USING DIGITAL TWINS IN HEALTH CARE

Adriana BOATĂ[1], Radu ANGELESCU[2], Radu DOBRESCU[3]

This paper aims to summarize the progress made in the use of the Digital Twin (DT) concept for choosing the right drug for a person, as well as presenting a subnet model that predicts existing drugs for a particular pathology, based on genetic expressions involved in the disease, and genes addressed by therapy. Finally, as a result of the research, the paper discusses how the proposed method can be applied to search for a medication for the disease caused by the Covid 19 virus, with reference to the results of some clinical trials conducted after the onset of the pandemic.

Keywords: Digital Twin, healthcare, molecular networks, gene expression

1. Introduction

The concept of Digital Twin (DT), which aims to create in a virtual environment a digital avatar of a real entity belonging to the physical environment, appears in the early 2000s as a solution for optimizing manufacturing processes and product life cycle. The use of DT is favored by advances in information technology, especially by the rapid evolution of IoT (Internet of Things), especially in its industrial version - IIoT (Industrial Internet of Things). In the use of DT, the essential is the real-time dual exchange of information between the physical entity and its virtual avatar, which requires a permanent communication between the virtual environment and the physical environment, with severe time restrictions (high data processing speed, low latency data block transfer) and requires a specialized interface for connecting the two environments, usually based on cyber-physical principles.

Gradually, the DT concept began to be used in other fields, including medicine and biology. In these two areas the use of DT has been mainly in scientific research, where DT has been perceived only as a simulation tool. The essential difference is that in this perspective the severe restrictions of real-time data exchange disappear, but it appears that the essential problem is to achieve a

[1] Master student, Department of Automatic Control and Industrial Informatics, Faculty of Automatic Control and Computer Science, University POLITEHNICA of Bucharest, Romania, E- mail: boata.adriana@yahoo.com

[2] PhD Eng., Department of Automatic Control and Industrial Informatics, Faculty of Automatic Control and Computer Science, University POLITEHNICA of Bucharest, Romania, E-mail: raduangelescu@gmail.com

[3] Professor, Department of Automatic Control and Industrial Informatics, Faculty of Automatic Control and Computer Science, University POLITEHNICA of Bucharest, Romania, E-mail: rd_dobrescu@yahoo.com

synchronization between the behavior of the physical entity (for example, an organism or a physiological system) and the dynamics of its virtual representation, which implies a precise tuning of the time scales at which the physiological processes take place. On the other hand, in medical applications DT must respect an important characteristic of the behavior of its industrial correspondent, that of the permanent exchange of data between the physical environment and the virtual environment.

In most situations, DT is used as a development environment that integrates and enables the running of digital processing algorithms borrowed from Artificial Intelligence techniques (eg. Machine Learning) and Exploratory Data Analysis (eg. Software Analytics) by creating digital simulation models that are updated as its counterparts in the physical world undergo changes. In this paper we used DT for graph representation of spatial networks, used as models for various types of biological networks (molecular networks, genetic networks) and we insisted on the advantages offered by DT in processing large volumes of data extracted from huge medical databases and to highlight the interactions between the constituent elements of these networks.

2. Integration of the Digital Twin concept in medicine

There are several instances in which DT is perceived in the medical field, but three of them can be considered truly representative:

1. The use of DT as a support for running medical software programs, both for assisting the diagnosis process (DT in the position of *virtual doctor*) and for therapeutic care (DT in the position of *virtual nurse*). The most difficult issue is the certification of such a support so that it becomes a regulated medical product. For this reason, in this case DT is used only in triage and screening processes, to facilitate selection in binary decision processes (eg. healthy / sick; positive test / negative test) before deciding to send a suspicious patient for examination by a specialist doctor. In addition, with the help of DT, statistical tests can be performed to assess the degree of confidence in the method used in triage.
2. The use of DT as a virtual modeling environment (VME) in terms of compatibility with the requirements of the Virtual Physiological Human (VPH) initiative, for the generation and integration of personalized knowledge about a particular patient who has been diagnosed with a particular disease. Specifically, in this case DT allows the clinical data provided by the current medical information systems of hospitals to be integrated with biometric and behavioral data from the patient's environment. It can be said that DT thus contributes to the

creation of an VPH Info structure that ultimately provides rule-based diagnoses to provide prediction regarding the evolution of the disease and personalized (patient-oriented) treatment.

3. The use of DT as a virtual framework that allows both legacy and sharing resources at different levels, as a foundation for the development of translational medicine. In this case DT is a platform located on one of the levels of a multi-level structure of hierarchical IT networks, contributing to offering IT support to ensure standardization and interoperability of information contained in large medical data repositories, based on evolved HPC cloud and grid computing and semantic web technologies.

The use of DT, from the perspective of the third position mentioned above, in medicine and biology research is the spearhead of *in silico* analysis techniques. Although studies *in silico* represent a relatively new research pathway, over the last decade they have been noted for predicting how drugs interact with the body and with pathogens. There are a wide variety of techniques in silico, but three of them are the most discussed: 1) Bacterial sequencing techniques; 2) Molecular modeling; 3) Simulation of the behavior of living cells, including the exchange of intra- and inter-cellular information.

In the thematic area of the paper (interactions in biological networks), the number of published papers is low, so we have nominated only three in which it is suggested to use of DT as an analysis tool. Thus, the paper [1] discusses the integration of DTs with Multi-Agent Systems (MAS) technologies and presents an application of agent-based DT to the management of severe traumas. Paper [2] explains how the performance of a model of the interactions of major organ systems can be tested by comparing the expected response predicted by DT and the observed patient response. The only paper that mentions that DT can produce predictive simulations of the spread of viral infection and the correspondent immune response to diseases caused by coronaviruses is [3], but the authors state "no current tools can predict… the most appropriate treatment for an individual COVID-19 patient", which is basically the objective of our work.

3. Use of DT in prescribing a medication scheme

DTs are currently the most important tools used in research related to the third direction. In this regard, in the following are presented details on the aspects regarding the use of DT as a tool for analyzing biological networks for establishing the appropriate medication for a disease with unknown treatment. The strategy recommended in the literature follows the methodology presented in [4] which is based on the creation of several virtual copies (DT entities) of the relevant factors in the treatment process of a certain disease which is then

simulated in the virtual environment. Fig. 1 shows this process, involving patient DT entities, disease DT entities and drug DT entities, which once introduced into the virtual environment, allow running a very large number of virtual treatment scenarios, with the hope that one of these scenarios will lead to the identification of the better performing drug. The method capitalizes on the advantages of DT in the ability to run a large number of scenarios, but does not guarantee a definite positive response, and the assessment of the performance of various drugs is qualitative (and implicitly subjective) and not quantitative

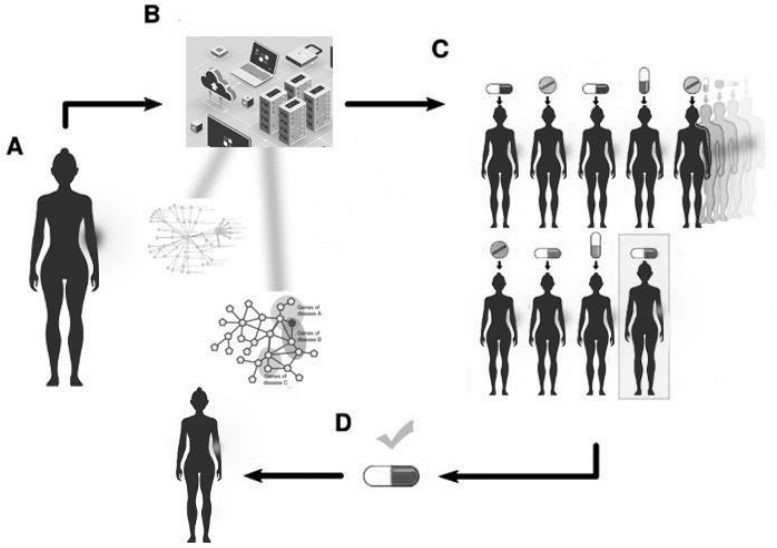

Fig. 1 The digital twin concept for personalized medicine (adapted after [4])

On the contrary, the strategy adopted in this paper proposes to resort to the techniques of modeling through biological networks of some interactions at molecular level, borrowing established procedures in the modeling through networks of complex systems. This approach allows to capitalize the results of the research carried out by using DT in the processing of biological networks that concentrate huge, massive amounts of data (Big Data), and thus fall into the category of complex systems.

Having in mind the complexity of human biological system, an appropriate model is likely to be multivariable, multidimensional and nonlinear. It is almost unanimously accepted that complex systems can be described and analyzed through network-type models that cover most of the requirements mentioned above. Several types of biological networks have been defined to cover the complexity of relationships and interactions between diseases, disease symptoms and drugs that can combat these diseases.

The Human Disease Network (HDN) is a powerful tool for establishing the association between disease pairs based on similarity criteria. Another

important network is the one that represents the diseases through the associated symptoms, called HSDN (Human Symptoms Disease Networks). In HSDN, the diseases represent nodes of the network, and the weight of a connecting line between two nodes (respectively two diseases) quantifies the similarity of the symptoms associated with each disease [5].

Significant progress in modeling interactions in biological networks has been made by clinical studies that have shown that genes that share similar phenotypes or that are characteristic of the same disease tend to encode proteins that interact with each other. Therefore, analysis of protein-protein interactions (PPIs) may clarify the relationships between diseases with similar (overlapping) clinical phenotypes. We can thus construct an extended HDN (eHDN) by combining information about a disease gene with information about PPI.

The paper [5] defined conceptually and formulated an algorithm for building a HSDN network. This is done in several steps, which will be briefly presented below, along with bibliographic references that signal significant changes that allow the improvement of the initial algorithm:

i) Extracting from the PubMed database the association relations between a certain disease and a symptom.

In [6] it is recommended to use for this purpose co-occurrence in the Medical Subject Headings (MeSH) metadata fields of PubMed.

ii) A HDN network is built, in which the nodes represent diseases, and the share of a link represents the similarity criterion between diseases.

iii) The integration of gene-disease associations is performed with the information extracted from the PPI databases in order to obtain associations between genes and PPIs that can be shared between different diseases. If the interaction takes place between directly connected proteins the binding pathway is of length 1 and is called order I link, and if the proteins are connected through an intermediate, the binding pathway is of length 2 and is called order II link [7].

iv) HDN is restored by adding as shares of the shared gene / PPi association and thus eHDN is obtained [8].

v) The backbone of HSDN is constructed by creating clusters in which a group of diseases share the same shared gene / PPI associations. By integrating eHDNs with weights of both disease-gene associations and given PPI, correlations can now be established between the degree of similarity of symptoms and the degree of joint exploitation of PPI and genes.

From our point of view, the further exploitation of the construct called the backbone of HSDN on the principle of a Scientific Data Pipeline (SDP) [9]. SDP consists of a sequence of steps that process sequentially through various functional units of input data sets to solve a specific well-defined problem. Applying SDP to the previously defined molecular biological networks we can both the interdependence of different diseases and the way of repositioning drugs.

At the same time, SDP ensures the reproducibility of experiments performed in various phases of research and facilitates the reuse of functional units by recombination, which can lead to the development of an improved treatment scheme, either by selecting a more effective drug or by defining a combination of drugs.

4. Proof of concept – example of a sub-network integration in HSDN

The algorithm presented above for generating the disease network (part of HSDN) can be replaced by DisGeNET version 7.0, a platform containing large collections of genes associated to human diseases [10] and also LINCS L1000 dataset [11], containing gene-expression signatures for various drugs and small molecules to generate the prediction system.

We provided an algorithm that does disease prediction and one algorithm for drug recommendation that takes a set of gene expression profiles as input. We then test the hypothesis of disease prediction and drug recommendation over a set of gene expression data from GEO (Gene Expression Omnibus) database, using results from existing papers.

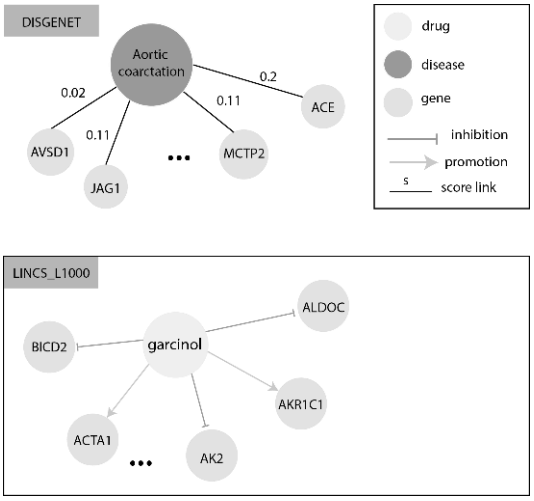

Fig. 2 Example of basic information from the two sources

In fig. 2 we present the basic information extracted and used from the two data sources with an example for the disease Aortic coarctation and the drug garcinol. Note that the scores present in this DisGeNET illustration are the scores computed in [10]: seven weights are used. For example, $W_{UniProt}$ is 0.3 if the association is reported in the UniProt database (https://www.uniprot.org/) and 0 if not. The last three weights correspond to the literature (GAD, LHGDN / BeFree). As it can be observed, the drug inhibits the activity of BICD2, ALDOC and AK2 and promote the activity of ACTA1 and AKR1C1.

After constructing the two algorithms, we test them on a set of data that contains gene expressions from breast tumor tissues – a model downloaded from GEO that was created using data from tumor and normal tissue, sequenced using DNA microarray. We split the input data in two batches: perturbed and control gene expressions to identify the DEG (differentially expressed genes). To be able to use the information provided in data sources on gene expression data collected from humans in real conditions, we need to convert gene expressions in a list of significant genes. We applied filters and LIMMA [12] to sort the genes into down regulated genes and up regulated genes. There are multiple algorithms available, but this one is considered the most efficient in eliminating the noise from data. Inputting the altered data into LIMMA we get a list of significant genes with their p-values. We then use the p-values in conjunction with their logFC (log2foldchange) values. It should be mentioned that from a biological point of view, the activation process of genetic products has two hypostases: downregulation and upregulation. Downregulation is the process by which a cell decreases, in response to an external stimulus, the quantity of a cellular component, such as protein or RNA. The complementary upregulation process involves the increase of such cellular components. The majority of differentially expressed genes are downregulated during malignant transformation. Describing the whole process of preparing the data does not constitute the subject of this paper and will be detailed in another one.

To generate the best drug list, that may bring the gene expression back to the normal state, we apply a scoring mechanism where the basic idea of the score is to match as many down regulated DEG with genes that are up regulated from LINCS, and up regulated DEG with downregulated genes for drugs. After picking the highest score drug, we remove the genes it regulates and rerun the scoring method, picking the next highest score on the remaining genes. This can be applied any number of times, based on how many drugs we want our treatment to have. Computing the drug score is straight forward, after splitting the DEG in 2 sets, one for up regulated genes.

To test the methods presented we used ‘GSE15852’ dataset created by IB et al. (2010) from GEO T et al. (2013) which is formed from expression data from human breast tumors and their paired normal tissues. This dataset provides gene expression for mammary tissues suffering from cancer and their healthy counterparts. The summary states that they identified a set of 33 significant differentially expressed gene expression patterns for 43 breast tumors. After filtering the data and applying LIMMA we do the volcano plot (Fig.3) and mark the top 30 most significant genes (picked a number like what the paper found).

The plot in Fig.3 shows the statistical significance, known as P value, versus the fold change FC (magnitude of change). It is possible to easy select by visual identification statistically significant genes, which have higher FC. These

plots are commonly used in *omics* experiments such as genomics or proteomics, in order to highlight the significant changes. Using the logFC values alone has been demonstrated to not be reliable in detecting significance, because of noise present in data, and using only p-values does not help us in determining the way the gene expression varied.

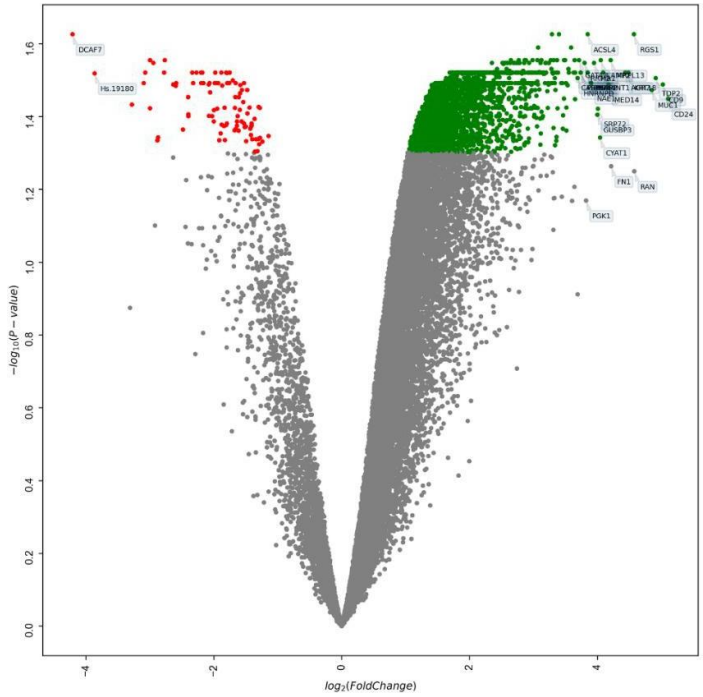

Fig. 3 GSE15852 experiment volcano plot

Running the disease prediction, we get the top 3 most likely disease results: Malignant neoplasm of breast (score: 118.57), Glycogen Storage Disease Type IIb (score: 116.0) and Breast carcinoma (96.36). So, the most-likely disease is Malignant neoplasm of breast. Running the second method to generate a two-drug combination therapy gives us vorinostat (score 9) + trichostatin (score 8) – two drugs already used with good results in the treatment of breast cancer, according to literature available.

5. Using DT for analysis of molecular networks in COVID-19 context

Without being directly linked to the DT concept, many recent research have focused on molecular network analysis to urgently find a suitable treatment for the fight against coronavirus 2019-nCoV/SARS-CoV-2, considering that such solution may appear much faster than developing and accepting a vaccine (which we now know it was not the case). We will present below how the method

described above for using DT in HDSN construction for the selection of a suitable drug based on IIP analysis. First, however, we will present how the protein-to-protein interactions are carried out through the multilayer modules, in order to identify and validate disease modules. To standardize the scientific language, we will refer to the methodology proposed by the research team coordinated by A-L. Barabasi [13], who considers that in the molecular networks of interactome type can be found three types of modules: topological, functional and disease. Topological modules contain nodes that tend to interact with each other with priority, and not with nodes outside the module. The functional modules correspond to the network neighborhoods in which the nodes located in the same network that are involved in strongly connected cellular functions tend to interact with each other with priority. Disease modules are groups of nodes in which a change (eg mutation or change in gene expression) is always caused by a phenotype of the disease. The hypothesis accepted in this paper is that in complex biological networks these three types of models can overlap, in the sense that pair associations can be created between a topological and a functional module, and that a disease module can cause alteration or even destruction of a functional module. We can now explain how a human interactome developed under DT technology can be used in the particular case of Covid-19 infection. We will consider three types of protein-protein interactions: viral-human PPI (occurring in the viral interactome), human-human PPI, and drug-human PPI (both occurring in the human interactome). The working hypothesis is that there is no connection between the viral interactome and the human interactome, and as such there is no interaction between drug and virus, the connection between virus and drug module being made indirectly through the human interactome

In addition to the work already mentioned ([13]), we point out 3 other papers that use this methodology to find a suitable medication for Covid-19. Thus, in [14] the implementation of a platform based on a pharmaceutical drug network, which quantifies the interactions of the host virus with target drugs in the human PPI network. The authors of the paper [15] propose the selection of a subset of human proteins that can bind to approved drugs, based on information about the interaction of human proteins and viruses. Finally, in [16] results are published on the efficacy of highly popular drugs in clinical trials (Remdesivir and the association Lopinavir / Ritonavir). Referring to these papers, we specify that the main original contribution of our paper is how the DT platform uses the criterion of network proximity between the drug target and HCoV-associated proteins

6. Conclusions

This paper aims to extract some of the international achievements over the last decade in medical research based on information networks integrated in a formal framework that allows the use of a series of quantitative approaches and

predictive tools to study pathogen - human host interactions, aimed at revealing the molecular mechanisms of infection, identifying comorbidities, and rapidly detecting drug candidates for appropriate treatment.

In particular, we presented the most effective possible drug selection scheme through a mechanism for highlighting the proximity of the nodes of a network of protein-protein interactions that integrates two subnetworks in a complex network of human interactions (called *human interactom*). The limited resources did not allow to obtain quantifiable results and moreover a concrete action in this respect can only succeed through the work of a large team, in a specialized research center.

R E F E R E N C E S

[1]. *Croatti, A., Gabellini, M., Montagna, S., Ricci, A.*, "On the Integration of Agents and Digital Twins in Healthcare", Journal Medical Systems, 44, 161, pp. 1-8, 2020

[2]. *Lal, A., et al.*, "Development and Verification of a Digital Twin Patient Model to Predict Specific Treatment Response During the First 24 Hours of Sepsis", Critical care explorations, vol. 2, iss.11, e0249, 2020

[3]. *Laubenbacher, R., Sluka, J., Glazier, J.*, "Using digital twins in viral infection: Personalized computer simulations of infection could allow more effective treatments", Science, Vol. 371, Iss. 6534, pp. 1105-1106, 2021

[4]. *Björnsson, B. et al., "*Digital twins to personalize medicine, Genome Medicine", 12, pp. 1-4, 2020

[5]. *Zhou, X., Menche, J., Barabási, A-L., Sharma, A., "*Human symptoms–disease network", Nature Communication, 5:4212, 2014.

[6]. U.S. National Library of Medicine, MESH website, url: https://meshb.nlm.nih.gov/search, 22019.

[7]. *Saha, N. et al.,* "miRwayDB: a database for experimentally validated microRNA-pathway associations in pathophysiological conditions", Database, Vol. 1-2018, 2018.

[8]. *Barabási, A-L., Gulbahce, N., Loscalzo, J.*, "Network medicine: a network-based approach to human disease", Nature Reviews Genetics, 1, pp. 56–68, 2011.

[9]. *Angelescu, R., Dobrescu, R.*, "Hyperbolic embedding model for a class of microRNA-disease networks, U.P.B. Sci. Bull., Series A, Vol. 72, Iss. 1, 2020

[10]. *Pinero, J. et al.,*The disgenet knowledge platform for disease genomics. Nucleic Acids Research, 2020

[11]. *Subramanian, A. et al.,* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles", Cell, 171(6), pp. 1437-1452, 2017

[12] *Sendama, W.* "L1000 connectivity map interrogation identifies candidate drugs for repurposing as SARS-CoV-2 antiviral therapies", Computational and Structural Biotechnology Journal, 18, pp. 3947-3949, 2020

[13]. *Gysi, D. M. et al.*, "Network medicine framework for identifying drug-repurposing opportunities for COVID-19", Proceedings of the National Academy of Sciences, 118 (19), 2021

[14]. *Zhou, Y., Hou, Y., Shen, J. et al.* "Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2", Cell Discovery, 6(14), 2020

[15]. *Habibi, M., Taheri, G.*, "Topological network based drug repurposing for coronavirus 2019", PLoS ONE, 16(7): e0255270, 2021

[16]. *Senger, M. R. et al.*, "COVID-19: molecular targets, drug repurposing and new avenues for drug discovery", SciELO Brazil, 2020, accessible at https://www.scielo.br/j/mioc/a/FT4WXwhbCPCkm7W6XdX4GHm/?lang=en