Irena Roterman-Konieczna (Ed.)

Simulations in Medicine

Irena Roterman-Konieczna (Ed.)

Simulations in Medicine

Pre-clinical and Clinical Applications

DE GRUYTER

Editor
Prof. Dr. Irena Roterman-Konieczna
Jagiellonian University
Medical College, Department of Bioinformatics and Telemedicine
Ul. Sw. Lazarza 16, 31-530 Krakow, Poland
e-mail: myroterm@cyf-kr.edu.pl

This book has 247 figures and 5 tables.

ISBN 978-3-11-040626-9
e-ISBN (PDF) 978-3-11-040634-4
e-ISBN (EPUB) 978-3-11-040644-3

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Cover image: Eraxion/istock/thinkstock
Typesetting: PTP-Berlin, Protago-TEX-Production GmbH, Berlin
Printing and binding: CPI books GmbH, Leck
♾ Printed on acid-free paper
Printed in Germany

www.degruyter.com

Preface

"Simulations in medicine" – typing these words into Google produces a long list of institutions that train students in all practical aspects of medicine using phantoms. The student may learn to perform a variety of procedures and surgical interventions by interacting with a simulated patient. Such centers perform a great range of tasks related to medical education; however, medical simulations are not limited to manual procedures.

The very word "simulation" is closely tied to computer science. It involves recreating a process which occurs over a period of time. The process may include actions performed manually by a student but it can also comprise events occurring in virtual space, under specific conditions and in accordance with predetermined rules – including processes occurring on the molecular (Chapter 1) or cellular (Chapter 2) level, at the level of a communication system (Chapter 3) or organs (Chapters 4 and 5) or even at the level of the complete organism – musculoskeletal relations (Chapter 6). "Simulations in medicine" also involve recreating the decision-making process in the context of diagnosis (Chapters 7, 8, 9), treatment (Chapter 10, 11), therapy (Chapter 12), as supported by large-scale telecommunication (Chapter 13) and finally in patient support (Chapter 14).

This interpretation of the presented concept – focusing on understanding of phenomena and processes observed in the organism – is the core subject of our book and can, in fact, be referred to as "PHANTOMLESS medical simulations".

The list of problems which can be presented in the form of simulations is vast. Some selection is therefore necessary. While our book adopts a selective approach to simulations, each simulation can be viewed as a specific example of a generic phenomenon: indeed, many biological events and processes can be described using coherent models and assigned to individual categories. This pattern-based approach broadens the range of interpretations and facilitates predictions based on the observable analogies. As a result, simulation results become applicable to a wide category of models, permitting further analysis.

One such universal pattern which we will refer to on numerous occasions is the concept of an "autonomous entity". The corresponding definition is broad, encompassing all systems capable of independent operation, ensuring their own survival and homeostasis. This includes individual organisms, but also complex social structures such as ant colonies, beehives or even factories operating under market conditions. The structures associated with the autonomous operation of these entities share certain common characteristics – they include e.g. construction structures which fulfill the role of "building blocks" (Fig. 1 (a)), function-related structures which act in accordance with automation principles (Fig. 1 (b)) and, finally, structures responsible for sequestration of materials, making them compact and durable while also ensuring that they can be easily accessed when needed (Fig. 1 (c)).

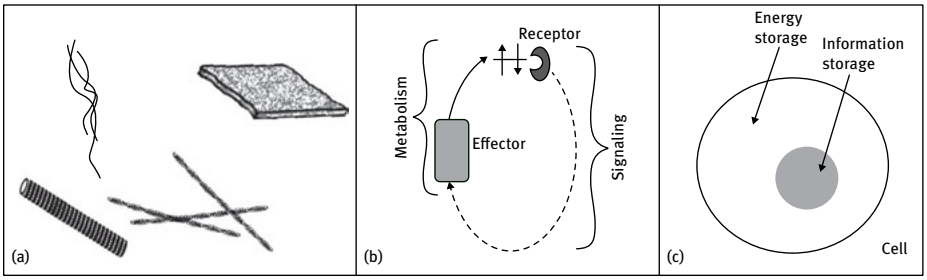


Fig. 1: Symbolic depiction of the structural and functional characteristics of the organism as an autonomous entity, comprising three basic types of components (a, b, c) corresponding to specific aims: (a) construction; (b) function; (c) storage.

Living organisms conform to the above described model. Each problem, when expressed in the form of a simulation, has its place in a coherent system – much like a newly acquired book in a library collection.

The division presented also helps explain common issues and problems relevant to each group of models. Afflictions of the skeletal system, metabolic diseases or storage-related conditions can all be categorized using the above presented schema (although some of them may affect more than one category).

Even randomly selected simulations follow this generalized model, contributing to proper categorization of biological phenomena. This fact underscores the importance of simulation-based imaging.

Journal "Bio-Algorithms and Med-Systems" published by de Gruyter invites all Readers to submit papers concerning the wildly understood spectrum of PHANTOM-LESS simulations in medicine.

You are invited to visit: http://www.degruyter.com/view/j/bams

Krakow, March, 2015 Irena Roterman-Konieczna

Contents

Part VII: **Support of therapy**

List of authors

Dr. Peregrina Arciaga
Charles R. Drew/UCLA University,
School of Medicine
1731 East 120th Street,
CA 90059 Los Angeles, USA
e-mail: peregrinaarciaga@cdrewu.edu
Chapter 8

Dr. Marian Bubak
AGH – Cyfronet
Nawojki 11, 30-950 Krakow, Poland
e-mail: bubak@agh.edu.pl
Chapter 2

Katarzyna Chłopaś – Student
Jagiellonian University – Medical College
Sw. Anny 12, 31-008 Krakow, Poland
email: katarzyna.chlopas@uj.edu.pl
Chapter 2

Prof. Jacek Cieślik
AGH – University of Science and Technology
Al. A. Mickiewicza 30, 30-059 Kraków, Poland
e-mail: cieslik@agh.edu.pl
Chapter 11

Dr. Łukasz Czekierda
AGH – University of Science and Technology
Kawiory 21, 30-055 Krakow, Poland
e-mail: luke@agh.edu.pl
Chapter 13

Jakub Dąbroś
AGH – University of Science and Technology
Łazarza 16, 30-530 Krakow, Poland
e-mail: kubadabros@gmail.com
Chapter 12

Dr. Dick Davies
Ambient Performance
43 Bedford Street, Suite 336,
Covent Garden, London WC2E 9HA, UK
e-mail: dick.davies@ambientperformance.com
Chapter 8

Dr. Parvati Dev
Innovation in Learning Stanford, USA
12600 Roble Ladera Rd,
CA 94022 Los Altos Hills, USA
e-mail: parvati@parvatidev.org
Chapter 8

Dr. Piotr Dudek
AGH – University of Science and Technology
Al. A. Mickiewicza 30, 30-059 Kraków, Poland
e-mail: pdudek@agh.edu.pl
Chapter 11

Jacek Dygut MD
Canton Hospital – Wojewodzki Hospital
Monte Casino 18, 37-700 Przemyśl, Poland
e-mail: jacekdygut@gmail.com
Chapter 12

Prof. Marc Ebner
Ernst-Moritz-Arndt University Greifswald
Institute for Mathematics and Informatics
Walther-Rathenau-Str. 47,
17487 Greifswald, Germany
e-mail: marc.ebner@uni-greifswald.de
Chapter 5

Prof. Andrzej Gackowski
Jagiellonian University – Medical College,
Cardiology Hospital
Prądnicka 80, 31-202 Krakow, Poland
e-mail: agackowski@szpitaljp2.krakow.pl
Chapter 13

Prof. Ewa Grabska
Jagiellonian University
Łojasiewicza 11, 30-348 Krakow, Poland
e-mail: ewa.grabska@uj.edu.pl
Chapter 14

Prof. Stuart Hameroff
Departments of Anesthesiology and Psychology
and Center for Consciousness Studies
The University of Arizona Tucson
Arizona 85724, USA
e-mail: hameroff@u.arizona.edu
Chapter 5

Dr. Inga Hege
Ludwig-Maximilians-University München
Ziemssenstr. 1, 80336 München, Germany
e-mail: inga.hege@med.uni-muenchen.de
Chapter 7

Dr. LeRoy Heinrichs
Stanford University School of Medicine, USA
8 Campbell Lane, CA 94022 Menlo Park, USA
e-mail: whl@stanford.edu
Chapter 8

Dr. Joanna Jaworek-Korjakowska
AGH – University of Science and Technology
Al. A. Mickiewicza 30, 30-059 Krakow
e-mail: jaworek@agh.edu.pl
Chapter 9

Dr. Wiktor Jurkowski
The Genome Analysis Centre,
Norwich Research Park
Norwich NR4 7UH, UK
e-mail: wiktor.jurkowski@tgac.ac.uk
Chapter 1

Tomasz Kawa MSc
Jagiellonian University
Łojasiewicza 11, 30-348 Krakow, Poland
Chapter 4

Prof. Leszek Konieczny
Jagiellonian University – Medical College
Kopernika 7, 31-034 Krakow, Poland
e-mail: mbkoniec@cyf-kr.edu.pl
Chapter 2

Marek Konieczny
AGH – University of Science and Technology
Kawiory 21, 30-055 Krakow, Poland
e-mail: marekko@agh.edu.pl
Chapter 13

Dr. Andrzej Kononowicz
Jagiellonian University – Medical College
Łazarza 16, 31-530 Kraków, Poland
e-mail: mykonono@cyf-kr.edu.pl
Chapter 7

Filip Malawski
AGH – University of Science and Technology
Kawiory 21, 30-055 Krakow, Poland
e-mail: fmal@agh.edu.pl
Chapter 13

Piotr Nowakowski MSc
AGH – University of Science and Technology
Nawojki 11, 30-950 Krakow, Poland
e-mail: ymnowako@cyf-kr.edu.pl
Chapter 2

Dr. Monika Piwowar
Jagiellonian University – Medical College
Łazarza 16, 31-530 Krakow, Poland
e-mail: myroterm@cyf-kr.edu.pl
Chapter 1

Prof. Irena Roterman-Konieczna
Jagiellonian University – Medical College
Łazarza 16, 31-530 Krakow, Poland
e-mail: myroterm@cyf-kr.edu.pl
Chapters 2, 3, and 14

Kornel Skałkowski MSc
AGH – University of Science and Technology
Kawiory 21, 30-055 Krakow, Poland
e-mail: skalkow@agh.edu.pl
Chapter 13

Dr. Anna Sochocka
Jagiellonian University
Łojasiewicza 11, 30-348 Krakow, Poland
e-mail: a.sochocka@uj.edu.pl
Chapter 4

Dr. Tomasz Szydło
AGH – University of Science and Technology
Kawiory 21, 30-055 Krakow, Poland
e-mail: tszydlo@agh.edu.pl
Chapter 13

Prof. Ryszard Tadeusiewicz
AGH – University of Science and Technology,
Chair of Automatics and Bioengineering
Al. A. Mickiewicza 30, 30-059 Kraków, Poland
e-mail: rtad@agh.edu.pl
Chapters 6, 9, and 10

Jakub Wach MSc
AGH – Cyfronet
Nawojki 11, 30-950 Krakow, Poland
e-mail: j.wach@cyfronet.pl
Chapter 2

Prof. Paweł Węgrzyn
Jagiellonian University
Łojasiewicza 11, 30-348 Krakow, Poland
e-mail: pawel.wegrzyn@uj.edu.pl
Chapter 14

Zdzisław Wiśniowski MSc
Jagiellonian University – Medical College
Łazarza 16, 30-530 Krakow, Poland
e-mail: mywisnio@cyf-kr.edu.pl
Chapter 12

Prof. Krzysztof Zieliński
AGH – University of Science and Technology,
Informatics Institute
Kawiory 21, 30-055 Krakow, Poland
e-mail: kz@ics.agh.edu.pl
Chapter 13

Part I: **Molecular level**

Monika Piwowar and Wiktor Jurkowski

1 Selected aspects of biological network analysis

1.1 Introduction

Much has been made of the Human Genome Project's potential to unlock the secrets of life [1, 2]. Mapping the entire human DNA was expected to provide answers to unsolved problems of heredity, evolution, protein structure and function, disease mechanisms and many others. The actual outcome of the project, however, differed from expectations. It turned out that coding fragments – genes – constitute only a minute fraction (approximately 2 %) of human DNA. Furthermore, comparative analysis of human and chimpanzee genomes revealed that despite profound phenotypic differences the DNA of these species differs by only 1.5 %. Despite being an undisputed technological tour de force, the Human Genome Project did not live up to the far-reaching hopes of the scientific community. It seems that genes alone do not convey sufficient information to explain phenotypic uniqueness – indeed, additional sources of information are required in order to maintain a coherent system under which the expression of individual genes is strictly regulated [3].

Cellular biology has historically been dominated by the reductionist ("bottom-up") approach. Researchers studied specific components of the cell and drew conclusions regarding the operation of the system as a whole [4, 5]. Structural and molecular biology reveals the sequential and structural arrangement of proteins, DNA and RNA chains. In recent years efficient technologies have emerged, enabling analysis of entire genomes (genomics) [6, 7], regulation of transcription processes (transcriptomics) [8], quantitative and qualitative properties of proteins (proteomics) [9] as well as the chemical reactions which form the basis of life (metabolomics) [10, 11]. Specialist literature is replete with breadth-first data analysis studies which are often jointly referred to as "omics" (e.g. lipidomics) [12]. The common factor of all these disciplines is the application of modern experimental methods to study changes which occur in a given cell or tissue [12].

The ongoing evolution of IT methodologies enables efficient processing of vast quantities of data and, as a result, many specialist databases have emerged. Progressive improvements in computational sciences facilitates more and more accurate analysis of the structure and function of individual components of living cells. Yet, despite the immense effort invested in this work, it has become evident that biological function cannot – in most cases – be accurately modeled by referring to a single molecule or organelle. In other words, the cell is more than merely a sum of its parts and it is not possible to analyze each part separately and then to assemble them together (like a bicycle). The fundamental phenomena and properties of life fade from focus when such a reductionist approach is applied. While an organism can be said to "operate" as determined by the laws of physics, and while it is composed of a wide variety of chem-

ical elements, it cannot be analyzed using the same tools which are successfully applied in other disciplines (e.g. linearization, extrapolation etc.) where our knowledge of the target system is complete [13, 14]. Molecules interact with one another forming a fantastically complex web of relationships. Hundreds of thousands of proteins are encoded by genes which themselves fall under the supervision of additional proteins. Genes and proteins act together to drive innumerable processes on the level of individual cells, tissues, organs and entire organisms. The end result is an enormously complicated, elastic and dynamic system, exhibiting a multitude of emergent phenomena which cannot be adequately explained by focusing on its base components [15].

The knowledge and data derived from efficient experimentation allow us to begin explaining how such components and their interactions affect the processes occurring in cells – whether autonomous or acting within the scope of a given tissue, organ or organism. This approach, usually referred to as "systems biology" has been gaining popularity in recent years. It is based on a holistic ("top-down") approach which attributes the properties of biological units to the requirements and features of systems to which they belong [3]. While a comprehensive description of the mechanism of life – even on the basic cellular level – is still beyond our capabilities, ongoing developments in systems biology and biomedicine supply ample evidence in support of this holistic methodology. Barbasi et al. [16] have conducted several studies which indicate that biological networks conform to certain basic, universal laws. Accurately describing individual modules and pathways calls for a marriage between experimental biology and other modern disciplines, including mathematics and computer science, which supply efficient means for the analysis of vast experimental datasets. This formal (mathematical) approach can be applied to biological processes, yielding suitable methods for modeling the complex interdependencies which play a key role in cells and organisms alike [17]. Such a "network-based" view of cellular mechanisms provides an entirely new framework for studies of both normal and pathological processes observed in living organisms [16, 18].

Network analysis is a promising approach in systems biology and produces good results when the target system has already been accurately described (e.g. metabolic reactions in mitochondria; well-studied signaling pathways etc.). While such systems are scarce – as evidenced by the interpretation of available results – network methods are also good at supplying hypotheses or singling out candidates for further study (e.g. interesting genes).

Existing mathematical models that find application in biology can be roughly divided into two classes based on their descriptive accuracy: continuous models, where the state of a molecule (its concentration, degree of activation etc.) and its interaction with other molecules (chemical reactions) can be formally described using ordinary differential equations (ODEs) [19, 20] under a specific kinetic model, and discrete models, where molecules exist in a limited number of states (typically two) interlinked in a directionless or directed graph. This second class includes Boolean networks, where each vertex assumes a value of 0 or 1 depending on the assumed topology and logic

[21, 22], and Bayesian networks, where the relations between molecules are probabilistic [23, 24]. As networks differ in terms of computational complexity, selecting the appropriate tool depends on the problem we are trying to solve. Boolean networks are well suited to systems which involve "on/off" switches, such as gene transcription factors which can be either present or absent, while continuous models usually provide a more accurate description of reaction kinetics where the quantities of substrates and products vary over time.

1.2 Selected biological databases

Formulating more and more precise theoretical descriptions of protein/protein or protein/gene interactions would not have been possible without experimental data supplied by molecular biology studies such as sequencing, dihybrid crossing, mass spectrometry and microarray experiments. From among these, particular attention has recently been devoted to the so-called vital stain techniques. Their application in the study of cellular processes is thought to hold great promise since they enable analysis of dynamic changes occurring in a living cell without disrupting its function. As a result, this approach avoids the complications associated with cell death and its biochemical consequences. Vital stains provide a valuable source of information which can be exploited in assembling and annotating relation networks. Such efforts are often complemented by microarray techniques which "capture" the state of the cell at a particular point in its life cycle. Microarray experiments carried out at predetermined intervals, while imperfect, provide much information regarding the relations between individual components of a cell, i.e. proteins. Such detailed data describing specific "members" of interaction networks along with their mutual relations is typically stored in specialized repositories, including:

- genomes
  - Ensembl (http://www.ensembl.org/index.html)
  - UCSD (http://genome.ucsc.edu/)
- protein data
  - Protein (http://www.ncbi.nlm.nih.gov/protein/)
  - Uniprot (http://www.uniprot.org/)
  - PDB (http://www.rcsb.org)
- microarray and NGS data
  - GEO (http://www.ncbi.nlm.nih.gov/geo/)
  - ArrayExpress (http://www.ebi.ac.uk/arrayexpress/)

1.2.1 Case study: Gene Expression Omnibus

GEO (Gene Expression Omnibus; http://www.ncbi.nlm.nih.gov/geo/) is a database which aggregates publicly available microarray data as well as data provided by next generation sequencing and other high-throughput genomics experiments. GEO data is curated and annotated so that users do not need to undertake complex preprocessing steps (such as noise removal or normalization) when they wish to e.g. review gene expression levels in patients with various stages of intestinal cancer. Additionally, the database provides user-friendly query interfaces and supports a wide range of visualization and data retrieval tools to ensure that gene expression profiles can be readily located and accessed.

Owing to its structure, GEO permits comparative analysis of results e.g. for different patients, applying statistical methods such as Student's t-test (comparison of average values in two groups) or ANOVA (comparison of a larger number of groups). Graphical representation of microarray data with color maps or charts depicting the expression of selected genes in several different experiments facilitates preliminary assessment and enables researchers to pinpoint interesting results. The database also hosts supplementary data: primary datasets obtained directly from scanning microarrays and converting fluorescence intensity into numerical values, as well as raw microarray scans (see Gene Expression Omnibus info; http://www.ncbi.nlm.nih.gov/geo/info/.)

The information present in the GEO database may be retrieved using several types of identifiers; specifically:

- GPLxxx: requests a specific *platform*. Platform description files contain data on matrices or microarray sequencers. Each platform may include multiple *samples*.
- GSMxxx: requests a specific *sample*. The description of a sample comprises the experiment's free variables as well as the conditions under which the experiment was performed. Each sample belongs to one platform and may be included in multiple *series*.
- GSExxx: requests a specific *series*. A series is a sequence of linked samples supplemented by a general description of the corresponding experiment. Series may also include information regarding specific data items and analysis steps, along with a summary of research results.

The identifiers of samples, series and platforms are mutually linked – thus, by querying for a specific microarray sample we may also obtain information on the platforms and series to which it belongs.

The GEO homepage offers access to gene expression profiles as well as sets of individual microarray samples obtained using identical platforms and under identical conditions. The repository also publishes its data via the National Center of Biotechnology Information (http://www.ncbi.nlm.nih.gov), with two distinct collections: GEO DataSet and Geo Gene Profiles. This division is due to practical reasons and a brief summary of the NCBI databases which aggregate GEO data is presented below.

GEO DataSet

The Geo DataSet database comprises data from curated microarray experiments carried out with the use of specific platforms under consistent conditions. It can be queried by supplying dataset identifiers (e.g. GDSxxx), keywords or names of target organisms. ID-based queries produce the most accurate results – keywords and names are ambiguous and may result in redundant data being included in the result set. An example of a microarray dataset (comprising a number of samples) is GDS3027 which measures gene expression levels in patients suffering from early-stage Duchenne muscular dystrophy. The study involved a control group as well as a group of patients of varying age (measured in months) (Fig. 1.1).

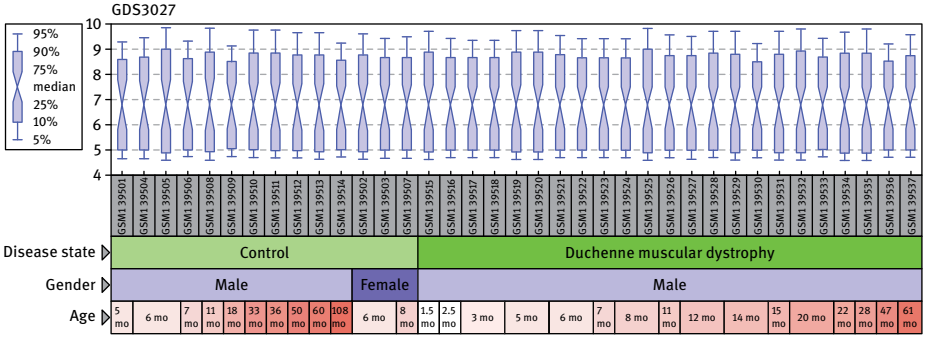


Fig. 1.1: Results of a microarray experiment involving a group of patients afflicted with Duchenne muscular dystrophy, along with a control group. GSMxxx identifiers refer to specific samples.

Graphical representation of GDS3027 results reveals the expression levels of individual genes (Fig. 1.2). Purple markers indicate high expression, green markers correspond to poor expression and grey areas indicate that no expression could be detected.

In addition, the repository aggregates data in clusters depending on the correlation between expression profiles with regard to specific samples (columns) and genes (rows).

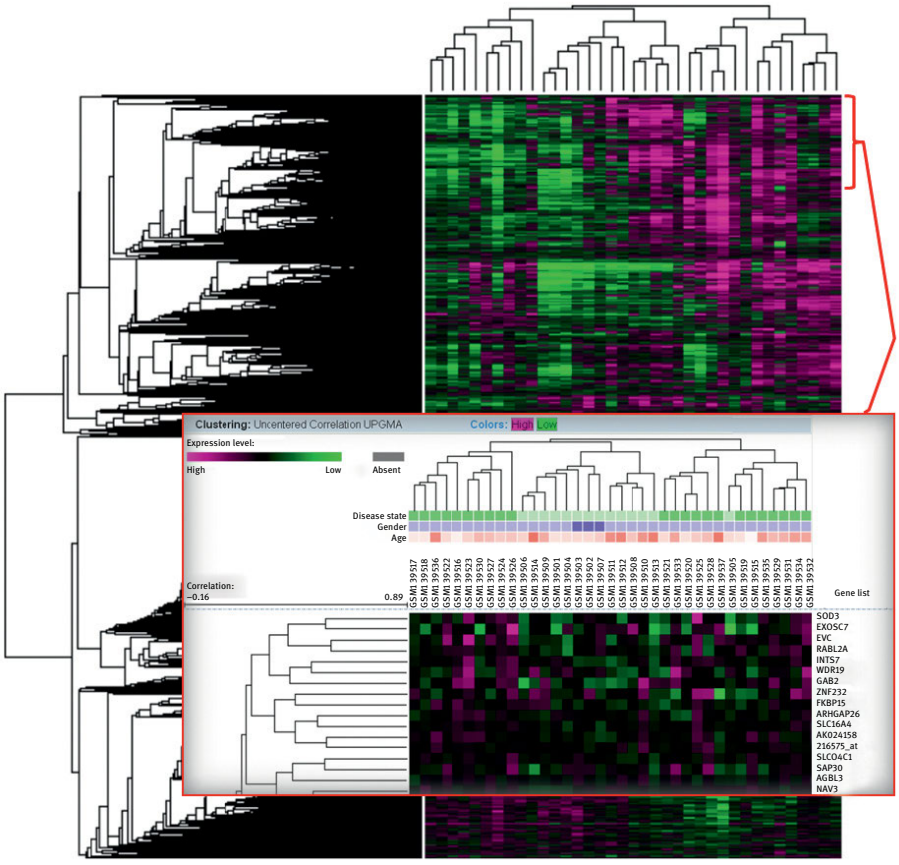


Fig. 1.2: Graphical representation of gene expression levels in the GDS3027 microarray dataset. The inset frame shows a magnified fragment of the GDS matrix. Colors correspond to expression levels: purple – high expression; green – poor expression; grey – no expression.

GEO Gene Profiles

Unlike GEO DataSet, this repository deals with expression of specific genes across a number of microarray experiments.

Gene expression levels may be "observed" under a given set of experimental conditions (such as time of study, gender or other concomitant variables) to quickly determine whether there is a connection between expression levels and any of these variables. Additionally, the database supplies links to genes with similar expression profiles. Queries can be forwarded to other databases aggregated by NCBI, e.g. to obtain additional data regarding the target sequence or protein structure. GEO Gene Profile search interfaces are roughly similar to those provided by GEO DataSet.

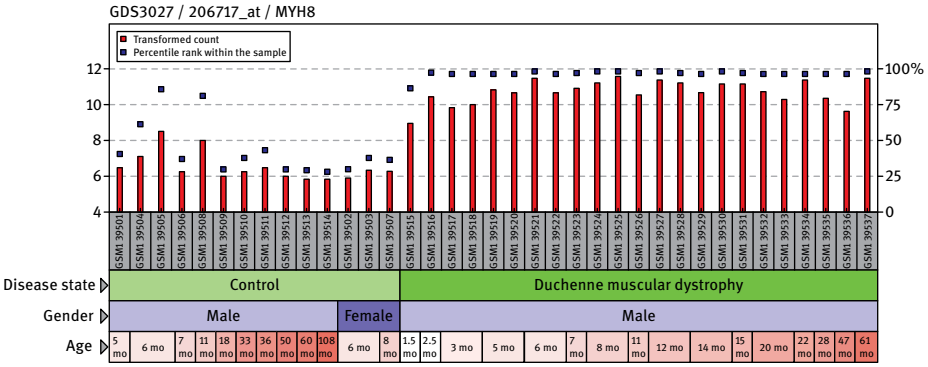


Fig. 1.3: MYH8 (myosin, heavy chain) expression profile. As shown, the expression levels of this gene are higher in the test group than in the control group.

The GDS3027 dataset includes (among others) myosin, whose expression in the test group is higher than in the control group. The corresponding GEO Gene Profile data is presented as a bar graph (Fig. 1.3).

Similar techniques can be applied to other genes. The database enables researchers to quickly discover homologues and genes with similar expression profiles (referred to as "profile neighbors"). Links to GEO DataSet profiles are also provided.

1.2.2 RegulonDB

RegulonDB is a database that focuses on the gene regulatory network of *E. coli* – arguably the most studied regulatory network [25]. The database portal provides a range of online (browser accessible) tools that can be used to query the database, analyze data and export results including DNA sequences and biological interdependence networks.

In conjunction with the *E. coli* microarray experiment results (which can be obtained from GEO), RegulonDB supports validation of regulatory network simulation algorithms.

Using RegulonDB to determine the efficiency of network construction algorithms

The main page of RegulonDB (http://regulondb.ccg.unam.mx/index.jsp) provides links to a set of search engines facilitating access to gene regulation data. The most popular engines are briefly characterized below.

- *Gene*: this interface returns data on a given gene, its products, Shine-Dalgarno sequences, regulators, operons and all transcription units associated with the gene. It also supplies a graphical depiction of all sequences present in the gene's neighborhood, including promoters, binding sites and terminators (in addition to loci which do not affect regulation of the target gene).
- *Operon*: the operon is commonly defined as a set of neighboring genes subject to co-transcription. The database introduces a further distinction between operons and transcription units, treating the operon as a set of transcription units that are shared by many genes. In RegulonDB a gene may not belong to more than one operon. A transcription unit (TU) is a set of one or more genes which are transcribed from a common promoter. TU may also provide binding loci for regulatory proteins, affecting its promoter and terminator. The search engine returns all information related to a given operon, its transcription units and the regulatory elements present in each unit. Graph visualization is provided, showing the placement of all regulatory elements within the target region. A complete set of known TUs (with detailed descriptions) is also listed below each operon.
- *Regulon*: this search interface provides basic and detailed information concerning regulons, i.e. groups of genes regulated by a single, common transcription factor. In addition to such "simple" regulons, RegulonDB introduces the notion of a complex regulon where several distinct transcription factors regulate a set of genes, with each factor exerting equal influence upon all genes from its set. The Regulon interface also shows binding sites and promoters grouped by function.

1.3 Types of biological networks

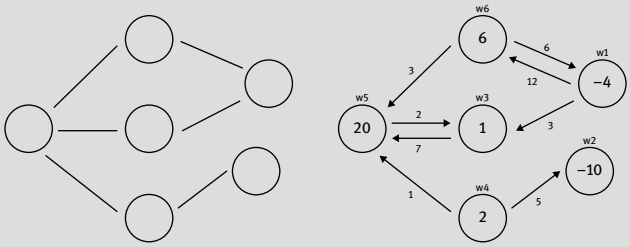
1.3.1 Relations between molecules and types of networks

Biological networks are composed of molecules: proteins, genes, cellular metabolites etc. These building blocks are linked by various types of chemical reactions. Among the simplest biological networks is the gene regulatory network (GRN) showing which genes activate or inhibit other genes. Networks are usually depicted as graphs (see inset); however this representation should not be confused with the graphical layout of networks stored in KEGG databases or wikipathways.

Graphs as a representation of networks

A **graph** is a collection of elements (called **vertices**) linked by mutual relationships (called **edges**). The interpretation of vertices and edges may vary – in gene regulatory networks vertices represent genes while edges correspond to activation/inhibition effects.

In a **simple graph** there are no **loops** (edges which connect a vertex with itself) and only one edge may appear between each pair of vertices. The maximum number of edges in a simple graph with N vertices is $N(N-1)/2$. In a **directed graph** each edge has a specific direction but there is no limit on the number of edges between each pair of vertices.



Protein-protein interaction networks are represented by simple graphs while signaling networks and gene regulatory networks usually rely on directed graphs. Metabolic networks describing reversible chemical reactions may use graphs with weighed edges – in these types of graphs each edge carries a numerical value which corresponds e.g. to its reaction rate constant. Graphs have many applications in information technology: for example they can be used for traffic modeling or Internet routing.

The most common types of vertices are genes, proteins and other molecules which participate in biochemical processes. Some networks also include cellular organelles (e.g. mitochondria, vacuoles etc.) viewed as "targets" of specific processes. The set of potential elements may be further extended with abstract concepts: UV radiation intensity, pH, ROS and any other phenomena which need to be taken into account when performing network analysis.

Relations between elements can be **direct** – e.g. a simple chemical reaction between two molecules – or **indirect** where a number of intervening reactions are necessary. An example of an indirect relationship is mutual regulation of genes. Simply observing that "gene A regulates the expression of gene B" conceals the existence of a complicated chain where the product of gene A acts upon the transcription factor or other mechanisms which, in turn, regulate the expression of gene B.

When the character of the relation is unknown, the relation is said to be **directionless**, i.e. we cannot determine which of the two interacting elements is the effector and which one is the receptor. This phenomenon occurs in many nonspecific protein-protein interactions: we may know that two proteins bind to each other but the purpose of the reaction is not known – unlike, for example, **directed** activation of adrenergic receptors via hormone complexation leading to release of protein G which, in turn, binds to its dedicated receptor. In some cases we possess knowledge not just of the relation's direction but also of its positive or negative effects.

A **positive effect** may involve upregulation of a chemical reaction by an enzyme, activation of gene expression or an increase in the concentration of some substrate. A **negative effect** indicates inhibition or simply a reduction in the intensity of the above mentioned processes.

This complex interplay of directionless and directed reactions underscores the fundamental difference between protein-protein interaction (PPI) networks which fo-

cus on nonspecific interactions between proteins, and signaling networks (SN) which provide detailed insight into biochemical processes occurring in the cell. As shown, the types of network elements and their mutual relations are directly related to the scope of our knowledge regarding biological mechanisms and the accuracy of experimental data.

1.3.2 Biochemical pathways

Several online databases store manually-validated process relationship data and visualize it by means of interaction diagrams:

- KEGG (http://www.genome.jp/kegg/)
- Reactome Pathways Database (http://www.reactome.org)
- Wikipathways (http://www.wikipathways.org/)

KEGG (*Kyoto Encyclopedia of Genes and Genomes – GenomeNet*; http://www.kegg.jp/kegg/) is a database dedicated to researchers who study the properties of molecular interaction networks on the level of cells, organisms or even entire ecosystems [26, 27].

Among the most popular features of KEGG is the presentation of molecular interactions as activity pathways (KEGG PATHWAY). The relationships between individual molecules (typically proteins) are represented as block diagrams with directed or directionless links indicating the flow of information. The number of activity pathways has grown so large that attempts are currently being made to assemble a global network consisting of various interlinked pathways (Fig. 1.4).

KEGG also includes a set of relations between individual pathway components (KEGG BRITE). This database is a set of hierarchical classifications representing our knowledge regarding various aspects of biological systems. KEGG DISEASE is an interesting database that stores molecular interaction data associated with various pathological processes in humans (http://www.genome.jp/kegg/) (Fig. 1.5).

The ability to visualize individual proteins and other molecules, along with references to detailed information regarding their properties, provides substantial help in creating network models for analysis of disease-related processes.

All KEGG databases are interlinked, permitting easy navigation between datasets.

Although KEGG is popular as a source of gene-centric information applied for instance to overrepresentation and Gene Set Enrichment analysis, KEGG has limited applicability for network analysis. The main hurdle is the heterogeneity in the style applied to represent particular pathways arising from the incompleteness of available knowledge and missing annotations. Interactions represented as a graph are often accompanied by disjoined boxes describing phenotypes or states. Some pathways are described by a set of chemical reactions and some are just lists of genes.

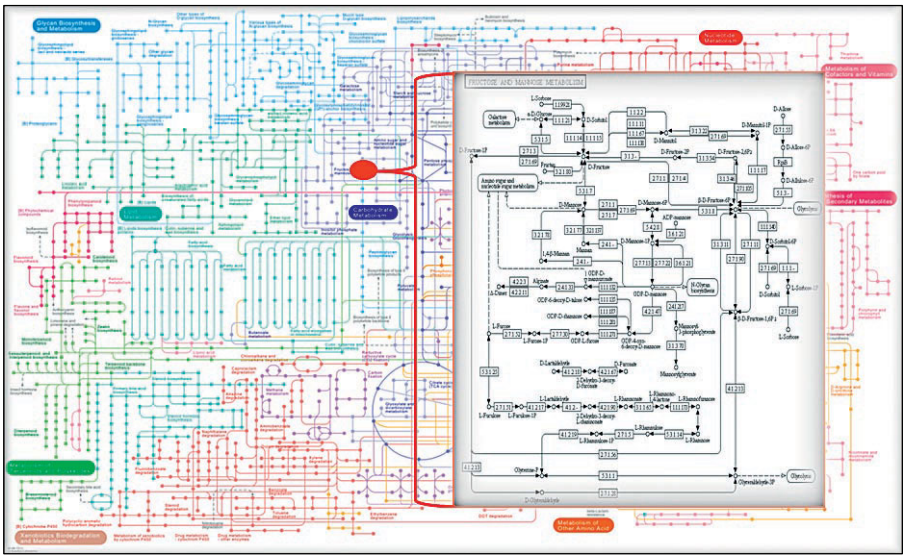


Fig. 1.4: Global activity network consisting of multiple pathways. Each dot indicates (in most cases) a single pathway. A more detailed view of a representative pathway is shown in the central part of the image, indicating stages of fructose and mannose metabolism.

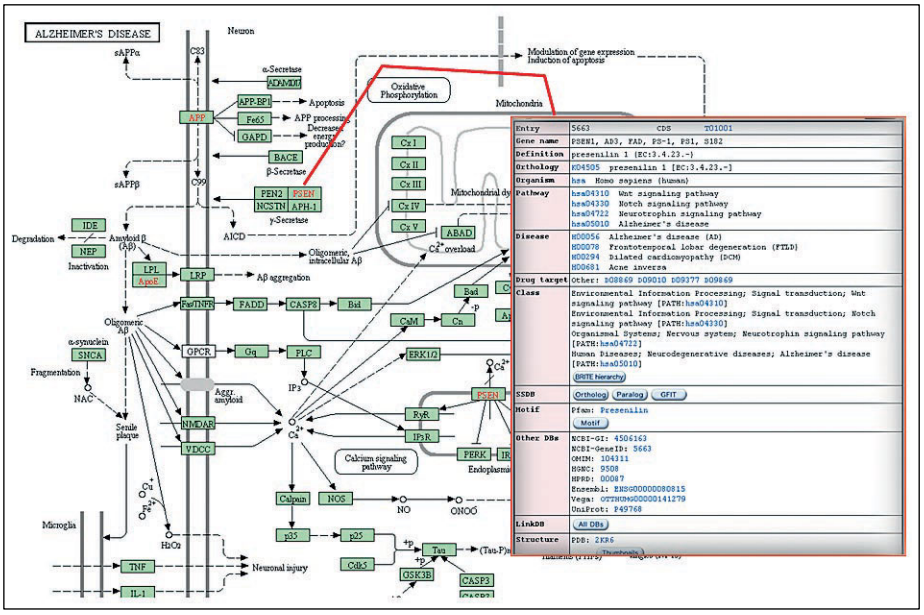


Fig. 1.5: KEGG interaction diagram corresponding to Alzheimer's disease. The red-framed inset contains detailed information concerning the protein labeled "PSEN".

Both Wikipathways and Reactome are focusing on gathering information that can be described in the form of biochemical reactions, therefore escaping the above-mentioned problems. They are much more straightforward in defining simulation models or interaction graphs.

1.4 Network development models

1.4.1 Selected tools for assembling networks on the basis of gene expression data

Assembling gene regulatory networks remains an open problem. Existing methods are not equally efficient in processing diverse datasets and it is often difficult to select the optimal algorithm for a given task. As few regulatory networks have been experimentally validated, assessment of the accuracy of hypothetical networks also poses a significant challenge. The DREAM (Dialogue for Reverse Engineering Assessments and Methods) consortium attempts to address these issues by organizing regular events where the efficiency of various network construction algorithms is independently validated (see http://www.the-dream-project.org/). This section discusses the fundamental aspects of the construction of regulatory networks based on gene expression data.

Gene Network Weaver – gene regulatory network processing software

Gene Network Weaver (GNW) provides an efficient way to determine the validity of gene regulatory network construction algorithms. This software package can read input datasets created for the purposes of the DREAM project. The first analysis step involves construction of a realistic regulatory network from known fragments of real-life interaction networks. This is followed by generation of simulated gene expression data. GNW is bundled with a number of preassembled datasets (*Escherichia coli* and *Staphylococcus* gene regulation networks along with several customized DREAM databases). The program enables users to select subnetworks in order to carry out operations on smaller and more convenient sets of data. In addition to providing its own datasets, GNW can import and parse user-generated networks [28].

Cytoscape

While Cytoscape will be presented further on in this chapter, we should note that it includes the CyniToolbox extension which can derive gene regulation networks from gene expression data [29]. Data analysis proceeds by detecting simple correlations on the basis of information theory concepts, such as mutual information and Bayesian networks. Additionally, CyniToolbox can fill in missing data and perform input discretization (as required by most processing algorithms). Similar tasks are handled by

another Cytoscape plug-in – MONET (http://apps.cytoscape.org/apps/monet). Each new version of Cytoscape comes with a range of plug-ins – up-to-date information can always be obtained on the toolkit's homepage.

GenePattern – gene expression analysis features

GenePattern is a comprehensive toolset for analyzing genomics data. Its feature analysis of genetic sequences, gene expression, proteomics and flow cytometry data. Tools can be chained into workflows to automate complex analyses. GenePattern is an open-source project and can be used free of charge for scientific purposes. User registration is required. Tools can be downloaded from the project's website and users may either set up local copies of the software or connect to one of the available public servers.

ARACNE – one of many GenePattern modules – can reconstruct cellular networks by applying the ARACNE algorithm. A thorough description of the data input format is available and data can also be imported from other modules using appropriate converters. The GEOImporter tool can download data directly from the GEO database (see Section 1.2.1). GenePattern also provides a server which recreates gene regulatory networks on the basis of selected DREAM methods, and implements a meta-algorithm assembled from the three highest ranked algorithms submitted to the most recent edition of DREAM.

1.4.2 Selected tools for reconstruction of networks via literature mining

Networks can be reconstructed by analyzing peer-reviewed publications. This process involves specification of target elements (e.g. gene symbols) and relation types (genetic regulation, protein complexation, etc.) The resulting network can be exported to a file which may then serve as input for another software package, or visualized with a GUI to enable further analysis of a specific graph edge or to prepare a presentation. The methods described in this section can be roughly divided into two groups. The first group comprises event-centric methods, e.g. searching for information on physical interactions between two proteins. This approach offers a great advantage since by focusing on the description of a biological event we avoid potentially incorrect interpretation of experiment results – although on the other hand the interpretation task is left entirely to the user. The second group covers methods which attempt to determine causative factors in intermolecular relations. This approach offers a shortcut to useful results since – in most cases – correct interpretations may have already been obtained and can aid in the reconstruction of cellular networks.

In both cases we should be mindful of the limitations inherent in combining the results of experiments carried out in various models (animals, tissues, cell lines) under differing conditions and with the use of dissimilar experimental techniques. The final

outcome of the process should be viewed with caution until it can be independently validated by a consistent series of experiments (e.g. differential gene expression analysis).

IntAct and MINT

IntAct (http://www.ebi.ac.uk/intact/) and MINT (http://mint.bio.uniroma2.it/mint/Welcome.do) contain validated interaction data for a broad set of proteins, genes and other micromolecules in various organisms, including humans. All data is traced to peer-reviewed publications presenting experimental results, and the databases only provide information on direct interactions without attempting to interpret their outcome.

The databases can be queried by publication, author credentials and proteins set, and additionally by the quality of the applied experimental methods and target organisms. Networks can be displayed or saved in one of the popular network file formats. Each relation can be traced by supplying the corresponding PubMed ID.

Pathway Studio and Ingenuity Pathway Analysis

Pathway Studio (Ariadne Genomics, http://ariadnegenomics.com) and Ingenuity Pathway Analysis (Ingenuity Systems, http://www.ingenuity.com) represent a different approach to literature mining: they subject publications to lexical analysis and submit preliminary results to a panel of experts in order to reduce the likelihood of mistakes.

Query results indicate which publications discuss the specific relation and provide information on the organisms, tissues and cells analyzed in the context of these publications.

1.5 Network analysis

The typical systems biology research process is a cycle comprising preliminary bioinformatics analysis generating new hypotheses concerning the operation of a given system followed by subsequent experiments to verify initial assumptions which can then be subjected to further analysis. The analysis of biological networks may be approached from various angles such as pathway analysis, which concerns itself with assembling rich gene ontology datasets and finding genes or biological processes overrepresented in the data under study; analysis of the flow of substrates in chemical reaction chains that allows precise quantification of perturbation; graph analysis, which seeks vertices of particular importance for a given process or cellular phenotype. Many software packages support the interpretation of biochemical data with the use of network analysis tools. This section introduces some of most popular tools.

1.5.1 Selected tools

From among the multitude of open-source and commercial network analysis packages, the following tools are particularly noteworthy: Cytoscape (www.cytoscape.org/), COPASI (www.copasi.org), Cell Illustrator (www.cellillustrator.com), and igraph (http://igraph.org/redirect.html). They permit the user to trace (among others) metabolic pathways, signaling cascades, gene regulatory networks and many other types of interactions between biologically active molecules (DNA, RNA and proteins). They also support statistical analysis and visualization of results as well as of the networks themselves. COPASI and Cell Illuminator base their simulations on a broad knowledge base which describes many important reactions in terms of differential equations. In Cytoscape and igraph biological networks are represented by graphs – in these cases the underlying reactions are not described in detail and simulations are instead based on the existence (or lack of) directed links between various molecules.

COPASI

COPASI is a noncommercial software package capable of analyzing and simulating biochemical reactions as well as any other processes which can be expressed in terms of mutual relations between entities [30]. It supports the SBML model description standard and can perform simulations using ordinary differential equations (ODEs) or Gillespie's stochastic algorithms acknowledging arbitrary discrete events.

COPASI can be used to simulate and study the kinetics of chemical reactions occurring in various zones (e.g. organelles) of the cell (Fig. 1.6). Biochemical processes are expressed as sets of reactions, using a standardized notation, with parameters such as reaction rate, stoichiometry and location taken into account. This functionality enables users to integrate various processes – chemical reaction, molecular aggregation, transport etc. The software comes with a rich set of metadata describing common reactions and, in most cases, the user only needs to select a given reaction from a list. In more complex scenarios users can define custom biochemical functions describing nonstandard reactions, along with a kinetic model expressing the relation between reagent concentrations and reaction rate. The tool also enables the user to determine where a given element can be found, which reactions it participates in and which kinetic models should be applied when simulating these reactions. Finally, COPASI can be used to define entirely new models describing phenomena other than chemical reactions.

Each reaction is assigned to a differential equation which will be used to simulate its progress. In theory this permits the user to simulate highly complex processes comprising many different reactions. In practice, however, dealing with a large set of differential equations forces the user to provide the values for many distinct parameters (e.g. on the basis of experimental data) and incorrect values may lead to nonsensical