EMBnet.journal
Bioinformatics in Action

A genomic data mining pipeline for 15 species of the genus *Olea*

Genomic big data hitting the storage bottleneck

NOTCH3 and CADASIL syndrome: a genetic and structural overview

AND MORE...

24
2019

EMBnet.journal
Bioinformatics in Action

Contents

Editorial

The articles published by EMBnet.journal during 2018 are a clear proof of the broad acceptance of Bioinformatics as an important and crucial research subject in the understanding of different aspects in the Life Sciences field.

Bioinformatics based research is not anymore, a science field limited to industrialised countries but is on a very explosive expansion in developing countries. This is shown in the article "Training Workshop on Mycobacterium Whole Genome Sequence Data Analysis organised in Addis Ababa, Ethiopia".

Other important aspects in modern Life Sciences research are the issues of handling the so-called Big Data. Training and knowledge in that field are issues that concern many research projects in Life Sciences and is imperative to work with computer scientists, mathematicians and statisticians to be able to handle and analyse the humungous amounts of data produced by new biotechnological technologies as NGS. In this issue these aspects can be exemplified on a report that shows how these issues can be discussed and how the students can be trained to obtain the necessary skills to fulfil their goals. The article describes the topics, detailed content and timeline of events during the CHARME ATS on Big Data for Life Sciences, which took place at the Technical University of Catalonia (UPC) in Barcelona, September 3-7, 2018.

EMBnet is a global organisation and to emphasize this the EMBnet board decided to organise the 2018 Annual General meeting together with the ISCB and the SoiBio networks (Nov 5 – 9, 2018). The AGM and Conference took place in Vina del Mar, Chile. The Conference was announced as the "Fifth International Society for Computational Biology Latin America, SOIBIO and EMBnet Joint Bioinformatics Conference 2018 (ISCB-LA SOIBIO EMBnet 2018). The Conference was well attended and successful, a report will be published on the 2019 issue.

Last but not least the EMBnet network in 2018 celebrated 30 years since its foundation, on next issue a 30 years chronicle will also be published.

Erik Bongcam-Rudloff
Editor-in-Chief
erik.bongcam@slu.se

http://dx.doi.org/10.14806/ej.24.0.929

EMBnet.journal
Bioinformatics in Action

Schulze et al. (2018) EMBnet.journal 24, e912
http://dx.doi.org/10.14806/ej.24.0.912

Reports

Proceedings of the "Think Tank Hackathon", Big Data Training School for Life Sciences Follow-up, Ljubljana 6th – 7th February 2018

Sabrina K. Schulze[1], Živa Ramšak[2], Yen Hoang[3], Eftim Zdravevski[4], Juliane Pfeil[5]✉, Ariel Duarte-López[6], Uwe Baier[7], Maja Zagorščak[2]

[1] Cell2Fab (Synthetic Biology, Faculty of Biochemistry and Biology), University of Potsdam, Potsdam, Germany

[2] Department of Biotechnology and Systems Biology, National Institute of Biology (NIB), Ljubljana, Slovenia

[3] Department of Signal Transduction, German Rheumatism Research Center, Berlin, Germany

[4] Department of Information Systems, Faculty of Computer Science and Engineering, Sts. Cyril and Methodius University in Skopje, Skopje, Macedonia

[5] Division Molecular Biotechnology and Functional Genomics, Technical University of Applied Sciences, Wildau, Germany

[6] DAMA-UPC, Department of Computer Architecture, Technical University of Catalonia, Barcelona, Spain

[7] Institute of Theoretical Computer Science, Institute of Theoretical Computer Science, Ulm University, Ulm, Germany

Competing interests: SKS none; ŽR none; YH none; EZ none; JP none; ADL none; UB none; MZ none

Abstract

On 6th and 7th February 2018 a Think Tank took place in Ljubljana, Slovenia. It was a follow-up of the "Big Data Training School for Life Sciences" held in Uppsala, Sweden, in September 2017. The focus was on identifying topics of interest and optimising the programme for a forthcoming "Advanced" Big Data Training School for Life Science, that we hope is again supported by the COST Action CHARME (Harmonising standardisation strategies to increase efficiency and competitiveness of European life-science research - CA15110). The Think Tank aimed to go into details of several topics that were - to a degree - covered by the former training school. Likewise, discussions embraced the recent experience of the attendees in light of the new knowledge obtained by the first edition of the training school and how it comes from the perspective of their current and upcoming work. The 2018 training school should strive for and further facilitate optimised applications of Big Data technologies in life sciences. The attendees of this hackathon entirely organised this workshop.

Background and course

In September 2017 the first "Big Data Training School for Life Sciences"[1] as organised by the COST Action CHARME[2] and EMBnet[3] and took place in Uppsala, Sweden (Pfeil *et al.*, 2018a). Although the academic background of young bioinformaticians participating in this training school was somewhat diverse, we all concluded that it is a necessity to learn more about handling Big Data in life sciences. It was denoted how it would be a great deal if additional useful topics could have been described, with ones already presented in Uppsala but shown in more detail. Out of these constructive discussions the idea of additional training came up, and it was proposed to the members of CHARME and EMBnet. This idea was welcomed with interest and approved by the organisers in the form of a short pre-meeting, which was organised by us. As a result, the Think Tank Hackathon[4,5] was organised at the Medical Faculty of the University of Ljubljana, Slovenia on 6th and 7th February 2018 with the support of ELIXIR.SI Node[6].

The hackathon was only open to attendees of the original training school in Uppsala, with eight of them enthusiastically participating (Figure 1). The organisation of this event was made under the leadership of Maja Zagorščak and Živa Ramšak, with the extra help of all the other attendees. As all the participants share the trait *"don't just tell me how to do it, I want to figure it out really"*,

[1] http://astrocyte.com/COST-CHARME/COST-CHARME/Home.html

[2] http://www.cost-charme.eu/

[3] http://www.embnet.org/

[4] http://conferences.nib.si/BigData/

[5] http://www.cost-charme.eu/events/follow-up-training-school

[6] https://www.elixir-europe.org/about-us/who-we-are/nodes/slovenia

Article history
Received: 19 February 2018
Published: 19 April 2018

Schulze et al. (2018) EMBnet.journal 24, e912
http://dx.doi.org/10.14806/ej.24.0.912

Figure 1. Attendees of the Hackathon joined by COST Action CHARME and ELIXIR.SI representatives (image kindly provided by Sabrina Schulze).

the primary focus of the meeting was the exchange of substantial knowledge about leading technologies and *in situ* troubleshooting by solving some real problems.

Before the training school, with the help of a questionnaire regarding group expertise and interests, and taking advantage of various collaboration technologies (*e.g.*, Slack, GitHub and Dropbox), we identified and agreed on several feasible topics, such as *Code, Parallelize, Containerise!, Image analysis with OpenCV - from leaf shape to chords, Machine learning - do it yourself* and *Deep learning without a PhD*.

The attendees, each knowledgeable in their subfield, took the complementary lead on diverse topics and were solely responsible for preparing presentation materials, code examples and tasks for the hackathon. During the Think Tank in Ljubljana, all issues were covered by the active participation of all attendees. Additionally, we brainstormed (hence the name Think Tank) and all together collaborated to the drafting of a proposal for a potential advanced training school.

The first day was opened by Maja Zagorščak with some introductory words about the Think Tank Hackathon goals, organisation and schedule. Her introduction was followed by Eftim Zdravevski, on the topic of Docker technology and its Big Data applications. He discussed parallelisation using Apache Spark™ (Zaharia *et al.*, 2010) and demonstrated k-means clustering and classification examples using the MLLIB (Meng *et al.*, 2016), Spark's scalable machine learning library. Therefore, he guided us through the process of algorithm parallelisation using the data-parallel computing paradigm, while discussing its performance and computational acceleration, using an approach similar to the ones presented Zdravevski et al., 2015a and in Zdravevski et al., 2015b.

Juliane Pfeil filled the second part of the day with a lecture on OpenCV, a computer vision library. She guided us through basic concepts of image analysis (*i.e.*, noise reduction, segmentation, structural pattern recognition and prediction) and demonstrated which functions and parameters are optimal for specific predetermined tasks (for an example see Pfeil *et al.*, 2018b). Lastly, she showed an interesting example of how detection of the centre of mass and radial function application can be used to transform bioimages into chords. All source code developed during the Think Tank hackathon is available online at the GitHub repository[7] Big Data Think Tank.

The second day began with *Deep learning without a PhD*, where Yen Hoang introduced Google's TensorFlow open-source library, which among others is also used for

[7]https://github.com/zagorGit/BigDataThinkTank

Schulze et al. (2018) EMBnet.journal 24, e912
http://dx.doi.org/10.14806/ej.24.0.912

Figure 2. The attendees during the deep learning session (image kindly provided by Sabrina Schulze).

machine learning applications (*e.g.*, neural networks). She guided us through various detailed presentations by Martin Görner (GitHub repository: TensorFlow MNIST Tutorial[8]) and moderated the discussion about neural network theory and the underlying code to increase our comprehension. For this section, the entire infrastructure was set up in advance by the ELIXIR.SI's member Andrej Kaštrin (users, applications, dependencies), which resulted in no time loss for the participants once the workshop started. Additionally, this led to an informal discussion over Docker usability in R, to ease issues with transparent data sourcing, availability and traceability of published results. These concerns are also of vital importance considering principles of FAIR (findable, accessible, interoperable, reusable) scientific data management (Wilkinson *et al.*, 2016). The presentations were concluded by an impromptu presentation given by Uwe Baier, regarding his ongoing work in data compression with examples. Among them the magic of the Burrows–Wheeler transform (BWT; Baier *et. al.*, 2016). Aside from standard BWT application in text compression, it was shown advantageous in many next-generation sequencing (NGS) alignment algorithms in the context of memory reduction.

[8]https://github.com/martin-gorner/tensorflow-mnist-tutorial

Both days were rounded off by discussions and brainstorming sessions regarding the organisation and structure of a potential follow-up training school to be organised later in 2018. We discussed about a possible organising committee, suitable dates and location of the training school; all these details will be formulated into an official proposal that will be submitted to CHARME for approval.

The "Advanced" Big Data Training School for Life Sciences was proposed to last 5 days, with various topics suggested to be focused on: feature extraction, deep learning, with both free of charge and payable software tools and platforms (*e.g.*, services offered by Microsoft Azure), or the backbone computational infrastructure required for tasks in Big Data analysis. All of these topics represent attractive options, given the current developments in these fields, especially if the knowledge learned is connected to real-world problems. It was also agreed to invite experts on these topics, which would explain the theory behind these applications and assist during the practical parts of the workshop. These ideas were finalised on the evening of the second Think Tank day, where preliminary time slots were agreed upon, including time for theory sessions and practical parts. The majority of practical sessions were proposed to be group tasks, where the participants would jointly work on some pre-selected datasets. The results of practical sessions

would then be presented to the remaining groups, thus optimising the time constraints of such tasks, and at the same time allowing for a continual learning improvement via the collaboration of the workshop's attendants. On that note, social activities were also taken into account in the schedule, again fostering future cooperation of participants. We finally decided that participants of the previous training school should be preferred as attendees for the advanced training school. The remaining slots should then be filled by other applicants, provided they are eligible given their background.

At the end of the Think Tank, a biweekly journal club was proposed, where related papers will be read beforehand and discussed in a one-hour session. In the beginning, this should be held only with the participants at the Uppsala training school. After some routine, collaborators could join the video conference via Skype or similar. This club would allow a progressive alignment of the participants' level regarding machine learning and deep learning elements.

Summary

This Think Tank event turned out in two exciting and productive days. We established the knowledge base on four different topics and identified topics for a possible advanced Big Data training school. The attendees who gave presentations on the hackathon noticed how it is sometimes easier to suggest improvements than to apply them in practice. However, the intellectual freedom in the organisation of future events gave us the opportunity to think outside the box. We consider the results of these days as a useful starting point leading towards a proposal for the Advanced Training School. Additionally, the idea of a biweekly journal club would help to bring all participants on the same level and to strengthen the network.

Fortunately, there was also some time scheduled for social events and networking. One was on the evening before the first session, where we spent some time playing abstract games in a board-game tavern, thus getting acquainted with other participants mindsets in a relaxed and casual manner. This idea was well accepted. Therefore, we suggest that this type of social activity is also included in other follow-up events. It offers a unique opportunity of catching up or getting to know potential new participants better and in advance to the starting of the official training event. The other social event we attended was on the evening of the first day. It was the dinner with CHARME WG5's meeting attendants. Some of them were in Uppsala the last September and to meet them in Ljubljana helped us to set up foundations for future collaborations and ideas on how to get the wheels in motion with the organisation of the "Advanced Training School".

Acknowledgement

We would like to thank ELIXIR.si for allocating rooms for us at the University of Ljubljana (Faculty of Medicine) and for infrastructural support. We also thank the COST Action CHARME (CA15110) that is supported by the EU framework program H2020, for providing us with the necessary funding for to organise and attend this event. A particular note of acknowledgement and gratitude goes to Maja Zagorščak and Živa Ramšak for preparing the collaboration tools and environments, for to schedule and leading the organisation of the event in the highest of standards.

References

1. Baier U, Beller T, Ohlebusch E (2016) Graphical pan-genome analysis with compressed suffix trees and the Burrows–Wheeler transform, Bioinformatics **32**, 497–504. http://dx.doi.org/10.1093/bioinformatics/btv603
2. Meng X, Bradley J, Yavuz B, Sparks E, Venkataraman S *et al.* (2016) Mllib: Machine learning in apache spark. J Mach Learn Res **17**, 1–7. http://jmlr.org/papers/v17/15-237.html
3. Pfeil J, Schulze SK, Zdravevski E, Hoang Y (2018a) Report on the "Big Data Training School for Life Sciences", 18-22 September 2017, Uppsala, Sweden. EMBnet.journal **23**, e905. http://dx.doi.org/10.14806/ej.23.0.905
4. Pfeil J, Frohme M, Schulze K (2018b). Mobile Microscopy and Automated Image Analysis: The ease of cell counting and classification. Optik & Photonik **13**, 36-39. http://dx.doi.org/10.1002/opph.201800002
5. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3**. http://dx.doi.org/10.1038/sdata.2016.18
6. Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I (2010) Spark: Cluster computing with working sets. Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing (HotCloud'10). http://www.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-53.html
7. Zdravevski E, Lameski P, Kulakov A, Jakimovski B, Filiposka S and Trajanov D (2015a) Feature ranking based on information gain for large classification problems with mapreduce. IEEE, Proceedings of the Trustcom/BigDataSE/ISPA. http://dx.doi.org/10.1109/Trustcom.2015.580
8. Zdravevski E, Lameski P, Kulakov A, Filiposka S, Trajanov D and Jakimovski B (2015b) Parallel computation of information gain using Hadoop and MapReduce. IEEE, Proceedings of the FedCSIS. http://dx.doi.org/10.15439/2015F89

EMBnet.journal
Bioinformatics in Action

Hirutu et al. (2019) EMBnet.journal 24, e920
http://dx.doi.org/10.14806/ej.24.0.920

Reports

Training workshop on *Mycobacterium* whole-genome sequence data analysis

Yonas Kassahun Hirutu[1]✉, Mesert D Bayeleygne[2], Adey F Desta[3], Tewodros Tariku[1], Markos Abebe[1]

[1]Armauer Hansen Research Institute (AHRI), Ethiopia
[2]Ethiopian Biotechnology Institute (EBTI), Ethiopia
[3]Addis Ababa University (AAU), Ethiopia

Competing interests: YKH none; MDB none; AFD none; TT none; MA none

Abstract

Basic bioinformatics training workshop conducted at Armauer Hansen Research Institute (AHRI), Addis Ababa, Ethiopia. This report describes a bioinformatics training initiative started at AHRI aiming to support life science researchers and postgraduates in handling next-generation sequence data.

Introduction

Institutional initiatives strengthening capacity at Armauer Hansen Research Institute (AHRI) focuses on building a bioinformatics training centre, a next-generation sequencing (NGS) facility and a computing platform to support researchers and postgraduates in Ethiopia. The faculties benefit both ongoing and new project initiatives such as on pathogen evolution, virulence determinants and epidemiology of important pathogens, including *M. tuberculosis*.

The workshop aimed at delivering a practical introduction to NGS data analysis of the *M. tuberculosis* complex (MTBC) genome.

Every workshop day included 40 minutes presentation, three hours hands-on practical and 20 minutes discussion. The presentation topics were on next-generation sequencing technologies, examples of

Article history
Received: 15 January 2019
Accepted: 21 September 2019
Published: 15 October 2019

sequence data file formats and stepwise description of each NGS data analysis bioinformatics workflow.

The workshop was held at the Armauer Hansen Research Institute (AHRI), Bioinformatics unit, 24-28 November 2018. There were 12 participants from health and biotechnology research institutes in Ethiopia (Figure 1).

Implementation of the workshop

The workshop started with a lecture on Linux and command-line tools. AHRI's computing room with Bio-Linux workstations was used for demonstrations and practical sessions. Each workstation was assigned for two participants during the workshop. The first-day activity was focused on giving adequate practice time and guidance on command line environment.

Illumina paired-end reads of three Ethiopian *M. tuberculosis* complex (MTBC) strains were selected for the practical sessions. The strains were subsets of fastq files stored at AHRI for which NGS runs were outsourced as part of previous studies in the institute. Nine MTBC genome sequences were downloaded from NCBI[1]. *M. tuberculosis H37Rv* (NC_000962) was used as a reference genome for mapping, variant calling and annotation of MTBC strains.

The practical workflow analysis included quality control of reads with the FastQC tool (Andrews, 2010). Duplicates were removed using a custom Python script. The reads were aligned by Burrows-Wheeler Alignment Tool (BWA) with default parameters for each sample (Li and Durbin, 2009). The aligned results were piped to SAMtools for the conversion of BWA output format to BAM format (Li *et al.*, 2009). Finally, a consensus variant file (VCF) was generated with BCFtools for subsequent SNP analysis (Danecek *et al.*, 2011). The variants identified with SAMtools/ BCFtools were inspected using the Integrative Genomics Viewer (IGV) (Thorvaldsdottir *et al.*, 2013).The reads were also analysed with the MTBseq tool which is a comprehensive pipeline for whole-genome sequence analysis of M. tuberculosis complex isolates with a full workflow functionality implemented in Perl modules (Kohl *et al.*, 2018).

MTBseq analysis tool is designed for MTBC NGS data for reference mapping, variant detection, variant annotation for drug resistance and comparative analysis among the samples. MTBseq tool was used to demonstrate additional approaches and summarise a consolidated view of all the steps followed during the practical workshop. MTBseq annotation outputs of amino acid changes for a known association to antibiotic resistance were used to characterise the samples. Variants information file was used to construct phylogenetic tree using FastTree (Price *et al.*, 2010) (Figure 2).

Finally, a presentation was made on NGS analysis of metagenomics data. The workshop participants had an opportunity to reflect on their training experience and comment on different aspects. Participants' feedback highlighted the need to improve computational power of the workstations and address fluctuations of internet connectivity. The participants also expressed a satisfactory level for acquiring basic skills in handling NGS data.

Figure 2. Workshop participants during the hands-on session.

In conclusion, the training coordinators and leaders of the institute have appreciated the success of the workshop. The trainers acknowledged the participants for their active engagement throughout the training week.

[1]fttp://ftp.ncbi.nih.gov/genomes/Bacteria/

Acknowledgements

The workshop was supported by training and research directorate at AHRI with a funding source from the Swedish International Development Cooperation Agency (SIDA). The authors would also like to acknowledge support from the African Academy of Sciences (AESA) through Tuberculosis Genetics Network in Africa (TBGENAfrica) research project at AHRI.

References

1. Andrews S. (2010)FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed 20 October 2018)
2. Li H and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform.Bioinformatics **25**: 1754-1760 http://dx.doi.org/10.1093/bioinformatics/btp324
3. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078-2079. http://dx.doi.org/10.1093/bioinformatics/btp352
4. Danecek P, Auton A, Abecasis G, Albers CA, Banks E et al. (2011) The variant call format and VCFtools. Bioinformatics **27**: 2156-2158. http://dx.doi.org/10.1093/bioinformatics/btr330
5. Thorvaldsdottir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. BriefBioinform. **14**: 178-192. http://dx.doi.org/10.1093/bib/bbs017
6. Kohl TA, Utpatel C, Schleusener V, De Filippo MR, Beckert P, Cirillo DM, Niemann S. (2018) MTBseq: a comprehensive pipeline for whole genome sequence analysis of Mycobacterium tuberculosis complex isolates. PeerJ6:e5895.http://dx.doi.org/10.7717/peerj.5895
7. Price MN, Dehal PS, Arkin AP (2010) FastTree 2 - Approximately Maximum-Likelihood Trees for Large Alignments. PLoSONE5:e9490. http://dx.doi.org/10.1371/journal.pone.0009490

EMBnet.journal
Bioinformatics in Action

Hoang et al. (2019) EMBnet.journal 24, e917
http://dx.doi.org/10.14806/ej.24.0.917

The CHARME "Advanced Big Data Training School for Life Sciences": an example of good practices for training on current bioinformatics challenges

Yen Hoang[1], Juliane Pfeil[2], Maja Zagorščak[3]✉, Axel Y. A. Thieffry[4], Eftim Zdravevski[5], Živa Ramšak[3], Petre Lameski[5], Sabrina K. Schulze[6], Eleni Papakonstantinou[7], Louis Papageorgiou[7,8], Tarry Singh[9], Ariel Duarte-López[10], Marta Pérez-Casany[11]

[1] German Rheumatism Research Center Berlin, A Leibniz Institute, Germany

[2] Division Molecular Biotechnology and Functional Genomics, Technical University of Applied Sciences, Wildau, Germany

[3] Department of Biotechnology and Systems Biology, National Institute of Biology (NIB), Ljubljana, Slovenia

[4] Department of Biology, Biotech Research & Innovation Centre, University of Copenhagen, Denmark

[5] Faculty of Computer Science and Engineering, Sts. Cyril and Methodius University in Skopje, Skopje, Macedonia

[6] Institute of Biochemistry and Biology, Cell2Fab, University of Potsdam, Potsdam, Germany

[7] Genetics and Computational Biology Group, Laboratory of Genetics, Department of Biotechnology, Agricultural University of Athens, Greece

[8] Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Greece

[9] CEO and AI Neuroscience Researcher of deepkapha.ai Labs, Deepkapha.ai, Amsterdam, Netherlands

[10] Department of Computer Architecture, Technical University of Catalonia, Barcelona, Spain

[11] Department of Statistics and OR, Technical University of Catalonia, Barcelona, Spain

Competing interests: YH none; JP none; MZ none; AYAT none; EZ none; ŽR none; PL none; SKS none; EP none; LP none; TS none; ADL none; MPC none

Abstract

The CHARME "Advanced Big Data Training School for Life Sciences" took place during 3-7 September 2018, at the Campus Nord of the Technical University of Catalonia (UPC) in Barcelona (ES). The school was organised by the Data Management Group (DAMA) of the UPC in collaboration with EMBnet as a follow-up of the first CHARME-EMBnet "Big Data Training School for Life Sciences", held in Uppsala, Sweden, in September 2017. The learning objectives of the school were defined and agreed during the CHARME "Think Tank Hackathon" that was held in Ljubljana, Slovenia, in February 2018.
This article explains in detail the step forward organisation of the training school, the covered contents and the interaction/relationships that thanks to this school have been established between the trainees, the trainers and the organisers.

Introduction

The "Advanced Big Data Training School for Life Sciences"[1] (ATS-LS) was a joint activity of the EU COST Action CHARME[2] (Harmonising standardisation strategies to increase efficiency and competitiveness of European life-science research - CA15110) and EMBnet[3] (The Global Bioinformatics Network), the former providing financial and the latter organisational support.

After the first "Big Data Training School for Life Sciences", held in September 2017 in Uppsala, Sweden (Pfeil *et al.*, 2018), individual participants showed interest in organising a more advanced version of the event. This idea, supported by the COST Action CHARME's management committee members, who collaborated to the organisation of the first school edition, was finalised at the CHARME "Think Tank Hackathon" that took place in Ljubljana, Slovenia (Schulze *et al.*, 2018) in February 2018.

Out of 48 received applications, 23 were selected (including five local participants) to attend the school and in the end 16 were selected by the scientific committee[4] to be awarded a stipend (CHARME grant). This selection

[1] http://dama-advancedbigdataschool.ac.upc.edu/

[2] http://www.cost-charme.eu/

[3] http://www.embnet.org/

[4] http://dama-advancedbigdataschool.ac.upc.edu/apply

Article history
Received: 02 October 2018
Accepted: 07 January 2019
Published: 05 February 2019

was based on specific criteria that were: i) the prior knowledge of programming and statistical methods in life sciences; ii) the participation of candidates at the first Big Data training school in Sweden; iii) the research background, and iv) candidate's motivation. Additional criteria were ethnicity, gender and national affiliation that were considered for keeping an overall balanced diversity.

The training school officially commenced on the 3rd of September, with opening words and a welcoming introduction from the vice-Chair of the CHARME action, Prof. Erik Bongcam-Rudloff, and the local organisers, Prof. Marta Pérez-Casany and Ariel Duarte-López. The programme encompassed three main frameworks composed of interactive lectures accompanied by hands-on sessions and roundtables in the categories of:

1. Feature Selection (FS)
2. Machine Learning (ML) using the Microsoft Azure platform and Virtual Machines (VMs)
3. Deep Learning (DL) and Artificial Intelligence (AI) applied to life science and healthcare data, leveraging the power of graphics processing units (GPUs)

The rest of this article is organised in six sections. The first three sections describe the organisation and scope of special lectures whereas the other ones present other essential aspects of the school, such as the "Trainers' Perspectives", a "Miscellaneous" section, that includes diverse information and a "Summary".

The section "Availability of data and materials" provides links to all materials, including presentations, code repositories, and datasets.

Topic I: Feature Selection

Andrzej Janusz, Assistant Professor, Institute of Informatics, University of Warsaw, Poland, opened the ATS with a one-day session, exhaustively introducing the topic of Feature Selection (FS) that covered:

- the terminology of feature and its synonyms;
- how FS reduces complexity, improves interpretability either for explicative as well as predictive models and generally contributes to having more robust models;
- basic FS tools that are used in linear models as the t-statistic;
- common misconceptions regarding FS;
- importance of understanding problem space, relevance measure, redundancy, and optimality;
- underlying assumptions, checklist and decision making, as well as the concepts of scalability, stability, generalisation, and overfitting;
- curse of dimensionality; the difference between dimensionality reduction methods and FS;
- categorisations of FS techniques from a perspective of data type, learning task and selection strategy; filter methods (univariate and multivariate), wrapper methods, embedded methods;
- decision reducts, bottom-up vs top-down approaches;
- optimisation algorithms, iterative methods, metaheuristics;
- discussion about the value and importance of FS in the dawn of DL.

At the end of the forenoon, students were presented with detailed walkthrough methods, several FS "traps" and "recipes". The afternoon topic was reserved for hands-on sessions followed by questions and answers, held in a round-table manner. Live coding was conducted in R (R Core Team, 2017) version 3.4.3 or higher using IRkernel and Jupyter Notebook (or Markdown, arbitrarily) with R packages of interest: ggplot2_3.0.0, FSelector_0.31, caTools_1.17.1.1, dprep_3.0, ROCR_1.0-7, data.table_1.11.4, digest_0.6.17, uuid_0.1-2, devtools_1.13.6, repr_0.15.0. Some of the packages above were archived from the CRAN repository due to uncorrected problems and required manual installation, which turned out to be a cumbersome process for an unpractised R user.

The first dataset of interest was a benchmark dataset related to breast cancer diagnostics. Acknowledging the importance of data visualisation before analysis, dimensionality reduction using principal component analysis was performed. The dataset was pre-processed, and usefulness of individual features was checked using various filtering approaches, *e.g.* t-test, Wilcoxon-test, and chi^2-test. Amongst others, the predictive power of individual attribute was assessed, by computing the area under the ROC curve (AUC) scores. To explore multi-class problems, multivariate feature ranking algorithms were applied. Those were implemented in several R libraries and fine-tuned according to exercise requirements. The decision on which attributes to choose was made upon multiple strategies (Riza *et al.*, 2017). One of the strategy in the form of R package, RmRMR, is publicly available from Andrzej Janusz's Github repository (https://github.com/janusza/). Furthermore, various wrapper and embedded methods (*e.g.*, decision trees and neural networks) were explored. The wrappers search heuristics covered: sequential forward search, recursive feature elimination, hill-climbing, simulated annealing, and genetic algorithms.

To raise awareness of false consequences, random FS was applied to synthetic data and resulted in a good set of predictors, that however lead to misinterpretation of the model's generalisation. Students were motivated to repeat the experiment using a proper methodology, make predictions and compare the outcomes. Likewise, using synthetic datasets with different noise manipulations, the task was to find a suitable FS method with high performance.

The last dataset of interest was a real-world example: an acute lymphoblastic leukaemia microarray dataset. The purpose of the exercise was to independently classify cases of leukemia into subtypes using some of the discussed techniques. At the end of the hands-on session, appropriate solutions, broken down into steps and sub goals, were shown and discussed.

Hoang et al. (2019) EMBnet.journal 24, e917
http://dx.doi.org/10.14806/ej.24.0.917

Topic II: Microsoft Azure data science and Machine Learning services of interest

This two-day topic began with a lecture by Dimitrios Vlachakis, Assistant Professor, Genetics Laboratory, Department of Biotechnology, Agricultural University of Athens, Greece, accompanied by two teaching assistants of his research group, Eleni Papakonstantinou, and Louis Papageorgiou. In the preface, Dimitrios gave a brief introduction to the Big Data topics and the way they are encountering in the fields of genomics, molecular dynamics, development of anticancer and antiviral agents, immunoinformatics and neurodegenerative diseases investigations. He justified the necessity of cloud computing for Big Data analysis and the various services offered by Microsoft Azure (*e.g.*, virtual machines), cloud services, websites, and mobile services. In preparation for this topic, all students were asked to sign in to Microsoft Azure Cloud. The local Organising Committee considered three ways to get free Azure credits. First-time users received 170 EUR credits and free 12-month use[5] to explore the different services on the Azure Cloud. Also, Alberto Marcos González, Higher Education Account Executive for Microsoft Spain, provided 25 vouchers with 100 USD (85 EUR) credit. As a third safety measure, Eftim Zdravevski, one of the participants, kindly offered the students to use some credits from his research grant.

Under the technical supervision of Eleni Papakonstantinou, the students went through a practical introduction to Windows Azure Data Services, followed by a hands-on session of SQL database creation and data upload procedures. After the topic of handling big data with Azure, the idea was to work with Azure Machine Learning Studio (AMLS)[6]. However, the Microsoft servers were on a temporary shutdown due to external environmental influences. Meanwhile, Raik Otto, technical assistant for Topic III, attempted to lead the students through preparation of Azure GPU NC-series VM instances, mandatory for the upcoming tasks in Topic III. This was also affected by the servers' outage, so Dimitrios' fast improvisation leads to his pharmacogenomics presentation allowing an uncompromised continuation of the learning experience.

Later that night, it was publicly announced that Azure was accessible again. According to the official webpage[7] section "9/4/2018 South Central US", there was an internal error of all Azure resources, which made it impossible to work with Azure. It was caused by a high energy storm that hit the Azure data centre before it could safely shut down. A significant number of storage servers were damaged, as well as a small number of network devices and power units.

At the beginning of the next day, Raik Otto once more led the way in preparations for Topic III. This time the alternative Azure account creation went through successfully. Thereupon, the student pairs applied for an Azure pass[8] receiving an additional 85,00 EUR of free credits and the ability to request for an NC6 series instance, *i.e.* a GPU optimised virtual machine suited for deep learning problems. Some participants with capabilities to run the tasks on their resources also prepared those by installing Linux Ubuntu (16.04 or higher); the NVIDIA graphic card driver (user specific), additional Nvidia libraries (cuDNN, NCCL), conda, python 3.6, openjdk, bazel, TensorFlow (from sources), Keras and Jupyter Notebook. To avoid potential problems and challenges, Eftim Zdravevski and Petre Lameski, Assistant professors at the Faculty of Computer Science and Engineering, University Sts. Cyril and Methodius in Skopje, Macedonia, prepared computational resources of their faculty, to be shared with participants in case of a need.

Successful infrastructure setup was smoothly boosted by a short talk about MS Azure by Alberto Marcos González, Higher Education Account Executive, Microsoft, and followed by the hands-on session under Eleni's supervision. She introduced AMLS and few "how-to's", graphic user interface that toggles focus from coding to interactive data analysis workflows and visualisation as the machine learning procedures adapted to novice and experienced. Machine learning experiment was built from scratch using "Wisconsin Breast Cancer Diagnostics" dataset. Data were accordingly processed, visualised and summarised. Feature selection procedure was applied, and outputs from different ML algorithms were compared. Models were scored and evaluated, which was followed by Web service deployment to test the model further on. After Eleni, Louis took over with ML topic on text corpus. The aim was to find optimal keywords to classify two diseases: neurodegenerative and heart disease. PubMed dataset was cleaned and preprocessed. Numeric feature vectors were extracted, and the classification and regression model was trained. Again, the model was scored, validated and deployed. Besides, Louis conducted a short tutorial on regular expression in the context of text mining.

The fruitful hands-on session was further enriched by Dimitrios 'tips-and-tricks' on how to apply for Azure and Horizon grants. The discussion about Azure grants was extended by other Azure grant holders, Petre Lameski and Eftim Zdravevski, who shared their experiences for making a successful application.

Topic III: Deep Learning

The last two days were guided by Tarry Singh, Founder/CEO of deepkapha.ai, with teaching and technical assistance of Raik Otto. They facilitated a workshop on Deep Learning (DL), which aims were:

[5]https://azure.microsoft.com/en-us/free/

[6]https://studio.azureml.net/

[7]https://azure.microsoft.com/en-us/status/history

[8]https://www.microsoftazurepass.com/

- to introduce the history and concept of DL;
- to grasp the power of artificial intelligence and what tremendous effect it could have on health care;
- to get participants acquainted with the use of Tensorflow and Keras on Python and its difference;
- to address DL practical issues with publicly available image datasets (MNIST and ISIC, respectively);
- to compare the performance with the different setups: using local CPU, local GPU (if available) and Microsoft Azure's NC6 GPU-backed instances.

Tarry Singh started with the history of ML and gave an overview of DL frameworks and libraries. He moved on to short introductions on several ongoing DL projects in healthcare and showed one active DL project from his company[9], under the hypothesis that (against all current statements) the thalamus is not just a passive relay centre. Furthermore, he exposed - amongst others - the pros and cons of the activation function ReLU (Rectified Linear Unit), which represents a significant step and is often used in DL frameworks. Tarry also stated that Aria2 (Adaptive Richard's Curve Weighted Activation) outperforms state-of-the-art activation functions on publicly available data such as MNIST, CIFAR10, and CIFAR100. He also mentioned the difficulties that computers have to differentiate between two similar looking things – and why humans don't have this problem as well as a possible solution to solve this problem with DL (Togootogtokh *et al.*, 2018).

Fortunately, for the most students support ticket was processed on time, meaning that the DL VMs were created and accessed at the beginning of the hands-on topic. Participants that were not able to create NC6 instances via their accounts were given access either by Eftim Zdravevski and Petre Lameski, from their Azure grant subscription or by Alberto Marcos González team, thus allowing the creation of additional NC6 instances to be used during the second session of DL.

After the theoretical introduction, the participants were instructed with a hands-on tutorial about using Tensorflow to build DL models. The practical tutorial was performed using hands-on coding approach, where the participants were introduced to how to code and where then expected to finish or improve the task themselves.

In the first part, after the NC6 instance was created, the participants set up a secure connection to the machine with support for Jupyter Notebook so that the code could be executed in a more explanatory environment and not directly via the console. Tarry and Raik provided initial notebooks. They were cloned on the NC6 instances and then loaded from each participants computer using a local browser. The notebooks were partially filled with explanations and know-how, and with the helpful lead of Tarry and Raik, the blanks in the notebooks were filled by the participants.

First, the tutorial included how to load a dataset using Python and Tensorflow. For the first days exercise the participants were loading the MNIST dataset. The MNIST dataset consists of 60.000 training images of handwritten numbers and 10.000 testing images. After that, the layers of the neural network were added and configured (first convolutional layers and then the fully connected layer). For the convolutional layers the RELU activation function was used, and for the fully connected layer, we used the softmax activation function.

After the network was set, the configuration of the training session was performed with code adding the type of optimisation and the other hyperparameters such as the number of epochs, dropout rate, learning rate etc. The participants were encouraged to tune the parameters of the network to obtain the best results on the MNIST dataset. This resulted in an unofficial competition which was won by Juliane Pfeil and Sabrina K. Schulze who managed to tune the parameters of the DL network and achieve an accuracy of 99.5 %. The participants were encouraged to try to improve their score as homework.

At the end of the day, the participants were able to load a dataset, design and configure a DL network, train a model and test its performance on the test set using Tensorflow and python.

On the last day of the training school, Tarry explained the difference between Tensorflow and Keras. Keras is a simple high-level neural networks library which works as a wrapper to Tensorflow, which makes DL programming easier and faster. To illustrate this, during the practical presentation, the same was performed with Keras, emphasising on the short and easy approach that Keras uses. The training was once again performed for the MNIST dataset with remarkably little effort for the coding in comparison of using Tensorflow alone.

Then, the transfer learning approach was described and applied to train a DL neural network for skin cancer classification from mole image dataset (International Skin Imaging Collaboration - ISIC). Namely, the Inception V3 (Szegedy *et al.*, 2016) pre-trained weights were used, which were originally obtained when training on the ImageNet dataset. The ISIC dataset contains thousands of images from benign and malign skin cancer, and the task was to differentiate between them. All coding was performed together with Tarry, who explained all the steps and parameters on-the-fly. The participants were also introduced with the data generators capabilities that can load images from a folder, process them (resize, rotate, shift, etc.) to enrich the train or test batches. Another feature that was presented was the freezing and unfreezing of layers for training and adding early stopping criteria while training. All these options and some fine tuning of the network were used to improve the accuracy of cancer classification. Finally, the visualisation of the loss and the results was performed.

The last day was closed by a short closing remark from Tarry Singh and Marta Pérez-Casany.

[9]https://deepkapha.ai

Hoang et al. (2019) EMBnet.journal 24, e917
http://dx.doi.org/10.14806/ej.24.0.917

Trainers' Perspectives

Eleni Papakonstantinou and Louis Papageorgiou

The ATS in Barcelona was an excellent opportunity for interacting with life science researchers and providing novel techniques and pipelines related to big data management and analysis through a well organised and successful workshop. Our aim through the 2-day session was to introduce the students to several of our projects in which cloud solutions have been applied and to implement a standardised pipeline for machine learning experiments using user-friendly and efficient platforms such as AMLS. We organised a hands-on session that all students could follow up regardless of their background, and experiment with different options for processing and analysing big data. Examples of predictive experiments were applied in two different directions, an analysis of the "Wisconsin Breast Cancer Diagnostics" dataset and a text learning approach for data mining. Even though there was a big setback with the Azure platform during the first day, students were able to discover the potential of cloud computing through AMLS in the next day and walk through a set of options provided to facilitate their research, regarding databases organisation, data processing, statistical analysis, feature extraction, machine learning algorithms and the deployment of a web service for predictive experiments.

Throughout the ATS we had the opportunity to interact with the other trainers and trainees, and have insightful discussions regarding highly important topics in our field of interest. In combination with the training session, which was a fruitful experience through the enthusiasm of the students, we consider this workshop as a valuable experience from which we can get important feedback on the real-life problems of both young and experienced researchers. We would also like to personally thank all the participants for contributing to the success of this productive and creative workshop, the fellow lecturers for setting their hearts and minds in organising the teaching sessions, and the trainees for their passion and enthusiasm. Many thanks to the organising committee (Ariel Duarte-López, Prof. Marta Pérez-Casany, and Prof. Erik Bongcam-Rudloff) for their warm welcome and all the efforts they made in front of and behind the scenes to accomplish this successful outcome. We strongly propose that this CHARME ATS. LS is a precursor for the progressive development of a unified scientific community able to address emerging scientific issues including the big data management.

Tarry Singh

The course and the logistics of the ATS were planned very well, and I wish to thank Ariel Duarte-López, Erik Bongcam-Rudloff, and especially my teaching assistant Raik Otto, who worked tirelessly to make this a successful workshop. It is the first time that we were able to do a full-scale DL bioinformatics workshop with real-world datasets. So, thanks not only to the organisers, but also the learners, especially Eftim Zdravevski for providing support as we struggled with the Microsoft Azure service support.

I was pleasantly surprised with the level of understanding and experience as well as enthusiasm most candidates displayed in understanding the concepts, conceptualising and reimagining into their projects and following the training diligently. It is always little time given there is so much to learn in this expanding field of DL, where new models, algorithms and network architectures are being updated on a regular basis. Having said that, it would be great to see a student/learner community emerge that pushes us to develop more advanced modules. In this way we can focus on solving more complex and challenging tasks, such as using best model with limited or smaller datasets, using Bayesian- or PGM (probability graph models) and eventually also experimenting with other advanced models such as GAN's (generative adversarial networks), Autoencoders, Capsule Networks and more.

My learning and take-away from this CHARME ATS was to collaborate with my peers (both professors, as well as students). There is a good scope of improvement when it comes to making a robust infrastructure ready for future exercises. I would suggest evaluating existing cloud computing as well as on-premise models and eventually choose for the right fit that allows learners to start immediately quickly. This also applies to trainers (myself included) that we must come with — for as much possible, a template-based approach so we can deploy systems before we start, and learners get on to learning from the first hour. Second, it should be good to consider planning a proper deep dive bioinformatics DL workshop where we spend at least three full days to learn and train the networks.

Apart from that, I am very grateful for all the generosity, flexibility and kindness that flowed throughout the ATS and I would be delighted to conduct more such workshops with you and perhaps even consider adjunct professor roles for helping build curriculums as there is a considerable need in medical institutions as well as the healthcare industry for such exercises. We keep getting bombarded by professors as heads of institutions who have a lot of data but lack skills and techniques to use DL.

Thanks once again and please stay connected as we are doing some excellent research as much as we are working with leading medical colleges and healthcare companies to make a more meaningful impact with DL and bioinformatics.

Miscellaneous

The first meetup was scheduled for Sunday, before the official start of the training school, where many participants as well as scientific committee and local organisers met and greeted at Palau de la Generalitat in the city centre of Barcelona to discover unique 'ir de

Hoang et al. (2019) EMBnet.journal 24, e917
http://dx.doi.org/10.14806/ej.24.0.917

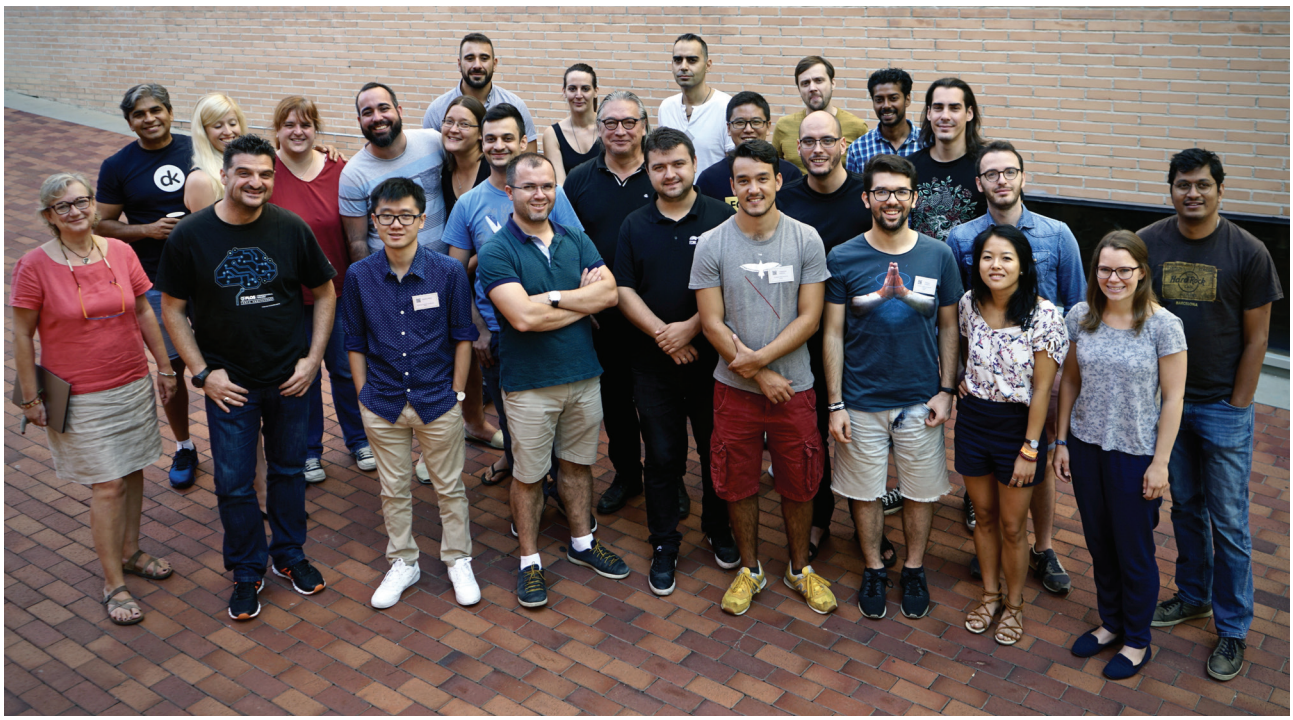

Figure 1. Trainees and trainers of the CHARME Advanced Training School on Big Data for Life Science (image provided by Sabrina Schulze).

tapeo' experience. The second meetup was scheduled on Wednesday after the hands-on sessions from Topic II to assure some 'intellectual break' and further enhance interpersonal relations. For this gathering, called 'potluck', all participants were encouraged to bring snacks, food or beverages from their countries to share. These social events arose from the suggestions of the Think Tank by the EMBnet network tradition. Many informal social events took place where participants, lecturers as well as assistants exchanged research experiences.

The success of this ATS can be measured to some extent by the average impression rates obtained through the participant evaluation forms. 'Content of course' scored on average 4.57 and 'course organisation' 4.86 (response rate of 14/23 and Likert scale [1,5]).

Conclusions

This paper described the topics, detailed content and timeline of events during the CHARME ATS on Big Data for Life Sciences, which took place at the Technical University of Catalonia (UPC) in Barcelona, September 3-7, 2018. All participants appreciated the supporting and inspiring environment in Barcelona. For this school, a more homogenous group was selected, and participants had the required level of prerequisite knowledge about the school's topics, thus streamlining the learning process. All participants seized the opportunity for scientific exchanges and discussions about ongoing projects. All lecturers motivated the participants to deep dive into this exciting topic of big data in bioinformatics. The training school was unanimously regarded as a well-organised platform to enhance future collaborations.

Abbreviations

AI	Artificial Intelligence
AMLS	Azure Machine Learning Studio
ATS	Advanced Training School
DL	Deep Learning
FS	Feature Selection
ML	Machine Learning
VM	Virtual Machines

Trainers' information

Speakers

Dr Andrzej Janusz, Assistant Professor at the Department of Mathematics Informatics and Mechanics, University of Warsaw, Warsaw, Poland (MIM UW), data scientist at eSensei[10], leading the SENSEI project co-funded by NCR&D and Co-founder of the Knowledge Pit platform[11]
Dr Dimitrios Vlachakis, Assistant Professor and group leader of the Genetics and Computational Biology group at the Department of Biotechnology, School of Food, Biotechnology, and Development at the Agricultural University of Athens in Greece
Dr Tarry Singh, CEO and founder of deepkapha.ai and AI neuroscience researcher, Amsterdam, Netherlands

Trainers' Assistants

Eleni Papakonstantinou, PhD (c), Department of Biotechnology, School of Food, Biotechnology and Development, Agricultural University of Athens, Athens, Greece

[10]http://esensei.net/

[11]https://knowledgepit.fedcsis.org/

Louis Papageorgiou, PhD (c), Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece, and Genetics and Computational Biology Group, Laboratory of Genetics, Department of Biotechnology, Agricultural University of Athens, 75 Iera Odos, 11855, Athens, Greece
Raik Otto, PhD (c), Institute for Mathematics and Computer Sciences, Humboldt University, Berlin, Germany

Declarations

Author's contributions

YH, JP, MZ, ŽR, AT, EZ, PL, TS, EP, LP contributed to writing the paper and conducted a critical revision. SKS contributed the notes. ADL and MPC proof-read and approved the final manuscript.

Key Points

- Description of a successfully organized training school for participants with diverse background.
- Illustration of a proactive organization of events within COST Actions by the potential participants.
- Outlining technical challenges and potential solutions to training school that use cloud and on-premises resources.
- Description and links to materials (presentations, source codes, datasets) for hands-on deep learning and feature selection tutorials.

Acknowledgement

A special thanks to Domenica D'Elia, Erik Bongcam-Rudloff, Marcus Frohme and Kristina Gruden for their assurance and full support. A special note of recognition and gratitude goes to Marta Pérez-Casany and Ariel Duarte-López for leading the local organisation of the training school. Other thanks to all three lecturers and assistants who made this school absorbing and exciting. We want to thank the COST Action CHARME for providing us with the necessary funding for this event (COST Action CA15110 is supported by the EU framework program H2020). EZ and PL also acknowledge the support of Microsoft Azure for Research.

Availability of data and materials

Slides, examples, and exercises

FS:
http://bit.ly/2QHP2yA
https://github.com/janusza/
ML on Azure:
http://dimitrislab.com/azure/
DL:
http://bit.ly/2PNn6YS
https://github.com/RaikOtto/Barcelona
https://github.com/TarrySingh/

Publicly available datasets

Wisconsin breast cancer:
https://bit.ly/2pkdRnS
MNIST:
http://yann.lecun.com/exdb/mnist/
CIFAR:
https://www.cs.toronto.edu/~kriz/cifar.html
ISIC:
https://isic-archive.com/api/v1

References

1. Pfeil J, Schulze SK, Zdravevski E, Hoang Y (2018) Report on the "Big Data Training School for Life Sciences", 18-22 September 2017, Uppsala, Sweden. EMBnet.journal **23**, e905. http://dx.doi.org/10.14806/ej.23.0.905
2. Schulze SK, Ramšak Ž, Hoang Y, Zdravevski E, Pfeil J, Duarte-López A, Baier U, Zagorščak M. Proceedings of the "Think Tank Hackathon", Big Data Training School for Life Sciences Follow-up, Ljubljana 6th–7th February 2018. EMBnet.journal **24**, e912. http://dx.doi.org/10.14806/ej.24.0.912
3. Riza LS, Janusz A, Bergmeir C, Cornelis C, Herrera F, Ślezak D, Benítez JM. (2014) Implementing algorithms of rough set theory and fuzzy rough set theory in the R package "roughsets". Information Sciences **287**:68-89. http://dx.doi.org/10.1016/j.ins.2014.07.029
4. R Core Team (2017). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (Accessed 03 September 2018)
5. Togootogtokh E, Amartuvshin A. (2018) Deep Learning Approach for Very Similar Objects Recognition Application on Chihuahua and Muffin Problem. arXiv preprint arXiv:1801.09573.
6. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. (2016) Rethinking the inception architecture for computer vision. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 2818-2826).

EMBnet.journal
Bioinformatics in Action

Bouzin et al. (2019) EMBnet.journal 24, e925
http://dx.doi.org/10.14806/ej.24.0.925

Technical Notes

PyFuncover: full proteome search for a specific function using BLAST and PFAM

Yoan Bouzin[1], Benjamin Thomas Viart[1], María Moriel-Carretero[2], Sofia Kossida[1] ✉

[1]IMGT®, IGH, Univ Montpellier, CNRS, Montpellier, France
[2]CRBM, Univ Montpellier, CNRS, Montpellier, France
Competing interests: YB none; BTV none; MMC none; SK none

Abstract

Python Function uncover (PyFuncover) is a new bioinformatic tool able to search proteins with a specific function in a full proteome. The pipeline coded in python uses BLAST alignment and the sequences from a PFAM family as the search seed. We tested PyFuncover using the fatty acid-binding family (FABP) Lipocalin_7 from PFAM (version 32.0, September 2018) against the Homo sapiens NCBI proteome. After applying the scoring function in all the BLAST results, the data were classified and submitted to a GO-TERM analysis using bioDBnet. Analyses showed that all families of FABPs were ranked within the top scores. Included within this category were also families able to bind to hydrophobic molecules similar to fatty acids such as the retinol acid transporter and the cellular retinoic acid-binding protein.
Availability: PyFuncover source code is freely available at **https://github.com/Tuisto59/PyFuncover/** under the GPL licence.

Introduction

High-throughput technologies produce massive amount of data and bioinformatics approaches help predict and annotate protein function using increasingly complex and precise methods. One example is the NCBI annotation pipeline (Thibaud-Nissen *et al.*, 2016). The human genome sequence was released in 2003 but the annotation of the human proteome in January of 2018 (GRCh38.p12) still contains 2,404 uncharacterised proteins (out of 113,620). Protein families for which the relationship between sequence and function is more complex pose the most significant challenges. The enzymes are particularly tricky because only a small part of the protein is responsible for its function. Moreover, specific binding motifs for which knowledge is still partial and poorly annotated add up to this category.

In 2011 a tool called Ada-BLAST was published and used to predict a fatty acid-binding motif in the human protein BRCA1 (Hedgepeth *et al.*, 2015) and the horse Oxy-myoglobin (Patterson *et al.*, 2011), revealing in those already well-known proteins a new property. Today, this tool is no longer available. Inspired by the methodology explained in (Hong *et al.*, 2009; Patterson *et al.*, 2011; Dae Ko *et al.*, 2011; Hedgepeth *et al.*, 2015; Chintapalli *et al.*, 2015), we created PyFuncover.

PyFuncover is a pipeline able to rank each protein from a proteome according to a specific Protein FAMily (PFAM) (El-Gebali *et al.*, 2019). As a proof of concept, we used this tool to find proteins with putative fatty acid-binding property in the human proteome. We used as a seed the Lipocalin_7 domain family (PF14651[1]).

Workflow

To study a specific activity and to identify other proteins with potentially similar function, the first step is to recover a large set of protein sequences using as a seed the protein annotated with the desired function. Each chosen sequence will make ten iterations (PSI-BLAST accepts a list of multiple sequences, but only the first sequences are used) (see Figure 1, blue box).

Specific family sequences can be downloaded from the PFAM database as a multiple sequences alignment (MSA) from NCBI[2] or UniProt[3] using various formats. Each sequence has a header containing the protein accession, followed by a slash and the domain boundary. The accession of all PSI-BLAST reports is compiled, and each PFAM accession is checked if it is included in the PSI-BLAST results. Other sequences can be from a close family to the chosen one or belong to the same PFAM family. Gaps from the MSA are removed (see Figure 1, green boxes), and a BLAST database is made (see Figure 1, black box).

A whole proteome dataset can be downloaded or any set of proteins in FASTA format (see Figure 1, orange

Article history
Received: 04 February 2019
Accepted: 02 March 2019
Published: 25 April 2019

[1]https://pfam.xfam.org/family/Lipocalin_7
[2]https://www.ncbi.nlm.nih.gov/
[3]https://www.uniprot.org/

Bouzin et al. (2019) EMBnet.journal 24, e925
http://dx.doi.org/10.14806/ej.24.0.925

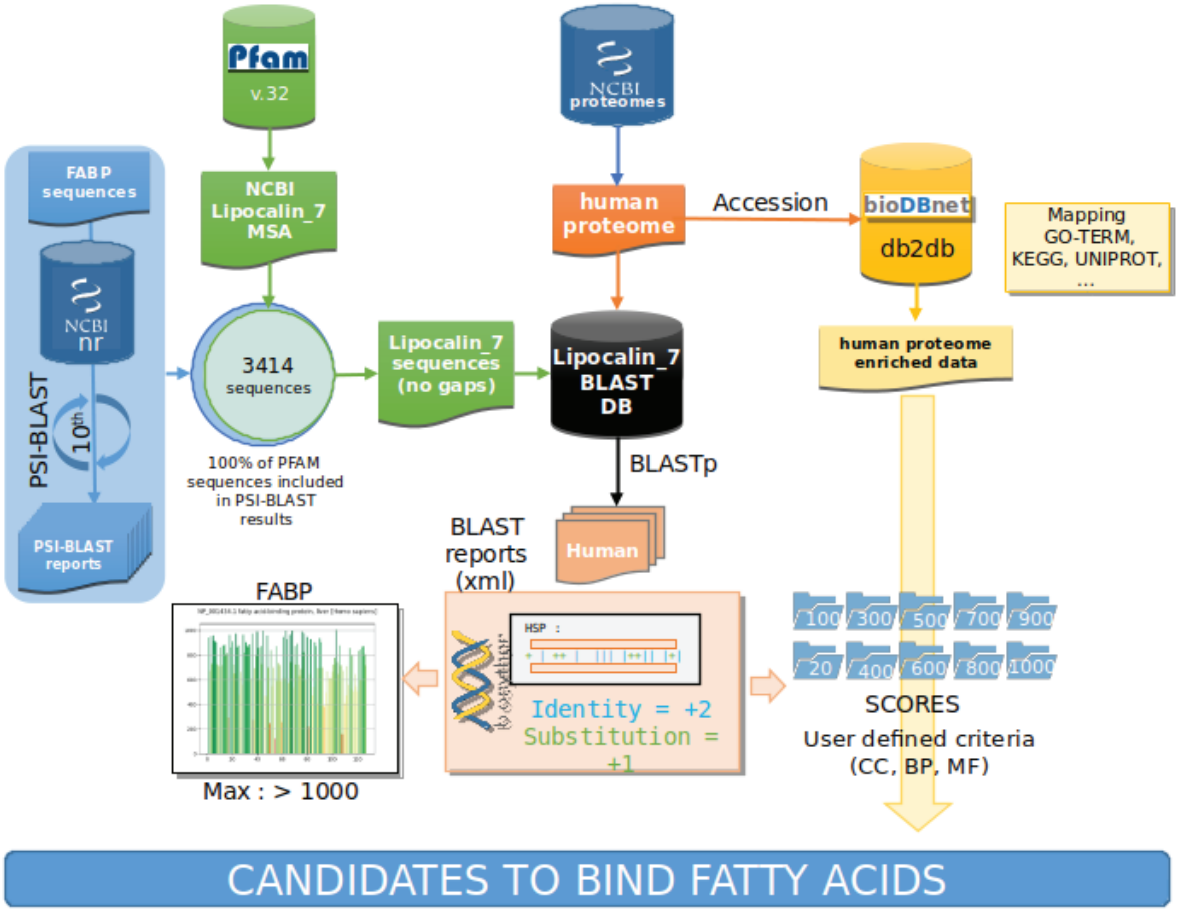

Figure 1. Workflow of PyFuncover.

box). For each sequence, a BLASTp is performed against the PFAM BLAST DB (see Figure 1, black box). For each protein (subject) that matches our sequence (query), BLAST produces alignments, called High Scoring Pairs (HSPs). A score of two, for all the identities, and a score of one, for all positive substitutions, is computed for each alignment. Accession numbers from NCBI are used to retrieve data from different databases (GO-Terms, UniProt, KEGG, PDB, BioCyc, Ensembl, GenBank…) (Ashburner *et al.*, 2000; Berman *et al.*, 2000; Clark *et al.* , 2016; Kanehisa et al., 2019; Karp et al., 2017; UniProt Consortium, 2018; Zerbino *et al.*, 2018) using the cross-reference database web application BioDBnet (db2db) (Mudunuri *et al.*, 2009) and compiled into a biologist-friendly table. This makes the results easy to open and parse using a spreadsheet software such as Excel.

Proof of concept

To test PyFuncover, we used a selection of human Fatty Acid-Binding Proteins (FABPs) (Table 1). The FABPs are part of the lipocalin_7 family (PF14651). The accession numbers of the 3414 sequences from the MSA of NCBI were compared with all the PSI-BLAST results. All the sequences were included into the PSI-BLAST results, and MSA were used to make the BLAST database. Using CDD-Search (Marchler-Bauer *et al.*, 2017), we checked the accessions of the PSI-BLAST reports. The accessions corresponded to the PFAM Lipocalin_7 or to the lipocalins 4 and 5 as expected since all three are members of the Calycin superfamily. The human proteome was downloaded to perform a BLASTp against the database made from the MSA. The XML reports were parsed using BioPython (Cock *et al.*, 2009).

Each amino-acid of each protein obtains a score. Scores can be represented as a barplot for visual analysis (Figure 2). Proteins were split into ten folders (from 100 up to 1000) based on its highest scored amino acid (Figure 2). For the FABPs input set, the highest score was 1052 for FABP7 (isoform X4, NP_001305971). Human

Table 1. List of the FABP used for the PSI-BLAST run.

FABP	UNIPROT Accession
FABP1	P07148
FABP2	P12104
FABP3 (FABP11)	P05413
FABP4	P15090
FABP5	Q01469
FABP6	P51161
FABP7	O15540
FABP8 (PMP2)	P02689
FABP9	Q0Z7S8
FABP12	A6NFH5

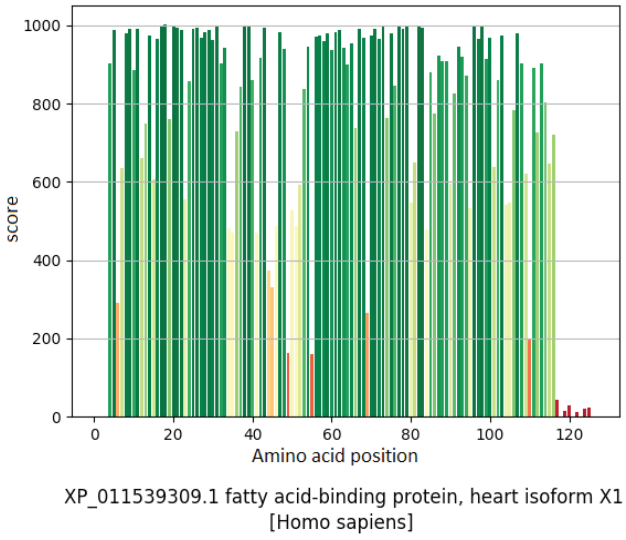

Figure 2. Score histogram per amino acid along the sequence of the FABP1 (isoform X1) of Homo sapiens. Colour range stands from less in red to high in green.

proteomes accession numbers were crossed with the GO-TERM database, using BioDBnet (see Figure 1, yellow box).

Considering the proteins with a score above 900 (arbitrarily chosen), we found members of all the nine FABPs families (Table 2). Above this threshold, we also found five (Cellular) Retinol-Binding Proteins (CRBPs) and two (Cellular) Retinoic Acid-Binding Proteins (CRABPs). This is remarkable, because FABPs, CRBPs and CRABPs are all three subfamilies of the intracellular Lipid-Binding Proteins (iLBPs) family. Moreover both retinol and retinoic acid display a partially similar structure to that of fatty acids (Smathers and Petersen, 2011). As expected the FABP1 family is ranked first using highest mean amino-acid score reaching 735 (Figure 3).

Conclusions

The dataset with a score above 900 contains the top one per cent of the input or 1,530 proteins. This number dramatically exceeds that described above as a proof of concept. This tool aims at helping biologists investigate their favourite set of proteins with a simple sequence-function scoring method. PyFuncover output table combines protein identification, score and several useful

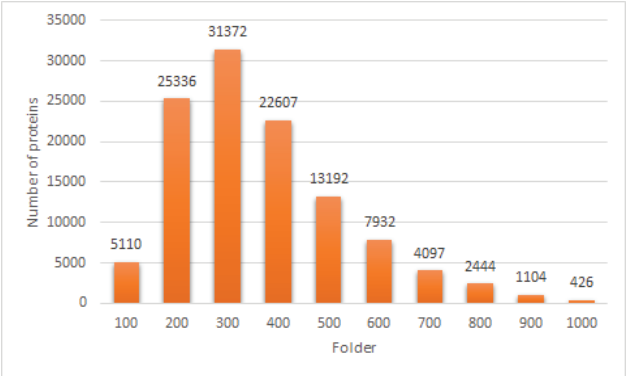

Figure 3. Number of protein in each folder.

Table 2. Proteins implicated in the binding of fatty acids and related hydrophobic molecules from Homo sapiens found in the 900 and 1000 folders.

Protein	Score 1000	Score 900
FABPs	FABP1, 3, 7, 8, 12	FABPS2, 3, 4, 5, 6, 7, 9
RBPs	RBP1, 5, 7	RPB2, 5
CRABPs		CRABP1, 2

databases cross-references for handy investigation. Additionally, while we used it here to detect putative fatty acids-binding motifs, PyFuncover can be tailored to search other functional features matching the user's wishes.

Key Points

- PyFuncover is a new bioinformatic tool to search proteins with a specific function in a full proteome.
- Using the Lipocalin 7 family as input we observed in the top-ranked proteins all families of FABPs as well as families able to bind to hydrophobic molecules similar to fatty acids.
- PyFuncover output table combines protein identification, score and several useful databases cross-references for handy investigation.
- This tool aims at helping biologists investigate their favorite set of proteins with a simple sequence-function scoring method.

Acknowledgements

This work was supported by Merck Sharp and Dohme Avenir (GnoSTic) to S. Kossida and by the ATIP-Avenir program to M. Moriel-Carretero.

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. **25** (1), 25–29. http://dx.doi.org/10.1038/75556
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, *et al.* (2000) The Protein Data Bank. Nucleic Acids Res. **28** (1), 235–242. http://dx.doi.org/10.1093/nar/28.1.235
3. Chintapalli SV, Bhardwaj G, Patel R, Shah N, Patterson RL, *et al.* (2015) Molecular dynamic simulations reveal the structural determinants of Fatty Acid binding to oxy-myoglobin. PLoS One **10** (6), e0128496. http://dx.doi.org/10.1371/journal.pone.0128496
4. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, and Sayers EW (2016) GenBank. Nucleic Acids Res. **44** (D1), D67–72. http://dx.doi.org/10.1093/nar/gkw1070
5. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics **25** (11), 1422–1423. http://dx.doi.org/10.1093/bioinformatics/btp163
6. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, *et al.* (2019) The Pfam protein families database in 2019. Nucleic Acids Res. **47** (D1), D427–D432. http://dx.doi.org/10.1093/nar/gky995
7. Hedgepeth SC, Garcia MI, Wagner LE 2nd, Rodriguez AM, Chintapalli SV, *et al.* (2015) The BRCA1 tumor suppressor binds to inositol 1,4,5-trisphosphate receptors to stimulate apoptotic

calcium release. J. Biol. Chem. **290** (11), 7304–7313. http://dx.doi.org/10.1074/jbc.M114.611186
8. Hong Y, Chalkia D, Ko KD, Bhardwaj G, Chang GS, *et al.* (2009) Phylogenetic Profiles Reveal Structural and Functional Determinants of Lipid-binding. J. Proteomics Bioinform. **2**, 139–149. http://dx.doi.org/10.4172/jpb.1000071
9. Kanehisa M, Sato Y, Furumichi M, Morishima K, and Tanabe M (2019) New approach for understanding genome variations in KEGG. Nucleic Acids Res. **47** (D1), D590–D595. http://dx.doi.org/10.1093/nar/gky962
10. Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, *et al.* (2017) The BioCyc collection of microbial genomes and metabolic pathways. Brief. Bioinform. http://dx.doi.org/10.1093/bib/bbx085
11. Kyung Dae Ko, Chunmei Liu, Rwebangira MR, Burge L, and Southerland W (2011) The development of a proteomic analyzing pipeline to identify proteins with multiple RRMs and predict their domain boundaries. In: 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW). IEEE, pp. 374–381 http://dx.doi.org/10.1109/BIBMW.2011.6112401
12. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, *et al.* (2017) CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. **45** (D1), D200–D203. http://dx.doi.org/10.1093/nar/gkw1129
13. Mudunuri U, Che A, Yi M, and Stephens RM (2009) bioDBnet: the biological database network. Bioinformatics **25** (4), 555–556. http://dx.doi.org/10.1093/bioinformatics/btn654
14. Patterson RL, Hong Y, Chintapalli SV, Bhardwaj G, Zhang Z, *et al.* (2011) Adaptive-BLAST: A User-defined Platform for the Study of Proteins. J. Integr. OMICS **1** (1) http://dx.doi.org/10.5584/jiomics.v1i1.33
15. Smathers RL and Petersen DR (2011) The human fatty acid-binding protein family: Evolutionary divergences and functions. Hum. Genomics **5** (3), 170. http://dx.doi.org/10.1186/1479-7364-5-3-170
16. Thibaud-Nissen F, DiCuccio M, Hlavina W, Kimchi A, Kitts PA, *et al.* (2016) P8008 The NCBI Eukaryotic Genome Annotation Pipeline. J. Anim. Sci. **94** (suppl_4), 184–184.
17. UniProt Consortium T (2018) UniProt: the universal protein knowledgebase. Nucleic Acids Res. **46** (5), 2699. http://dx.doi.org/10.1093/nar/gkv1189
18. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, *et al.* (2018) Ensembl 2018. Nucleic Acids Res. **46** (D1), D754–D761. http://dx.doi.org/10.1093/nar/gkx1098

EMBnet.journal
Bioinformatics in Action

Papageorgiou et al. (2018) EMBnet.journal 24, e910
http://dx.doi.org/10.14806/ej.24.0.910

Reviews

Genomic big data hitting the storage bottleneck

Louis Papageorgiou[1,2], Picasi Eleni[1], Sofia Raftopoulou[1,3,4], Meropi Mantaiou[3], Vasileios Megalooikonomou[5], Dimitrios Vlachakis[1,4,5] ✉

[1] Laboratory of Genetics, Department of Biotechnology, School of Food, Biotechnology and Development, Agricultural University of Athens, Athens, Greece

[2] Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece

[3] Sotiria Chest Diseases Hospital, Athens, Greece

[4] Division of Endocrinology and Metabolism, Center of Clinical, Experimental Surgery and Translational Research, Biomedical Research Foundation of the Academy of Athens, Athens, Greece

[5] Computer Engineering and Informatics Department, School of Engineering, University of Patras, Patras, Greece

Competing interests: LP none; PE none; SR none; MM none; VM none; DM none

Abstract

During the last decades, there is a vast data explosion in bioinformatics. Big data centres are trying to face this data crisis, reaching high storage capacity levels. Although several scientific giants examine how to handle the enormous pile of information in their cupboards, the problem remains unsolved. On a daily basis, there is a massive quantity of permanent loss of extensive information due to infrastructure and storage space problems. The motivation for sequencing has fallen behind. Sometimes, the time that is spent to solve storage space problems is longer than the one dedicated to collect and analyse data. To bring sequencing to the foreground, scientists have to slide over such obstacles and find alternative ways to approach the issue of data volume. Scientific community experiences the data crisis era, where, out of the box solutions may ease the typical research workflow, until technological development meets the needs of Bioinformatics.

Introduction

Since 1956, but mainly in the last decades, storage space needs have grown spectacularly. The problem is that, as time flows, the storage funding issue has increased more than sequencing. That is a big problem that the modern scientist has to face. Sequencing has become more troubling because this issue makes the whole procedure difficult. The motivation for sequencing and producing new data has started to fall away (De Silva and Ganegoda, 2016).

Such data comes in the form of short sequencing reads, i.e. short character strings (typically having lengths in the range 75-150). Each character represents a nucleotide (which is also called a "base"), and can assume the values of A (adenine), C (cytosine), G (guanine), T (thymine), or N (failure in the base calling process) (Langmead, 2010). The nucleotide string is usually accompanied by a corresponding string of ASCII characters, encoding the "quality" (that is, the error probability of the base calling) of each of the nucleotides. This is a representative case of how a typical sequencing setup works when a resequencing problem is considered. In such a case, a reference (possibly not 100% accurate) for the genome/transcriptome of the organism being sequenced is already known. One has to map the DNA/RNA sequence reads to the reference (*i.e.*, understand where such reads come from in the reference) and find variants present in the genetic code of the specific organisms compared to the reference (Xu *et al.*, 2014).

Depending on the biological application at hand, one might need to perform several tasks on the data, possibly in several steps, with both per-read and global computations required (Libbrecht and Noble, 2015). A typical workflow corresponding to the above use case might be as follows:

- store the reads in compressed searchable form (necessary to avoid excessive storage consumption);
- retrieve (a subset of) the reads based on some criterion, possibly depending on the experiment metadata (for instance, select all the sequencing reads derived from a given tissue subject to a specific biological condition);
- select/process the reads, for example: identify all the reads containing long stretches of low-quality nucleotides, and trim/eliminate them;
- pattern/match the surviving data, read by read, onto a reference genome;
- store the reads and their alignments to the reference genome (that is, the matches found in the genome for each read) in compressed searchable form again.

In the meantime, the Cern data centre has upgraded storage capacity on 200 petabytes, breaking the previous record of 100 petabytes. Information produced every day is one petabyte per second. This leads to lack of space

Article history
Received: 14 February 2018
Accepted: 07 March 2018
Published: 19 April 2018

Papageorgiou et al. (2018) EMBnet.journal 24, e910
http://dx.doi.org/10.14806/ej.24.0.910

capacity within 3 minutes. Then all this information has to be filtered for any findings which are stored for later use, after three minutes everything is deleted and three minutes is a very short period to trace back all this information (Britton and Lloyd, 2014).

All this data that need to be retrieved and handled is being held up in I/O traffic because of slow processing power (Fan *et al.*, 2014). Even if process power isn't still satisfying for such needs, there are other ways to slide over this obstacle. Technology and science go on hand by hand, and someone has to think out of the box to solve any occurring problem, without being stuck conventionally. The other suggested path is the information packings. By limiting, not only the data space needed for the information that we already have but also the new information we get, we can go further in a less chaotic and more organised environment by throwing away unnecessary information (repeats) (Fan *et al.*, 2014).

The important thing is to compress information without losing data that is needed. One should keep in mind that not only huge amounts of data will need to be processed each day, but also that some operations might need to be performed incrementally. For instance, the data produced at some point might be used to refine the results obtained from some other data generated previously, implying the reprocessing of a possibly much bigger dataset. For these reasons the development of a robust and extensible high-throughput storage/matching/processing system is necessary. Many other workflows might be envisaged, but most of them share the same skeleton structure, that is storage, retrieval, filtering/processing, and final storage of the results.

Clustering information based on a representative model (in some permissible limits) is an interesting way to approach the problem (Slonim *et al.*, 2005). For instance, when information is recorded in output, the ones that don't differ from our first recorded ones should not be referred. The differences are the essential information for our search.

To some extent, sequencing data are intrinsically noisy (they depend on chemical reactions which are stochastic in nature) (Alvarez et al., 2015). On the one other hand, high-throughput sequencing techniques have now reached a high degree of reliability, so sequencing errors are relatively rare (Pareek *et al.*, 2011). Also, as mentioned above, sequencing machines provide a quantification of the sequencing error at each nucleotide regarding "qualities", which can be used to pinpoint problematic nucleotides/regions in the read.

Storage state of the art

Since several years, under the pressure of increasing volumes of data and due to reduced hardware costs, the view of databases as centralised data access points has become vaguer (Sreenivasaiah and Kim, 2010). Fundamental paradigms of data organisation and storage have been revised to accommodate parallelisation, disreputability and efficiency. The storage mechanics, the querying methods and the analysis and aggregation of the results follow new models and practices. Search has gone beyond the boolean match, being directly linked to efficient indexes allowing approximate matching in domains ranging from string to graph matching (Pienta *et al.*, 2016). The main points of this progress can be summarised as follows.

From row-oriented representation, nowadays the trend is to move to column-oriented representation and database systems (Abadi *et al.*, 2009), which are the evolution of what was called "large statistical databases" in earlier literature (Corwin *et al.*, 2007; Turner *et al.*, 1979). Column-oriented database systems allow high compressibility per column (Abadi *et al.*, 2008), by direct application of existing ratio-optimised compression algorithms (Abadi *et al.*, 2006). Furthermore, several threads are pulling current database practices away from the relational paradigm. Large-scale storage and access may include dynamic control over data layout. Peer-topeer (P2P) overlays are also used in distributed stores, exchanging, e.g., index information to contributing nodes in distributed data warehouses (Doka *et al.*, 2011), where even the queries can be executed in a peerbased fashion spreading the processing load. Another alternative, related to large-scale analysis is the case of Pig Latin (Gates *et al.*, 2009), where a SQL-like syntax is used to provide the data flow requirements for analysis over a map-reduce infrastructure. Other efforts offer partial SQL support, as is the case of Hive (Ashish *et al.*, 2010) and the corresponding query language, named HiveQL.

Recently, parallel databases (*e.g.*, Oracle Exadata, Teradata) allowed high efficiency at the expense of failure recovery and elasticity (Pavlo *et al.*, 2009). Newer approaches and versions of these parallel databases integrate a map-reduce approach into the systems to alleviate these drawbacks, see (Abouzeid *et al.*, 2009) for more information.

The increased availability of low-cost, legacy computers has brought cloud computing settings to the front line. Shared-nothing architectures, implying selfsufficient storage or computation nodes, are applied to storage settings (O'Driscoll *et al.*, 2013). There exist also alternative clouds based on active data storage (Delmerico *et al.*, 2009; Fan *et al.*, 2014) where part of the computational database effort is distributed among the processing units of storage peripherals. Such an example is the case of DataLab (Moretti *et al.*, 2010) where data operations, both read and write, are based on "sets" - essentially named collections of files - distributed across several active storage units (ASUs).

Finally, task-focused storage solutions are devised to face problems in bioinformatics (Hsi-Yang Fritz *et al.*, 2011), social networks (Ruflin *et al.*, 2011) and networkmonitoring and forensics (Giura and Memon, 2010), showing how much data requirements drive the need for research on storage systems. Especially in bioinformatics, there exist approaches that combine compressed storage and indexing under a common

approach, based on sequence properties and works on indexed string storage (Arroyuelo and Navarro, 2011; Ferragina and Manzini, 2005). There are cases where the system provides tunable parameters that allow a balance between data reuse and space recovery (Hsi-Yang Fritz *et al.*, 2011), by keeping only the data that may be reused shortly. At this point it must be stressed that there still exist relational databases that are used for high-throughput data storage, an example being the NCBI GEO archive (Barrett *et al.*, 2009) which supports the submission of experimental outputs and provides a set of tools to retrieve, explore and visualise data. However, even in the case of NCBI GEO, the relational nature of the underlying database is used to identify specific datasets and not specific sequences (*i.e.*, instances). Further analysis tools are used to locate sequences and aggregate information from them. In time series and sensor networks, storage can be a severe problem. In the literature, there are methods such as Sparse Indexing (Lillibridge *et al.*, 2009), where sampling and backup streams are used to create indexes that avoid disk bottlenecks and storage limitations.

Beyond the full-text indexing - combined with compressed storage, as explained above - often met in bioinformatics, there are several works on time series indexing and graph indexing. These two types of indexes, together with the string (and, thus, sequence) indexes, provide full artillery of methods that can cope with a great variety of problems and settings. Graph indexing is under massive research, due to its applicability on such cases as chemical compounds, protein interactions, XML documents, and multimedia.

Graph indexes are often based on frequent subgraphs (Yan *et al.*, 2005), or otherwise "semantically" interesting (Jiang *et al.*, 2007). There exist hierarchical graph index methods (Abello and Kotidis, 2003), and hash-based ones. A related recent work (Schafer *et al.*, 2017) relies on "fingerprints" of graphs - derived from hashing on cycles and trees within a graph - for efficient indexing. The method is part of an open source software, named "Scaffold Hunter", for visual analysis of chemical compound databases.

In the case of time series, to efficiently process and analyse large volumes of data, one must consider operating on summaries (or approximations) of these data series. Several techniques have been proposed in the literature (Anguera *et al.*, 2016), including Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT), Piecewise Aggregate Approximation (PAA) , Discrete Wavelet Transform (DWT) , Adaptive Piecewise Constant Approximation (APCA) , Approximation (SAX), and others. Recent works (Emil Gydesen *et al.*, 2015) based on the iSAX (Shieh and Keogh, 2009) algorithm have focused on the batch update process of indexing very large collections of time series and have proposed highly efficiency algorithms with optimised disk I/O, managing to index "one billion time series" very efficiently on a single machine. Another system, Cypress (Reeves *et al.*, 2009), applies multi-scale analysis to decompose time series and to obtain sparse representations in various domains, allowing reduced storage requirements. Furthermore, this method can answer many statistical queries without the need to reconstruct the original data.

Conclusions

The life sciences are becoming a "big data business". Modern science needs have changed, and lack of storage space has become of great interest among the scientific community. There is an urgent need for computational ability and storage capacity development. In a short period, several scientists are finding themselves unable to extract full value from the large amounts of data becoming available. The revolution that happened in next-generation sequencing, bioinformatics and biotechnology are unprecedented. Sequencing has to come first in priority but, because of technical problems during this process, the time spent to solve space problems is longer than the one dedicated to the part of collecting and analysing data. During this problem, a huge amount of data produced every day is being lost. As we understand, the scientist must overcome some hurdles, from storing and moving data to integrate and analysing it, which will require a substantial cultural shift. Moreover, similar problems will appear in many other fields of life science. As an example, the challenges that neuroscientists have to face in the future will be even greater than those we nowadays deal with the next generation sequencing in genomics. The nervous system and the brain are far more complicated entities than the genome. Today, the whole genome of a species can fit on a CD, but in the future how we will handle the brain which is comparable to the digital content of the world. Therefore, new technological methods more effective and efficient must be found, to serve the needs of scientific search. Solving that "bottleneck" has enormous consequences for human health and the environment.

Acknowledgements

The research reported in the present paper was partially supported by the FrailSafe Project (H2020-PHC-21-2015 - 690140) "Sensing and predictive treatment of frailty and associated co-morbidities using advanced personalized models and advanced interventions", co-funded by the European Commission under the Horizon 2020 research and innovation programme.

References

1. Abadi D, Boncz P and Harizopoulos S (2009) Column-oriented database systems, Proceedings of the VLDB Endowment, **2**(2), 1664-1665. http://dx.doi.org/10.14778/1687553.1687625
2. Abadi D, Madden S and Ferreira M (2006) Integrating Compression and Execution in Column-Oriented Database Systems, Proceedings of the 2006 ACM SIGMOD international conference on Management of data, **1**, 671-682. https://doi.org/10.1145/1142473.1142548

3. Abadi D, Madden S and Hachem N (2008) Column-stores vs. row-stores: how different are they really?, Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 67-980. http://dx.doi.org/10.1145/1376616.1376712
4. Abello J and Kotidis Y (2003) Hierarchical graph indexing. Proceedings of the twelfth international conference on Information and knowledge management. ACM, New Orleans, LA, USA, **1**, 453-460. https://doi.org/10.1145/956863.956948
5. Abouzeid A, Pawlikowski K, Abadi D, Saliberschatz A and Rasin A (2009) HadoopDB: an architectural hybrid of MapReduce and DBMS technologies for analytical workloads, Journal Proceedings of the VLDB Endowment, **2**(1), 922-933. http://dx.doi.org/10.14778/1687627.1687731
6. Alvarez J R, Skachkov D, Massey S E, Kalitsov A and Velev J P (2015) DNA/RNA transverse current sequencing: intrinsic structural noise from neighboring bases, Frontiers in genetics, **6**(213), 1-11. https://doi.org/10.3389/fgene.2015.00213
7. Anguera A, Barreiro J M, Lara J A and Lizcano D (2016) Applying data mining techniques to medical time series: an empirical case study in electroencephalography and stabilometry, Computational and structural biotechnology journal, **14**, 185-199. https://doi.org/10.1016/j.csbj.2016.05.002
8. Arroyuelo D and Navarro G (2011) Space-efficient construction of Lempel–Ziv compressed text indexes, Information and Computation, **209**(7), 1070-1102. https://doi.org/10.1016/j.ic.2011.03.001
9. Ashish T, Joydeep S, Namit J, Zheng S, Prasad C, et al. (2010) Hive A Petabyte Scale Data Warehouse Using Hadoop, Proceedings of the 26th International Conference on Data, **1**, 996-1005. http://dx.doi.org/10.1109/ICDE.2010.5447738
10. Barrett T, Troup D B, Wilhite S E, Ledoux P, Rudnev D, et al. (2009) NCBI GEO: archive for high-throughput functional genomic data, Nucleic acids research, **37**, 885-890. https://doi.org/10.1093/nar/gkn764
11. Britton D and Lloyd S L (2014) How to deal with petabytes of data: the LHC Grid project, Reports on progress in physics. Physical Society. https://doi.org/10.1088/0034-4885/77/6/065902
12. Corwin J, Silberschatz A, Miller P L and Marenco L (2007) Dynamic tables: an architecture for managing evolving, heterogeneous biomedical data in relational database management systems, Journal of the American Medical Informatics Association : JAMIA, **14**, 86-93. https://doi.org/10.1197/jamia.M2189
13. De Silva P Y and Ganegoda G U (2016) New Trends of Digital Data Storage in DNA, BioMed research international. http://dx.doi.org/10.1155/2016/8072463
14. Doka K, Tsoumakos D and Koziris N (2011) Online Querying of Dimensional Hierarchies, Journal of Parallel and Distributed Computing, **71**(3), 424-437. http://dx.doi.org/10.1016/j.jpdc.2010.10.005
15. Delmerico J A, Byrnes N A, Bruno A E, Jones M D, Gallo S M, et al. (2009) Comparing the performance of clusters, Hadoop, and Active Disks on microarray correlation computations. 2009 International Conference on High Performance Computing (HiPC), 378-387. http://dx.doi.org/10.1109/HIPC.2009.5433190
16. Emil Gydesen J, Haxholm H, Sonnich Poulsen N, Wahl S and Thiesson B (2015) HyperSAX: Fast Approximate Search of Multidimensional Data. http://dx.doi.org/10.5220/0005185201900198
17. Fan J, Han F and Liu H (2014) Challenges of Big Data Analysis, National science review, **1**, 293-314. https://doi.org/10.1093/nsr/nwt032
18. Ferragina P and Manzini G (2005) Indexing compressed text, J. ACM, **52**, 552-581. http://dx.doi.org/10.1145/1082036.1082039
19. Gates A, Natkovich O, Chopra S, Kamath P, Narayanamurthy S, et al. (2009) Building a high-level dataflow system on top of Map-Reduce: the Pig experience, Proceedings of the VLDB Endowment, **2**(2), 1414-1425. http://dx.doi.org/10.14778/1687553.1687568
20. Giura P and Memon N (2010) NetStore: An Efficient Storage Infrastructure for Network Forensics and Monitoring., International Workshop on Recent Advances in Intrusion Detection, **6307** 277-296. https://doi.org/10.1007/978-3-642-15512-3_15
21. Hsi-Yang Fritz M, Leinonen R, Cochrane G and Birney E (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression, Genome research, **21** (5), 734-740. http://dx.doi.org/10.1101/gr.114819.110
22. Jiang H, Wang H, Yu P S and Zhou S (2007) GString: A Novel Approach for Efficient Search in Graph Databases. 2007 IEEE 23rd International Conference on Data Engineering. http://dx.doi.org/10.1109/ICDE.2007.367902
23. Langmead B (2010) Aligning short sequencing reads with Bowtie, Current protocols in bioinformatics. http://dx.doi.org/10.1002/0471250953.bi1107s32
24. Libbrecht M W and Noble W S (2015) Machine learning applications in genetics and genomics, Nature reviews. Genetics, **16**(6), 321-332. http://dx.doi.org/10.1038/nrg3920
25. Lillibridge M, Eshghi K, Bhagwat D, Deolalikar V, Trezise G, et al. (2009) Sparse indexing: large scale, inline deduplication using sampling and locality. Proccedings of the 7th conference on File and storage technologies. USENIX Association, San Francisco, California, 111-123.
26. Moretti C, Bui H, Hollingsworth K, Rich B, Flynn P, et al. (2010) All-Pairs: An Abstraction for Data-Intensive Computing on Campus Grids, IEEE Transactions on Parallel and Distributed Systems, **21**(1), 33-46. http://dx.doi.org/10.1109/TPDS.2009.49
27. O'Driscoll A, Daugelaite J and Sleator R D (2013) 'Big data', Hadoop and cloud computing in genomics, Journal of biomedical informatics, **46**(5), 774-781. https://doi.org/10.1016/j.jbi.2013.07.001
28. Pareek C S, Smoczynski R and Tretyn A (2011) Sequencing technologies and genome sequencing, Journal of applied genetics, **52**(4), 413-435. https://doi.org/10.1007/s13353-011-0057-x
29. Pavlo A, Paulson E, Rasin A, Abadi D, Dewitt D, et al. (2009) A comparison of approaches to large-scale data analysis, ACM SIGMOD, International Conference on Management of data. http://dx.doi.org/10.1145/1559845.1559865
30. Pienta R, Navathe S, Tamersoy A, Tong H, Endert A, et al. (2016) VISAGE: Interactive Visual Graph Querying, AVI : proceedings of the Workshop on Advanced Visual Interfaces. AVI. http://dx.doi.org/10.1145/2909132.2909246
31. Reeves G, Liu J, Nath S and Zhao F (2009) Managing massive time series streams with multi-scale compressed trickles, Proc. VLDB Endow., **2**(1), 97-108. http://dx.doi.org/10.14778/1687627.1687639
32. Ruflin N, Burkhart H and Rizzotti S (2011) Social-data storagesystems. Databases and Social Networks. ACM, Athens, Greece. http://dx.doi.org/10.1145/1996413.1996415
33. Schafer T, Kriege N, Humbeck L, Klein K, Koch O, et al. (2017) Scaffold Hunter: a comprehensive visual analytics framework for drug discovery, Journal of cheminformatics, **9**(1), 28-32. https://doi.org/10.1186/s13321-017-0213-3
34. Shieh J and Keogh E (2009) iSAX: disk-aware mining and indexing of massive time series datasets, Data Min. Knowl. Discov., **19**, 57. https://doi.org/10.1007/s10618-009-0125-6
35. Slonim N, Atwal G S, Tkacik G and Bialek W (2005) Informationbased clustering, Proceedings of the National Academy of Sciences of the United States of America, **102**(51), 18297-18302. https://doi.org/10.1073/pnas.0507432102
36. Sreenivasaiah P K and Kim D H (2010) Current trends and new challenges of databases and web applications for systems driven biological research, Frontiers in physiology. http://dx.doi.org/10.3389/fphys.2010.00147
37. Turner M, Hammond R and Cotton P (1979) A DBMS for large statistical databases, Proceeding VLDB '79 Proceedings of the fifth international conference on Very Large Data Bases, **5**, 319-327. http://dx.doi.org/10.1109/VLDB.1979.718147
38. Xu P, Zhang X, Wang X, Li J, Liu G, *et al.* (2014) Genome sequence and genetic diversity of the common carp, Cyprinus carpio, Nature genetics, **46**(11), 1212-1219. http://dx.doi.org/10.1038/ng.3098
39. Yan X, Yu P S and Han J (2005) Graph indexing based on discriminative frequent structure analysis, ACM Trans. Database Syst., **30**(4), 960-993. http://dx.doi.org/10.1145/1114244.1114248

EMBnet.journal
Bioinformatics in Action

Papakonstantinou et al. (2019) EMBnet.journal 24, e921
http://dx.doi.org/10.14806/ej.24.0.921

Reviews

NOTCH3 and CADASIL syndrome: a genetic and structural overview

Eleni Papakonstantinou[1,2], Flora Bacopoulou[3], Dimitrios Brouzas[4], Vasileios Megalooikonomou[5], Domenica D'Elia[6], Erik Bongcam-Rudloff[7], Dimitrios Vlachakis[1,2,8] ✉

[1]Laboratory of Genetics, Department of Biotechnology, School of Food, Biotechnology and Development, Agricultural University of Athens, Athens, Greece

[2]Lab of Molecular Endocrinology, Center of Clinical, Experimental Surgery and Translational Research, Biomedical Research Foundation of the Academy of Athens, Athens, Greece

[3]Center for Adolescent Medicine and UNESCO Chair on Adolescent Health Care, First Department of Pediatrics, Medical School, National and Kapodistrian University of Athens, Aghia Sophia Children's Hospital, Athens, Greece

[4]1st Department of Ophthalmology, National and Kapodistrian University of Athens, Athens, Greece

[5]Computer Engineering and Informatics Department, School of Engineering, University of Patras, Patras, Greece

[6]CNR Institute for Biomedical Technologies, Bari, Italy

[7]SLU-Global Bioinformatics Centre, Department of Animal Breeding and Genetics Science, University of Agricultural Sciences, Uppsala, Sweden

[8]Department of Informatics, Faculty of Natural and Mathematical Sciences, King's College London, London, United Kingdom

Competing interests: EP none; FB none; DB none; VM none; DD none; EBR none; DV none

Abstract

CADASIL syndrome is a rare disease that belongs to a group of disorders called leukodystrophies. It is well established that NOTCH3 gene on chromosome 19 is primarily responsible for the development of the CADASIL syndrome. Herein, an attempt is made to shed light on the actual molecular mechanism underlying CADASIL syndrome, through insights extracted from comprehensive evolutionary studies and in silico modelling on Notch 3 protein. In particular, we suggest the use of optical coherence tomography angiography for the detection of early signs of small vessel diseases, which are the major precursors to a repertoire of neurodegenerative conditions, including CADASIL.

Introduction

The CADASIL (Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy) syndrome is a hereditary dominant rare disease caused by NOTCH3 gene mutations, affecting adults over the middle-age and leading to dementia and disability.

At systemic level CASADIL syndrome is characterised by a series of damages at the central nervous system produced by recurrent ischemic strokes accompanied by diffuse white matter lesions and subcortical infarcts. Among all known rare diseases, CASADIL is one of the most common form of hereditary stroke disorder that primarily affects small blood vessels in the white matter of the brain (Guruharsha, 2012), and is distinguished from other vascular diseases by the characteristic accumulation of granular osmiophilic material in brain vasculature (Tikka *et al.*, 2009).

Article history
Received: 14 January 2019
Accepted: 27 January 2019
Published: 22 May 2019

The most recent updating (Last Update: March 14, 2019) from GeneReviews[1], an international point-of-care resource for clinicians, only provides some recommendations to supportive treatment of strokes and to alleviate/limit the extent of symptoms such as frequent migraine, mood disturbance, apathy and the progressive cognitive decline to dementia. The efficacy of prevention of primary manifestations such as stroke/TIA has not still been proven, and surveillance is demanded to clinician's specialists. The instruments available for diagnosis are the genetic analysis and magnetic resonance imaging (MRI) evaluation. As for any rare disease, the number of studies is still too small (Figure 1), whereas to elucidate the molecular mechanisms underlining CASADIL syndrome is crucial to provide people affected by this disease with a hope for effective therapy.

First studies identifying in mutations of the NOTCH3 gene the genetic origin of the CASADIL syndrome were published by (Joutel *et al.*, 1996), after a previous study that mapped, through genetic linkage analysis in two unrelated families, the disease locus to

[1]https://www.ncbi.nlm.nih.gov/books/NBK1500/

Papakonstantinou et al. (2019) EMBnet.journal 24, e921
http://dx.doi.org/10.14806/ej.24.0.921

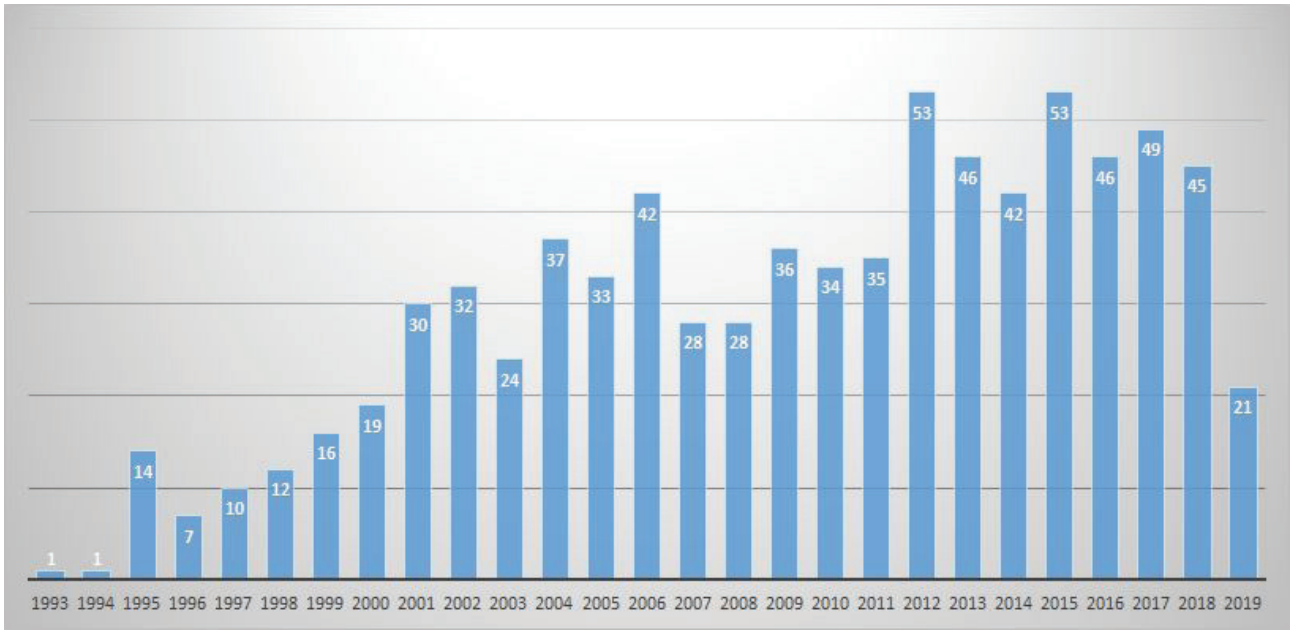

Figure 1. PubMed research results for articles published on the CASADIL syndrome. Search term "Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts AND Leukoencephalopathy" in the [Title/Abstract] fields.

chromosome 19q12 (Tournier-Lasserve *et al.*, 1993). Since then, the majority of efforts have been focused on the study of NOTCH3 and more than 200 mutations have been reported. Some of these mutations give out a phenotype while others remain silent. Extensive analysis for grouping, organising and mapping these mutations is essential for a straightforward linkage of genotype-phenotype.

NOTCH3 mutations in CASADIL syndrome

Notch3 is a large type I transmembrane receptor, mainly expressed in vascular smooth muscle cells and pericytes close to the local blood vessels. It has been reported that if Notch switches off becoming inactive, epidermal precursors kick in that convert normal cells to neuroblasts (Artavanis-Tsakonas and Muskavitch, 2010; Louvi and Artavanis-Tsakonas, 2012; Poulson, 1937; Vlachakis *et al.*, 2014). Neuroblasts differentiate and produce embryos which are characterised by nervous system hypertrophy and epidermal structure deficiencies. Accumulation and deposition of Notch3 extracellular domain within vessel walls is a key pathological feature in CADASIL patients and is believed to be responsible for the formation of granular osmiophilic material (GOM) on the surface of vascular smooth muscle cells and pericytes. For this reason, GOM has diagnostic value for CASADIL syndrome when observed in ultrastructural analyses of skin biopsies.

So far, a series of Notch3 pathogenic mutations affecting the number of cysteine residues in the extracellular domain of the receptor and causing protein misfolding and receptor aggregation have been identified. Some of these mutations that lead to the loss or gain of a cysteine residue in 1 of the 34 epidermal growth factor-like repeat (EGFr) domains of the protein. These mutations are highly distinctive for the CASADIL syndrome. Several studies have been carried out to determine their penetrance and frequency. A recent work reports that NOTCH3 EGFr 1–6 pathogenic variants are much more frequent in CASADIL patients than EGFr 7–34 pathogenic variants, which instead predominate in the population. NOTCH3 EGFr 1–6 pathogenic variants are also associated with a more severe phenotype, are characterized by a 12-year earlier onset of stroke, and by a lower survival of CASADIL patients compared with EGFr 7–34 pathogenic variants (Rutten *et al.*, 2019). However, also cysteine-sparing NOTCH3 mutations have recently been identified and do not follow the characteristic pathology and pattern of the disease. Some of these cysteine-sparing NOTCH3 missense mutations also cause GOM whereas some others do not (Muiño *et al.*, 2017). The pathogenic role of cysteine-sparing NOTCH3 mutations in patients with typical clinical CADASIL syndrome is still unknown.

Linking known Notch3 mutations to structure

The specific aim of this study is to provide insights into the structural properties of the Notch3 protein which promote CADASIL. This can be achieved by analyzing the 3D structural properties of the Notch3 protein. Through our preliminary genetic and proteomic analysis of the Notch3 protein (Dr Baumann group), we are aware of a series of point mutations of the Notch3 protein which lead to loss of structure and therefore promote CADASIL (Figure 2) (Polychronidou *et al.*, 2015).

The link between the CADASIL manifestation and the Notch3 3D structure is mainly based on the partially known X-ray crystallography Notch3 3D structure.

Papakonstantinou et al. (2019) EMBnet.journal 24, e921
http://dx.doi.org/10.14806/ej.24.0.921

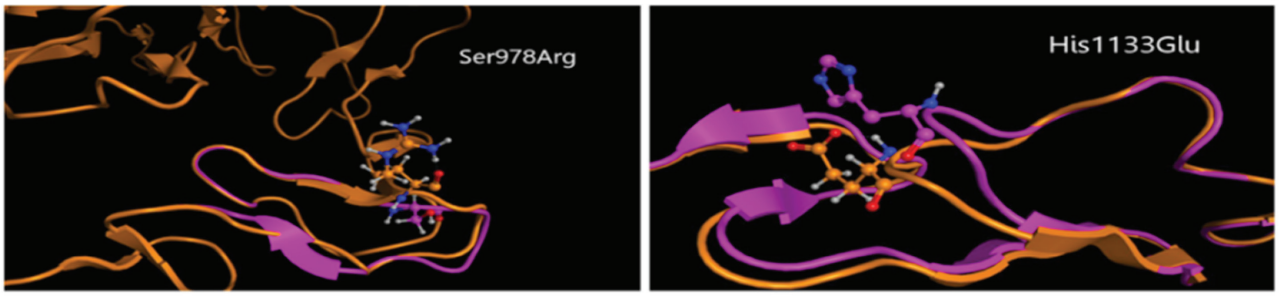

Figure 2. 3D molecular modelling studies on point mutations of Notch 3 indicates a significant loss of 3D structure, thus resulting in a partially unfolded protein with compromised functionality. The study was performed in the Molecular Operating Environment (MOE 2011.10). Montreal, Quebec, Canada: Chemical Computing Group ULC[2]

Correlation of the mutant Notch3 proteins to the wild type model has provided us with all the vital structural information concerning the extent and nature of Notch3 structural loss. The structurally and conformationally essential sites on the Notch3 structure have been identified so this information could lead to future in silico and in vitro experiments to be conducted to analyse their effect and role in CADASIL (Ioannidou *et al.*, 2013).

Evolutionary study of Notch 3

During this first phase of the phylogenetic analysis, motif construction and codon usage have been conducted (Vlachakis *et al.*, 2014). The Notch3 protein and nucleotide sequences have been retrieved from the literature and the publicly available genomic databases. To determine the phylogenetic status of the Notch3 protein, detailed and comprehensive phylogenetic analysis of protein sequences has been conducted. Subsequently, sequence motifs were excised from the alignments. It has been demonstrated that different genomes have their characteristic patterns of codon usage. Based on the above, we have investigated the variations in codon usage among the different NOTCH3 genomes. We have evaluated if there are any characteristic patterns of codon usage and how these patterns are related to the CADASIL disease. Towards this direction, we have combined the genes from all NOTCH3 genomes and calculated the relative codon usage for each genome and each gene per genome by using the Codon Adaptation Index (CAI) (Sharp and Li, 1987).

Genetic study of NOTCH3

Statistical analysis was employed to explore all sequence variations and Single Nucleotide Polymorphisms (SNP) for NOTCH3 (Polychronidou *et al.*, 2015). Sequence variation was further analysed by exploring any relationship or patterns among SNPs within the species. To identify the individual nucleotide positions which most contribute to the NOTCH3 differentiation, a list of all informative phylogenetic sites from the codon-aligned multiple genome alignment was assembled. Then, a consensus sequence was created by the "majority rule". Subsequently, a Bayesian partition model was employed to identify groups of more than two SNPs in the data exhibiting similar behaviour, estimating their optimum number and distributional properties in the genomes considered. By using the latter approach, we aimed to enhance our understanding of the dependencies between the various motifs examined for all the populations and consequently their involvement in the CADASIL. Past studies on NOTCH3 genetics have employed standard statistical techniques or methodologies which allow exploring genetic data and identify natural groupings of most similar sequences; however, here we suggest a likelihood-based model approach where whole sequence interactions are considered without restricting our search to pair-wise or threshold restricted groups of SNPs similarities.

Structural study of Notch 3

The next step included the structural analysis task of the multiple known mutations of Notch3 (Vlachakis *et al.*, 2014). Notch3 structural features were explored, parameterised and prepared as input information for our pipeline. We induced the various known mutations of Notch3 on the wild type 3D structure of Notch3 which has been made available by X-ray crystallography experiments. Then, using energy minimisation and molecular dynamics simulations, the structural significance of the induced mutations was evaluated in silico. Investigation of the resulting molecular conformations of the mutants shed light to the structural properties of the Notch3 3D organisation which lead to the activation or deactivation of the Notch3 protein and eventually to CADASIL disease.

Moreover, we already know that the 3D structural arrangement of Notch 3 protein is very similar to that of Fibrillin. A series of many Fibrillin mutations which lead to the Marfan syndrome have been published in literature so far. We mimicked those Fibrillins mutations to the Notch3 protein. Therefore, mapping all the known mutations on the Notch 3 structure and then subsequently projecting them on the 3D fold of Fibrillin have yielded invaluable results since the latter's structure and properties have been well studied and reported in tha literature.

Next, we performed homology modelling of the various mutant constructs of Notch3 protein. The homology modelling algorithm performed an initial

Papakonstantinou et al. (2019) EMBnet.journal 24, e921
http://dx.doi.org/10.14806/ej.24.0.921

partial geometry alignment for the sequence of the template proteins with the Notch3 sequence. All available substrates and ligands of Notch3 were added to the model to establish the specific interaction patterns and "key" residues involved in the ligand – Notch3 interaction. Similarities or complementarities between the surfaces of Notch3 and other X-ray determined structures have provided crucial insights on structural conservation among these proteins that relate to their mechanism of action and function.

High throughput virtual screening techniques were used to screen all available substances with orphan drug designation from EMA. To date, the EMA orphan drug database contains 1727 entries. Molecular dynamics and molecular docking simulations were applied to evaluate the association and binding efficacy of each one of the orphan drugs to the wild type Notch3 protein and the models of Notch3 mutants. Using molecular dynamics, we were able to monitor the motion of atoms and molecules in a computerised biological system.

Notch 3 structure – phenotype map

Finally, a 3D interactive Notch3 specific prediction structure –phenotype map was established (Vlachakis *et al.*, 2014). Each amino acid position on the Notch3 protein was linked to the phenotype it produces, so that the phenotype of new/future mutations may be predicted in silico. This final part of the study aims to combine all the information gathered from all previous steps in an easy to comprehend graphic representation of Notch3 protein. This way it is possible firstly to get updated with the current research on NOTCH3 and CADASIL, as well as to predict the effect that a new mutation may have, based on the position of the mutation on the 3D structure and the physicochemical alteration which the mutation induces. Emphasis was paid to the Cysteine residue mutations which usually lead to unpaired Cysteine residues and promote the CADASIL disease. Even though CADASIL is a rare disease, its linkage to NOTCH3 makes it very interesting from a basic research point of view. We believe that the establishment of an interactive map of Notch3 genotype/phenotype on the 3D structure of Notch3 can bring the scientific community up to speed with current developments in this exciting field.

CADASIL and optical coherence tomography angiography

Abnormalities on brain imaging, usually MRI, often exist in CASADIL patients long before symptoms occur, and seem to have a faster rate of progression than other clinical outcome measures, such as cognitive impairment or stroke incidence. For this reason MRI can be used as surrogate biomarkers of the disease sensitive to change in time. Using digital imaging of the retinal area has been growing increasingly common, and can be used for the analysis of vascular topography, including the width of retinal microvessels (Patton *et al.*, 2005). Retinal vasculature can be visualised in vivo and photographed in 2D. Hence, it has great potential of being used as an index, given the anatomical correlation between both the macrovascular and the microvascular blood supply, to the retina and the brain. Fractal analysis has been used to measure the complexity, or density, of the retinal vessel branching, expressed by the mean fractal dimension (mean-D) value, based on the hypothesis that reduced mean-D in CADASIL patients reflects the cerebral microvascular changes associated with the disease progression (Cavallari *et al.*, 2011).

Similarly, optical coherence tomography angiography (OCT-A), a new non-invasive imaging technique which generates volumetric angiography images of the retina, may be used as a surrogate outcome marker, though less validated compared with MRI markers (de Carlo *et al.*, 2015). Using OCT-A, it is possible to detect changes in the retinal microcirculation and generate a blood flow map. A recent work by (Nelis *et al.*, 2018), reports about a study in which they found a significant decrease in macular vessel density in the deep retinal plexus in CADASIL patients compared to healthy control, therefore supporting the use of this technology to detect the disease in asymptomatic individuals and to monitor the progression of the disease in patients.

Conclusions

Based on our already demonstrated results (via exhaustive molecular dynamics simulations) it was found that non-Cys mutations trigger significant loss of structure in the Notch3 protein, compared to the wild type. To identify the underlying mechanism of Notch3 role and implications in cell signal transduction, an investigation was performed to the nature, extent, physicochemical and structural significance of the mutant species. Preliminary in silico studies revealed a rather complex molecular mechanism on the structural level. Even though there are mainly point mutations, the effect of each one of them on the three-dimensional structure of the Notch3 protein is significant. However, in some cases, although local rearrangements in structure are observed, the overall 3D structural conformation of Notch3 remains quite unchanged. Finally, the structural similarity of Notch3 and Fibrillin was explored to transfer knowledge regarding structural characteristics and ligands from the well-studied field of Fibrillin to the Notch3 research domain.

Currently, there is none therapeutic treatment available for CADASIL and thereof no drug which can act specifically on the Notch3 protein receptor. Medical practitioners prescribe aspirin, dipyridamole, or clopidogrel, or a combination of these, which are found to limit the symptoms of the disease and to relatively slow it down. Given that all conventional attempts have failed in identifying a disease-modifying treatment, an extensive in silico analysis of the Notch3 mutations and

of the resulting angiogenic plasticity of the CADASIL phenotype on small vessels, could potentially lead to a radical early detection pipeline. The latter coupled by recent breakthroughs in OCT-A technology, image analysis and computational biology are steadily gaining ground in neurodegenerative disease treatments under the emerging prism of preventive and precision medicine.

Key Points

- The CADASIL syndrome is a hereditary dominant rare disease, caused by NOTCH3 gene mutations.
- A series of Notch 3 pathogenic mutations have been identified to cause protein misfolding and receptor aggregation.
- Accumulation and deposition of Notch 3 extracellular domain within vessel walls is a key pathological feature in CADASIL patients.
- Optical coherence tomography angiography (OCT-A) represents the most recent tool in ophthalmic imaging for the detection of early signs of small vessel diseases.
- OCT-A is a very effective and non-invasive tool for the investigation of CASADIL in asymptomatic individuals and to monitor the progression of the disease in patients.

Acknowledgements

The research was supported by a Microsoft Azure for Genomics research Grant (CRM:0740983) and by the FrailSafe Project (H2020-PHC-21-2015 - 690140) "Sensing and predictive treatment of frailty and associated co-morbidities using advanced personalized models and advanced interventions", co-funded by the European Commission under the Horizon 2020 research and innovation program. EP was supported by the State Scholarships Foundation (IKY) - European Union (European Social Fund - ESF) and Greek national funds through the action entitled "Strengthening Human Resources Research Potential via Doctorate Research" in the framework of the Operational Program "Human Resources Development Program, Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) 2014 – 2020.

References

1. Artavanis-Tsakonas, S. and M.A. Muskavitch (2010) Notch: the past, the present, and the future. Curr Top Dev Biol, **92**: p. 1-29, http://dx.doi.org/10.1016/S0070-2153(10)92001-2
2. Cavallari, M. *et al.* (2011). Fractal analysis reveals reduced complexity of retinal vessels in CADASIL. PloS one, **6**(4), e19150. http://dx.doi.org/10.1371/journal.pone.0019150
3. de Carlo, Talisa E *et al.* (2015) A review of optical coherence tomography angiography (OCTA). Int J Retina Vitreous vol. **1**:5. http://dx.doi.org/10.1186/s40942-015-0005-8
4. Guruharsha, K. G., Kankel, M. W., & Artavanis-Tsakonas, S. (2012) The Notch signalling system: recent insights into the complexity of a conserved pathway. Nat Rev Genet. **13**(9), 654–666. http://dx.doi.org/10.1038/nrg3272
5. Ioannidou, K., Vlachakis, D. *et al.* (2013) Neuroscience and Symptoms Related to the CADASIL Disease. International Journal of Systems Biology and Biomedical Technologies (IJSBBT) **2**(4):17-23
6. Joutel A, Corpechot C, Ducros A, *et al.* (1996) Notch3 mutations in CADASIL, a hereditary adult-onset condition causing stroke and dementia. Nature. **383**(6602):707-710. http://dx.doi.org/10.1038/383707a0
7. Louvi, A., & Artavanis-Tsakonas, S. (2012) Notch and disease: a growing field. Semin Cell Dev Biol. **23**(4), 473–480. http://dx.doi.org/10.1016/j.semcdb.2012.02.005
8. Muiño, E. *et al.* (2017). Systematic Review of Cysteine-Sparing NOTCH3 Missense Mutations in Patients with Clinical Suspicion of CADASIL. Int J Mol Sci. **18**(9), 1964. http://dx.doi.org/10.3390/ijms18091964
9. Nelis, P. *et al.* (2018). OCT-Angiography reveals reduced vessel density in the deep retinal plexus of CADASIL patients. Sci Rep. **8**(1), 8148. http://dx.doi.org/10.1038/s41598-018-26475-5
10. Patton, Niall *et al.* (2005) Retinal vascular image analysis as a potential screening tool for cerebrovascular disease: a rationale based on homology between cerebral and retinal microvasculatures. J Anat. **206** (4): 319-48. http://dx.doi.org/10.1111/j.1469-7580.2005.00395.x
11. Polychronidou E., Vlachakis D., Vlamos P. *et al.* (2015) Notch Signaling and Ageing. GeNeDis 2014. Adv Exp Med Biol., vol **822**. Springer, Cham. http://dx.doi.org/10.1007/978-3-319-08927-0_6
12. Poulson D. F. (1937) Chromosomal Deficiencies and the Embryonic Development of Drosophila Melanogaster. Proc Natl Acad Sci USA, **23**(3), 133–137. http://dx.doi.org/10.1073/pnas.23.3.133
13. Rutten, Julie W. *et al.* (2019) The effect of NOTCH3 pathogenic variant position on CADASIL disease severity: NOTCH3 EGFr 1–6 pathogenic variant are associated with a more severe phenotype and lower survival compared with EGFr 7–34 pathogenic variant. Genet Med. **21**(3):676-682. http://dx.doi.org/10.1038/s41436-018-0088-3
14. Sharp PM, Li WH (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15**(3): 1281–1295. http://dx.doi.org/ 10.1093/nar/15.3.1281
15. Tikka, S. *et al.* (2009) Congruence between NOTCH3 mutations and GOM in 131 CADASIL patients. Brain **132**,Pt 4: 933-9. http://dx.doi.org/10.1093/brain/awn364
16. Tournier-Lasserve E. *et al.* (1993). Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy maps to chromosome 19q12. Nat Genet. **3**:256–9. http://dx.doi.org/10.1038/ng0393-256
17. Vlachakis D., *et al.* (2014). A series of Notch3 mutations in CADASIL; insights from 3D molecular modelling and evolutionary analyses. JMolBiochem **3**(1): p. 97-105.

EMBnet.journal
Bioinformatics in Action

Salis et al. (2019) EMBnet.journal 24, e922
http://dx.doi.org/10.14806/ej.24.0.922

Research Papers

A genomic data mining pipeline for 15 species of the genus *Olea*

Constantinos Salis[1], Eleni Papakonstantinou[1], Katerina Pierouli[1], Athanasios Mitsis[1], Lia Basdeki[1], Vasileios Megalooikonomou[2], Dimitrios Vlachakis[1,3,4]✉, Marianna Hagidimitriou[1]

[1]Laboratory of Genetics, Department of Biotechnology, School of Food, Biotechnology and Development, Agricultural University of Athens, Athens, Greece

[2]Computer Engineering and Informatics Department, School of Engineering, University of Patras, Patras, Greece

[3]Lab of Molecular Endocrinology, Center of Clinical, Experimental Surgery and Translational Research, Biomedical Research Foundation of the Academy of Athens, Athens, Greece

[4]Department of Informatics, Faculty of Natural and Mathematical Sciences, King's College London, London, United Kingdom

Competing interests: CS none; EP none; KP none; AM none; LB none; VM none; DV none; MH none

Abstract

In the big data era, conventional bioinformatics seems to fail in managing the full extent of the available genomic information. The current study is focused on olive tree species and the collection and analysis of genetic and genomic data, which are fragmented in various depositories. Extra virgin olive oil is classified as a medical food, due to nutraceutical benefits and its protective properties against cancer, cardiovascular diseases, age-related diseases, neurodegenerative disorders, and many other diseases. Extensive studies have reported the benefits of olive oil on human health. However, available data at the nucleotide sequence level are highly unstructured. Towards this aim, we describe an *in-silico* approach that combines methods from data mining and machine learning pipelines to ontology classification and semantic annotation. Fusing and analysing all available olive tree data is a step of uttermost importance in classifying and characterising the various cultivars, towards a comprehensive approach under the context of food safety and public health.

Introduction

The "Big Data" era is here and now. The amount of digitised data produced in modern society is increasing at an exponential rate and is estimated to account for five Tb (terabytes) for every human by 2020 (Egan 2013). Large-scale data is being generated each second in a wide range of areas, such as social networks, business and finance, and biosciences, posing a great challenge for data collection, storage, processing and analysis. In life sciences, the revolution following next-generation sequencing (Bahassi and Stambrook, 2014; Hui, 2014; van Dijk *et al.*, 2014) the Human Genome Project (Collins *et al.*, 2003; Green *et al.*, 2015), the advances in protein structure determination (Giege, 2013; Hekmat, 2015; Gavira, 2016), the development of biomedical and health informatics and of imaging informatics (Andreu-Perez *et al.*, 2015; Binder and Blettner, 2015) have inevitably led to an unprecedented data explosion. Consequently, biological data generated by genomics, proteomics, transcriptomics and metabolomics are characterised by a higher order complexity.

The advances in bioinformatics over the last decades has dramatically empowered researchers in handling omics information. An extensive set of computational tools, algorithms and databases have been developed for data analysis (Berger *et al.*, 2013). Still, at the rate at which data is generated and the ever increasing needs for storage, processing and meaningful analysis the spotlights are on the realm of bioinformatics. Moore's law predicts that computing power and storage capacity doubles every 15 years, whereas genomic data have grown tenfold every year since 2002 (Moore, 1965; Kahn, 2011). Storage space availability and computational power cannot keep up and fulfil the needs for rapidly expanding data-driven research domains (Papageorgiou *et al.*, 2018). Genomic raw data are not always useful as they come out from NGS and Illumina pipelines. The extraction, analysis and collection of data or the way they are annotated in databases, is far from to be standardised. Furthermore, genomic datasets are packed with noise or erroneous information (Fan *et al.*, 2014).

High-performance computing, smarter and faster algorithms and parallelisation for storage and processing seem to be the answer for data handling. As an example, column-oriented database systems have outmatched raw-oriented representation for data storage, enabling higher compressibility (Abadi *et al.*, 2009). Moreover,

Article history
Received: 17 January 2019
Accepted: 27 January 2019
Published: 22 May 2019

compressive algorithms have been developed which enable direct processing of the compressed data (Loh *et al.*, 2012; Berger *et al.*, 2016). Cloud computing infrastructures either driven by major software companies such as MS Azure and AWS, or joined multi-national initiatives, such as the Elixir[1] programme in Europe, strive towards the larger goal of unified and standardised metagenomics. The present study is focused on organising and mapping all available and dispersed olive tree nucleotide sequences to characterised regions on the reference genome of the recently published wild olive tree variant (*Olea europaea var sylvestris*) (Unver *et al.*, 2017).

The olive tree is one of the most ancient plants on earth and is primarily cultivated in the Mediterranean region which produces 90% of the olive oil consumed worldwide and controls almost 80% of the market share in exports (Bartolini and Petruccelli, 2002; Vasto *et al.*, 2014). Olive oil is the principal source of healthy fatty acids of the Mediterranean cuisine and is perceived as "superfood" rich in beneficial compounds (Vasto *et al.*, 2014; Gerber and Hoffman, 2015; Martinez-Gonzalez *et al.*, 2015). Extra virgin olive oil, rich in phenolic components, such as polyphenols (Barbaro *et al.*, 2014; Rigacci and Stefani, 2016), has been extensively studied for its antimicrobial, antioxidant and anti-inflammatory effect (Cicerale *et al.*, 2012). Additional to its nutraceutical benefits, the consumption of olive oil has been associated with reduced risk of various diseases, establishing it as a medical food. Indeed, many studies have denoted the protective effects of extra virgin olive oil for cardiovascular disease (Estruch *et al.*, 2006; Estruch *et al.*, 2013), diabetes (Salas-Salvado *et al.*, 2011, Salas-Salvado *et al.*, 2014), age-related and neurodegenerative diseases (Khalatbary 2013; Rodriguez-Morato *et al.*, 2015). Olive oil phenols have also been observed in several cancer cell lines to inhibiting proliferation and promote apoptosis, thus impending tumour aggregation.

The olive tree has been the subject of intensive research, whereas little is known about the phylogenetic relationships with other species. However, the molecular bases which conceal the differences between cultivars remain poorly understood. A resourceful pipeline for the analysis of the olive tree genetic and genomic information is essential towards the extraction of reliable conclusions about the molecular mechanisms of action of the olive tree and its beneficial effects on human health. On top of that, humanity will have to deal with the impact of climate change in the following years. Species in the plants' kingdom are profoundly affected, especially the olive tree, and climate change is posing a significant risk in olive cultivars. The potentiality to cultivate in different climate conditions, and expand in non-traditional continents, is highly dependent on the genetic profile of the species. The present study is an important precursor for handling and analysing raw genomics and genetics data from plants. The aim is to fill in the gaps in such analysis through filtering, clustering and classification with the use of ontology terms to discover the relational nodes of the available information. A data mining pipeline was performed on available genomic data of several species of the *Olea* genus, and we have developed an approach that may help to annotate plant genomic sequences better.

Methods

Data Collection

The dataset of genomic sequences was built by collecting data from the Nucleotide database of the NCBI. Keywords used for the retrieval and extraction of data were: "*Olea europaea*", "*europaea*", "protein", "dna", "nucleotide", "genome", "clone", "cultivar", "wild species", "propagating material", "subspecies", "*Oleaceae*", "olive", "gene", "protein" and "*Olea*". The analysis of the collected sequences was performed on three basic layers interacting with each other: the size of sequences, ontologies and nucleotide sequence similarities.

Data Filtering – First level of analysis

The dataset of nucleotide sequences obtained was filtered using the MATLAB platform and programming language. To reduce the noise, partial and variant sequences were removed from the dataset using a set of regular expressions. The new dataset, containing only full sequences, was then split into three sub-datasets by sequences' length as follow:

Group A: sequence length ≤ 1,000 bases
Group B: 10,000 bases ≤ sequence length ≥ 1000 bases
Group C: sequence length ≥ 10,000 bases

The dataset was split by sequence length because the goal was to isolate and focus on areas which correspond to protein sequences. As a result, Group A and Group B were used in the second level of analysis.

Data Mining and Semantic – Second level of analysis

Different groups of datasets were characterised by ontologies using clustering and classification algorithms. In this direction, the Bioinformatics Toolbox[2] was mainly employed for the computation, development, acquisition and modelling as a high-performance language for computing and programming, in a user-friendly operating environment (Cai, Smith *et al.*, 2005). In this direction, on the basis of the second level of analysis, a new database was created containing individual sub-datasets including: a) chloroplast, b) mitochondria, c) microsatellite, d) cultivars, e) protein, f) helicase, g) ATPase, h) plastid, i) trn gene, j) enzyme, k) species (Figure 1). Besides, the protein dataset was further categorised into smaller individual datasets, as follows: a) ribosomal, b) phosphatase, c) E3, d) FAR, e) Fbox, f) kinase, g) zinc-finger, h) pentatricopeptide.

[1]https://elixir-europe.org/

[2]https://www.mathworks.com/products/bioinfo.html

Salis et al. (2019) EMBnet.journal 24, e922
http://dx.doi.org/10.14806/ej.24.0.922

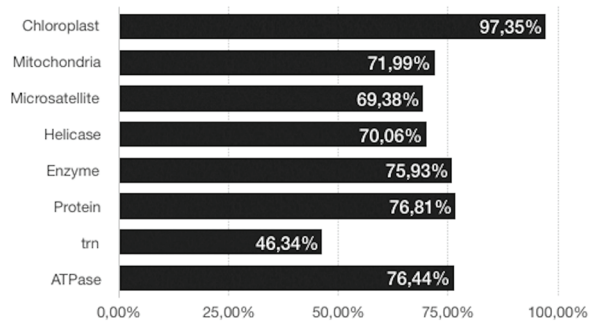

Figure 1. Percentage of *Olea europaea* nucleotide sequences in the individual sub-groups.

Analysis of genetic information – Third level of analysis

The third level of analysis consisted of grouping data obtained by the second analysis level by strict correlations of gene information. A classification function was created with the BLASTClust algorithm, in the Bio Linux operating system, to identify genetic similarity/dissimilarity between each genomic sequence BLASTClust inputs were nucleotide sequences that were analysed with the following parameters' values: coverage over 90% of the length of each sequence, with a 95% similarity cut off, and for the full-length (100% query cover) sequence, a 70% similarity cut off.

Results and Discussion

Data collection

During the data collection stage, we were able to put together more than 420000 nucleotide sequences from NCBI, which were then mapped on the genome of the wild olive tree, called "oleaster", which was assembled and annotated by Unver *et al.* (2017) (Unver, Wu *et al.*, 2017).

Data Filtering – First level of analysis

From the original pool of sequences, 8871% were classified as complete sequences and worthy of further investigation, while the remaining sequences (1129%) were partial or incomplete. The composition of the dataset was extremely heterogeneous; we indeed identified several genome regions of 15 different species of the genus *Olea* (Table 1), genome regions of 17 species of the plant kingdom and genome regions of 21 microorganisms, most of them affecting directly of indirectly the olive tree phenotype. From the filtered full-length sequences, 8674% referred to the genus *Olea* and particularly to the ontologies *europaea* and *oleaster*. In more detail, within the *Olea europaea* dataset were identified sequences with the ontologies "*europaea*", "*cuspidata*", "*laperrinei*", "*cerasiformis*", "*guanchika*" and "*maroccana*", which represent the subspecies of *Olea europaea* species, and 74% of the sequences were uncharacterised and represented as "*orphan*" sequences within *Olea europaea* species. In the remaining filtered data set, nucleotide sequences of several species of the genus *Olea* were identified, including *Olea exasperata*, *Olea capensis* with the ontologies "*hochstetteri*", "*macrocarpa*", "*enervis*", "*capensis*" and "*welwitschii*". In total, 72 cultivars were identified in the filtered dataset and another 20 cultivars discovered in the noise dataset with the partial and variant.

Regarding the split of the sequence pool by sequence length, 65,521% of the filtered sequences were in length Group A, 29,345% were in length Group B, while the rest 5,132% belonged to the length Group C.

Data mining and Semantics on Olea europaea - Second level of analysis

After the collection of all the available genetic and genomic information on the *Olea europea*, the possible relationships between the nucleotide sequences had to be identified. To this aim we needed to determine the integral nodes inside the selected dataset. Individual subgroups based on ontologies were filtered against the thousands of entries in Groups A and B. As an example,

Table 1. Ontologies per species identified in the dataset of the genus *Olea*

A/A	Species	Ontologies
1	*Olea europaea*	*Sylvestris*
		Europaea
		cuspidata/africana/indica/ferruginea
		Laperrinei
		cerasiformis
		guanchika
		maroccana
2	*Olea exasperata*	-
3	*Olea capensis*	*hochstetteri*
		macrocarpa
		enervis
		capensis
		welwitschii
4	*Olea lancea*	-
5	*Olea paniculata*	-
6	*Olea salicifolia*	-
7	*Olea rosea*	-
8	*Olea borneensis*	-
9	*Olea neriifolia*	-
10	*Olea brachiata*	-
11	*Olea javanica*	-
12	*Olea tsoongii*	-
13	*Olea schliebenii*	-
14	*Olea chimanimani*	-
15	*Olea woodiana*	*woodiana*

in the sub-dataset under the ontology "chloroplast", 1439 nucleotide sequences were clustered with a sequence average length of about ≈3841 bases 97,35% of the sequences referred to *Olea europaea*, while in the same group 12 species of the genus *Olea* were identified, three other species of the plant kingdom and four cultivars of the species *Olea europaea europaea*.

Similarly, the sub-dataset under the ontology "mitochondrial" contained 1439 sequences whose average length was about 1,241 bases. Among them, we identified two species of the genus *Olea*, 71,9% *Olea europaea* and 2,08% *Olea exasperata*, two other species of the plant kingdom and ten cultivars of the species *Olea europaea europaea*. Also, the sub-group under the ontology "micro satellite", contained 343 sequences with mean sequence length ~270 bases, was composed of 96,5% of *Olea europaea* sequences and the two subspecies, europaea and cuspidata. What is more, in the sub-group under the ontology "helicase", the 70,06% of sequences referred to *Olea europaea* and in the sub-dataset under the ontology "enzyme", *Olea europaea* sequences covered 74,27% of the dataset. The dataset under the ontology "trn" included 29 species, among which, two were *Olea europaea*, with six subspecies, and *Olea capensis*, with four subspecies.

On top of the above, in the dataset related to protein regions, 49 keywords with remarkable repeatability were identified, and 76,81% of the whole pool of sequences referred to *Olea europaea*, variant *Sylvestris*. Among the 49 keywords, eight became distinct and isolated from the dataset under the ontology "protein". The keyword "ribosomal", representing 15,27% of the dataset, "kinase" 30,06%, "E3" 9,85%, "phosphatase" 8,99%, "pentatricopeptide" 7,01%, "Fbox" 4,53% and "fatty acid- and retinol-binding protein (FAR)" 2,89% of the entire dataset, respectively. Lastly, based on the genetic information, we were able to identify nucleotide chains which bear the zinc-finger motif, representing the 8,67% of the protein dataset.

Analysis of genetic information – Third level of analysis

We were able to correlate nucleotide sequences of sub-datasets into clusters based on their genetic similarity. Clusters were made by considering that a cluster should have at least five nucleotide sequences with genetic similarity above the predefined threshold to be annotated as a cluster. Most of the sequences belonging to the length Group A, in the sub-datasets under the ontologies "trn" and "microsatellite", formed the highest number of clusters, seven and six respectively. The sub-group under the ontology "chloroplast" revealed four clusters and the sub-group under the ontology "mitochondrial" were all grouped in one cluster. The other sub-groups did not reveal any cluster. In total, the length Group A marked 19 clusters. Results showed that genomic sequences annotated as unknown sequences were clustered in an equal percentage as appeared in the initial dataset of *Olea* genus. Ultimately, the clustering results revealed that the hybrid pipeline could work well as a prediction tool.

Conclusions

This work represents the first attempt to cluster and identify olive tree cultivars based on their genome and genetic information. To date, cultivar assignment is done on the merit of morphology and pedigree in known breeds of the olive tree. However, since only recently the full genome of the *Olea europaea* variant was made public, it is now feasible to map on it all fragmented sequences of the olive tree genera and produce a set of genes or a gene panel that will be used to identify each cultivar with high accuracy genetically. Olive tree genetic fingerprinting holds great promise in the future for advanced control of olive tree breeding and olive oil that is consumed by the masses under the prism of food safety and public health.

Key Points

- Data mining and machine learning pipelines for the classification of olive tree cultivars.
- Olive tree genetic fingerprinting under the context of food safety and public health.
- Nutraceutical bioinformatics for olive oil as a medical food.

Acknowledgements

Research was supported by a Microsoft Azure for Genomics research Grant (CRM:0740983) and by the FrailSafe Project (H2020-PHC-21-2015 - 690140) "Sensing and predictive treatment of frailty and associated co-morbidities using advanced personalized models and advanced interventions", co-funded by the European Commission under the Horizon 2020 research and innovation program. EP was supported by the State Scholarships Foundation (IKY) - European Union (European Social Fund - ESF) and Greek national funds through the action entitled "Strengthening Human Resources Research Potential" via Doctorate Research in the framework of the Operational Program Human Resources Development Program, Education and Lifelong Learning of the National Strategic Reference Framework (NSRF) 2014 – 2020.

References

1. Abadi DJ, Boncz PA, Harizopoulos S (2009) Column-oriented database systems. Proceedings of the VLDB Endowment **2**(2): 1664-1665.
2. Andreu-Perez J, *et al.* (2015) Big data for health. IEEE J Biomed Health Inform **19**(4): 1193-1208, http://dx.doi.org/10.1109/JBHI.2015.245036
3. Bahassi el M and Stambrook PJ (2014) Next-generation sequencing technologies: breaking the sound barrier of human genetics. Mutagenesis **29**(5): 303-310. http://dx.doi.org/10.1093/mutage/geu031

4. Barbaro BG, *et al.* (2014) Effects of the olive-derived polyphenol oleuropein on human health. Int J Mol Sci **15**(10): 18508-18524. http://dx.doi.org/10.3390/ijms151018508
5. Bartolini G, Petruccelli R (2002) Classification, origin, diffusion and history of the olive. Food & Agriculture Org Book: ISBN-13: 978-9251048313
6. Berger BN, Daniels NM, Yu YW (2016) Computational Biology in the 21st Century: Scaling with Compressive Algorithms. Commun ACM **59**(8): 72-80. http://dx.doi.org/10.1145/2957324.
7. Berger B, Peng J, Singh M (2013) Computational solutions for omics data. Nat Rev Genet **14**(5): 333-346. http://dx.doi.org/10.1038/nrg3433.
8. Binder H, Blettner M (2015) Big data in medical science--a biostatistical view. Dtsch Arztebl Int **112**(9): 137-142. http://dx.doi.org/ 10.3238/arztebl.2015.0137.
9. Cai, J. J., Smith, D. K., Xia, X., & Yuen, K. Y. (2005). MBEToolbox: a MATLAB toolbox for sequence data analysis in molecular biology and evolution. BMC bioinformatics, **6**:64. http://dx.doi.org/10.1186/1471-2105-6-64
10. Cicerale S, Lucas LJ, Keast RS (2012) Antimicrobial, antioxidant and anti-inflammatory phenolic activities in extra virgin olive oil. Curr Opin Biotechnol **23**(2): 129-135. http://dx.doi.org/10.1016/j.copbio.2011.09.006.
11. Collins FS, Morgan M, Patrinos A (2003) The Human Genome Project: lessons from large-scale biology. Science **300**(5617): 286-290. http://dx.doi.org/10.1126/science.1084564
12. Cushman JC, Bohnert HJ (2000) Genomic approaches to plant stress tolerance. Curr Opin. Plant Biol. 3:117–124.
13. Egan BM (2013) Prediction of incident hypertension Health implications of data mining in the 'Big Data' era. J Hypertens **31**(11): 2123-2124. http://dx.doi.org/10.1097/HJH.0b013e328365b932.
14. Estruch R, Martinez-Gonzalez MA, Corella D *et al.* (2006) Effects of a Mediterranean-style diet on cardiovascular risk factors: a randomized trial. Ann Intern Med **145**(1): 1-11
15. Estruch R, *et al.* (2018) Primary prevention of cardiovascular disease with a Mediterranean diet. N N Engl J Med. **378**(25):e34. http://dx.doi.org/10.1056/NEJMoa1800389
16. Fan J, Han F, Liu H (2014) Challenges of Big Data Analysis. Natl Sci Rev **1**(2): 293-314. http://dx.doi.org/10.1093/nsr/nwt032
17. Gavira JA (2016) Current trends in protein crystallization. Arch Biochem Biophys **602**: 3-11. http://dx.doi.org/10.1016/j.abb.2015.12.010.
18. Moore GE (1965) Cramming more components onto integrated circuits. Electronics **38**(4).
19. Gerber M, Hoffman R (2015) The Mediterranean diet: health, science and society. Br J Nutr **113** Suppl 2: S4-10. http://dx.doi.org/101017/S0007114514003912.
20. Giege R (2013) A historical perspective on protein crystallization from 1840 to the present day. FEBS J **280**(24): 6456-6497. http://dx.doi.org/10.1111/febs.12580
21. Green ED, Watson JD, Collins FS (2015) Human Genome Project: Twenty-five years of big biology. Nature **526**(7571): 29-31. http://dx.doi.org/10.1038/526029a.
22. Hekmat D (2015) Large-scale crystallization of proteins for purification and formulation. Bioprocess Biosyst Eng **38**(7): 1209-1231. http://dx.doi.org/10.1007/s00449-015-1374-y.
23. Hui P (2014) Next generation sequencing: chemistry, technology and applications. Top Curr Chem **336**: 1-18. http://dx.doi.org/10.1007/128_2012_329.
24. Kahn SD (2011) On the future of genomic data. Science **331**(6018): 728-729. http://dx.doi.org/10.1126/science.1197891.
25. Khalatbary A R (2013) Olive oil phenols and neuroprotection. Nutr Neurosci **16**(6): 243-249. http://dx.doi.org/10.1179/1476830513Y.0000000052.
26. Loh PR, Baym, Berger B (2012) Compressive genomics. Nat Biotechnol **30**(7): 627-630. http://dx.doi.org/10.1038/nbt.2241.
27. Martinez-Gonzalez MA, *et al.* (2015) Benefits of the Mediterranean Diet: Insights From the PREDIMED Study. Prog Cardiovasc Dis **58**(1): 50-60. http://dx.doi.org/10.1016/j.pcad.2015.04.003.
28. Papageorgiou L, *et al.* (2018) Genomic big data hitting the storage bottleneck. EMBnet.journal **24**, e910. http://dx.doi.org/10.14806/ej.24.0.910
29. Ponti L, Gutierrez AP, Ruti PM, Dell'Aquila D (2014) Fine-scale ecological and economic assessment of climate change on olive in the Mediterranean Basin reveals winners and losers. Proc Natl Acad Sci U S A. **111**(15): 5598-5603. http://dx.doi.org/10.1073/pnas.1314437111.
30. Rigacci S, Stefani M (2016) Nutraceutical Properties of Olive Oil Polyphenols an Itinerary from Cultured Cells through Animal Models to Humans. Int J Mol Sci **17**(6). http://dx.doi.org/10.3390/ijms17060843.
31. Rodriguez-Morato J, Xicota L, Fito M, Farre M, Dierssen M, *et al.* (2015) Potential role of olive oil phenolic compounds in the prevention of neurodegenerative diseases. Molecules **20**(3): 4655-4680. http://dx.doi.org/10.3390/molecules20034655.
32. Salas-Salvado J, *et al.* (2011) Reduction in the incidence of type 2 diabetes with the Mediterranean diet: results of the PREDIMED-Reus nutrition intervention randomized trial. Diabetes Care **34**(1): 14-19. http://dx.doi.org/10.2337/dc10-1288.
33. Salas-Salvado J, Bullo N, Estruch R *et al.* (2014) Prevention of diabetes with Mediterranean diets: a subgroup analysis of a randomized trial. Ann Intern Med **160**(1): 1-10. http://dx.doi.org/10.7326/M13-1725.
34. Unver T, Wu Z, Sterck L *et al.* (2017) Genome of wild olive and the evolution of oil biosynthesis. Proc Natl Acad Sci U S A **114**(44): E9413-E9422. http://dx.doi.org/10.1073/pnas.1708621114.
35. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology. Trends Genet **30**(9): 418-426. http://dx.doi.org/10.1016/j.tig.2014.07.001.
36. Vasto S, Barera A, Rizzo C, Di Carlo M, Caruso C, *et al.* (2014) Mediterranean diet and longevity: an example of nutraceuticals? Curr Vasc Pharmacol **12**(5): 735-738.

more to it

Vivienne Baillie Gerritsen

Multitasking is not limited to computers. On a day-to-day basis, humans frequently deal with more than one thing at a time – for the sake of speed, convenience and no doubt productivity. We brush our teeth while taking a shower, eat a sandwich while answering mails, wash the dishes while calling a relative. Though we are pretty good at it, humans are far from the only multitaskers on this planet. We also harbour a few inside us. Just consider cells… One cell can synthesize proteins, while secreting others, repairing its cytoskeleton and maintaining its membrane. And we now know that the odd protein is also able to juggle with more than one task – thus applying yet another layer of obsolescence to the not-so-old "one gene, one protein, one function" hypothesis. Such is the case for a protein known as dual function macrocyclase-peptidase, or POPB. POPB is involved in making amatoxins, which are very small cyclic peptides found in some mushrooms and particularly poisonous when ingested.

Black Relationship (1924)

Wassily Kandinsky (1866-1944)

Poisonous mushrooms have been known – and avoided – by animals and human beings for thousands of years. Among them: *Galerina marginata*, a small-sized and rather plain brownish mushroom that feeds off decaying softwood and hardwood in forests of the Northern Hemisphere – from North America to Europe and Asia. *Galerina marginata* was first described in 1789 by August Batsch, a German naturalist who was a renowned mycologist. Batsch discovered almost 200 new species of mushroom, which he described in a book, "Discussion of Fungi" – a reference in the field to this day. Why are mushrooms toxic in the first place, you may ask? The most obvious answer would be for their protection: by keeping animals away, they have time to disperse their spores and proliferate. But why, then, are some mushrooms toxic while others are not? Chance, no doubt. Evolution seems to have given some mushrooms the opportunity to develop toxins, which means that they have an advantage over others.

The toxins found in *Galerina marginata* are known as amatoxins. These are small cyclic eight amino-acid peptides, whose varying side groups define the variety of toxins found within a given mushroom. Despite their modest size, amatoxins are invariably lethal. That is because their structure is particularly rigid and stable, and they are able to squeeze through membranes with surprising ease while being resistant to proteases. When ingested, amatoxins are rapidly absorbed into the bloodstream which they use to reach the liver. There, they inhibit RNA polymerase II – a polymerase directly involved in translating DNA into RNA, and hence in protein expression. This is bad news for the liver, whose vital activities are gradually hindered and shut down, leading to death unless the poison has been countered.

Where does POPB come in? *Galerina marginata* synthesizes α-amanitin. Like all

amatoxins, α-amanitin is a cyclic octapeptide. It begins as a 35 amino-acid precursor peptide that undergoes two processes – proteolysis and cyclization, and in that order – both of which are accomplished by POPB though, surprisingly, not in two successive steps. So, in effect, POPB is a sort of deferred multitasker… The intermediate severed peptide is released before it binds again to POPB – and not necessarily the same molecule – to be cyclized. Why would it do this? Researchers think it is a question of space, and that the intermediate peptide doesn't have enough room to move and present the part that needs to be cyclized. So it leaves the peptidase altogether, to come back and position itself in the right way.

POPB has two structural domains – a catalytic domain, and a seven-bladed ß-propeller domain. In the absence of substrate, the two domains rest in an open conformation, similar to the way the two shells of an open oyster would stay apart. In the presence of substrate, the propeller domain moves towards the catalytic domain, positions itself on top of it while clamping the substrate inside. Once bound to POPB, the N-terminal 10 amino-acid leader is removed from the 35 amino-acid peptide precursor and discarded. This produces a 25 amino-acid peptide with a newly exposed N-terminal and the original C-terminal tail, which is subsequently released from POPB. When the 25 amino-acid intermediate binds again to POPB, the N-terminal 8 amino acids are cyclized. As a result, both the 35 amino-acid precursors and the 25 amino-acid intermediates bind to POPB via their C-terminal tails, which sink deep into POPB's propeller domain. Which begs the question: how does POPB know that it has to cyclize part of the 25 amino-acid intermediate and not simply cut a bit off, as it does with the 35 amino-acid precursor? Because there is a short linker region between the C-terminal tail and the N-terminal octapeptide that angles the two substrates differently in POPB's active site, thus promoting proteolysis or cyclization.

The subtle and clever ways Nature has of performing various activities has inspired many a life scientist. Cyclic peptides, like amatoxins, are small, structurally varied, sturdy, resistant to proteases and oblivious to membrane permeability thus making them great candidates for designing drugs. Associated with antibody drugs, cyclic peptides can be used as powerful warheads in targeting specific molecules; as an example, when associated with antibodies against colorectal and prostrate cancer, amanita has proved to be particularly effective in this way.

In the past ten years, nine cyclic peptides have actually been approved in the fight against bacterial infection, fungal infection, cancer and gastrointestinal disorders. And there are more on the way. The thing is, cyclic peptides are more expensive to synthesize than linear peptides are and, to date, the only source for amanita is still in the wild. But if POPB could be expressed in *S.cerevisiae* – which seems to be kinetically possible – then it could cyclize all sorts of novel cyclic peptides in which could also be included unusual amino acids which would add yet other chemical properties, structures and functions to potential drugs. The possibilities seem to be not only promising, but seemingly boundless.

Cross-references to UniProt

Dual function macrocyclase-peptidase POPB, *Galarina marginata*: H2E7Q8

References

1. Czekster C.M., Ludewig H., McMahon S.A., Naismith J.H
Characterization of a dual function macrocyclase enables design and use of efficient macrocyclization substrates
Nature Communications 8:1045-1045(2017)
PMID: 29051530

2. Luo H., Hallen-Adams H.E., Scott-Craig J.S., Walton J.D.
Ribosomal biosynthesis of α-amanatin in *Galerina marginata*
Fungal Genetics and Biology 49:123-129(2012)
PMID: 22202811

proteinspotlight

Protein Spotlight (ISSN 1424-4721) is a montly review written by the **Swiss-Prot** team of the **SIB Swiss Institute of Bioinformatics**. Spotlight articles describe a specific protein or family of proteins on an informal tone. http://web.expasy.org/spotlight/

tainted

Vivienne Baillie Gerritsen

It has happened to all of us. You are seated in a good restaurant and the waiter has just brought you the wine you ordered. He solemnly shows you the label. You nod, and he proceeds to slit open the lead seal with the tip of his corkscrew. Pealing the seal off the bottle neck, he then screws the screw into the cork which he extracts with a muffled pop. He may bring the cork to his nostrils and sniff it, then with one hand behind his back, he will carefully fill the glass of the person who is to inform him whether the wine tastes fine, or not. And on rare occasions, it does not. But you're never quite sure. So you ask the person sitting opposite whether they would care to try. And they're not sure either, so you both say to the waiter that the wine is lovely, thank you. Yet, over dinner, there's a slight musty taste each time you take a sip of wine. This characteristic faint off-taste – coined cork taint – is caused by the presence of chemical molecules known as chloroanisoles in the wine. Scientists recently characterized an enzyme, that they named chlorophenol *O*-methyltransferase, from the filamentous fungus *Trichoderma longibrachiatum* that is responsible for their production.

by Edward Linley Sambourne, 1890

"Phylloxera, a true gourmet […]"

© Punch, Source: Wikipedia

This may not sound like news. For years now, wine producers have known that microbes are to blame for cork taint. However, it turns out that the microbes themselves are less to blame than the presence of chlorophenols on the cork itself. Chlorophenols are toxic chemical substances that derive from human economic activities. Among those responsible for introducing them into the environment are the pharmaceutical industry and numerous pesticides used for agricultural purposes. As the years went by, the accumulated presence of chlorophenols in our surroundings began to modify the quality of our water, soil and air, and their use in biocides has since been restricted in many countries – save the use of trichlorophenols (TCPs) in fungicides for wood, and hence cork oak (*Quercus suber*). Most corks for wine bottles are made out of the bark of cork oak. If TCPs are present on the cork stopper before bottling, and *Trichoderma longibrachiatum* is lurking there too, then the fungus will pick up the TCPs and turn them into trichloroanisoles, which is what ends up tainting the wine.

No one knows the exact circumstances that led to the discovery of wine. It is likely that one of our ancestors picked berries which were kept in a bowl a little too long, and began to ferment. Dipping a finger in the alcohol lying in the bottom will have dealt with the rest. The oldest archaeological traces of wine – understand by this fermented fruit – were found in China and date back to 7000 BC. The first evidence of wine production dates back to 4000 BC in Armenia where wild grapes still grow today. Producing wine, however, means that you have to find a means of preserving it, and this was done by storing the wine in wooden barrels lined with pine resin. The taste for wine gradually spread westwards to Europe. By the middle ages, it had become a religious symbol for the Catholic Mass, and the Benedictine monks began to produce it on an industrial scale in France and Germany. The wine industry flourished up to the late 19th Century when grape *phylloxera* devastated all of Europe's wine vineyards – save for those in the Balkans where local varieties still survive today. Since then, the art of wine-making has changed radically, and spread to other continents, such as North and South America, South Africa and Australia.

Chlorophenol O-methyltransferase, or CPOMT, is the enzyme that transforms trichlorophenols into trichloroanisoles (TCAs) in *Trichoderma longibrachiatum* through *O* methylation. CPOMT produces different kinds of TCAs, among which 2,4,6-TCA which is the essence of cork taint. CPOMT is actually induced by chlorophenols themselves. This implies that these particular chemical compounds – which stem from human industrial activities in the first place – are able to cross the fungal membrane and stimulate the synthesis of CPOMT which, in turn, transforms the toxic pollutants into the less harmful trichloroanisoles. There has not been enough time for evolution to design an enzyme whose goal is to clean up our mess. So, the methylation of TCPs into TCAs must be due to a very fortunate set of circumstances, and CPOMT's main activity must be the methylation of substrates that still remain to be identified. Nevertheless, CPOMT is indeed expressed to counter the toxic effect of TCPs and therefore represents a form of resistance to man-made pesticides.

So, in a roundabout way, CPOMT is doing us a favour. Chlorophenols (CP) can indeed do a lot of damage to humans, not to mention other animals and no doubt plants. As an illustration, scientists have been able to associate CPs with lung cancer, asthma and heart disease in people who work in rural communities, pesticide manufacturing industries, textile industries or petrochemical industries. Toxicity is due mainly to the high reactivity of the hydroxyl group in CPs that can react with proteins and nucleic acids (DNA) in the cell, thus resulting in serious cellular damage. When the hydroxyl group is blocked by *O* methylation, almost all toxicity is lost; all you get is that mildewy taste in drinking water, certain foods, dried fruit, and… Brazilian coffee…

Like their substrates, though far less harmless, chloroanisoles are also significant environmental pollutants and found in the soil, in rivers and marine sediments, in lakes and, of course, in the bark of cork oak. Though chemically very stable, they can be biodegraded by bacteria and ligninolytic fungi. Between the 1950s and 1970s, many houses in Sweden were built with wood that had been treated with pesticides that contained chlorophenols. This gradually created a major indoor environment problem with houses smelling musty and causing adverse health effects. And you only need 2-4 ng of TCAs per litre of wine for humans to perceive cork taint. However, at each sniff (which is why you end up saying to the waiter that the wine is okay after all), we smell it less because of our olfactory system's rapid habituation to TCA.

Trichoderma longibrachiatum or CPOMT may present a way of getting rid of some of these harmful chemicals we have pumped into our environment. It would be helpful to develop reliable toxicity tests to detect the presence – and levels – of chlorophenols at a specific location. The wine-making industry loses 2 to 7% of its bottles every year – which is a lot of money too – and would benefit greatly from tests that could evaluate the presence of chlorophenols in cork planks before they are used to make cork stoppers or that could discern the presence of TCAs on the corks themselves before they are inserted into bottles. Cork taint may also result from materials other than the cork stoppers themselves. Wine cellars are full of wood, from the barrels the wine is kept in to the cellar's ceiling constructions, and TCAs are particularly volatile and could easily contaminate part of the wine-producing equipment. In a way, cork taint is something of a backlash. And microbes seem to have found a way to deal with it faster than we have.

Cross-references to UniProt

Chlorophenol O-methyltransferase, *Trichoderma longibrachiatum* : D3H5H5

References

1. Coque J.-J. R., Álvarez-Rodríguez M. L., Larriba G.
Characterization of an inducible chlorophenol *O*-methyltransferase from *Trichoderma longibrachiatum* involved in the formation of chloroanisoles and determination of its role in cork taint of wines
Applied and Environmental Microbiology 69:5089-5095(2003)
PMID: 12957890

2. Feltrer R., Álvarez-Rodríguez M. L., Barreiro C., Godio R.P., Coque J.-J. R.
Characterization of a novel 2,4,6-trichlorphenol-inducible gene encoding chlorophenol *O*-methyltransferase from *Trichoderma longibrachiatum* responsible for the formation of chloranisoles and detoxification of chlorophenols
Fungal Genetics and Biology 47:458-467(2010)
PMID: 20144725

proteinspotlight

Protein Spotlight (ISSN 1424-4721) is a montly review written by the **Swiss-Prot** team of the **SIB Swiss Institute of Bioinformatics**. Spotlight articles describe a specific protein or family of proteins on an informal tone. http://web.expasy.org/spotlight/

on the right track

Vivienne Baillie Gerritsen

Left only to the passage of time, everything gravitates towards chaos. Gardens become overgrown. Roads gather potholes and cracks. Relationships wither, and teeth rot. We have ways of dealing with this however. Gardeners look after the lawns, engineers inspect the roads, therapists have a go at unravelling relationships, and dentists tend to our mouths. Life, too, has its keepers. Left unattended, the very essence of life – our DNA – will collect unfortunate mutations that have the power to wreak havoc inside us. Over time, our cells have found ways of coping with this by promoting, for example, self-destruction so as not to propagate what has become unhealthy, or by repairing damaged DNA. As a result, cells are kept on the right track. Who, though, is the keeper? Different keepers are summoned at different stages and depending on the cell's fate. One protein, however, seems to be the orchestrator. Its name is p53, and it has been studied extensively since the 1970s because when it goes wrong, life is at stake.

"Poetic Faith", by Glynis C. Tinglof

Courtesy of the artist

p53 is one of the most studied proteins ever because it is at the heart of many forms of one the most widespread diseases in human beings: cancer. In the early 1970s, a link between p53 and the formation of tumours had already been acknowledged but the viral origin of cancer was still a popular assumption among researchers. By the late 1970s, however, it became apparent that p53 was not a viral protein that had been injected into its host but rather a protein that belonged to the host itself. Though some forms of cancer are indeed caused by viruses, this particular discovery shifted the attention of scientists to the insides of humans as opposed to those of viruses. Research on p53 plodded on until the late 1980s when it literally exploded. This is because p53's role in vital activities such as cell growth and survival had been unveiled, as had its crucial involvement in DNA repair. p53 is central in spotting DNA that has been spoiled, and setting off pathways that will fix the damage. It is because of this protective role that p53 is sometimes referred to as "the guardian of the genome".

Though, paradoxically perhaps, life and evolution depend on changes that occur in DNA, some parts are best left untouched. This is no doubt why Nature designed a protein such as p53. One of p53's major roles is to protect the integrity of the genome, i.e. to fix damaged DNA that may well cause downstream harm. When DNA needs fixing, cell growth is temporarily interrupted while p53 calls for repair. If the damage is too bad, the cell will either be kept in this arrested – and harmless – state (senescence), or it will be prompted to commit suicide (apoptosis). This can be compared to a broken-down car that has been taken off the road a few days while a mechanic attends to it. If the car is beyond repair, there are chances it will be set aside in a remote corner of the garage or taken away for demolition.

In the recent past however, researchers discovered that p53 is also involved in other less life-threatening circumstances, such as cell homeostasis, tissue growth, the nestling of the blastocyst in the uterus and

old age. However, whichever way you look at it, p53 certainly seems to be central to what keeps life going, so its own regulation and activation have to be exquisitely tuned. Too much p53, and tissue growth or longevity could be touched. Not enough of it, and tumours may emerge or cells be requested to die unnecessarily. p53 must be activated at the correct time – when metabolic homeostasis is threatened, for example, or DNA is damaged. Stress signals, such as the lack of nutrients or irradiation, upregulate the synthesis of p53, while hordes of transcriptional and post-transcriptional modifications like acetylation, phosphorylation, ubiquitination and glycosylation modulate p53 stability and activity. Recently, scientists discovered that micro RNAs (miRNAs) also seemed to have an effect on p53 activity.

So what does p53 do exactly? p53 is a nuclear transcription factor whose active site has one Zn atom. Once called in for action, tetrameric p53 regulates the expression of various genes by binding to their promoters. The products of these genes then relay signals to set off metabolisms whose ultimate job is to repair damaged DNA, plunge a cell into senescence, promote cell death or assist tissue growth, cell homeostasis and longevity. It is not difficult to grasp that if p53 is dysfunctional, many important processes in a cell are hindered. In particular, cells are left to proliferate unimpeded and their DNA to accumulate mutations – which are the fodder of cancer. This is why p53 is known as a tumour suppressor, which is perhaps an anthropocentric way of saying that when p53 is healthy, it keeps an eye on the state of our DNA and makes sure cell division is normal.

What, do life scientists define as “cancer”? To cut a very complex and long story short, cancer is caused by cells that belong to us and have acquired the unfortunate faculty of multiplying in an uncontrolled manner. If the unchecked growth begins in the lungs, it causes lung cancer. If it emerges in the breast, it causes breast cancer. This unrestrained growth creates tumours – or clumps of cells – that can shed smaller tumours, or metastases, which find ways of entering the blood stream or the lymph system to invade other parts of our body where they will also grow unceasingly. At the same time, tumours gradually acquire a life of their own by stimulating the growth of tiny blood vessels that provide nutrients to keep them “alive”. Tumours become harmful to an organism because they begin to hinder the organs in which they are growing. If the organs are vital – such as our lungs, our kidneys, our liver or our brain – our life is at risk.

p53 was elected molecule of the year in the December issue of Science in 1993. Exactly 25 years ago. The tumour suppressor activity of p53 and hence its capacity to become oncogenic once mutated – lending it a Jekyll and Hyde nature – was then a novelty. A quarter of a century on, though far more is known on the molecular level, cancer still continues to kill many people around the globe. If doctors were able to check the malfunction of mutated p53 in patients, or pump functional p53 into them, would this not be a way of treating – or at least – slowing down the progression of cancer? Perhaps, yes. The fact is, mutated p53 is found in only 50% of all tumours. What is more, dysfunctional p53 is modified in different ways, which makes it all the more difficult for drugs to target. Nonetheless, there is great hope for p53 personalised therapeutics in the years to come. Thousands of years ago, human beings lived to the average age of 30, which would not have given p53 the opportunities it has today to become dysfunctional. We seem to be standing on a curious set of scales where, on one side, we have created conditions that make life longer and, on the other, we are forging an environment whose tendency is to make it shorter.

Cross-references to UniProt

Cellular tumor antigen p53, *Homo sapiens* (Human): P04637

References

1. Farnebo M., Bykov V.J.N., Wiman K.G.
The p53 tumor suppressor: A master regulator of diverse cellular processes and therapeutic target in cancer
Biochemical and Biophysical Research Communications 396:85-89(2010)
PMID: 20494116

2. Soussi T.
The history of p53 – a perfect example of the drawbacks of scientific paradigms
EMBO Reports 11:822-826(2010)
PMID: 20930848

proteinspotlight

SIB Swiss Institute of Bioinformatics

Protein Spotlight (ISSN 1424-4721) is a montly review written by the **Swiss-Prot** team of the **SIB Swiss Institute of Bioinformatics**. Spotlight articles describe a specific protein or family of proteins on an informal tone. http://web.expasy.org/spotlight/

on mar and motion

Vivienne Baillie Gerritsen

Movement is what sustains life. Organisms need to move to find food, seek shelter and to reproduce. Mobility is also essential inside organisms where cells are continuously dividing and migrating. There is also unceasing movement inside every cell where myriads of molecules are being trafficked, and cellular compartments of all shapes and sizes shifted. What keeps things moving? Years ago, scientists discovered a protein they coined actin. Actin is a small globular protein that has many different roles in eukaryotic cells. One characteristic feature is its capacity to polymerize into microfilaments that stretch from one end of a cell to another to form the cell's cytoskeleton – which speaks for itself. Though the formation of a cell's cytoskeleton is perhaps considered as actin's fundamental role in the cytoplasm, the protein is also involved in many other activities, one of which is mobility. Actin is also present in the nucleus but, until recently, scientists believed that microfilaments did not form there. It turns out that they do: damaged DNA seems to be oriented towards repair centres thanks to actin microfilaments whose growth is prompted by a protein complex known as Arp2/3.

watercolour by Giorgiana Houghton (ca 1868)

We tend to think of cells as entities with many compartments and macromolecules floating in a biological fluid, and navigation from A to B – if necessary – is merely a question of swimming there. But it is not so. As for all vehicles, energy is required to create motion, and this is what actin provides. A cell's nucleus is characterised by its DNA, usually stored in the form of chromosomes. There are also hosts of proteins that fold the DNA, protect it, transcribe it, translate it, replicate it and so on. DNA repair is also an important task that occurs in the nucleus. Damage occurs to an organism's DNA all the time, which is why cells must have repair mechanisms. Small 'point' mutations are the most frequent type of damage, but greater damage can occur when the DNA double helix snaps altogether and some bits may even get lost in the process. The cell has two ways of repairing this: either it simply sticks the loose ends back together or, first, it fills in what may have gone amiss, and then joins the ends. It is this second kind of repair that demands more craft, and where nuclear actin microfilaments are involved.

Actin is one of the most versatile proteins in eukaryotes. Its sequence has changed very little over the course of evolution and it is found in organisms as diverse as algae and humans – as a consequence, scientists consider its structure optimised. Its fundamental role is to hydrolyse ATP – the biological currency of energy – thereby releasing power. It is hardly surprising, then, that it is involved in so many different activities in the cell, i.e. cell shape, cell robustness, cell plasticity, cell adhesion, cell division and tissue stabilisation. It is also part of many cell-signalling pathways while providing a scaffold for transport inside the cell and a means to organize the cell's contents in space. Actin also belongs to more specialised structures such as flagella and cilia, and is an integral part of muscle contraction.

Actin is a small globular protein that can polymerise and depolymerise at a surprising pace. In doing so, it creates dynamic structures known as microfilaments of varying lengths and durability – it all depends on the microfilaments' function, where they are

necessary in the cell and when. Actin is also active in the nucleus where it is involved in DNA transcription and gene expression for instance. However, no one had observed nuclear actin microfilaments, and the ongoing belief was that actin acted only in its monomeric form in this part of the cell. Until recently, when researchers noticed that bits of damaged DNA were relocated to distinct parts of the nucleus for repair – and relocation was achieved by way of actin microfilaments.

Arp2/3 is an actin nucleation factor, and its presence is necessary to initiate the growth of actin microfilaments both in the cell's cytoplasm and in its nucleus. Arp2/3 is a complex of seven protein subunits, all of which have specific roles in keeping the complex together, increasing nucleation efficiency, tethering one subunit to another and so on. Phosphorylation may also be a way of fine-tuning Arp2/3 activity, and the concept of multiple versions of Arp2/3 that coexist in a cell is beginning to emerge, as opposed to only one as has been thought for the past 20 years. Arp2 and Arp3, in particular, have been dubbed "unconventional actins" because they adopt the same three-dimensional fold and form the first two subunits of a nascent microfilament.

As mentioned above, there are two ways of repairing damaged DNA, and these are non-homologous end joining (NHEJ) or homology-directed repair (HDR). NHEJ is the most straightforward and widespread way of repairing DNA: the broken ends of both DNA strands are 'simply' stuck back together. In the process, however, little bits may have fallen off the part that snapped and the repaired DNA is not identical to what it was before damage. Try snapping a chocolate bar in half and look at all the crumbs that fall to the ground. As for HDR, the nucleus makes sure that all the missing parts are added before the DNA double strands are joined together again. In this way, not only is the DNA break repaired but no information is lost either. Actin microfilaments seem to grow in the nucleus especially for HDR. Though very little is yet known on the molecular level, it could be that the HDR machinery stimulates actin polymerization, and DNA double strand breaks are hurried to the "repair centre" – although who stimulates what is difficult to unjumble.

The British physiologist W.D. Haliburton discovered actin experimentally in 1887. It took another half century for the laboratory of the Hungarian biochemist Albert Szent-Györgyi to extract pure actin from muscle in 1942. Ever since, actin has been studied extensively, yet it has taken a further 80 years to realise that actin microfilaments also exist in the nucleus. In a way, this shouldn't be surprising. Why would a protein as versatile as actin not polymerize in the nucleus too? Why would microfilaments be specific only to the cytoplasm? Is movement not as fundamental to the nucleus as it is to the cytoplasm? Why would Nature imagine a different system between two cellular compartments when it has used the same system throughout eukaryotes? Actin certainly seems to share its secrets sparingly. Or perhaps the obvious is sometimes difficult to see.

Cross-references to UniProt

Actin-related protein 2, *Homo sapiens* (Human): P61160
Actin-related protein 3, *Homo sapiens* (Human): P61158
Actin, cytoplasmic 1, *Homo sapiens* (Human): P60709

References

1. Schrank B.R., Aparicio T., Li Y., Chang W., Chait B.T., Gundersen G.G.
Nuclear ARP2/3 drives DNA break clustering for homology-directed repair
Nature 559: 61-66(2018)
PMID: 29925947

2. Pizzaro-Cerdà J., Chorev D.S., Geiger B., Cossart P.
The diverse family of Arp2/3 complexes
Trends in Cell Biology 27:93-100(2017)
PMID: 27595492

proteinspotlight

Protein Spotlight (ISSN 1424-4721) is a montly review written by the **Swiss-Prot** team of the **SIB Swiss Institute of Bioinformatics**. Spotlight articles describe a specific protein or family of proteins on an informal tone. http://web.expasy.org/spotlight/

EMBnet.journal
Bioinformatics in Action

ISSN 1023-4144

DEAR READER,

If you have any comments or suggestions regarding this journal we would be very glad to hear from you. If you have a tip you feel we can publish then please let us know. Before submitting your contribution read the "Authors guidelines"[1] and send your manuscript and supplementary files using our on-line submission system[2].

Past issues are available as PDF files from the web archive[3].

Visit EMBnet website for more information: www.journal.embnet.org

EMBNET.JOURNAL EXECUTIVE EDITORIAL BOARD

[1]http://journal.embnet.org/index.php/embnetjournal/about/submissions#authorGuidelines
[2]http://journal.embnet.org/index.php/embnetjournal/author/submit
[3]http://journal.embnet.org/index.php/embnetjournal/issue/archive

PUBLISHER

EMBnet Stichting p/a
CMBI Radboud University
Nijmegen Medical Centre
6581 GB Nijmegen
Th e Netherlands

Email: erik.bongcam@slu.se
Tel: +46-18-67 21 21