BOHR

BOHR Journal of Biocomputing and Nano Technology
2024, Vol. 2, No. 1, pp. 11–14
DOI: 10.54646/bjbnt.2024.14
www.bohrpub.com

RESEARCH

Integrating induction and deduction in order to complement experiments and priors for the understanding, forecast, control, and foresight of apparently complex processes in molecular biology's interactomics

Diego Liberati*

National Research Council of Italy @ Politecnico di Milano University, Milan, Italy

***Correspondence:**
Diego Liberati,
diego.Liberati@cnr.it

Received: 22 October 2023; **Accepted:** 04 May 2024; **Published:** 05 August 2024

A developed battery of Systems and Control tools to manage complex processes is discussed, with application to a molecular biology example, namely, genes–proteins interactomic analysis within the cell: a kind of molecular complex bio-social networked communicating system. Such Systems and Control tools are shown to be a great help to understand the apparently hidden relationship behavior among agents–namely, genes and proteins–in order to provide consistent help to molecular biologists and physicians who may have to decide how to impact the interactomic.

Keywords: interactomic regulation networks, community analysis, hybrid dynamic-logic systems, principal component analysis, cancer modeling

1. Introduction

Biomedicine is less and less a matter of evidence—letting the so-called "clinical eyes" implicitly infer causal reasons behind epiphenomena—than just a matter of axiomatic deduction.

In fact, each of us gets old, which is itself an illness (as already Lucius Annaeus Seneca, Nero's preceptor, was stating: but the alternative is to die young—one of my maternal grandfather's wisdom citations!), and get old in one's own peculiar way—not just as a corollary of the above-mentioned "Seneca's theorem."

When an illness happens, each individual, with her/his own DNA, does try to repair what is wrong. Even monozygotic twins, sharing identical DNA, exhibit epigenetic differentiation ever since, being, for instance, the position in the uterus physically different—no matter the equal identification that a mother does feel with each of her children.

On average, illness does occur more often when getting older, forcing the ill person to start malfunctioning in her/his own way, thus, needing a kind of "LEGO co-player with God" (as a metaphor of the famous Einstein's quotation "God does not play dice") to help to repair.

In order to hope to do that, the therapist should have a kind of design, like a scientist, or at least an engineer, would employ: just employing a simple trial and error approach would often not be sufficient! Inductive—deductive data mining and modeling.

Thus, a model of the analyzed system is needed, describing in mathematical terms the main interacting features without going into too many details—that would not be useful for the specific purpose—letting the dimension of the model be still tractable; traditional physiological studies are much like that.

On the other side, one could resort to this century's fashion of inference from data, because of the growing amount of available data. The main drawback of such an approach is that often results are able to discriminate cohorts, thus helping in differential diagnosis, but do not exhibit any intelligible knowledge of what is happening inside the investigated system, nor they are usually able to mimic our natural "wet" neural approach of learning by examples how to control the underlying processes towards the desired goals. Moreover, they usually need heavy computation power, at least in the learning phase. Briefly, often even the fashionable and powerful deep learning does learn how to do it but is much less good than us in then explaining why it acted like it did.

Logical networks (1) do overcome the said drawbacks by inferring in the canonical OR of ANDs form of electrical circuits fashion (2), namely, in deduction form, then easily keen to introduce priors just *a posteriori*, by modifying inferred rules (think about inferring the need to run at no more than 45.96 km/h to avoid tickets, it could probably be *a posteriori* set at 50 by enforcing an obvious prior to whom the inference was asymptotically tending).

When data are available not just as single shots but in time course, a Piece-Wise Affine identification approach—within a hybrid dynamic and logic framework—does generalize the above by optimally cutting even complex non-linear multivariable processes with hysteresis in time periods of almost linear behavior among almost stationarity borders (3). Then, within each linearly identified region, even simple approaches like principal components—carefully revisited in conjunction with k-means clustering—could then allow the identification of a hierarchy among the conditioning factors by considering each salient factor's influence on the main discriminating principal component (4). Such a portfolio of tools thus does appear to be keen to help study the complex molecular interplay within the cells. In particular, we are here interested in the genes–proteins called interactomics: genes are known to codify for proteins, while, in turn, proteins are regulating gene expression in the powerful feedback system on which our life is grounded. Such a powerful, quite general-purpose portfolio of tools has even proved to be able, for instance, to help in fostering research on quite complex scenarios, like, for instance, attenuation in mobile communications by rain interference, providing on the other hand, a toll to improve local weather forecasting (5).

Even popular pioneering discriminations about pathologies (6) could then easily be outperformed (4), even discovering a classification error in the data repository.

Hypotheses could then be tested as not falsified by key experiments, as in confirming a new WNT pathway in leukemia discrimination (7). And, finally, to downsize at the molecular level scale, inter-domain competition analysis and simulation could even predict not yet discovered mutant of Sos oncosuppressor, then discovered after the paper was submitted. Such a portfolio of tools thus does appear to be keen to help study the complex molecular interplay within the cells. In particular, we are here interested in the genes–proteins called interactomics: genes are known to codify for proteins, while, in turn, proteins are regulating gene expression in the powerful feedback system on which our life is grounded.

It is worth here to remind that the powerful though simplistic idea that every gene just codifies for its own protein—then free to interact within the cell—is nowadays over ever since. It would not account for most of the well-known epigenetic properties, modulating gene expression to the different contexts, within, for instance, different organs and/or different individuals. This is why the investigated feedback system is at least multivariable, usually non-linear, and sometimes it does even exhibit hysteresis.

Networks of interactomic actors, including genes and codified proteins themselves, are nowadays known, though not yet all completely identified, as responsible for their exhibited complex quite para-social interaction, resulting in the beautiful diversity, within similarity, of key factors of life. By the way, the same happened years ago, at a different scale, within the central nervous system's neuroscience, when the so-called "grand mother" neuron, formerly believed to be responsible for memorizing the beloved, was substituted by the task recruited by the natural "wet" neural network, including several actors, each of which was, in turn, still available to contribute to other tasks within (sometimes only partially) different other natural neural networks. Still the same, at an even bigger scale, does in fact appear—under this respect—everyday everybody's social multi-interactions experience of each homo oeconomicus of us.

2. Methodology

In order to investigate such a kind of interactomic network, a simple but powerful idea, as proved by Google's usefulness and consequential success, is to investigate the ranking of each actor's relationships to each other, being such actor either an internet page, as in the original Page's algorithm, or a gene (or codified protein) in our case.

It is worth noticing, as pointed out in a seminal review by Vidyasagar (8), that the recent randomized approach introduced by Ishii and Tempo (9) would be the technological key to drastically reducing, at the cost of a limited loss of precision, the overwhelming computational complexity that would prevent applying Page ranking to the analysis of every significant interactomic network, besides the almost-toy subnetworks already investigated.

A possible complementary so-called "community approach," proposed by Landi and Piccardi (10), could also be taken into account for our interactomic regulation networks.

As a publicly available benchmark, the data (11) can be used in order to investigate which features of the alternatively proposed approaches are possibly useful as a complement, in the hope to even improve the already powerful Page approach on one side or eventually be able to surrogate it in a less costly way, even taking into account the recalled randomized economy. A probably even better solution will be, then, some improvement on randomized Page ranking that we could even make with Ideaky Ishiii, also in memory of our beloved colleague Roberto Tempo, both coauthors of the seminal paper introducing randomized Page ranking (9).

3. Results

A first trial, to test alternative approaches to Page ranking, was obviously done over a couple of public data sets, as described in Gavin et al. (11), already used investigating them via the randomized Page ranking. Such data sets describe protein interactions in *Saccharomyces cerevisiae*, namely:

1. PPI-D1 data set, obtained via mass spectrophotometry, including 1430 proteins linked by 6531 interactions;
2. PPI-D2 data set, combining six different experiments with hybrid techniques, including 3869 proteins linked by 23,399 interactions.

The recorded interactions are weighted in order to sharpen information, by pruning the non-trustable information via adjusted weights, under the assumption that proteins with the same neighborhoods usually should share similar functions. For the two considered data sets, such correction yields: for PPI-D1 to 990 proteins with 4687 interactions, while for PPI-D2, the network results in 1443 proteins with 6993 interactions, thus now of comparable dimension and, probably, complexity.

4. Discussion

A few approaches fully described elsewhere have been considered. For the sake of completeness, some of their salient features are nevertheless recalled within the next few sentences.

Performances of the various approaches are evaluated and compared in the bi-dimensional space of so-called "Recall and Precision," taking into account a combination of positive and negative false rejections, namely, the usual type A and type B statistical errors.

By considering Precision and Recall for several approaches on the used data sets, among the four approaches reminded in the following, one of them [in-out- and pseudo-community analysis by (10)] generally deserves good results, yielding the subsets of salient communities confirming our priors.

Interactomic regulation networks can thus be seen as a special case of the more general network community analysis, quite popular nowadays in the interdisciplinary field of complex systems, at the edge, among other disciplines such as physics, control theory, and operating research.

Four approaches have mainly been taken into account: two of them, Louvain and LMC, are based on network partition, while the one, IOPC, proposed by Landi and Piccardi (10)—and based on a quasi-local search allowing partial superposition of communities—looks more general, usually outperforming the two other approaches as for both Recall and Precision. The fourth approach is, of course, the discussed Page ranking, possibly randomized, sometimes comparable to IOPC, as, for instance, on one of the two used test networks, while, for the other one, the loss of Recall with Page rank with respect to IOCP is dramatic. Among such four approaches, considered the best ones among the many other approaches available in the literature, the very best, on the pair of benchmarks employed, appears to be IOPC, very strict for false positives, thus probably even improvable by pruning or merging the smallest found communities.

Other further possible improvements may imply taking into account the time evolution of the interaction graph (like PRISM) in order to also take into account the dynamics of interactions. This would allow us to deal not only with the identification of the salient variables but also with their dynamical interplay, like in (3), thus leading to the so-called path defining the destiny of the considered interactomic interaction.

Improvements in either Louvain or LMC, in the direction of not forcing partitions anymore, could make their performances comparable with IOPC, while Page ranking, already often not worse than IOPC, is the simplest to be randomized, offering thus advantages in speed as well as being keen to be improved, as discussed with its still-lived inventor Ideaki Ishii.

All the offered results are thus in some sense qualitative, or at most semi-quantitative, offering, at the present stage, only figures of merit to judge their usefulness: they are just able to provide the involved actors, but not yet the (dynamical) weights balancing their interaction. To become fully quantitative, one should probably resort to ideas like the ones able to mine, for instance, linear, binary (1), inferential, or even hybrid (3) quantitative models learned from richer data, weighting, even dynamically, the arcs of the interactomic graphs, and thus becoming able to generalize.

5. Conclusion

We have proposed just a simple example at the molecular scale for the considered Systems and Control tools: obviously we would not dare to aim to vicariate the powerful traditional approaches based on Schroedinger equations for solving atomic interactions in molecules, but complementing such a powerful approach with a battery of little instruments could be helpful, especially when molecules are investigated not as deeply in detail as until their atomic level, but at a higher scale, namely, for instance:

either their subdomains;

or their catalytic and/or allosteric interaction with their neighborhoods;

or their two ways of interaction with the genome in cells, as here of interest, being the genome a code to have the cell-making proteins, but, in turn, being proteins, in this context, facilitators or inhibitors of the genome task within the cell cytoplasm.

We are then confident that similar approaches, possibly revisited, could help to complement others in deeply investigating fundamental interaction in different contexts, from human–robot interaction to high energy physics like LCC does allow, to cite just a couple of examples in which we are involved.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

Carlo Piccardi introduced us to community analysis, Mariani and Nicoletta provided preliminary results in fulfillment of their Master's thesis in Mathematical Engineering at Politecnico di Milano University.

References

1. Muselli M, Liberati D. Binary rule generation via Hamming Clustering. *IEEE Trans Knowledge Data Eng.* (2002) 14:1256–68.
2. Muselli M, Liberati D. Training digital circuits with Hamming Clustering. *IEEE Trans Circuits And Systems I.* (2000) 47:513–27.
3. Ferrari Trecate G, Muselli M, Liberati D, Morari M. A clustering technique for the identification of piecewise affine systems. *Automatica.* (2003) 39:205–17.
4. Garatti S, Bittanti S, Liberati D, Maffezzoli A. An unsupervised clustering approach for leukemia classification based on DNA micro-arrays data. *Intelligent Data Analy.* (2007) 11:175–88.
5. Formentin S, Luini L, Capsoni C, Nebuloni R, Liberati D. Modeling rain fields for earth space propagation applications by an autoregressive modeling approach. *Proceedings of the 8th Advanced Satellite Multimedia Systems Conference and 14th Signal Processing for Space Communications Workshop, ASMS/SPSC.* Mallorca: (2016).
6. Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science.* (1999) 286(5439):531–7.
7. Grassi S, Palumbo S, Mariotti V, Liberati D, Guerrini F, Ciabatti E, et al. The WNT Pathway is relevant for the BCR-ABL-1 independent resistance in Chronic Myeloid Leukemia. *Front Oncol.* (2019) 9:e532. doi: 10.3389/fonc.2019.00532
8. Vidyasagar M. Probabilistic methods in cancer biology. *Eur J Control.* (2011) 17:483–511.
9. Ishii I, Tempo R. Distributed Randomized Algorithms for the PageRank Computation. *IEEE Trans Autom Control.* (2010) 55:1987–2002.
10. Landi P, Piccardi C. Community analysis in directed networks: In-, out-, and pseudocommunities. *Phys Rev E.* (2014) 89:012814.
11. Gavin A, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature.* (2006) 440:631–6.
12. Abada A, Liberati D, et al. FCC-ee: the lepton collider: future circular collider conceptual design report volume 2. *Eur Phys J Spec Top.* (2019) 228:261–623.
13. Babiloni F, Carducci F, Cerutti S, Liberati D, Urbano A, Babiloni C. Comparison between human and artificial neural network detection of Laplacian-derived electroencephalographic activity related to unilateral voluntary movements. *Comput Biomed Res.* (2000) 33:59–74.
14. Fantini P, Maffezzoni P., Daniel L., Liberati D, Taisch M: in preparation, 2024
15. Hanna EM, Zaki N. Detecting protein complexes in protein interaction networks using a ranking algorithm with a refined merging procedure. *BMC Bioinformatics* (2014) 15:204.
16. Pagani M, Mazzuero G, Ferrari A, Liberati D, Cerutti S, Vaitl D, et al. Sympathovagal interaction during mental stress. A study using spectral analysis of heart rate variability in healthy control subjects and patients with a prior myocardial infarction. *Circulation.* (1991) 4:43–51.