OK, DNA!: A Conversational Interface to Explore Genomic Data

Pietro Crovari
Fabio Catania
Pietro Pinoli[*]
<name>.<surname>@polimi.it

Philipp Roytburg[†]
Asier Salzar[‡]
philipp.roytburg@rwth-aachen.de
asiers49@opendeusto.es

Franca Garzotto
Stefano Ceri[§]
<name>.<surname>@polimi.it

ABSTRACT

OK, DNA! is an intuitive conversational interface designed to assist biologists and clinicians, who generally have little computer science expertise, in retrieving data from genomic databases. Through this novel technology, complex bioinformatics queries are abstracted to a dialogue interface, that does not require any specific knowledge on querying languages. The power of OK, DNA! is that it exploits the conversation with two goals – as a guide to assist biologists and clinicians during the creation of the entire search process, and as a natural language interface to formulate queries on multiple heterogeneous genomic datasets.

CCS CONCEPTS

• **Applied computing → Bioinformatics**; • **Human-centered computing → Natural language interfaces**; Web-based interaction.

KEYWORDS

Conversational Interface, Data retrieval, Database, Bioinformatics

ACM Reference Format:
Pietro Crovari, Fabio Catania, Pietro Pinoli, Philipp Roytburg, Asier Salzar, Franca Garzotto, and Stefano Ceri. 2020. OK, DNA!: A Conversational Interface to Explore Genomic Data. In *2nd Conference on Conversational User Interfaces (CUI '20), July 22–24, 2020, Bilbao, Spain.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3405755.3406163

1 MOTIVATION AND CONTEXT

Bioinformatics is a multidisciplinary subject that aims to study genomic data (mainly DNA and RNA) through the lens of computer science. In particular, Big Data methods and tools, Artificial Intelligence, Machine Learning, and Data Visualization techniques are used to integrate heterogeneous data to answer complex and biological questions.

The figure of the bioinformatician, though, is intrinsically multidisciplinary. A researcher in this field must have a profound knowledge of the biological mechanisms that rule living beings and the molecular interactions that happen inside cells, a good level of expertise in data analysis and some competence in computer science, in order to retrieve and explore data from genomic databases and process them. This needed amount of heterogeneous preliminary knowledge is often an obstacle to exploit the potential of bioinformatics research [13]. As a consequence, there is the urge to find ways to minimize this barrier. Human-Computer Interaction plays a fundamental role, providing a set of tools that support the design of such tools to maximize their usability ed efficiency [1, 11].

Data Retrieval is a crucial step of the bioiformatics analysis pipeline. Indeed, it is essential to obtain the correct data to work on in the later research phases. The selection of an incomplete or incorrect dataset will unavoidably compromise the analysis and, consequently, the results. Today, data retrieval is done through two major families of applications:

- Visual interfaces: these tools allow data retrieval through a high-level visual interaction. To obtain a good usability, these applications have limited functionalities, and are not flexible in term of query formulation. cBioPortal, UCSC Genome browser are some examples [2, 4, 7];
- Query-based interfaces: These tools give a total expressive freedom to the user, but require a good understanding of querying languages, usually SQL- or graph- based, to be used. GMQL, GROK, and GORpipe are some examples [5, 8, 9].

To fill the gap among these two families of data retrieval tools, Conversational Agents started to be used as usable interfaces to obtain data that are flexible and effective [10, 12], but does not require the knowledge of any specific query language. In bioinformatics domain, some efforts have been done, but these interfaces require the knowledge of the underlying database, or do not support the user during the retrieval process.

We propose OK, DNA!, a web-based written conversational interface, or chatbot, that guides and assist the users in the search of genomic data in multiple heterogeneous biologic datasets. The user can express through the natural language (i.e., English sentences) to work on data in a database, without the necessity of knowing any querying language and, consequently, without worrying of their typically rigid and complex syntax [3, 6]. The CA has been designed in cooperation with bioinformaticians, to tailor the conversation around the final user's requirements. OK, DNA! exploits the dialogue with the conversational interface on two levels: (i) *abstracts* structure of etherogeneous data sources to help the user in retrieving the data in a transparent way, and (ii) *supports* the user during the whole process.

The power of OK, DNA! is in the underlying translation process from natural to query language and vice versa: the conversational

[*] Department of Electronics, Information and Bio-engineering – Politecnico di Milano
[†] RWTH Aachen University
[‡] University of Deusto
[§] Department of Electronics, Information and Bio-engineering – Politecnico di Milano

interface is just a facilitator for query creation to improve the usability of the query tool, and does not reduce the flexibility and expressive power of the query language.

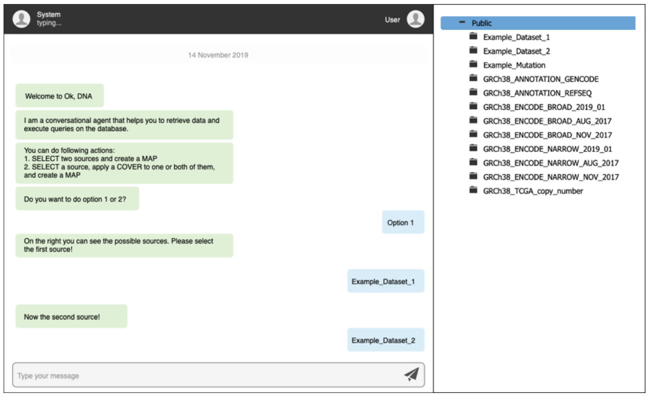
2 OK, DNA!

OK, DNA! is a chat-bot running on a web application that guides people in an intuitive and easy way for the selection of the desired data in a genomic database. Users, in particular biologists and clinicians, do not have to know any querying language to explore the data. In fact, OK,DNA! acts as a proxy translating the information collected through the dialogue into a SQL-like query. In taxonomy, our dialogue system is task-oriented, domain restricted, and proactive. The system has been developed to be used on top of GenoMetric Query Language (GMQL), a declarative querying language that provides in addition to the traditional relational operators, domain-specific operators useful in genomic applications [8]. The development process involved a group of bioinformaticians that where interviewed to elicit the user requirements. A set of semi-structured explorative interviews were held, were researchers described what they would expected from such an interface.

The user accesses OK, DNA! via a web interface, meant to be available on desktop devices. The front-end is connected to a server, that acts as a gateway to the conversational agent engine and as Web-hook to the database. When the user loads the application, the chat page is opened. On the left, there is the core of the application: the chat. Users can interact with the chatbot by writing the messages and pushing the "send" button, or by pressing "enter" on the keyboard. The agent replies guiding the user through the process that ends with the retrieval of the data. The agent actively asks questions to the user to: (i) suggest the next steps, and (ii) assess if there are any problems due to the lack of knowledge of the system.

The chat interface is enhanced by the panel present on the right. In fact, during the interview sessions with the bioiformaticians, two main issues were highlighted. First, they had difficulties in remembering the complex names of the dataset sources, being composed by the composition of multiple acronyms. Second, the bioformaticians wanted a summarization of all the possible commands they can use, with a short explanation associated. As a consequence, we designed a set of tabs thought to help the user in the whole process. In particular, *Actions* tab, if expanded, provides the descriptions of the operations a person can do with OK, DNA!. The *Dataset* tab, instead, presents the list of all the datasets available in the system with a short description. If the user clicks on the name of a table of the dataset, the table name is automatically copied into the message box, such that the user does not have to copy it. Even if the user can expand the two tabs at any time, the system automatically expands them according to the operation that is being done.

Through a dialogue interface, OK, DNA! minimizes the cognitive effort the domain experts have to do to generate the query, making them able to concentrate on what really matters: describing the characteristics of the desired data. The system automatically translates the natural language expression stated (e.g., about the chosen data sets and the desired properties and attributes of the data) by the user into a formal query to be processed on the genomic database, and returns the data corresponding to the query specification. The conversational interface of OK, DNA! makes an innovative usage of conversational technology, exploiting it on three levels, guiding the user during the whole search process, assisting the bioinformatician in all the decisions taken, and allowing the user describing needed data in a transparent way.



14 November 2019

Welcome to Ok, DNA

I am a conversational agent that helps you to retrieve data and execute queries on the database.

You can do following actions:
1. SELECT two sources and create a MAP
2. SELECT a source, apply a COVER to one or both of them, and create a MAP

Do you want to do option 1 or 2?

Option 1

On the right you can see the possible sources. Please select the first source!

Example_Dataset_1

Now the second source!

Example_Dataset_2

− Public
- Example_Dataset_1
- Example_Dataset_2
- Example_Mutation
- GRCh38_ANNOTATION_GENCODE
- GRCh38_ANNOTATION_REFSEQ
- GRCh38_ENCODE_BROAD_2019_01
- GRCh38_ENCODE_BROAD_AUG_2017
- GRCh38_ENCODE_BROAD_NOV_2017
- GRCh38_ENCODE_NARROW_2019_01
- GRCh38_ENCODE_NARROW_AUG_2017
- GRCh38_ENCODE_NARROW_NOV_2017
- GRCh38_TCGA_copy_number

Figure 1: Web interface of OK, DNA!. On the left, there is the conversation pane, whereas on the right available datasets are listed

3 CONCLUSIONS AND FUTURE WORKS

Genomic Research is potentially going to change our lives. Through the analysis of genomic material, we will have a complete understanding of the mechanisms that rule life, what causes the most severe diseases, and, possibly, how to face them. To achieve this result, though, we need to provide bioiformaticians computing tools that empowers them, without introducing further complexity. OK, DNA is a first step in this direction. OK, DNA! is a novel tool aimed to facilitate bioinformaticians in retrieving data from genomic databases, guaranteeing freedom of action, expressiveness of the query language, and reducing the need of knowing querying languages.

To further improve OK, DNA!, we want to face the challenge of tailoring the conversation such that it mimics the epistemological process of research, and the use of data in this discipline. In this way, we are able to provide bioinformaticians a tool that is always more similar to their approach to data search, and not to the underlying querying process.

We worked side-by-side with bioinformaticians during the whole design process of the system. Still, we did not perform a systematic empirical evaluation yet to understand the strengths and weaknesses of such our tool: this activity is the first one in our research agenda. Through the analysis of the interactions of real users with OK, DNA! we will iteratively refine both the formulation of the individual phrases and the flow of the dialogue, to improve the usability and acceptability of the system.

OK, DNA! is a first attempt to face bioinformatics research from an HCI perspective, creating a set of tools that are modeled on the researchers' way of reasoning and not on the requirements of the underlying technology. We believe this experiment is paving the ground to the adoption of new, intelligent interfaces in bioinformatics, opening this recent discipline to radically new interaction paradigms, that will play a key role in genomic research.

REFERENCES

[1] Davide Bolchini, Anthony Finkelstein, Vito Perrone, and Sylvia Nagl. 2008. Better bioinformatics through usability analysis. *Bioinformatics* 25, 3 (2008), 406–412.

[2] Arif Canakoglu, Anna Bernasconi, Andrea Colombo, Marco Masseroli, and Stefano Ceri. 2019. GenoSurf: metadata driven semantic search system for integrated genomic datasets. *Database* 2019 (2019).

[3] Ethan Fast, Binbin Chen, Julia Mendelsohn, Jonathan Bassen, and Michael S Bernstein. 2018. Iris: A conversational agent for complex tasks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[4] Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, et al. 2013. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6, 269 (2013), pl1–pl1.

[5] Hákon Guðbjartsson, Guðmundur Fr Georgsson, Sigurjón A Guðjónsson, Ragnar Þór Valdimarsson, Jóhann H Sigurðsson, Sigmar K Stefánsson, Gísli Másson, Gísli Magnússon, Vilmundur Pálmason, and Kári Stefánsson. 2016. GORpipe: a query tool for working with sequence data based on a Genomic Ordered Relational (GOR) architecture. *Bioinformatics* 32, 20 (2016), 3081–3088.

[6] Rogers Jeffrey Leo John, Navneet Potti, and Jignesh M Patel. 2017. Ava: From Data to Insights Through Conversations.. In *CIDR*.

[7] Donna Karolchik, Robert Baertsch, Mark Diekhans, Terrence S Furey, Angie Hinrichs, YT Lu, Krishna M Roskin, Matthias Schwartz, Charles W Sugnet, Daryl J Thomas, et al. 2003. The UCSC genome browser database. *Nucleic acids research* 31, 1 (2003), 51–54.

[8] Marco Masseroli, Pietro Pinoli, Francesco Venco, Abdulrahman Kaitoua, Vahid Jalili, Fernando Palluzzi, Heiko Muller, and Stefano Ceri. 2015. GenoMetric Query Language: a novel approach to large-scale genomic data management. *Bioinformatics* 31, 12 (2015), 1881–1888.

[9] Kristian Ovaska, Lauri Lyly, Biswajyoti Sahu, Olli A Janne, and Sampsa Hautaniemi. 2012. Genomic region operation kit for flexible processing of deep sequencing data. *IEEE/ACM transactions on computational biology and bioinformatics* 10, 1 (2012), 200–206.

[10] Majdi Owda, Zuhair Bandar, and Keeley Crockett. 2011. Information extraction for SQL query generation in the conversation-based interfaces to relational databases (C-BIRD). In *KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications*. Springer, 44–53.

[11] Katrina Pavelin, Jennifer A Cham, Paula de Matos, Cath Brooksbank, Graham Cameron, and Christoph Steinbeck. 2012. Bioinformatics meets user-centred design: a perspective. *PLoS computational biology* 8, 7 (2012).

[12] Karen Pudner, Keeley A Crockett, and Zuhair Bandar. 2007. An Intelligent Conversational Agent Approach to Extracting Queries from Natural Language.. In *World Congress on Engineering*, Vol. 1. 305.

[13] Shoba Ranganathan. 2005. Bioinformatics education—perspectives and challenges. *PLoS computational biology* 1, 6 (2005), e52.