queries in Gremlin, a query language, in order to extract biological data from BioGraphDB. Ok DNA! wants to help biologists and clinicians in extracting genomic data without having knowledge of query languages. All these have the final goal to help and improve the bioinformatics process by means of human-computer interaction tools.

To design usable and efficient platforms, it is fundamental to have a clear idea of the tasks to support—[3]. Human-Computer Interaction research spent a considerable effort to produce many frameworks for task model elicitation [1]. The GOMS framework is one of the most known and adopted ones [13]. GOMS have been designed to describe task analysis in User Interaction through the means of four fundamental elements of the interaction: *Goals*, *Operators*, *Methods*, and *Selection Rules*. In the years, many variants have been developed. Keystroke-Level Model (KLM) [6], NGOMSL [15], and CPM-GOMS [12] are some examples of frameworks that try to simplify or extend GOMS. MECANO [23], MOBI-B [24],

TRIDENT [4] and TADEUS [26] are other spread framework; in these cases, though, the models preclude the integration of the knowledge of the user [1].

Among the available framework, a considerable number adopted a tree-based representation for the elicited model. Most of them follow the assumption that tasks are not atomic and individuals but can be decomposed into sub-tasks, therefore originating a hierarchy. This representation has the advantage of facilitating the elicitation process; on top of that, produced trees enable the comparison of different approaches to the support of the same task, both in terms of the types and the number of steps the approaches require. One of the most popular tree-based frameworks is ConcurTaskTree [20], which is capable embeds inside the topology of the model also the temporal dependencies of the tasks. To the scope of this paper, we will use Hierarchical Task Analysis [27], given its suitability for being adopted in user studies.

In order to construct the conceptual model of the bioinformatics tertiary analysis research process, we ran a user study to understand researchers' work routines. We conducted interview sessions that terminated in the construction of a hierarchical task tree. Then, we confronted and integrated the outcomes of the various sessions to create a unique tree, able to generalize the research process of all the individuals.

Population We interviewed eight bioinformatics expert, recruited on a volunteer basis. The population was balanced in gender (4M, 4F) and heterogeneous in the role covered in Academia (Ph.D. students, postdoctoral researchers, and assistant professors). All the volunteers were recruited through emails. No sensitive information of the participants was stored to guarantee the anonymity of the collected data.

Setting The study took place in a room where the participants and the interviewer could sit around a table. Both the interviewed and the researcher had their personal computer in front of them. Due to the current pandemic emergency outbreak, the final interviews were held with the same protocol, but online, through a videoconferencing software.

Protocol The study consisted of a semi-structured individual interview divided into three phases. During the whole process the volunteers could use a virtual whiteboard to help themselves in the process with sketches.

In the beginning, the participants had to describe the steps that constitute their typical research process. No constraints were given about the granularity of the steps, neither on their number. The researcher intervened to ask for a description of the steps or to ask for any clarification. Once the pipeline was formulated, the volunteers were asked to classify the process elements according to their abstraction level. Finally, participants had to create a hierarchical task tree starting from the elicited actions and integrating them with parents and child tasks to complete the hierarchy.

Results All the participants successfully concluded the interview session. Comparing the outcomes of the first phase, we notice that the research flow is similar for most participants. Despite different abstraction levels that the interviewed people adopted, a comparison of the results shows that in the different pipelines, similar actions have the same ordering. In particular, four common macro-phases results: the retrieval of the data, their exploration, the data analysis, and the visualization and validation of the results, both from a computational and biological perspective. On the other hand, different participants focused more on different sections of the workflow, therefore providing complementary perspectives on the tertiary analysis.

During the second phase, the volunteers provided comparable classifications, showing similar perceptions on the abstraction of the various operations.

The final phase resulted in a set of trees with very similar topology. The comparison of the trees created few conflicts, all of them in the deepest nodes of the trees, showing how the researchers implicitly agree on how the research process is carried out. As in the first phase, different areas of the trees were stressed during the interviews.

3 Ontology

From the outcomes of the user study, we create a hierarchical task tree as an ontology to describe the bioinformatics tertiary research process.

The tree-based representation has many advantages. First, the research process can be analyzed at different levels of abstraction according to the need for specificity. Consequently, a unique hierarchical ontology can describe tools that operate at a different level. At the same time, the hierarchy embeds essential information, like the *part of* relationship, useful for the design and the specification of a tool functionalities. For example, if a tool predicates on a task, it must allow the user to perform all the children operations.

To elicit the ontology, we iteratively integrated the trees described by the study participants in a unique structure. When conflicts arose, we opted for the solution adopted by the majority of the participants. In the case of a tie, we asked an expert bioinformatician not involved in the interviews to resolve the conflict, providing his point of view.

Figure 1 illustrates the final tree. Tertiary analysis is composed of four main phases: *defining objectives*, *data extraction*, *data analysis*, and *results analysis*.

Defining objectives is the first step performed by every interviewed person. It is characterized by the definition of the research question, followed by a State of the Art analysis to understand the existing literature and researches on the same topic. In this step, a bioinformatician also wants to understand which are the deliverables needed to answer the research question, such as data, tables, or plots.

After having defined the goal of the research, a bioinformatics analysis proceeds with the data extraction. This phase is, in turn, composed of three parts. The first one is the data retrieval: after having looked at the different public

Fig. 1. Resulting tree that represents the ontology of the typical bioinformatics tertiary analysis.

available biological data, a bioinformatician selects the datasets of interest. The chosen data are explored: first, their format and meaning are studied, then there is a first preprocessing. The data are analysed to assess their quality, and if needed to clean them, removing the noisy or wrong ones, they are normalized according to the most suitable metric. When the user selects more than one dataset, the data needs to be integrated into a single data structure.

The third phase is the core of the process. Data analysis is where the extracted data are studied and passed through the algorithm selected to answer the biological question. First, some preliminary analyses are computed to choose the algorithm, then the data has to be prepared and organized into training and test sets to be passed to the chosen algorithm. The analysis is executed, using a parameter tuning and optimizing it, if needed.

The last step is the analysis of the results. The results of a computational biology analysis are divided into computational and biological ones. The first ones are evaluated through performances and robustness, to then compute on them a comparative analysis and tests. The biological results are instead validated through commonly used analysis, for example, using an enrichment analysis. Often, relevant features are extracted and also they are validated.

Even if we represent the tertiary analysis as a streamlined process, we do not have to forget that this process is a continuous iteration between the phases. Indeed, researchers iteratively refine their hypothesis to draw conclusions that are scientifically significant.

4 An example of Ontology Application

The research process ontology discussed in the previous section facilitates the design of interfaces that can be easily used by bioinformaticians and biologists. Our case study builds a conversational agent that allows us to extract the required genomic data, i.e., the Data Extraction task. As the ontology suggests, the tool must allow users to perform three macro operations, the retrieval of the datasets, the exploration of the retrieved data, and, if necessary, their integration. As a consequence, we design a conversation formed by three main moments, that represent the three macro operations. The same reasoning must be applied iteratively at every part of the conversation, breaking each operation in its sub-tasks, until the leaves of the ontology are reached and mapped to dialogue units.

Following the Data Extraction branch, the resulting agent starts by providing all the public available data and the user decides which datasets to select. Users can apply filtering operations to refine the research. Then, the chatbot allows us to explore the data starting from providing to the user the data meaning and format and proceeding with asking if the user wants to compute some quality assessment, data cleaning and normalization. In particular, for this last part, the agent must ask the user which metrics she would like to use. The agent has also to understand if the user wants to integrate the extracted datasets, if yes, it has to do it.

Figure 2 shows an instance of the conversation. The agent, knowing the underlying process flow, does not limit itself to execute the operations requested by the users, but actively support and guide them through the pipeline. In fact, at every moment the chatbot exactly knows not only what the user is doing in that moment, but also what they will have to do to accomplish their goals.

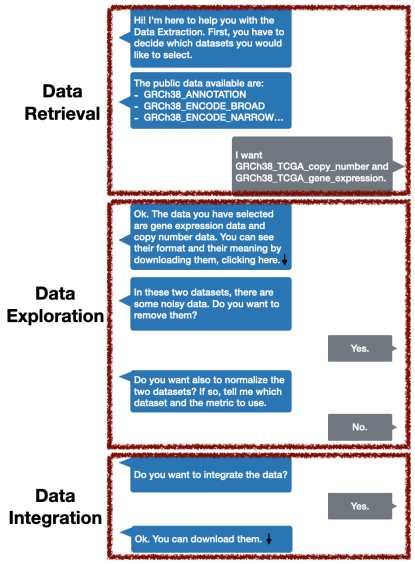

Fig. 2. Example of a dialogue built using our ontology.

To allow users to smoothly pass from an operation to the other, the output of one operation must be the input accepted by the following one. Indeed, the selected datasets of the Data Retrieval phase are passed to the Data Exploration and after the cleaning and normalization to the Data Integration one.

5 Conclusion

In this work, we presented an user centered approach to systematically elicit an ontology of the bioinformatics tertiary analysis process. With our research, we bring two major contributions:

- an ontology focusing on tertiary analysis,
- an example of exploitation of the elicited ontology for the design and development of a new tool,namely, an conversational agent that help researchers in bioinformatics to use advanced data retrieval and analytics functionalities

The elicited ontology has profound implications for the bioinformatics research panorama. Such a model allows tools designers to have a complete overview

of the process in which the tool is inserted, therefore seeing the tools not as stand-alone pieces of software, but as part of a broader pipeline. In this way, the ontology can provide the platform requirements from a functional perspective (such as input data format, expected output, and required operations) and facilitate the integration of tools with complementary capabilities. A tree-based model has the advantage of being usable at every level of abstraction, resulting in more flexibility for its adoption. On top of that, having the pipeline in mind, the designer can truly understand the problem from the users' point of view, therefore having a complete overview of the motivations that push a user to use such a tool.

A clear model of the research process provides a clear way of stating tool capabilities, improving the clarity of the software specifications. At the same time, a standard nomenclature can support in the description of the platform capabilities, removing any ambiguity in the terminology.

Even if promising, our work is not exempt from limitations. The current pandemic emergency prevented us from interviewing bioinformatics researchers on a larger scale. On the other hand, the similarity of the collected responses allowed us to converge to a unified model.

In the future, we aim at continuing the user study to further validate the ontology. In the same way, we want to validate it technically, verifying if it can be used as a to classify the existing tools.

References

1. Abed, M., Tabary, D., Kolski, C.: Using formal specification techniques for the modelling of tasks and generation of hci specifications. The handbook of task analysis for human computer interaction pp. 503–529 (2003)
2. Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A., et al.: The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic acids research **46**(W1), W537–W544 (2018)
3. Benyon, D., Murray, D.: Applying user modeling to human-computer interaction design. Artificial Intelligence Review **7**(3-4), 199–225 (1993)
4. Bodar, F., Hennebert, A.M., Leheureux, J.M., Provot, I., Vanderdonckt, J., Zucchinetti, G.: Key activities for a development methodology of interactive applications. In: Critical Issues in User Interface Systems Engineering, pp. 109–134. Springer (1996)
5. Bolchini, D., Finkelstein, A., Perrone, V., Nagl, S.: Better bioinformatics through usability analysis. Bioinformatics **25**(3), 406–412 (2009)
6. Card, S.K., Moran, T.P., Newell, A.: The keystroke-level model for user performance time with interactive systems. Communications of the ACM **23**(7), 396–410 (1980)
7. Crovari, P., Catania, F., Pinoli, P., Roytburg, P., Salzar, A., Garzotto, F., Ceri, S.: Ok, dna!: A conversational interface to explore genomic data. In: Proceedings of the 2st International Conference on Conversational User Interfaces (To be published)
8. Demšar, J., Curk, T., Erjavec, A., Črt Gorup, Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L.,

Žagar, L., Žbontar, J., Žitnik, M., Zupan, B.: Orange: Data mining toolbox in python. Journal of Machine Learning Research **14**, 2349–2353 (2013), http://jmlr.org/papers/v14/demsar13a.html
9. Diniz, W., Canduri, F.: Bioinformatics: an overview and its applications. Genet Mol Res **16**(1) (2017)
10. Fast, E., Chen, B., Mendelsohn, J., Bassen, J., Bernstein, M.S.: Iris: A conversational agent for complex tasks. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–12 (2018)
11. Goldman, M.J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A.N., et al.: Visualizing and interpreting cancer genomics data via the xena platform. Nature Biotechnology pp. 1–4 (2020)
12. John, B.E., Gray, W.D.: Cpm-goms: an analysis method for tasks with parallel activities. In: Conference companion on Human factors in computing systems. pp. 393–394 (1995)
13. John, B.E., Kieras, D.E.: The goms family of user interface analysis techniques: Comparison and contrast. ACM Transactions on Computer-Human Interaction (TOCHI) **3**(4), 320–351 (1996)
14. John, R.J.L., Potti, N., Patel, J.M.: Ava: From data to insights through conversations. In: CIDR (2017)
15. Kieras, D.: A guide to goms model usability evaluation using ngomsl. In: Handbook of human-computer interaction, pp. 733–766. Elsevier (1997)
16. Luscombe, N.M., Greenbaum, D., Gerstein, M.: What is bioinformatics? a proposed definition and overview of the field. Methods of information in medicine **40**(04), 346–358 (2001)
17. Madduri, R.K., Sulakhe, D., Lacinski, L., Liu, B., Rodriguez, A., Chard, K., Dave, U.J., Foster, I.T.: Experiences building globus genomics: a next-generation sequencing analysis service using galaxy, globus, and amazon web services. Concurrency and Computation: Practice and Experience **26**(13), 2266–2279 (2014)
18. Masseroli, M., Canakoglu, A., Pinoli, P., Kaitoua, A., Gulino, A., Horlova, O., Nanni, L., Bernasconi, A., Perna, S., Stamoulakatou, E., et al.: Processing of big heterogeneous genomic datasets for tertiary analysis of next generation sequencing data. Bioinformatics **35**(5), 729–736 (2019)
19. Messina, A., Augello, A., Pilato, G., Rizzo, R.: Biographbot: A conversational assistant for bioinformatics graph databases. In: International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing. pp. 135–146. Springer (2017)
20. Mori, G., Paternò, F., Santoro, C.: Ctte: support for developing and analyzing task models for interactive system design. IEEE Transactions on software engineering **28**(8), 797–813 (2002)
21. Paixão-Côrtes, W.R., Paixão-Côrtes, V.S.M., Ellwanger, C., de Souza, O.N.: Development and usability evaluation of a prototype conversational interface for biological information retrieval via bioinformatics. In: International Conference on Human-Computer Interaction. pp. 575–593. Springer (2019)
22. Pereira, R., Oliveira, J., Sousa, M.: Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. Journal of clinical medicine **9**(1), 132 (2020)
23. Puerta, A.R.: The mecano project: enabling user-task automation during interface development. In: Proceedings of AAAI. vol. 96, pp. 117–121 (1996)
24. Puerta, A.R., Maulsby, D.: Management of interface design knowledge with mobid. In: Proceedings of the 2nd international conference on Intelligent user interfaces. pp. 249–252 (1997)

25. Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., Mesirov, J.P.: Genepattern 2.0. Nature genetics **38**(5), 500–501 (2006)
26. Schlungbaum, E.: Support of task-based user interface design in tadeus. Universitat Rostock (1998)
27. Stanton, N.A.: Hierarchical task analysis: Developments, applications, and extensions. Applied ergonomics **37**(1), 55–79 (2006)