Universitext

Jürgen Jost

Mathematical Methods in Biology and Neurobiology

Springer

Universitext

Universitext

Universitext is a series of textbooks that presents material from a wide variety of mathematical disciplines at master's level and beyond. The books, often well class-tested by their author, may have an informal, personal, even experimental approach to their subject matter. Some of the most successful and established books in the series have evolved through several editions, always following the evolution of teaching curricula, into very polished texts.

Thus as research topics trickle down into graduate-level teaching, first textbooks written for new, cutting-edge courses may make their way into *Universitext*.

For further volumes:
http://www.springer.com/series/223

Jürgen Jost

Mathematical Methods in Biology and Neurobiology

Springer

Jürgen Jost
Max Planck Institute for Mathematics
in the Sciences
Leipzig
Germany

ISSN 0172-5939 ISSN 2191-6675 (electronic)
ISBN 978-1-4471-6352-7 ISBN 978-1-4471-6353-4 (eBook)
DOI 10.1007/978-1-4471-6353-4
Springer London Heidelberg New York Dordrecht

Library of Congress Control Number: 2014931395

Mathematics Subject Classification: 05Axx, 05Cxx, 05C05, 05C50, 05C75, 05C80, 05C82, 31B05, 34A12, 34A34, 34C05, 34C11, 34C15, 34C23, 34C25, 34C26, 34C28, 34C29, 34C60, 34D20, 34D23, 34D45, 34E10, 34E13, 35A01, 35B36, 35B50, 35B51, 35C07, 35G05, 35G10, 35G15, 35J05, 35K57, 35K05, 35L05, 35Q83, 35Q92, 49J05, 60G07, 60G55, 60J60, 60J70, 60J80, 60J85, 70K70, 92B05, 92C05, 92C40, 92C42, 92D15, 92D25, 92E20.

Mathematica® is the registered trademark of Wolfram Research, Inc., http://www.wolfram.com/

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Structures and processes studied in biology range from molecules in cells to populations or ecosystems, or to brains consisting of billions of interacting neurons, and the formal models employed in biology range from graphs as abstract representations of pairwise interactions to complicated systems of partial differential equations that try to capture all details of some biological system. Therefore, also the mathematical methods and tools employed in biology and neurobiology are quite diverse and heterogeneous. A student wanting to learn and apply mathematical techniques in biology might be confronted with the problem that she or he does not possess an overview of the available mathematical tools and does not know which method could be appropriate for a specific biological problem. A biological structure, in fact, can be modeled at various levels of details, and it is not necessarily the case that a more detailed and precise model yields better quantitative predictions.

In that situation, this book presents a spectrum of mathematical methods that are relevant and important for biology and neurobiology. Thereby, the student should be equipped with an overview and a working knowledge of the most important mathematical tools. These methods fall into three categories: First of all, there are the discrete methods, from combinatorics and graph theory. Graphs can be used to model the structure of pairwise interactions between elements in some network, whatever their precise biological nature might be. They can also be utilized to analyze empirical network data. A particular class of graphs, the trees, plays a special role in biology because they model descendence relations. The second class of models comprises the stochastic ones. Much of biology, in fact, is modeled in stochastic terms, be it the firing of neurons in the brain or the random forces of evolution. Therefore, I provide a systematic introduction to stochastic processes. Finally, there are the analytical methods from the theory of differential equations, like dynamical systems or partial differential equations, that are used to explain the formation of biological patterns, ranging from the molecular scale to that of interacting species. Often, such models are derived from optimization principles, the theoretical rationale being that evolutionary competition has produced structures that best perform certain functions. Therefore, we also devote a chapter to optimization schemes. A final chapter then deals with a particular area of mathematical biology, population genetics. That has been the field of biology where mathematical methods first have been applied in a very

systematic manner. It continues to be alive today, and I present a new geometric approach to population genetics that will, as I hope, clarify the underlying mathematical structure.[1] These last two chapters are thus both concerned with issues of evolutionary biology, but from two different perspectives, that of optimization versus that of random processes. For a mathematical understanding of evolution, the combination of these two perspectives is essential.

The exercises are concerned with both the mathematical techniques developed in this book and their application in biological modeling. While some exercises are more of the traditional drill type that is needed to master some technique, others are more open in order to stimulate and encourage your own thinking.

In this book, I try to explain the underlying mathematical concepts and to prove the easier statements so that the reader can develop some feeling for the abstract mathematical structures. Throughout the text, I also develop applications to biology, from intracellular structures or the dynamics of neurons to those of populations. Thus, the applications span many physical orders of magnitude, but perhaps somewhat surprisingly, often the same mathematical structures turn out to yield useful models at several rather different levels. In any case, the systematic arrangement of the material is according to mathematical and not to biological principles. This seemed the natural choice for the material to be presented here, but in order to compensate for that, I am writing a companion volume "Biology and Mathematics" [2] where I attempt a systematic presentation according to biological principles and structures. Actually, I have recently also written a book entitled "Mathematical Concepts" [3] where the mathematical structures are developed at a much more abstract level. I believe that this may be relevant for biology because theoretical biology needs to develop more abstract and encompassing concepts in order to organize and understand the multitude of biological structures and processes and the increasing wealth and heterogeneity of biological data more deeply. This aspect, however, is not addressed in the present book which rather concentrates on established mathematical methods and their biological applications. In contrast to such an abstract systematic treatment, this book emphasizes the richness and diversity of the applications of mathematics to biology.

The literature in mathematical biology is too extensive to be adequately covered in this book. Therefore, the references are very selective, and you should consult the monographs and survey articles listed in the bibliography for further or more precise references. I apologize to any authors whose work is not, or not correctly, referenced in this book.

In any case, while this book certainly aims at teaching a range of mathematical concepts and methods that are relevant for the modeling and analysis of biological structures and processes, it also wants to stimulate your curiosity about biological phenomena and your independent thinking about how to model and analyze them with mathematical tools.

[1] A more detailed exposition of this theory will be given in [1].

This book is based on graduate courses at Leipzig (in a joint program between the Max Planck Institute for Mathematics in the Sciences and the Department of Mathematics and Computer Science of Leipzig University, the International Max Planck Research School "Mathematics in the Sciences," directed by Stephan Luckhaus) and at the Ecole Normale Superieure in Paris (organized by Benoît Perthame).

Thus, a student who would like to use this book should have some basic mathematical knowledge, including in particular calculus. Some background in biology might help to appreciate the significance of the mathematical methods, but is not indispensable for reading this book. In fact, the book can also be taken as a survey over a rather wide range of mathematical structures, for any student of mathematics or the sciences.

I thank Ugur Abdullah, Fatihcan Atay, Nihat Ay, Anirban Banerjee, Frank Bauer, Nils Bertschinger, Pierre-Yves Bourguignon, Olaf Breidbach, Andreas Dress, Bernhard Englitz, Boris Gutkin, Julian Hofrichter, Danijela Horak, Bobo Hua, Martin Kell, François Kèpés, Ilona Kosziuk, Michael Lachmann, Shiping Liu, Stephan Luckhaus, Vic Norris, Eckehard Olbrich, John Pepper, Benoît Perthame, Johannes Rauh, Christian Rodrigues, Thimo Rohlf, Areejit Samal, Klaus Scherrer, Susanne Schindler, Peter Schuster, Bärbel Stadler, Peter Stadler, Angela Stevens, Tat Dat Tran, Henry Tuckwell, Leo van Hemmen, and others whom unfortunately I may have forgotten to mention, for various discussions about the mathematical and/or biological aspects. Tat Dat Tran also checked the entire manuscript, pointed out several corrections, and created some of the figures, in particular the simulations of the FitzHugh-Nagumo and van der Pol equations, with the help of Mathematica®. For several of the diagrams, I have used the latex supplement DCpic of Pedro Quaresma.

The author also gratefully acknowledges the support from the ERC Advanced Grant FP7-267087, the Volkswagenstiftung and the Klaus Tschira Stiftung.

References

1. Hofrichter, J., Jost, J., Tran, T.D.: Information geometry and population genetics, in preparation
2. Jost, J.: Biologie und Mathematik, to appear
3. Jost, J.: Mathematical Concepts, to appear

Contents

Chapter 1
Introduction

Abstract
Questions:

- What can mathematics contribute to biology, and which mathematical theories are useful for that purpose?

Biology does not have the clear structure of mathematics. Nevertheless, it possesses some fundamental concepts. The *gene* is the unit of coding, function, and inheritance. It contains the information for a phenotypic trait that is realized in interaction with contributions from the environment and transmitted to offspring. The *cell* is the basic unit within which metabolic processes can take place. The *species* is the dynamic pool for genetic recombination. An *organism* is a carrier of genes, an organized ensemble of cells and a member of a population or species. Mathematical methods to study biological phenomena can be taken from algebra, analysis, stochastics, or geometry, but should always be developed with a clear vision of the biological problems to be addressed.

1.1 Theses About Biology

Thesis 1. *Biological structures are aggregate structures. Therefore, biological laws are not basic ones that do not admit exceptions, but rather emerging from some lower scale.*

Thesis 2. *Biological entities are discrete, but biological structures are situated in continuous space and biological processes take place in continuous time.*

Thesis 3. *Biological processes intertwine stochastic effects and deterministic dynamics. Randomness can support order while deterministic processes can be unpredictable, chaotic. The question then is at which level regularities emerge.*

J. Jost, *Mathematical Methods in Biology and Neurobiology*, Universitext,
DOI: 10.1007/978-1-4471-6353-4_1,

Thesis 4. *Large populations of discrete units can be described by continuous models and, conversely, invariant discrete quantities can emerge from an underlying continuous substrate.*

Thesis 5. *Fundamental biological concepts, like fitness or information, are relative and not absolute ones.*

Thesis 6. *Fundamental biological quantities do not satisfy conservation laws. Those rather appear as external constraints.*

Thesis 7. *Biological systems interact with their environments and are thermodynamically open. Biological structures sustain the processes that reproduce them and are therefore operationally closed.*

Thesis 8. *Biological structures are results of historical processes. It is the task of biological theory to distinguish the regularities from the contingencies.*

Thesis 9. *The abstract question posed to mathematics by biology is structure formation. This needs to be understood as a process because living structures are not at thermodynamic equilibrium.*

Thesis 10. *Gathering biological data without guiding concepts and theories is useless.*

1.2 Fundamental Biological Concepts

1. The **gene** is the unit of coding, function, and inheritance. As such, it links molecular biology and evolutionary biology. The Neodarwinian Synthesis combined Mendel and Darwin. Modern molecular biology seems to offer a more basic perspective.
2. The **cell** is the unit of metabolism. It constitutes the basic operationally closed, autopoietic system in biology. Modern biology struggles to understand cells on the basis of their molecular constituents, DNA, RNA, and polypeptides (proteins). Multicellular organisms emerge through a partial suppression of the autonomy of the constituting cells.
3. The **species** represents the balance between the diverging effects of genetic mutations and selection at the organismic or other levels and the converging mechanism of sexual recombination. It is the arena of population biology, a child of the Neodarwinian Synthesis and the first success of mathematical models in biology. It is also important in ecology.

The **organism**, in fact, is the carrier of genes, the organization of cells, and the member of a species. It thus links the three fundamental biological concepts. It is also a, but not the exclusive, unit of selection.

It seems that neurobiology has not yet identified such a fundamental concept, but perhaps the **spike** can be considered as the basic event of information transmission, and the **synapse** as the basic structure supporting this.

1.3 A Classification of Mathematical Methods

The following is a somewhat incomplete list, arranged partly with relevance for biology in mind.

1. Discrete structures → **Algebra**
   (a) Static structures
      i. Algebraic concepts: Combination and composition of objects
      ii. Graphs and networks, including phylogenetic trees
      iii. Information
      iv. Discrete invariants of continuous structures and dynamical processes
   (b) Discrete processes (Cellular automata, Boolean networks, finite state machines,...)
   (c) Game theory as the formalization of competition
2. Spatial relations → **Geometry**
   (a) Geometry of (three-dimensional) physical space
   (b) Abstract notions of space for expressing relationships (discrete ones like graphs and continuous ones like Hilbert spaces; state spaces of dynamical systems)
   (c) Symmetries and invariances
3. Continuous methods → **Analysis**
   (a) Deterministic dynamical processes
      i. Continuous states enable phase transitions and bifurcations, that is, qualitative structural changes resulting from small underlying variations
      ii. Continuous states and time: Ordinary differential equations and other dynamical systems
      iii. Continuous spatial structures: Partial differential equations (example: Reaction-diffusion equations)
   (b) Stochastic analysis
      i. Stochastic processes (while stochastic processes may also operate on discrete quantities, the concept of probability is a continuous one)
      ii. Population processes: averaging over stochastic fluctuations in lower level dynamics
      iii. Optimization schemes with stochastic ingredients: Genetic and other evolutionary algorithms, swarm algorithms for distributed search, certain neural networks,...
      iv. Statistical methods for the analysis of biological data

4. Hybrid models
   (a) Difference equations (continuous states, but discrete time)
   (b) Dynamical networks (dynamical systems coupled by a graph), in particular neural networks
5. System theory as a global unifying perspective?

According to the preceding list, not all mathematical subjects seem to be relevant for biology. Classical algebraic structures occur in a cursory manner at best, and one of the deepest branches, number theory and arithmetics, is entirely absent. Three-dimensional physical space constitutes an important constraint for biological organization. Organisms and their constitutive biological structures like cells are living and interacting in space, and are defining and shaping their own spaces like architectural structures which is constitutive for morphology. Symmetries and invariances, the merging ground of algebra and geometry, are important issues for the neurobiology underlying cognition, as well as for many classification purposes. In any case, the branches of algebra, geometry and analysis are often interwoven.

Chapter 2
Discrete Structures

Abstract
Questions:

- How can the cells of an organism which all share the same genes can fulfill so many different functions?
- Are there good mathematical tools to identify the important features in all those networks that modern biological data collection produces?
- How long ago did the last common ancestor of two species or two individuals live?

A model of combinatorial gene regulation shows the power of combinatorics. Graphs are useful tools for network analysis, and their spectral theory is developed. Phylogenetic relationships between species are modeled by particular types of graphs, the trees. Descendence relations between individuals involve two parents and lead to genealogies. Coalescents treat the question of common ancestors. Such structures also naturally lead to the stochastic processes treated in the Chap. 3.

2.1 Introductory Example: Gene Regulation and the Power of Combinatorics

In this section, I present an example of a combinatorial scheme in molecular biology. This is meant to show that even elementary mathematical reasoning can help us to clearly understand a biological situation that may initially look rather complicated. First, however, I shall sketch the most basic principles of molecular biology. More details can be found in standard textbooks, like [1] or [93].
Metabolism and other fundamental functions of the cell are essentially carried out by proteins. The building blocks of proteins are polypeptides, sequences of typically a few hundred amino acids that fold into particular three-dimensional shapes according to attractive forces between different amino acids and interactions with water molecules in the cell. A protein consists of one or several such polypeptides, and

J. Jost, *Mathematical Methods in Biology and Neurobiology*, Universitext,
DOI: 10.1007/978-1-4471-6353-4_2, © Springer-Verlag London 2014

its three-dimensional shape determines its function. The information for the particular sequence of each polypeptide is contained in the DNA of the cell. The DNA is a sequence itself, consisting of nucleotides instead of amino acids, and the DNA is inherited by the daughter cells under cell division and the germ cells in sexual recombination. This will now be described with some more details and precision.
The fundamental process of molecular biology then is gene expression, that is, the production of polypeptides, the building blocks of proteins, according to the genetic information contained in the DNA of a cell. The DNA (desoxyribonucleic acid) is a long string of base pairs, arranged in the shape of a double helix, as discovered by Watson and Crick. There are four different nucleotide bases, labelled A, C, G, and T (we are not concerned here with their precise chemical identity, and so, these letters may suffice for our purposes). Thus, each of the two strands of the double helix is a long sequence composed of these 4 "letters". Each strand determines the identity of the complementary strand, because C in one strand is paired with G in the other, and A with T. Therefore, when the double helix is split apart, each strand contains the complete information for assembling a new such double helix. This is the principle underlying genetic inheritance. Here, however, we are not concerned with inheritance, but rather with gene expression. The first step of gene expression, called transcription, then consists in copying the information in a segment of one of the strands into another macromolecule, RNA (ribonucleic acid), which is chemically more active and flexible. It also consists of sequences composed of 4 letters, A, C, G as in the DNA and a new letter U taking the place of T. Again, this copying works according to the above complementarity principle. Which segments of the DNA are thus copied under a given cellular condition is controlled by certain proteins, the transcription factors that typically bind to locations in the DNA nearby those to be copied and that can then trigger, enhance or block the transcription process [26]. Of course, one and the same stretch of DNA can be repeatedly transcribed, and the regulation of the number of such transcripts is essential, but we shall not emphasize this aspect in the sequel. The resulting RNA is then further processed, through interactions with itself or with other RNAs or with certain proteins again. The final mRNA (m standing for "messenger") can then be translated into a polypeptide, in a certain complex called the ribosome, with the help of some other auxiliary RNA, the rRNA (r standing for "ribosomal"). The principle of the translation is that the unit of translation in the mRNA is a triplet of nucleotides, like *ACG* or *UAA*, also called a codon. Each such triplet is translated into a specific amino acid, and the resulting polypeptide thus is a sequence of amino acids. Since there are 64 possible triplets, but only 20 amino acids, several different triplets can correspond to the same amino acid. This fact is called the degeneracy of the genetic code, although redundancy might be the more accurate word. (Actually, the triplet *UGA* has a special role: It serves as the stop codon, that is, when this triplet is encountered in the ribosomal complex, the polypeptide is released, and a new translation can start.) In fact, the relation between such triplets and amino acids is mediated by another type of RNA, called tRNA (t for "transfer"). Chemically, this relation, called the genetic code, that is, which triplet is translated into which amino acid, is arbitrary, and so the question emerges why the translation rules are as they are, instead of being different. That is, why is for instance

GCC translated into the amino acid alanine, instead of, say, cysteine? Is that simply a historical accident, an arbitrary rule that all living creatures have inherited from their common ancestor who had adopted these translation rules by chance? Or are there some chemical or formal principles behind this, like symmetry considerations or coding efficiency? There have been many different speculations about this issue, but none so far has met with general approval.

One or more polypeptides then are combined into a protein. An important point is that a protein is not simply an amino acid sequence, but that for its molecular function, it assumes a specific three-dimensional shape. This shape, is determined by chemical attraction and repulsion between different pieces, but the details are very intricate, and the problem of computing the three-dimensional shape of a protein, or better, the process, called protein folding, by which it acquires this shape from its constituting amino acid sequence is not yet fully solved, despite considerable attempts by many mathematicians and physicists.

The fundamental question for a cell then is which genes to express when, under which circumstances. The mechanism of the cell for answering this question is gene regulation. I have already described that specific proteins, the transcription factors, trigger or inhibit the transcription of DNA segments. In eukaryotic cells (the cells that we are made of, those containing a nucleus, in contrast to prokaryotic cells, without nucleus, like bacteria), the most important part of gene regulation, however, seems to take place at the level of RNA rather than DNA. First of all, the transcribed RNA, called pre-mRNA, is reassembled in a process called splicing into mRNA. Here, on one hand, certain segments, the so-called introns, are cut out whereas the remaining ones, the exons, can then be assembled possibly in different ways, so as to produce different results from one and the same stretch of DNA [13], or pieces of different origins can be put together or interact in other ways. The processing on one hand is based on the spatial configuration assumed by an RNA molecule, on the basis of bindings between complementary nucleotides (*A* with *U* or *C* with *G*), no longer between different strands as in the DNA, but now between bases in one and the same RNA sequence [59]. On the other hand, it results from interactions with certain other small RNAs, the so-called miRNAs (mi for "micro") or siRNAs (si for "small interfering" or "silencing") or with specific proteins. These proteins bind to RNA molecules to form so-called RNP complexes (where P stands for "protein") [119]. Much of this RNA regulation works as repression, that is, preventing the mRNA from being translated. The biological rationale for this is that on one hand, the production of RNA is energetically cheap, and on the other hand, with mRNA already around, it is much faster to produce the corresponding proteins than if the process had to start anew from the DNA level. Thus, the cell can respond much quicker to new circumstances. (For a systematic analysis, see [103, 104] and the subsequent discussion in the journal Theory in Biosciences, see [105].)

After the genome of humans (and several other species) has been sequenced, that is, the the identity of all the 3 billion letters in the DNA sequence has been established [63], now the ENCODE project systematically records and catalogues all the different RNA molecules that can be present in human (and other) cells [34, 38, 49]. The genetic sequence contains both coding information that can be potentially

activated and utilized in a cell with the assistance of specific proteins, and important structural elements. But we need to identify all the different RNAs and understand their interactions with other RNAs and proteins in order to understand the regulation of gene expression in the active cell.

Now, obviously, the scheme described offers many possibilities for combinatorial reasoning as a formal description of the rules governing those processes. Here, as an example I shall discuss a model that arises from my work with the molecular biologist Klaus Scherrer, see [76]. The important point here is that the nucleotides in an mRNA can assume two different roles simultaneously. On one hand, they are parts of coding triplets (except for certain portions at the beginning or end of an mRNA sequence). On the other hand, stretches of about 30 nucleotides can function as binding sites for specific proteins which then regulate the fate of the mRNA, as explained (see [103, 104]). We call such a regulatory stretch of nucleotides an oligomotif. In the basic version of the model, there then is a one-to-one correspondence between such oligomotives and mRNA binding proteins. That is, there is a second, regulatory, code superimposed upon the first code, the genetic code governing translation. In both cases, however, the chemical identity of the nucleotides involved is crucial. An average mRNA may then possess about 20 such oligomotives. The ground state then is when the corresponding proteins are attached to all those 20 oligomotives. In this state, the mRNA is repressed and not translated. It only becomes available for translation when at least 3 of those proteins are removed. (We shall call such a set of 3 oligomotives, or equivalently, of 3 mRNA binding proteins, a triple, not to be confused with the triplet of the genetic code.) That is, when a signal arrives in the cell that causes the release of 3 such binding proteins, the corresponding mRNA gets translated, and a specific polypeptide is produced. Now, however, in a given situation, a cell needs not only one type of polypeptide, but a suitable combination of perhaps hundreds of polypeptides. The preceding structure now offers an elegant scheme for the coordinated expression of groups of genes, that is, the coordinated production of specific combinations of polypeptides and proteins. First of all, there are then $\binom{20}{3} = 1,140$ different possibilities for such triples of oligomotives. The key point now is that different mRNAs will share some, but not all of their oligomotives. That is, whenever we identify 3 proteins for removal, that is, select 3 oligomotives, we then get a specific set of mRNAs that contain those 3 oligomotives and that will then get translated, whereas the remaining ones will stay repressed. And when we select a different set of 3 oligomotives, we obtain a different combination of mRNAs to be translated, hence a different combination of proteins in the cell. This set may partially overlap with the preceding one, depending on the distribution of oligomotives across the different RNAs. In fact, one estimates that there are about 3,000 different mRNA binding proteins, hence also about 3,000 different oligomotives according to the model. We thus have $\binom{3,000}{20}$ different possibilities to distribute the oligomotives across the mRNAs (there are perhaps around 10,000 different mRNAs in a typical mammalian cell).

Let us now look into this scheme in more numerical detail. As explained, in order that several mRNAs participate in the same condition, they need to share at least 3 oligomotives. And when some mRNAs share m oligomotives ($3 \leq m \leq 20$), they can

simultaneously participate in $\binom{m}{3}$ conditions. This number varies from 1 (for $m = 3$) to 1,140 (for $m = 20$). However, when $m = 20$, that is, when the mRNAs share all their oligomotives, they can no longer be distinguished in this scheme. Let us consider some numerical examples, on the basis of the general scheme. For K oligomotives, there are $\binom{K}{20}$ different possibilities to choose 20 among them. This means that we can distinguish that many mRNAs through their different endowments with 20 out of these K oligomotives. As explained, a condition for translation is achieved by the selection of 3 (or more) out of these K oligomotives. Every choice of $\binom{K}{3}$ yields a different condition. Precisely those mRNAs will participate in such a condition that carry all those 3 oligomotives. Thus, 3 out of their 20 oligomotives are fixed, and 17 remain for free choice. That is, we have $\binom{K-3}{17}$ different possibilities. Thus, assuming that all the above $\binom{K}{20}$ possibilities are realized, by selecting 3 oligomotives, we select $\binom{K-3}{17}$ different mRNAs. Here are simple numerical examples.

- Distribute 21 oligomotives among 21 mRNAs (20 oligomotives/mRNA) so that each mRNA is identified by which oligo it does not contain. By specifying 3 oligomotives, any of the possible $\binom{21}{3} = \binom{21}{18} = 1,330$ combinations of 18 mRNAs can then be selected. Here, we have only relatively few different mRNAs.
- Distribute 23 oligomotives among $\binom{23}{3} = 1,771$ mRNAs (20 oligomotives/mRNA) so that each mRNA is identified by which 3 oligomotives it does not contain. By specifying 3 oligomotives, any of the possible $\binom{23}{3} = \binom{23}{20} = 1,771$ combinations of $\binom{20}{3} = 1,140$ mRNAS can be selected. Here, we obtain a large collection of selected mRNAs.
- Distribute 22 oligomotives among $\binom{22}{2} = 231$ mRNAs (20 oligomotives/mRNA) so that each mRNA is identified by which 2 oligomotives it does not contain. By specifying 3 oligomotives, any of the possible $\binom{22}{3} = \binom{22}{19} = 1,440$ combinations of $\binom{19}{2} = 171$ mRNAs can be selected. This is a biologically reasonable number.

Obviously, the number 3,000 of different mRNA binding proteins, that is, of different oligomotives is far larger than needed in our model. This indicates that, in reality, gene regulation at mRNA level is more complex than captured by the model. Nevertheless, the model should describe a core principle of regulation. Moreover, there is an interesting combinatorial problem suggested by this model: How to distribute K labels among N units so that each unit receives k of them so that by selecting $\kappa < k$ of them (for which we have $\binom{k}{\kappa}$ different possibilities), we identify the maximal number of different subsets of those N units? We may here wish to constrain those subsets to be of some fixed size n, or to be within a certain size range, say between n_1 and n_2.

In order to understand the mathematical structure of this problem better, it is helpful to translate it into a combinatorial design problem. We consider an $N \times K$ matrix with entries 1 or 0 where each of the N rows has precisely k 1s, and hence $K - k$ 0s. For $\kappa < k$, we then want to find collections of rows that have (at least) κ 1s in common. The question then is how to distribute the 1s in the rows so as to find as many such collections as possible within a given size range.

No full solution seems to be known for this problem. In any case, the example is meant to show that by elementary mathematical reasoning, we can come up with clever ways of how a cell could regulate its genes so that in one situation, in a single stroke, it can co-activate specific groups of genes, and in another situation, again in a single stroke, it can activate another set of genes, perhaps partly overlapping with the first one, without having to address all these genes individually. This is the power of combinatorics.

2.2 Graphs and Networks

2.2.1 Graphs in Biology

A graph is the mathematical structure representing binary relationships between discrete elements. These elements are the vertices of the graph, and the relationships are encoded as connections or edges between vertices. Such a graph can then be a network, that is, the substrate of dynamical interactions carried by the edges between processes located at the vertices. Biological applications abound.

In neural networks, the vertices stand for neurons, and the edges for synaptic connections between them. The interaction is the electrochemical transmission of pulsed dynamical activity, the spikes generated in the neurons. This activity is considered to be the carrier of information, enabling cognitive processes, but the precise identification of the information inside that dynamical activity remains unclear at present.

At smaller scales, the vertices can represent molecules like proteins, and the edges again interactions between them. The vertices can also stand for genes, and the edges for correlations in expression patterns indicating functional interactions.

At larger scales, the vertices can be the members of a population, and the edges social or other interactions, like mating. For a population with separate sexes, we then have a bipartite graph, that is, one with two distinct classes of elements such that edges exist only between members of opposite classes, but not inside one class.

At the still larger scale of ecosystems, the vertices can represent species, and the edges stand for trophic interactions. The graph then encodes a food web.

Another important class of biological graphs are the phylogenetic trees that turn genetic or other similarities between species into descendence relations from common ancestors. For individual descendence relations inside a sexually recombining species we rather have pedigrees because each individual then has two parents which in turn may have more than one offspring.

For detailed studies of biological networks and their properties, the reader can consult [94] and [111] and the many references therein.