EMBnet.journal
Bioinformatics in Action

In memory of Dr. Allan Orozco

ML4Microbiome workshop 2021 - Statistical and Machine Learning Techniques for Microbiome Data Analysis

Deep Learning concepts for genomics

AND MORE...

27
2022

EMBnet.journal
Bioinformatics in Action

Contents

Editorial

The last year was a year that presented completely new challenges to the bioinformatics research community. Not only at the level of development of new algorithms and workflows for the analysis of the always increasing data produced by the life sciences community but also at the level of human interactions. Since 2020 most of our conferences and meetings were held virtually leading to a lost in human touch and social interactions at coffee breaks, occasions that usually result in new collaborations and synergies. The most used phrases in our communications became: "Can you hear me" or "you are muted". Lecturing could easily become boring also for a speaker when all participant windows started the black window march into darkness creating a communication and interaction failure. But most of us adapted and we continued our research work, efforts that resulted on publications. The articles and reports published in Volume 27 of EMBnet.journal are a small prove of that.

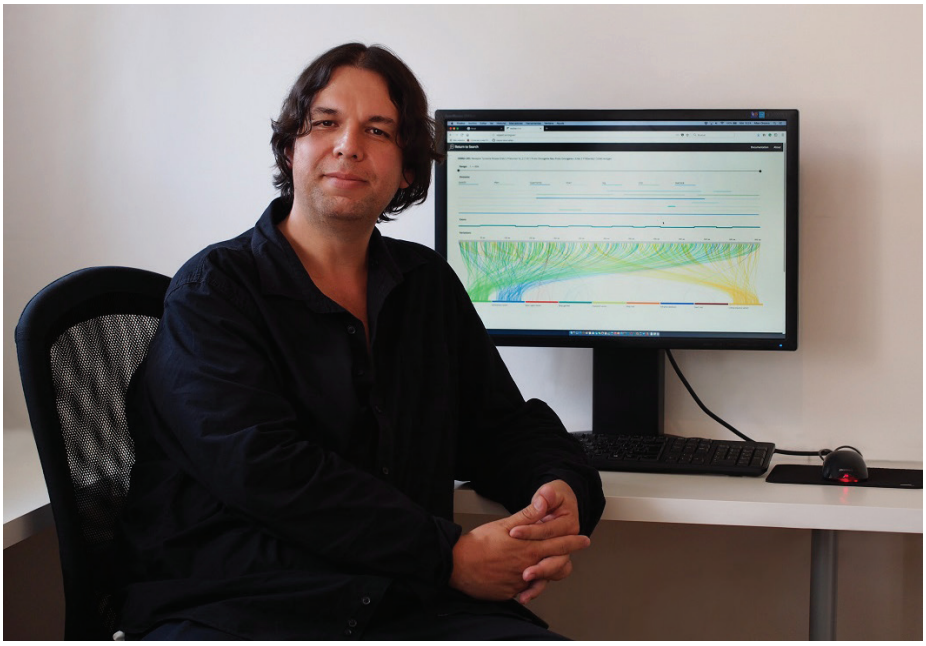
Erik Bongcam-Rudloff
Editor-in-Chief
erik.bongcam@slu.se

http://dx.doi.org/10.14806/ej.27.0.1032

EMBnet.journal
Bioinformatics in Action

Barreto-Hernández (2022) EMBnet.journal 27, e1017
http://dx.doi.org/10.14806/ej.27.0.1017

In memory of Dr. Allan Orozco



Dr. Orozco always fought against terminal diseases such as cancer (Picture: ©Simone Ecker)

Dr Allan Orozco Solano, passed away last September 26, 2021. He was our colleague and friend, a member of EMBnet for many years as EMBnet Node Manager of Costa Rica.

Dr Allan Orozco had a bachelor's degree in Engineering, a master's degree in Nanotechnology and Bioinformatics, and a Ph.D. in Bioinformatics and Biomedicine from the Universidad Autónoma de Madrid. He was an outstanding and recognised Professor at the University of Costa Rica (Faculty of Medicine and Engineering) and the National Academy of Sciences of Costa Rica (international talent) awarded him for his contributions to the scientific development of Costa Rica.

For more than 15 years, Dr Orozco was dedicated to teaching and research associated with Bioinformatics and its applications, mainly in health and biodiversity. He promoted the creation of multiple networks such as the National Network of Genomic Sequencing and Supercomputing of Costa Rica as a National Coordinator and the Central American Bioinformatics Network in 2012. At the time of his death, he was Director of the Information and Knowledge Society Program of the University of Costa Rica.

His interest in supporting the development of Costa Rica in the area of supercomputing, especially for the development of bioinformatics tools, led him to seek collaborations and partnerships with scientists and institutions in several countries. He was Co-Founder of the Iberoamerican Society for Bioinformatics (SOIBIO), and Scientific Coordinator of the National Bioinformatics Institute (INB) in Spain.

Costa Rica and we, as colleagues and friends, lament the loss of this humanistic scientist and leader dedicated to collaborating in Covid's research during his last days. However, his contributions and teachings remain to the hundreds of Costa Rican students who were fortunate enough to receive his classes. He sought to help from his field the Costa Rican community.

Prof. Emiliano Barreto-Hernández
Colombian EMBnet node manager

See also: https://www.crhoy.com/obituarios-crhoy/fallece-destacado-investigador-en-bioinformatica-dr-allan-orozco/

Article history
Received: 07 March 2022
Published: 22 July 2022

EMBnet.journal
Bioinformatics in Action

Ibrahimi et al. (2022) EMBnet.journal 27, e1012
http://dx.doi.org/10.14806/ej.27.0.1012

Reports

Report of the ML4Microbiome workshop 2021 - Statistical and Machine Learning Techniques for Microbiome Data Analysis

Eliana Ibrahimi[1,2], Ilze Elbere[3,4], Magali Berland[5], Domenica D'Elia[6]✉

[1]Department of Biology, University of Tirana, Tirana, Albania
[2]Biomedical Research Institute, Hasselt University, Hasselt, Belgium
[3]Latvian Biomedical Research and Study Centre, Riga, Latvia
[4]University of Latvia, Riga, Latvia
[5]Université Paris-Saclay, INRAE, MGP, 78350 Jouy-en-Josas, France
[6]CNR, Institute for Biomedical Technologies, Bari, Italy
Competing interests: EI none; IE none; MB none; DD none

Abstract

The bacteria living in and on us, over 100 trillion, are essential for human development, immunity and nutrition and ultimately for human health. Researchers awareness of the importance of the microbiome for human wellbeing brought in the latest decade to the flourishing of new and numerous projects focused on the human microbiome, increasing the number of large microbiome datasets available. As a result, statistical and machine learning techniques to solve microbiome data analysis challenges are in high demand. The workshop "Statistical and Machine Learning Techniques for Microbiome Data Analysis" was organised by the COST Action ML4Microbiome to introduce the main concepts of study design and statistical/machine learning techniques used in human microbiome studies to a broad community of researchers and to foster connections between the discovery-oriented microbiome and statistical/machine learning researchers inside and outside the ML4Microbiome COST Action. To this aim, the workshop was organised as part of the training programme of the GOBLET & EMBnet Annual General Meeting 2021. This allowed ML4Microbiome to reach a broad and multidisciplinary community of scientists working in Bioinformatics and Computational Biology research and education from all over the world. The workshop attracted more than 80 participants from 40 different countries. In this paper, we report about the main topics treated and discuss the attendants' feedback regarding their level of satisfaction and their specific needs/demands for possible improvement of the training offer of ML4Microbiome in the microbiome research field. Under the authors' permission, presentations were recorded and are available as an ML4Microbiome playlist on YouTube (https://www.youtube.com/channel/UCiFVD1SsJzIdPRLdJ79SNcA). The programme and presentations files are available at the ML4Microbiome website (https://www.ml4microbiome.eu/activities-and-events/ml4microbiome-workshop-on-statistical-and-machine-learning-techniques-for-microbiome-data-analysis/).

Workshop programme, aims and content

ML4Microbiome[1] - Statistical and machine learning techniques in human microbiome studies - is a COST Action (CA18131)[2] started in 2019 as an initiative of 24 European countries (currently 34) to establish a network aiming to improve the impact of microbiome research in life science research. To this aim, CA18131 specifically works to optimise and standardise statistical and machine learning techniques to analyse microbiome data (Moreno-Indias *et al.*, 2021; Marcos-Zambrano *et al.*, 2021). The correct usage of these techniques will allow researchers to better identify predictive and discriminatory 'omics' features, improve study reproducibility, and provide mechanistic insights into possible causal or contributing roles of the microbiome in human health and diseases.

[1]https://www.ml4microbiome.eu/ml4-microbiome-overview/
[2]https://www.cost.eu/actions/CA18131/

Article history
Received: 11 January 2022
Published: 21 March 2022

The workshop "Statistical and Machine Learning Techniques for Microbiome Data Analysis"[3] was held on October 15th and organised in the context of the GOBLET & EMBnet Annual General Meeting - Bioinformatics Education & Training from 2012 and beyond[4]. The organisation of the workshop in the context of this main event allowed ML4Microbiome to reach a vast audience of scientists in different disciplines and from many other countries. The GOBLET & EMBnet AGM 2021 was organised in collaboration with also the ELIXIR Italian Node[5].

Moreover, another event organised jointly with the leading conference was an ML4Microbiome Symposium entitled "Grand Challenges of Data-Intensive Science in microbiome & metagenome data analysis and training"[6].

The Symposium was held the day before the workshop (October 14th) and, the aims were, in addition to bringing to a large audience the latest advancements in the application of ML techniques to microbiome data in different research domains, also to serve as an introduction to the themes planned to be illustrated and discussed during the workshop.

The workshop's programme[3] was conceived to be of interest to a broad audience. The specific objective was to introduce attendants to the main challenges and possible solutions for optimising the experimental design and computational microbiome data analysis.

Domenica D'Elia, Leader of the ML4Microbiome Working Group 4[7], and co-organiser of the workshop along with Eliana Ibrahimi, introduced the workshop programme, the trainers' role and expertise and provided a brief description of ML4Microbiome[8].

Ilze Elbere, a researcher in molecular biology at the Latvian Biomedical Research and Study Centre (Latvia) and lecturer at the University of Latvia (Latvia) who is also currently coordinating the Latvian Microbiome project (citizen science-based initiative), presented and discussed challenges and potential pitfalls of microbiome study design, sampling and wet-lab key steps. Various aspects of each topic were presented, including a short overview of the pros and cons of currently the most often used microbiome analysis methods, as well as examples from the lecturer's practical experience and available published data. The lecture also contained information on additional resources helpful when planning and conducting microbiome studies. The slides of the Ilze presentation are available on the ML4Microbiome website[9].

Eliana Ibrahimi, lecturer and researcher in biostatistics at the University of Tirana (Albania) and research affiliate at Hasselt University (Belgium), discussed the applications of statistics in microbiome data analysis using R statistics software (Statistical Analysis of Microbiome Data with R[10]). The first part of the lecture addressed the microbiome data structure and exploration. In this part, several exploration techniques applied to explore the microbiome were discussed with the R packages that implement them, such as Phyloseq (McMurdie and Holmes, 2013). This part was followed by the univariate (parametric and nonparametric) and multivariate community analysis applications in microbiome studies. Particular attention in this part was given to the multivariate analysis of variance with permutation (PERMANOVA) (Anderson, 2017) and the analysis of group similarities (ANOSIM). Then the compositional nature of microbiome data and ways to deal with it are briefly discussed. The last part of the lecture introduced several statistical models that can successfully be applied to model microbiome data. This part discussed the application of over-dispersed and zero-inflated models (Xia *et al.*, 2018), Dirichlet-multinomial models, zero-inflated longitudinal models (Fang *et al.*, 2016), and multivariate Bayesian mixed-effects models (Grantham *et al.*, 2020) in microbiome data modelling. The implementation in R and practical examples are introduced for each model above.

Magali Berland, leader of the ML4Microbiome Working Group 3 'Optimisation and Standardization', conducted a lecture on "How Artificial Intelligence is Enhancing Human Microbiome Research"[11]. Magali Berland is a research scientist in artificial intelligence and data science for the microbiome working at MetaGenoPolis[12], an INRAE unit composed of experts in gut microbiome research applied to health and nutrition. Her lecture introduced the three types of machine learning (unsupervised learning, supervised learning and reinforcement learning) and their main mechanisms. The current application of these techniques on microbiome data has been illustrated with many examples drawn from the recent literature. Finally, the current challenges and active research areas have been outlined.

Workshop facilities, participation and evaluation

Due to the COVID-19 pandemic, the workshop was held as an online event using the Zoom conference web platform. A Google Drive shared folder was dedicated to the event to provide participants free access to the workshop material (programme and presentations) before and during the workshop. After the workshop,

[3]https://www.ml4microbiome.eu/activities-and-events/ml4microbiome-workshop-on-statistical-and-machine-learning-techniques-for-microbiome-data-analysis/
[4]https://embnet.eu/blog.html
[5]elixir-europe.org/about-us/who-we-are/nodes/italy
[6]https://www.ml4microbiome.eu/activities-and-events/ml4microbiome-symposium-grand-challenges-of-data-intensive-science-in-microbiome-metagenome-data-analysis-and-training-14-oct-2021/
[7]https://www.ml4microbiome.eu/working-group-descriptions/
[8]https://www.ml4microbiome.eu/wp-content/uploads/2021/10/ML4-welcome-at-ML4-WORKSHOP-15-oct-2021-DDElia.pdf
[9]https://www.ml4microbiome.eu/wp-content/uploads/2021/10/Microbiome_data_overview_ML4Microbiome_symposium_15.10.21_I.Elbere.pdf
[10]https://www.ml4microbiome.eu/wp-content/uploads/2021/11/Statistical-Analysis-of-Microbiome-Data-with-R-Eliana-Ibrahimi.pdf
[11]www.ml4microbiome.eu/wp-content/uploads/2021/10/AI-for-microbiome-research-Magali_Berland.pdf
[12]mgps.eu

Ibrahimi et al. (2022) EMBnet.journal 27, e1012
http://dx.doi.org/10.14806/ej.27.0.1012

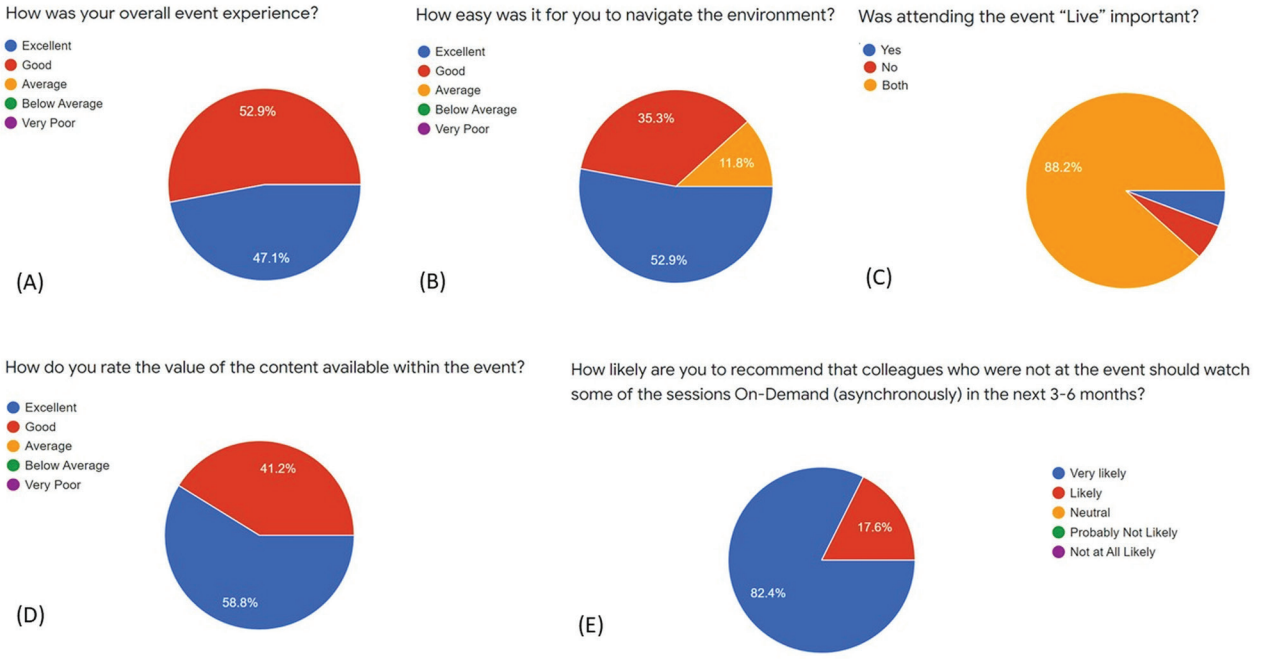


Figure 1. The figure shows the pie charts of post-workshop survey results. The survey was created using Google Drive Survey Forms facilities.

the speaker presentations were made public throughout the ML4Microbiome website[3] for to be visualised online or downloaded. Presentations from speakers were also recorded (under their permission) during the workshop, and videos are currently available as a YouTube playlist (ML4Microbiome Workshop 2021[13]) and as ML4Microbiome training material on the Action's website[14].

The registration to the event was free of charge but mandatory for organisational reasons, such as the need for moderation of access to the Zoom platform and apriori evaluation of the type of audience the workshop was going to have.

The number of people attending the workshop was 84 (including the speakers). ML4Microbiome members constituted 25%, others were from ELIXIR, EMBnet, GOBLET, and diverse Universities. Looking at attendants' geographical distribution and academic title, we have observed significant participation of young people (PhD or PostDoc) and broad geographical distribution. Many attendants were from Europe, but consistent involvement was also observed by Australia, Brazil, Ghana, India, Japan, Morocco, Mauritius, Pakistan, Perù, Sri Lanka and the USA.

After the workshop, attendants were invited to participate in a survey to collect their comments on the workshop's content and format and suggestions for improvements. People who participated in the survey were 20% of the total attendants. The main results are shown in Figure 1. The workshop experience was quoted as "Excellent" by 47,1% and as "Good" by 52,9% of attendants (Fig. 1-A). Rating the workshop content, 58,8% of participants judged it "Excellent", and 41,2% "Good" (Fig. 1-D); more than 80% claimed to be favourable to suggest colleagues not attending the workshop, to look at recorded sessions on-demand (Fig. 1-E). At the question: "From a content perspective, what could be improved?" (not shown in Fig. 1), the majority asked for more hands-on workshops.

Conclusions

From the point of view of the organisation of training initiatives, this event has provided valuable suggestions for the organisation of similar events in the upcoming year that can offer more hands-on sessions in experimental design and statistical analysis. On the contrary to what was believed in the pre-COVID era, researchers prefer online training events and appreciate the availability of recorded sessions (high demand) to be re-watched or first accessed after the event has taken place live. There were indeed many requests before and after the workshop for the availability of recorded presentations. If users' advantages are obvious, we learned with this experience that this training format also represents an advantage for trainers whose work is not just exhausted in the space of a few hours but can have a long-term value for the whole community of researchers.

[13] https://www.youtube.com/playlist?list=PL2aQNGWO1UUUmK7GJBK29yZr8Hnq6xFSD
[14] https://www.ml4microbiome.eu/training-material-2/

Acknowledgements

This article is based upon work from COST Action CA18131 ML4Microbiome, supported by COST (European Cooperation in Science and Technology).

References

AAnderson MJ (2017) Permutational Multivariate Analysis of Variance (PERMANOVA). In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels). http://dx.doi.org/10.1002/9781118445112.stat07841

Fang R, Wagner BD, Harris JK, Fillon SA (2016) Zero-inflated negative binomial mixed model: an application to two microbial organisms important in oesophagitis. Epidemiol Infect. **144**(11):2447-55. http://dx.doi.org/10.1017/S0950268816000662

Grantham NS, Guan Y, Reich BJ, Borer ET, Gross K (2020) MIMIX: A Bayesian Mixed-Effects Model for Microbiome Data From Designed Experiments. Journal of the American Statistical Association, **115**(530):599-609. http://dx.doi.org/10.1080/01621459.2019.1626242

Marcos-Zambrano LJ, Karaduzovic-Hadziabdic K, Loncar Turukalo T, Przymus P, Trajkovik V *et al.* (2021) Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. Frontiers in microbiology, **12**, 313. http://dx.doi.org/10.3389/fmicb.2021.634511

McMurdie PJ, Holmes S (2013) Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One **8**(4):e61217. http://dx.doi.org/10.1371/journal.pone.0061217

Moreno-Indias I, Lahti L, Nedyalkova M, Elbere I, Roshchupkin G *et al.* (2021) Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. Front Microbiol., **12**:635781. http://dx.doi.org/10.3389/fmicb.2021.635781

Xia Y, Sun J, Chen DG (2018) Statistical Analysis of Microbiome Data with R. ICSA Book Series in Statistics. Springer, Singapore.

EMBnet.journal
Bioinformatics in Action

Papakonstantinou et al. (2022) EMBnet.journal 27, e1005
http://dx.doi.org/10.14806/ej.27.0.1005

Reviews

The medical cyborg concept

Eleni Papakonstantinou[1], Thanasis Mitsis[1], Konstantina Dragoumani[1], Flora Bacopoulou[2], Vasilis Megalooikonomou[3], George P. Chrousos[2,4], Dimitrios Vlachakis[1,2,4]✉

[1]Laboratory of Genetics, Department of Biotechnology, School of Applied Biology and Biotechnology, Agricultural University of Athens, Athens, Greece

[2]University Research Institute of Maternal and Child Health & Precision Medicine, and UNESCO Chair on Adolescent Health Care, National and Kapodistrian University of Athens, "Aghia Sophia" Children's Hospital, Athens, Greece

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

Competing interests: EP none; TM none; KD none; FB none; VM none; GPC none; DV none

Abstract

Medical technology has made significant advances in the 21st century and, at present, medicine makes use of information technology, telecommunications, and state-of-the-art engineering to provide the best possible healthcare services. Electronic sensors provide health practitioners with the ability to constantly monitor their patients' health, to streamlines a number of medical processes, and to increase patients' access to health services. Mobile phones also empower patients and play a major role in their health's monitoring. The use of cybernetics technology can now help patients overcome even serious disabilities, enabling many disabled patients to live their lives similarly to their non-disabled fellow men through the use of artificial organs and implants. All these advances have paved the way for a more personalized type of healthcare that provides individualized solutions to each patient. Once a number of hurdles are overcome, medical technology will bring forth a new era of more precise and enabling medicine.

Introduction

The term medical technology refers to the application of scientific knowledge and skills in the form of devices, medicine, vaccines, procedures, and systems made to solve a health problem and improve patients' quality of life (Ten Haken *et al.*, 2018). The technological developments observed in the 21st century have led to tremendous changes in healthcare (Thimbleby, 2013). What could be thought of as science fiction a couple of decades prior is, now, everyday medical practice. From the use of robots (Hockstein *et al.*, 2007) to tailoring therapy to each patient's needs (Mathur and Sutton, 2017), modern medicine could be considered an advanced technological profession (Mesko, 2018). This notion is further reinforced by the introduction of modern engineering, information technology, and wireless communications in everyday medical practice.

The above technological developments have led to the emergence of novel medical fields and concepts. This review summarizes a number of ways technology has influenced current medicine and displays a number of promising future prospects.

eHealth

Electronic health (eHealth) refers to the medical field, where information and communications technologies are used in health-related services and processes. eHealth includes a variety of applications, such as electronic health records, electronic medication overviews, medical data collection, and telemedicine services (Wernhart *et al.*, 2019).

Electronic health records

Electronic health records (EHRs) refer to the digital forms of patient care records that include information, such as personal contact information, medical history, medication orders, diagnostic and laboratory test results, allergies, and treatment plans (Kruse *et al.*, 2018; Ratwani, 2017). EHRs bring multiple benefits to medical practitioners. First, EHRs allow the use of computerized clinical decision support systems (CDSSs) (Menachemi and Collum, 2011). Such systems consist of software designed to aid in clinical decision-making, where an individual patient's characteristics are

Article history
Received: 25 April 2021
Accepted: 21 May 2021
Published: 04 April 2022

Papakonstantinou et al. (2022) EMBnet.journal 27, e1005
http://dx.doi.org/10.14806/ej.27.0.1005

matched to a clinical knowledge database and patient-specific assessments or recommendations are given to the clinician for the betterment of her/his decision (Sutton *et al.*, 2020). Another benefit of EHRs is the incorporation of computerized physician order entry (CPOE) (Menachemi and Collum, 2011). CPOE systems are computer applications that offer clinicians the ability to enter electronic orders for medications, laboratory tests, imaging examinations, medical procedures, and referrals (Amiri *et al.*, 2018). The digitization of these processes reduces medication errors potentially caused by clinicians' poor penmanship and makes ordering more efficient, because nursing and pharmacy staff do not need to pursue clarifications or retrieve missing information from illegible or incomplete orders (Menachemi and Collum, 2011).

EHRs allow for the secure exchange of medical information among different health organizations and providers, promoting synergy and cooperation. This exchange's results are ideally the improved speed, accuracy, cost, and safety of medical decisions (Devine *et al.*, 2017). Last, information from EHRs can be used for secondary research purposes also. A prime example is that data in EHRs can be mined to identify previously unknown drug interactions or adverse events, which are essential research topics in pharmacology (Carroll *et al.*, 2015). Although the use of EHRs seems to be constantly increasing among healthcare providers and organizations, a number of obstacles need to be overcome. First, although legal protections have been implemented, EHRs are prone to breaches that may harm patient privacy (Shenoy and Appel, 2017). Second, although EHRs are generally thought to decrease the risk of medical errors, simple actions made possible through computerized record-keeping, like copy and pasting, may cause repeated typing errors that can potentially lead to a medical error (Palabindala *et al.*, 2016). Last, the cost of implementation and maintenance is quite high, which disincentivizes hospitals and healthcare providers from using EHRs (Palabindala *et al.*, 2016).

Wearables

A significant method of medical data collection, analysis, and storing is through wearables that accumulate real-time patient data. Wearables provide helpful information to prevent, diagnose, monitor, and manage chronic diseases and conditions (Uddin and Syed-Abdul, 2020). These devices include wristbands, smartwatches, wearable sensors, and mobile hub medical devices that collect data such as heart rate, skin temperature, galvanic skin response, skin temperature, peripheral capillary oxygen saturation, plus geolocation information and ambient environmental variables (Heikenfeld *et al.*, 2018; Witt *et al.*, 2019). The ability to provide real-time data is one of the most important assets presented by wearables. Most diagnostic tools provide information that is 'a snapshot in time' while wearables allow continuous monitoring of physiological and biochemical information under natural physiological conditions and in any environment (Aliverti, 2017).

Naturally, the information received in the 'snapshot' period may not be representative of a patient's health status, while monitoring a week's worth, or longer, of data can improve the analysis of the patient's health and help elucidate the progress of an existing disorder. The constant monitoring of an individual's current health condition by clinicians is vital in patients with chronic diseases, such as Alzheimer or Parkinson disease, or patients who are in a critical condition (Hasan *et al.*, 2019). In these cases, information received by wearables can be used to adjust medication dose, manage possible adverse events, and check patient's adherence to medical advice (Izmailova *et al.*, 2018).

In addition, wearables can be used by sports physicians. Wearables can provide information on an individual athlete's movement and physical activity, allowing sports physicians to design more efficient training programs for optimal performance (Li *et al.*, 2016). A prime example is providing real-time feedback to swimmers and runners in an effort to better their technique (Adesida *et al.*, 2019). The constant monitoring of an athlete's workload and biological parameters can help minimize the potential of injury or mitigate the effects of any existing one (Seshadri *et al.*, 2019).

Although wearables are quite promising devices when it comes to their application in eHealth, they do have a number of disadvantages. A recurring theme in the current review regarding the shortcomings of modern medical technology is privacy concerns. Wearables accumulate a large number of personal health data, where potential breaches can lead to the exploitation of patients' medical information (Cilliers, 2020). Moreover, wearable devices may lead to unintentional behavioral changes, like patients becoming overly anxious about their health and display an "addiction" to the wearable device (Schukat *et al.*, 2016).

Telemedicine

The term telemedicine refers to the delivery of healthcare services at a distance through the use of electronic technology (Serper and Volk, 2018). Telemedicine includes various practices such as over-the-phone consultation, medical advice via video-calls or e-mails, access to and sharing of medical data, and telesurgery (Zhang and Zhang, 2016). Numerous medical branches make use of telemedicine, such as radiology, pathology, dermatology, and psychology (Kruse *et al.*, 2017). Medical practitioners provide three distinct types of telemedicine. The first one is synchronous telemedicine, where patient and provider have a live interaction. The second one is asynchronous telemedicine, where a patient or physician stores medical history, images, and pathology reports and then forwards them to a specialist for diagnostic and treatment expertise. Last, there is remote patient monitoring, where a physician continuously monitors a patient's health through direct video monitoring or

reviewing continuous data received remotely (Mechanic *et al.*, 2020). Interest in telemedicine is rising since it extends the services of healthcare providers to remote areas and increases the availability of experts in specific medical fields (Kruse *et al.*, 2017).

Telemedicine showcases a number of disadvantages such as privacy concerns, but the most important is the occasional data transmission delay. In this case, the lack of dedicated and reliable networking infrastructure may lead to delays in information exchange, potentially postponing critical diagnoses and interventions (Gang *et al.*, 2017).

As mentioned above, an important application of telemedicine is telesurgery, which utilizes wireless networking and state-of-the-art robotics to allow surgeons to operate on patients who are distantly located (Choi *et al.*, 2018). Telesurgery is really important for patients who cannot afford to move out of their residence for various reasons, including risky travel, economic burden, and health reasons. Despite its' promising nature, telesurgery still remains at a halt due to lack of training programs, equipment expense, and legal issues among countries (Choi *et al.*, 2018). Thus, currently, the main method encompassing eHealth into surgical procedures is telemonitoring, where an expert surgeon guides another surgeon in a different geographical location by watching a real-time feed of the operation (Hung *et al.*, 2018).

mHealth

The advent of smartphones and tablets, mobile personal devices with computing abilities and access to the internet, seem to influence every aspect of modern life (Panova and Carbonell, 2018). These devices use microprocessors that provide computing power similar to desktop personal computers but on a smaller size and with a lower power budget (Furber, 2017). This increase in computing power and mobile connectivity has led to the emergence of a new technological field called mobile health (mHealth) (Steinhubl *et al.*, 2015). Mobile health refers to the use of mobile devices and applications (apps) to deliver healthcare services (Wilson, 2018b). These devices and applications are used both by healthcare providers and patients and have an essential role in healthcare democratization. Regarding their role in healthcare providers' work, mHealth applications allow mobile devices to function as wearable sensors or health communication hubs, becoming an integral part of the eHealth field.

Furthermore, it is expected that through modern advancements in microfluidics and microelectronics, mHealth devices could potentially act as mobile labs with diagnostic capabilities (Steinhubl *et al.*, 2015). Regarding these devices and apps' effect on patient life, the use of mHealth apps seems to have an empowering role and can help these patients manage their health. These applications are cost-effective, help patients track their health status, increase patient adherence to medical advice, and promote healthier lifestyle decisions (Mahmood *et al.*, 2019). These characteristics of mHealth make this technology very intriguing, however, there are several obstacles that deter mobile health from reaching its' full potential. Apart from the disadvantages present in wearables technology, which are also characteristics of mobile devices, health apps add new complications to the application of mHealth. These apps are currently not adequately regulated, with many of them overpromising on their healthcare potential or even being outright deceitful, thus endangering an individual's health (Steinhubl *et al.*, 2015).

The cyborg concept

The modern advancements of medical technology are more evident than ever when nowadays, the term 'cyborg' does not feel like a far-fetched science fiction concept, but a word describing everyday people (Quigley and Ayihongbe, 2018). Cyborg, short for cybernetic organism, refers to an organism that includes both biological and electronic parts (Li and Zhang, 2016). Current bionic technologies like bionic hands, leg prostheses, exoskeletons, retina-implants, and cochlear-implants have helped numerous individuals with physical disabilities (Meyer and Asbrock, 2018). People who were once perceived as 'good intentioned but lacking in physical abilities' are now thought to be equally competent -or at times- more competent than non-disabled individuals (Meyer and Asbrock, 2018).

Modern-day bionic hands allow motor control of prostheses, while rigorous research is being conducted on adding the ability of intricate sensory feedback (Bumbaširević *et al.*, 2020). These so-called myoelectric prostheses use embedded electromyography (EMG) electrodes that record the muscle's electrical activity and use it to control the prosthetic limb (Aman *et al.*, 2019). Although currently commercially available myoelectric prostheses do not offer any intentional sensory feedback, experimental approaches that use neural interfaces to stimulate peripheral nerves have been shown to elicit sensations such as pressure and pain (Aman *et al.*, 2019). Moreover, advancements in embedded systems technology appear to be quite promising regarding the betterment of control and feedback of such prostheses (Mastinu *et al.*, 2017). On the other hand, lower-limb prostheses (LLPs), like a prosthetic foot, use mechanical joint axes, compressive foams, and bumpers (Stevens *et al.*, 2018). Lower limb extremities are less complicated than the upper limbs ones, and thus patients receiving LLPs can walk, dance, or participate in sports on a level similar to that of non-disabled individuals (Bumbaširević *et al.*, 2020).

Exoskeletons, *i.e.*, wearable robotic units controlled by computer boards that power an intricate mechanical system to restore locomotion, are mainly used for rehabilitation purposes in medical centers or home use (Gorgey, 2018). Research has shown that the use of exoskeletons may have beneficial effects on gait function

Papakonstantinou et al. (2022) EMBnet.journal 27, e1005
http://dx.doi.org/10.14806/ej.27.0.1005

and walking independence in a mixed population of neurological disorders (Palermo *et al.*, 2017). Electronic retinal implants are a step towards artificial vision. Artificial vision attempts to enable some blind people to see through the electrical stimulation of the retina (Mills *et al.*, 2017). Retinal implants have shown a number of promising results, such as partial visual restoration and better performance in everyday tasks (Bloch *et al.*, 2019).

Cochlear implants are used to treat children and adults with severe to profound sensorineural hearing loss (Deep *et al.*, 2019). These implants transduce acoustic energy to an electrical signal that they later use to stimulate the auditory nerve's surviving spiral ganglion cells (Deep *et al.*, 2019). Cochlear implants are some of the most successful prostheses and have helped a tremendous number of people around the world (Wilson, 2018a). Lastly, implantable biosensors provide accurate real-time health assessment of a patient, with a prime example being implantable glucose monitors that can assess glycemia in real-time in diabetes (Waldman and Terzic, 2011).

The next step for cybernetics and medicine would be the construction of fully artificial internal organs. Several events have taken place in the last couple of decades towards the accomplishment of this goal. 3D printing, a relatively new technology that generates three-dimensional constructs from digital information, is an essential part of the scientific effort to produce artificial internal organs (Aimar *et al.*, 2019). 3D-bioprinting particularly allows printing different cell types, biomolecules, and biomaterials simultaneously (Pati *et al.*, 2016). Potential fully bio-printed organs could save lives by reducing the waiting list of patients in need of organ transplantation (Aimar *et al.*, 2019). Currently, synthetic organ devices like artificial hearts, such as the SynCadia Total Artificial Heart, are used to support patients' circulatory system, thus allowing sufficient time to find appropriate live heart transplants (Chung *et al.*, 2020).

In any case, bionic implants still showcase a number of disadvantages that should be taken into consideration. Just like regular transplantation, there is a possibility of infection (Bumbaširević *et al.*, 2020; Lewis *et al.*, 2016). This possibility makes the already highly intricate attachment of a bionic part even more difficult and demands constant monitoring after the procedure. Another potential problem is the interference caused by electronic devices. Multiple studies implicate novel devices like wireless charging systems that can generate electromagnetic fields in the intermediate frequency which could interfere with cardiac electronic implants (Driessen *et al.*, 2019). Last, many patients may themselves reject such technology for reasons such as high cost, durability, appearance, and the already mentioned lack of intrinsic feedback functions (Godfrey *et al.*, 2018).

Towards personalized medicine?

The so-called holy grail of medicine has always been to provide 'the right treatment to the right patient at the right time' (Kravitz, 2014). Recent technological advancements have set the basis for the development of more personalized medicine, one which uses both modern nucleic acid sequencing techniques and some of the aforementioned monitoring and implantation methods to individualize treatment (Goetz and Schork, 2018).

The emergence of next-generation sequencing techniques has allowed the quick and cheap sequencing of a complete human genome (Garrido-Cardenas *et al.*, 2017). By sequencing a patient's whole genome, a genomic portfolio could be constructed that showcases their possible predispositions and vulnerabilities to certain diseases (Mathur and Sutton, 2017). This information could later be included in their EHR, giving a clinician a more precise image of the patient's health (Williams *et al.*, 2019). A patient may then be monitored for specific biomarkers or behaviors that are characteristic of the pathophysiology to which they are susceptible. The monitoring could be undertaken by wearables, mobile health applications, implants, or telemedicine sessions. The above procedures may allow a clinician to propose an intervention completely aimed that specific patient for optimal outcome.

Conclusions

Medical technology has evolved by leaps and bounds in the 21st century. Information technology, telecommunications, and the ever-advancing power of computer processors have granted clinicians the ability to monitor an individual's health for great periods of time, or even continuously, as opposed to the 'snapshot' visits at a physician's office. Moreover, the inclusion of electronic health records and genomic profiles in modern hospitals allows health practitioners to tailor their therapeutic approach to each individual patient. The use of bionic technology can now overcome even life-altering disabilities. These advancements have led to the emergence of new and more precise healthcare services (Figure 1).

Just like all technological advancements, a number of hurdles should be overcome. The most important one is privacy. A patient's terminal disease, sexually transmitted disease, or even genetic profile could be used as a means of extortion by criminals. The storage and sharing of sensitive medical information should be a high priority in the coming years. Another problem is the cost of research. Medical technology is a complicated high-risk field that requires large investments. Moreover, the application of modern technology also has a high cost, both by health organizations and by customers. On the other hand, expensive technology is a bargain if it can improve quality of life, preserve economic productivity, and prevent the high cost of disability. Last, medical technology should aim to help as many people as

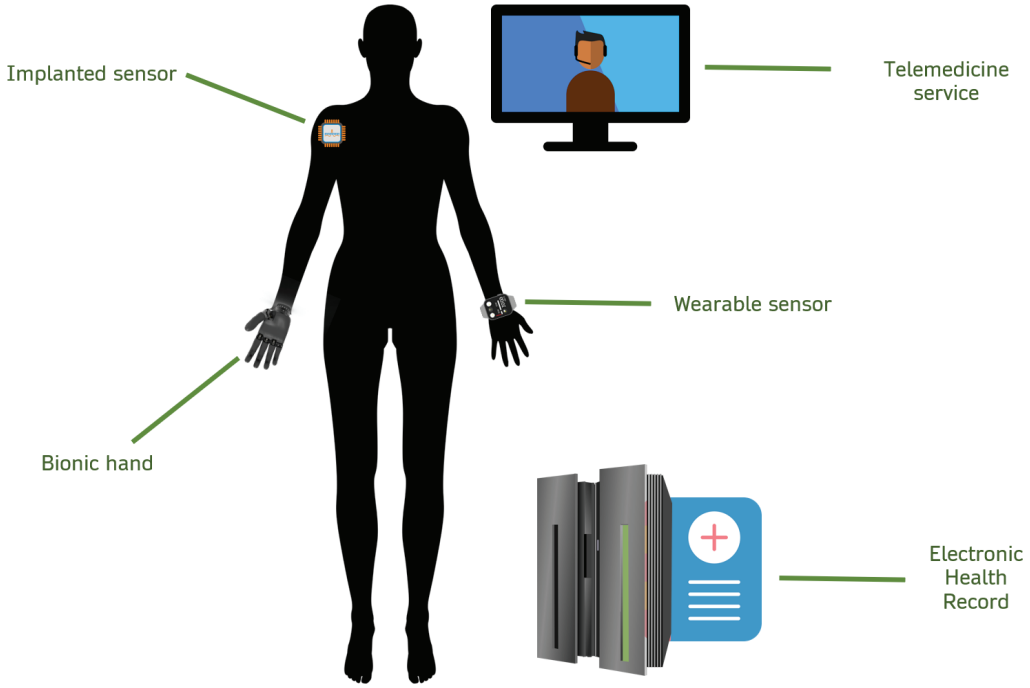


Figure 1. Modern healthcare as influenced by recent advancements in medical technology. The use of bionic limbs could potentially overcome serious disabilities. Implanted sensors can provide real-time information regarding complex biomarkers like glucose levels, while wearables can provide information on everyday fluctuations of health markers, such as body temperature and heart rate. This information could then be included in an extensive HER. Telemedicine services could then provide a patient with the appropriate advice based on the aforementioned information.

possible, thus, attributes like ease of use should be taken into account in order to provide the best possible service.

Key Points

- Cybernetics technology will enable the holistic medical, preventive and personalized perspective of health care.
- Electronic health records (EHRs) are the digital form of patient care records that set the ground for a digitalized decision support system.
- Bionic technologies like bionic hands, leg prostheses, exoskeletons, retina-implants, and cochlear-implants have helped numerous individuals with physical disabilities.
- The concepts of eHealth and mHealth inclusion in the healthcare delivery system pave the way for a tailored therapeutic approach.

References

Adesida Y, Papi E and McGregor AH (2019) Exploring the Role of Wearable Technology in Sport Kinematics and Kinetics: A Systematic Review. Sensors (Basel, Switzerland) 19(7), 1597. http://dx.doi.org/10.3390/s19071597

Aimar A, Palermo A and Innocenti B (2019) The Role of 3D Printing in Medical Applications: A State of the Art. Journal of healthcare engineering 2019, 5340616-5340616. http://dx.doi.org/10.1155/2019/5340616

Aliverti A (2017) Wearable technology: role in respiratory health and disease. Breathe (Sheffield, England) **13**(2), e27-e36. http://dx.doi.org/10.1183/20734735.008417

Aman M, Festin C, Sporer ME, Gstoettner C, Prahm C *et al.* (2019) Bionic reconstruction : Restoration of extremity function with osseointegrated and mind-controlled prostheses. Wiener klinische Wochenschrift **131**(23-24), 599-607. http://dx.doi.org/10.1007/s00508-019-1518-1

Amiri P, Rahimi B and Khalkhali HR (2018) Determinant of successful implementation of Computerized Provider Order Entry (CPOE) system from physicians' perspective: Feasibility study prior to implementation. Electronic physician **10**(1), 6201-6207. http://dx.doi.org/10.19082/6201

Bloch E, Luo Y and da Cruz L (2019) Advances in retinal prosthesis systems. Therapeutic advances in ophthalmology **11**, 2515841418817501-2515841418817501. http://dx.doi.org/10.1177/2515841418817501

Bumbaširević M, Lesic A, Palibrk T, Milovanovic D, Zoka M *et al.* (2020) The current state of bionic limbs from the surgeon's viewpoint. EFORT open reviews **5**(2), 65-72. http://dx.doi.org/10.1302/2058-5241.5.180038

Carroll RJ, Eyler AE and Denny JC (2015) Intelligent use and clinical benefits of electronic health records in rheumatoid arthritis. Expert review of clinical immunology **11**(3), 329-337. http://dx.doi.org/10.1586/1744666X.2015.1009895

Choi PJ, Oskouian RJ and Tubbs RS (2018) Telesurgery: Past, Present, and Future. Cureus **10**(5), e2716-e2716. http://dx.doi.org/10.7759/cureus.2716

Chung JS, Emerson D, Megna D and Arabia FA (2020) Total artificial heart: surgical technique in the patient with normal cardiac anatomy. Annals of cardiothoracic surgery **9**(2), 81-88. http://dx.doi.org/10.21037/acs.2020.02.09

Cilliers L (2020) Wearable devices in healthcare: Privacy and information security issues. Health Inf Manag **49**(2-3), 150-156. http://dx.doi.org/10.1177/1833358319851684

Deep NL, Dowling EM, Jethanamest D and Carlson ML (2019) Cochlear Implantation: An Overview. Journal of neurological surgery. Part B, Skull base **80**(2), 169-177. http://dx.doi.org/10.1055/s-0038-1669411

Devine EB, Totten AM, Gorman P, Eden KB, Kassakian S *et al.* (2017) Health Information Exchange Use (1990-2015): A Systematic Review. EGEMS (Washington, DC) **5**(1), 27-27. http://dx.doi.org/10.5334/egems.249

Driessen S, Napp A, Schmiedchen K, Kraus T and Stunder D (2019) Electromagnetic interference in cardiac electronic implants caused by novel electrical appliances emitting electromagnetic fields in the intermediate frequency range: a systematic review. Europace : European pacing, arrhythmias, and cardiac electrophysiology : journal of the working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology **21**(2), 219-229. http://dx.doi.org/10.1093/europace/euy155

Furber S (2017) Microprocessors: the engines of the digital age. Proceedings. Mathematical, physical, and engineering sciences **473**(2199), 20160893-20160893. http://dx.doi.org/10.1098/rspa.2016.0893

Gang W, Shan L, Mullen-Fortino M, Sokolsky O and Insup L (2017) Transmission delay performance in telemedicine: A case study. Annu Int Conf IEEE Eng Med Biol Soc 2017, 3723-3727. http://dx.doi.org/10.1109/embc.2017.8037666

Garrido-Cardenas JA, Garcia-Maroto F, Alvarez-Bermejo JA and Manzano-Agugliaro F (2017) DNA Sequencing Sensors: An Overview. Sensors (Basel, Switzerland) **17**(3), 588. http://dx.doi.org/10.3390/s17030588

Godfrey SB, Zhao KD, Theuer A, Catalano MG, Bianchi M *et al.* (2018) The SoftHand Pro: Functional evaluation of a novel, flexible, and robust myoelectric prosthesis. PloS one **13**(10), e0205653-e0205653. http://dx.doi.org/10.1371/journal.pone.0205653

Goetz LH and Schork NJ (2018) Personalized medicine: motivation, challenges, and progress. Fertility and sterility **109**(6), 952-963. http://dx.doi.org/10.1016/j.fertnstert.2018.05.006

Gorgey AS (2018) Robotic exoskeletons: The current pros and cons. World journal of orthopedics **9**(9), 112-119. http://dx.doi.org/10.5312/wjo.v9.i9.112

Hasan MK, Shahjalal M, Chowdhury MZ and Jang YM (2019) Real-Time Healthcare Data Transmission for Remote Patient Monitoring in Patch-Based Hybrid OCC/BLE Networks. Sensors (Basel, Switzerland) **19**(5), 1208. http://dx.doi.org/10.3390/s19051208

Heikenfeld J, Jajack A, Rogers J, Gutruf P, Tian L *et al.* (2018) Wearable sensors: modalities, challenges, and prospects. Lab on a chip **18**(2), 217-248. http://dx.doi.org/10.1039/c7lc00914c

Hockstein NG, Gourin CG, Faust RA and Terris DJ (2007) A history of robots: from science fiction to surgical robotics. Journal of robotic surgery **1**(2), 113-118. http://dx.doi.org/10.1007/s11701-007-0021-2

Hung AJ, Chen J, Shah A and Gill IS (2018) Telementoring and Telesurgery for Minimally Invasive Procedures. J Urol **199**(2), 355-369. http://dx.doi.org/10.1016/j.juro.2017.06.082

Izmailova ES, Wagner JA and Perakslis ED (2018) Wearable Devices in Clinical Trials: Hype and Hypothesis. Clinical pharmacology and therapeutics **104**(1), 42-52. http://dx.doi.org/10.1002/cpt.966

Kravitz RL (2014) Personalized medicine without the "omics". Journal of general internal medicine **29**(4), 551-551. http://dx.doi.org/10.1007/s11606-014-2789-x

Kruse CS, Krowski N, Rodriguez B, Tran L, Vela J *et al.* (2017) Telehealth and patient satisfaction: a systematic review and narrative analysis. BMJ open **7**(8), e016242-e016242. http://dx.doi.org/10.1136/bmjopen-2017-016242

Kruse CS, Stein A, Thomas H and Kaur H (2018) The use of Electronic Health Records to Support Population Health: A Systematic Review of the Literature. Journal of medical systems **42**(11), 214-214. http://dx.doi.org/10.1007/s10916-018-1075-6

Lewis PM, Ayton LN, Guymer RH, Lowery AJ, Blamey PJ *et al.* (2016) Advances in implantable bionic devices for blindness: a review. ANZ journal of surgery **86**(9), 654-659. http://dx.doi.org/10.1111/ans.13616

Li G and Zhang D (2016) Brain-Computer Interface Controlled Cyborg: Establishing a Functional Information Transfer Pathway from Human Brain to Cockroach Brain. PloS one **11**(3), e0150667-e0150667. http://dx.doi.org/10.1371/journal.pone.0150667

Li RT, Kling SR, Salata MJ, Cupp SA, Sheehan J *et al.* (2016) Wearable Performance Devices in Sports Medicine. Sports health **8**(1), 74-78. http://dx.doi.org/10.1177/1941738115616917

Mahmood A, Kedia S, Wyant DK, Ahn S and Bhuyan SS (2019) Use of mobile health applications for health-promoting behavior among individuals with chronic medical conditions. Digital health **5**, 2055207619882181-2055207619882181. http://dx.doi.org/10.1177/2055207619882181

Mastinu E, Doguet P, Botquin Y, Hakansson B and Ortiz-Catalan M (2017) Embedded System for Prosthetic Control Using Implanted Neuromuscular Interfaces Accessed Via an Osseointegrated Implant. IEEE Trans Biomed Circuits Syst **11**(4), 867-877. http://dx.doi.org/10.1109/tbcas.2017.2694710

Mathur S and Sutton J (2017) Personalized medicine could transform healthcare. Biomedical reports **7**(1), 3-5. http://dx.doi.org/10.3892/br.2017.922

Mechanic OJ, Persaud Y and Kimball AB (2020) Telehealth Systems (Eds.) StatPearls. StatPearls Publishing, Copyright © 2020, StatPearls Publishing LLC., Treasure Island (FL).

Menachemi N and Collum TH (2011) Benefits and drawbacks of electronic health record systems. Risk management and healthcare policy **4**, 47-55. http://dx.doi.org/10.2147/RMHP.S12985

Mesko B (2018) Health IT and digital health: The future of health technology is diverse. Journal of clinical and translational research **3**(Suppl 3), 431-434.

Meyer B and Asbrock F (2018) Disabled or Cyborg? How Bionics Affect Stereotypes Toward People With Physical Disabilities. Frontiers in psychology **9**, 2251-2251. http://dx.doi.org/10.3389/fpsyg.2018.02251

Mills JO, Jalil A and Stanga PE (2017) Electronic retinal implants and artificial vision: journey and present. Eye (London, England) **31**(10), 1383-1398. http://dx.doi.org/10.1038/eye.2017.65

Palabindala V, Pamarthy A and Jonnalagadda NR (2016) Adoption of electronic health records and barriers. Journal of community hospital internal medicine perspectives **6**(5), 32643-32643. http://dx.doi.org/10.3402/jchimp.v6.32643

Palermo AE, Maher JL, Baunsgaard CB and Nash MS (2017) Clinician-Focused Overview of Bionic Exoskeleton Use After Spinal Cord Injury. Topics in spinal cord injury rehabilitation **23**(3), 234-244. http://dx.doi.org/10.1310/sci2303-234

Panova T and Carbonell X (2018) Is smartphone addiction really an addiction? Journal of behavioral addictions **7**(2), 252-259. http://dx.doi.org/10.1556/2006.7.2018.49

Pati F, Gantelius J and Svahn HA (2016) 3D Bioprinting of Tissue/Organ Models. Angew Chem Int Ed Engl **55**(15), 4650-4665. http://dx.doi.org/10.1002/anie.201505062

Quigley M and Ayihongbe S (2018) Everyday Cyborgs: On Integrated Persons and Integrated Goods. Medical law review **26**(2), 276-308. http://dx.doi.org/10.1093/medlaw/fwy003

Ratwani R (2017) Electronic Health Records and Improved Patient Care: Opportunities for Applied Psychology. Current directions in psychological science **26**(4), 359-365. http://dx.doi.org/10.1177/0963721417700691

Schukat M, McCaldin D, Wang K, Schreier G, Lovell NH *et al.* (2016) Unintended Consequences of Wearable Sensor Use in Healthcare. Contribution of the IMIA Wearable Sensors in Healthcare WG. Yearbook of medical informatics(1), 73-86. http://dx.doi.org/10.15265/IY-2016-025

Serper M and Volk ML (2018) Current and Future Applications of Telemedicine to Optimize the Delivery of Care in Chronic Liver Disease. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association **16**(2), 157-161.e158. http://dx.doi.org/10.1016/j.cgh.2017.10.004

Seshadri DR, Li RT, Voos JE, Rowbottom JR, Alfes CM *et al.* (2019) Wearable sensors for monitoring the internal and external workload of the athlete. NPJ digital medicine **2**, 71-71. http://dx.doi.org/10.1038/s41746-019-0149-2

Shenoy A and Appel JM (2017) Safeguarding Confidentiality in Electronic Health Records. Camb Q Healthc Ethics **26**(2), 337-341. http://dx.doi.org/10.1017/s0963180116000931

Steinhubl SR, Muse ED and Topol EJ (2015) The emerging field of mobile health. Science translational medicine **7**(283), 283rv283-283rv283. http://dx.doi.org/10.1126/scitranslmed.aaa3487

Stevens PM, Rheinstein J and Wurdeman SR (2018) Prosthetic Foot Selection for Individuals with Lower-Limb Amputation: A Clinical Practice Guideline. Journal of prosthetics and orthotics : JPO **30**(4), 175-180. http://dx.doi.org/10.1097/JPO.0000000000000181

Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN *et al.* (2020) An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ digital medicine **3**, 17-17. http://dx.doi.org/10.1038/s41746-020-0221-y

Ten Haken I, Ben Allouch S and van Harten WH (2018) The use of advanced medical technologies at home: a systematic review of the literature. BMC public health **18**(1), 284-284. http://dx.doi.org/10.1186/s12889-018-5123-4

Thimbleby H (2013) Technology and the future of healthcare. Journal of public health research **2**(3), e28-e28. http://dx.doi.org/10.4081/jphr.2013.e28

Uddin M and Syed-Abdul S (2020) Data Analytics and Applications of the Wearable Sensors in Healthcare: An Overview. Sensors (Basel, Switzerland) **20**(5), 1379. http://dx.doi.org/10.3390/s20051379

Waldman SA and Terzic A (2011) Bionic technologies transforming the science of healthcare delivery. Clinical and translational science **4**(2), 84-86. http://dx.doi.org/10.1111/j.1752-8062.2011.00271.x

Wernhart A, Gahbauer S and Haluza D (2019) eHealth and telemedicine: Practices and beliefs among healthcare professionals and medical students at a medical university. PloS one **14**(2), e0213067-e0213067. http://dx.doi.org/10.1371/journal.pone.0213067

Williams MS, Taylor CO, Walton NA, Goehringer SR, Aronson S *et al.* (2019) Genomic Information for Clinicians in the Electronic Health Record: Lessons Learned From the Clinical Genome Resource Project and the Electronic Medical Records and Genomics Network. Frontiers in genetics **10**, 1059-1059. http://dx.doi.org/10.3389/fgene.2019.01059

Wilson BS (2018a) The cochlear implant and possibilities for narrowing the remaining gaps between prosthetic and normal hearing. World journal of otorhinolaryngology - head and neck surgery **3**(4), 200-210. http://dx.doi.org/10.1016/j.wjorl.2017.12.005

Wilson K (2018b) Mobile cell phone technology puts the future of health care in our hands. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne **190**(13), E378-E379. http://dx.doi.org/10.1503/cmaj.180269

Witt D, Kellogg R, Snyder M and Dunn J (2019) Windows Into Human Health Through Wearables Data Analytics. Current opinion in biomedical engineering **9**, 28-46. http://dx.doi.org/10.1016/j.cobme.2019.01.001

Zhang X-Y and Zhang P (2016) Telemedicine in clinical setting. Experimental and therapeutic medicine **12**(4), 2405-2407. http://dx.doi.org/10.3892/etm.2016.3656

EMBnet.journal
Bioinformatics in Action

Elhassani et al. (2022) EMBnet.journal 27, e990
http://dx.doi.org/10.14806/ej.27.0.990

Reviews

Deep Learning concepts for genomics: an overview

Merouane Elazami Elhassani[1,2]✉, Loic Maisonnasse[2], Antoine Olgiati[2], Rey Jerome[2], Majda Rehali[3], Patrice Duroux[1], Veronique Giudicelli[1], Sofia Kossida[1]

[1]IMGT®, The International ImMunoGeneTics Information System®, Centre National de la Recherche Scientifique (CNRS), Institut de Génétique Humaine (IGH), Université de Montpellier (UM), Montpellier, France

[2]ATOS Montpellier, River Ouest, Bezons, France

[3]Artificial Intelligence, Data Science and Emerging Systems Laboratory, National School of Applied Sciences, Sidi Mohamed Ben Abdellah University, Fez, Morocco

Competing interests: MEE none; LM none; AO none; RJ none; MR none; PD none; VG none; SK none

Abstract

Nowadays, Deep Learning is taking the world by a storm, known as a technology that makes use of Artificial Neural Networks to automatically extrapolate knowledge from a training data set, then uses this knowledge to give predictions for unseen samples. This data driven paradigm gained a widespread adoption in many disciplines, from handwriting recognition, driving an autonomous car to cracking the 50-year-old protein folding problem. With this review, we shed some light on the concepts of Deep Learning and provide some visualizations, skim over the different architectures such as Deep Neural Network (DNN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and touch upon the modern architectures such as Transformers and BERT. We also provide various examples targeting the genomics field, reference utilities, libraries useful for newcomers and disseminate our feedback.

Introduction

In few years, Deep Learning (LeCun *et al.*, 2015), a subset of Machine Learning (Figure 1), has become one of the the most successful and promising technologies, as it was able to outperform the countless methods and approaches across many fields and in diversified tasks. At the core of this paradigm shift are the "Artificial Neurons". Initially, they were conceptualized to understand and mimic the physiology and functioning of the human brain, in a computational manner (McCulloch and Pitts, 1943). Interconnecting those neurons produced Artificial Neural Networks (ANNs). Over the past decade, a wide variety of Neural Network architectures were designed (DNN, RNN, CNN, GAN...). So far, they are contentiously flourishing (Transformer, GPT, BERT...). Consequently, new state-of-the-art performances are continuously achieved with endless opportunities.

Machine Learning

Machine Learning algorithms are a subset of Artificial Intelligence. They are suitable for problems such as (Computer Vision, Voice Recognition, Face Recognition ...) that traditional programming paradigm may not solve, due to their complexities and high variability. In general, Machine Learning entails four categories of learning: Supervised Learning, Unsupervised Learning, Semi-supervised Learning and Reinforcement Learning (Figure 2).

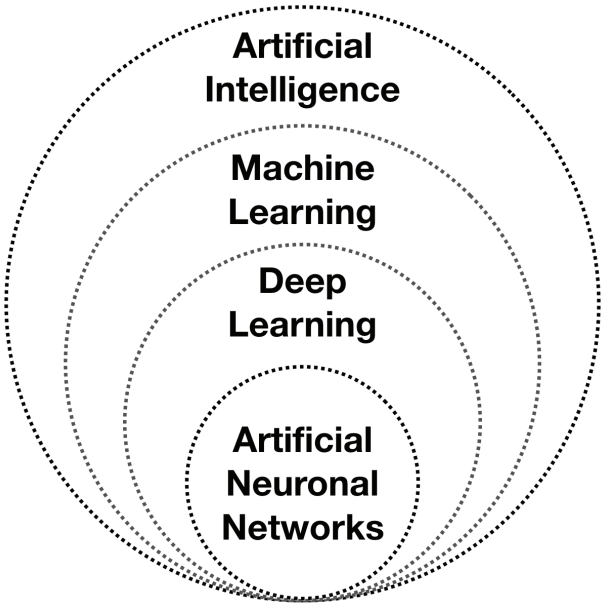


Figure 1. Machine Leaning techniques fall under Artificial Intelligence umbrella. Deep Learning is a category of Machine Learning that relays heavily on Artificial Neural Networks.

Article history
Received: 11 March 2021
Accepted: 29 April 2021
Published: 03 June 2022

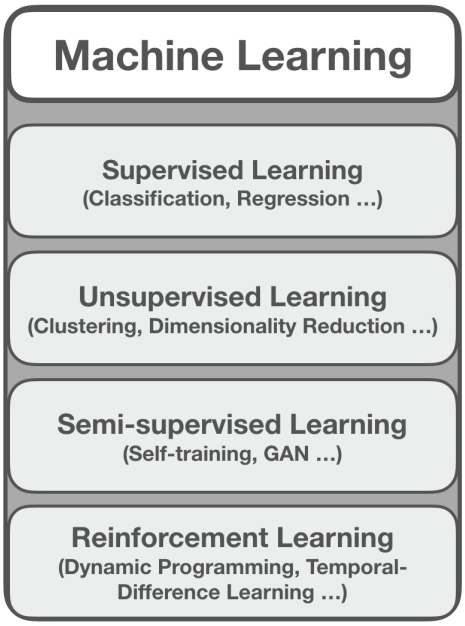


Figure 2. Machine Learning comprises four categories of learning: Supervised Learning, Unsupervised Learning, Semi-supervised Learning and Reinforcement Learning.

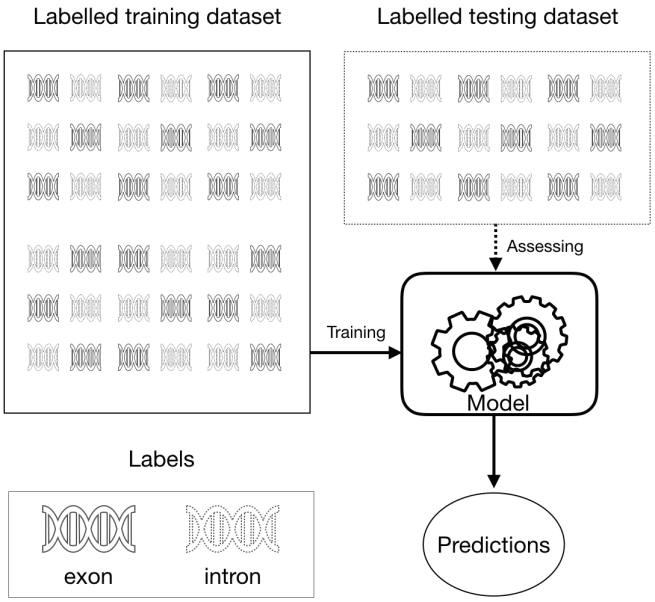


Figure 3. An example of Supervised Learning, in which the model is thought to classify the type of sequences (exon or intron). The model is trained using a labelled training dataset, then assessed using the testing dataset to validate the model performance.

Supervised Learning

A Supervised Learning (Kotsiantis *et al.*, 2006) approach consists of using the past experience data to train a Machine Learning Model to predict future ones, with respect to their classes. For instance, DECRES (Li *et al.*, 2018) is an example of Supervised Learning approach for genome-wide prediction of cis-regulatory elements. It delineated locations of 300,000 candidate enhancers genome wide and 26,000 candidate promoters (0.6% of the genome). MPRA-DragoNN (Movva *et al.*, 2019) is another example, used for deciphering regulatory DNA sequences and noncoding genetic variants. It employs a (CNN)-based architecture to predict and interpret the regulatory activity of DNA sequences as measured by MPRAs. For a pedagogical purpose and to explain this concept, let us imagine a fictional system that tries to predict, for a provided DNA sequence as input, it is either an exon or an intron. This kind of problematic is called a "Binary classification problem", in which the model tries to predict from a given input, which class of output it belongs to. With enough labelled data available, one can train a Deep Learning Model in supervised manner to resolve this classification problem, (Figure 3). A Supervised Learning approach relies heavily on the labelled samples within the dataset, they teach the Deep Learning model how to differentiate between the various classes of the output to make accurate predictions.

Unsupervised Learning

An Unsupervised Learning (Ghahramani, 2004) occurs when the available data are not labelled at all. Its key concept is to find clues or features to cluster data and reveal their hidden relationships. This category of learning comprises different learning families: Clustering (Figure 4), Dimensionality Reduction and Generative algorithms. For instance, clustering the single-cell RNA-seq data (Kiselev *et al.*, 2019) discusses the challenges and strategies used to cluster the RNA-seq.

Semi-supervised Learning

Semi-supervised Learning lays in the middle between Supervised and Unsupervised Learning. It is often used when the available data are partially labelled. For instance, the Semi-supervised Learning was employed to classify microRNA in order to maximize the utility of both labelled and unlabelled data (Sheikh Hassani and Green, 2019). The results outperformed the state-of-the-art miRDeep2 (Friedländer *et al.*, 2012) and miPIE (Peace *et al.*, 2018) methods, with an improved performance of 8.3% and 4.2% in average AUPRC.

Reinforcement Learning

In contrast to the other learning algorithms, the Reinforcement Learning (RL) (Sutton and Barto, 2018) has the particularity to be suited for unknown environments. RL was used for drug design (Popova *et al.*, 2018), genome Assembly (Xavier *et al.*, 2020), protein interaction network constructing (F *et al.*, 2015), where an infinite number of possibilities exist. RL formalizes the problem as a learning agent that interacts with its environment through a series of selected actions. Each action generates a new state that may impact the environment. The agent tries to learn how to interact as well as possible with this environment by selecting the best possible actions, which are rewarded as good or

Elhassani et al. (2022) EMBnet.journal 27, e990
http://dx.doi.org/10.14806/ej.27.0.990

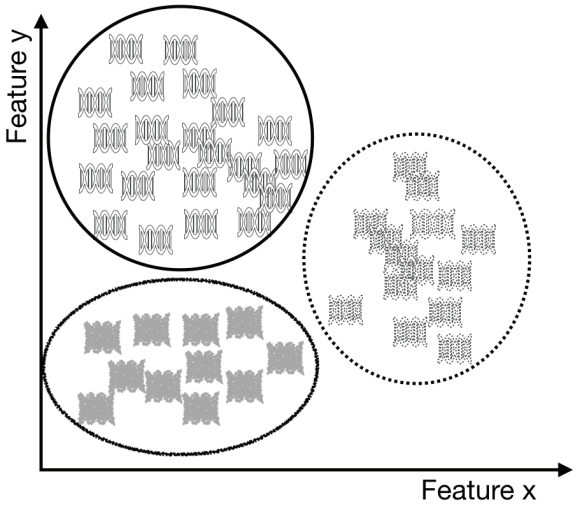


Figure 4. An example of Unsupervised Learning (Clustering), in which the model does not have any clues about the input, but it was able to distinguish 3 distinct clusters of sequence data. The samples from each group share a certain number of features, thus they appear close to each other.

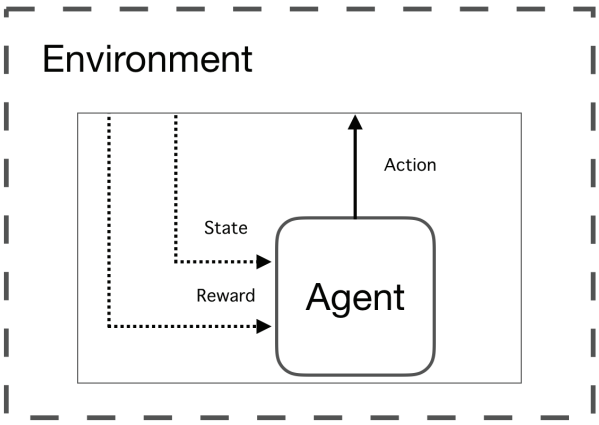


Figure 5. Reinforcement Learning agent interacting with its environment. As the agent emits an action (A_i), the environment produces a new state (S_i+1) and a new reward (R_i+1).

bad, depending on the newly generated state and their impacts on this environment, (Figure 5).

Deep Learning

Deep Learning is a subset of Machine Learning that demonstrated a great potential in several areas. In genomics for instance, it was used to learn and represent the hierarchical organization of yeast transcriptomics machinery (Chen *et al.*, 2016), to understand gene regulation (Singh *et al.*, 2017), to predict enhancer-promoter interaction from genomic sequence (Singh *et al.*, 2019), to create artificial human genomes using generative models (Yelmen *et al.*, 2019), to predict cell type specific transcription factor binding from nucleotide-resolution sequential data (Quang and Xie, 2019), to model and design protein structure (Gao *et al.*, 2020) and so on. In essence, DL is characterized by the use of Artificial Neurons that serve as the building blocks for the different Neural Network architectures.

A brief history

Deep Learning is not new, many of its concepts date back to more than half a century. Initially, (McCulloch and Pitts, 1943) put the first brick and proposed a simplified version of a neuron, as an attempt to understand how the brain could produce very complex patterns using only simple interconnected cells (biological neurons). (Rosenblatt, 1957) proposed the Perceptron, a simplified neuron which had true learning abilities for doing a binary classification. Afterwards, came the first Feedforward Multilayered Neural Networks (Ivakhnenko and Lapa, 1966) (interconnected layered neurons). A breakthrough was made by (Le Cun *et al.*, 1989), who was able to train a Convolutional Neural Network (CNN) named LeNet to recognize handwritten digits. Decades later, a Deep Convolutional Neural Networks model "AlexNet" (Krizhevsky *et al.*, 2012) made a quantum leap in computer vision. It won the ImageNet Large Scale Visual Recognition Challenge 2012 with a phenomenally great margin. Also, it demonstrated for the first time, the automatic feature learning aspect. Recently, DeepMind (Senior *et al.*, 2020) was able to crack the 50-year-old Protein Folding Prediction problem, which is another milestone for understanding biology using Artificial Intelligence.

Underlying Concepts

Artificial Neurons

Artificial Neurons or simply neurons are the backbone of the Deep Learning technology. A neuron is comprised of a single input layer and one output node, the input layer contains n nodes that transmit n features. The output y is calculated using the inputs and their weights (*i.e.* the weighted sum), where each x_i from the input vector $X = [x_1,\ldots,x_n]$ is multiplied respectively with its w_i from the learning weight vector $W = [w_1,\ldots,w_n]$, an additional bias variable b is added to capture the invariant part of the prediction. An activation function A is applied to the weighted sum which decides whether the output of this neuron should be activated or not, (Figure 6).

$$\hat{y} = A(W.X + b) = A\left(\sum_{i=1}^{n} w_i . x_i + b\right) \qquad (1)$$

Activation function

The main idea behind the activation function (Nwankpa *et al.*, 2018) is adding a non-linearity to the neural network, which allows the network to capture more complex pattern inherent within the data. The choice of this function impacts profoundly the design of a neural network, it is a task specific. For instance, a binary classification will definitely have a different activation function then a multi-class probability classification. A multitude of activation functions exist, (Nwankpa *et al.*, 2018) compiles majority of them and outlines the current trends in the applications and usage of these functions in practical deep learning deployments against the state-of-the-art research results.

Loss function

One of the most important questions in designing a Neural Network architecture is how to gauge its performance. The loss or cost function is the answer. It quantifies how the Neural Network model is performing by calculating the difference between the output y and the predicted output ŷ. In other words, it measures how far or close the predictions are from the excepted values. Many loss functions exist, selecting a particular one impacts profoundly the learning process of the Neural Network (Hennig and Kutlukaya, 2007). Advanced analyses were conducted by (Wang *et al.*, 2020) that summarize and analyse 31 classical loss functions in Machine Learning.

Optimization

At the outset, most of the Neural Network models will not have the best performances (a poor prediction and very high Loss). During the training, an optimizer algorithm is configured with the model. Its goal is to minimize the loss function and maximize the prediction by tuning parameters of the model in response to the output of the loss function, (Figure 7). Several optimization algorithms exist, (Choi *et al.*, 2020) empirically compared them.

Neural Networks

A single neuron may not be enough to learn all the necessary features for complex tasks. Interconnecting those neurons produces advanced architectures called Neural Networks, as they are capable of learning more complex patterns than a single neuron, (Figure 8). A Neural Network is organized in form of layers, each layer contains a number of neurons. The layers in between the input and output layers are called the hidden layers. A Neural Network that has only one or two hidden layers is named a “Shallow Neural Network” as opposed to a “Deep Neural Network” which has several hidden layers. DeeplyEssential (Hasan and Lonardi, 2020) is an example of the Deep Neural Network, conceived to predict the critical genes for the survival and reproduction of bacteria and microbes. It consists of an input followed by six hidden layers, in which the rectified linear unit (ReLU) is used as the activation function. The output has two classes (binary classification) where Sigmoid is used

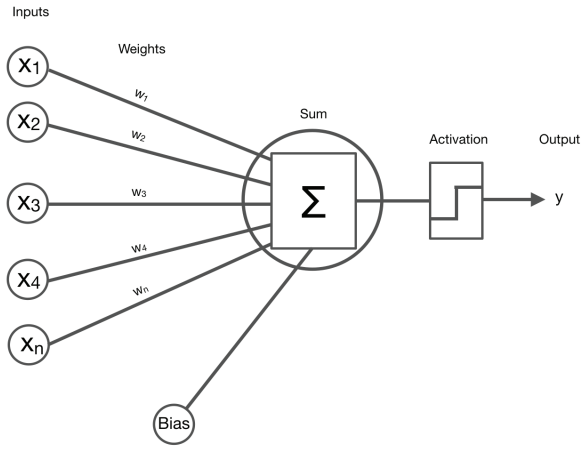


Figure 6. A single neuron comprised of inputs and their respective weights, the bias, the arithmetic operation, the activation function and the output, mathematically expressed with the formula (1]).

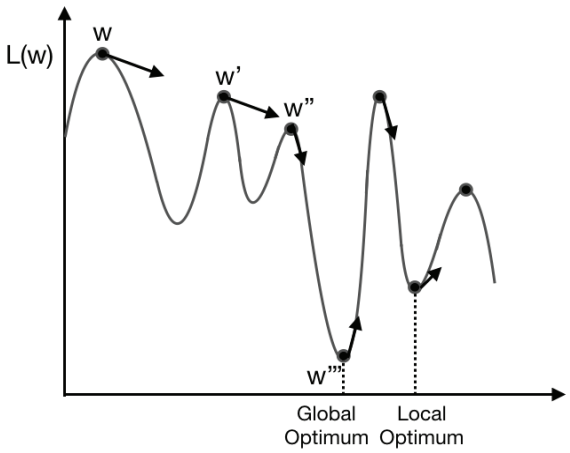


Figure 7. The optimizer looks to minimize the loss function by tuning the weights of the neurons.

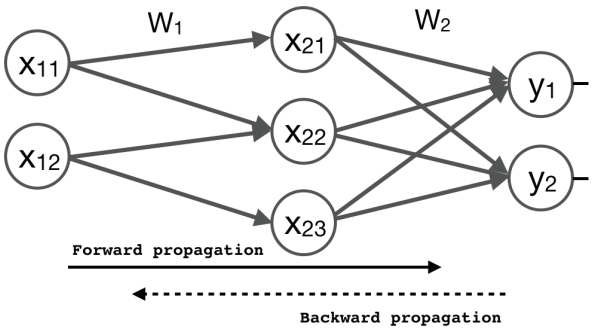


Figure 8. An example of a Shallow Neural Network comprised of an input layer (two neurons [x_{11},x_{12}]), a hidden layer in the middle, (three neurons [x_{21},x_{22},x_{23}]) and the output layer (2 neurons [y_1,y_2]).

Elhassani et al. (2022) EMBnet.journal 27, e990
http://dx.doi.org/10.14806/ej.27.0.990

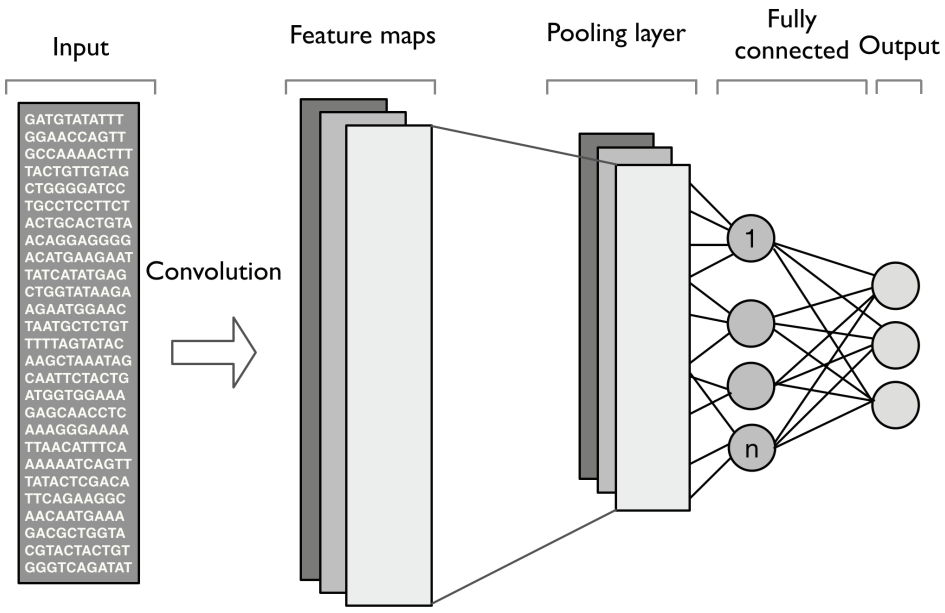


Figure 9. A example of a Convolution Neural Network comprised of the input, a Convolution layer, Feature maps, Pooling layer, Fully connected layer and the output.

as the activation function. The binary cross-entropy is chosen as the loss function.

Deep Learning architectures

Convolutional Neural Network

Convolutional Neural Network (CNN) is one of the most successful Neural Network architectures. Originally, it was inspired from the work of (Hubel and Wiesel, 1962) to understand the cat's visual cortex. CNN became so famous and has a wide variety of applications. For instance, it was successfully implemented to detect COVID-19, given a chest X-ray image. It predicts if the patient has the COVID-19 or not with 98.92% average accuracy (Irmak, 2020). CNN was also used to learn the functional activity of DNA sequences from genomics data. It was trained on a compendium of accessible genomic sites mapped in 164 cell types by DNase-seq and demonstrated greater predictive accuracy than previous methods (Kelley *et al.*).

CNN architecture consists of an input layer, followed by a Convolution layer that produces feature maps, then hidden layers (the Pooling layers, Normalization) and finally the output, (Figure 9).

Convolution layer

The outstanding capacity of CNN is owed to its ability to analyse spatial information and automatically extract features with the help of a convolution operation. The convolution operation is a mathematical operation between the input and the kernels, (Figure 10). Numerous convolution operations exist, such as standard convolution, dilated convolution, transposed convolution and separable convolution.

Pooling layer

The Pooling layer aims to reduce the resolution of the feature maps produced by the Convolution layer. The most famous pooling layers are: Average and Max Pooling layers, (Figure 11). Average Pooling takes the average of a range of numbers. Max Pooling takes the maximum number within a range of numbers.

Padding

Padding is simply the process of adding layers of zeros to the input, as to avoid the problem of losing values on corners or shrinking the input, (Figure 12).

Stride

While convolving, the stride describes how the convolution window moves over the input. By default, it slides by one at each step, (Figure 13).

Fully connected layer

This layer often appears at the end of the CNN architectures to sum the features produced by the previous layers and make predictions for the output, (Figure 9).

Recurrent Neural Network

Recurrent Neural Network (RNN) architecture is specially designed for sequential data, such as genomics or text. It is able to model space-temporal structures. Thanks to its hidden states that serve as a memory and keep track of the previous state. They provide a context for the current prediction, (Figure 14). For instance, RNN-VirSeeker (Liu *et al.*, 2020) successfully used RNN to outperform three widely used methods: VirSorter, VirFinder and DeepVirFinder in identifying short viral sequences, by obtaining 92% precision for sequences of 500bp.

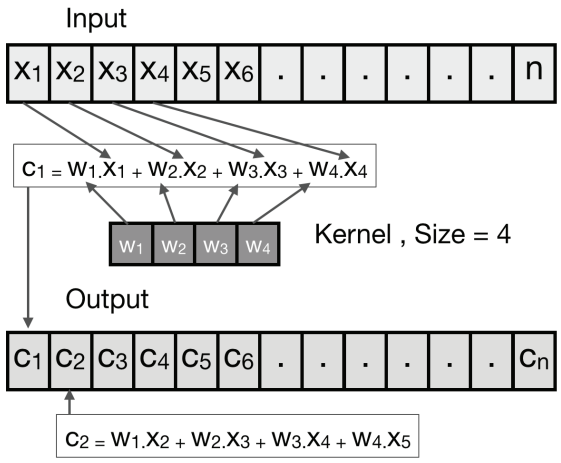


Figure 10. An example of a standard convolution calculation. The window size and kernel is 4. The output c_i is calculated based on the input and the kernel.

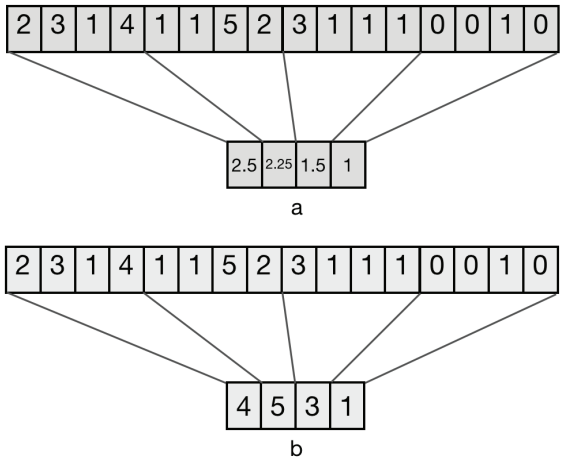


Figure 11. (a) Average Pooling operation calculates the average value for a given range. **(b)** Max Pooling operation looks for the maximum number in a given range.

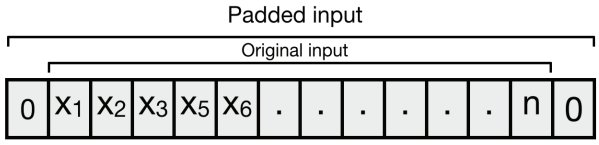


Figure 12. An example of zero padding, where zeros are appended and prepended to the original input to avoid losing x_1 and x_n values.

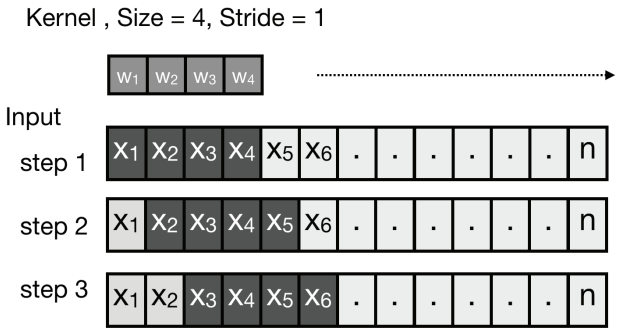


Figure 13. An example of a convolutional filter that convolves over the input. Its window size is four and it strides by one at each step.

One drawback of the vanilla RNNs is the long-term dependencies problem, formulated as Exploding and Vanishing Gradients. As a remedy, special variant of RNNs named respectively, "Long Short Term Memory"(LSTM) and "Gated Recurrent Unit"(GRU) are introduced.

Long Short Term Memory

Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) is a variant of the RNN architecture that tries to solve the Vanishing Gradient Problem. Internally, the LSTM cell differs from the RNN cell. It has three control gates: forget, update and output gates, (Figure 15). The Forget gate allows the cell to forget information in the cell state.

The Update gate allows the cell to place a new value in the memory (cell state).

The Output gate applies the Sigmoid activation function and produces the output of the current timestep. ProLanGO (Cao *et al.*, 2017) is an example of LSTM that predicts protein functions. It converts the protein sequences into a language space "ProGO" based on the frequency of k-mers (around 500,000 protein sequences were employed). The Gene Ontology terms are encoded into a language space "LanGO", as well as a neural machine translation model is built based on Recurrent Neural Networks that translates "ProLan" language to "GOLan" language.

Gated Recurrent Unit

Gated Recurrent Unit (GRU) (Cho *et al.*, 2014) is another variant of the RNN architecture. It simplifies the LSTM by having only two gates: the Update and Reset gates, (Figure 16).

The GRU was successfully used for pan-specific prediction of HLA-I-binding peptides (Human leukocyte antigens) (Heng *et al.*, 2020). The model performance was very good after 31 epochs. The prediction accuracies of the training and validation sets are, respectively, 87% and 85%.

Generative Adversarial Network

The particularity of the Generative Adversarial Network (GAN) (Goodfellow *et al.*) architecture is the competition aspect between two Neural Networks. One agent is the Generator (G), the other one is the Discriminator (D),

Elhassani et al. (2022) EMBnet.journal 27, e990
http://dx.doi.org/10.14806/ej.27.0.990

they contest with each other. The Generative tries to synthesize fake data that resemble to real ones, whereas the Discriminative tries to distinguish between the real and the fake ones, (Figure 17). G is trained in a such way to maximize the probability of D making a mistake. For instance, GAN was used for gene expression inference to approximate the joint distribution of landmark for the target genes and to learn their conditional distribution given the landmark gene (Ghasedi Dizaji *et al.*, 2018).

Transformers

One of the modern Deep Learning architectures are Transformers (Vaswani *et al.*, 2017). They are game changers as they have outperformed the previous architectures in many tasks. Initially, they were designed for textual data. Recently, it was shown that they can work with any kind of data. The transformer model relies heavily on the Attention mechanism. Basically, the Attention mechanism tries to learn and score the part of the data that is more important within a context. A transformer is comprised of two parts, Encoders and Decoders, (Figure 18). The stacked encoders are responsible for encoding the information while the stacked decoders decode it. The size of the stack is related to the architecture design. For instance, Transformers are used for improving the compound–protein interaction prediction by sequence-based Deep Learning with self-attention mechanism and label reversal experiments (Chen *et al.*, 2020). Transformers were also used to learn the protein language (facebookresearch/esm, 2021), in which the Unsupervised Learning was employed to train a deep contextual language model on 86 billion amino acids across 250 million protein sequences spanning evolutionary diversity. The resulting model contained information about biological properties. The model learned the representation space in a multi-scale organization reflecting structure from the level of biochemical properties of amino acids to remote homology of proteins.

BERT

Bidirectional Encoder Representations from Transformers (BERT) (Devlin *et al.*, 2019) is based on the Transformer architecture. It caused a stir in the Deep Learning community by achieving new state-of-the-art results on several tasks and in diversified disciplines. BERT is trained in two steps: pre-training and fine-tuning. During the pre-training phase, Unsupervised Learning is employed to train the model on unlabelled data. Consequently, a general purpose pre-trained model is obtained that can be fine-tuned for a specific problem using its labelled data. BERT uses two learning strategies: Masked Language Model (MLM) and Next Sentence Prediction (NSP). Regarding the MLM, the model randomly masks some tokens from the input sequence, then it tries to predict them based on the information provided by unmasked tokens in the sequence. For the

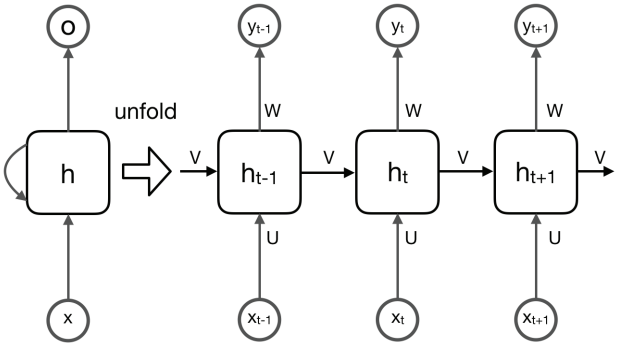


Figure 14. An unfold RNN cell. It has an input x_t, a hidden state h_t and an output y_t at a timeslot t. V, U and W are respectively the hidden state, input and weight matrices.

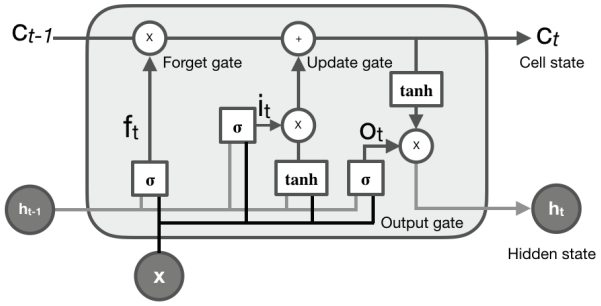


Figure 15. The LSTM cell internal structure. At time t, the cell reads the input x_t, updates the cell state c_t and the hidden state h_t using three gates that control the signal workflow.

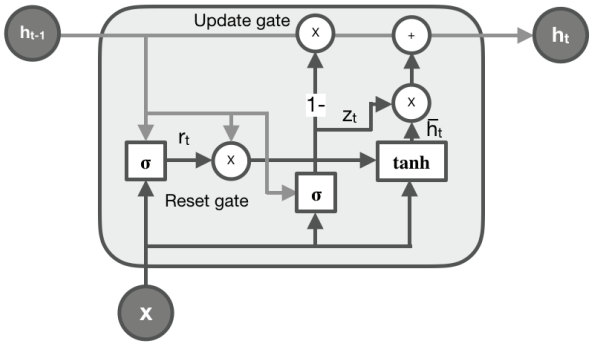


Figure 16. The GRU cell simplifies the architecture of the LSTM. It has only two gates: the Update and Reset gates.

NSP, the model needs to predict whether a given sentence is the subsequent sentence to the current one or not.

BERT technology is brand new. In genomics, for instance, it is used to decipher the language of non-coding DNA (Ji *et al.*, 2020). It was able to simultaneously

Elhassani et al. (2022) EMBnet.journal 27, e990
http://dx.doi.org/10.14806/ej.27.0.990

Figure 17. An illustrative example of how the GAN architecture is trained. The Generator tries to synthesize sequences, while the Discriminator tries to distinguish between the real and fake ones.

achieve the state-of-the-art performance on many sequence predictions tasks, such as: identifying the transcription factor binding sites also predicting the proximal and the core promoter regions.

Genomic data

Deep Learning requires huge amount of data. Luckily, genomic data were collected, organized and stored in open access databases for several decades. For instance, GenBank® (Benson *et al.*, 2013) is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequence. Gene database (Ostell, 2013) is another example that integrates information from a wide range of species where it contains over 17 million entries. Such databases can serve as data sources to extract genomic data to train Deep Learning models in order to solve specific problems. Conventionally, datasets can be split into three subsets:

- *training set* is a subset of the original data used to train the model,
- *test set* is a subset of the original data used to test the trained model,
- *validation set* is used to optimize the model during the development process.

Training

Training a Deep Learning model refers to the process of searching the best parameters that fit the model to the data set. While being trained, a Neural Network looks for the best values to tune its weights and obtain the best performances with the help of an optimization algorithm. The role of the optimizer is to minimize the loss function by reaching global minima. Neural Networks are trained iteratively. Each training iteration consists of a Forward propagation and Backward propagation passes in which a subset of the dataset named "mini-batch" is passed to the network. When the entire training set is consumed, it is called "Epoch". The performance of a Deep Learning model depends on a multitude of hyperparameters. Hyperparameters refer to the parameters whose values are used to control the learning process. Tuning those hyperparameters refers to the process of deciding about their values that determine the network structure such as the number of hidden layers ..., also those that determine how the network is trained, such as the learning rate, epochs ... Setting them is one of the difficulties of the Deep Learning approach due to their considerable numbers and their empirical attitudes (Makwe and Rathore, 2021).

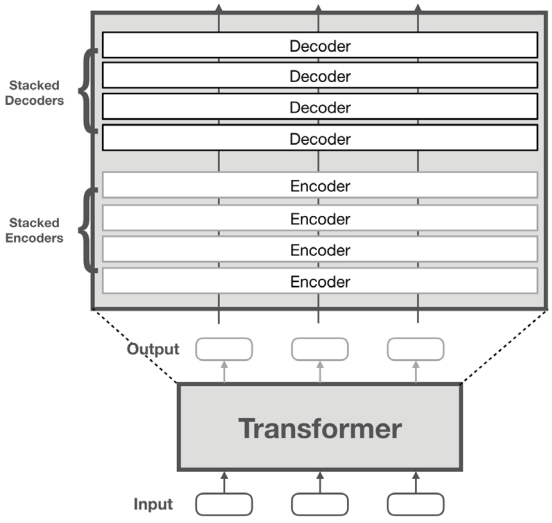


Figure 18. The transformer inner architecture comprised of stacked encoders in which the input is encoded, followed by stacked decoders that decode the information. The size of the stack is related to the architecture design.

Elhassani et al. (2022) EMBnet.journal 27, e990
http://dx.doi.org/10.14806/ej.27.0.990

Table 1. Details about some useful Deep Learning frameworks.

Framework	Core Language	License	Creator
TensorFlow	C++, Python	Apache 2.0	Google
PyTorch	C, Lua	BSD	Facebook
Keras	Python	MIT	François Chollet
Caffe	C++	BSD	Berkeley
Deeplearning4j	C++, Java	Apache 2.0	Deeplearning4j community
Theano	Python	BSD	University of Montréal
MXNet	C++	Apache 2.0	Apache Foundation
CNTK	C++	MIT	Microsoft
Janggu	Python	GPL-v3	(Kopp et al., 2019)
DragoNN	Python	MIT	Kundaje Lab
Kipoi	Python	MIT	(Avsec et al., 2019)
Flax	Python	Apache-2.0	Google

Table 2. Some useful tools and libraries.

Name	Description
Biopython	A biological computation library.
Scikit-bio	Library providing data structures, algorithms, and educational resources for bioinformatics.
PyEnsembl	Interface to Ensembl reference genome metadata.
Pandas	data analysis and manipulation tool.
NumPy	A library to operate large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions.
Matplotlib	A library that provides visualizations.
JAX	NumPy on the CPU, GPU, and TPU, with great automatic differentiation for high-performance machine learning research.
Scikit-learn	A library for Machine Learning.
bioconda	A platform and language independent package manager that sports easy distribution, installation and version management of software, it provides bioinformatics related packages

Forward propagation

Forward propagation pass is when data flow from the input towards the output. It comprises all the calculations of the Neural Network weights for the prediction in a forward direction.

Backward propagation

Historically, training the Neural Networks was one of the big challenges. Hence, the Backpropagation (Rumelhart and McClelland, 1987) algorithm was introduced to fulfil this duty. The backward propagation pass is when data flow from the output to the input with the purpose to tune the model. The Neural Network gradients are calculated in the backward direction. The Backprop abstracts the extensive computations used to calculate and update the weights of the network in a backward propagation.

Hyperparameters

This section will discuss some important hyperparameters.

Initialization

Initialization refers to the strategy selected to initialize the weights of the network when it starts the learning process. A good initialization strategy may reduce training time and computational costs.

Regularization

Regularization represents the techniques put in place to fight the Overfitting. They are concerned with adjusting the prediction function.

Dropout

Dropout is a method where the not needed neurons are dropped from the network. It is used to reduce the overfitting.

Appendix. A non exhaustive list of Deep Learning methodology applications targetting genomics.

Name	Domain	Architecture	References	Year
DECRES	Genomics	MLP	http://dx.doi.org/10.1186/s12859-018-2187-1	2018
DFS		MLP	http://dx.doi.org/10.1089/cmb.2015.0189	2016
PEDLA		MLP	http://dx.doi.org/10.1038/srep28517	2016
lincRNA predict		AE	http://dx.doi.org/10.1186/s12859-017-1922-3	2017
NeuSomatic	Variant calling	CNN	http://dx.doi.org/10.1038/s41467-019-09027-x	2019
seq2species		CNN	http://dx.doi.org/10.1101/353474	2019
Deep Variant		CNN	http://dx.doi.org/10.1038/nbt.4235	2018
Clairvoyante		CNN	http://dx.doi.org/10.1101/310458	2018
Clair		RNN	http://dx.doi.org/10.1101/865782	2019
CNNScoreVariants		CNN	http://dx.doi.org/10.1093/bioinformatics/btz901	2020
scvis	Transcriptomics	AE	http://dx.doi.org/10.1038/s41467-018-04368-5	2018
MRCNN		CNN	http://dx.doi.org/10.1186/s12864-019-5488-5	2019
DeepCpG		CNN & RNN	http://dx.doi.org/10.1186/s13059-017-1189-z	2017
DeepImpute		MLP	http://dx.doi.org/10.1186/s13059-019-1837-6	2019
scIGain		GAN	http://dx.doi.org/10.1093/nar/gkaa506	2020
scDeepCluster		AE	http://dx.doi.org/10.1038/s42256-019-0037-0	2019
DeepSEA	Epigenetics	CNN	http://dx.doi.org/10.1038/nmeth.3547	2015
DeepBind		CNN	http://dx.doi.org/10.1038/nbt.3300	2015
DanQ		CNN & LSTM	http://dx.doi.org/10.1093/nar/gkw226	2016
DeepLift		CNN	http://dx.doi.org/10.1101/737981	2020
DeepHistone		CNN	http://dx.doi.org/10.1186/s12864-019-5489-4	2019
AutoImpute	Metagenomics	AE	http://dx.doi.org/10.1038/s41598-018-34688-x	2018
DeepMicrobes		LSTM	http://dx.doi.org/10.1093/nargab/lqaa009	2020
Meta2		AE	https://arxiv.org/abs/1909.13146	2020
scScope		AE	http://dx.doi.org/10.1101/315556	2018
GeNet		CNN	https://arxiv.org/abs/1901.11015	2019
AlphaFold	Proteomics	Residual CNN	http://dx.doi.org/10.1038/s41586-019-1923-7	2020
DeepCDpred		Multi-stage FFNN	http://dx.doi.org/10.1371/journal.pone.0205214	2019
trRosetta		Residual CNN	http://dx.doi.org/10.1073/pnas.1914677117	2020
DeepInterface		CNN	http://dx.doi.org/10.1101/617506	2019
MaSIF		GNN	http://dx.doi.org/10.1038/s41592-019-0666-6	2020
DRREP		DNN	http://dx.doi.org/10.1186/s12864-017-4024-8	2017
ESM		Transformer	http://dx.doi.org/10.1101/622803	2019

Normalization

Normalization is concerned with feature scaling techniques for data adjustment.

Common problems

Overfitting

Overfitting is the situation when the model learns too much on the used dataset, thus it gives good accuracy on the training data but does not generalize well on new data. It often appears when working with finite samples or limited datasets.

Underfitting

Underfitting is the scenario when the model has not learn enough from the data, thus it was not able to generalize.

Vanishing gradient

The Vanishing Gradient problem (Hochreiter, 1998) may happen when the network is comprised from several neural layers, in particular the Recurrent Neural Networks. In essence, Vanishing Gradient occurs when gradients are very small or zero. Thus, little to no training can take place and a poor predictive performance is noticed.

Exploding gradient

Exploding gradient is a problem where large error gradients accumulate resulting very large updates to Neural Network model weights during training. Consequently, the model becomes unstable and unable to learn from the training data.

Frameworks

Recently, Deep Learning approach has witnessed a wide adoption. One of the many reasons is the plethora of available libraries and frameworks that emerged to support this trend. They save time and offer the necessary toolkits for rapid prototyping of new concepts. The Tables 1 and 2 summarize some of the widely used ones.

Limitations

Deep Learning faces many challenges in genomics, from which interesting to note:

The curse of dimensionality

Genomics is considered as a Big Data science. Taking in account the volume of the available datasets (Gigabytes), their heterogeneity (sequencing of coding or non-coding genes, gene variants...) and their variety, which can pose challenges for this approach.

Lack of data

Deep Learning requires a huge amount of data. Sometimes and for a specific problem, the available data are not enough to obtain good performance.

Imbalanced classes

Usually, the collected genomics data suffer from imbalanced ratio of instances per class. Thus, the DL model may fail to generalize about certain classes.

Model interpretation

Generally, it is considered to be a major problem of the Deep Learning approach. Sometimes, it is difficult for the model designer to understand and interpret the learned patterns. This problem is known as the black box.

Conclusion

With the advancement in processing power, availability of toolbox for practitioners and abundance of genomics data, Deep Learning is delivering impressive results in various fields including genomics. In this work, we introduced the different concepts of this technology and supplied various use case examples, also pointed out some of its advantages, difficulties and challenges. Deep Learning can be a real opportunity for researchers to tackle various genomics problems in a data driven approach.

Acknowledgements

IMGT® was funded in part by the BIOMED1 (BIOCT930038), Biotechnology BIOTECH2 (BIO4CT960037), 5th PCRDT Quality of Life and Management of Living Resources (QLG2-2000-01287), and 6th PCRDT Information Science and Technology (ImmunoGrid, FP6 IST-028069) programmes of the European Union (EU). IMGT® received financial support from the GIS IBiSA, the Agence Nationale de la Recherche (ANR) Labex MabImprove (ANR-10-LABX-53-01), the Région Occitanie Languedoc-Roussillon (Grand Plateau Technique pour la Recherche (GPTR), BioCampus Montpellier. IMGT® is currently supported by the Centre National de la Recherche Scientifique (CNRS), the Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI), the University of Montpellier, and the French Infrastructure Institut Français de Bioinformatique (IFB) ANR-11-INBS-0013. IMGT® is a registered trademark of CNRS. IMGT® is member of the International Medical Informatics Association (IMIA) and a member of the Global Alliance for Genomics and Health (GA4GH). IMGT is granted access to the High Performance Computing (HPC) resources of Meso@LR and of Centre Informatique National de l'Enseignement Supérieur (CINES) and to Très Grand Centre de Calcul (TGCC) of the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) under the allocation 036029 (2010-2021) made by GENCI (Grand Equipement National de Calcul Intensif).

We are grateful to Fotis Psomopoulos for helpful scientific discussions.

Key Points

- Deep Learning is a subset of Machine Learning methods that makes use of interconnected Artificial Neurons "Neural Networks" to automatically learn features from raw data.
- Learning can be supervised, semi-supervised, unsupervised or reinforcement.
- A wide variety of Deep Learning architectures exist such as CNN, RNN, GAN, Transformers…
- Training a Deep Learning model may require a huge amount of data and fine-tuning the hyperparameters certainly impacts the learning process and the performance.
- Deep Learning is a promising technology that demonstrated its supremacy in various tasks and has various application in different domains including the genomics.

References

Avsec Ž, Kreuzhuber R, Israeli J, Xu N, Cheng J, *et al.* (2019) The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. Nat. Biotechnol. **37** (6), 592–600. http://dx.doi.org/10.1038/s41587-019-0140-0

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, *et al.* (2013) GenBank. Nucleic Acids Res. **41** (Database issue), D36-42. http://dx.doi.org/10.1093/nar/gks1195

Cao R, Freitas C, Chan L, Sun M, Jiang H, *et al.* (2017) ProLanGO: Protein Function Prediction Using Neural Machine Translation Based on a Recurrent Neural Network. Molecules **22** (10), 1732. http://dx.doi.org/10.3390/molecules22101732

Chen L, Cai C, Chen V, and Lu X (2016) Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. BMC Bioinformatics **17** (1), S9. http://dx.doi.org/10.1186/s12859-015-0852-1

Chen L, Tan X, Wang D, Zhong F, Liu X, *et al.* (2020) TransformerCPI: improving compound–protein interaction prediction by sequence-

based deep learning with self-attention mechanism and label reversal experiments. Bioinformatics **36** (16), 4406–4414. http://dx.doi.org/10.1093/bioinformatics/btaa524

Cho K, van Merrienboer B, Bahdanau D, and Bengio Y (2014) On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. ArXiv14091259 Cs Stat. http://dx.doi.org/10.48550/arXiv.1409.1259

Choi D, Shallue CJ, Nado Z, Lee J, Maddison CJ, *et al.* (2020) On Empirical Comparisons of Optimizers for Deep Learning. ArXiv191005446 Cs Stat. http://dx.doi.org/10.48550/arXiv.1910.05446

Devlin J, Chang M-W, Lee K, and Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv181004805 Cs. http://dx.doi.org/10.48550/arXiv.1810.04805

F Z, Q L, X Z, and B S (2015) Protein interaction network constructing based on text mining and reinforcement learning with application to prostate cancer. IET Syst Biol **9** (4), 106–112. http://dx.doi.org/10.1049/iet-syb.2014.0050

facebookresearch/esm (2021) Facebook Research,.

Friedländer MR, Mackowiak SD, Li N, Chen W, and Rajewsky N (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. **40** (1), 37–52. http://dx.doi.org/10.1093/nar/gkr688

Gao W, Mahajan SP, Sulam J, and Gray JJ (2020) Deep Learning in Protein Structural Modeling and Design. Patterns N 1 (9), 100142. http://dx.doi.org/10.1016/j.patter.2020.100142

Ghahramani Z (2004) Unsupervised Learning. In: BousquetO, von LuxburgU, and RätschG (eds) Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, Berlin, Heidelberg,pp. 72–112

Ghasedi Dizaji K, Wang X, and Huang H (2018) Semi-Supervised Generative Adversarial Network for Gene Expression Inference. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '18. Association for Computing Machinery, New York, NY, USA, New York, NY, USA,pp. 1435–1444

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, *et al.* Generative Adversarial Nets. , 9.

Hasan MA and Lonardi S (2020) DeeplyEssential: a deep neural network for predicting essential genes in microbes. BMC Bioinformatics **21** (14), 367. http://dx.doi.org/10.1186/s12859-020-03688-y

Heng Y, Kuang Z, Huang S, Chen L, Shi T, *et al.* (2020) A Pan-Specific GRU-Based Recurrent Neural Network for Predicting HLA-I-Binding Peptides. ACS Omega **5** (29), 18321–18330. http://dx.doi.org/10.1021/acsomega.0c02039

Hennig C and Kutlukaya M Some thoughts about the design of loss functions. REVSTAT–Statistical J., 2007.

Hochreiter S (1998) The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. Int J Unc Fuzz Knowl Based Syst **06** (02), 107–116. http://dx.doi.org/10.1142/S0218488598000094

Hochreiter S and Schmidhuber J (1997) Long Short-Term Memory. Neural Comput **9** (8), 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735

Hubel DH and Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol **160**, 106–154. http://dx.doi.org/10.1113/jphysiol.1962.sp006837

Irmak E (2020) Implementation of convolutional neural network approach for COVID-19 disease detection. Physiol. Genomics **52** (12), 590–601. http://dx.doi.org/10.1152/physiolgenomics.00084.2020

Ji Y, Zhou Z, Liu H, and Davuluri RV (2020) DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. bioRxiv, 2020.09.17.301879. http://dx.doi.org/10.1101/2020.09.17.301879

Kelley DR, Snoek J, and Rinn JL Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. , 35.

Kiselev VY, Andrews TS, and Hemberg M (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet **20** (5), 273–282. http://dx.doi.org/10.1038/s41576-018-0088-9

Kopp W, Monti R, Tamburrini A, Ohler U, and Akalin A (2019) Janggu - Deep learning for genomics. bioRxiv http://dx.doi.org/10.1101/700450

Kotsiantis SB, Zaharakis ID, and Pintelas PE (2006) Machine learning: a review of classification and combining techniques. Artif Intell Rev **26** (3), 159–190. http://dx.doi.org/10.1007/s10462-007-9052-3

Krizhevsky A, Sutskever I, and Hinton GE (2012) ImageNet Classification with Deep Convolutional Neural Networks. In: PereiraF, BurgesCJC, BottouL, and WeinbergerKQ (eds) Advances in Neural Information Processing Systems. Curran Associates, Inc., Vol**25**,pp. 1097–1105 http://dx.doi.org/10.1145/3065386

Le Cun Y, Boser B, Denker JS, Henderson D, Howard RE, *et al.* (1989) Handwritten digit recognition with a back-propagation network. In: Proceedings of the 2nd International Conference on Neural Information Processing Systems. NIPS'89. MIT Press, Cambridge, MA, USA, Cambridge, MA, USA,pp. 396–404

LeCun Y, Bengio Y, and Hinton G (2015) Deep learning. Nature **521** (7553), 436–444. http://dx.doi.org/10.1038/nature14539

Li Y, Shi W, and Wasserman WW (2018) Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. BMC Bioinformatics **19** (1), 202. http://dx.doi.org/10.1186/s12859-018-2187-1

Liu F, Miao Y, Liu Y, and Hou T (2020) RNN-VirSeeker: a deep learning method for identification of short viral sequences from metagenomes. IEEEACM Trans Comput Biol Bioinform PP http://dx.doi.org/10.1109/TCBB.2020.3044575

Makwe A and Rathore AS (2021) An Empirical Study of Neural Network Hyperparameters. In: BhatejaV, PengS-L, SatapathySC, and ZhangY-D (eds) Evolution in Computational Intelligence. Advances in Intelligent Systems and Computing. Springer, Singapore, Singapore,pp. 371–383. http://dx.doi.org/10.1007/978-981-15-5788-0_36

McCulloch WS and Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. **5** (4), 115–133. http://dx.doi.org/10.1007/BF02478259

Movva R, Greenside P, Marinov GK, Nair S, Shrikumar A, *et al.* (2019) Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. bioRxiv, 393926. http://dx.doi.org/10.1101/393926

Nwankpa C, Ijomah W, Gachagan A, and Marshall S (2018) Activation Functions: Comparison of trends in Practice and Research for Deep Learning. ArXiv181103378 Cs. http://dx.doi.org/10.48550/arXiv.1811.03378

Ostell J (2013) What's in a Genome at NCBI? National Center for Biotechnology Information (US),.

Peace RJ, Hassani MS, and Green JR (2018) miPIE: NGS-based Prediction of miRNA Using Integrated Evidence. bioRxiv, 405357. http://dx.doi.org/10.1101/405357

Popova M, Isayev O, and Tropsha A (2018) Deep reinforcement learning for de novo drug design. Sci. Adv. **4** (7), eaap7885. http://dx.doi.org/10.1126/sciadv.aap7885

Press TM Reinforcement Learning \textbar The MIT Press.

Quang D and Xie X (2019) FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. Methods 166, 40–47. http://dx.doi.org/10.1016/j.ymeth.2019.03.020

Rosenblatt F (1957) The Perceptron, a Perceiving and Recognizing Automaton Project Para Cornell Aeronautical Laboratory,.

Rumelhart DE and McClelland JL (1987) Learning Internal Representations by Error Propagation. In: Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations. MIT Press, pp. 318–362

Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, *et al.* (2020) Improved protein structure prediction using potentials from deep learning. Nature **577** (7792), 706–710. http://dx.doi.org/10.1038/s41586-019-1923-7

Sheikh Hassani M and Green JR (2019) A semi-supervised machine learning framework for microRNA classification. Hum Genomics **13** (Suppl 1), 43. http://dx.doi.org/10.1186/s40246-019-0221-7

Singh R, Lanchantin J, Sekhon A, and Qi Y (2017) Attend and Predict: Understanding Gene Regulation by Selective Attention on Chromatin. ArXiv170800339 Cs. http://dx.doi.org/10.48550/arXiv.1708.00339

Singh S, Yang Y, Póczos B, and Ma J (2019) Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. Quant Biol **7** (2), 122–137. http://dx.doi.org/10.1007/s40484-019-0154-0

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, *et al.* (2017) Attention Is All You Need. ArXiv170603762 Cs. http://dx.doi.org/10.48550/arXiv.1706.03762

Wang Q, Ma Y, Zhao K, and Tian Y (2020) A Comprehensive Survey of Loss Functions in Machine Learning. Ann Data Sci http://dx.doi.org/10.1007/s40745-020-00253-5

Xavier R, de Souza KP, Chateau A, and Alves R (2020) Genome Assembly Using Reinforcement Learning. In: KowadaL and de OliveiraD (eds) Advances in Bioinformatics and Computational Biology. Lecture Notes in Computer Science. Springer International Publishing, Cham, Cham,pp. 16–28. http://dx.doi.org/10.1007/978-3-030-46417-2_2

Yelmen B, Decelle A, Ongaro L, Marnetto D, Tallec C, *et al.* (2019) Creating Artificial Human Genomes Using Generative Models. bioRxiv, 769091. http://dx.doi.org/10.1101/769091

EMBnet.journal
Bioinformatics in Action

Papakonstantinou et al. (2022) EMBnet.journal 27, e1004
http://dx.doi.org/10.14806/ej.27.0.1004

Research Papers

Molecular fusion events in carcinogenic organisms: a bioinformatics study for the detection of fused proteins between viruses, bacteria and eukaryotes

Eleni Papakonstantinou[1], Kalliopi Io Diakou[1], Thanasis Mitsis[1], Konstantina Dragoumani[1], Flora Bacopoulou[2], Vasilis Megalooikonomou[3], Sophia Kossida[4], George P. Chrousos[2,5], Dimitrios Vlachakis[1,2,5]✉

[1]Laboratory of Genetics, Department of Biotechnology, School of Applied Biology and Biotechnology, Agricultural University of Athens, Athens, Greece

[2]University Research Institute of Maternal and Child Health & Precision Medicine, and UNESCO Chair on Adolescent Health Care, National and Kapodistrian University of Athens, "Aghia Sophia" Children's Hospital, Athens, Greece

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

[4]IMGT, The International ImMunoGeneTics Information System, Université de Montpellier, Laboratoire d'ImmunoGénétique Moléculaire and Institut de Génétique Humaine, University of Montpellier, Montpellier, France

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

Competing interests: EP none; KID none; TM none; KD none; FB none; VM none; SK none; GPC none; DV none

Abstract

Molecular fusion events have a prominent role in the initial steps of carcinogenesis. In this study, a bioinformatics analysis was performed between four organisms that are known to induce cancer development in humans: two viruses, *Human Herpesvirus 4*, and *Human T-cell leukaemia virus*, one bacterium, *Helicobacter Pylori*, and one trematode, *Schistosoma mansoni*. The annotated proteomes from these organisms were analysed using the SAFE software to identify protein fusion events, which may provide insight into protein function similarities and possible merging events during the course of evolution. Based on the results, five fused proteins with very similar functions were detected, whereas proteins with different functions that might act in the same molecular complex or biochemical pathway were not found. Thus, this study analysed the above four well-known cancer-related organisms with *de novo* bioinformatics programs and provided useful information on protein fusion events, hopefully leading to deeper understanding of carcinogenenesis.

Introduction

Cancer is one of the leading causes of death in the world with 9.6 million cancer related deaths in 2018 and a projection of more than 16 million by the year 2040 (World Health Organisation). Carcinogenic effects appear with the transformation of a normal cell into a tumour cell and its unrestrained proliferation, with potential to invade beyond normal tissue boundaries and metastasize to distant organs (Tomasetti *et al.*, 2017). Genetic factors are involved in cancer, but external agents can also play a key role in the disruptive cellular changes. These external factors are separated into three main categories, the physical carcinogens, the chemical carcinogens and the biological carcinogens (Vineis *et al.*, 2010). This study focuses on the biological type of carcinogens, with the goal to provide new perspectives on cancer-generating events. The term biological carcinogens can refer to infections from various organisms, such as certain viruses, bacteria or parasites. Herein, possible molecular fusion events and probably linked proteins are studied between organisms including bacteria (*Helicobacter pylori I*), eukaryotes (*Schistosoma flatworms*) and viruses (*Human Herpesvirus 4* and *Human T-cell leukaemia virus*).

Helicobacter Pylori is a non-invasive, gram-negative bacterium which colonizes the stomach and is present in almost half of the human population. It has been classified as a member of the group I carcinogenic agents by the International Agency for Research on Cancer (Thorell *et al.*, 2017). The transmission of these bacteria is interhuman and infection often occurs during childhood. In most cases, the bacteria can live in the human stomach for years without causing any problem

Article history
Received: 25 April 2021
Accepted: 21 May 2021
Published: 04 April 2022

(Burucoa and Axon, 2017; Lopes *et al.*, 2014). *H. pylori* is considered to be the most important risk factor for gastric cancer (Polk and Peek, 2010). Infection by this bacterium induces chronic gastritis that can progress to intestinal dysplasia and, through complex mechanisms, can ultimately lead to gastric cancer (Parkin *et al.*, 2005). Studies *in vivo* showed that *H. pylori* infection induced double-stranded breaks, a type of DNA damage (Xie *et al.*, 2014). Endogenous DNA damage is inherently linked to genomic instability, which in turn is one of the most prevalent onsets of tumorigenesis. Meta-analysis of randomized control trials showed that eradication of the bacterium may reduce the risk of gastric cancer (Liou *et al.*, 2020), while the safety and efficacy of the eradication was reviewed in a recent study (Liou *et al.*, 2019).

Schistosoma (blood flukes) is a genus of parasitic flatworms that are responsible for schistosomiasis (Utzinger *et al.*, 2009). This disease affects millions of people, especially in developing countries (Africa and South America) and is considered the second most socioeconomically devastating parasitic disease after malaria (World Health Organization). Out of the five species that infect humans, *Schistosoma mansoni* is the most common and generally the one used in laboratory studies. Humans are their primary host, but the larvae need to pass through an intermediate freshwater snail host to infect another mammalian host (Aguiar *et al.*, 2017; Protasio *et al.*, 2012). Clinical manifestations of the disease pass through various acute, sub-acute and chronic stages. Schistosomiasis has been associated with development of malignancy in the rectum, bladder and lymphoid tissue (Palumbo, 2007; Barsoum *et al.*, 2013). Recently, glutathione transferase from *Schistosoma japonicum* was studied as a potential drug design target towards this parasite (Platis *et al.*, 2020).

Human Herpesvirus 4 (HHV-4), also called *Epstein–Barr virus* (EBV), is one of the most common viruses in humans, affecting nearly 90% of the adult population in the world. The virus is transmitted by saliva and, in the majority of cases, it causes infectious mononucleosis (Smatti, Al-Sadeq *et al.*, 2018). Additionally, the virus is aetiologically linked to two pre-malignant lymphoproliferative diseases (LPDs) and up to nine distinct human tumours. The LPDs include B-cell origin diseases, such as Burkitt lymphoma, Hodgkin lymphoma and immune impairment associated tumour pathologies, such as Plasmablastic lymphoma (Kanda *et al.*, 2019). Furthermore, the virus is involved in various tumour pathologies like nasopharyngeal cancer, gastric carcinoma and leiomyosarcoma (Abe, Kaneda *et al.*, 2015) (Hall *et al.*, 2015).

Human T-cell leukaemia virus is the first pathogenic human retrovirus to have been discovered (Coffin, 2015). *Human T-cell leukaemia virus type I* (HTLV-I) is the most common one, having infected an estimated 5-20 million individuals. Among other disorders, HTLV-1 causes a form of T-cell lymphoproliferation, characterized as leukemia or lymphoma, and termed adult T-cell leukemia (ATL) or adult T-cell leukemia-lymphoma (ATLL) (Cook *et al.*, 2017; Meissner *et al.*, 2017). Therapies that block HTLV-1 replication include integrase inhibitors and nucleoside reverse transcriptase inhibitors, while a combination of interferon alpha and zidovudine seems to have significant effects on the chronic/acute forms of ATLL and not on the lymphoma sub-type of ATLL (Nasr, El-Hajj *et al.*, 2011). *Human T-cell leukaemia virus type II* (HTLV-II) has been related to T-cell variant of hairy cell leukaemia, and like HTLV-I, it is capable of transforming normal human peripheral blood into lymphocytes in vitro, however its mechanism remains elusive (Murphy 2016; Shima *et al.*, 1986).

Proteins control almost all biological systems in a cell, and while some proteins work independently, the vast majority perform their actions in collaboration with other proteins. Protein-protein interactions are established in all cellular processes, such as in the cell cycle and intracellular signalling, and are also prognostic factors of diseases such as cancer (Chen, Sam *et al.*, 2010). As the same proteins may be involved in different cellular processes, the study of their interactions allows finding correlations between different diseases or cellular conditions, and promotes research in the field of drug design (Tsaniras SC *et al.*, 2015). The analysis of protein-protein interaction is, therefore, critical for a more profound understanding of biological processes (Papageorgiou *et al.*, 2014; Steinhauf *et al.*, 2014). Although there are experimental methods towards determining protein-protein interactions, the techniques are labour-intensive, time consuming and not necessarily easily performed in high-throughput analyses (Berggard *et al.*, 2007). A useful alternative in the face of those hindrances is a bioinformatic approach, which can take multiple, equally interesting forms, as evidenced by various studies (Li, Wang *et al.*, 2018;Chen, Wang *et al.*, 2019)

Functional links between proteins can be examined through various ways. The occurrence of a fusion event between two proteins can suggest the likelihood of a functional connection between them, for example by either being part of the same protein complex, or by acting in the same pathway or biological network (Tsagrasoulis *et al.*, 2012). A bioinformatic analysis of protein fusion events between two species can, therefore, enable the detection of functional links between proteins (Enright *et al.*, 1999).

Molecular fusion events have a prominent role in the initial steps of carcinogenesis, where chromosome translocations and gene fusions result in the deregulation of physiological molecular mechanisms (Yu *et al.*, 2019). Gene fusions have been detected in all types of human neoplasias, with a varying proportion among different types (Mitelman *et al.*, 2007). The analysis of fusion events employs genomic structure and sequence analysis of two or more genomes to detect possibly connected protein pairs without any prior knowledge, which would not have been necessarily detected by experimental analysis (Dimitriadis *et al.*, 2011). Nowadays, the development of techniques, such as next generation sequencing (NGS)

Papakonstantinou et al. (2022) EMBnet.journal 27, e1004
http://dx.doi.org/10.14806/ej.27.0.1004

Table 1. List of selected organisms.

Organism common name	Proteome Name	UniProt Proteome ID
Human herpesvirus 4	*Epstein-Barr virus* (strain B95-8)	UP000007640
Helicobacter Pylori	*Helicobacter pylori* (strain ATCC 700392 / 26695) (*Campylobacter pylori*)	UP000000429
Human T-cell leukaemia virus	*Human T-cell leukaemia virus 2*	UP000009254
Schistosoma mansoni (Blood fluke)	*Schistosoma mansoni Puerto Rican*	UP000008854

or transcriptome analysis, and of various bioinformatics algorithms, has led to the creation of several databases containing information on gene fusion events and their functions (Panigrahi *et al.*, 2018). A recent example of such an accomplishment is the establishment of the ChimberDB, an extensive database of fusion genes (Jang, Jang *et al.*, 2020). The SAFE software (Tsagrasoulis *et al.*, 2012) was used in the past towards the successful detection of protein interactions in several studies (Dimitriadis *et al.*, 2011; Trimpalis *et al.*, 2013; Vlachakis *et al.*, 2013). SAFE is an application used to identify, filter and visualize fusion events. It enables the analysis and representation of fusion proteins by performing pairwise alignments of protein sets, permitting an independent research on fusion proteins and their subsequent imaging in this specific application, thus simplifying the analysis and providing optimum results (Tsagrasoulis *et al.*, 2012).

Methods

Database sequence search

The proteome is the entire set of proteins expressed by a specific organism at a certain time (Jensen, 2006). In contrast to the genome, the proteome is not static and continually changes in response to external and internal events.

In order to run the comparison with SAFE software and detect fusion events, the proteome sequences of the four organisms were used. The FASTA files of the proteomes were downloaded by UniProtKB[1], a freely accessible protein database containing protein sequences and biological information. More specifically, its "Proteomes" subsection allows access to the whole proteome of numerous organisms. It includes both manually reviewed (UniProtKB/Swiss-Prot) and unreviewed (UniProtKB/TrEMBL) entries. The carcinogenic organisms included in their study and their respective UniProt Proteome ID are listed in Table 1.

Fusion events analyses

Fusion events analyses were performed with SAFE software (Software for the Analysis of Fusion Events 3). SAFE is designed to search for fusion events between a set of organisms given as input in a FASTA file format, one file for each organism to analyse.

Its algorithm is implemented as follows: The first task conducted is the removal of potentially duplicated proteins from each sequence file. This is done according to a user-defined parameter, which represents the percentage of minimal identities to consider two proteins as duplicated ones. Therefore, for each FASTA proteome file, each protein is blasted against its respective organism proteome, considering the user parameter "Max Blast identities" to identify duplicates. Within this step, each shorter duplicated protein over the max blast identity threshold is removed from the results. The proteins that are found above threshold are then saved in a new FASTA file called [initial_file_name] _reduced.txt.

The second task to perform is the actual analysis of fusion events from this reduced proteomes, using the following parameter for the fusion detection processes: Min. Domain Length: 70; Min. Blast Identities: 27; Min. Fused Protein coverage: 70; Max. overlaps Region in Domains: 0; Multiple Proteins cut-off: 5; E-value: 9 .10−3. The main SAFE algorithm is described in Figure 1.

Results

The proteomes of the two viruses, *Human Herpesvirus 4* and *Human T-cell leukaemia virus 2*, that can cause cancer in humans, as well as the proteomes of *Helicobater pylori*, a gram-negative microaerophilic bacterium of the human stomach that is correlated to gastric cancer, and *Schistosoma mansoni*, a human parasite that is responsible for intestinal schistosomiasis, were all analysed for potential protein fusion events. The following analysis pairs were carried out; the proteome of *Helicobacter pylori* against the proteome of *Schistosoma mansoni*, the proteome of *Helicobacter pylori* against the proteome of *Human T-cell leukaemia virus 2*, the proteome of Human T-cell leukaemia virus 2 against *Schistosoma mansoni* and the proteome of *Human T-cell leukaemia virus 2* against *Human Herpesvirus 4*, as well as the backward BLAST analysis for each of these sets.

No putative protein fusion events in any comparison within the two viruses were identified. 20 fusion events were detected in the comparison of *Helicobacter pylori* and *Schistosoma mansoni* proteomes. Out of those 20 fusion events, only 5 of them are proteins that satisfy the unique protein threshold. 19 were detected in the proteome of *Helicobater pylori* (4 of them above the unique protein cut-off) and 1 was detected in the proteome of *Schistosoma mansoni*, which also satisfies the unique protein cut-off (Table 2).

[1]http://www.uniprot.org

Papakonstantinou et al. (2022) EMBnet.journal 27, e1004
http://dx.doi.org/10.14806/ej.27.0.1004

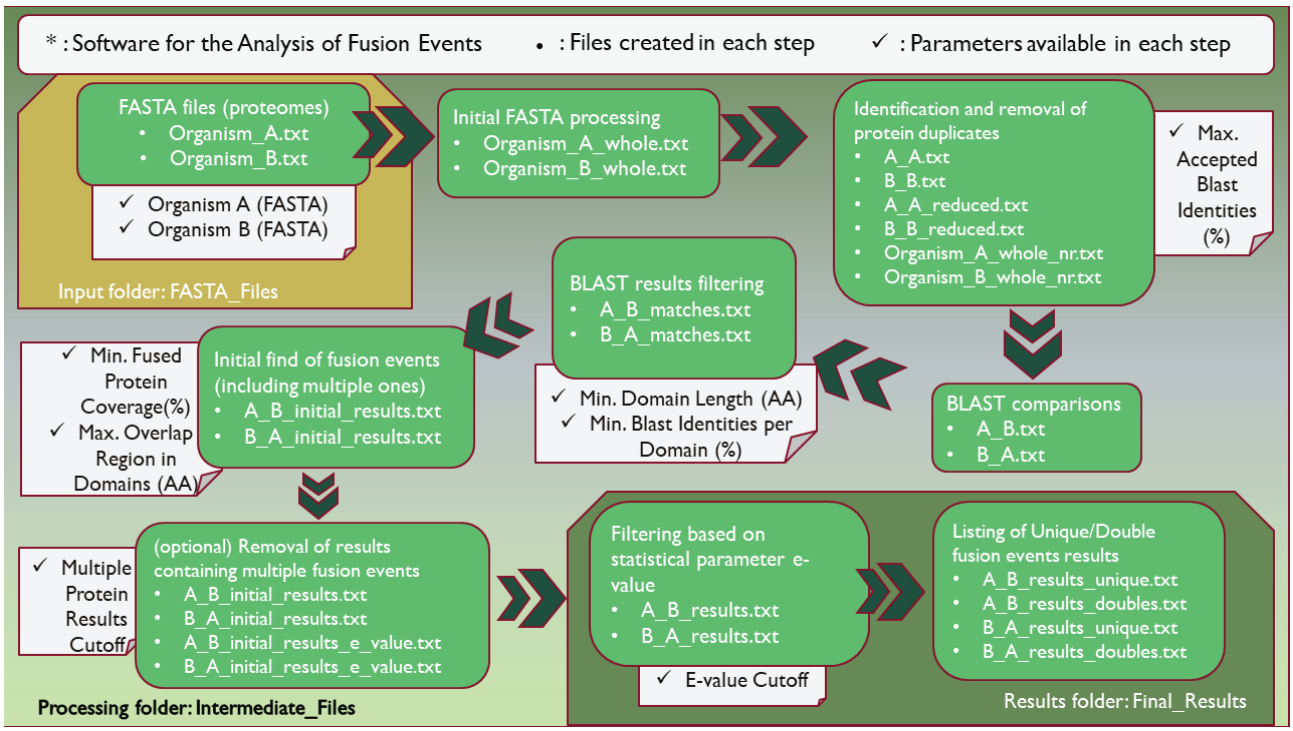


Figure 1. SAFE software, the main algorithm steps.

The fusion event that took place in *Schistosoma mansoni* represents the fusion of two ATPase domains, one from a ATP-dependent zink metalloprotease FtsH and one from a cell division protein FtsH from the AAA family (ATPases Associated with diverse cellular Activities) from *Helicobater pylori* into a cell division control protein 48 from the AAA family in *Schistosoma mansoni*. In this case, only parts of the proteins from *Schistosoma mansoni* were used to build domains in the fused protein in *Helicobater pylori* (Figure 2A). The fused protein has the same predicted enzymatic function as the two original proteins.

In all four fusion events detected in *Helicobater pylori* it was visible that almost the full length of the two proteins from *Schistosoma mansoni* fused to form the majority of the protein in *Helicobater pylori* (Figure 2). In three of these cases, the proteins from *Schistosoma mansoni* had the same function, leading to a protein with similar function in *Helicobater pylori*. In one case, the two original proteins seem to have slightly different functions, which lead to a fused protein that incorporates both functions.

In the first fusion event, two putative Radical SAM proteins containing one $[4Fe\text{-}4S]^{+}$ cluster each, involved in RNA modifications, fused almost entirely to form tRNA-2-methylthio-N(6)-dimethylallyladenosine synthase (miaB) in *Helicobacter pylori*, containing two $[4Fe\text{-}4S]^{+}$ clusters (Figure 2). In the second fusion event, two Phosphoglycerate kinases from *Schistosoma mansoni* fused entirely to form a Phosphoglycerate kinase in *Helicobater pylori* (Figure 2). In the third fusion event, the C-terminal 3/4 part of a threonyl-tRNA synthetase, including the threonyl-tRNA synthetase-domain, was fused with the whole of an uncharacterized protein that carries an anticodon-binding domain (prediction carried out by InterPro[2]. This fusion built a threonyl-tRNA synthetase in *Helicobater pylori* (Figure

[2]http://www.ebi.ac.uk/interpro

Table 2. Detected fusion events between the analysed organisms. The number in brackets indicates the fusion events without considering the unique protein cut-off.

	Human Herpesvirus 4	*Human T-cell leukaemia virus 2*	*Helicobater pylori*	*Schistosoma mansoni*
Human Herpesvirus 4	-	**0**	**0**	**0**
Human T-cell leukaemia virus 2	**0**	-	**0**	**0**
Helicobater pylori	**0**	**0**	-	**1 (1)**
Schistosoma mansoni	**0**	**0**	**4 (19)**	-

Papakonstantinou et al. (2022) EMBnet.journal 27, e1004
http://dx.doi.org/10.14806/ej.27.0.1004

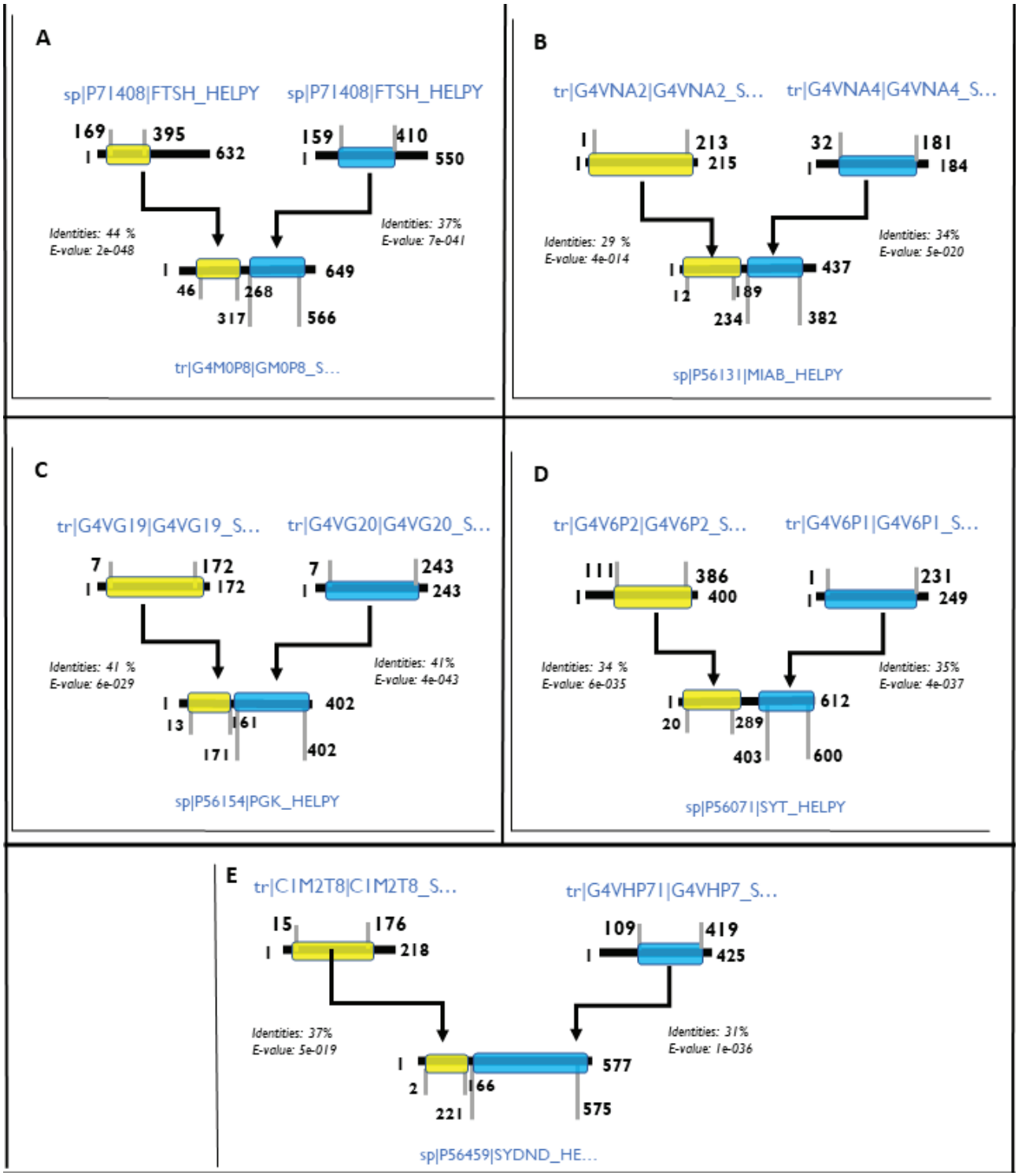


Figure 2. Graphical representation of the discovered fusion events, that passed the unique cutoff filter, recuperated from the SAFE software. **(A)** The fusion event detected leading to a cell division control protein 48 from the AAA family in *Schistosoma mansoni*, containing ATPase domains from the proteins in *Helicobater pylori*. (**B-E**) The fusion events detected in *Helicobater pylori*. **(B)** The tRNA-2-methylthio-N(6)-dimethylallyladenosine synthase (miaB) in *Helicobacter pylori* containing two $[4Fe\text{-}4S]^{+}$ cluster created through the fusion of two putative radical SAM proteins with each one $[4Fe\text{-}4S]^{+}$ cluster from *Schistosoma mansoni*. **(C)** The Phosphoglycerate kinase in *Helicobater pylori* created through the fusion of two Phosphoglycerate kinase from *Schistosoma mansoni*. **(D)** A threonyl-tRNA synthetase in *Helicobater pylori* created through the fusion of the C-terminal 3/4 part of a threonyl-tRNA synthetase, including the threonyl-tRNA synthetase-domain are fused with an uncharacterized protein that carries an anticodon-binding domain from *Schistosoma mansoni*. **(E)** An Aspartate-tRNA(Asp/Asn) ligase in *Helicobacter pylori* that it is able to aspartylate aspertat-tRNA and asparagine-tRNA created through the fusion of two threonyl-tRNA synthetase, one with a putative asparagine-tRNA ligase activity and one with a putative aspartate-tRNA ligase activity.

Papakonstantinou et al. (2022) EMBnet.journal 27, e1004
http://dx.doi.org/10.14806/ej.27.0.1004

2). In the last fusion event, two proteins with slightly different function fused to form an enzyme with both functions. An Aspartyl-tRNA synthetase with a putative asparagine-tRNA ligase activity was almost entirely fused with the C-terminal 3/4 of an Aspartyl-tRNA synthetase with a putative aspartate-tRNA ligase activity, to form the majority of an Aspartate-tRNA (Asp/Asn) ligase in *Helicobacter pylori*. The *H. pylori* ligase presents a relaxed tRNA specificity since it can aspartylate not only its cognate aspertat-tRNA, but also asparagine-tRNA (Figure 2).

Discussion

No fusion events were discovered when the two viruses, *Human Herpesvirus 4* and *Human T-cell leukaemia virus 2*, were compared. This may be an expected result when one considers that viruses do not share a lot with other evolution branches. Moreover, the chosen viruses have a very small genome (Vlachakis *et al.*, 2013), which inherently limits the chances of finding fusion events amongst them and against other organisms. On the other hand, the fusion events detected through the comparison of *Helicobater pylori* and *Schistosoma mansoni* were between proteins of the same or similar function. The performed analysis was not able to detect potential interactions of proteins with different function that may act in the same complex, or the same pathway. This does not mean that the fusion events that were detected could not point to a potential interaction of proteins with similar function, rather that the fusion events may have been stabilised during the course of evolution. For example, a Phosphoglycerate kinase with two phosphoglyceratekinase domains may be more efficient than a Phosphoglycerate kinase with only one.

The fusion events that were found by comparing the proteomes of *Schistosoma mansoni* and *Helicobater pylori* were almost all found in *Helicobater pylori* (19 out of 20) and only one was found in *Schistosoma mansoni*. *Schistosoma mansoni* is a worm and thereby higher up in the "tree of life" than *Helicobater pylori*, which is a prokaryote. These fusion events seem to be in opposite direction of evolution and could also represent protein fission events.

The detected fusion proteins were a cell division control protein 48 from the AAA family in *Schistosoma mansoni*, containing ATPase domains from the proteins in *Helicobater pylori*. A tRNAsynthase in Helicobacter pylori containing two $[4Fe\text{-}4S]^{+}$ cluster created through the fusion of two proteins with one $[4Fe\text{-}4S]^{+}$ cluster each from *Schistosoma mansoni*. A Phosphoglycerate kinase in *Helicobater pylori* created through the fusion of two Phosphoglycerate kinase from *Schistosoma mansoni*. A threonyl-tRNA synthetase in *Helicobater pylori* created through the fusion a threonyl-tRNA synthetase, including the threonyl-tRNA synthetase-domain fused with an uncharacterized protein that carries an anticodon-binding domain from *Schistosoma mansoni.* An Aspartate-tRNA (Asp/Asn) ligase in *Helicobacter pylori* that it is able to aspartylate aspertat-tRNA and asparagine-tRNA created through the fusion of two threonyl-tRNA synthetase, one with a putative asparagine-tRNA ligase activity and one with a putative aspartate-tRNA ligase activity.

The present study conducted amongst these four organisms revealed a limited number of fusion events, thus it is difficult to make any statement on whether the proteins that were detected could be in any relation with a possible cancer induction stemming from the presence of the organism in humans. Replicated in a larger scale and for an increased number of carcinogenic organisms, the study of putative fusion events could reveal potentially interacting proteins and connection to cancer by analysing GO terms and involved pathways.

Conclusions

Chromosome translocations, chromosomal interstitial deletion and inversion, and the resultant fusion events frequently underlie cancer development, through deregulation of molecular mechanisms and the generation of fused gene products, which often possess oncogenic properties. Four fusion events and one linked protein were identified in the present study from the comparison analysis of four prominent cancer-related organisms using the SAFE software. Identification of such fusion events may provide a useful basis for the discovery of novel potential therapeutic targets against cancer and other diseases.

Acknowledgements

DV would like to acknowledge funding from: i) AdjustEBOVGP-Dx (RIA2018EF-2081): Biochemical Adjustments of native EBOV Glycoprotein in Patient Sample to Unmask target Epitopes for Rapid Diagnostic Testing. A European and Developing Countries Clinical Trials Partnership (EDCTP2) under the Horizon 2020 'Research and Innovation Actions' DESCA, and ii) "MilkSafe: A novel pipeline to enrich formula milk using omics technologies", a research co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T2EDK-02222).

Key Points

- Molecular fusion events have a prominent role in the initial steps of carcinogenesis.
- Identification of fusion events provide a useful basis for the discovery of novel potential therapeutic targets against cancer and other diseases.
- A bioinformatics analysis was performed between four organisms that are known to induce cancer development in humans.
- Four fusion events and one linked protein were identified in the present study from the comparison analysis of four prominent cancer-related organisms.

References

Abe H, Kaneda A and Fukayama M (2015) Epstein-Barr Virus-Associated Gastric Carcinoma: Use of Host Cell Machineries and Somatic Gene Mutations. Pathobiology **82**(5), 212-223. http://dx.doi.org/10.1159/000434683

Aguiar PHN, Fernandes NMGS, Zani CL and Mourão MM (2017) A high-throughput colorimetric assay for detection of Schistosoma mansoni viability based on the tetrazolium salt XTT. Parasites & vectors **10**(1), 300-300. http://dx.doi.org/10.1186/s13071-017-2240-3

Barsoum RS, Esmat G and El-Baz T (2013) Human Schistosomiasis: Clinical Perspective: Review. Journal of Advanced Research **4**(5), 433-444. http://dx.doi.org/https://doi.org/10.1016/j.jare.2013.01.005

Berggård T, Linse S and James P (2007) Methods for the detection and analysis of protein-protein interactions. Proteomics **7**(16), 2833-2842. http://dx.doi.org/10.1002/pmic.200700131

Burucoa C and Axon A (2017) Epidemiology of Helicobacter pylori infection. Helicobacter **22** Suppl 1. http://dx.doi.org/10.1111/hel.12403

Chen J, Sam L, Huang Y, Lee Y, Li J *et al.* (2010) Protein interaction network underpins concordant prognosis among heterogeneous breast cancer signatures. Journal of Biomedical Informatics **43**(3), 385-396. http://dx.doi.org/https://doi.org/10.1016/j.jbi.2010.03.009

Chen K-H, Wang T-F and Hu Y-J (2019) Protein-protein interaction prediction using a hybrid feature representation and a stacked generalization scheme. BMC Bioinformatics **20**(1), 308. http://dx.doi.org/10.1186/s12859-019-2907-1

Coffin JM (2015) The discovery of HTLV-1, the first pathogenic human retrovirus. Proc Natl Acad Sci U S A **112**(51), 15525-15529. http://dx.doi.org/10.1073/pnas.1521629112

Cook L, Melamed A, Yaguchi H and Bangham CR (2017) The impact of HTLV-1 on the cellular genome. Curr Opin Virol **26**, 125-131. http://dx.doi.org/10.1016/j.coviro.2017.07.013

Dimitriadis D, Koumandou VL, Trimpalis P and Kossida S (2011) Protein functional links in Trypanosoma brucei, identified by gene fusion analysis. BMC Evolutionary Biology **11**(1), 193. http://dx.doi.org/10.1186/1471-2148-11-193

Enright AJ, Iliopoulos I, Kyrpides NC and Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature **402**(6757), 86-90. http://dx.doi.org/10.1038/47056

Hall LD, Eminger LA, Hesterman KS and Heymann WR (2015) Epstein-Barr virus: dermatologic associations and implications: part I. Mucocutaneous manifestations of Epstein-Barr virus and nonmalignant disorders. J Am Acad Dermatol **72**(1), 1-19; quiz 19-20. http://dx.doi.org/10.1016/j.jaad.2014.07.034

Jang YE, Jang I, Kim S, Cho S, Kim D *et al.* (2020) ChimerDB 4.0: an updated and expanded database of fusion genes. Nucleic Acids Res **48**(D1), D817-d824. http://dx.doi.org/10.1093/nar/gkz1013

Jensen ON (2006) Interpreting the protein language using proteomics. Nature Reviews Molecular Cell Biology **7**(6), 391-403. http://dx.doi.org/10.1038/nrm1939

Kanda T, Yajima M and Ikuta K (2019) Epstein-Barr virus strain variation and cancer. Cancer Sci **110**(4), 1132-1139. http://dx.doi.org/10.1111/cas.13954

Li LP, Wang YB, You ZH, Li Y and An JY (2018) PCLPred: A Bioinformatics Method for Predicting Protein-Protein Interactions by Combining Relevance Vector Machine Model with Low-Rank Matrix Approximation. Int J Mol Sci **19**(4). http://dx.doi.org/10.3390/ijms19041029

Liou JM, Lee YC, El-Omar EM and Wu MS (2019) Efficacy and Long-Term Safety of H. pylori Eradication for Gastric Cancer Prevention. Cancers (Basel) **11**(5). http://dx.doi.org/10.3390/cancers11050593

Liou JM, Malfertheiner P, Lee YC, Sheu BS, Sugano K *et al.* (2020) Screening and eradication of Helicobacter pylori for gastric cancer prevention: the Taipei global consensus. Gut **69**(12), 2093-2112. http://dx.doi.org/10.1136/gutjnl-2020-322368

Lopes D, Nunes C, Martins MC, Sarmento B and Reis S (2014) Eradication of Helicobacter pylori: Past, present and future. J Control Release **189**, 169-186. http://dx.doi.org/10.1016/j.jconrel.2014.06.020

Meissner ME, Mendonça LM, Zhang W and Mansky LM (2017) Polymorphic Nature of Human T-Cell Leukemia Virus Type 1 Particle Cores as Revealed through Characterization of a Chronically Infected Cell Line. J Virol **91**(16). http://dx.doi.org/10.1128/jvi.00369-17

Mitelman F, Johansson B and Mertens F (2007) The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer **7**(4), 233-245. http://dx.doi.org/10.1038/nrc2091

Murphy EL (2016) Infection with human T-lymphotropic virus types-1 and -2 (HTLV-1 and -2): Implications for blood transfusion safety. Transfus Clin Biol **23**(1), 13-19. http://dx.doi.org/10.1016/j.tracli.2015.12.001

Nasr R, El Hajj H, Kfoury Y, de Thé H, Hermine O *et al.* (2011) Controversies in targeted therapy of adult T cell leukemia/lymphoma: ON target or OFF target effects? Viruses **3**(6), 750-769. http://dx.doi.org/10.3390/v3060750

Palumbo EM (2007) Association Between Schistosomiasis and Cancer: A Review. Infectious Diseases in Clinical Practice **15**, 145-148.

Panigrahi P, Jere A and Anamika K (2018) FusionHub: A unified web platform for annotation and visualization of gene fusion events in human cancer. PLoS ONE **13**(5), e0196588. http://dx.doi.org/10.1371/journal.pone.0196588

Papageorgiou L, Loukatou S, Koumandou VL, Makałowski W, Megalooikonomou V *et al.* (2014) Structural models for the design of novel antiviral agents against Greek Goat Encephalitis. PeerJ **2**, e664. http://dx.doi.org/10.7717/peerj.664

Parkin DM, Bray F, Ferlay J and Pisani P (2005) Global cancer statistics, 2002. CA Cancer J Clin **55**(2), 74-108. http://dx.doi.org/10.3322/canjclin.55.2.74

Platis M, Vlachakis D, Foudah AI, Muharram MM, Alqarni MH *et al.* (2021) The Interaction of Schistosoma Japonicum Glutathione Transferase with Cibacron Blue 3GA and its Fragments. Med Chem **17**(4), 332-343. http://dx.doi.org/10.2174/1573406416666200403074742

Polk DB and Peek RM (2010) Helicobacter pylori: gastric cancer and beyond. Nature Reviews Cancer **10**(6), 403-414. http://dx.doi.org/10.1038/nrc2857

Protasio AV, Tsai IJ, Babbage A, Nichol S, Hunt M *et al.* (2012) A systematically improved high quality genome and transcriptome of the human blood fluke Schistosoma mansoni. PLoS neglected tropical diseases **6**(1), e1455-e1455. http://dx.doi.org/10.1371/journal.pntd.0001455

Shima H, Takano M, Shimotohno K and Miwa M (1986) Identification of p26Xb and p24Xb of human T-cell leukemia virus type II. FEBS Lett **209**(2), 289-294. http://dx.doi.org/10.1016/0014-5793(86)81129-2

Smatti MK, Al-Sadeq DW, Ali NH, Pintus G, Abou-Saleh H *et al.* (2018) Epstein-Barr Virus Epidemiology, Serology, and Genetic Variability of LMP-1 Oncogene Among Healthy Population: An Update. Front Oncol **8**, 211. http://dx.doi.org/10.3389/fonc.2018.00211

Steinhauf D, Rodriguez A, Vlachakis D, Virgo G, Maksimov V *et al.* (2014) Silencing motifs in the Clr2 protein from fission yeast, Schizosaccharomyces pombe. PLoS ONE **9**(1), e86948. http://dx.doi.org/10.1371/journal.pone.0086948

Thorell K, Lehours P and Vale FF (2017) Genomics of Helicobacter pylori. Helicobacter **22**(S1), e12409. http://dx.doi.org/https://doi.org/10.1111/hel.12409

Tomasetti C, Li L and Vogelstein B (2017) Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science **355**(6331), 1330-1334. http://dx.doi.org/10.1126/science.aaf9011

Trimpalis P, Koumandou VL, Pliakou E, Anagnou NP and Kossida S (2013) Gene fusion analysis in the battle against the African endemic sleeping sickness. PLoS ONE **8**(7), e68854. http://dx.doi.org/10.1371/journal.pone.0068854

Tsagrasoulis D, Danos V, Kissa M, Trimpalis P, Koumandou VL *et al.* (2012) SAFE Software and FED Database to Uncover Protein-Protein Interactions using Gene Fusion Analysis. Evol Bioinform Online **8**, 47-60. http://dx.doi.org/10.4137/ebo.s8018

Tsaniras S, Vlachakis D and Taraviras S (2015) The Nucleophosmin-Pin1 interaction links the cell cycle, cancer and pluripotency. J Mol Biochem **4**(3), 50-51.

Utzinger J, Raso G, Brooker S, De Savigny D, Tanner M *et al.* (2009) Schistosomiasis and neglected tropical diseases: towards integrated and sustainable control and a word of caution. Parasitology **136**(13), 1859-1874. http://dx.doi.org/10.1017/s0031182009991600

Vineis P, Schatzkin A and Potter JD (2010) Models of carcinogenesis: an overview. Carcinogenesis **31**(10), 1703-1709. http://dx.doi.org/10.1093/carcin/bgq087

Vlachakis D, Champeris Tsaniras S, Karozou A and Kossida S (2013a) An update on virology and emerging viral epidemics. J Mol Biochem **2**(2), 80-84.

Vlachakis D, Pavlopoulou A, Kazazi D and Kossida S (2013b) Unraveling microalgal molecular interactions using evolutionary and structural bioinformatics. Gene **528**(2), 109-119. http://dx.doi.org/10.1016/j.gene.2013.07.039

Xie C, Xu LY, Yang Z, Cao XM, Li W *et al.* (2014) Expression of γH2AX in various gastric pathologies and its association with Helicobacter pylori infection. Oncol Lett **7**(1), 159-163. http://dx.doi.org/10.3892/ol.2013.1693

Yu Y-P, Liu P, Nelson J, Hamilton RL, Bhargava R *et al.* (2019) Identification of recurrent fusion genes across multiple cancer types. Scientific Reports **9**(1), 1074. http://dx.doi.org/10.1038/s41598-019-38550-6

EMBnet.journal
Bioinformatics in Action

Papakonstantinou et al. (2022) EMBnet.journal 27, e1009
http://dx.doi.org/10.14806/ej.27.0.1009

Research Papers

A rational structure-based drug design strategy for the discovery of novel antiviral agents against the Yellow Fever Virus helicase

Eleni Papakonstantinou[1], Katerina Pierouli[1], George N Goulielmos[2], Elias Eliopoulos[1]✉

[1]Laboratory of Genetics, Department of Biotechnology, School of Applied Biology and Biotechnology, Agricultural University of Athens, Athens, Greece

[2]Section of Molecular Pathology and Human Genetics, Department of Internal Medicine, School of Medicine, University of Crete, Heraklion, Greece

Competing interests: EP none; KP none; GNG none; EE none

Abstract

Yellow Fever is a viral hemorrhagic disease that is transmitted mainly through arthropods with high mortality rates. Yellow Fever Virus (YFV) is an enveloped positive sense single-stranded RNA virus, member of the *Flaviviridae* family and the *Flavivirus* genus, and is endemic in countries of Africa and South America. However, recent cases of infection in North America, Asia and Europe are highlighting the potential risk of an outbreak with no effective treatment available and the urgent need to develop potent antiviral agents against the YFV. In this direction, a range of specific modulators were designed and *in silico* evaluated in an effort to hinder the enzymatic activity of the YFV helicase as a prominent pharmacological target. Following a structure-based rational drug design pipeline, a phylogenetic analysis of *Flaviviridae* viruses and an in-depth evolutionary study on the Yellow Fever Virus helicase has provided invaluable insights into structural conservation and structural elements and features that are vital for the viral helicase function. Using comparative modelling and molecular dynamics simulations the YFV helicase-ssRNA complex was established, and the specific molecular interactions and physicochemical properties of the complex could be analyzed and used towards the designing and elucidation of a specific YFV 3D pharmacophore model. A high throughput virtual screening simulation was conducted to assess a set of in-house maintained low molecular weight compounds as bioactive inhibitors of the YFV helicase enzyme. The *in-silico* study described herein, could pave the way towards the designing and more efficient screening of potential novel modulator compounds against the YFV as well as attest and designate the NS3 helicase as an antiviral pharmacological target of uttermost value and potential.

Introduction

The viral family *Flaviviridae* comprises the genera *Flavivirus, Hepacivirus, Pestivirus*, and the Unclassified genus, and includes numerous important human and animal pathogens. The most common pathogens of the genus Flavivirus are Tick-borne encephalitis viruses, Dengue virus (DENV), Yellow fever virus (YFV), West Nile virus (WNV) and Zika virus (ZIKV) and are transmitted mainly by arthropods. YFV is transmitted by the mosquitoes *Aedes aegypti* and *Haemagogus leucocelaenus*. The treatment and management of flaviviruses' infection is not 100% effective, even in cases where vaccines are available. It is thus highly important to study these viruses in terms of their genetic material and mechanisms of infection and reproduction in order to find efficient ways for their constrain in case of an outbreak (Best, 2016).

The small, enveloped virions of the different members of the *Flaviviridae* family contains a single-stranded, positive-sense RNA genome of about 9.5–12.5 kb. Their genome consists of a single, long open reading frame (ORF), flanked by untranslated regions (UTRs) at 5' and 3' ends. Extensive studies on sub-genomic Flavivirus RNA replicons have revealed that the non-structural (NS) proteins, which are encoded by the C-terminal part of the polyprotein, play a crucial role in viral RNA replication. Accordingly, these proteins are assumed to form replication complexes in conjunction with genomic RNA and possibly with other cellular factors (Best, 2016; Chambers *et al.*, 1990). Inhibition of viral proteins, mainly NS3 helicase and NS5 polymerase,

Article history
Received: 05 November 2021
Accepted: 12 November 2021
Published: 07 July 2022

Papakonstantinou et al. (2022) EMBnet.journal 27, e1009
http://dx.doi.org/10.14806/ej.27.0.1009

is becoming increasingly popular (Papageorgiou *et al.*, 2016; Vlachakis, 2021). More specifically, the NS3 protein is a multifunctional polypeptide and encodes three enzymes with different functions including a serine protease, a NTPase and an RNA helicase. The RNA helicase is encoded by the C-terminal domain of the NS3 protein (aa 180-618) and belongs to the helicase superfamily 2 (SF2). Its structure consists of three subdomains, where subdomains 1 and 2 contain eight conserved motifs essential for RNA binding, ATP hydrolysis and structural stability (Fairman-Williams *et al.*, 2010; Pyle, 2008) adopting the RecA-like fold (Rao and Rossmann, 1973), and subdomain 3 forms the single-stranded RNA binding tunnel. Subdomain 3 also mediates the interaction between NS3 and NS5, and the disruption of this interaction is also considered a powerful and effective strategy for designing antiviral compounds (Tay *et al.*, 2015). Viruses carrying an impaired NS3 helicase gene cannot reproduce properly, proving the essential role of NS3 helicase activity in virus replication.

Viral helicase activity is essential for the virus during its reproductive process. NS3 protein appears to be a potential pharmacological target for inhibiting YFV replication (García *et al.*, 2017) and antiviral strategies aiming for flavivirus helicase inhibition has been implemented in various cases (Lim *et al.*, 2013; Luo *et al.*, 2015). Although rare reports of NS3 helicase inhibitors have been reported to date, the presence of halogenated benzenes that inhibit WNV helicase (Sampath and Padmanabhan, 2009) and ivermectin, an antiparasitic drug for helminths, that inhibits JEV and YFV helicases, are reported (Lai *et al.*, 2017; Mastrangelo *et al.*, 2012). In addition, ST-610 and suramin have been reported as DENV helicase inhibitors (Basavannacharya and Vasudevan, 2014; Lim *et al.*, 2013). Compound ML283, which has been shown to act as an inhibitor of HCV and DENV helicases and pyrrolone, acts as a helicase inhibitor for both DENV and WNV (Sweeney *et al.*, 2015). Although increasingly more studies are being conducted at the development of inhibitors for viral helicases, the lack of specific pockets in RNA and NTP binding sites poses a significant problem in the process as significant toxicity may occur, as compounds targeting these sites may also to target many similar cellular proteins with helicase or NTPase functions. In addition, another problem in the development of high affinity and potency inhibitors is the inherent flexibility of motor proteins, however allosteric inhibition still remains an attractive idea for inhibitor design (Li *et al.*, 2013). In conclusion, to date no helicase inhibitor has been approved for clinical usage, which may be due to the above limitations in the process of *in-silico* drug development.

Sequence alignments of the Yellow Fever viral helicase identified several conserved sequence motifs that are important for biological functions. So far, the crystal structures of helicases from various RNA viruses have been determined, including the helicases from Yellow Fever Virus (Wu *et al.*, 2005), Hepatitis C virus (Yao *et al.*, 1997), Dengue virus (Luo *et al.*, 2008a), Zika virus (Tian *et al.*, 2016), and Kunjin virus (Mastrangelo *et al.*, 2007). In the present work, the three-dimensional structure of the helicase enzyme of Yellow Fever virus in complex with a ssRNA molecule was predicted through comparative modelling and a 3D pharmacophore was developed in order to scan and detect specific helicase inhibitors with antiviral potential.

Methods

Sequence Alignment and Phylogenetic Analysis

The amino acid sequence of Yellow Fever viral helicase was obtained from the GenBank database (accession no.: NC_002031, entry name: Yellow Fever virus, complete genome). All available sequences of *Flaviviridae* NS3s were collected from the NIAID Virus Pathogen Database and Analysis Resource (ViPR) (Pickett *et al.*, 2011) and the NCBI RefSeq database. Representative sequences were selected and sequence alignment was performed using the ClustalO algorithm in the Jalview program (Waterhouse *et al.*, 2009a). The phylogenetic trees were constructed with the Neighbor Joining algorithm (Saitou and Nei, 1987) and visualization was performed using iTol[1] and Jalview software (Waterhouse *et al.*, 2009b).

Energy Minimisation

Initially, available structures of Flaviviridae helicases were queried in the RSCB Protein Data Bank and a total of 110 structures were identified. 17 representative structures from each species were selected and structural studies were performed to optimize and evaluate the three-dimensional (3D) structure of the X-ray determined YFV helicase (PDB ID: 1YKS) and the other viral helicase structures. Energy minimization was used to remove any residual geometrical strain in each molecular system, using the CHARMM27 forcefield (Foloppe and MacKerell, 2000). Sequence alignments and structural superpositions were performed using the ClustalO algorithm (Sievers *et al.*, 2011) and the MOE software (Group, 2019) respectively.

Molecular electrostatic potential (MEP)

Electrostatic potential surfaces were calculated by solving the nonlinear Poisson–Boltzmann equation using the finite difference method as implemented in the PyMOL Software (Schrödinger, 2020). The potential was calculated on grid points per side (65, 65, 65) and the grid fill by solute parameter was set to 80%. The dielectric constants of the solvent and the solute were set to 80.0 and 2.0, respectively. An ionic exclusion radius of 2.0 Å, a solvent radius of 1.4 Å and a solvent ionic strength of 0.145 M were applied. Amber99 charges and atomic radii were used for this calculation.

[1]https://itol.embl.de

Model Optimization

Energy minimization was done in MOE initially using the CHARMM27 forcefield (Foloppe and MacKerell, 2000) implemented into the same package, up to a RMSD gradient of 0.0001 to remove the geometrical strain. The model was subsequently solvated with SPC water using the truncated octahedron box extending to 7 Å from the model and molecular dynamics were performed after that at 300K, 1 atm with 2 second step size and for a total of ten nanoseconds, using the NVT ensemble in a canonical environment. NVT stand for Number of atoms, Volume and Temperature that remain constant throughout the calculation. The results of the molecular dynamics simulation were collected into a database by MOE and can be further analyzed.

Model Evaluation

The produced models were initially evaluated within the MOE package (Group, 2019) by a residue packing quality function, which depends on the number of buried non-polar side chain groups and on hydrogen bonding.

High-Throughput Virtual Screening and in-silico de novo drug design

A 3D pharmacophore model was constructed using the Pharmacophore tool in MOE (Group, 2019) and representative pharmacophoric features were selected based on the ssRNA-helicase interactions. High-throughput virtual screening simulations were consequently performed using the pharmacophore query tool in MOE. Novel molecules were *in-silico* designed based on the chemical structures of the WO/2009/125191 patent for HCV helicase inhibitors as scaffolds using the MOE BREED module and were consequently *in-silico* evaluated based on their binding free energies.

Results and Discussion

Description of the Yellow Fever virus helicase structure

The Yellow Fever virus helicase model exhibits the structural features of known *Flaviviridae* helicases, and its structure has been experimentally determined by Wu *et al.* at 1.80Å resolution. Namely, the three distinct domains of helicases as well as the various motifs are structurally similar. The GxGKT/S motif in domain 1 is one of the most crucial motifs in *Flaviviridae* helicases, which is conserved to the same loop in kinases. It is a Walker A motif and binds the β-phosphate of ATP (Saraste *et al.*, 1990). The importance of this motif is highlighted in site directed mutagenesis studies by the fact that the mutant protein is inactive. Furthermore, the DExH motif is another crucial motif for the helicase function, which is responsible for the binding of the Mg^{2+}–ATP substrate. According to studies in adenylate and thymidine kinases, an aspartate (Asp170) has been revealed that binds the Mg^{2+} helping in the establishment of the ATP optimum orientation for nucleophilic attack (Ruff *et al.*, 1991). Finally, QRxGRxGR motif is also a crucial motif, the role of which is exceptionally crucial to the *Flaviviridae* helicase function as it is involved in nucleic acid binding (Gross and Shuman, 1996; Vlachakis, 2009).

The *Flaviviridae* helicases have three domains in total, which are separated by two channels. The first and third domains are more interacting together in contrast with domain two. During the unwinding of double-stranded nucleic acids, domain two undergoes significant movements compared to the other two domains. The channel between domains 3 and 1-2 accommodate the ssRNA during the viral unwinding process. The second domain contains an arginine-rich site where RNA binds to the helicase. The ATP and ssRNA sites were found to have been conserved on the Yellow Fever Virus helicase model (Luo *et al.*, 2008b).

Comparative Modelling

The NS3 domain of Flaviviridae contains both the protease and the helicase coding regions. For this study, all available helicase structures of the *Flaviviridae* family were retrieved and filtered to remove redundant and duplicate structures (17 out of 110 in total) and were consequently minimized to remove geometrical strains in each molecular system. The optimized structures were aligned and superposed against the Yellow Fever virus helicase sequence. All the major helicase motifs, characteristic of the *Flaviviridae* family (Garg *et al.*, 2013) were found to be conserved both in sequence and structure and the constructed phylogenetic tree represents the relations between these viral species (Figure 1).

The overall alignment showed a sequence identity that ranged from 18.2% (HCV helicase, PDB ID: 1A1V) to 50.0% (Dengue virus 4 helicase, PDB ID: 5XC6), whereas sequence similarity of the sequences ranged from 28.4% to 66.8%, respectively. The alpha-carbon structural superposition of the Flaviviridae helicases against the Yellow Fever helicase exhibited major differences in their domain orientations and features and the resulting RMSD ranged from 1.663 Å to 7.970 Å. The most similar structures identified were the flavivirus helicases of Zika virus (PDB ID: 5MFX) and Dengue virus 4 (PDB ID: 5XC6) (Figure 2).

To construct the ssRNA-helicase complex of the Yellow Fever virus, three available Flaviviridae helicase-ssRNA complexes were used as templates (HCV PDB ID:1A1V; ZV PDB ID:5MFX; DV4 PDB ID:5XC6). The coordinates of the ssRNA substrates were transferred to the Yellow Fever helicase structure according to the superposed structures and the three resulting model complexes were evaluated. The models were subjected to energy minimization and molecular dynamics (MD) simulations in the presence of the ssRNA substrate. Based on structural stability and superposition of the consequent binding sites, the Zika virus helicase-

Papakonstantinou et al. (2022) EMBnet.journal 27, e1009
http://dx.doi.org/10.14806/ej.27.0.1009

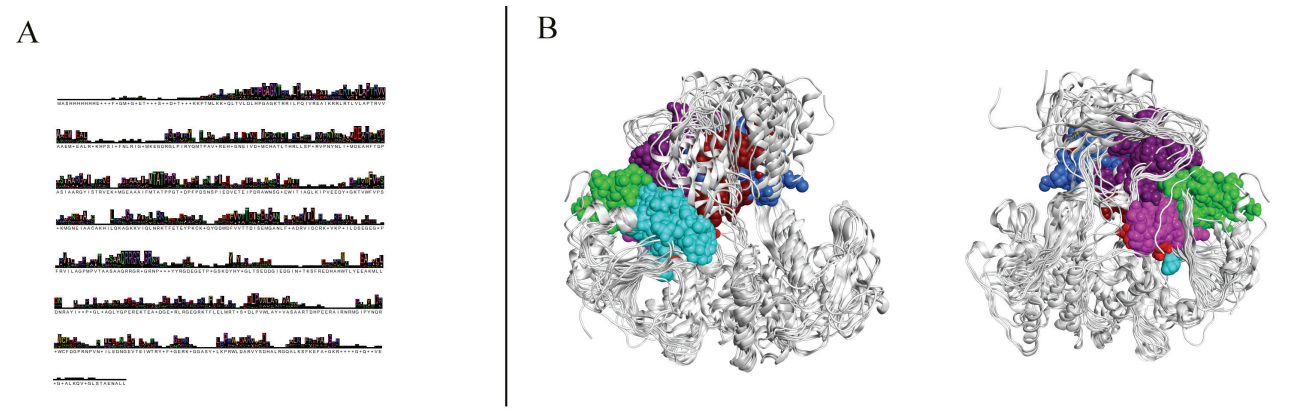


Figure 1. A: Sequence alignment between the representative Flaviviridae helicase sequences where all seven motifs are identified. **B:** A representative phylogenetic tree of the Flaviviridae helicase enzymes. **C:** Structural superposition of the 3-D structures of the representative Flaviviridae helicase enzymes. All seven major conserved motifs of Flaviviridae helicases have been color-coded and represented in CPK format.

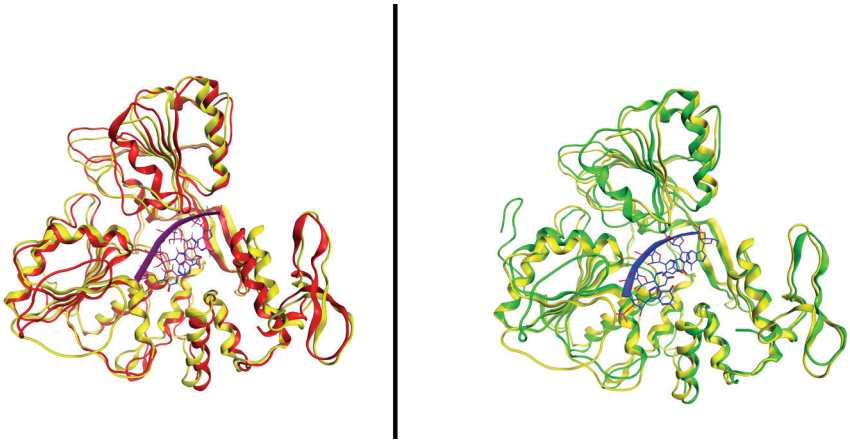


Figure 2. A: The Yellow Fever virus helicase in yellow (PDB ID: 1YKS) superimposed with the Zika virus helicase in red (PDB ID: 5MFX) and the ssRNA substrate in magenta. **B:** The Yellow Fever virus helicase in yellow (PDB ID: 1YKS) superimposed with the Dengue virus 4 helicase in green (PDB ID: 5XC6) and the ssRNA substrate in blue.

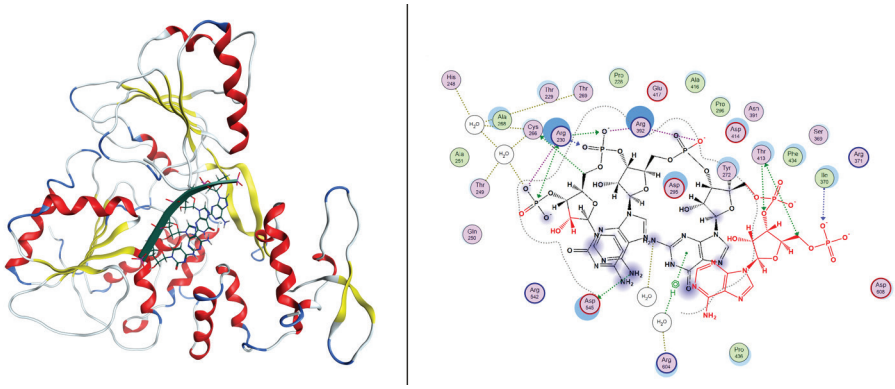


Figure 3. The Yellow Fever helicase-ssRNA complex model. **A:** The 3-D model of the YF helicase in cartoon representation color coded by structural elements, with the modelled ssRNA fragment in ribbon and stick representation in dark green. **B:** The interaction map of the per-residue ssRNA interaction pattern from Ligplot for Yellow Fever virus helicase.

ssRNA complex was chosen as the most appropriate template (Figure 3). Invariant residues of numerous motifs in the vicinity of the substrate in the Zika virus template structure were conserved in the Yellow Fever virus helicase structural model. Interactions of Yellow Fever helicase-ssRNA fragment were established with the backbone of the ssRNA fragment, that create non-specific protein–nucleic acid interactions. The bases in the middle of the ssRNA do not appear to interact with the protein. The contacts of the enzymatic receptor emerge mostly from domains one and two of the Yellow Fever helicase and, specifically, from loops between secondary

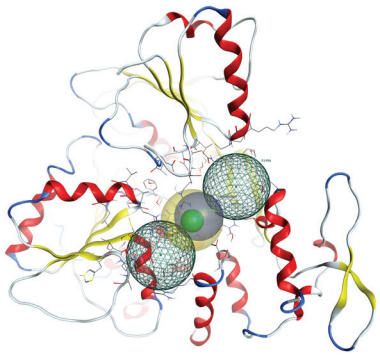


Figure 4. The pharmacophore model for the RNA binding site of the Yellow Fever helicase structure. 5 pharmacophoric sites are represented, two projected ring features in both ends of the channel in dark green mesh, and three dynamic sites in the center of the channel, one aromatic or hydrophobic centroid feature in light green sphere, one site of H-donor or H-acceptor in blue sphere, and one site of H-bond or anion feature in yellow sphere.

structure elements of the latter domains. LigPlot, which is a built-in module of MOE, was used for the drawing of more detailed (i.e. per residue) comparison of the ssRNA interaction pattern between the Yellow Fever virus helicase model and the Zika virus helicase structure.

The Yellow Fever virus helicase-ssRNA model reached a conformational equilibrium similar to that of Zika virus complex based on the 10ns MD simulations revealed. Thus, the viability of the comparative modelling of the Yellow Fever complex model was illustrated by these observations. The Yellow Fever complex model was compared with its template structure by calculating the root mean square deviations (RMSD) between equivalent atoms for the full MD course for evaluation. Large values of RMSD are indicative of systems of poor quality. The Cα RMSD of the Yellow Fever virus helicase model from the equivalent domains of the template structures was less than 0.65. This low value of RMSD reflected the high similarity of this structures since it seems to remain conformationally close to the template structure upon the minimization and the molecular dynamics simulation course that followed.

The electrostatic potential surface was calculated to analyze the molecular surface of the simulated Yellow Fever virus complex. In order to compare directly the template structure used in this study, electrostatic potential surfaces were also calculated for the Zika virus helicase. According to the results, the two helicases exhibited almost identical electrostatic surfaces and shared common features such as a negatively charged ssRNA entrance to the helicase tunnel verifying the validity of the model, which was found to share a similar electrostatic surface to its X-ray crystal structure complex template.

Pharmacophore modelling and in-silico de novo drug design

Following the establishment of the YFV helicase-ssRNA model, the specific molecular interactions and physicochemical properties of the complex were analyzed. Based on the interactions identified, a 3-D dynamic pharmacophore model was created, to represent the interaction sites and nucleotide binding channel (Figure 4). The pharmacophore consists of 5 sites representing two projected ring features, one site that represents an annotation of aromatic or hydrophobic centroid feature, one site of H-donor or acceptor feature, and one site of H-bond or anion features. The aromatic ring features are located at the edges of the binding site, whereas the rest of the features are found in the core of the channel. Novel compounds were *in-silico* designed based on evaluated molecules included in the WO/2009/125191 patent, that encompasses molecular structures suitable for use in the treatment of HCV infection against the viral helicase. These compounds are symmetrical in their chemical structure and features accounting for the non-directionality of the nucleotide substrates. Based on their features and interaction properties, these molecules were used as scaffolds in the MOE BREED module and novel structures were generated. The module implements the crossover operator of genetic algorithms by evaluating important features of the original structures and combines fragments to produce novel new energy structures with similar orientation. The designed structures preserve significant intramolecular interactions of the lead compounds. Based on the pharmacophoric representation of the binding site, the collection of the in-house maintained designed compounds were evaluated based on the London dG scoring function of the free energy of binding for each ligand. The pharmacophore-based high-throughput virtual screening identified top compounds that could act as bioactive inhibitors of the YFH helicase, disrupting the binding of single stranded RNA and obstructing the enzyme function.

Conclusions

Computer-based methodologies have become an integral part of the process of developing new drugs and repurposing of approved drugs against numerous diseases, as well as discover potential pharmacological targets. Developed applications for computational drug design are continuously upgrading and transforming traditional methodologies and pipelines, speeding up the research in antiviral strategies. In contrast to the traditional drug development methods which are time consuming and costly, computational drug design methods are widely used in the development of antivirals (Shaker *et al.*, 2021). These state-of-the-art techniques dock small molecules into macromolecular targets and predict the affinity and activity of small molecules (Dalkas *et al.*, 2012). Interestingly, information technologies

Papakonstantinou et al. (2022) EMBnet.journal 27, e1009
http://dx.doi.org/10.14806/ej.27.0.1009

Table 1. Smiles representation and London dG scoring values for the top 10 novel designed molecules.

mol	smiles	London dG
1	O=C(Nc1ccc(-c2[nH]c3c(n2)cccc3)cc1)c1c(-c2ccc(-c3[nH]c4c(n3)cccc4)cc2)cc(C(=O)Nc2ccc(-c3[nH]c4c(n3)cccc4)cc2)cc1	-7.6665
2	O=C(Nc1cc(-c2[nH]c3c(n2)cccc3)c(-c2[nH]c3c(n2)cccc3)cc1)CCCCC(=O)Nc1ccc(-c2[nH]c3c(n2)cccc3)cc1	-6.9248
3	O=C(Nc1ccc(-c2[nH]c3c(n2)cccc3)cc1)c1c(-c2ccc(-c3[nH]c4c(n3)cccc4)cc2)cc(-c2ccc(C(=O)Nc3ccc(-c4[nH]c5c(n4)cccc5)cc3)cc2)cc1	-6.0711
4	O=C(Nc1ccc(-c2[nH]c3c(n2)cccc3)cc1)CCCCCCCCNc1ccc(-c2[nH]c3c(n2)cccc3)cc1	-6.0541
5	O=C(Nc1ccc(-c2[nH]c3c(n2)cccc3)cc1)CCCCCCCCC(=O)N	-5.9514
6	O=C(Nc1ccc(-c2[nH]c3c(n2)cccc3)cc1)CC[C@H]1[C@H](C(=O)Nc2ccc(-c3[nH]c4c(n3)cccc4)cc2)CCCC1	-5.6812
7	O=C(Nc1ccc(-c2[nH]c3c(n2)cccc3)cc1)C(=O)c1ccc(C(=O)Nc2ccc(-c3[nH]c4c(n3)cccc4)cc2)cc1	-5.4822
8	O=C(Nc1ccc(-c2[nH]c3c(n2)cccc3)cc1)CCCCCc1c(NC(=O)c2c(C(=O)Nc3ccc(-c4[nH]c5c(n4)cccc5)cc3)cccc2)ccc(-c2[nH]c3c(n2)cccc3)c1	-5.4674
9	O=C(Nc1ccc(-c2[nH]c3c(n2)cccc3)cc1)CCCCCCCC=O	-5.4051
10	O=C(Nc1ccc(-c2[nH]c3c(n2)cccc3)cc1)c1ccc(C(=O)c2ccc(-c3[nH]c4c(n3)cccc4)cc2)cc1	-5.3922

and machine learning algorithms are almost inevitably implemented in these new approaches to improve the efficacy of the prediction.

Especially nowadays that we are going through a period of public health crisis due to the SARS-CoV-2 virus, the need to develop a quick and efficient pipeline for identifying potential antiviral drugs for future health risks is imperative (Basu *et al.*, 2021). Members of the Flaviviridae family, are already endemic to African and South American countries. In addition, several cases of *Flaviviridae* outbreaks are being reported in Southern Europe and America (WHO, 7 May 2019). Numerous studies are being performed to detect effective drug targets in several viruses of this family, with nonstructural proteins, such as the viral protease, helicase and polymerase enzymes, being among the most prominent pharmacological targets (Vlachakis, 2021). Thus, computational methods for homology modelling and prediction of viral protein structures, such as bovine viral diarrhea virus (BVDV) (Xu *et al.*, 1997), Classical Swine Fever virus (Li *et al.*, 2018), Dengue virus and Zika virus (Ekins *et al.*, 2016; Jain *et al.*, 2016), are used to develop new promising viral inhibitors.

In this study, the three-dimensional structure of the Yellow Fever virus helicase in complex with a single stranded RNA molecule was established based on available templates of helicase enzymes cocrystallized with ssRNA of the *Flaviviridae* family. An extensive comparative analysis was performed and the produced model using the structure of Zika virus helicase as template was optimized. The evaluation of the model was carried out successfully in terms of geometry, fold recognition as well as in terms of the criteria required for members of the viral *Flaviviridae* family. In addition, the Yellow Fever virus complex model was evaluated by molecular dynamics simulations and used to design a 3-D pharmacophore, indicative of the RNA binding site properties. Novel chemical structures were designed through an *in-silico* approach that combines significant features of evaluated structures against viral helicases through the implementation of genetic algorithms. A pharmacophore-based screening was performed, and potent molecules were evaluated and recognized as potential inhibitors of the activity of the Yellow Fever virus helicase. Our applied methodology is paving the way towards the designing and more efficient screening of potential novel modulator compounds against the YFV, as well as attest and designate the NS3 helicase as an antiviral pharmacological target of uttermost value and potential.

Key Points

- Yellow Fever is a viral hemorrhagic with high mortality rates.
- Designing and screening for a potential novel modulator against the YFV helicase is of great value.
- A rational structure-based drug design was performed for the evaluation of YFV helicase inhibitora.
- Computer-based methodologies are an integral part of the process of novel drug development and the discovery of potential pharmacological targets.

Funding

This work is funded by the ESPA Young Researchers Support, «Rational Drug Design of Novel Antiviral Agents against the Helicase Enzyme of the Yellow Fever Virus», MIS 5048546, NSRF 2014 – 2020.

References

Basavannacharya C and Vasudevan SG (2014) Suramin inhibits helicase activity of NS3 protein of dengue virus in a fluorescence-based high throughput assay format. Biochem Biophys Res Commun **453**(3), 539-544. http://dx.doi.org/10.1016/j.bbrc.2014.09.113

Basu S, Ramaiah S and Anbarasu A (2021) In-silico strategies to combat COVID-19: A comprehensive review. Biotechnol Genet Eng Rev **37**(1), 64-81. http://dx.doi.org/10.1080/02648725.2021.1966920

Best SM (2016) Flaviviruses. Curr Biol **26**(24), R1258-r1260. http://dx.doi.org/10.1016/j.cub.2016.09.029

Chambers TJ, Hahn CS, Galler R and Rice CM (1990) Flavivirus genome organization, expression, and replication. Annu Rev Microbiol **44**, 649-688. http://dx.doi.org/10.1146/annurev.mi.44.100190.003245

Dalkas GA, Vlachakis D, Tsagkrasoulis D, Kastania A and Kossida S (2012) State-of-the-art technology in modern computer-aided drug design. Briefings in Bioinformatics **14**(6), 745-752. http://dx.doi.org/10.1093/bib/bbs063

Ekins S, Liebler J, Neves BJ, Lewis WG, Coffee M *et al.* (2016) Illustrating and homology modeling the proteins of the Zika virus. F1000Res **5**, 275. http://dx.doi.org/10.12688/f1000research.8213.2

Fairman-Williams ME, Guenther U-P and Jankowsky E (2010) SF1 and SF2 helicases: family matters. Current Opinion in Structural Biology **20**(3), 313-324. http://dx.doi.org/https://doi.org/10.1016/j.sbi.2010.03.011

Foloppe N and MacKerell A (2000) All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. Journal of Computational Chemistry **21**, 86-104. http://dx.doi.org/10.1002/(SICI)1096-987X(20000130)21:2%3C86::AID-JCC2%3E3.0.CO;2-G

García LL, Padilla L and Castaño JC (2017) Inhibitors compounds of the flavivirus replication process. Virol J **14**(1), 95. http://dx.doi.org/10.1186/s12985-017-0761-1

Garg H, Lee RTC, Tek NO, Maurer-Stroh S and Joshi A (2013) Identification of conserved motifs in the Westnile virus envelope essential for particle secretion. BMC Microbiology **13**(1), 197. http://dx.doi.org/10.1186/1471-2180-13-197

Gross CH and Shuman S (1996) The QRxGRxGRxxxG motif of the vaccinia virus DExH box RNA helicase NPH-II is required for ATP hydrolysis and RNA unwinding but not for RNA binding. J Virol **70**(3), 1706-1713. http://dx.doi.org/10.1128/jvi.70.3.1706-1713.1996

Group CC (2019) Molecular Operating Environment (MOE). 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2021.

Jain R, Coloma J, García-Sastre A and Aggarwal AK (2016) Structure of the NS3 helicase from Zika virus. Nat Struct Mol Biol **23**(8), 752-754. http://dx.doi.org/10.1038/nsmb.3258

Lai JH, Lin YL and Hsieh SL (2017) Pharmacological intervention for dengue virus infection. Biochem Pharmacol **129**, 14-25. http://dx.doi.org/10.1016/j.bcp.2017.01.005

Li K, Frankowski KJ, Hanson AM, Ndjomou J, Shanahan MA *et al.* (2013) Hepatitis C virus NS3 helicase inhibitor discovery. Probe Reports from the NIH Molecular Libraries Program [Internet].

Li W, Wu B, Soca WA and An L (2018) Crystal Structure of Classical Swine Fever Virus NS5B Reveals a Novel N-Terminal Domain. J Virol **92**(14). http://dx.doi.org/10.1128/jvi.00324-18

Lim SP, Wang QY, Noble CG, Chen YL, Dong H *et al.* (2013) Ten years of dengue drug discovery: progress and prospects. Antiviral Res **100**(2), 500-519. http://dx.doi.org/10.1016/j.antiviral.2013.09.013

Luo D, Vasudevan SG and Lescar J (2015) The flavivirus NS2B-NS3 protease-helicase as a target for antiviral drug development. Antiviral Res **118**, 148-158. http://dx.doi.org/10.1016/j.antiviral.2015.03.014

Luo D, Xu T, Hunke C, Grüber G, Vasudevan SG *et al.* (2008a) Crystal structure of the NS3 protease-helicase from dengue virus. J Virol **82**(1), 173-183. http://dx.doi.org/10.1128/JVI.01788-07

Luo D, Xu T, Watson RP, Scherer-Becker D, Sampath A *et al.* (2008b) Insights into RNA unwinding and ATP hydrolysis by the flavivirus NS3 protein. Embo j **27**(23), 3209-3219. http://dx.doi.org/10.1038/emboj.2008.232

Mastrangelo E, Milani M, Bollati M, Selisko B, Peyrane F *et al.* (2007) Crystal structure and activity of Kunjin virus NS3 helicase; protease and helicase domain assembly in the full length NS3 protein. J Mol Biol **372**(2), 444-455. http://dx.doi.org/10.1016/j.jmb.2007.06.055

Mastrangelo E, Pezzullo M, De Burghgraeve T, Kaptein S, Pastorino B *et al.* (2012) Ivermectin is a potent inhibitor of flavivirus replication specifically targeting NS3 helicase activity: new prospects for an old drug. J Antimicrob Chemother **67**(8), 1884-1894. http://dx.doi.org/10.1093/jac/dks147

Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB *et al.* (2011) ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Research **40**(D1), D593-D598. http://dx.doi.org/10.1093/nar/gkr859

Pyle AM (2008) Translocation and unwinding mechanisms of RNA and DNA helicases. Annu Rev Biophys **37**, 317-336. http://dx.doi.org/10.1146/annurev.biophys.37.032807.125908

Rao ST and Rossmann MG (1973) Comparison of super-secondary structures in proteins. Journal of Molecular Biology **76**(2), 241-256. http://dx.doi.org/10.1016/0022-2836(73)90388-4

Ruff M, Krishnaswamy S, Boeglin M, Poterszman A, Mitschler A *et al.* (1991) Class II aminoacyl transfer RNA synthetases: crystal structure of yeast aspartyl-tRNA synthetase complexed with tRNA(Asp). Science **252**(5013), 1682-1689. http://dx.doi.org/10.1126/science.2047877

Saitou N and Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution **4**(4), 406-425. http://dx.doi.org/10.1093/oxfordjournals.molbev.a040454

Sampath A and Padmanabhan R (2009) Molecular targets for flavivirus drug discovery. Antiviral Res **81**(1), 6-15. http://dx.doi.org/10.1016/j.antiviral.2008.08.004

Saraste M, Sibbald PR and Wittinghofer A (1990) The P-loop--a common motif in ATP- and GTP-binding proteins. Trends Biochem Sci **15**(11), 430-434. http://dx.doi.org/10.1016/0968-0004(90)90281-f

Schrödinger L, & DeLano, W. (2020). "PyMOL." from http://www.pymol.org/pymol

Shaker B, Ahmad S, Lee J, Jung C and Na D (2021) In silico methods and tools for drug discovery. Comput Biol Med **137**, 104851. http://dx.doi.org/10.1016/j.compbiomed.2021.104851

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol **7**, 539. http://dx.doi.org/10.1038/msb.2011.75

Sweeney NL, Hanson AM, Mukherjee S, Ndjomou J, Geiss BJ *et al.* (2015) Benzothiazole and Pyrrolone Flavivirus Inhibitors Targeting the Viral Helicase. ACS Infect Dis **1**(3), 140-148. http://dx.doi.org/10.1021/id5000458

Tay MYF, Saw WG, Zhao Y, Chan KWK, Singh D *et al.* (2015) The C-terminal 50 Amino Acid Residues of Dengue NS3 Protein Are Important for NS3-NS5 Interaction and Viral Replication*. Journal of Biological Chemistry **290**(4), 2379-2394. http://dx.doi.org/10.1074/jbc.M114.607341

Tian H, Ji X, Yang X, Xie W, Yang K *et al.* (2016) The crystal structure of Zika virus helicase: basis for antiviral drug design. Protein Cell **7**(6), 450-454. http://dx.doi.org/10.1007/s13238-016-0275-4

Vlachakis D (2009) Theoretical study of the Usutu virus helicase 3D structure, by means of computer-aided homology modelling. Theoretical Biology and Medical Modelling **6**(1), 9. http://dx.doi.org/10.1186/1742-4682-6-9

Vlachakis D (2021) Genetic and structural analyses of ssRNA viruses pave the way for the discovery of novel antiviral pharmacological targets. Molecular Omics **17**(3), 357-364. http://dx.doi.org/10.1039/D0MO00173B

Waterhouse AM, Procter JB, Martin DM, Clamp M and Barton GJ (2009a) Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics **25**(9), 1189-1191. http://dx.doi.org/10.1093/bioinformatics/btp033

Waterhouse AM, Procter JB, Martin DMA, Clamp M and Barton GJ (2009b) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics **25**(9), 1189-1191. http://dx.doi.org/10.1093/bioinformatics/btp033

WHO (7 May 2019). "Yellow fever." from https://www.who.int/news-room/fact-sheets/detail/yellow-fever.

Wu J, Bera AK, Kuhn RJ and Smith JL (2005) Structure of the Flavivirus helicase: implications for catalytic activity, protein interactions, and proteolytic processing. J Virol **79**(16), 10268-10277. http://dx.doi.org/10.1128/jvi.79.16.10268-10277.2005

Xu J, Mendez E, Caron PR, Lin C, Murcko MA *et al.* (1997) Bovine viral diarrhea virus NS3 serine proteinase: polyprotein cleavage sites, cofactor requirements, and molecular model of an enzyme essential for pestivirus replication. J Virol **71**(7), 5312-5322. http://dx.doi.org/10.1128/jvi.71.7.5312-5322.1997

Yao N, Hesson T, Cable M, Hong Z, Kwong AD *et al.* (1997) Structure of the hepatitis C virus RNA helicase domain. Nat Struct Biol **4**(6), 463-467. http://dx.doi.org/10.1038/nsb0697-463

EMBnet.journal
Bioinformatics in Action

Avramouli et al. (2022) EMBnet.journal 27, e1011
http://dx.doi.org/10.14806/ej.27.0.1011

Research Papers

Structural analysis on mutations related to Alzheimer's disease

Antigoni Avramouli, Eleftheria Polychronidou, Panayiotis Vlamos✉

BiHELab – Bioinformatics and Human Electrophysiology Lab, Department of Informatics of Ionian University, Corfu, Greece
Competing interests: AA none; EP none; PV none

Abstract

Proteins have a significant role in all biological processes. The functional properties of proteins rely upon their three-dimensional structures. Over the last twenty years substantial advances in genomic technologies have enhanced our knowledge of the genetics of Alzheimer's disease. To that end, the identification of mutations pathogenicity is still of vital importance. The methodology of the present research work focuses on the structural analysis of proteins related to Alzheimer's disease and the comparative study to create groups with clear structural similarity and pathogenicity. To achieve that, three-dimensional descriptors (fpfh, rsd and 3dsc) were applied along with supervised machine learning classification methods. In total, 62 APP, 286 PSEN1, 68 PSEN2 and 25 MAPT variants were evaluated in our study. The output of the methodology characterised thirty mutations that were unclear at the point of the data collection.

Introduction

Alzheimer's disease (AD), the most common neurodegenerative disease, is identified by an insidious decline in cognitive and memory function. Furthermore, the number of AD is growing rapidly with the increase in the aging population (Saez-Atienzar *et al.*, 2020). AD has a long prodromal phase, thus the onset of the pathogenetic changes until the appearance of clinical symptoms makes early diagnosis and treatment of this disease more demanding. The diagnosis criteria to define AD based on biomarker evidence currently include deposits of extracellular senile plaques in the cerebral cortex, the formation of neurofibrillary tangles, and neurodegeneration [AT(N)] classification system. As a result, there is still on-going research for improving the identification and classification of AD patients. To that end, the current research approach aims to deliver a methodology that can predict the classification of AD through the pathogenicity of the mutation. The pathogenicity is further mapped to clinical phenotypes that can support the patients' stratification and the prediction of disease progression. The current methodology was applied to four proteins: APP, MAPT, PSEN1 and PSEN2. Three of them are associated with autosomal-dominant AD, amyloid precursor protein (APP) (OMIM 104760), presenilin 1 (PSEN1) (OMIM 104311) and presenilin 2 (PSEN2) (OMIM 600759), while microtubule associated protein tau (MAPT) (OMIM 157140) encodes the tau protein that is aberrantly phosphorylated in AD (Neuner *et al.*, 2020).

Background

Based on the age of onset, AD is divided into two classes: early-onset AD (EOAD) with onset before 65, and late-onset AD (LOAD) (Cuyvers and Sleegers, 2016); EOAD comprises about 5% to 10% of all AD patients and has strong patterns of familial inheritance (Zhu *et al.*, 2015). EOAD has been linked to pathogenic mutations in one of three causative genes: APP, PSEN1, PSEN2. Up to now, over 400 known mutations on these genes have been described, while PSEN1 mutations are responsible for approximately 75% of genotyped families positive for a mutation, whereas APP and PSEN2 mutations account for 13% and 12%, respectively. Aβ peptides result from the cleavage of APP by β- and γ-secretases while PSEN1 and PSEN2 are components of the γ-secretase complex (Haass and De Strooper, 1999).

APP encodes for the amyloid precursor protein, a transmembrane protein whose cleavage forms amyloidogenic Aβ peptides, key components of amyloid plaque. Most APP mutations are missense or nonsense. They are normally localised either within the domain that encodes the Aβ peptide, (amino acids 692–705) (93% of total mutations) or near the cleavage sites of secretases (amino acids 670–682 and 713–724) (Cacace *et al.*, 2016; Dai *et al.*, 2017). The overall effect of APP

Article history
Received: 13 December 2021
Accepted: 15 December 2021
Published: 07 July 2022

Avramouli et al. (2022) EMBnet.journal 27, e1011
http://dx.doi.org/10.14806/ej.27.0.1011

mutations alters the processing by secretases and leads to increased generation and/or aggregation of amyloid, and/or a change in the ratio of specific Aβ peptides.

Presenilin-1 and presenilin-2 proteins are critical subunits of the γ-secretase complex responsible for processing of APP. PSEN2 is about 60% homologous to PSEN1, thus it is possible that they also have overlapping or similar activities. Similar to some mutations in APP, mutations in PSEN1 and PSEN2 typically result either in the overproduction of Aβ or an increased ratio of Aβ42 over Aβ40 (Loy *et al.*, 2014), triggering the formation of amyloid plaques and leading to the development of AD (Sun *et al.*, 2017). Mutations in PSEN1 are the most common cause of EOAD; as of September 2021, over 350 mutations (some of unclear pathogenicity) have been identified (www.Alzforum.org). PSEN1 mutations are estimated to contribute to around 80% of monogenic AD with complete penetrance and early age of onset (Giri *et al.*, 2016). The exact mechanism through which mutations in PSEN1 result in dementia and neurodegeneration in EOAD remains unknown. In addition to their role in γ-secretase activity, PSEN1 mutations may compromise neuronal function, affecting γ secretase activity and kinesin-I-based motility, thus leading to neurodegeneration (Giri *et al.*, 2016). To date, 341 pathogenic mutations have been identified in PSEN1, most of whom are missense, while most of them occur in exons 5, 6, 7, and 8.

PSEN2 mutations are much rarer, with only around 30 mutations identified in EOAD families (Cacase *et al.*, 2016). Mutations in PSEN2 alter the γ-secretase activity and lead to elevation of Aβ42/40 ratio in a similar manner to the PSEN1 mutation. Though PSEN2 is homologous to PSEN1, less amyloid peptide is produced by PSEN2 mutations. In some people with PSEN2 mutations, neuropathological changes appear as neuritic plaque formation and neurofibrillary tangle accumulation (Giri *et al.*, 2016). Furthermore, β-secretase activity is enhanced by PSEN2 mutation, through reactive oxygen species-dependent activation of extracellular signal-regulated kinase (Park *et al.*, 2012). PSEN2 mutations are very rare, and to date 84 pathogenic PSEN2 mutations have been detected worldwide. Moreover, in the pathogenic/likely pathogenic variants, missense variants are more common in PSEN1 than those in PSEN2. In general, most of the pathogenic AD mutations are located in exons 16–17 of the APP, exons 3–12 of PSEN1 and PSEN2 genes (An *et al.*, 2016). The localisation of mutations in AD causing genes leads to the assumption that the above exons are variant hotspots and need to be given priority when performing DNA sequencing (Zhao and Liu, 2017).

MAPT encodes the microtubule associated protein tau, a protein crucial to AD neuropathology. Even though MAPT mutations are not linked to familial forms of AD, SNPs near the MAPT locus are associated with AD risk. Interestingly SNPs in exon 3 act protectively against AD through decreased aggregation of tau protein (Neuner *et al.*, 2020). Up to now 15 no disease-causative mutations have been linked with AD.

Methodology

The implementation methodology for this research work follows the approach of the Automated shape-based clustering of 3D immunoglobulin protein structures that was evaluated in the use case of chronic lymphocytic leukemia (Polychronidou *et al.*, 2018).

The proteins described were used as target proteins due to their important role in Alzheimer's disease progression. The first step of the analysis was to identify the mutations related to the proteins. This information was extracted by the Alzforum - Mutations public database (Alzforum, 2021). This database is a repository of variants in genes linked to Alzheimer's disease (AD). The database includes the three genes associated with autosomal-dominant AD (APP, PSEN1, PSEN2) and two genes associated with AD by way of genetics or the neuropathology of the encoded protein (TREM2 and MAPT). TREM2 was excluded from the analysis as the identification of the primary structure wasn't feasible by the selected sources.

Evaluation of Protein Structures

Since the three-dimensional shape of most of the related proteins is not determined through experimental methodologies, established servers and online databases like Uniprot (UniProt Consortium, 2015), PolyPhen-2 (Adzhubei *et al.*, 2013), iTASSER (Yang *et al.*, 2015) and PDBeFold (Krissinel, 2007) were evaluated for predicting the mutated structures and estimate the impact of the mutations to the 3-dimensional structure. A list of the selected methodologies is presented on the Table 1

These methodologies were used to better understand the structures and determine the structures for the mutated proteins. However, during the study, AlphaFold (AlQuraishi, 2019) was published as the latest state-of-the-art method for the prediction of protein structures. AlphaFold (Pereira, 2021), a neural network-based model, was validated in the challenging 14th Critical Assessment of Protein Structure Prediction (CASP14) and was vastly more accurate than competing methods. The four protein structures were identified in the AlphaFold database and used as the target structures of this analysis.

The following step in the pipeline was to create the structures for the mutations by protein. To achieve that the DynaMut server was used (Rodrigues *et al.*, 2018). DynaMut implements two distinct, well established normal mode approaches, which can be used to analyse and visualise protein dynamics by sampling conformations and assess the impact of mutations on protein dynamics and stability resulting from vibrational entropy changes. DynaMut integrates our graph-based signatures along with normal mode dynamics to generate

a consensus prediction of the impact of a mutation on protein stability.

Through this approach the mutated structures were predicted for the four proteins of interest. Specifically, 25 structures were retrieved by MART mutations, 62 structures by APP mutations, 286 structures by PSEN1 mutations and 68 structures by PSEN2 mutations. For each mutation, the description on Pathogenicity was also extracted from ALZforum and normalised into four categories: (1) Unclear, (2) Benign, (3) Pathogenic, (4) Not classified.

The objective of the study was to classify the resulted structures by Unclear mutations, through an AI/ML approach derived by the protein structures. To further analyse the mutated structures, an established methodology from the field of 3D object recognition was applied. The individual examination and combination of the local descriptors was applied to the 3D structures to extract the appropriate features for the comparison.

Three distance matrices were created by applying the FPFH, RSD and 3DSC descriptors. These matrices were the input of hierarchical clustering. The methodology was selected to supervise the separation of structures into clusters of structures with high similarity.

Indicative examples of hierarchical clustering output are presented here for the APP structure (Figure 1), PSEN2 structure (Figure 2), and the 3DSC descriptor. The optimal number of clusters for each protein-descriptor combination was determined through Silhouette analysis (Figure 3) and the clusters were analyzed based

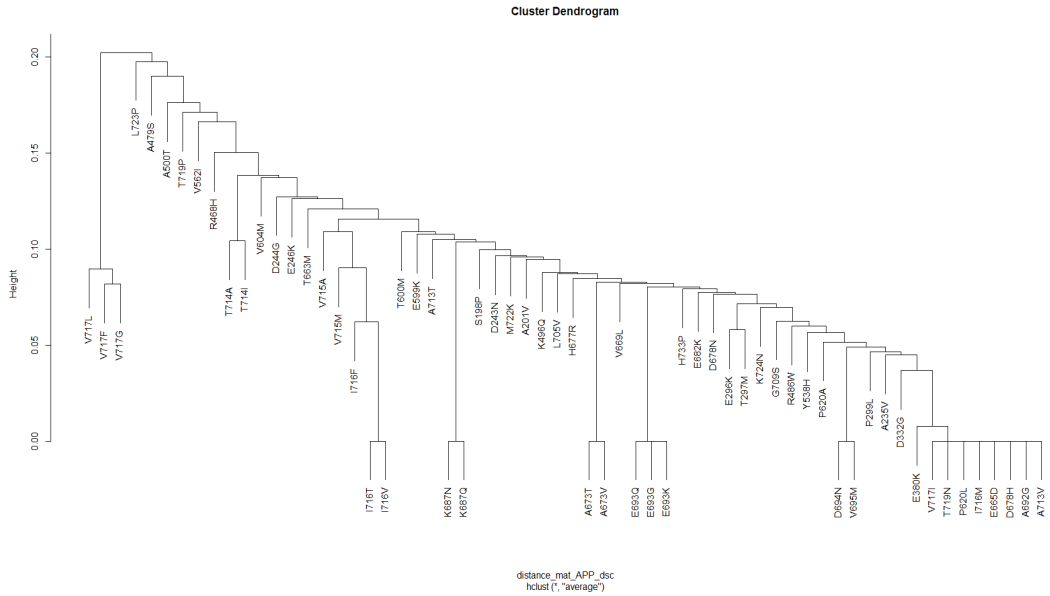


Figure 1. Dendrogram resulted from the hierarchical clustering in APP protein using the 3DSC descriptor.

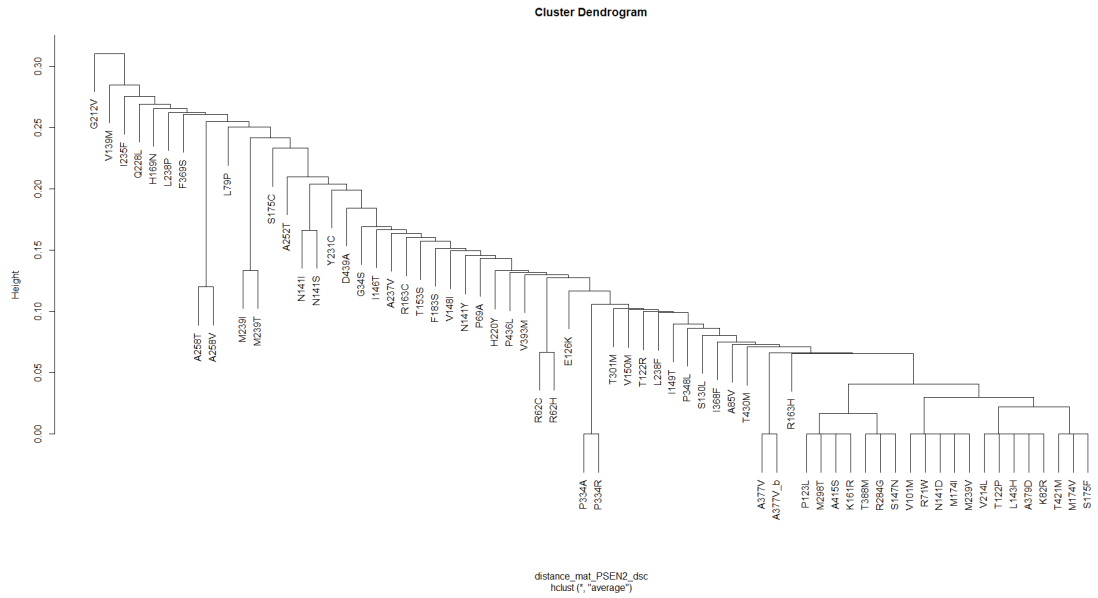


Figure 2. Dendrogram resulted from the hierarchical clustering in PSEN2 protein using the 3DSC descriptor.

Avramouli et al. (2022) EMBnet.journal 27, e1011
http://dx.doi.org/10.14806/ej.27.0.1011

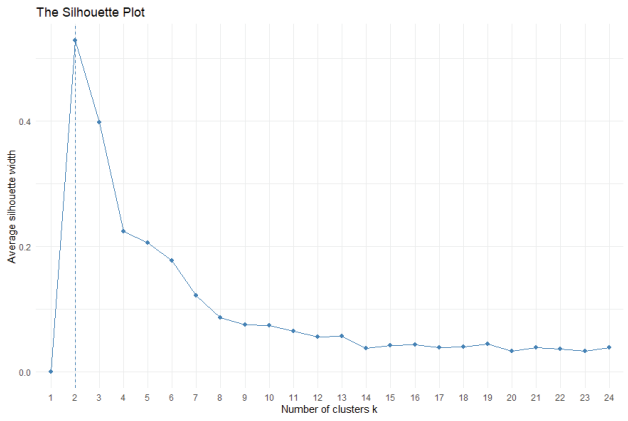


Figure 3. Silhouette analysis on PSEN1 protein using the FPFH descriptor, to determine the optimal number of clusters.

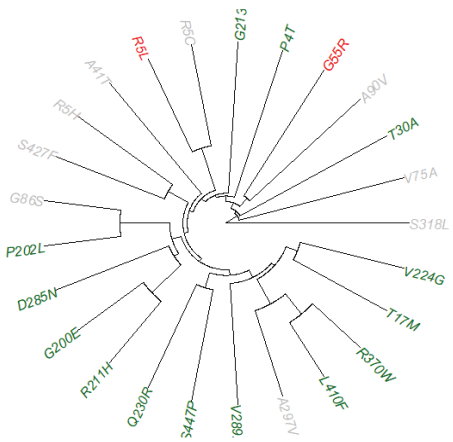


Figure 4. Fan dendrogram of MAPT protein using the 3DSC descriptors.

on their pathogenicity. The results from this analysis classified the PSEN1 mutations into two main clusters, Pathogenic and Non-pathogenic. Based on this analysis, 110 mutated structures originally derived by Unclear pathogenic mutations, were classified as pathogenic.

To further analyse the output, fan dendrograms were produced by also using the colors of the pathogenic types (Figure 4). Through this low-level cluster visualisation, the lowest height of the cluster was identified, and the groups of protein structures were analysed. In detail, six unclear structures were characterised for APP, nine unclear structures were characterised for PSEN2 and four unclear structures were characterised for MAPT.

3DSC descriptor supported the characterisation of the nine unclear or not classified mutations related to PSEN2. On MAPT, two mutations were characterised as Pathogenic (A90V, R5C) by RSD and 3DSC while R5C was classified as Benign from FPFH. R5H was classified as pathogenic by RSD and unclear by FPFH. Finally, A297V was classified as Benign by all methods and G86S as Benign only by 3DSC as the other descriptors didn't reveal any specific cluster. In MAPT case FPFH and RSD didn't perform as 3DSC in the cases of established pathogenicity (unclear cases), thus 3DSC is the descriptor that performed best in this protein. By following the output of the 3DSC descriptor, four new mutations can be characterised.

In the case of APP, FPFH was the descriptor with the highest confidence and through this approach the method characterised six Not Classified mutations. This

Table 1. Results of PSEN2 pathogenicity prediction. The numbers on the descriptors column describes the groups that the proteins were grouped with while the color describes the type of pathogenicity (green = Benign, red=Pathogenic). The new prediction of the not classified or unclear structures is included in the corresponding cell.

Original Protein	Mutation	Pathogenicity	3DSC
PSEN2	A379D	Not Classified -> Pathogenic	4
PSEN2	A415S	Not Classified -> Pathogenic	1
PSEN2	K161R	Not Classified-> Pathogenic	1
PSEN2	K82R	Not Classified-> Pathogenic	4
PSEN2	L143H	Not Classified-> Pathogenic	4
PSEN2	M174I	Not Classified-> Pathogenic	3
PSEN2	M174V	Benign	5
PSEN2	M239V	Likely Pathogenic	3
PSEN2	M298T	Uncertain Significance	1
PSEN2	N141D	Not Classified-> Pathogenic	3
PSEN2	P123L	Likely Pathogenic	1
PSEN2	S175F	Uncertain Significance	5
PSEN2	T122P	Likely Pathogenic	4
PSEN2	T421M	Benign	5
PSEN2	V101M	Unclear Pathogenicity -> Pathogenic	3
PSEN2	V214L	Unclear Pathogenicity-> Pathogenic	4

Table 2. Results of APP pathogenicity prediction

Original Protein	Mutation	Pathogenicity	FPFH
APP	A235V	Likely Benign	5
APP	A692G	Pathogenic	6
APP	E296K	Not Classified -> Benign	5
APP	E380K	Uncertain Significance	7
APP	E665D	Benign	1
APP	G709S	Not Classified	9
APP	H733P	Not Classified -> Pathogenic	6
APP	I716M	Not Classified	9
APP	K496Q	Not Classified-> Benign	5
APP	L705V	Pathogenic	8
APP	M722K	Pathogenic	2
APP	P299L	Not Classified -> Pathogenic	2
APP	P620L	Uncertain Significance	7
APP	R486W	Not Classified-> Pathogenic	3
APP	T297M	Uncertain Significance	8
APP	T719N	Pathogenic	4
APP	V562I	Uncertain Significance	9
APP	V669L	Not Classified-> Benign	1
APP	V717F	Pathogenic	3
APP	V717I	Pathogenic	4
APP	V717L	Pathogenic	8

number corresponds to ~20% of the total not classified APP mutations.

To support the analysis of the methodology, evidence for clinical phenotype, pathogenicity, neuropathology, and biological effect were also taken into consideration. For example, in APP - H733P mutation has not been classified, but the in-silico analysis suggests damaging effect (Guerreiro *et al.*, 2010). This mutated structure was classified as pathogenic by this process as well. Hence, additional evidence beyond the experimental evaluation is generated by our suggested methodology.

Discussion

In summary, the phenotype of APP, PSEN1 and PSEN2 mutation carriers is heterogeneous. Applying pathogenicity prediction methodology to variants of unknown significance, we classified many of them as probably pathogenic. Variants of unknown significance were mainly identified in single individuals' phenotype clinically with AD. Data from families with a monogenic form of AD or patients with a known causative mutation provide the opportunity to identify mutation-specific effects and to correlate genotypic changes with clinical and pathophysiological manifestations of the disease. Asymptomatic carriers of mutations can also serve as candidates for disease-modifying treatment or prevention trials. In the future, different genetic causes of AD should be targeted with specific interventions.

Studies involving mapping pathogenic mutations to tertiary structural domains are required to show the vital relationships between structure and function. Since the amino acid position can, in fact, predict pathogenicity we analysed mutations in AD causative genes and compared these changes to available clinical data. To the best of our knowledge, this is the first study of its kind performing comparative and ab initio prediction of protein structure for mutated APP, PSEN1, PSEN2 and MAPT proteins. In this study we used prediction tools to elucidate how mutations in the causative genes change the tertiary structure of the proteins. We aim in the identification of common structural issues, and in the relation between structure and function through the deleterious effects of the loss of tertiary structure in EOAD causative genes.

Key Points

- There is still on-going research for improving the identification and classification of Alzheimer's disease patients.
- Structural similarity of mutated proteins supports the evidence generation for characterisation of mutations pathogenicity.
- The applied implementation uses three-dimensional descriptors to identify the distance between the structures.
- The methodology was very effective and successfully generated a new dimension in the pathogenicity determination process.

Acknowledgment

This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning 2014-2020» in the context of the project "Analysis of the tertiary protein structure and correlation of mutations with the clinical characteristics of Alzheimer's disease", Project no. 5067210.

References

Adzhubei I, Jordan DM, Sunyaev SR (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Current protocols in human genetics, **76**(1), 7-20. http://dx.doi.org/10.1002/0471142905.hg0720s76

ALZFORUM (2021) Mutations https://www.alzforum.org/mutations Accessed 20 Sep 2021

AlQuraishi M (2019) AlphaFold at CASP13. Bioinformatics, **35**(22), 4862-4865. http://dx.doi.org/10.1093/bioinformatics/btz422

An SS, Park SA, Bagyinszky E, Bae SO, Kim YJ, Im JY *et al.* (2016) A genetic screen of the mutations in the Korean patients with early-onset Alzheimer's disease. Clin Interv Aging **15**, 1817-1822. http://dx.doi.org/10.2147/CIA.S116724

Cacace R, Sleegers K, Van Broeckhoven C (2016) Molecular genetics of early-onset Alzheimer's disease revisited. Alzheimers Dement **12**(6), 733–748. http://dx.doi.org/10.1016/j.jalz.2016.01.012

Cuyvers E and Sleegers K (2016) Genetic variations underlying Alzheimer's disease: evidence from genome-wide association studies and beyond. Lancet Neurol **15**(8), 857–868. http://dx.doi.org/10.1016/S1474-4422(16)00127-7

Dai MH, Zheng H, Zeng LD, Zhang Y (2017) The genes associated with early-onset Alzheimer's disease. Oncotarget **9**(19), 15132–15143. http://dx.doi.org/10.18632/oncotarget.23738

Giri M, Zhang M, Lü Y (2016) Genes associated with Alzheimer's disease: an overview and current status. Clin Interv Aging **11**, 665–681. http://dx.doi.org/10.2147/CIA.S105769

Guerreiro RJ, Baquero M, Blesa R, Boada M, Brás JM *et al.* (2010) Genetic screening of Alzheimer's disease genes in Iberian and African samples yields novel mutations in presenilins and APP. Neurobiol aging **31**(5), 725–731. http://dx.doi.org/10.1016/j.neurobiolaging.2008.06.012

Haass C and De Strooper B (1999) The presenilins in Alzheimer's disease--proteolysis holds the key. Science **286**(5441), 916–919. http://dx.doi.org/10.1126/science.286.5441.916

Krissinel E (2007). On the relationship between sequence and structure similarities in proteomics. Bioinformatics **23**, 717-723. http://dx.doi.org/10.1093/bioinformatics/btm006

Loy CT, Schofield PR, Turner AM, Kwok JB (2014) Genetics of dementia. Lancet **383**(9919), 828–840. http://dx.doi.org/10.1016/S0140-6736(13)60630-3

Neuner SM, Tcw J, Goate AM (2020) Genetic architecture of Alzheimer's disease. Neurobiol Dis **143**, 104976. http://dx.doi.org/10.1016/j.nbd.2020.104976

Park MH, Choi DY, Jin HW, Yoo HS, Han JY *et al.* (2012) Mutant presenilin 2 increases β-secretase activity through reactive oxygen species-dependent activation of extracellular signal-regulated kinase. J Neuropathol Exp Neurol **71**(2), 130-139. http://dx.doi.org/10.1097/NEN.0b013e3182432967

Pereira J, Simpkin AJ, Hartmann MD, Rigden DJ, Keegan RM *et al.* (2021). High-accuracy protein structure prediction in CASP14. Proteins: Structure, Function, and Bioinformatics, **89**(12), 1687-1699. http://dx.doi.org/10.1002/prot.26171

Polychronidou E, Kalamaras I, Agathangelidis A, Sutton LA, Yan XJ *et al.* (2018) Automated shape-based clustering of 3D immunoglobulin protein structures in chronic lymphocytic leukemia. BMC Bioinformatics **19**(14), 67-81. http://dx.doi.org/10.1186/s12859-018-2381-1

Rodrigues CH, Pires DE, Ascher DB (2018) DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res **46**(W1), W350-W355. http://dx.doi.org/10.1093/nar/gky300

Saez-Atienzar S and Masliah E (2020) Cellular senescence and Alzheimer disease: the egg and the chicken scenario. Nat Rev Neurosci **21**(8), 433–444. http://dx.doi.org/10.1038/s41583-020-0325-z

Sun L, Zhou R, Yang G, Shi Y (2017) Analysis of 138 pathogenic mutations in presenilin-1 on the in vitro production of Aβ42 and Aβ40 peptides by γ-secretase Proc Natl Acad Sci U S A **114**(4), E476–E485. http://dx.doi.org/10.1073/pnas.1618657114

UniProt Consortium (2015) UniProt: a hub for protein information. Nucleic acids research **43**(D1), D204-D212. http://dx.doi.org/10.1093/nar/gku989

Yang J, Yan R, Roy A, Xu D, Poisson J *et al.* (2015). The I-TASSER Suite: protein structure and function prediction. Nature methods **12**(1), 7-8. http://dx.doi.org/10.1038/nmeth.3213

Zhao GH and Liu XM (2017) Clinical features and genotype-phenotype correlation analysis in patients with ATL1 variants: A literature reanalysis. Transl Neurodegener **6**:9. http://dx.doi.org/10.1186/s40035-017-0079-3

Zhu XC, Tan L, Wang HF, Jiang T, Cao L *et al.* (2015) Rate of early onset Alzheimer's disease: a systematic review and meta-analysis. Ann Transl Med **3**(3), 38. http://dx.doi.org/10.3978/j.issn.2305-5839.2015.01.19

wrong place

Vivienne Baillie Gerritsen

When you reach a certain age, one question arises on a painfully regular basis. It begins with a "Where are my…?" or a "Where is my…" Reading glasses are a constant. Frequently, they are not where they ought to be. Having relocated them, you may well remark that they are not where you put them. But they are. The thing is, in a moment of distraction, you left them where you would not normally: on the garden wall, in your coat pocket, on the clothes washing machine, perhaps even in the fridge. All in all, they were inadvertently mislocated. On a far smaller scale, the same kind of thing can happen to proteins. There are times when proteins end up where they should not be – which is a source of stress both for their unusual environment and the one they have not reached. Over time, cells have developed various quality control systems to correct all sorts of mistakes – one of them being mislocation. As an illustration, lodged in the endoplasmic reticulum membrane, the enzyme P5A-ATPase is able to spot mislocated transmembrane mitochondrial proteins, grab hold of them and fling them back into the cellular cytosol.

'The Visitor' by Shaun Tan

with permission, www.shauntan.net

Every cell is a metropolis in itself – with motorways running from one end to the other, bodies travelling along them, buildings going up and some coming down, and scaffolds being erected while myriads of entities get on with the business of breaking down, assembling, inhibiting, stimulating, folding, transporting, denaturing, translating, binning, sorting, orchestrating… Proteins constitute, without a doubt, one of the cell's most precious components in that they do most, if not all, of the work. However, every single protein must go to the right place to get on with what it has been programmed to do. There would be little advantage in sending a cobbler to work in a bank. So how do proteins end up in the right place? It all has to do with 'protein signaling' or 'protein targeting'.

For many years, scientists wondered whether the ribosomes sent the newly synthesized proteins to their destinations, or whether the proteins themselves knew the address. It was an issue of great biological importance, which was clarified in the 1970s following studies undertaken by the German-American biologist Günter Blobel. Targeted proteins know themselves where to go thanks to a 15 to 30 amino-acid sequence they carry and which forms what has been dubbed a 'signal peptide' or 'transit peptide'. In essence, a signal/transit peptide is a tag on which is written the protein's destination. It turns out that there are three kinds of tags. The first – and the most frequent – is situated on the N-terminal or the C-terminal end of a nascent protein. The second kind of tag requires additional specifications because the address held within it is not sufficient – there lacks a postal code, or the street number. This occurs by way of protein modifications, such as glycosylation for instance. The third type of peptide signal is formed thanks to an assembly of 'signal patches'. In this case, a number of incomplete, so to speak, signal peptides are scattered within the protein. The address is completed once the protein folds into its 3D conformation and the patches join, as in a puzzle, to complete the tag.

Though the destination they must reach is written on them one way or another, proteins need other molecules – sometimes many others – with whose help they are finally delivered to the correct port of call. Like our own mail, the chances of a protein ending up at the wrong address are bound to occur from time to time, and if there is no kind of quality control, things will gradually go wrong: parts of a cell will begin to stutter, and perhaps even end up failing completely. So how do cells correct mistakes? In a variety of ways; one illustration is the transmembrane helix dislocase P5A-ATPase.

P5A-ATPase is lodged in the endoplasmic reticulum (ER) membrane and belongs to the P-type ATPases family of active transporters whose primordial function is to drive ions or lipids across membranes. P5A-ATPase, however, does not seem to transport ions or lipids, and has a particularly large substrate-binding pocket suggesting that it transports something else. Like other members of the family, the pocket has an opening towards the ER lumen or to the cytosol depending on its dynamic state, but it also has a lateral opening onto the ER lipid bilayer. This suggests that its role may be to transfer molecules located within the actual membrane to the cytosol. Mitochondria and the ER interact closely with each other, which is why mitochondrial proteins may sometimes end up at the wrong address. As ER proteins may too. In particular, P5A-ATPases seem to be able to spot, specifically, tail-anchored transmembrane mitochondrial proteins which have, despite their natural fate, managed to slip into the ER membrane.

Like all P-type ATPases, our dislocase is rather a bulky protein which crosses the ER membrane a dozen times, from which protrudes a characteristic arm-like domain. During ATP-driven ion and lipid transfer, P-type ATPase substrate pockets undergo an important conformational change alternating between a V-shape and a U-shape. Such a change also occurs in our dislocase but the lateral opening is not affected and swings between the cytosol and the ER membrane. This swing is interpreted as the mechanism P5A-ATPase uses to transfer mislocated proteins in the ER's lipid bilayer out into the cytosol. How exactly? By grabbing hold of a short segment of the mislocated protein that dangles in the ER lumen, and literally pulling on it to fling the protein back into the cytosol, much in the way you would sling a ball to the other side of a street by the end of a short rope attached to it. It is a very elegant way of dealing with unwanted material, and it may be that the mislocated protein is actually re-targeted to its correct destination, i.e. the mitochondrion itself.

Cells have a variety of much-needed quality control systems which have developed over time and throughout all kingdoms of life. Take protein synthesis, for instance, which makes use of such systems at every single step to ensure that a gene is correctly transcribed and then translated, and that the product is not only processed the way it should be but also targeted to the right place. A cell's organelles – its nucleus, the mitochondria, the endoplasmic reticulum, the Golgi apparatus to name but four – each require myriads of different proteins to function properly, both within their membranes and in their lumens. It is not difficult to understand, then, that not only must these proteins function properly but that they need to be in the right place. It is a question of organ and cell homeostasis. And, in the end, life itself.

Cross-references to UniProt

Endoplasmic reticulum transmembrane helix translocase, *Homo sapiens* (Human) : Q9HD20
Endoplasmic reticulum transmembrane helix translocase, *Saccharomyces cerevisiae* (Baker's yeast) : P39986

References

1. McKenna M.J., Sim S.I., Ordureau A., Wei L., Harper J.W., Shao S., Park E.
The endoplasmic reticulum P5A-ATPase is a transmembrane helix dislocase
Science 369: DOI: 10.1126/science.abc5809
PMID: 32973005

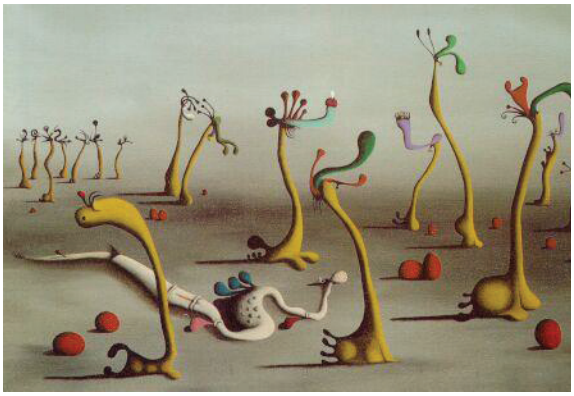
proteinspotlight

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/

versatile

Vivienne Baillie Gerritsen

You cannot beat versatility. Whichever way you look at it, versatility strengthens, opens doors, widens horizons. Many notorious people have been endowed with multiple talents. Author of the hugely popular book on human behaviour "The Naked Ape", Desmond Morris has not only spent a life as a zoologist and a writer but also as a surrealist painter. Dora Maar, in her days, was a well-known photographer as she was a poet and a painter. Le Corbusier, too, gained recognition for his architecture as he did for his furniture design and sculptures. Leonardo da Vinci also comes to mind – as many others do too. Of course, you do not need to be famous to be multi-talented. You do not need to be human, either. You can even be a protein. While many proteins, through the course of their existence, are quite content to have one role, others may be endowed with more. One such protein is transglutaminase 2, whose achievements are so varied that it even ends up being involved in opposing events, such as cell growth and cell death.



Disturbance In The Colony, by Desmond Morris

Courtesy of the artist, writer & zoologist

Transglutaminases were discovered in New York in the 1950s by the neurochemist Heinrich Waelsch who, at the time, was interested in neurotransmitters – namely, glutamates – and how they bind to proteins at the postsynaptic level. Shortly before the structure of DNA had been deciphered and the intimacy of proteins had been understood, Waelsch, and a few fellow researchers, not only described how transglutaminases probably acted on the molecular level but also how they may be involved in neural pathological processes. To sum things up briefly, transglutaminases create bridges, or crosslinks, between proteins via a process known as transamidation, during which glutamine residues (in one protein) are linked to lysine residues (in the other) by way of a particularly solid bond. One transglutaminase, however, is proving to be far more resourceful than its family members: transglutaminase 2, or TG2.

The great majority of transglutaminases are indeed involved in a variety of biological activities that they promote by crosslinking proteins. One example is coagulation factor XIII, a transglutaminase responsible for crosslinking proteins known as fibrins, ultimately leading to the clotting of blood at the site of a wound. Transglutaminase 2 also happily crosslinks proteins via transamidation, but it can also act as a GTPase. Thus armed, TG2 has been shown to be involved in biological processes as diverse as cell adhesion, cell motility, cell signalling as well as the erection of cellular scaffolds – all of which happen to be essential steps in cell growth and development. But also in cell damage and cell death.

As could be expected from an enzyme involved in such a wide span of biological activities, besides being present in almost all types of tissue, TG2 can be either extracellular or intracellular where it is found in the cell cytosol, the nucleus, the mitochondrion and endosomes. The omnipresence of TG2 both in tissues and

various cell compartments could help to explain the multiplicity of its talents – which could be influenced, or brought about, by its surroundings.

Although TG2 multiplicity continues to confound researchers, the discovery that the enzyme takes part in events as opposing as cell growth and cell death is perhaps even more astounding. As an illustration, during apoptosis, extracellular TG2 is known to promote the crosslinking of two proteins – fibronectin and integrin – to form a strong extracellular scaffold which surrounds the dying cell while preventing its insides from leaking out. Another surprising find is TG2's involvement in modifying histones – more precisely histone H3 – and its subsequent influence on gene expression. Histones are proteins which participate in condensing and protecting DNA in the nucleus. When certain genes need to be expressed, the protective histones must release their grip, so to speak, so that gene transcription can occur. TG2 unfastens H3's hold by way of a chemical modification known as serotonylation, where serotonin is popped onto a glutamine residue on H3.

Could TG2's involvement in so many biological activities be explained on a structural level? Perhaps. TG2 exists in two distinct conformations – open and closed – depending on the presence and the concentration of Ca^{2+} and GTP. When Ca^{2+} levels are low, TG2 binds GTP causing the enzyme to fold up onto itself and adopt the closed form. In this conformation, TG2 is unable to bind to its substrate – glutamine – to follow through with transamidation. Despite this, the enzyme is able to perform GTP-dependent functions such as phosphorylation for example. When Ca^{2+} levels increase, under environmental stress for example, the affinity for GTP-binding weakens and TG2 opens up thus presenting its substrate-binding site, ready for transamidation.

TG2 could thus be capable of multiple roles thanks to the uncharacteristically large structural difference between its closed and open forms, its concomitant binding to Ca^{2+} and GTP, and the capacity to remain active in both conformations. Besides this surprising state of events, scientists have also had to come to terms with the almost disturbing notion that TG2 displays opposing effects – like cell death and cell survival – within the same physiological system, so much so that the case of this enzyme has been compared to that of Robert Louis Stevenson's Dr Jekyll and Mr Hyde. Endowed with so many talents and ubiquitous, it is hardly surprising to learn that TG2 is expected to take part in several diseases – among which, as Waelsch had surmised, neurodegenerative diseases like Alzheimer, Parkinson, Huntington and perhaps even multiple sclerosis, besides certain forms of cancer and autoimmune disorders such as celiac disease. Understanding the molecular structure of TG2 in each of its two conformations will help design drugs which, much like the multi-faceted enzyme, are expected to have more than just one pharmacological effect.

Cross-references to UniProt

Protein-glutamine gamma-glutamyltransferase 2, *Homo sapiens* (Human): P21980

References

1. Beninati S., Piacentini M., Bergamini C.M.
Transglutaminase 2, a double face enzyme
Amino Acids 49:415-423(2017)
PMID: 28204961

2. Tatsukawa H., Furutani Y., Hitomi K., Kojima S.
Transglutaminase 2 has opposing roles in the regulation of cellular functions as well as cell growth and death
Cell Death and Disease (2016) / doi: 10.1038/cddis.2016.150
PMID: 27253408

3. Pinkas D.M., Strop P., Brunger A.T., Khosla C.
Transglutaminase 2 undergoes a large conformational change upon activation
PLOS Biology (2007) / doi: 10.1371/journal.pbio.0050327
PMID: 18092889

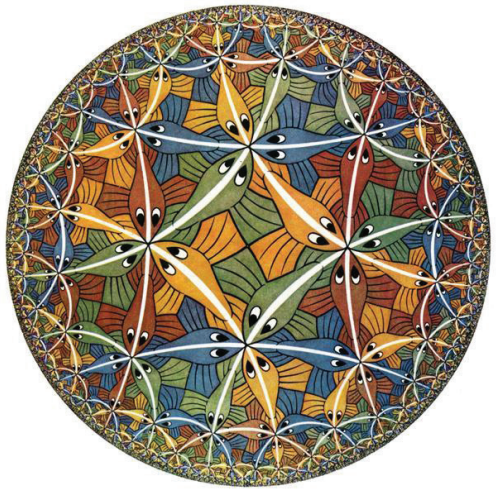
proteinspotlight

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/

a peculiar architecture

Vivienne Baillie Gerritsen

The space humans evolve in is divided into parts. It makes life easier. Each part is dedicated to a certain activity. We have homes to live in, pools to swim in, restaurants to socialise in, trains to travel on, roads to drive along and offices to work in. Cells, too, are divided into parts, and these are known as organelles or compartments. Mitochondria and chloroplasts produce energy, the nucleus transcribes DNA, the endoplasmic reticulum is the seat of protein trafficking, and vacuoles are destined for breaking down cellular waste. Like all cells, bacteria also have their compartments. Carboxysomes are a particularly intriguing example. Why? Because not only is their architecture reminiscent of crystals – sporting layers, straight lines, tips and angles – but their shells are built with proteins. One of the major shell proteins, dubbed CcmK2, assembles into cyclic hexamers which link to one another to form a twenty-facetted polyhedron.



Circle Limit III by M.S. Escher, 1959

Wikiart, Fair Use

Carboxysomes are just one example of bacterial compartments whose shell is made out of protein – as opposed to lipid bilayers in organelles like mitochondria for example. The point of microcompartments is to isolate what is needed to perform certain metabolic pathways. Carboxysomes are where inorganic carbon – CO_2 – is fixed to ultimately produce sugar, and they were first observed in the cyanobacterium *Phormidium uncinatum* in the late 1950s. It took a further twenty years, however, to purify them and understand them in more molecular detail. Inside, as may have been expected, researchers discovered the presence of ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO), the enzyme that fixes CO2 to produce an intermediate molecule (3-phosphoglyceric acid, or 3-PGA) which is subsequently converted to energy (ATP).

Why sequester this particular step of energy production within a microcompartment? Some suggest carboxysomes may have appeared as the level of O_2 rose in the earth's atmosphere. It is a fact that RuBisCO quite happily fixes O_2 – an unfortunate circumstance since this actually costs the bacterium energy by way of a pathway known as photorespiration. If CO_2 and RuBisCO are sequestered somewhere together, then RuBisCO can concentrate on CO_2 and thus provide energy (in the intermediate form of 3-PGA) instead of consuming it. This may well be what actually occurred because most – if not all – of the cyanobacterial RuBisCOs are found in carboxysomes, and very few in the cytosol.

The overall architecture of carboxysomes seems almost out of place in the world of cells. When imagining things biological, the first thought that springs to mind is flow, smoothness and supple contours. The last structure you would think of is a polyhedron, i.e. a three-dimensional form made out of polygonal faces, straight edges and sharp corners. Carboxysomes are in fact icosahedrons, that is to say polyhedrons with no less than twenty faces. When compared to a cell membrane, the shell itself is relatively thin and the inside is packed with

concentric layers of RuBisCO. In between are found other proteins; the most important being carbonic anhydrase that processes bicarbonate to produce CO_2 – RuBisCO's substrate.

As an illustration, the main building block of the carboxysome shell in *Synechococcus elongatus strain PCC 7942* is the structural protein CcmK2. CcmK2 is 102 amino acids long, adopts an alpha/beta fold and assembles into hexamers where one side is concave, and the other convex. The hexamers then bind to one another along their sides, gradually crafting each of the icosahedron's twenty faces. The tips of each face, or triangle, join to form what is known as an icosahedral vertice, and are usually capped by a pentameric shell protein. All in all, about 4,800 proteins are needed to build the carboxysome shell which is roughly about 10,000 times smaller than a pinhead. Besides being involved in constructing the shell, CcmK2 may also play a part in carboxysome intracellular location and distribution during cell division.

On the whole, a tightly-arranged enzymatic core of RuBisCO and, to a lesser degree, carbonic anhydrase forms the inside of carboxysomes. Taken singly, Rubisco happens to be one of the most slothful enzymes in town. However, if thousands of RuBisCOs are gathered in one small space to do exactly the same job – as in an assembly line – then the rate of total production will be heightened. This is exactly what occurs. Furthermore, not only does carbonic anhydrase provide RuBisCO with CO_2 but it also happens to be one of the fastest enzymes! Carboxysomes are therefore like high-speed production lines, able to produce vast amounts of energy not only in record time but in environments that are poor in CO_2. The fact that RuBisCO is encased in carboxysomes also prevents it from fixing O_2, and thus wasting cell energy.

Besides producing energy, molecules of various nature and size have to be able to travel in and out of carboxysomes. Their shells are thought to be sufficiently porous to let small molecules diffuse. But how do larger metabolites such as bicarbonate and 3-PGA enter and exit the shell, respectively? Undoubtedly through pores. Each hexamer forms a pore at its centre. In *Halothiobacillus neapolitanus* for instance, studies have shown that dimers of trimers may further stack at this point, creating central chambers which could let larger metabolites – namely bicarbonate – through, via an 'airlock' type mechanism. The pentamers that cap the icosahedral vertices also form pores. Pores, too, are located at regular intervals along the edges of the icosahedral faces. Could it be, then, that the sheer variety of carboxysome proteins are able to tune shell permeability?

The role of carboxysomes was understood quite early on, but no one realised that an astounding 10-25% of CO_2 is fixed globally by them! This is why scientists are keen to find out whether carboxysomes can be genetically engineered to fix CO_2 within various biotechnological applications. This could prove to be helpful, for example, in the fight against the surplus of CO_2 that human industry continues to fling into the atmosphere. Carboxysomes could also be introduced into plant chloroplasts as part of a CO_2 concentrating mechanism which could help to improve crop yield. Gutted carboxysomes could perhaps be designed, too, to ship cargo other than RuBisCO and co. Carboxysomes may, one day, prove to be a genetic engineer's dream. However, a lot remains to be understood. Not the least the fact that, much in the way choreographers direct their dancers, to build a carboxysome you need to know how to orchestrate the ratio and flow of a large number of self-assembling proteins.

Cross-references to UniProt

Carboxysome shell protein CcmK2, *Synechococcus elongatus (*strain PCC 7942*)* : Q03511

References

1. Yeates T.O., Thompson M.C., Bobik T.A.
The protein shells of bacterial microcompartment organelles
Current Opinion in Structural Biology 21:223-231(2011)
PMID: 21315581

2. Borden J.S., Savage D.F.
New discoveries expand possibilities for carboxysome engineering
Current Opinion in Microbiology 61:58-66(2021)
PMID: 33798818

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/

small sacrifice

Vivienne Baillie Gerritsen

Martyrdom is not particular to humans. It is inherent to microbes too. A cell's answer to something which has gone dramatically wrong can be to self-destruct. It is a common approach to irretrievable damage, which cells frequently use. However, when suicide is chosen to save harm spreading further, the act is akin to self-sacrifice. Take viral infection for instance. When a virus finds its way into our body, our immune system calls up different means to fight it off. As a result, either essential components of the virus are successfully attacked and muted, or infected cells are simply wiped out. Such defence strategies are used across all kingdoms. In fact, living beings have devised astonishingly creative and cunning ways of dealing with infection – the most drastic of which is undoubtedly a form of suicide. The bacterium *Escherichia coli* recently revealed an immune strategy it uses, along with other microbes, which leads to its demise to save infection spreading throughout the colony. The strategy termed CBASS, for cyclic oligonucleotide-based antiphage signaling system, interrupts viral replication while also killing the infected host for good measure. One enzyme is at the very heart of this system, and its name is cyclic GMP-AMP synthase.

Tug of War

Ethel Louise Spowers (1890-1947)

Escherichia coli is one of today's most illustrious research model organisms. It takes its name from the man who first described it in 1885, the German-Austrian paediatrician Theodor Escherich. *E.coli* is usually quite harmless and lives in the gastrointestinal tract of warm-blooded animals – humans included. It is part of the normal microbiota of the gut where it plays several important roles such as warding off pathogenic bacteria or producing vitamin K that is vital for blood coagulation. In fact, it is so vital to the human gut that it takes barely forty hours for *E.coli* to colonize the gastrointestinal tract of a new-born.

A single-cell organism, though without a nucleus, *E.coli* shares metabolic pathways with organisms as distant as ourselves. This is not the reason it became one of the lab's star organisms however – almost seventy years ago. Of equal experimental interest are its size, its relative structural simplicity, the way it multiplies, the speed at which it does it, low pathogenicity, but, perhaps most important of all, its genetics – especially the existence of plasmids in the bacterial cytoplasm – compounded by the microbe's ability to be infected by viruses. Though not unique to bacteria, plasmids are small circular genetic elements that are not part of the bacterial genome but are able to replicate autonomously nevertheless. Ironically, over the past decades, the *E.coli* strain most favoured by labs – *E.coli strain* K12/MG1655 – has been so tamed that it is no longer able to colonize humans.

Short for cyclic oligonucleotide-based antiphage signaling system, CBASS is just one of the immune strategies used by bacteria to check viral infection, by sending out warning signals that trigger off various defence systems. The bacterium will also pay a price, however, since it too will be checked.

In brief, when a virus attacks a bacterium, it injects components that are recognized by the host as 'foreign'. If CBASS is the chosen strategy, the intrusion activates an enzyme – a cyclase – that produces cyclic dinucleotide or trinucleotide molecules, which go on to activate downstream cell-killing effector proteins. As there are different ways of annihilating something, there are, equally, different ways of putting an end to a cell. As such, effector proteins can be destined to destroy the cell's genome, to puncture its membrane or deplete cellular NAD+ levels for example – all actions that end up killing the cell.

In the case of *E.coli* strain TW11681 (which is not closely related to K12), scientists do not know which of the injected viral components are actually spotted – but when they are, the proteins they bind to are activated. In turn, these proteins then bind to cyclic GMP-AMP synthase – otherwise known as cGAS/DncV-like nucleotidyltransferase or CD-NTase – to activate it. This results in the CD-NTase forming phosphodiester bonds between nucleotides to produce cyclic nucleotidic signalling molecules which are then released. These signalling molecules go on to bind to cell-killing effector proteins such as phospholipases that degrade the bacterium's inner membrane or endonucleases that degrade DNA indiscriminately, hence *E.coli*'s DNA too. Both actions kill the bacterium quickly – so swiftly, in fact, that the virus does not have time to finish its replication cycle. As a result, infection is aborted and no progeny is released. The strategy, therefore, is to wipe out the infected cell before the virus has a chance to propagate.

The study of interactions between bacteria and their phages have unveiled myriads of immune strategies. What is more, in a system such as CBASS, scientists expect even more complexity with combinations at every level: viral detection, CD-NTase activation, cyclic nucleotide signal production and the nature of cell-killer effectors. In a way, it is almost surprising that CBASS was first discovered in the genomes of non-model organisms and not the model *E.coli* strain, but this is because of the very nature of model organisms: they are a good representation of their fellow beings but they remain just that. It is simply astonishing how complex an immune strategy can be at the level of one cell only, such as *E.coli*, and it is difficult to imagine the complexity of what occurs when a multicellular organism, such as ours, is infected by a virus.

Cross-references to UniProt

Cyclic GMP-AMP synthase, *Escherichia coli* : P0DTF0

References

1. Whiteley A.T., Eaglesham J.B., de Oliveira Mann C.C. et al.
Bacterial cGAS-ike enzymes synthesize diverse nucleotide signals
Nature 567:194-199(2020)
PMID: 30787435

2. Ye Q., Lau R.K., Mathews I.T. et al.
HORMA domain proteins and a Trip13-like ATPase regulate bacterial cGAS-like enzymes to mediate bacteriophage immunity
Molecular Cell 77:709-722(2020)
PMID: 31932165

3. Lopatina A., Tal N., Sorek R.
Abortive infection : Bacterial suicide as an antiviral immune strategy
Annual Review of Virology 7:371-384(2020)
PMID: 32559405

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