BIOTECHNO 2022

The Fourteenth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies

ISBN: 978-1-61208-971-3

May 22nd –26th, 2022

Venice, Italy

BIOTECHNO 2022 Editors

Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany

Manuela Popescu, IARIA, USA

BIOTECHNO 2022

Foreword

The Fourteenth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies (BIOTECHNO 2022), held between May 22 – 26, 2022, covered these three main areas: bioinformatics, biomedical technologies, and biocomputing.

Bioinformatics deals with the system-level study of complex interactions in biosystems providing a quantitative systemic approach to understand them and appropriate tool support and concepts to model them. Understanding and modeling biosystems requires simulation of biological behaviors and functions. Bioinformatics itself constitutes a vast area of research and specialization, as many classical domains such as databases, modeling, and regular expressions are used to represent, store, retrieve and process a huge volume of knowledge. There are challenging aspects concerning biocomputation technologies, bioinformatics mechanisms dealing with chemoinformatics, bioimaging, and neuroinformatics.

Biotechnology is defined as the industrial use of living organisms or biological techniques developed through basic research. Bio-oriented technologies became very popular in various research topics and industrial market segments. Current human mechanisms seem to offer significant ways for improving theories, algorithms, technologies, products and systems. The focus is driven by fundamentals in approaching and applying biotechnologies in terms of engineering methods, special electronics, and special materials and systems. Borrowing simplicity and performance from the real life, biodevices cover a large spectrum of areas, from sensors, chips, and biometry to computing. One of the chief domains is represented by the biomedical biotechnologies, from instrumentation to monitoring, from simple sensors to integrated systems, including image processing and visualization systems. As the state-of-the-art in all the domains enumerated in the conference topics evolve with high velocity, new biotechnologes and biosystems become available. Their rapid integration in the real life becomes a challenge.

Brain-computing, biocomputing, and computation biology and microbiology represent advanced methodologies and mechanisms in approaching and understanding the challenging behavior of life mechanisms. Using bio-ontologies, biosemantics and special processing concepts, progress was achieved in dealing with genomics, biopharmaceutical and molecular intelligence, in the biology and microbiology domains. The area brings a rich spectrum of informatics paradigms, such as epidemic models, pattern classification, graph theory, or stochastic models, to support special biocomputing applications in biomedical, genetics, molecular and cellular biology and microbiology. While progress is achieved with a high speed, challenges must be overcome for large-scale bio-subsystems, special genomics cases, bio-nanotechnologies, drugs, or microbial propagation and immunity.

We take here the opportunity to warmly thank all the members of the BIOTECHNO 2022 Technical Program Committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to BIOTECHNO 2022.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the BIOTECHNO 2022 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that BIOTECHNO 2022 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the fields of bioinformatics, biocomputational systems and biotechnologies.

We are convinced that the participants found the event useful and communications very open. We also hope that Venice provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city

BIOTECHNO 2022 Chairs:

BIOTECHNO 2022 Steering Committee
Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany
Hesham H. Ali, University of Nebraska at Omaha, USA

BIOTECHNO 2022 Publicity Chairs
Hannah Russell, Universitat Politècnica de València (UPV), Spain
Mar Parra, Universitat Politecnica de Valencia, Spain

BIOTECHNO 2022

Committee

BIOTECHNO 2022 Steering Committee

Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany
Hesham H. Ali, University of Nebraska at Omaha, USA

BIOTECHNO 2022 Publicity Chairs

Hannah Russell, Universitat Politècnica de València (UPV), Spain
Mar Parra, Universitat Politecnica de Valencia, Spain

BIOTECHNO 2022 Technical Program Committee

Behrooz Abbaszadeh, University of Ottawa, Canada
Antonino Abbruzzo, University of Palermo, Italy
A M Abirami, Thiagarajar College of Engineering, Madurai, India
Don Adjeroh, West Virginia University, USA
Aftab Ahmad, City University of New York, USA
Jens Allmer, Hochschule Ruhr West - University of Applied Sciences, Germany
Joel P. Arrais, University of Coimbra, Portugal
Simone Avesani, Università di Verona, Italy
Yoseph Bar-Cohen, Electroactive Technologies / NDEAA Lab - Jet Propulsion Laboratory (JPL), USA
Kais Belwafi, King Saud University, Saudi Arabia
Boubaker Ben Ali, University of Bordeaux, France / University of Manouba, Tunisia
Razvan Bocu, Transilvania University of Brasov, Romania
Vincenzo Bonnici, Università di Parma, Italy
Paolo Cazzaniga, University of Bergamo, Italy
Matthias Chung, Virginia Tech, USA
Peter Clote, Boston College, USA
Giovanni Cugliari, University of Turin, Italy
Adithi Deborah Chakravarthy, University of Nebraska at Omaha, USA
Maria Evelina Fantacci, University of Pisa, Italy
Abdolhossein Fathi, Razi University, Iran
Rosalba Giugno, University of Verona, Italy
Henry Griffith, University of Texas at San Antonio, USA
Chih-Cheng Hung, Kennesaw State University, USA
Asier Ibeas, Universitat Autònoma de Barcelona, Spain
Megha Khosla, L3S Research Center | Leibniz University, Hannover, Germany
Jan Kubicek, VSB - Technical University of Ostrava, Czech Republic
Antonio LaTorre, Universidad Politécnica de Madrid, Spain
Chen Li, Monash University, Australia
Yiheng Liang, Bridgewater State University, USA
Tatjana Lončar-Turukalo, University of Novi Sad, Serbia

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Customized Adaptative Neuro-Fuzzy Approach to pH Control on a Stirred Tank Bioreactor 1
Fernando Agustin Hernandez Gobertti, Carlos Heber Cigliutti Barilari, and Andre Luiz Fonseca De Oliveira

Simulating Solar Irradiance through AM Wave Equations for Metabolic Pathways 8
Rosa Maria Esquinas-Ariza, Javier Jimenez-Ruescas, Beatriz Diaz, Hermenegilda Macia, and Edelmira Valero

Heuristic Random Designs for Exact Identification of Positives Using Single Round Non-adaptive Group Testing and Compressed Sensing 10
Catherine Haddad-Zaknoon

A New Phased Array Magnetic Resonance Imaging Coil for Hbo2 Studies 17
Azma Mareyam, Erik Shank, Laurance Wald, Michael Qin, and Giorgio Bonmassar

Simpati: Network-based System for Patients' Classification Reveals Disease Specific Pathways Driven by Cohesive Communities 20
Luca Giudice, Claudia Mengoni, and Rosalba Giugno

Customized Adaptative Neuro-Fuzzy Approach to pH Control on a Stirred Tank Bioreactor

Fernando Agustin Hernandez Gobertti
Facultad de Ingenieria
Universidad ORT Uruguay
Montevideo, Uruguay
email: fhernandez@ieee.org

Carlos Heber Cigliutti Barilari
Facultad de Ingeniería
Universidad ORT Uruguay
Montevideo, Uruguay
email: cgcigliutti@gmail.com

Andre Luiz Fonseca De Oliveira
Facultad de Ingenieria
Universidad ORT Uruguay
Montevideo, Uruguay
email: fonseca@fi365.ort.edu.uy

***Abstract*—Bioreactors are complex sets of tubes, sensors and actuators embodied in recipients of different shapes and sizes, used thoroughly in biotechnical and chemical investigative and commercial environments for hours on end. Stirred tank bioreactors are widely used and available globally, whose design is intentioned for in-batch and continuous operations. They serve as closed controlled systems for specific organic compounds reaction examination when treated with agitation changes and temperature shifts, as well as oxygen saturation and viscosity variation for both aerobic and anaerobic processes. In particular, the rapidness and magnitude at which a substance changes its associated pH affect its solubility and molecular structure, possibly reaching its denaturalization. Hence, analyzing and acting upon these systems pH, where several parameters interact with each other, is crucial for avoiding compound stressing and arriving to the desired products in addition to coherent investigation conclusions. The pH level management, commonly done manually by scientists, serves the purpose for applying fuzzy logic principles where historical data, as well as human expertise and experience, can be best utilized in designing the controller sets of inference rules and membership functions. Thus, this paper focuses on experiments design and proven tuned applications with limited microorganisms capacities of adaptative neuro-fuzzy process trained with custom genetic algorithms for automatic pH control in 5 litres stirred tank bioreactors, joined with practical comparisons between other control engineering formulations. While earlier related research focused on simulated reactions and theoretical control, this empirical procedure gives promising results on 10 seconds cycles of sensing and actuating, achieving an average 0.1 pH error margin on stability when utilized on an agitation and temperature controlled environment. This study provides scientists with an extendable and configurable procedure so to successfully and efficiently control pH on closed systems using an affordable master-slave micro-controllers architecture.**

***Keywords*—fuzzy logic, automatic control, pH controller, bioreactor design, optimization engineering**

I. Introduction

Bioreactors are meticulously designed and carefully manipulated closed systems, crucial for biotechnological and chemical procedures in both academical and commercial contexts [1]. Its research usages usually aim to analyze biochemical active organic compounds behaviour and response when affected by different cycles of natural conditions variations such as temperature [2], oxygen saturation [3] and viscosity [4]. Its main market-oriented utilization ranges from large scale production of consumables or custom alcoholic beverages and milk processing [5] to more contained and closely examined small scale on vaccines fabrication and proteins synthesis [6]. These systems standard capacity vary between contents of approximately 5 litres for investigation purposes and 200 litres for mass production [7]. This paper proposes and compares approaches to bioreactor pH control guided by classical and fuzzy logic, which are general and independent of the systems dimensions, but dependent on its substances concentrations.

Controlling the pH of compounds formed by microorganisms in a liquid medium is paramount for the proper study and analysis of the biomolecular processes that occur, as well as for the correct and expected nutrients development and biological functions availability [8]. Nonetheless, given the strict system requirements for accurate usage, this mandatory regulation mechanism needs to act in conjunction with other relevant controlled properties (e.g., external vest temperature, agitation rotor speed) whose variance produce changes on pH and vice-versa [9]. This intrinsic unavoidable feedback joined by the microorganisms behavior unpredictability demands generally complex solutions through classical means. Using fuzzy logic, however, trained human experience and deductive thinking can be emulated on a robust, reliable and efficient controller [10].

The presented approaches are tested with multiple compounds on bioreactors with working regular agitation speeds from 120 to 220 rpm for molecular oxygenation and temperature variation between 18 and 42 Celsius grads. Further sterilization ranges, i.e., microorganisms cleansing, are not considered for pH control.

Figure 1 shows the used system architecture and communication sequencing per cycle regarding devices related to pH control. It corresponds to a centralized master-slave structure where the master has the control logic of all regulatory properties of the system and the slave communicates and mandates over the peripherals sensor and actuators, consisting on one peristaltic pump for each acid and base drops.

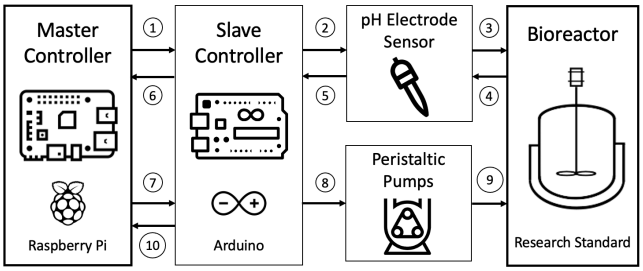


Figure 1. Proposed architecture and sequencing of control system.

 ISBN: 978-1-61208-971-3

This paper proceeds with a brief review of related bioreactor control literature followed by considerations of relevant preliminary experiments necessary for particular system understanding. Then, it focuses on proposed classical, fuzzy and neuro-fuzzy perspectives on pH control in order to later compare its testing results and applications.

II. Literature Review

Bioreactor design's research and analysis is a discipline whose beginnings date back several decades while being continuously encompassed and updated with evolving practices and techniques in its search for generality, scalability and efficiency [11]. Indeed, recent viral outbreaks not only promoted the technical relevance of the biotechnological field but also demanded sophisticated tuned precision and fault tolerant equipment to fulfill the increasing vaccines requirements [12].

Currently, the majority of bioreactor control observations consists on specification and hierarchical delineation of interested magnitudes sets whose regulation may provide greater dominion over part of the system processing. These properties are then subjected to traditional ON-OFF or PID controls (usual for agitation and temperature procedures) [13], as well as neural networks on supervised reinforcement learning algorithms [14]. Albeit useful in practice, these perspectives stand as too complex to the general public, with little availability for further customization.

Contemporary remarks on bioreactor pH control remain traditional variations of classical loop stresses on particular fermentation compounds using stepwise aggregation procedures [15] and/or CO_2 sparge feedback commands [16]. Although tested optimal within their specific environment and processes, these studies depict certain lack of generality and flexibility for more broaden and global scenarios.

Recent related researches on fuzzy control of bioreactors describe specialized designs of a predictive model and fuzzy supervisory controllers for anaerobic processes [17], as well as adaptive PI controller with fuzzy-based parameter selection for fed-batched procedures [18]. Although these methodologies are innovative, they are focused on simulated reactions of scarce microorganisms types with theoretical specifications.

III. Preliminary Experiments

This section focuses on initial experimentation with selected peripherals and associated operational fault managements, necessary in every control design. The proposed procedure also requires a historical analysis and abstraction of the system's time evolution over iterated pH homogeneous variation.

A. Peripherals Examination

The utilized glass two-electrode sensor measures pH with a precision of 0.002 in the entire range 0-14 at usual aforementioned conditions, with reference electrode fixed at 6.86 [19]. It communicates with slave controller using a communication module through I^2C data protocol, industrially used between integrated circuits on environments with little signal interference [20]. This enables additional automatic functionalities of connectivity check and calibration verification.

The low-pressure electrical peristaltic pumps are integrated with DC engines and generates drops of tested approximate 69μL through silicon tubes of 1 cm diameter [21]. These are activated by slave controller using dual h-bridge motor drivers and Pulse Width Modulation (PWM) as shown in Figure 2.

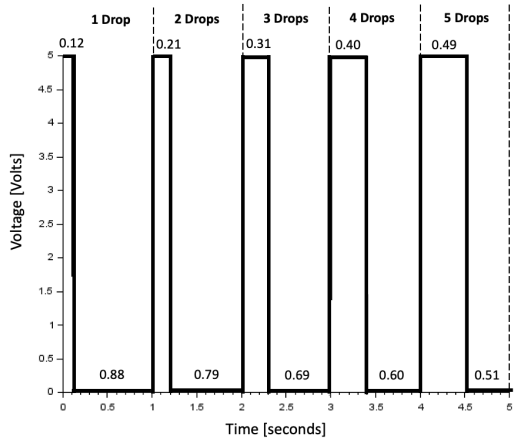


Figure 2. PWM signaling to peristaltic pumps for 1 to 5 drops expulsion.

Thus, sensors and actuators are commanded using affordable Arduino slave controllers, which are connected to a Raspberry Pi master controller where the procedural control decisions are defined. Master and slaves controllers communicate through Modbus ASCII serial protocol [22], which redundantly corroborates message information on both ends. This system design enables the usage of multiple sensors and actuators with a correspondent latency increase.

B. Operational Considerations

Automatic sequential systems functioning over long periods of time demand in practice to define actions consequence of possible defects and events. Due to its overarching simplicity, procedures based on GEMMA framework [23] are structured for pH control, segmenting strategies as functioning, stop and failing processes shown in Figure 3.

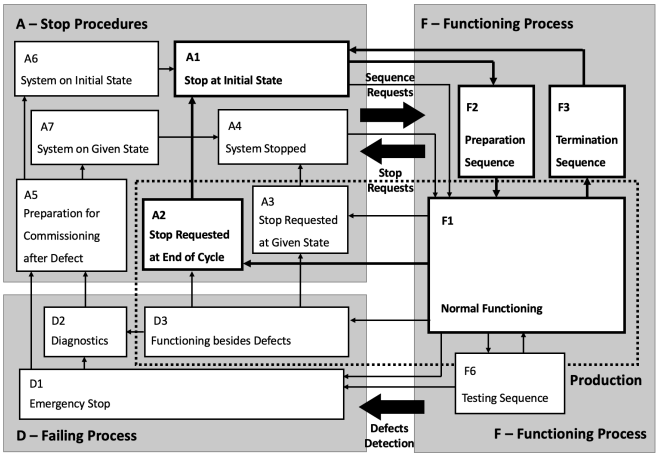


Figure 3. Depiction of automation processes based on GEMMA framework.

Predominant faults are classified as measurement defects, communication errors and actuators malfunction, all not mutually exclusive. Measurement defects are non blocker faults result of incorrect sensing that require message repetition. Communication errors are blocker faults consequence of bad frame reading or writing on a given endpoint, which obligate system stop and physical connector checking. Actuators malfunction provoke unexpected expulsions of acid or base drops, or none entirely.

Bidirectional communication between master and slave controllers, as well as with sensors, provides with acknowledgment notification possibility after each action, enabling repetitions or cancellations. However, communication with actuators remains unidirectional as there are no guarantees of consistent requested drops amount. This is considered on control logic, as incoherent system interaction generates corrective maneuvers on subsequent iterations.

C. System Analysis

The actuators interact with the bioreactor by the expulsion of standard acid HCl with pH 1.8 and base NaOH with pH 11.6, both on 70% concentrated solutions. When analyzing regular pH increments and decrements on organic compounds with constant homogenization, an immediate effect is identified on the pH measurement. Moreover, no apparent inertia is observed on the response as is illustrated in Figure 4, even when using multiple continuous drops. These factors cause that the usual working range is limited by the pH of acid and base expelled by actuators.

Considering these observations and given the need of prolonged continuous system functioning of days at a time, as well as the limited quantity of actuator solutions and engines life expectancy before replacement, extra requirements such as the usage of minimal drops with loose intervals over time are imposed.

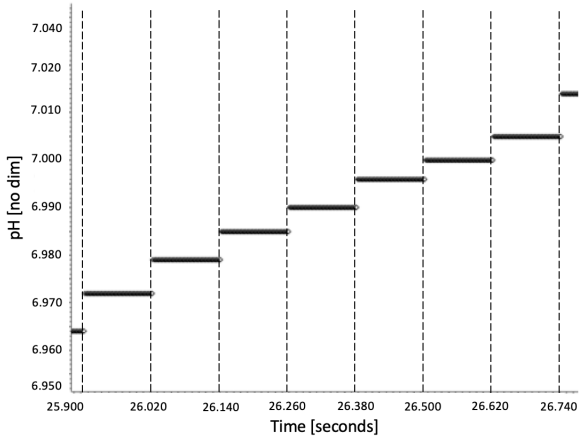


Figure 4. Incremental variation on pH result of periodic drop expulsions.

The addition of a buffered solution as an intent to maintain the pH level at a certain value provokes the natural differentiation of the system response into zones, distinguishable for the steepness of the pH variation. Figure 5 illustrates this,

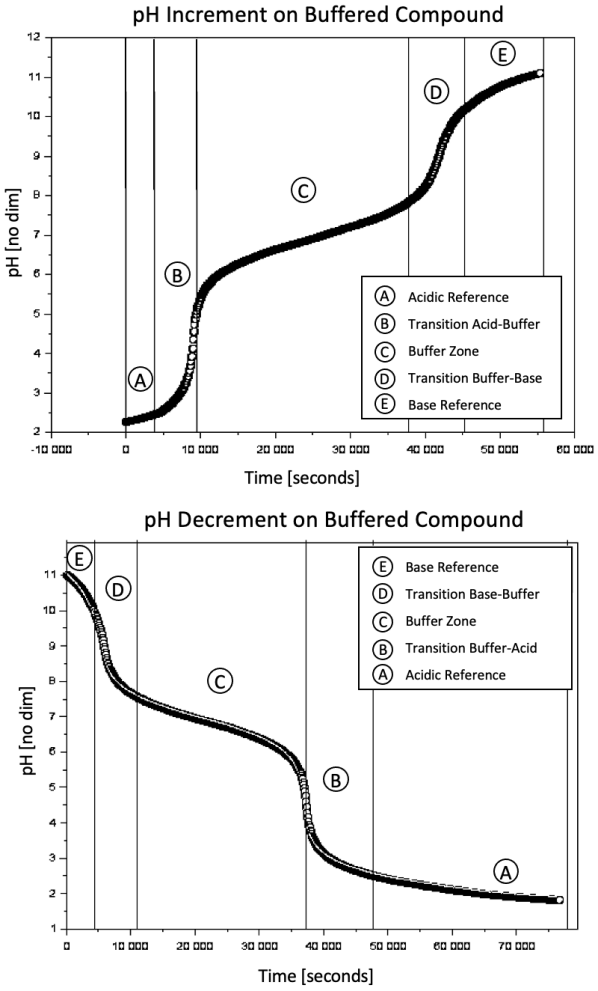


Figure 5. Variations of pH using three drops in a buffered KH_2PO_4 solution of *Streptococcus Thermophilus* in lacteus medium, with zones distinctions.

where five distinct zones are defined and named based on the nearness to buffer control or border limits and transitions between each state, for separate increments and decrements. These temporal behaviour is performed for traditional buffers KH_2PO_4, $C_2H_3NaO_2$, $C_2H_4O_2$ and $Ca_3(BO_3)_2$ on sets of 1, 3 and 5 continuous drops, resetting the system setup for each iteration. In order to avoid excessive compound waste, this analysis can be miniaturized on smaller recipients with same conditions and adequate substances concentration.

Several limitations are also required for consideration. First, bioreactors are usually used nearing maximum capacity with relevant compounds, and exceeded inclusions of actuator solutions are undesirable. Second, the system is needed for prolonged continuous system functioning of multiple days and it must endure limited quantity of actuator solutions and engines life expectancy before replacement. These imposes an extra requirement of minimal drops usage with loose intervals over time, granting a 6 to 10 seconds idle cycles in between consecutive control measurements and conditional expulsions of acid or base.

 ISBN: 978-1-61208-971-3

IV. Control Observations

Previous experiences compendium and its analysis enables the structuring of custom classical, fuzzy and neuro-fuzzy controllers. These logic seek simplification and customization that fulfills requirements of counteracting biochemical reactions due to pH variations lesser than 0.2 per 10 seconds cycles.

A. Classical Controller

Given the observed non-inertial and immediate characteristics of pH permutations, a ON-OFF controller with decisions based on pH sensibilities and an error deadband for both increments and decrements of 0.05 from objective is used.

Figure 6 illustrates the general ON-OFF sequential logic, having as output the acid or base drops quantity to expel on system in the current control cycle.

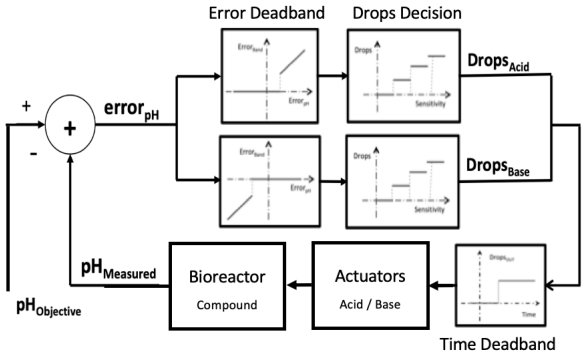


Figure 6. Diagram of utilized sensibility-based ON-OFF control logic.

The general idea, albeit compound specific, requires for significant drops expulsion on buffer zone (e.g., 5 drops for 0.02 variation) and considerably less on transition and reference zones (e.g., 1 and 3 drops respectively for same variation).

B. Fuzzy Perspective

An alternate broader approach is using fuzzy logic controllers, looking for a soft system response given ambiguous inputs and avoiding non-trivial mathematical modelling.

Figure 7 shows the global fuzzy control sequencing loop with the same drops amount parameter as decision output.

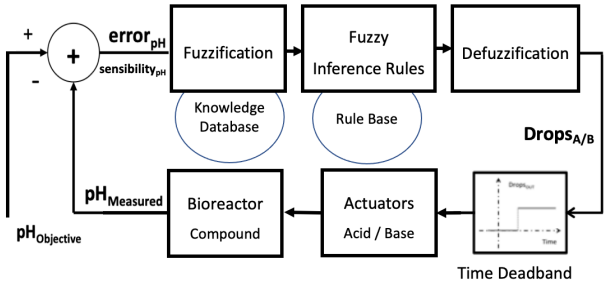


Figure 7. Diagram of proposed fuzzy-based control logic within the system.

In this case, input functions refer to error and sensibility, while output function allude to drops quantity expulsion for acid and base. Figure 8 exhibit a combination of triangular and trapezoidal membership functions that characterize fuzzy sets elements of both inputs expressed by linguistic variables result of a support fuzzification process, defining its universe of discourse. Through this method, a mapping of the crisp input values to the defined membership functions and truth values is performed. Then, these variables are used among max-min inference rules resulting in output linguistic variables, which conclude on the drops quantity after a defuzzification process guided by discrete centroid method, thus favoring the rule with the output of greatest area. In centroid defuzzification the truth values result of each rule are OR'd, i.e., the maximum value is used and the results are then combined using a centroid calculation.

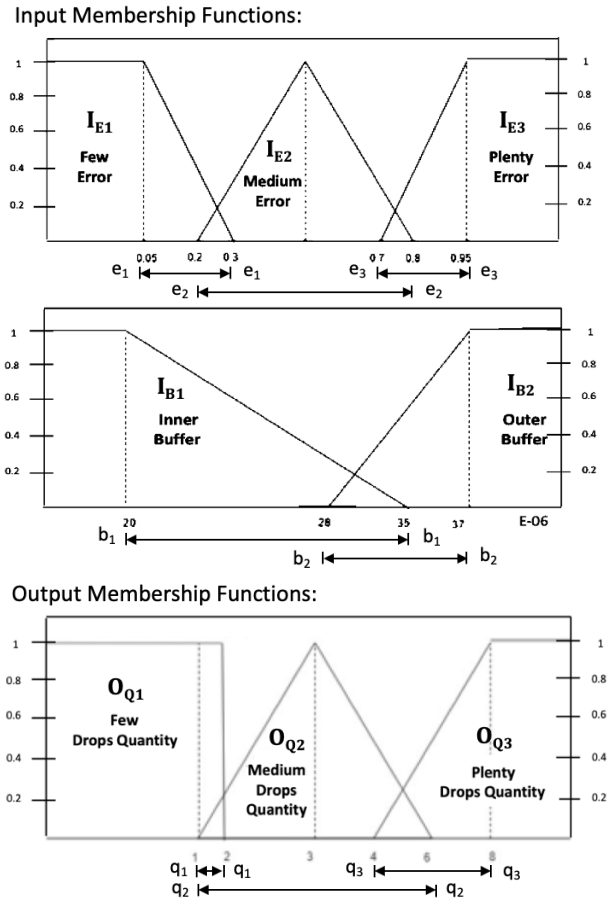


Figure 8. Fuzzy knowledge database, with membership functions definitions and linguistic variables indications.

The conditional rule base is described as follows:

- IF$[(I_E = I_{E1} \wedge I_B = I_{B2})] \Rightarrow O_{Q1} = O_{Q1}^{acid} \vee O_{Q1}^{base}$
- IF$[(I_E = I_{E1} \wedge I_B = I_{B1})] \Rightarrow O_{Q2} = O_{Q2}^{acid} \vee O_{Q2}^{base}$
- IF$[(I_E = I_{E2} \wedge I_B = I_{B2})] \Rightarrow O_{Q1} = O_{Q1}^{acid} \vee O_{Q1}^{base}$
- IF$[(I_E = I_{E2} \wedge I_B = I_{B1})] \Rightarrow O_{Q2} = O_{Q2}^{acid} \vee O_{Q2}^{base}$
- IF$[(I_E = I_{E3} \wedge I_B = I_{B2})] \Rightarrow O_{Q2} = O_{Q2}^{acid} \vee O_{Q2}^{base}$
- IF$[(I_E = I_{E3} \wedge I_B = I_{B1})] \Rightarrow O_{Q3} = O_{Q3}^{acid} \vee O_{Q3}^{base}$

 ISBN: 978-1-61208-971-3

The selection of acid or base drops expulsion is decided implicitly by the rules based on the error differential sign.

When processing these inference rules using max-min, if an AND relationship is specified, then their minimum value is used as the combined truth value, occurring analogously with OR relationships and their maximum value.

Given the carefulness needed for these systems variable conditions, traditional common practices for pH regulation are the manual addition of acidic and alkaline solutions. Thus, practical human experience is mandatory for the definition of mentioned placement and usual ranges of membership functions, as well as for distinguishing each linguistic variable and truth values. In consequence, the proposed solution is diagrammed by empirical methods using a trial-and-error approach on *Streptococcus Thermophilus*, *Escherichia Coli*, *Myxococcus Xanthus* and *Deinococcus Radiodurans* while testing grade fuzzification methods together with weighted-average and mean-max defuzzification processes.

C. Neuro-Fuzzy Approach

Another alternative consists in joining fuzzy logic with customizable learning methodologies, thus providing with adaptative responses over natural changes on system behaviour. In this case, again, input functions refer to error and sensibility, while output function allude to drops quantity expulsion for acid and base. This perspective, diagrammed in Figure 9, enables constant feedback between the fuzzy neural network and the genetic algorithm, which selects the amplitude sets $\{e_1, e_2, e_3\}$, $\{b_1, b_2\}$ and $\{q_1, q_2, q_3\}$ of the predefined membership functions based on historical pH variations with given buffer. Consequent permutations over neural formulations adjusts dynamic responses over persistent system modifications.

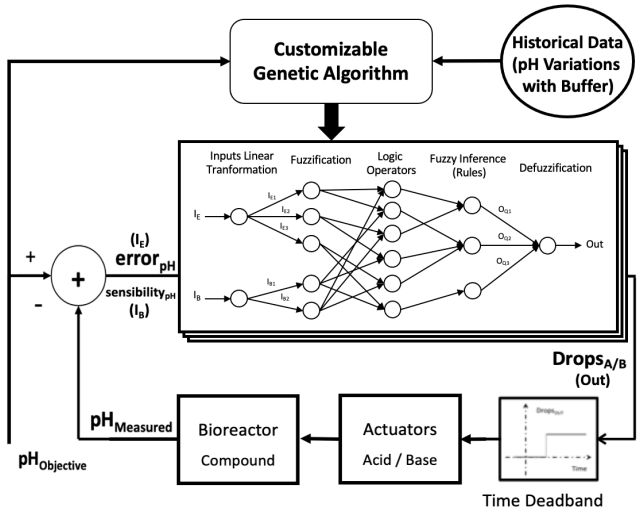


Figure 9. Neuro-fuzzy flow representation, including neural network iterations trained using a genetic algorithm helped by historical data.

Selected Mamdani-based feed-forward neural network is a 5-layer sequence with two inputs and one output that resembles aforementioned traditional fuzzy flow. Transitions from layers 1 to 2, as well as layers 4 to 5, contain [0,1] weights with equal average values in order to ensure symmetrical answer distributions on error input (I_E) and drops output (Out).

A genetic algorithm is used for training and selection of adequate amplitude sets and neural network configuration on each t cycle iteration. Defined selection rules dynamically choose past contiguous iteration's sets as parents of future generations. Custom crossover rules linearly combine these parents using proportional [0,1] parameters $\{\alpha, \beta, \gamma\}$ based on historical data, as shown in below equations.

$$\begin{pmatrix} e_1^t & e_2^t & e_3^t \end{pmatrix} = \begin{pmatrix} \alpha_I & \alpha_{II} & \alpha_{III} \end{pmatrix} \begin{pmatrix} e_1^{t-1} & e_2^{t-1} & e_3^{t-1} \\ e_1^{t-2} & e_2^{t-2} & e_3^{t-2} \\ e_1^{t-3} & e_2^{t-3} & e_3^{t-3} \end{pmatrix}$$

$$\begin{pmatrix} b_1^t & b_2^t \end{pmatrix} = \begin{pmatrix} \beta_I & \beta_{II} \end{pmatrix} \begin{pmatrix} b_1^{t-1} & b_2^{t-1} & b_3^{t-1} \\ b_1^{t-2} & b_2^{t-2} & b_3^{t-2} \end{pmatrix}$$

$$\begin{pmatrix} q_1^t & q_2^t & q_3^t \end{pmatrix} = \begin{pmatrix} \gamma_I & \gamma_{II} & \gamma_{III} \end{pmatrix} \begin{pmatrix} q_1^{t-1} & q_2^{t-1} & q_3^{t-1} \\ q_1^{t-2} & q_2^{t-2} & q_3^{t-2} \\ q_1^{t-3} & q_2^{t-3} & q_3^{t-3} \end{pmatrix}$$

Customization attributes enable constant or evolutionary proportional parameters indication, as well as fixed or variable parents selection. Particularly, setting constant proportional parameters provide equal pondering on membership variations per cycle, while using evolutionary variations generates dynamic functions for reaching certain behaviour at a given point in time. Moreover, configuring fixed parenting promotes constant and stable considerations of parenthood relationships, while selecting variable parenting allows for suppressing or emphasizing set behaviours caused by expected disturbances. Both evolutionary proportional parameters and variable parenting involves preliminary optimization steps with specific distributions, which resolve primarily on fewer or lower buffer usage and consequent error in regime. Possible combinations of aforementioned approaches broaden the system's response and behaviour for a given experiment context, which might deliver further research and production possibilities for in-batch microorganisms growth. Furthermore, initial iterations are defined mirroring aforementioned traditional fuzzy perspective, thus aiming at overcome natural system hysteresis and early reactions. Seeking simplification, no specific mutation rules are currently determined or deemed necessary.

D. Results Comparison

Generalizing outcomes are complex for systems with non-identical repeatable experiences, even more when considering different combinations of input parameters values and process cycles through ever-changing environmental conditions. However, certain particularities can be observed for most use cases that enable objective control results distinctions.

Figure 10 shows examples of the system evolution with active pH controls that illustrates the comparable similarities between all the examined approaches.

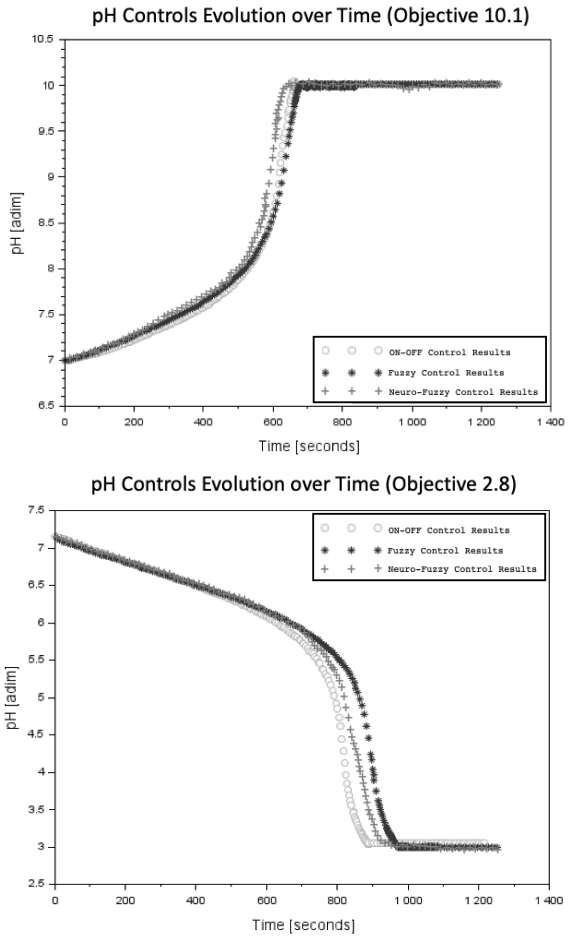


Figure 10. System response to pH controls at different objectives setpoints on *Streptococcus Thermophilus* solution with buffer KH_2PO_4.

According to these tuning and verification experiments results, the ON-OFF, fuzzy and neuro-fuzzy controllers stimulate the system to successfully reach pH levels with less than required 0.1 error margin without considerable overdraft nor oscillations when stationary for both increments and decrements. In fact, the fuzzy control gets to higher precision results than the classical control, i.e., closer to pH objective at regime, at a similar variation speed but at a greater transition time. This differentiation can be clearly appreciated on decrements and when a more extended pH variation is needed. Indeed, the fuzzy and neuro-fuzzy controllers are empirically more robust to noisy data caused by sensor malfunctions or circumstantial system behavioral spikes and levels reactions variations in setpoint vicinity in lesser approximation cycles quantities.

The neuro-fuzzy approach applied here defines balanced and constant proportional parameters while assuming fixed parents selection of three previous cycle iterations. As shown in Figure 10, it provides with sharper transition periods on both pH increment and decrement compared to traditional fuzzy perspective, with softer albeit slower adaptability changes and smaller divergences from objective. In particular, more accurate results with faster transitions can be obtained when setting evolutionary parameters along with same fixed parents selection, or variable parenting with extended preceding cycles considerations for further adaptability possibilities.

V. Conclusion

In this paper, three different approaches to pH control in bioreactors were proposed and its results compared through tuning and application of compounds with *Streptococcus Thermophilus*, *Escherichia Coli*, *Myxococcus Xanthus* and *Deinococcus Radiodurans* with standard actuators NaOH and HCl together with usual buffers KH_2PO_4, $C_2H_3NaO_2$, $C_2H_4O_2$ and $Ca_3(BO_3)_2$. Motivated by the natural relevance of pH property on the growth and survival of different microorganisms, and sought of general customized and flexible procedures for its control, all perspectives achieved an acceptable functioning with variable precision within the system characteristics and set requirements using affordable and scalable devices. Improvements related to diminishing transition times and increasing selection of membership classes can be further pursued for more meticulous or precise control and expanding current action ranges with additional limited drops quantities.

While the classical ON-OFF controller presented a more standard and direct logic sequence, the fuzzy and neuro-fuzzy propositions aimed at a more generic, customized and adaptable scheme to uncertain biochemical reaction changes with nonlinear behaviour. Due to its successful empirical testing and customization capabilities, the neuro-fuzzy approach is recommended to use on standard stirred tank bioreactors, with possible further investigation related to variants on other systems (e.g., other bioreactor types), as well as studied influence of different actuators concentrations and biological compounds characteristics (e.g., distinct buffers solutions and microorganisms combinations).

Potential real world use cases of this procedure involve commercial consumable fermentation and composition (e.g., milk derivatives preparation), as well as vaccine components concoction and manufacturing (e.g., antivirus processing for different animals) for small and large scale aerobic or anaerobic production. This is justified by the main dependence of the proposed procedure on compounds concentrations and independence of the system dimensions or capacities. Also, other academic use cases consist on studying certain microorganisms behaviour under stressing contexts in addition to genetic codes examinations, apart from traditional teaching and cultivation of recombinant DNA on proteins and bacteria. Thus, the complete process focused on reducing manual control and automating the simultaneous managing of multiple system properties, which is a contemporary trending practice on general bioreactors with long-term processes.

Acknowledgment

We would like to share our appreciation with the Biotechnological Department of Universidad ORT Uruguay, whose tools and expertise were paramount for the realization and application of this research project.

 ISBN: 978-1-61208-971-3

REFERENCES

[1] S. Connelly, S. G. Shin, and R. J. Dillon, "Bioreactor scalability: laboratory-scale bioreactor design influences performance, ecology, and community physiology in expanded granular sludge bed bioreactors," *Frontiers in microbiology*, vol. 4, pp. 664, May 2017.

[2] Z. Chen, et al., "Performance of a novel multiple draft tubes airlift loop membrane bioreactor to treat ampicillin pharmaceutical wastewater under different temperatures", *Chemical Engineering Journal*, vol. 380, pp. 122521, 2020.

[3] M. J. Rahimi, H. Sitaraman, D. Hambird, and J. J. Stickel, "Computational fluid dynamics study of full-scale aerobic bioreactors: Evaluation of gas–liquid mass transfer, oxygen uptake, and dynamic oxygen distribution", *Chemical Engineering Research and Design*, vol. 139, pp. 283–295, 2018.

[4] B. Chezeau, and C. Vial, "Combined effects of digestate viscosity and agitation conditions on the fermentative biohydrogen production", *Biochemical Engineering Journal*, vol. 142, pp. 105–116, 2019.

[5] J. Ritonja, A. Gorsek, and D. Pecar, "Control of milk fermentation in batch bioreactor", *Elektronika ir Elektrotechnika*, vol. 26, no. 1, pp. 4-9, 2020.

[6] M. Vandermies, and P. Fickers, "Bioreactor-scale strategies for the production of recombinant protein in the yeast Yarrowia lipolytica", *Microorganisms*, vol. 7, no. 2, pp. 40, 2019.

[7] S. Arora, R. Rani, and S. Ghosh, "Bioreactors in solid state fermentation technology: Design, applications and engineering aspects", *Journal of Biotechnology*, vol. 269, pp. 16–34, 2018.

[8] M. A. Delavar, and J. Wang, "Numerical investigation of pH control on dark fermentation and hydrogen production in a microbioreactor", *Fuel*, vol. 292, pp. 120355, 2021.

[9] H. Parangusan, J. Bhadra, and N. Al-Thani, "A review of passivity breakdown on metal surfaces: Influence of chloride-and sulfide-ion concentrations, temperature, and pH", *Emergent Materials*, vol. 4, no. 5, pp. 1187–1203, 2021.

[10] J.S. Roger Jang, C.T. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*, Simon & Schuster/A Viacom Company, New Jersey, USA: Prentice Hall, 1997, pp. 1–90.

[11] J. A. Asenjo, "Bioreactor Design" in *Bioreactor System Design*, J. C. Merchuk, Ed., New York, USA: CRC Press, 1995, pp. 139–256.

[12] D. M. Berrie, R. C. Waters, C. Montoya, A. Chatel and E. M. Vela, "Development of a high-yield live-virus vaccine production platform using a novel fixed-bed bioreactor", *Vaccine*, vol. 38, no. 20, pp. 3639–3645, 2020.

[13] F. A. Hernandez, C. H. Cigliutti, and A. L. Fonseca, "Automatic Control and Remote Supervision of a Stirred Tank Bioreactor", *Memoria Investigaciones en Ingenieria*, vol. 18, pp. 34–46, 2020.

[14] B. Jaganathan and M. Mithra, "Control of a bioreactor using a new partially supervised reinforcement learning algorithm", *Journal of Process Control*, vol. 69, pp. 16–29, 2018.

[15] X. Wang, et al., "Stepwise pH control to promote synergy of chemical and biological processes for augmenting short-chain fatty acid production from anaerobic sludge fermentation", *Water Research*, vol. 155, pp. 193–203, 2019.

[16] W. Xing, et al., "pH control and microbial community analysis with HCl or CO2 addition in H2-based autotrophic denitrification", *Water Research*, vol. 168, pp. 115200, 2020.

[17] M. A. Ghanavati, E. Vafa, and M. Shahrokhi, "Control of an anaerobic bioreactor using a fuzzy supervisory controller", *Journal of Process Control*, vol. 103, pp. 87–99, 2021.

[18] M. Butkus, J. Repšytė, and V. Galvanauskas, "Fuzzy Logic-Based Adaptive Control of Specific Growth Rate in Fed-Batch Biotechnological Processes. A Simulation Study", *Applied Sciences*, vol. 10, no. 19, pp. 6818, 2020.

[19] H. J. Park, J. H. Hoon, K. G. Lee, and B. G. Choi, "Potentiometric performance of flexible pH sensor based on polyaniline nanofiber arrays", *Nano Convergence*, vol. 6, pp. 1–7, 2019.

[20] S. Rekha, B. Reshma, N. P. Dilipkumar, A. Ajai Crocier, and N. Mohankumar, "Logically Locked I2C Protocol for Improved Security", presented at International Conference on Communication, Computing and Electronics Systems, Singapore, pp. 707-716, 2020.

[21] G. Formato, et al, "Fluid–structure interaction modeling applied to peristaltic pump flow simulations", *Machines*, vol. 7, no. 3, pp. 50, 2019.

[22] K. Kim, and J. Lee, "Independent I/O Relay Class Design Using Modbus Protocol for Embedded Systems", *Journal of the Korea Society of Computer and Information*, vol. 25, no. 6, pp. 1–8, 2020.

[23] O. Yakrangi, et al., "An intelligent algorithm for decision making system and control of the GEMMA guide paradigm using the fuzzy petri nets approach", *Electronics*, vol. 10, no. 4, pp. 489, 2021.

 ISBN: 978-1-61208-971-3

Simulating Solar Irradiance through AM Wave Equations for Metabolic Pathways

Rosa María Esquinas-Ariza
Dep. de Química Física (UCLM)
02071 Albacete. Spain
email: Rosa.Esquinas@uclm.es

Javier Jiménez-Ruescas
Albacete Research Institute of Informatics (UCLM)
02071 Albacete. Spain
email: Javier.Jimenez21@alu.uclm.es

Beatriz Díaz
Dep. de Química Física (UCLM)
02071 Albacete. Spain
email: beads204@gmail.com

Hermenegilda Macià
Albacete Research Institute of Informatics (UCLM)
02071 Albacete. Spain
email: Hermenegilda.Macia@uclm.es

Edelmira Valero
Dep. de Química Física (UCLM)
02071 Albacete. Spain
email: Edelmira.Valero@uclm.es

***Abstract*—Climate change will have serious repercussions on the planet, including an increase in solar irradiance. The aim of this article was to study how this problem will affect the behaviour of plants. This is a topic of high relevance in Systems Biology and Metabolic Networks. In particular, we will study the metabolic pathway known as the ascorbate-glutathione cycle. For this purpose, solar irradiance over the course of a year was modelled using the Amplitude Modulation technique and will serve as an input to this metabolic pathway. The aim was to study how the plant behaves in the different seasons of the year and how much increase in solar irradiance the plant will be able to withstand.**

Keywords- Amplitude Modulation; Solar Irradiance; Metabolic Pathway.

I. Introduction

Light/dark cycles play an essential role in regulating plant growth and development. On the other hand, computer simulation is a key technology increasingly used in systems biology to analyze the dynamic behavior of plant metabolome. Therefore, solar irradiance becomes a relevant parameter for plant models. However, very few models contemplate this fact, and in the best of cases, they only simulate a few days.

Numerous algorithms have been developed to predict solar irradiance from meteorological data [1]. However, linking these calculations with model simulation can become difficult and many times such excessive precision is not necessary. Therefore, our aim was to address this question looking for a simple algorithm to simulate daily solar irradiance. For that, we have chosen the modulation technique known as Amplitude Modulation (AM), widely used in electronic communication in the transmission messages with a radio wave [2]. The corresponding equations have been adapted to design a basic model able to generate quasi-real solar irradiance data, which can be used as an input for metabolic pathways. Specifically, the ascorbate-glutathione redox pathway in chloroplasts has been studied [3][4] (see Figure 1).

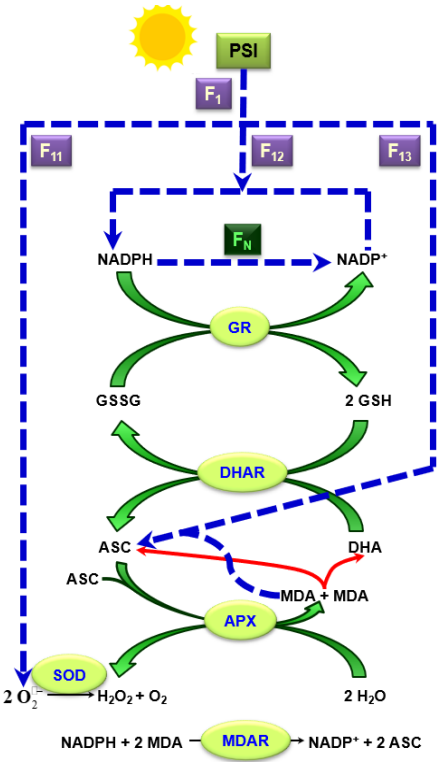


Figure 1. Scheme of the Ascorbate-Glutathione cycle.

II. Main Results

The first step is to have approximate information about solar irradiance data at the target location for a typical year. It is also necessary to take into account the duration of day and night for that place. As a proof of concept, we shall consider the city of Albacete (Spain) (Lat/Long 38.998/-1.853) for the case of global horizontal irradiance in a typical meteorological year [5].

The equations have been described in a normalised way (maximum solar irradiance is 1), and a random value will be used to simulate real fluctuations. This factor can be adapted

 ISBN: 978-1-61208-971-3

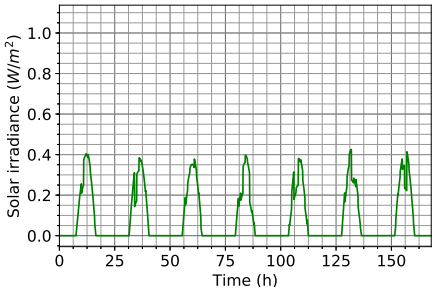


Figure 2. Normalised Solar Irradiance (7 days in winter).

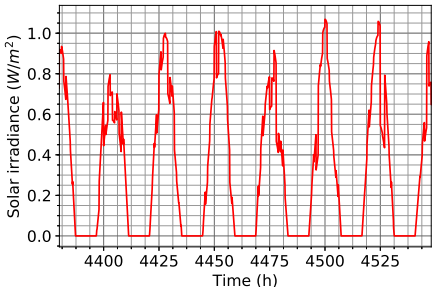


Figure 3. Normalised Solar Irradiance (7 days in summer).

and modified *ad-hoc*. Figure 2 shows a zoom for the 7 first days of winter and Figure 3, for summer. As can be seen, intensity of solar irradiance is different, and the length of light/dark periods has changed.

Figure 4 shows the normalised solar irradiance simulated through this technique for a year (8,760 h). This methodology is easily adaptable to other geographical locations.

Then, this yearly light/dark model is used to feed a complex metabolic pathway, in particular the ascorbate-glutathione cycle using Tellurium [6], a Python environment for reproducible dynamical modelling of biological networks. This cycle involves the photosensitive enzyme ascorbate peroxidase (APX) and Figure 5 shows the changes in the APX concentration over a year.

III. Conclusion

The model herein proposed can be easily used to simulate solar irradiance to input plant metabolic models. As a future work, we plan to increase the solar irradiance and study how it affects to the chloroplast; additionally we will consider other photosensitive pathways and geographic locations. All of them following the guidelines proposed in [7].

Acknowledgments

This work was supported by the Spanish Ministry of Science, Innovation and Universities (Grant RTI2018-093608-B-C32) and the JCCM (SBPLY/17/180501/000276/2), both of them cofounded by the European Union FEDER funds, and the UCLM group research grant with reference 2021-GRIN-30993 and 2021-GRIN-31073.

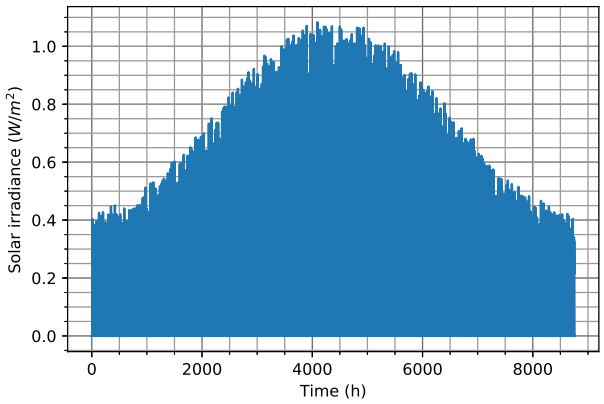


Figure 4. Normalised Solar Irradiance for a year.

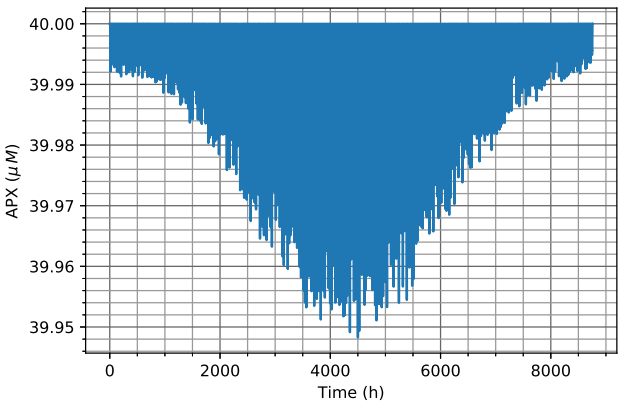


Figure 5. Fluctuations in APX concentration over a year.

References

[1] L. Huang et al., "Solar Radiation Prediction Using Different Machine Learning Algorithms and Implications for Extreme Climate Events", Frontiers in Earth Science, vol. 9, 2021.

[2] S. Ramo, J. R. Whinnery and T. Van Duzer, Fields and Waves in Communication Electronics. John Wiley & Sons, 2nd ed. 1984.

[3] E. Valero, M. I. González-Sánchez, H. Macià and F. García-Carmona, "Computer Simulation of the Dynamic Behavior of the Glutathione-Ascorbate Redox Cycle in Chloroplasts," Plant Physiology, vol. 149 (4), pp. 1858-1969, 2009.

[4] E. Valero et al., "Modeling the ascorbate-glutathione cycle in chloroplasts under light/dark conditions," BMC Syst. Biol., vol 10, pp. 1-11, 2016.

[5] European Commision, Photovoltaic Geographical Information System, https://re.jrc.ec.europa.eu (Accessed, Feb 2022).

[6] K. Choi et al., "Tellurium: An extensible python-based modeling environment for systems and synthetic biology," Biosystems, vol. 171, pp. 74-79, 2018.

[7] L. Sordo Vieira and R. C. Laubenbacher, "Computational models in systems biology: standards, dissemination, and best practices", Current Opinion in Biotechnology, vol. 75. 2022.

 ISBN: 978-1-61208-971-3

Heuristic Random Designs for Exact Identification of Positives Using Single Round Non-adaptive Group Testing and Compressed Sensing

Catherine A. Haddad-Zaknoon
dept. of Computer Science
Technion, Israel Institution of Technology
Haifa, Israel
email: catherine@cs.technion.ac.il

***Abstract*—Among the challenges that the COVID-19 pandemic outbreak revealed is the problem of reducing the number of tests required for identifying the virus carriers. To cope with this issue, a prevalence testing paradigm based on Group Testing and Compressive Sensing approach or *GTCS* was examined. In these settings, a non-adaptive group testing algorithm is designed to rule out sure-negative samples. Then, a compressive sensing algorithm is applied to decode the positives without requiring any further testing. The result is a single-round non-adaptive group testing - compressive sensing algorithm to identify the positive samples. In this paper, we propose a heuristic random method to construct the test design called $\alpha-$*random row design* or $\alpha-$RRD. In the $\alpha-$RRD, a random test matrix is constructed such that each test aggregates at most α samples in one group test or pool. The pooled tests are heuristically selected one by one such that samples that were previously selected in the same test are less likely to be aggregated together in a new test. We examined the performance of the $\alpha-$RRD design within the GTCS paradigm for several values of α. The experiments were conducted on synthetic data and sensitivity to noise was checked. Our results show that, for some values of α, a reduction of up to 10 fold in the tests number can be achieved when $\alpha-$RRD design is applied in the GTCS paradigm.**

***Index Terms*—Group Testing, Pooling Design, Compressive Sensing, COVID19-PCR**

I. Introduction

The problem of *group testing* is the problem of identifying a small amount of *items* or *subjects* known as *defective items* or *positive subjects* within a pile of elements using *group tests* or *pools*.

Denote the number of positive subjects by d and the total number of elements by n. A *group test* or a *pool* is a subset of subjects. A test result is *positive* if it contains at least one positive subject and *negative* otherwise. The objective of group testing algorithms is to find the set of positive subjects, denoted by I, with minimum number of group tests.

In this paper, we will examine *non-adaptive* group testing. In non-adaptive algorithms, tests are independent and must not rely on previous results. Therefore, all the tests can be performed in a single parallel step. The set of tests in any non-adaptive deterministic (resp. randomized) algorithm can be identified with an (resp. random) $m \times n$ *test design matrix* M (also called *pool design*) that its rows are all the assignments a that correspond to the group tests selected by the algorithm.

Group testing approach was first introduced during World War II [3], when Robert Dorfman, in 1943, suggested the method to reduce the expected number of tests needed to weed out all syphilitic soldiers in a specific unit. Among its recent applications, due to the recent pandemic outbreak, group testing approach for accelerating COVID-19 testing was widely applied across the globe. Due to severe shortages in testing kits supply, a number of researches adopted the group testing paradigm for COVID-19 mass testing not only to accelerate the testing process, but also to reduce the number of the tests required to reveal positive virus-carriers [4] [5] [6] [9] [12] [13] [16]. In many labs, COVID-19 detection was performed using *Polymerase Chain Reaction* tests or PCR tests for short. PCR-based machines can perform multiple parallel tests in single run, while each run can be several hours long. Driven by the process of PCR testing, non-adaptive group testing is most fit for these settings. In this context, the items in question are samples taken from potential patients and the *positive subjects* are samples that test positive to the virus.

While many researchers applied Dorfman's attitude with multi-stage PCR runs, some have examined designing single-PCR round tests instead. One of the promising directions is the Group Testing - Compressed Sensing paradigm (GTCS) used in [7] [8] [10] [14]. This method includes the following stages; initially, a test matrix M is designed for a single non-adaptive group testing round. Upon test results delivery, a two-stage decoding process is performed. The decoding process is purely combinatorial and does not involve any further sample testing. Using standard non-adaptive group testing decoding (e.g., Combinatorial Matching Pursuit or COMP algorithm [2]), a substantial amount of samples that tested negative to the virus are ruled out. Obviously, the main benefit of this phase is to reduce the dimension of the compressed sensing problem by cutting down the number of samples that need further decoding. This is crucial due to the

 ISBN: 978-1-61208-971-3

computational complexity of compressed sensing algorithms. In the next stage, compressive sensing techniques are used over the reduced problem (e.g., Orthogonal Matching Pursuit - OMP [15], Fast-OMP [11]), to identify real carriers.

The design of the test matrix M is crucial for both group testing phase and the compressive sensing phase that follows. In [14], the design matrix is constructed using Reed Solomon error correcting codes. The authors has checked their method on a set of $n = 384$ samples in which 5 samples are positive (about 1.3%). For pool size of 48 and using 48 group tests, they could recover all the 5 positive samples. In the work of Jirong, Mudumbai and Xu [10], the authors investigated two types of pooling designs. The first is Bernoulli random matrix where each entry is selected to be 1 or 0 with equal probability. The second design is obtained using expander graphs where each column has a fixed number of non-zero entries. The designs tested in [7] are based on Kirkman triples.

In this paper, we propose a heuristic random method to construct the test design M called $\alpha-$*random row design* or $\alpha-$RRD. In the $\alpha-$RRD, a random test matrix is constructed such that each test in M aggregates at most $\alpha < n$ samples in one group test. This model is useful in applications were tests reliability might be compromised if the pool size is large. We call α the *pool size*. The matrix rows are selected one by one. The main idea of the construction is to choose the non-zero entries of a new row according to two considerations: samples that belong to the same subject participate in similar number of tests on average (fairness); and samples that were previously selected in the same test are less likely to be aggregated together in a new test (sparsity). We perform experiments on noiseless and noisy synthetic data to examine the performance of the design, while applying Orthogonal Matching Pursuit or OMP as the compressive sensing algorithm. Practically, test designs need to be deterministic, meaning, they need to be predefined before the testing process. To use random test design, it is acceptable to make simulations of several random designs and choosing the design that performs best on some set of data. Then, this design is adopted to be used as a deterministic one for the real time tests.

The advantage of the $\alpha-$RRD design is first that it can be applied for any dimensions m and n. In many applications, the pool size is crucial for the accuracy of the testing process, therefore, it is highly recommended to use α as small as possible. In some applications there is an upper bound on the number samples that can be merged in one pool. Therefore, the $\alpha-$RRD design fits those settings when choosing α within the bounds of the pool size.

Our experiments results suggest that using the GTCS framework with $\alpha-$RRD design can reduce the number of tests dramatically. In the experiments, we tested the performance of the framework on designs with total number of tests $m = 96$ and the number of samples can be $n = 400, 600$ or 900. For each value of n, we tested on several pool sizes α that range between $\alpha = 12$ up to 48. The number of positives d ranges from 1 up to 20. For each n, α and d, we calculated the average error in restoring the positives subset over 200 random sets. The results imply that there is an evident correlation between the value of α and the performance of the process; choosing higher values of α can increase the success rate in identifying the positives. Moreover, the tests results show that the GTCS paradigm with the $\alpha - RRD$ matrix, can improve dramatically the total number of tests. In some settings, a $10-$fold improvement can be achieved compared to the single sample per test approach.

The paper is organized as follows. In Section II, we cover some definitions and preliminaries required for defining the problem of group testing and the compressive sensing in mathematical terms. Moreover, in this section, we define the group testing - compressive sensing (GTCS) paradigm. In Section III, we describe in details of the $\alpha-$RRD design and give a detailed algorithm for constructing such design. Section IV outlines experiments results designed to measure the performance of the $\alpha-$RRD design as part of the GTCS paradigm, and in Section V, we give some conclusions and future directions.

II. Definitions and Preliminaries

In this section, we define the mathematically of the problems of group testing and compressive sensing.

A. The group testing − GT problem

Let $X = [n] := \{1, \cdots, n\}$ be a set of n items or subjects, and let $I \subseteq X$ be the set of positive (defective) items such that $|I| = d \ll n$. A *group test* or a *pool* is a subset $Q \subseteq X$ of items. The quantity $\alpha := |Q|$ is called the *pool size*. The result of the test Q with respect to I is defined by $Q(I) := 1$ if $Q \cap I \neq \varnothing$ and $Q(I) := 0$ otherwise. Alternatively, we identify the test $Q \subseteq X$ with an *assignment* $a \in \{0, 1\}^n$ where $a_i = 1$ if and only if $i \in Q$.

The set of tests in any non-adaptive group testing algorithm can be identified with an $m \times n$ *test design matrix* M (pool design), where each row corresponds to an assignment $a \in \{0, 1\}^n$ that defines a group test selected by the algorithm. Upon performing the tests defined by M, each test of the m assignments in M yields the value 1 or 0 according to whether the tests contains at least one positive sample or not. Let $y \in \{0, 1\}^m$ denote the test results obtained by performing the tests of M, and let $x \in \{0, 1\}^n$ be a vector such that $x_i = 1$ if and only if $i \in I$. Formally,

$$y = M \odot x,$$

where the operation $\odot$ is defined as follows; for each $1 \leq i \leq m$,

$$y_i = \bigvee_{j=1}^{n} M_{i,j} \cdot x_j, \tag{1}$$

where the $\vee$ operation is the logic OR. It is easy to see that, the definition from (1) is equivalent to $y_i = 1$ if and only if $M_{(i)} \cap I \neq \varnothing$, where $M_{(i)}$ is the set that corresponds to the test defined by the ith row in M.

B. The compressive sensing − CS problem

Assume that each subject sample can be measured by a real valued number that expresses the *magnitude* or the *load* of the examined symptom (e.g., viral load in COVID-19 case). Let $\hat{x} \in R^n$ be a $n-$dimensional real-valued vector that signifies the symptom load of the subjects; i.e., for each $1 \leq i \leq n$, $\hat{x}_i$ indicates the symptom load of the subject i, where the value of $\hat{x}_i$ is directly proportional to the load. Symptom-free items will have their corresponding load measure equals to 0. We assume that the number of positives $d \ll n$, therefore, the load vector $\hat{x}$ is $d-$sparse; it includes only d non-zero entries. The objective is to restore the indexes of the non-zero entries in $\hat{x}$.

Similar to the definition of the result vector y from the GT settings, the design matrix M, also called the *sensing matrix* in the compressive sensing context, defines the load vector $\hat{y} \in R^n$ where each entry $\hat{y}_i$ correlates with the load of the ith pool in M. That is,

$$\hat{y} = M \cdot \hat{x}, \tag{2}$$

where the $(\cdot)$ operation is the standard matrix multiplication, therefore, for each $1 \leq i \leq m$,

$$\hat{y}_i = \sum_{j=1}^{n} M_{i,j} \cdot \hat{x}_j. \tag{3}$$

In this paper, we are interested in restoring the indexes of the non-zero entries of the vector $\hat{x}$, which is equivalent to restoring the binary vector x from the GT settings.

Formally, to find solutions for (2), we consider the following optimization problem (P_0):

$$\min_{\hat{x}} \quad \|\hat{x}\|_0 \quad \text{s.t.} \qquad \hat{y} = M \cdot \hat{x} \tag{4}$$

where $\|x\|_0$ denotes the zero norm, L_0, which is defined as the number of non-zero entries in x. The problem in (4) is NP-Hard. The main difficulty in solving (P_0) is that the constraint is highly non-smooth due to the L_0 penalty. Therefore, some relaxations are considered for approximating the solution. Even if the problem (P_0) has a unique solution, for slightly perturbed vector $\hat{y}$, the system $M \cdot \hat{x} = \hat{y}$ will no longer have a sparse solution as desired (a solution with at most d non-zero entries). Moreover, the L_0 measure is strict, and a small random noise in $\hat{x}$ causes the solution of (P_0) to be fully dense. To cope with these two problems, the following alternative problem (P_0^ϵ) is considered:

$$\min_{\hat{x}} \quad \|\hat{x}\|_0 \quad \text{s.t.} \qquad \|M \cdot \hat{x} - \hat{y}\|_2^2 \leq \epsilon^2. \tag{5}$$

It is well known that the $\epsilon-$deviation in the constraint in (P_0^ϵ) overcomes the two difficulties. Therefore, compressive-sensing algorithms designed to solve the problem (P_0^ϵ) have inherent robustness for noise. The OMP algorithm finds the best approximation for the solution of (5) using a greedy attitude. Other methods relax the L_0 norm via the L_1 norm. In our experiments, we choose $\epsilon = 10^{-3}$.

C. The group testing - compressive sensing paradigm - GTCS

The GTCS paradigm suggests a non-adaptive group testing generic algorithm for identifying the exact set of positives while using compressive-sensing based decoding techniques. The GTCS paradigm is composed of three basic phases.

1) **Create and perform the actual tests:** Create a test design M and perform the group tests defined by the design. Practically, the outcome of this stage is a vector $\hat{y} \in R^n$ as described in (2) and (3). This stage is followed by two-stage decoding process to exactly identify the test of positives. The vector y is derived from $\hat{y}$ by assigning each entry $y_i = 1$ if $\hat{y}_i > 0$, and $y_i = 0$ otherwise.
2) **Group testing decoding:** using standard group testing decoding methods (e.g., Combinatorial Maching Pursuit or COMP algorithm [2]) on the problem $y = M \odot x$, a subset $X_0 \subseteq X$ of items that are guaranteed by the GT algorithm to be negative samples is identified. This stage is used to reduce the size of the problem to be solved in the next stage. The rational behind this step is to exploit the fact that GT decoding algorithms like COMP has zero false negatives (i.e. all sample that were detected by the algorithm as negative ones are actually negative). Therefore, eliminating the set of sure-negative samples X_0 reduces the computational complexity of the step that follows, while keeping its decoding accuracy intact.
   The reduced compressive-sensing problem is established by applying the following enhancements. Given the set X_0 that includes the sure negatives, we define a new set $X_r := X \setminus X_0$. Let $Y_0 \subseteq [m]$ be the set of tests indexes that yielded the result 0 in the previous stage. The new test design matrix M_r is constructed from M, X_0 and Y_0 by projecting M on the columns that correspond to the samples in $X \setminus X_0$ and the rows that appear in the set $[m] \setminus Y_0$. Therefore, the resulting matrix is an $(m_r \times n_r)$ binary matrix where $m_r = m - |Y_0|$ and $n_r = n - |X_0|$. The reduced test result vector $\hat{y}_r$ is derived from $\hat{y}$ by deleting the entries that correspond to Y_0.
3) **Compressive sensing decoding:** by applying standard compressive sensing algorithms (e.g., OMP) on the reduced problem $\hat{y}_r = M_r \cdot \hat{x}_r$, and using the results from previous stage, the vector $\hat{x}_r$ and therefore, the vectors $\hat{x}$ and x can be restored.

III. Random Row Design - $\alpha - RRD$

In this section, we propose a random design for GT. For this design, we restrict the pool size to be at most $\alpha < n$. The design is constructed row by row. The main idea of the construction is to choose the non-zero entries in the new row according to two principles;

1) **Fairness:** Elements that participated in the minimum number of tests in previous rows, will be more likely to be chosen in the new test.

 ISBN: 978-1-61208-971-3

2) **Sparsity**: Elements that were previously selected in the same test, will be less likely to be assembled together in the new test.

For a vector $a = (a_1, \ldots, a_n) \in \{0,1\}^n$, recall that $\mathcal{H}(a) := \{i : a_i = 1\} \subseteq [n]$. The set $\mathcal{H}(a)$ is also called *the support of* a. The *Hamming weight* of a is denoted by $\omega(a)$ and is equal to $\omega(a) = |\mathcal{H}(a)|$. Let $\mathbf{0}^n$ ($\mathbf{1}^n$, ∞^n) denote the all zero (one, ∞ resp.) vector of length n. Let $A_{(m\times n)}$ be an $m \times n$ matrix over $\{0,1\}$. For all $1 \leq i \leq m$, denote by $A_{(i)} \in \{0,1\}^n$ the ith row of the matrix A, by $A^{(j)}$ the jth column and by $A_{i,j}$ the element in A that corresponds to the ith row and the jth column. The *columns weight vector* of a matrix $A_{m\times n}$ is a vector $w = (w_1, \ldots, w_n) \in R^n$ such that $w_j = \sum_{i=1}^{m} A_{i,j}$. Practically, the weight column vector indicates the number of tests each element participated in.

The procedure $\mathbf{RRD}(n, m, \alpha)$ from Fig. 1 describes the RRD strategy to choose a random design A with m rows and n columns where each row is of Hamming weight at most α. The algorithm starts by randomly choosing the first row in A from the set of binary vectors of length n and Hamming weight α. Assume that the first $\ell-1$ rows are already chosen, and let $A_{\ell-1}$ be the matrix defined by those rows. Let $\hat{w} = (w_1, \ldots, w_n) \in R^n$ be the columns weight vector of $A_{\ell-1}$. Then, the algorithm chooses the first non-zero entry in the ℓth row uniformly randomly from the set of indexes that correspond to the entries of minimal value in $\hat{w}$. This choice complies with the fairness principle.

Let $k < \alpha$ be the number of non-zero entries that algorithm already chose for the ℓth row. The $k+1$ entry is chosen as follows. Let Q_k be the set of indexes of the non-zero entires chosen so far in the current row. Let Z be the set of rows indexes i, such that $\mathcal{H}(A_{(i)}) \cap Q_k \neq \varnothing$, and let $\hat{w}$ be the weight vector of submatrix of A defined by the rows in Z. The algorithm evaluates $\hat{w}$ in steps (8) and (9) in Fig. 1. Then, the algorithm constructs the set of indexes $S \subseteq [n]$ that includes all the indexes j such that w_j is of minimum value among the entries in $\hat{w}$ and sums the corresponding columns. Let X be the set of column indexes with minimum value. Then, among the indexes in X, choose $s \in X$ uniformly at random and assign $A_{\ell,s} = 1$. These are steps (10) $-$ (16) in the algorithm in Fig. 1. This choice complies with the fairness principle.

Fig. 4 describes the procedure $\mathbf{CalcSelectedRows}$. The procedure outputs a set of row numbers $C \subseteq \{1, \ldots, \ell - 1\}$ such that $i \in C$ if and only if $\mathcal{H}(A_{(i)}) \cap Q_k \neq \varnothing$. The procedure $\mathbf{UpdateWeight}$ calculates the weight vector $w = \sum_{j\in C} A_{(j)}$ (See step no. 2 in Fig. 2). To ensure that the entries in Q_k are excluded from the selection of the next non-zero index, the weight vector w is updated to have the value ∞ in the corresponding indexes. (See step 8 in Fig. 2).

The selection of the set S complies with the sparsity principle. The set S is derived from the weight vector $\hat{w}$ by selecting the indexes with minimal weight, where the weight is evaluated over the rows that agree with one or more of the entries selected for the current test. The initialization step in $\mathbf{SumColumns}$ implies that the choice of the next non-zero entry of the current test will be from the indexes in S (See Fig. 3). Those indexes are the ones with minimum agreement with the current test, therefore, choosing the next non-zero entry from them is the best choice to keep the sparsity principle. The selection of the set X in step 14 of the algorithm in Fig. 1 complies with the fairness principle; among the best candidates from the indexes of S, the algorithm chooses s uniformly from those that appeared minimum number of times over all the previous tests.

The time complexity of generating $\alpha-$RRD design is polynomial in the dimensions of the design and can be easily generated for any dimensions m and n.

Procedure: m, n, α
Output: An $m \times n$ design matrix A
1: $A \leftarrow \{0\}_{(m\times n)}$.
2: Choose $a \in \{0,1\}^n$ uniformly at random from all vectors of weight α.
3: $A_{(1)} \leftarrow a$.
4: **for** $\ell = 2$ to m **do**
5: $\quad k \leftarrow 0, Q_0 \leftarrow \{\}$
6: $\quad$ **while** $k < \alpha$ **do**
7: $\quad\quad k \leftarrow k + 1$.
8: $\quad\quad C \leftarrow \mathbf{ClacSelectedRows}(n, Q_{k-1}, A, \ell - 1)$
9: $\quad\quad \hat{w} \leftarrow \mathbf{UpdateWeight}(n, Q_{k-1}, C, A)$
10: $\quad\quad \hat{w}_{\min} \leftarrow \min_{1\leq j\leq n} \hat{w}_j$
11: $\quad\quad S \leftarrow \{p : \hat{w}_p = \hat{w}_{\min}\}$
12: $\quad\quad \hat{z} \leftarrow \mathbf{SumColumns}(n, A, \ell - 1, S)$
13: $\quad\quad \hat{z}_{\min} \leftarrow \min_{1\leq j\leq n} \hat{z}_j$
14: $\quad\quad X \leftarrow \{t : \hat{z}_t = \hat{z}_{\min}\}$
15: $\quad\quad$ Select s uniformly at random from X.
16: $\quad\quad Q_k \leftarrow Q_{k-1} \cup \{s\}$.
17: $\quad\quad A_{\ell,s} \leftarrow 1$
18: $\quad$ **end while**
19: **end for**
20: Return A.

Fig. 1: The procedure $\mathbf{RRD}(m, n, \alpha)$

Procedure: $\mathbf{UpdateWeight}(n, Q, C, A)$
Output: An updated weight vector w
1: **if** $C = \varnothing$ **then**
2: $\quad w \leftarrow 1^n$
3: **else**
4: $\quad w \leftarrow \sum_{j\in C} A_{(}j)$
5: **end if**
6: **for** $i = 1$ to n **do**
7: $\quad$ **if** $i \in Q$ **then**
8: $\quad\quad w_i \leftarrow \infty$
9: $\quad$ **end if**
10: **end for**
11: Return w

Fig. 2: The procedure **UpdateWeight**

Procedure: n, A, ℓ, S.
Output: Sum all columns in A
1: $w \leftarrow \infty^n$
2: **for** each $i \in S$ **do**
3: $\quad w_i \leftarrow \sum_{j=1}^{\ell} A_{j,i}$
4: **end for**
5: Return w

Fig. 3: The procedure **SumColumns**(n, A, ℓ, S)

Procedure: n, A, ℓ, S.
Output: Calculate selected rows
1: $C \leftarrow \varnothing$
2: **for** each $i = 1$ to ℓ **do**
3: $\quad$ **if** $(\mathcal{H}(A_{(i)}) \cap Q \neq \varnothing)$ **then**
4: $\quad\quad C \leftarrow C \cup \{i\}$
5: $\quad$ **end if**
6: **end for**
7: Return C

Fig. 4: The procedure **CalcSelectedRows**(n, Q, A, ℓ)

IV. Experiments and Simulations

In this section, we outline tests results of the performance of the $\alpha-$RRD design when chosen as the test design in the GTCS generic paradigm. For the GT decoding, the COMP algorithm is selected to generate initial sure-negative set, while the OMP is used as the CS algorithm in the final stage.

A. Data generation

We test the performance of the $\alpha-$RRD design on both synthetic noisy and noiseless data, where an $\alpha-$RRD design matrix is generated for the dimensions $m = 96$ and $n = 400, 600$ and 900. Despite the fact that the construction from section III does not impose any limitation on the parameter m, the choice of the value of m in the experiments is derived from the number of tests that can be performed in parallel in most PCR machines used in the industry. We examine several values of α starting from $\alpha = 12$ up to 48. The minimum choice of α is derived from the applicability of the compressive sensing algorithm on the constructed matrix, while the maximum value of the pool size matches the maximum value tested for COVID-19 PCR pool designs [14]. The number of positive subjects d ranges from 1 up to 20. For each choice of n and α, an $m \times n$ $\alpha-$RRD matrix M was randomized according to the algorithm from Fig. 1 for several values of α. For each value d, we randomized 200 vectors $\hat{x} = (\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_n) \in \Re^n$ with d non-zero entries that signify the symptom load in the positive samples amongst the n samples. The vector $\hat{x}$ is chosen where the d non-zero entries are chosen uniformly at random, while the symptom load of each non-zero entry is chosen uniformly over the real range $[1, 2]$.

For each realization $\hat{x}$, the test result $\hat{y}$ is generated according to $\hat{y} = M \cdot \hat{x}$. In the noisy settings, a random noise vector v with energy $\|v\|_2 = 10^{-3}$ was added to $\hat{y}$ to generate a noisy version of $\hat{y}$, denoted by $\tilde{y} = \hat{y} + v$ (here, $\|\cdot\|_2$ denotes the L_2 norm of the vector v). In the noiseless case, we have $\tilde{y} = \hat{y}$. Given the design matrix M and $\tilde{y}$, we use the COMP algorithm followed by OMP to restore the support of $\hat{x}$. Denote by $\tilde{x}$ the result calculated after the OMP phase. Let S be the true support of $\hat{x}$, i.e. $S = \{i | \hat{x}_i > 0\}$, and $\tilde{S}$ be the support of $\tilde{x}$. The *support recovery error* is defined as

$$1 - \frac{|S \cap \tilde{S}|}{\max\{|\tilde{S}|, |S|\}}.$$

B. Tests results

Our experiments results suggest that using the GTCS framework with $\alpha-$RRD design dramatically reduces the number of tests. Fig. 5 shows the average support recovery error over all the 200 trials for each value of d and α for the noiseless case when $n = 400, 600$ and 900, where the number of positive samples is up to $d = 20$. In all these settings, an $\alpha-$RRD design matrix with total number of tests $m = 96$ is selected. For $n = 400$ and positives rate near 2.5% $(d = 10)$, the average error in restoring the correct support is less than 0.005 when α approaches 20. Moreover, the error drops to 0 for positives rate 1.5% $(d = 6)$ for $\alpha = 16$. This is $4-$fold improvement compared to the single test per sample settings. For $n = 900$ and positives rate up to 1% $(d = 9)$, for $\alpha = 48$ the error probability is less than 0.06. When the rate is 0.5% and $\alpha = 48$, the error probability drops to near 0 value. (See Fig. 5.(c)). This is $10-$fold improvement over single-test per sample method.

The results of the noiseless settings are reproduced also for the noisy case. For example, Fig. 6.(a) shows the average support recovery error over all the 200 trials for each value of d and α for the noisy case, when $n = 400$. It can be noticed that, for the same settings of the noiseless case, i.e, $\alpha = 20$ and $d = 10$, the error is bellow 0.005, and reaches 0 for $d = 6$ and $\alpha = 16$. For $n = 900$ and positives rate up to 1%, for $\alpha = 48$, the error probability is less than 0.06 and for positive rate 0.56% $(d = 5)$ the error rate is near 0 (See Fig. 6.(c)).

Moreover, the results imply that there is a correlation between the value of α and the performance of the process; choosing higher values of α can decrease the average error in identifying the positives. For example, in the noisy case, Fig. 6 shows that for $n = 400$ and $d = 10$ $(p = 2.5\%)$, the average error for $\alpha = 12$ is greater than 0.1 while it can be decreased bellow 0.005 for α between 20. For $n = 600$, and $p = 1\%$, when $\alpha = 16$, the error is about 0.069 while it drops to less than 0.001 when $\alpha = 44$. Similarly, for $n = 900$, $p = 0.56\%$ and $\alpha = 48$, the error probability is near zero, while for $\alpha = 20$, the error is higher than 0.1. This paradigm is reproduced in both noisy and noise free case too.

In practice, deciding the best value for α for the problem in-hand can be done while taking in consideration the limitations of the test process (for example, if there is some upper bound on the pool size). Once such limitations on the value of α are known, we can use computer simulation on synthetic data, similar to the ones described in this work, to decide on the best choices of α for each settings.

 ISBN: 978-1-61208-971-3

The results in Fig. 5 and Fig. 6 show that the error rate increases with d. This behavior is as expected from any group testing - compressive sensing algorithm, since those are designed to be used when the solutions x and $\hat{x}$ of the equations (1) and (2) are sparse vectors, meaning the number of the non-zero entries d is very small relative to n, more precisely when $d = O(\sqrt{n})$.

It is worth noticing that, the differences in the noisy case vs. the noise-free case are almost negligible. This behavior can be explained by two major factors. First, group testing decoding algorithms like COMP used in our simulations, are known for their robustness for *false negatives* (a false-negative is a sample classified by the algorithm as negative, but it is actually positive). That is, the set of samples classified by the GT algorithm as sure-negatives and therefore excluded from the decoding of the CS algorithm, does not include false-negatives. Therefore, it is highly unlikely to miss positive samples during the GT initial classification. The second factor is the noise-tolerance of the compressive sensing algorithm. Specifically, the OMP algorithm is known for its high accuracy in the presence of noise, while its drawback is its computational complexity.

V. Conclusion

In this paper, we suggested a new random pooling design $\alpha-$RRD. This design can be used as part of the GTCS paradigm in order to build a single-round non-adaptive group testing protocol to exactly identify positives within a large set of elements. The complexity of generating $\alpha-$RRD design is polynomial in the dimensions of the design and can be easily generated for any dimensions m and n. By its design, the $\alpha-$RRD pooling matrix is designed to restrict the size of the pool α which might be critical for test accuracy. If there is no practical restrictions on the size of α, then, given the parameters m, n and positives rate, the best choice for the parameter α can be concluded using computer simulations. Moreover, since random sensing matrices can perform well with compressive-sensing algorithms, the GTCS paradigm can be further tested with other well-known group testing random designs such as RID, RrSD, RsSD and Transversal design [1]. Similarly, other compressive sensing algorithms can be applied too. Besides being tested on synthetic data, it is worth examining the efficiency of the method and the design on real COVID-19 data or any other disease that follow the same paradigm.

References

[1] N. H. Bshouty, G. Haddad and C. A. Haddad-Zaknoon, "Bounds for the number of tests in non-adaptive randomized algorithms for group testing," SOFSEM 2020: Theory and Practice of Computer Science, Lecture Notes in Computer Science, Springer, vol. 12011, pp. 101–112, 2020.

[2] C. L. Chan, S. Jaggi, V. Saligrama and S. Agnihotri, " Non-adaptive group testing: Explicit bounds and novel algorithms," 2012 IEEE International Symposium on Information Theory Proceedings, pp. 1837-1841, 2012.

[3] R. Dorfman, "The detection of defective members of large populations," The Annals of Mathematical Statistics, vol. 14(4), pp. 436-440, 1943.

[4] A. Eis-Hübinger et al., 'Ad hoc laboratory-based surveillance of SARS-CoV-2 by real-time RT-PCR using minipools of RNA prepared from routine respiratory samples," Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology, vol. 127, 104381, 2020.

[5] J. Cabrera et al., "Pooling for SARS-CoV-2 control in care institutions," BMC infectious diseases, vol. 20(1), 745, 2020.

[6] R. Ben-Ami et al., "Large-scale implementation of pooled rna extraction and rt-pcr for sars-cov-2 detection," Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases, vol. 26(9), pp.1248–1253, 2020.

[7] S. Ghosh et al., "A Compressed Sensing Approach to Pooled RT-PCR Testing for COVID-19 Detection," IEEE Open Journal of Signal Processing, vol. 2, pp. 248-264, 2021.

[8] S. Ghosh et al., "Tapestry: A single-round smart pooling technique for covid-19 testing," Unpublished.

[9] C. Gollier and O. Gossner, "Group testing against covid-19.," Covid Economics, vol. 1, pp. 32–42, April 2020.

[10] J. Yi , R. Mudumbai and W. Xu, "Low-cost and high-throughput testing of COVID-19 viruses and antibodies via compressed sensing: system concepts and computational experiments,", unpublished.

[11] M. Yaghoobi, D. Wu and M. E. Davies, "Fast non-negative orthogonal matching pursuit," IEEE Signal Processing Letters, vol. 22(9), pp. 1229–1233, 2015.

[12] C. Mentus, M. Romeo, and C. DiPaola, "Analysis and applications of adaptive group testing methods for covid-19.," medRxiv, https://www.medrxiv.org/content/early/2020/04/07/2020.04.05.20050245, 2020.

[13] H. Shani-Narkiss, O. Gilday, N. Yayon, and I. Landau., "Efficient and practical sample pooling for high-throughput pcr diagnosis of covid-19," medRxiv, https://www.medrxiv.org/content/early/2020/04/14/2020.04.06.20052159, 2020.

[14] N. Shental et al., "Efficient high throughput sars-cov-2 testing to detect asymptomatic carriers," Science Advances, vol. 6, pp. eabc5961, 2020.

[15] Y. C. Pati, R. Rezaiifar and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," Proceedings of 27th Asilomar Conference on Signals, Systems and Computers, vol. 1, pp. 40–44, 1991.

[16] I. Yelin et al., "Evaluation of covid-19 rt-qpcr test in multi-sample pools," Infectious Diseases Society of America, vol. 71(16), pp. 2073–2078, 2020.

 ISBN: 978-1-61208-971-3

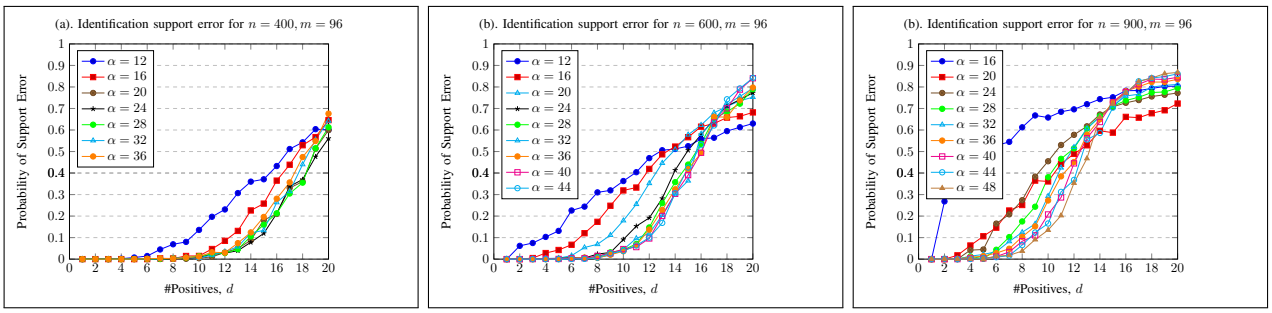


Fig. 5: Probability of support error for d up to 20, $n = 400, 600$, and 900, $m = 96$ and $\alpha \leq 48$ for the noise-free case.

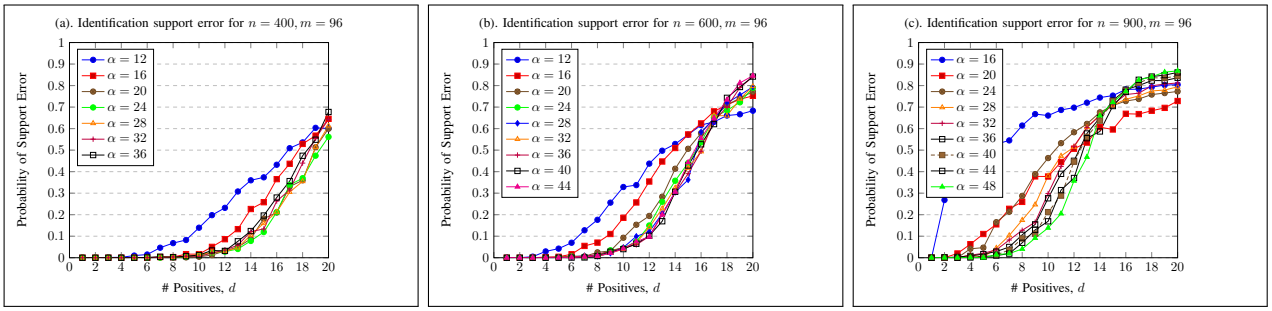


Fig. 6: Probability of support error for d up to 20, $n = 400, 600$, and 900, $m = 96$ and $\alpha \leq 48$ for the noisy case.

 ISBN: 978-1-61208-971-3

A New Phased Array Magnetic Resonance Imaging Coil For Hbo2 Studies

Azma Mareyam[1], Erik Shank[2], Laurance Wald[1], Michael Qin[3], and Giorgio Bonmassar[1]
[1] A. A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Boston, USA.
e-mail: giorgio.bonmassar@mgh.harvard.edu
[2] Department of Anesthesia, Massachusetts General Hospital, Harvard Medical School, Boston, USA,
e-mail: eshank@mgh.harvard.edu
[3] Office of Naval Research, United States Department of Defense · Navy, Arlington, USA.
e-mail: michael.qin@navy.mil

***Abstract*—Functional Magnetic Resonance Imaging (fMRI) measures brain activity by detecting blood flow changes as cerebral blood flow and neuronal activation are coupled. fMRI is noninvasive, is considered safe, and may allow studying the brain under hyperbaric conditions. This new coil may be necessary for studying decompression sickness and disorders of hyperbaricity, including nitrogen narcosis. This study focuses on the safety and technical details of building fMRI coils for human hyperbaric studies. One of the most remarkable properties of this novel technology is that the new multichannel arrays provide high-quality images at 3 Tesla (T) MRI, one of the highest strength magnets among the most common MRI scanners available in the market. The paper describes all the risks associated with simultaneous MRI and Hyperbaric oxygen therapy (HBO2) and discusses mitigation strategies and regulatory testing. One of the most significant risks for this type of study is a fire in the hyperbaric chamber caused by the sparking of the MRI coils due to high voltage RF arcs. RF pulses at 128MHz elicit signals from the human tissues, and RF sparking occurs commonly and is considered safe in normobaric conditions. We describe how we built a coil for HBO2-MRI studies by modifying an eight-channel phased-array MRI coil with all the mitigation strategies discussed. The coil was fabricated and tested with a unique testing platform that simulated the worst-case RF field of a 3 Tesla MRI in a Hyperlite hyperbaric chamber at 3 atm pressure. The coil was also tested in normobaric conditions for image quality in a 3 T scanner in volunteers and SNR measurement in phantoms. Further studies are necessary to completely characterize the coil safety for HBOT/MRI studies by following the Guidance for Industry and Food and Drug Administration titled "Testing and Labeling Medical Devices for Safety in the Magnetic Resonance (MR) Environment".**

Keywords-hyperbaric oxygen therapy; MRI; safety; diving medicine.

I. INTRODUCTION

Magnetic Resonance Imaging (MRI) and functional Magnetic Resonance Imaging (fMRI) have rapidly gained acceptance as "the gold standard" for diagnosing and evaluating neurologic conditions. This adoption comes from MRI's perceived non-invasiveness, avoidance of ionizing radiation, and ability to elucidate the human nervous system's fine anatomic and functional nuances. MRI and fMRI have, however, not been utilized in the study of Hyperbaric Oxygen Therapy (HBO2) due to the genuine dangers that these two challenging environments present.

HBO2 has been accepted for the treatment of many neurological conditions (e.g., decompression sickness, carbon monoxide poisoning, cerebral arterial gas emboli, etc.) and postulated to help in many others (stroke, cerebral palsy, and even autism); however, there has not been a safe method to utilize the tools of MRI and fMRI to evaluate mechanisms and efficacy of HBO2. The hyperoxic high-pressure environment of HBO chambers and the powerful magnetic fields of the MRI scanner are traditionally incompatible. There are genuine risks of mechanical injury to the integrity of most chambers by magnetic forces, space, and access challenges to the patients in both monoplane chambers and MRI scanners, and --likely the most concerning-- the risk of fire from radio frequency (RF) MRI generated arcing in a hyperoxic environment. For many years HBO2 has provided safe and effective treatments for many diseases. Hyperbaric chambers provide oxygen administration in a manner that has few side effects. Combining these two technologies (HBO2 and MRI) can reveal other illnesses that HBO2 and fine-tuning established HBO therapies can treat, such as stroke rehabilitation [1]. In this paper, we discuss the risks as well as strategies and technological approaches to mitigate these risks and enable MRI and fMRI studies to be performed under hyperbaric conditions. We hope that these technological breakthroughs will permit both an increased understanding of HBO2 mechanisms as well as a clinical tool to evaluate the efficacy of HBO2.

II. METHODS

A new type of receive array MRI coil was designed and built to minimize fire hazards by using a combination of electronic protection components and fire retardant epoxy for insulation. Figure 1 (left panel, Top), in which eight surface coils are connected each to an independent amplifier and receiver channel (see below). The outputs from the receiver channels are combined in an optimum manner with a phase correction dependent on the point in space from which the signal is originated. Technical issues related to the mutual inductance of the coils have been addressed by using partial loop overlap [2]. In contrast, we followed the

 ISBN: 978-1-61208-971-3

current state-of-the-art MRI systems by adopting a phased array coil design.

III. RESULTS

The new MRI coil for HBO2 was fabricated (Figure 1, left panel) and tested at 3 Tesla. The constructed array coil passed the safety tests [3] without additional adjustment beyond the bench adjustments. The design included an 8-channel phased array with extremely large area coil loops that have resulted in very high image quality. The head coil was composed of (Figure 1 left panel, middle): (A) lattice balun with a PIN diode for detuning during transmit and every soldering was inspected and photographed, (B) copper wires loops, (C) distributed capacitors, and (D) fuses.

The following image acquisitions were performed on a human volunteer for a total scan time of approximately 2 hours. Figure 1A (right panel) shows the amplitude and phase of a field map. A field map can help reconstruct high-fidelity images as it is typically acquired during the MRI system tuning. Furthermore, lipid suppression can be much more robust by measuring the field map and adjusting the acquisition parameters.

One of the most common MRI sequences is the T1-weighted scan which depicts differences in signal based upon intrinsic T1 relaxation time of various tissues. Figure 1B-C (right panel) shows the typical and high-quality T1-weighted images. In these images, fat tissue realigns its longitudinal magnetization with B0, which appears bright. Conversely, tissues predominately made out of the water, such as central spinal fluid, have a much slower longitudinal magnetization realignment after an RF pulse and appear dark.

Separate MRI and HBO2 studies have shown that imaging is a quantitative biomarker that can help guide and optimize HBO2 therapy [4]. However, only performing simultaneous fMRI and HBOT will enable studying the brain-altering effects of nitrogen narcosis or oxygen toxicity.

IV. CONCLUSIONS

To the best of our knowledge, this is the first work aiming at imaging the brain in extreme pressurized environments, like HBOT, which have significant repercussions for understanding the brain in conditions like decompression sickness, which are currently poorly understood. We expect to complete the safety studies and request IRB/IDE to complete the HBO2/fMRI studies in future work.

REFERENCES

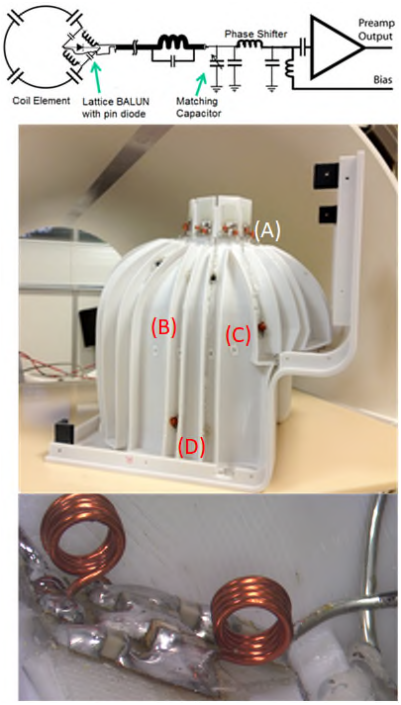
[1] R. Boussi-Gross, et al., "Improvement of memory impairments in poststroke patients by hyperbaric oxygen therapy," Neuropsychology, vol. 29 pp. 610-21, 2015.

[2] P. B. Roemer, et al., "The NMR phased array," Magn Reson Med, vol. 16, pp. 192-225, 1990.

[3] A. Ghotra, et al., "A size-adaptive 32-channel array coil for awake infant neuroimaging at 3 Tesla MRI," Magn Reson Med, vol. 86: pp. 1773-1785, 2021.

[4] C. Q. Li, S. Gerson, and B. Snyder, "Case report: hyperbaric oxygen and MRI findings in radiation-induced optic neuropathy," Undersea Hyperb Med, vol. 41, pp. 59-63, 2014.

 ISBN: 978-1-61208-971-3



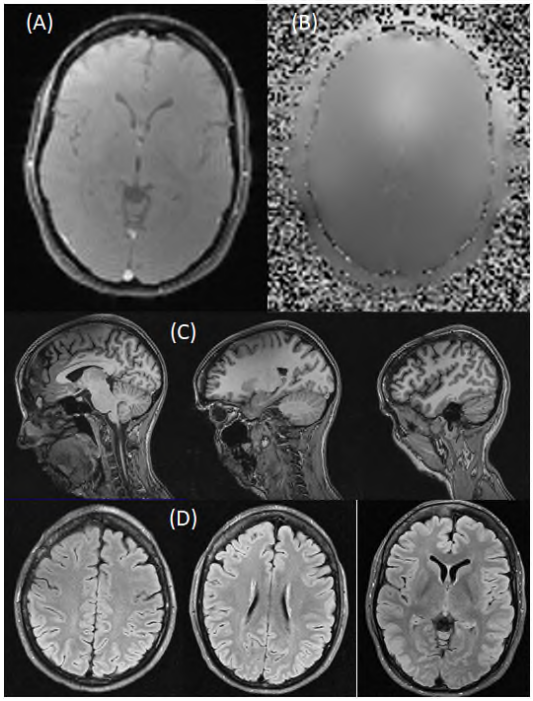


Figure 1 (LEFT PANEL) The MRI Coil. (Top) Design of the layout and schematics of the 8-channel Phased Array Coil System. (Middle) The new coil: (Bottom) lattice balun. (RIGHT PANEL) 3 Tesla normobaric MR images. (Top) Field map images of magnitude (A) and phase (B), sagittal (C), and axial (D) T1-weighted MPRAGE images.

 ISBN: 978-1-61208-971-3

Simpati: Network-based System for Patients' Classification Reveals Disease Specific Pathways Driven by Cohesive Communities

Luca Giudice
A.I. Virtanen Institute for Molecular Sciences
University of Eastern Finland
Kuopio, Finland
luca.giudice@uef.fi

Claudia Mengoni, Rosalba Giugno
Department of Informatics
University of Verona
Verona, Italy
claudia.mengoni@univr.it, rosalba.giugno@univr.it

***Abstract*—Patient classifiers should be able to rely on the strength of machine learning methodologies while not losing biological interpretability. So far, most of the developed methods lack in one of the two aspects. We propose Simpati, a pathway-based tool for patient classification, which enables accurate classification focusing on the detection of relevant biological features and patient cohesive communities. The tool makes it possible to classify patients and investigate the features which were mostly representative of each class. It presents ad-hoc algorithms for the processing of patient similarity networks and proposes an effective simulation strategy as a recommender system to predict a patient's class based on graph topology. Its computational performance, classification performance and biological validation were performed on genetic data from different types of cancer and compared favorably with state-of-the-art competitors.**

Pathway-based classification; Network-based propagation; Patient similarity network; Subgroup cohesive algorithm.

I. INTRODUCTION

High-throughput biological data provide valuable information to clinicians for the prognosis and treatment response of patients. They offer quantitative and qualitative evidences to biomedical scientists for developing a study or confirming wet-lab results. Pathway-based analysis is a technique to investigate these data and detect molecular mechanisms related to the patients [1][2]. The pathway space is more robust to noise than the single feature level, summarizes the information of multiple patient's molecules into the pathway activity (inhibited or activated), reduces the model complexity and maintains predictive accuracy [3][4]. Nowadays, pathway-based analysis is mostly performed through enrichment tools, fundamental methods which provide to clinicians understanding of the cellular functions affected in a patient, so that they can better define a disease phenotype and manually classify patients. Although some attempts have been made to couple pathway enrichment with classification [5], pathway-based classifiers that do not require pathway enrichment (i.e., supervised classifiers able to integrate simple pathway information to classify biological samples), are not yet strongly developed.
Among them there are two classifiers that exploit the idea of pathway. The first is PASNet [6], which incorporates biological pathways in a Deep Neural Network. The neural network is composed by an input gene layer, a pathway layer, a hidden layer that represents hierarchical relationships among biological pathways and an output layer that corresponds to the patient classes. The second is netDx [7] and represents pathways thanks to the Patient Similarity Network (PSN) paradigm. In a PSN, each node is an individual patient and an edge between two patients corresponds to pairwise similarity for a given patient's feature (e.g., gender, height, gene expression). All the user-provided data are converted into PSNs and molecular data can be converted into networks representing pathways. This made netDx a pioneer classifier able to combine multi-omics and pathway specific features. The decision system of the software relies on GeneMANIA [8], state-of-art gene function predictor, to select the best patient similarity networks and to use them in the classification. netDx revealed to be better than canonical machine learning algorithms and to provide a good level of interpretability based on the network's graphical representation. However, the software requires the user to define a similarity measures for each input data and manually tune hyper-parameters, making the results highly dependent on users choices. Additionally, netDx does not consider the topology of the networks for inferring the relationships between training and testing patients, providing a black box prediction difficult to interpret.
A classifier should be able to benefit both from the interpretability of pathway-based enrichment tools and the strength of machine learning methodologies [9]. We want to stand up to the challenge by proposing the pathway-based classifier Simpati. Our method provides a novel feature-selection strategy for classifiers based on patient similarity networks, implements a subgroup cohesive algorithm for extracting patient communities in PSNs and proposes an effective simulation strategy to predict a patient's class based on graph topology. Plus, the method introduces ad-hoc operations for genetic data to reduce the number of hyper-parameters, similarity measures, or external software that the user has to define or install, it naturally handles outliers and integrates a graphical user interface to allow the visualization of the networks.

This text is structured as follows: in the Methods section the general workflow of the tool is described and different

 ISBN: 978-1-61208-971-3

subsections detail the implementation of each step. These include all steps necessary for data preparation, feature selection and prediction, as well as a description of required input data and possible downstream analyses. In the Results section, Simpati performances are compared to those of two state-of-the-art competitors, both in terms of computational requirements, classification performance and biological interpretation. Finally, the Conclusions section remarks the impact of this classifier, its limitations, and its future development.

II. METHODS

In this section, a general overview of Simpati's workflow is given, then the other subsections detail the specific aspects of implementation of each step. The R package to use Simpati and its graphical interface can be found online [10][11].

A. Overview

Simpati is a binary patient classifier, which exploits the similarity of patients' molecular profiles at the pathway-level. An overview of the method is shown in Figure 1. It takes as input patients' genetic profiles similarly to a gene differential analysis setting where counts have been library normalized and two classes are to be compared. The method has to be provided also with a list of pathways and a gene interaction network. Simpati transforms the profile of each individual patient to take into account the interconnectivity of genes. Each profile is propagated over the interaction network and the transformed data are used in the downstream analysis. Next, Simpati creates, selects and cleans PSNs. For each set of genetic features falling into a pathway, Simpati creates a Pathway-Specific PSN (psPSN), tests if the two patient classes show separability and finds cohesive communities inside each class. A psPSN is retained if it shows a strong intra-similarity between patients of one class, while having at the same time a weak intra-class similarity in the other class and a weak inter-class similarity. Once a network is selected as significant, Simpati removes patients showing an outlier pathway activity as compared to the rest of patients in the same class. Signature pathways are then used to classify patients of unknown class, based on their similarity to labeled patients.

B. Network-based data preparation

The first step is the transformation of patients' biological profiles using a network-based propagation algorithm. Each single-level feature gets a new value based on its a priori

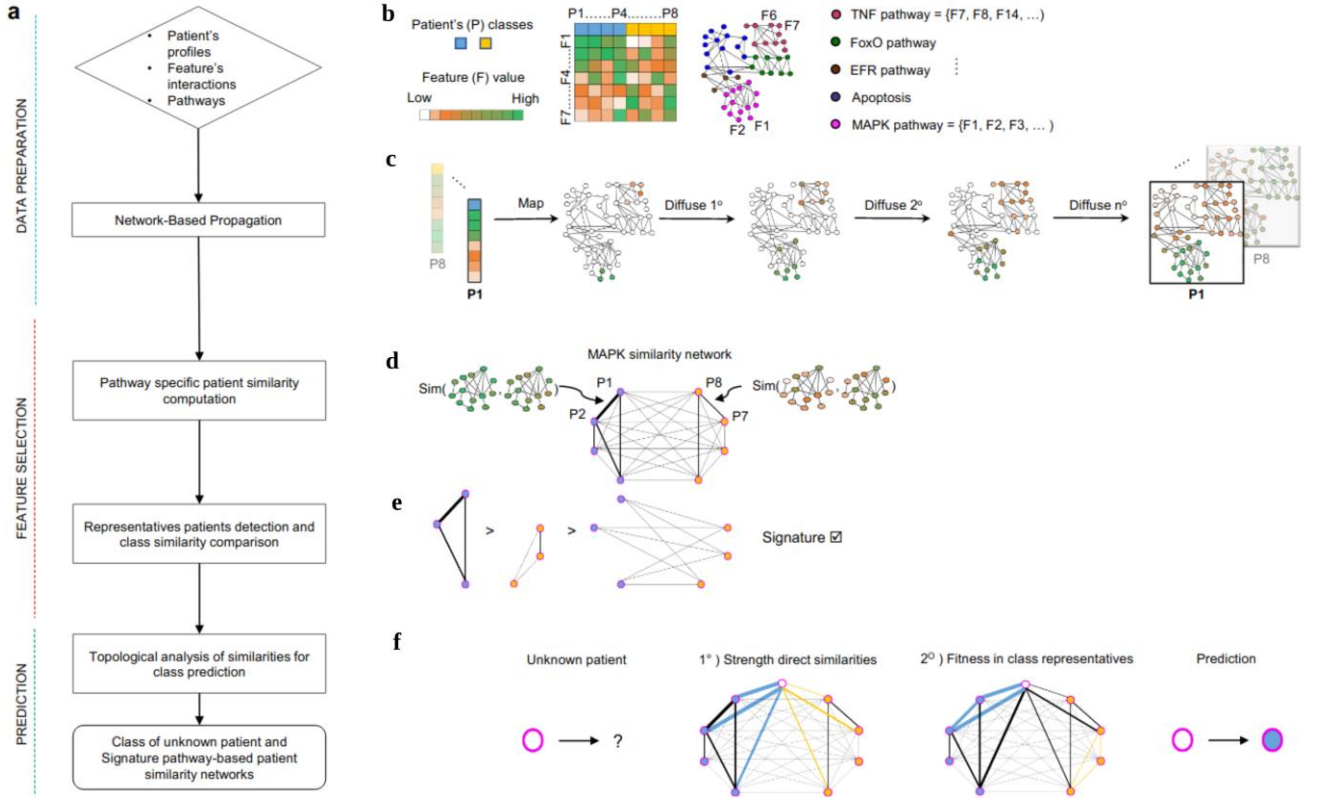


Figure 1. Simpati's workflow: (a) Overview of the main steps. (b) Input: matrix of features by patients, feature interaction network, features grouped by pathways. (c) Each patient's profile is propagated on the interaction network. (d) Within each pathway's subnetwork patient's similarity is computed. Patient's similarities are the edges of Pathway-Specific Patient Similarity Networks (psPSNs). (e) A psPSN is signature if intra-class similarities are stronger that intra-class similarities of the other class and inter-class similarities. (f) Unknown patients are classified based on their similarity to other patients and on how well they resemble class representatives.

 ISBN: 978-1-61208-971-3

information (e.g., gene expression) and on its associations with all the other molecules in the network. At first, the values of the patient's genetic features are mapped to their corresponding nodes in the provided interaction network, then Simpati propagates their values through the interactions. Each node, including the ones without a value, gets a score, which reflects its starting information and the amount given and received from its neighbors. Simpati propagates using the random walk with restart algorithm on the row-normalized network [12]. Propagating a patient's profile starting from the genetic single-value features allows us to obtain a genome-wide profile. This is relevant because the profile can be compared across patients and gives a genome-wide overview of all the genes. Moreover, this is particularly beneficial when we deal with sparse data (e.g., somatic mutation data) where fewer features are identified from the analysis [13].

C. Pathway specific patient similarity

Simpati computes a pairwise similarity between patients for each set of genetic features falling into a specific pathway. In this way, Simpati creates a database of psPSNs reflecting the similarity of patients in each pathway. The nodes of a psPSN are all the patients with known class and the edges are weighted to reflect the pairwise similarity of patients in the features belonging to the pathway.

The approach of measuring similarity on a pathway-level, not only allows to reduce the dimensionality of the features to be compared across patients, but it also creates a feature space, which is more robust to noise compared to single features, while still retaining predictive accuracy [14].

Pathway-specific patient similarity is computed as a linear combination score of three factors. The first one (1) is the Weighted Jaccard and determines how similar the propagated values between two profiles are; the second factor (2) determines how high or low the propagated values are, while the third factor is the opposite of their difference (3). The similarity increases as the two patients have similar values and at the same time high values for the same single-level feature. This is reflected in the final similarity measure, called Trending Matching (4):

$$WJ_p(P_a,P_b) = \frac{\sum_g \min(m_{g,a}, m_{g,b})}{\sum_g \max(m_{g,a}, m_{g,b})} \quad (1)$$

$$MG_p(P_a,P_b) = \frac{\sum_g (m_{g,a} + m_{g,b})/2}{|p|} \quad (2)$$

$$DIFF_p(P_a,P_b) = 1-|WJ_p(P_a,P_b)-MG_p(P_a,P_b)| \quad (3)$$

$$TM_p(P_a,P_b) = WJ_p(P_a,P_b)+MG_p(P_a,P_b) +DIFF(P_a,P_b) \quad (4)$$

where p is a pathway, P_a and P_b are two patients, g are all the features $| g \in p$, and m is the matrix of features by patients.

D. Feature selection and Best Friend Connector algorithm

Simpati evaluates which pathways are signatures for one of the classes. The members of one class must be more similar (strong intra-similarities of one class) than the members of the opposite class (weak intra-similarities of the other class) and the two classes are not similar (weak inter-similarities). In other words, the topology of the psPSN must reflect the presence of a clique of nodes belonging to the same class being more strongly connected than the rest of the patients. Despite this criterion being genetically intuitive, it is not easy to satisfy due to the complex structure of a patient similarity network where each patient is connected to any other member of the classes in comparison. One patient can easily be more similar to the patients of its opposite class in one specific pathway activity and decrease the separability of the groups. To account for this situation and making the feature selection more robust to outliers at the level of the single pathway, we developed an algorithm called Best Friends Connector algorithm (BFC). The latter is a cohesive subgroup detection algorithm implemented specifically for PSNs to find the strongest community of patients from each class in a network. The algorithm relies on the definition of the concepts of first order best friend (1BF), second order best friend (2BF) and outsiders. Given a root node, its 1BFs are its most similar nodes. 2BFs are the nodes that are not among the root's 1BFs but are 1BFs to one of the root's 1BFs. Outsiders do not belong to any of the previous definitions. The algorithm performs the following operations. It first adjusts the weights of the intraclass connections. Precisely, it increases the similarity of two patients when they both have a weak similarity with outsiders and it decreases it in the opposite case. Then, it iteratively considers one patient as root, it assesses the average of the intraclass connection weights of the subgroup composed by his 1BFs and 2BFs. When each patient has been considered, the algorithm retrieves the set of best friends who got the strongest connections. The cardinality of the 1BFs and 2BFs subgroups, as well as the size of the final subgroup, are customizable.

E. Classification

The signature pathways identified by Simpati are used to classify unknown patients. Each of them is compared to already annotated patients and assigned to the same class of who is most similar to. However, the only strength of similarity could be misleading. The unknown patient could have the strongest similarity with outlier members of the class. Therefore, we designed Simpati to consider also how much the unknown patient represents the class.

The patient to be classified undergoes the same preprocessing described for annotated patients: its profile is propagated in the interaction network and its pairwise TM similarity to each annotated patient is computed, so that the unclassified patient becomes itself a node in each signature psPSN. Then, Simpati associates the profile to one of the classes based on the results of two approaches. For the first, it determines the average similarity of the patient to the members of each class. The patient would be assigned the class to which it has the strongest similarity. For the second

 ISBN: 978-1-61208-971-3

approach, Simpati pretends that the patient belongs to one class and measures how far it is from being considered an outlier. The patient would be assigned the class in which it is considered less of an outlier with respect to the other members. More specifically, the patient is simulated to belong to one class and the BFC algorithm is run iteratively. At each run, the algorithm is asked to return a smaller number of strongly connected individuals. The iteration stops when the patient does not belong to the best subgroup. A large number of iterations reflects a strong similarity of the patient to the class representatives. Due to this, the patient would be a candidate to be assigned the class in which it survived the highest number of iterations. Simpati assigns the patient to the class that has been predicted by both the approaches. In case, the results are not concordant, then Simpati does not make the prediction and the pathway together with its PSN are removed from the downstream operations. This step is performed for all signature psPSNs, then the patient's definitive class is the one to which the patient has been most frequently assigned.

The classification performance are evaluated with a leave one out cross validation approach, such that iteratively one patient is considered unknown and composes the testing set, while the others are known and are used as training to determine which pathways are signature. The performance on the testing set are computed using area under the receiver operator characteristic curve (auROC) and area under precision recall curve (auPR) metrics.

F. Downstream Analysis

The signature pathways that are used to classify at least one patient are reported in the final output of Simpati and information about which class they were identified to be signature for. To further pinpoint the most relevant pathways and confirm their signature role for a class, an empirical probability value is computed. On each signature psPSN it is tested whether by randomly shuffling the patients between the two classes, the pathway is still predictive of the original signature class.

To improve the interpretability of the results some other information is computed. First, it has been established that signature pathways reflect strong similarity between members of one class. However, Simpati also reports whether the members are similar in having high values (e.g., high gene expression), reported as up-involved signature pathway, or low values (e.g., low gene expression), reported as down-involved signature pathway. Additionally, based on the BFC results, it is reported how many times a patient has been considered an outlier for its class.

When the features of the profiles provided as input to Simpati are genes and the classification aims to determine association to a disease, it is possible to validate the biological relevance of the identified pathways within Simpati. Queries to the gene-disease associations database (DisGeNet) [15] and to the Human Protein Atlas [16] allows detecting whether the features returned are already known to be associated with the disease being tested.

To obtain a graphical representation of the psPSNs of interest, Simpati offers a graphical interface, which allows to obtain a compact representation of the networks. Patients are grouped based on their similarity so that, instead of plotting all nodes, only some representatives are depicted, making the interpretation of the figure much more feasible.

G. Data preparation for testing

Simpati performances were tested by classifying patients from five cancer types, extracted from The Cancer Genome Atlas (TCGA) using the R packages curatedTCGAData (v1.1.38) [17] and TCGAutils [18]. Two types of biological omics were tested for each cancer type, gene expression from RNAseq data and somatic mutations. The classes assigned to the patients were based on disease stage progression binarized into Early (stage I and II) or Late (stage III and IV). Data preparation for the RNAseq followed the workflow defined by Law et al. [19], while somatic mutation data have been converted into a binary matrix, where a value equal to one was indicating a mutated gene in a patient and zero otherwise. Finally, the six datasets were composed of the following number of samples: 14 Liver hepatocellular carcinoma (LIHC) (7 Early, 7 Late), 21 Stomach adenocarcinoma (STAD) (8 Early, 13 Late), 37 Kidney renal clear cell carcinoma (KIRC) (24 Early, 13 Late), 45 Bladder Urothelial Carcinoma (BLCA) (8 Early, 37 Late), 75 Lung squamous cell carcinoma (LUSC) (60 Early, 13 Late) and 152 Esophageal carcinoma (ESCA) (91 Early, 61 Late) patients.

Pathways were collected from the major databases MSigDB [20] and GO [21] and KEGG [22], while a Biogrid network (v4.2.191) [23] was used to model the biological feature's interactions.

III. RESULTS

Simpati classification results and computational performance were compared to those obtained with netDx (v1.2.0 14-10-2020) for both gene expression and somatic mutations on the prepared TCGA datasets and with PASNet only for gene expression, as this tool does not handle sparse data. Additionally, a biological validation of the pathways retrieved was performed on Simpati and netDx. An online repository is available with a tutorial on how to replicate the results [24].

The classification comparison was performed on the metrics supported by both netDx and PASNet, the auROC and the auPR. These were obtained from a 10-fold cross-validation approach in netDx and a stratified 5-fold cross-validation repeated 10 timed in PASNet, based on the authors' vignette, while for Simpati it was obtained through the leave one out cross validation approach. Figure 2 shows how Simpati performs better than the competitors in both the measures and the biological omics. Simpati also proves

 ISBN: 978-1-61208-971-3

Figure 2. Classification performance comparison between methods. The top box shows performance on the RNAseq datasets, the bottom box on the somatic mutation datasets.

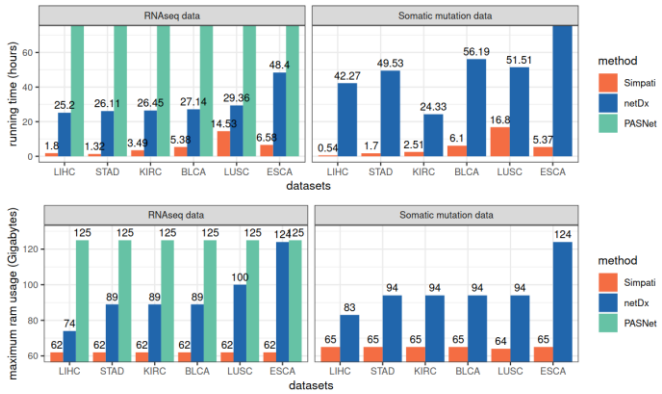


Figure 3. Computational performance comparison between methods. The top box shows running time in hours, the bottom box shows RAM usage in Gigabytes.

to be more reliable in each dataset with a standard error equal to zero due to its leave-one out cross-validation approach.

The patient similarity network paradigm used by Simpati and netDx brings many advantages both in the feature selection, in the classification phase and in the overall interpretability of the software. However, these pros come with a price, which is the software scalability already introduced as a challenge by Pai et al. [5]. A PSN is a complete graph that the methods build with all the patients and for every pathway. This means that an increment in the number of patients and in the number of annotated pathways lead the methods to require more computational resources. netDx and Simpati faced this point with different approaches. netDx is implemented in R and Java, uses the disk to save temporary files and applies a sparsification of the PSNs to decrease the number of edges and so the amount of information associated with them. Simpati is implemented completely in R, natively supports parallel computing and handles all the data of the workflow as sparse matrices or vectors. The RAM usage and the running time required to classify the TCGA datasets were monitored with the same hardware settings for all tools (32-Core Processor, 251 Gigabyte System memory). Simpati compared favorably in the usage of the resources, as reflected in Figure 3. On average across the datasets, Simpati it's ~ 16 times faster than netDx and requires ~ 1.5 times less Gb of RAM. Both netDx and Simpati outperformed PASNet performance.

Both Simpati and netDx provide the most relevant pathways they detect during the workflow. These pathways should help characterize patient's classes and improve the interpretability of the method. For this reason, Simpati integrates into its workflow a biological validation step exploiting DisGeNet and the Human Protein Atlas. For each dataset, a set of key words describing the disease are defined, then the percentage of key words associated with the pathway in DisGeNet at least once are reported. Additionally, Simpati reports the percentage of features in each pathway which are associated with the cancer type in the Human Protein Atlas. In order to compare the biological validity of the methods, these values were computed for netDx and Simpati signature pathways and only the most biologically relevant pathways were kept. Two criteria for retaining relevant pathways were tested: pathways having at least one key word associated in DisGeNet and pathways having more than 90% of features associated with the cancer type in the Human Protein Atlas. The number of pathways satisfying these constraints were compared and results are shown in Figure 4.

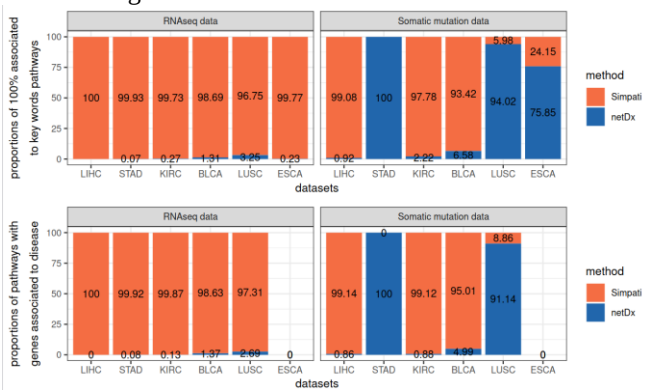


Figure 4. Biological validation comparison. The top box shows the relative proportion of signature pathways associated with relevant dataset key-words between the two methods and the bottom box shows the relative proportion of signature pathways associated with disease-type between the two methods.

This analysis highlights how Simpati is able to select biologically significant pathways directly associated with the patients it classifies and it performs better than the competitor.

IV. CONCLUSIONS

Simpati is a pathway-based classifier of patient classes for genetic data. It is the first classifier employing novel ad-hoc algorithms for PSNs to detect pathway-specific similarities. The tool is strongly centered around providing a good interpretability, as it provides signature pathways to unveil the altered biological mechanisms of a disease

phenotype. Thanks to a propagation algorithm that considers the interconnected nature of the cell's molecules, Simpati can classify dense, sparse, and nonhomogeneous genetic data. Future work will be focused on the development of strategies for the integration of multiple omics and on improving scalability for larger datasets.

ACKNOWLEDGMENTS

L.G developed the method, implemented and tested the software, wrote the text, and produced the images. C.M. tested the software and prepared the text. R. G. coordinated the project and revised the text.

This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement 814978 (TUBE), and by the JPND Personalized Medicine for Neurodegenerative Diseases project 334799 (ADAIR).

REFERENCES

[1] L. Jin et al., "Pathway-based analysis tools for complex disease: a review", Genomics Proteomics Bioinformatics, no. 12, pp.210-220, 2014.

[2] Y. Drier, M. Sheffer, and E. Domany, "Pathway-based personalized analysis of cancer", PNAS, no.110, pp.6388-6393, 2013.

[3] M. P. Segura-Lepe, H. C. Keun, and T. M. D. Ebbels, "Predictive modelling using pathway scores: robustness and significance of pathway collections", BMC Bioinformatics, no.20, pp.543, 2019.

[4] E. Lee, H. Y. Chuang, J. W. Kim, T. Ideker, and D. Lee, "Inferring pathway activity toward precise disease classification", PLoS Comput Biol, 2008.

[5] M. Yousef, E. Ülgen, and O. U. Sezerman, "CogNet: classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis". PeerJ Computer Science no7, pp336, 2021.

[6] J. Hao, Y. Kim, T. K. Kim, and M. Kang, "PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data", BMC Bioinformatics, no.19 pp.510, 2018.

[7] S. Pai, et al., "netDx: interpretable patient classification using integrated patient similarity networks", Mol Syst Biol no.15, 2019.

[8] S. Mostafavi, D. Ray, D. Warde-Farley, C. Grouios, and Q. Morris, "GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function", Genome Biology, vol.9, 2018.

[9] F. Fabris, D. Palmer, J. P. de Magalhães, and A. A. Freitas "Comparing enrichment analysis and machine learning for identifying gene properties that discriminate between gene classes", Brief Bioinform, no.21, pp.803–14, 2020.

[10] Simpati R package. [Online]. Available at: https://github.com/InfOmics/Simpati.

[11] Simpati GUI. [Online]. Available at: https://github.com/LucaGiudice/propaGUIation.

[12] D. H. Le, "Random walk with restart: A powerful network propagation algorithm in Bioinformatics field", 4th NAFOSTED Conference on Information and Computer Science, p. 242–247, 2017.

[13] M. Hofree, J. P. Shen, H. Carter, A. Gross, and T. Ideker, "Network-based stratification of tumor mutations", Nat Methods, no.10, pp.1108–1115, 2013.

[14] M. P. Segura-Lepe, H. C. Keun, and T. M. D. Ebbels, "Predictive modelling using pathway scores: robustness and significance of pathway collections", BMC Bioinformatics no.20, pp.543, 2019.

[15] J. Piñero et al., "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants", Nucleic Acids Res, no.45, pp.D833–9. , 2017.

[16] M. Uhlen et al., "A pathology atlas of the human cancer transcriptome", Science vol.357, 2017.

[17] Multiomic Integration of Public Oncology Databases in Bioconductor - PubMed n.d. https://pubmed.ncbi.nlm.nih.gov/33119407/ [retrieved May, 2021].

[18] M. Ramos, L. Schiffer, S. Davis, and L. Waldron. "TCGAutils: TCGA utility functions for data management", R package version 1.10.1, 2021.

[19] C. W. Law et al., "RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR", F1000Res vol.5, 2015

[20] A. Subramanian et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles", Proc Natl Acad Sci USA, no.102, pp.15545–50, 2005.

[21] M. Ashburner et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet, no.25, pp.25–29, 2000.

[22] M. Kanehisa M., and S. Goto. "KEGG: kyoto encyclopedia of genes and genomes", Nucleic Acids Res no.28, pp.27–30, 2000.

[23] R. Oughtred et al., "The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions", Protein Sci, no.30, pp.187–200, 2021.

[24] Simpati's Supplementary for results replication. [Online]. Available at: https://github.com/LucaGiudice/supplementary-Simpati.

 ISBN: 978-1-61208-971-3