SpringerBriefs in Mathematics

SpringerBriefs in Mathematics showcases expositions in all areas of mathematics and applied mathematics. Manuscripts presenting new results or a single new result in a classical field, new field, or an emerging topic, applications, or bridges between new results and already published works, are encouraged. The series is intended for mathematicians and applied mathematicians.

More information about this series at http://www.springer.com/series/10030

SBMAC SpringerBriefs

The **SBMAC SpringerBriefs** series publishes relevant contributions in the fields of applied and computational mathematics, mathematics, scientific computing, and related areas. Featuring compact volumes of 50 to 125 pages, the series covers a range of content from professional to academic.

The Sociedade Brasileira de Matemática Aplicada e Computacional (Brazilian Society of Computational and Applied Mathematics, SBMAC) is a professional association focused on computational and industrial applied mathematics. The society is active in furthering the development of mathematics and its applications in scientific, technological, and industrial fields. The SBMAC has helped to develop the applications of mathematics in science, technology, and industry, to encourage the development and implementation of effective methods and mathematical techniques for the benefit of science and technology, and to promote the exchange of ideas and information between the diverse areas of application.

http://www.sbmac.org.br/

Michel Eduardo Beleza Yamagishi

Mathematical Grammar of Biology

Springer

Michel Eduardo Beleza Yamagishi
Laboratório de Bioinformática Aplicada
Embrapa Informática Agropecuária
Campinas, SP, Brazil

ISSN 2191-8198 ISSN 2191-8201 (electronic)
SpringerBriefs in Mathematics
ISBN 978-3-319-62688-8 ISBN 978-3-319-62689-5 (eBook)
DOI 10.1007/978-3-319-62689-5

Library of Congress Control Number: 2017946683

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

With love to my great-grandparents, grandparents, mother, father, wife, daughter, son, grandchildren, and great-grandchildren.

Preface

> Few philosophers would now dispute this: the nihilist or irrationalist nature of Popperite philosophy of science is by now pretty much an open secret. (David Stove [88])

I used to be uncomfortable with the post-modernist outlook on science until I read one of the books [89] by Australian philosopher David Stove (1927–1994), and figured out the source of my uneasiness. Since then, I've realized that I'd rather cling to the classical view, in which, as Erwin Chargaff wrote

> Science is the attempt to learn the truth about those parts of nature that are explorable.

Man has always explored the natural world, building what I call knowledge heritage, without putting under question his own intellectual power to do so. Knowledge belongs to humanity, and for this reason, scientists[1] must try their best to disseminate their findings. Spreading new ideas seems to be an easy task, but it cannot be forgotten that human communication is susceptible to deficiencies. Depending on how a message is conveyed, its content may be completely obfuscated. Therefore, the scientific work is not restrained to reveal nature's secrets; it also demands a proper skill to share them. Otherwise, no matter how important one finding may be, it will remain unnoticed until somebody else rediscovers it and does a better job of presenting it. The history of science has innumerable poignant and sad cases like these. However there is an even more important reason to broadly communicate new results: science, like all human activities, is not an error-free enterprise. Its outcomes must be independently double checked by peers in order to identify mistakes and correct them.

Breaking new ground is usually difficult. Arguably, it is easier to present a consolidated matter than it is to introduce a new one. Indeed, novel findings ought to first be deeply understood before there is any serious attempt to publish them. The results I will present are quite seminal, and on top of that, they are inherently multidisciplinary, which makes their presentation even harder. For this reason,

[1] And whoever explores nature.

long before writing this book, I spent a considerable amount of time reflecting on the results themselves and on the best way to present them intelligibly. It is well established that creation is not a linear process. Most scientific books show a sequence of events that does not correspond to the facts. This artifice, however, is necessary in order to make it easier for others to understand the main ideas. Albeit reluctantly, I was forced[2] to adopt the same procedure, but I do regret the unavoidable side effect of misrepresenting the actual process of scientific discovery, which is full of deadlocks, retreats, and just few true advances.

Just to give a glimpse of how involved human understanding is, as awkward as it may sound, it is possible for someone to fail to fully grasp all of the subtleties of his *own* work. Based on my experience, I used to believe that I was the only one to recognize this embarrassing truth, but, fortunately for me, there are other researchers[3] who have publicly admitted this as well. For instance, a few years ago, I read the book *A Universe from Nothing* [60]. Its subtitle *Why there is something rather than nothing?* is an old philosophical question that has been addressed by humankind's greatest minds. Lamentably, no satisfactory answer has been proposed thus far, and because Philosophy[4] is one of my deepest interests, I could not help but buy the book. The book was enjoyable and at times even funny, which allowed me to keep reading, even after "nothing" was redefined as "empty space," and "why" was replaced by "how" in the ninth chapter, which had the following title: "Nothing Is Something."[5] I suppose if Chargaff could revise it, he would produce a piece entitled *"A quick descent from Mount Olympus."* [12] Though, in my humble opinion, the book did not deliver what it promised, its reading was important to me because I unexpectedly stumbled upon the following disconcerting confession:

> Indeed, there are several of my own most important papers that I only fully understood well after the fact.

Even in Mathematics, where reason plays a major role, human understanding is not a completely rational and conscious process. Commenting on mathematician modes of thought, G.H. Hardy (1877–1947) wrote

> ...that unconscious activity often plays a decisive part in discovery; that periods of ineffective effort are often followed, after intervals of rest and distraction, by moments of sudden illumination; that these flashes of inspiration are explicable only as the result of activities of which the agent has been unaware – the evidence for all this seems overwhelming. [45]

In other words, contrary to the popular belief, scientific insights may arise from pure intuition, as in the *arts*. Both Henri Poincaré (1854–1912) and Jacques

[2] I've tried several alternatives, but none were successful.

[3] Chargaff once said: "When I began to realize how unique were the regularities we had discovered, I tried, of course, to understand what it all meant, but I did not get very far." [14]

[4] Philosophy is not dead. Whoever believes the contrary has not kept up with authentic philosophy.

[5] I admit that if the book's title were "A Universe from *empty space*. *How* there is something rather than *something*?," I would have never bought it.

Hadamard (1865–1963) classified mathematicians as having either "logical" or "intuitive" minds. I'd rather represent any scientist as a *convex combination* of three components: Reason (R), Intuition (I), and Luck (L). Scientific achievements depend on all three, but I usually tell my students that only the latter is necessary and sufficient. Therefore, in my opinion, this is the main reason why, with the exception of geniuses, scientists in general should not be too proud. I would dare to represent three great mathematicians as follows:

$Ramanujan = 0.2R + 0.7I + 0.1L$[6];
$Hilbert = 0.7R + 0.29I + 0.01L$;
$Gödel = 0.9R + 0.099I + 0.001L$.[7]

The weight of the R component is correlated with the skill to *consciously* grasp complex problems. I do not have the ridiculous pretension of comparing myself to these giants; nevertheless, given that every scientist may be represented in that way, I would say that $Yamagishi = 0.005R + 0.1I + 0.895L$.

Despite all my earnest attempts to equally address both mathematicians and non-mathematicians in this book, I'm afraid that my inborn capacities are not enough to accomplish this feat. Therefore, my fellow mathematician will probably yearn for a better exposition of some biological aspects, while the biologists reading will find the book wanting for clarity on several mathematical concepts. No writer is completely immune to this sort of compromise when addressing two or more different audiences. However, for the interested reader, the relevant bibliography on the main topics may be found in the reference section. In an attempt to facilitate comprehension, I've tried to exhibit only the aspects of the fields that have truly contributed to my scientific work.

I believe that publications like the *SBMAC SpringerBriefs* are very important. There is always trouble when it comes to bringing interdisciplinary works to press. Manuscripts on applied mathematics may be either too mathematical to be accepted by applied periodicals or too applied to be published in Mathematics journals. There is a whole spectrum within the field of Applied Mathematics, ranging from Pure Mathematics to any other natural sciences, that lack appropriate means of publication. The work presented in this book is a good example of this, for there is no *"Mathematical Grammar of Biology Journal"* yet. In the absence of interdisciplinary compendiums similar to the *SBMAC SpringerBriefs*, several original applied works are likely to have been condemned to oblivion.

Finally, I would like to pay homage to the great scientist Erwin Chargaff. I first learned of him through his scientific papers, but I soon realized that he also left

[6]Without that 0.1 of *Luck*, Hardy would have never read and realized the importance of Ramanujan's former works, and they would have never worked together to reach even more interesting results. Maybe a little bit more *Luck* would have prevented him from dying so young.

[7]G.K. Chesterton (1874–1936) once said that *"Imagination does not breed insanity. Exactly what does breed insanity is reason. Poets do not go mad; but chess-players do. Mathematicians go mad."* [17] Indeed, Gödel paid a high price for being extremely rational: he starved to death.

behind an equally important humanistic opus. Unfortunately, his books in English[8] are prematurely out of print. But I've managed to read all of the publications I could buy from used and antique booksellers, and they were worth every penny. Chargaff had a classic erudition that made him a *sui generis* scholar, as can be observed in his memoirs [14].

> I cannot serve as an example for younger scientists to follow. What I can teach cannot be learned. I have never been a "100 percent scientist". My reading has always been shamefully nonprofessional.

Highly influenced by his "former high-school teacher"[9] Karl Krauss (1874–1936), Chargaff was often satirical (a delicious example may be found in [12]), which earned him some strains in his relationships.

> Nevertheless, if at one time or another I have brushed a few colleagues the wrong way, I must apologize: I had not realized that they were covered with fur.

However, even at the risk of incomprehension, Chargaff did not hesitate to express deep thoughts through simple sentences, as in his statement:

> We posit intelligence where we deny it. We humanize things, but we reify man.

Certainly, he belonged to a rare class of Scientists. There were few before him, and there have been none like him since his death.[10]

This is the last time that the pronoun "I" will appear in this manuscript. Although I am the only author of every subjective opinion or philosophical digression in the book, part of the scientific results reported within these pages were obtained in partnership with other colleagues, namely *Alex I. Shimabukuro* and *Roberto H. Herai*. Nevertheless, all inaccuracies or gross mistakes that may have reached the final revision should be credited to me alone.

Campinas, SP, Brazil — Michel Eduardo Beleza Yamagishi
July 29, 2017

[8] "The reason why I stopped publishing in English is very simple, because I couldn't find a publisher." [47]

[9] Actually, Karl Krauss was an editor and writer. Chargaff read his texts and attended his lectures.

[10] In Chargaff's own words:

> I would say that most of the great scientists of the past could not have arisen, that, in fact, most sciences could not have been founded, if the present utility-drunk and goal directed attitude had prevailed [14].

Culture is to human beings as soil is to plants, geniuses need a rich and fecund "soil" to flourish. Perhaps, in the West, the first historical example of such a rare soil occurred in the classical Athens that nourished Socrates, Plato, and Aristotle. The last instance took place in Vienna, between the late nineteenth and early twentieth century, where great minds either were born or lived for a while. Paradoxically, those "Viennese" giants ushered *the-man-without-qualities* era. Chargaff was one of them.

Contents

1 Introduction 1
- 1.1 Synthetic Biology 3
- 1.2 DNA's First Principles 4
- 1.3 The Science of Patterns 5
- 1.4 The Palimpsest 6
- 1.5 The Alphabet 6
- 1.6 The Birth of the Grammar of Biology 7

2 Modeling Human Nucleotide Frequencies 9
- 2.1 Sequencing DNA 10
  - 2.1.1 The Legacy of the Human Genome Project 11
  - 2.1.2 The ENCODE Project 12
  - 2.1.3 Pos-HGP: Next-Generation Sequencing Technologies 12
  - 2.1.4 Multidisciplinary Approach 13
- 2.2 Mathematical Modeling 14
  - 2.2.1 The Line 15
  - 2.2.2 The Premises 17
  - 2.2.3 Optimization Problem 23
  - 2.2.4 Experiment Follow-Up 24

3 Expanding the Grammar of Biology 29
- 3.1 The Right Question 30
- 3.2 Mathematical Definitions 33
- 3.3 Operators over $\mathscr{W}^k$ 36
  - 3.3.1 The Complement-Reverse Operator 37
  - 3.3.2 Induced k-Word Set Partition 38
  - 3.3.3 First Insight 40
  - 3.3.4 Generating Set 42
  - 3.3.5 The Riddle of the Symmetry Principle 43
  - 3.3.6 Palindromic Sequences 45

3.4 New Parity Rules 46
3.4.1 Second Insight 48
3.4.2 Third Insight 51
3.4.3 Alternative Path 52

4 "In God We Trust; All Others, Bring Data" 55
4.1 "To Be or Not to Be" 56
4.2 The Three Kingdoms of Life 59
4.2.1 A Plausible Explanation for the k2-Effect 62
4.3 The Synergy Between Mathematics and Biology 69

Postscript 71

References 73

Index 79