Interdisciplinary Relationships Between Biological and Physical Sciences

Paulo E. P. Burke\textsuperscript{1} and Luciano da F. Costa\textsuperscript{2}

\textsuperscript{1}Bioinformatics Graduate Program, University of São Paulo, São Carlos, SP, Brazil
\textsuperscript{2}São Carlos Institute of Physics, University of São Paulo, PO Box 369, São Carlos, 13560-970, SP, Brazil

Abstract

Several interdisciplinary areas have appeared at the interface between biological and physical sciences. In this work, we suggest a complex network-based methodology for analyzing the interrelationships between some of these interdisciplinary areas, including Bioinformatics, Computational Biology, Biochemistry, among others. This approach has been applied over respective data derived from Wikipedia. Related reviews from the scientific literature are also considered as a reference, yielding a respective bipartite hypergraph which can be used to gain insights about the interrelationships underlying the considered interdisciplinary areas. Several interesting results are obtained, including greater interconnection between the considered interdisciplinary areas with biological than with physical sciences. A good agreement was also found between the network obtained from Wikipedia and the interrelationships revealed by the literature reviews. At the same time, the former network was found to exhibit more intricate relationships than in the hypergraph derived from the literature review.

1 Introduction

Scientific research has become increasingly interdisciplinary \cite{1,2}. It also has been suggested that interdisciplinary research tends to be intensely cited \cite{3}. Among all combinations of disciplines, the relationship between biological and physical sciences holds particular interest because of its potential impact on society.

It is particularly challenging to trace back the origin of the interplay between biological and physical knowledge, perhaps as a consequence of the intrinsic relationship between these two areas, which tend to make almost any related research multidisciplinary. Nevertheless, we may consider as an important mark the application, in the 19th century, of chemistry techniques to the study of proteins\cite{4}, which had been just discovered. The term “Biochemistry” was then coined in order to name the intersection of chemistry and biology.

Another important starting-point that can be highlighted consists in the integration of mathematics and biology developed by Alfred J. Lotka and Vito Volterra’s work on population dynamics in 1926, commonly known today as the predator-prey model \cite{5}. Decades later, the now well-established field of Biochemistry relied on the integration of concepts from several other areas. For example, in the early 60s computational approaches powered the elucidation of amino-acid sequences and three-dimensional structures of proteins\cite{6}. At that moment, the most common term used to identify the application of mathematical and computational techniques to biological problems was “Computational Biology”. Though still adopted today, this terminology has been almost subdued by the term “Bioinformatics”, which was in great part motivated by the advent of DNA sequencing \cite{7} in 1977, and further, to the Human Genome Project\cite{8} in 1990. Nowadays, some
researches claim it is nearly impossible to achieve substantial scientific advances in biology without a strong computational basis\cite{9, 10}. In fact, the outcomes from such interactions have been central to answer fundamental questions of biology\cite{11, 12} and to push forward applied fields such as precision medicine\cite{13, 14}.

Given the fast growth and importance of the interface between biological and physical sciences through areas such as Biochemistry, Computational Biology, and Bioinformatics, it becomes important to understand how these areas interrelate one another. Though we could try understanding what is meant by each of these areas simplistically from their names (e.g., biochemistry as corresponding to chemistry in biology), the fast changes characterizing the development of modern science, and the appearance of new interfaces between branching subareas, makes a precise definition of the considered multidisciplinary areas very difficult. Even so, efforts toward understanding, even in approximate terms, the meaning of existing multidisciplinary areas remain a critical issue for several reasons, including the need to continuously organizing knowledge and results, properly identifying interfaces and sub-areas, more effectively index the related literature, orient programs of study, among many other possibilities.

Many are the works which aim at understanding how scientists manage interdisciplinarity\cite{15} and what is its impact on science\cite{16}. However, we lack approaches that can treat the organization of the interaction between different areas of knowledge per se. Fortunately, due to the development of new information science concepts and methods, such as in complex networks\cite{17} and scientometry\cite{18, 19, 20}, it becomes possible to develop quantitative methods capable of revealing some of the main aspects of several inter and multidisciplinary areas through data. In this work, we propose a network approach to estimate knowledge interaction between scientific areas based on a given literature database. More specifically, we create a citation network based on articles related to areas of interest from a given database and their respective references. Then, we generate a higher-level network which captures the relationships between these areas based on the former citation network. As a case example, we use Wikipedia articles relative to biological and physical topics, as well as interdisciplinary areas which derive from them, in order to understand how the knowledge of these areas interact one another. We also perform a traditional literature review of the mentioned topics in order to have a reference for comparison.

This paper is organized as follows. We first describe the general idea of constructing article networks and how we derive knowledge networks from it. Then, we present the Wikipedia database which is used to perform our analysis. The outcomes of applying our methodology to the Wikipedia database are then analyzed and discussed. At last, we present some conclusions about this work and possible future applications.

2 Methodology

Knowledge is mainly perpetuated through its register on some media that can be propagated and referred. Until now, the most efficient method found to store knowledge is to write it down in the form of books or articles, in paper or digitally. Also, it is very likely that most of the registered knowledge derives from, or relates to, some previous knowledge or information. In scientific literature, as well as in other areas, the relationship between pieces of literature are registered in the form of references between texts. Thus, the web formed by texts and references between them can unveil a higher-level structure of knowledge.

In order to assess this web structure of knowledge, let us consider a collection of texts associated with a given subject. These texts, additionally to their contents, have references to other related texts, belonging or not to this collection. With articles and references, we can then construct an article network, following the procedure detailed in Section\ref{sec:nets}. Then, we can estimate the knowledge interactions between those subjects by grouping their respective articles. To avoid some unwanted biases, we normalize and group the nodes following the procedure described in Section\ref{sec:aggreg}.
2.1 Network Construction

Networks have been used to model and analyze several kinds of data, from social interaction to transportation. Essentially, networks are composed of two elements: vertices and edges. Real-world data involving discrete objects can be mapped into networks by assigning entities to vertices and using links to represent relationships between the entities. In the case of scientific literature, we can map texts into vertices and respective hyperlinks or bibliographic references as edges. In mathematical terms, our network can therefore be described as $G = (v, e)$, where $N$ is the total of texts, $v = \{v_1, v_2, \ldots, v_N\}$ are the vertices, and $e = \{(v_i, v_j), (v_k, v_l), \ldots, (v_m, v_n)\}$ are edges indicating respective relationships between the texts.

2.2 Normalization of Categories’ Sizes by Sampling

The number of articles contained in each of the categories of interest can substantially vary in magnitude. This variance can be a consequence of several factors, including area specificity, existence time, and funding, among others. This variation implies several difficulties while analyzing relationships between the considered areas. For instance, the links between two larger areas would tend to appear in a more significant number than between two areas with a relatively smaller number of works. Thus, if one is interested in quantitatively comparing the relationship between any two different areas, it is necessary to normalize concerning the size of the areas. In order to minimize this potential bias, we sampled the network while selecting a fixed number of vertices for every one of its areas, as exemplified in Figure 2. Observe that the number of selected vertices must be smaller than the size of the category containing the smallest number of articles. Next, the articles so selected within each area were collapsed into a single vertex, resulting in a subsumed network where the number of vertices is equal to the number of areas. The frequency of connections between two areas, i.e. the external edges, is now represented as the weight of the edge which connects their respective vertices, while the frequency of connections within articles of the same area, the internal edges, is stored as a node weight. After sampling the network several times, an average of the obtained networks will result in the normalized network of knowledge. Finally, the weight of each edge/node was divided by the sum of the weights of all edges/nodes in order to the result become invariant to the number of selected nodes per area.

2.3 Article Database

Wikipedia is currently the most extensive collaborative encyclopedia. Nowadays, the English version contains around 5.6 million articles with subjects ranging from science to popular culture, history, art,
Figure 2: Illustration of the network sampling procedure. In this example, four nodes from each area are randomly selected while the links between them are kept. The resulting collapsed network has its links weighted by the total connections between areas yielded by the original nodes.

We selected a subset of articles from Wikipedia where their content refers to biological and physical sciences topics. More specifically, we selected articles originally assigned to the following categories: Chemistry, Mathematics, Applied mathematics, Dynamical Systems, Computer Science, Statistics, Engineering, Biomedical Engineering, Biology, Ecology, Medicine, Health Sciences, Molecular Biology, Bioinformatics, Biochemistry, Computational Ecology, Biotechnology, Systems Biology, and Computational Biology. Observe that these categories correspond to respective tags available directly from Wikipedia. Those categories were chosen in order to cover the main areas within biological and physical sciences as well as interdisciplinary areas, namely Bioinformatics, Computational Biology, Biomedical Engineering, Biochemistry, Computational Ecology and Systems Biology, which are intrinsically related to those two major areas. Given that Wikipedia’s categories are hierarchically organized, we also selected subtopics at one hierarchical level downwards from the categories mentioned above. The total number of articles available in each category, considering their respective subtopics, is depicted in Fig. 3.

Figure 3: Number of articles in each of the considered categories.

3 Results

Relationships between scientific areas can be estimated in several ways. Here, we used Wikipedia’s articles as a means of quantitatively measuring the interactions between biological and physical areas, as well as interdisciplinary fields which might be a consequence of the interplay between these two major areas. In order to do so, we obtained articles from Wikipedia’s database which were already assigned to labels referring to the scientific areas listed in Section 2.3 as well as their respective sub-categories within one level in depth. However, some categories, namely Biomedical Engineering, Systems Biology, Statistics, Computational Biology, and Computational Ecology, were excluded from the analysis due to their very small number of articles when compared to the others categories, as observed in Figure 3.

We constructed the article network which con-
tained 19,400 Wikipedia articles connected by 131,657 hyperlinks. To generate the knowledge network, we normalized the connections between areas following the procedure described in Section 2.2 by sampling 200 nodes of each area, and taking an average over 10,000 samples. We also sampled the network using 100 and 300 nodes per category, and no significant difference in the results was noticed (data not shown). The obtained knowledge network, depicted in Figure 4a, is fully connected. However, its strongest edges (highlighted in orange) show a well-bounded community structure forming two groups. The group on the left is composed solely of physical sciences, namely Mathematics, Applied Mathematics, Dynamical Systems, Computer Science and Engineering.

On the other hand, the group on the right mixes all biological sciences, interdisciplinary areas, and Chemistry. This emergence of two groups indicates a substantial level of separation between physical and biological sciences. Interestingly, despite all interdisciplinary areas being strongly connected to biological sciences, they display different connecting patterns. In Figure 4b, we can observe that Biochemistry has most of its connections pointing to Chemistry, Molecular Biology, Ecology, and Biotechnology. However, even Bioinformatics also being strongly connected to the second group, in Figure 4b we can observe that it has more connections with the physical group than the former area, being especially connected with Applied Mathematics and Computer Science.

Regarding the balance between internal and external edges of each area, all fields have more internal than external edges. However, a trend can be observed in Figure 4c where some areas have a higher percentage of external links than others. The first four areas with a lower percentage of external edges are all from physical sciences. Contrariwise, the last six areas which yield the highest external edges percentages comprise almost all biological areas plus Chemistry and Applied Mathematics. All interdisciplinary areas plus Health Sciences presented a midterm percentage of external edges in comparison to the others. This analysis suggests a more integrated environment within the biological than among physical sciences.

3.1 Interdisciplinary Relationships from Literature Review

In order to provide a comparative basis to our study, we gathered information about the disciplines that arose from the physical-biological interchange by performing a systematic literature review, looking for comprehensive texts which provide broad overviews or historical perspectives of their respective areas. We performed searches in Google Scholar using as keywords the areas’ names combined with the words “review”, “overview” and “history”. Texts referring only to specific topics from the areas were not considered. The selected manuscripts were then organized in Table 1 by row according to its interdisciplinary area and their reference in each column indicates that the article cites a relationship between the interdisciplinary area and the corresponding physical/biological field.

Regarding the considered biological and physical areas, we have from Table 1, Computer Science showed to be highly referred among almost all interdisciplinary sciences. It is intuitive to think that the vast majority of analyses that are carried out nowadays in any area make use of computational resources. On the biological side, Molecular Biology has a predominance in almost all interdisciplinary areas except Computational Ecology and Biomedical Engineering.

With respect to the interdisciplinary areas, although we can say that Bioinformatics is today the most known name among all considered interdisciplinary areas, few were the scientific manuscripts found that discuss the area as a whole rather than some specific topic. Nevertheless, the so found manuscripts suggest an almost exclusive relationship between Computer Science and Molecular Biology. A similar scenario can be observed when considering Systems Biology. Despite being a younger scientific field, it underwent fast development at the beginning of this millennium due to its power to integrate biological data, and therefore, provide more faithful and larger computational models of biological systems.

The data in Table 1 is represented as a diagram in Figure 5. It is easy to observe that Computational Biology extends through a wider range of areas, perhaps as a consequence of existing for a longer time. How-
Figure 4: (a) Scientific areas network visualization. Biological, Physical, and Interdisciplinary areas are represented by green, blue, and yellow vertices, respectively. The size of each vertex is proportional to the number of connections within articles of the same area. The thickness of the edges is proportional to the number of links between the two areas it connects. Highlighted connections between areas compose the smallest set of strongest edges which keep all nodes connected into a single component. (b) Percentage of external connections of each interdisciplinary area with the other biological and physical areas. (c) The proportion of the internal and external edges of each area.

ever, despite its historical importance, Biochemistry, our results suggest that it remained mostly within its original realm.

If we compare the results obtained from Wikipedia with those derived from the performed literature review, some differences become evident. First of all, four out of all six considered interdisciplinary areas could not be considered in the Wikipedia analysis because of their extremely low number of articles. The possible reasons for this relatively small number of include area specificity, existence time, and funding, among others. On the other hand, for the remaining areas with many related articles in the Wikipedia analysis, namely Bioinformatics and Biochemistry, we could not find more than three and two respective review articles.

It is worth noticing that the subject of Biotechnology, considered as a biological area, could not be found in any of the selected review articles. Nevertheless, we have from Figure 4a that Biotechnology has strong interactions with both Bioinformatics and Biochemistry areas.

4 Concluding Remarks

As implied by its own name, interdisciplinary sciences emerge from interplays between different disciplines, and their importance to scientific advance has grown steadily. Among all possible combinations between knowledge fields, the interaction between physical and biological sciences is one that deserves particular attention. The outcomes from this interaction are already affecting society as a whole, once they permeate vital sectors such as health and food production. Thus, it is essential to understand how the combinations of physical and biological sciences are establishing, evolving and influencing one another. To do so, we proposed a data-driven methodology where relationships between scientific areas can be estimated from the scientific literature. We used the Wikipedia’s database to obtain articles whose categories were related to several scientific areas under biological and physical sciences, as well as interdisciplinary fields that have emerged from them. Given that each of the considered areas had a substantially
Table 1: Literature review of interdisciplinary areas

| Physical Sciences | Biological Sciences | Total |
|-------------------|---------------------|-------|
| Mathematics       | Ecology             | 3     |
| Computer Science  | Biology             | 3     |
| Applied           | Biotechnology        | 3     |
| Dynamical Systems | Medicine            | 3     |
| Chemistry         | Health Sciences      | 3     |
| Engineering       |                      | 3     |

Figure 5: Diagram of relationships between biological and physical sciences derived from the data in Table 1.

Regarding the investigation of review in the related scientific literature, we have that Computer Science and Molecular Biology are found in most of the reviews. At the same time, it is interesting to observe that, unlike this hypergraph, the network obtained from the Wikipedia data is completely interconnected. This suggests that the interrelationships between the considered interdisciplinary areas could be more intricate, in the sense of being more intensely interconnected, between the interdisciplinary areas with the biological instead of with the physical areas. A possible explanation for this interesting structure could be that most of the interdisciplinary areas were originally motivated from biology, reflecting an increasing need for incorporating more and more concepts and tools from the physical sciences. In other words, the considered interdisciplinary areas would be mostly driven to biological applications. Other interesting results include the centrality of Biochemistry among the biological areas, as well as the particularly strong interconnection between the considered biological and physical realms implemented by Chemistry.
citate and integrated than that highlighted by respective reviews. Still, a good agreement can be observed between these two structures, with exception of the Biotechnology area, which appears well connected with other areas in the obtained network derived from Wikipedia while it was not highlighted in any of the considered reviews.

All in all, we have approached the problem of characterizing some of the main interdisciplinary areas between the biological and physical sciences in terms of both real data obtained from the Wikipedia, as well as through the consideration of some respective literature reviews. Though a putative network of interrelationships has been obtained, it is important to bear in mind that these results are, as yet, substantially incomplete and preliminary. The future availability of more data and literature reviews can provide a means for obtaining more complete and accurate characteristics by using the same, or similar, methodology. Another interesting prospect for future research regards deriving descriptions of how the interdisciplinary areas changed with time.

5 Acknowledgments

Paulo E. P. Burke thanks Filipi Nascimento Silva for providing the filtered Wikipedia data. Luciano da F. Costa thanks CNPq (grant no. 307085/2018-0) for sponsorship. This work has benefited from FAPESP grant 15/22308-2. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

[1] A. L. Porter and I. Rafols, “Is science becoming more interdisciplinary? Measuring and mapping six research fields over time,” *Scientometrics*, vol. 81, no. 3, pp. 719–745, 2009.

[2] R. Van Noorden, “Interdisciplinary research by the numbers,” *Nature*, vol. 525, no. 7569, pp. 306–307, 2015.

[3] S. Chen, C. Arsenault, and V. Larivière, “Are top-cited papers more interdisciplinary?,” *Journal of Informetrics*, vol. 9, no. 4, pp. 1034–1046, 2015.

[4] G. K. Hunter, *Vital forces: the discovery of the molecular basis of life*. Academic Press, 2000.

[5] A. A. Berryman, “The Orgins and Evolution of Predator-Prey Theory,” *Ecology*, vol. 73, pp. 1530–1535, oct 1992.

[6] J. B. Hagen, “The origins of bioinformatics,” *Nature Reviews Genetics*, vol. 1, no. 3, pp. 231–236, 2000.

[7] F. Sanger, S. Nicklen, and A. R. Coulson, “DNA sequencing with chain-terminating inhibitors,” *Proceedings of the National Academy of Sciences*, vol. 74, pp. 5463–5467, dec 1977.

[8] J. Watson, “The human genome project: past, present, and future,” *Science*, vol. 248, no. 4951, pp. 44–49, 1990.

[9] G. W. Brodland, “How computational models can help unlock biological systems,” *Seminars in Cell & Developmental Biology*, vol. 47-48, pp. 62–73, dec 2015.

[10] F. Markowetz, “All biology is computational biology,” *PLOS Biology*, vol. 15, p. e2002050, mar 2017.

[11] J. D. J. Han, Y. Liu, H. Xue, K. Xia, H. Yu, S. Zhu, Z. Chen, W. Zhang, Z. Huang, C. Jin, B. Xian, J. Li, L. Hou, Y. Han, C. Niu, and T. C. Alcon, “Developmental systems biology flourishing on new technologies,” *Journal of Genetics and Genomics*, vol. 35, no. 10, pp. 577–584, 2008.

[12] P. Nurse and J. Hayles, “The cell in an era of systems biology,” 2011.

[13] S. J. Aronson and H. L. Rehm, “Building the foundation for genomics in precision medicine,” 2015.
[14] C. Krittanawong, H. J. Zhang, Z. Wang, M. Aydar, and T. Kitai, “Artificial Intelligence in Precision Cardiovascular Medicine,” 2017.

[15] C. L. Palmer, “Structures and strategies of interdisciplinary science,” *Journal of the American Society for Information Science*, vol. 50, no. 3, pp. 242–253, 2002.

[16] E. Omodei, M. De Domenico, and A. Arenas, “Evaluating the impact of interdisciplinary research: A multilayer network approach,” *Network Science*, vol. 5, no. 2, pp. 235–246, 2017.

[17] F. N. Silva, D. R. Amancio, M. Bardosova, L. d. F. Costa, and O. N. Oliveira, “Using network science and text analytics to produce surveys in a scientific topic,” *Journal of Informetrics*, vol. 10, no. 2, pp. 487–502, 2016.

[18] D. R. Amancio, M. G. Nunes, O. N. Oliveira, and L. F. da Costa, “Using complex networks concepts to assess approaches for citations in scientific papers,” *Scientometrics*, vol. 91, no. 3, pp. 827–842, 2012.

[19] F. N. Silva, F. A. Rodrigues, O. N. Oliveira, and L. Da, “Quantifying the interdisciplinarity of scientific journals and fields,” *Journal of Informetrics*, vol. 7, no. 2, pp. 469–477, 2013.

[20] C. Mund and P. Neuhausler, “Towards an early-stage identification of emerging topics in science: The usability of bibliometric characteristics,” *Journal of Informetrics*, vol. 9, no. 4, pp. 1018–1033, 2015.

[21] J. Helly, T. Case, F. Davis, S. Levin, and W. Michener, “The State of Computational Ecology,” *San Diego Super Computer Center*, 1995.

[22] M. Pascual, “Computational Ecology: From the Complex to the Simple and Back,” *PLoS Computational Biology*, vol. 1, no. 2, p. e18, 2005.

[23] S. Petrovskii and N. Petrovskaya, “Computational ecology as an emerging science,” *Interface focus*, vol. 2, pp. 241–54, apr 2012.

[24] J. Cohen, “Bioinformatics—an introduction for computer scientists,” *ACM Computing Surveys*, vol. 36, pp. 122–158, jun 2004.

[25] N. M. Luscombe, D. Greenbaum, and M. Gerstein, “What is bioinformatics? An introduction and overview,” *Methods in Information in Medicine*, pp. 346–358, 2001.

[26] H. Stevens, *Life out of sequence: a data-driven history of bioinformatics*. The University of Chicago Press, 2013.

[27] L. Hood, “Systems biology: integrating technology, biology, and computation,” *Mechanisms of Ageing and Development*, vol. 124, pp. 9–16, jan 2003.

[28] H. Kitano, “Computational systems biology,” *Nature*, vol. 420, pp. 206–210, nov 2002.

[29] H. Kitano, “Systems Biology: A Brief Overview,” *Science*, vol. 295, pp. 1662–1664, mar 2002.

[30] T. Ideker, T. Galitski, and L. Hood, “A New Approach to Decoding Life: Systems Biology,” *Annual Review of Genomics and Human Genetics*, vol. 2, pp. 343–372, sep 2001.

[31] P. Kohl, E. J. Crampin, T. A. Quinn, and D. Noble, “Systems Biology: An Approach,” *Clinical Pharmacology & Therapeutics*, vol. 88, pp. 25–33, jul 2010.

[32] M. S. Waterman, *Introduction to computational biology: maps, sequences, and genomes: interdisciplinary statistics*. Chapman & Hall/CRC, 2000.

[33] B. M. Slepchenko, J. C. Schaff, J. H. Carson, and L. M. Loew, “Computational Cell Biology: Spatiotemporal Simulation of Cellular Events,” *Annual Review of Biophysics and Biomolecular Structure*, vol. 31, pp. 423–441, jan 2002.

[34] D. Noble, “The rise of computational biology,” *Nature Reviews Molecular Cell Biology*, vol. 3, no. 6, pp. 459–463, 2002.
[35] J. D. Bronzino, *The biomedical engineering handbook*. CRC Press, 2000.

[36] J. D. J. D. Enderle and J. D. Bronzino, *Introduction to biomedical engineering*. Elsevier/Academic Press, 2012.

[37] F. Nebeker, “Golden accomplishments in biomedical engineering,” *IEEE Engineering in Medicine and Biology Magazine*, vol. 21, pp. 17–47, May 2002.

[38] P. Singh, H. S. Batra, and M. Naithani, “History of biochemistry.” *Bulletin of the Indian Institute of History of Medicine (Hyderabad)*, vol. 34, no. 1, pp. 75–86, 2004.

[39] D. Voet and J. G. Voet, *Biochemistry*. Wiley, 4 ed., 2010.