BRNO UNIVERSITY OF TECHNOLOGY

Faculty of Electrical Engineering
and Communication

MASTER'S THESIS

Brno, 2019

Bc. Jana Musilová

BRNO UNIVERSITY OF TECHNOLOGY

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

FAKULTA ELEKTROTECHNIKY
A KOMUNIKAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF BIOMEDICAL ENGINEERING

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

SIGNALING PATHWAY FOR BUTANOL PRODUCTION IN SOLVENTOGENIC CLOSTRIDIUM BACTERIA

SIGNÁLNÍ DRÁHA PRODUKCE BUTANOLU BAKTERIÍ RODU CLOSTRIDIUM

MASTER'S THESIS
DIPLOMOVÁ PRÁCE

AUTHOR AUTOR PRÁCE	Bc. Jana Musilová
SUPERVISOR VEDOUCÍ PRÁCE	Mgr. Ing. Karel Sedlář

BRNO 2019

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

Master's Thesis

Master's study field **Biomedical Engineering and Bioinformatics**
Department of Biomedical Engineering

Student:	Bc. Jana Musilová	***ID:***	173572
Year of study:	2	***Academic year:***	2018/19

TITLE OF THESIS:

Signaling Pathway for Butanol Production in Solventogenic Clostridium Bacteria

INSTRUCTION:

1) Prepare a literature review of signaling pathways modelling by means of systems biology, including commonly used tools, databases, and data formats. 2) Study the possibility of using lab data for inference of new models of signaling pathways or for refinement of known models. Aim on lab techniques for measuring gene expression. 3) Using a suitable tool, e.g. Cell Colective, and the knowledge gathered from literature and public databases, design a basic model of a signaling pathway involved in butanol production in solventogenic Clostridia. 4) Refine the model with results of gene expression analysis of the strain Clostridium beijerinckii NRRL B-598. 5) Perform a static and dynamic analysis of modelled signaling pathway. 6) Discuss the results.

RECOMMENDED LITERATURE:

[1] HELIKAR, Tomáš, Bryan KOWAL, Sean MCCLENATHAN, et al. The Cell Collective: Toward an open and collaborative approach to systems biology. BMC Systems Biology. 2012, 6(1), 96-.

[2] SEDLAR, Karel, Pavlina KOSCOVA, Maryna VASYLKIVSKA,et al. Transcription profiling of butanol producer Clostridium beijerinckii NRRL B-598 using RNA-Seq. BMC Genomics. 2018, 19(1), 415-.

Date of project specification:	4.2.2019	***Deadline for submission:***	17.5.2019

Supervisor: Mgr. Ing. Karel Sedlář
Consultant:

prof. Ing. Ivo Provazník, Ph.D.
Subject Council chairman

WARNING:

The author of the Master's Thesis claims that by creating this thesis he/she did not infringe the rights of third persons and the personal and/or property rights of third persons were not subjected to derogatory treatment. The author is fully aware of the legal consequences of an infringement of provisions as per Section 11 and following of Act No 121/2000 Coll. on copyright and rights related to copyright and on amendments to some other laws (the Copyright Act) in the wording of subsequent directives including the possible criminal consequences as resulting from provisions of Part 2, Chapter VI, Article 4 of Criminal Code 40/2009 Coll.

ABSTRAKT

Diplomová práce se zabývá studiem signální dráhy produkce butanolu bakterií rodu *Clostridium*. V první části pojednává o modelování signálních drah pomocí metod systémové biologie. Navazuje popisem zisku dat pro tvorbu a úpravu modelů signálních drah s hlavním zaměřením na techniky pro zjištění genové exprese, produkce a fenotypu. Třetí sekcí je získání základního modelu signální dráhy zapojené do produkce butanolu u solventogenních klostridií. Posledním bodem a zároveň hlavním cílem je vytvoření dynamického modelu signální dráhy produkce butanolu kmene *Clostridium beijerinckii* NRRL B-598, jeho vyhodnocení pomocí statické a dynamické analýzy a srovnání s biologickými daty.

KLÍČOVÁ SLOVA

signální dráhy; dynamický model; klostridie; butanol

ABSTRACT

The diploma thesis is dedicated to studying signaling pathway for butanol production in *Clostridium* bacteria. The first part addresses the signaling pathways modeling by means of systems biology. The thesis follows the description of the data acquisition for signaling pathways modeling and modifying with the main focus on techniques for the detection of gene expression, products and phenotype. The third section is to obtain a basic model of a signaling pathway involved in butanol production in solventogenic clostridia. The final point and main goal is to create a dynamic model of butanol production signaling pathway in the *Clostridium beijerinckii* NRRL B-598 strain, its evaluation by static and dynamic analysis and comparison with biological data.

KEYWORDS

signaling pathways; dynamic model; clostridium; butanol

MUSILOVÁ, J. *Signaling Pathway for Butanol Production in Solventogenic Clostridium Bacteria*. Brno, 2019. Available from: https://www.vutbr.cz/studenti/zav-prace/detail/118355. Master's Thesis. Brno University of Technology, Faculty of Electrical Engineering and Communication, Department of Biomedical Engineering. Supervisor Karel Sedlář.

DECLARATION

I declare that I have elaborated my diploma's thesis on the theme of "Signaling Pathway for Butanol Production in Solventogenic Clostridium Bacteria" independently, under the supervision of the diploma's thesis supervisor and with the use of technical literature and other sources of information which are all quoted in the thesis and detailed in the list of literature at the end of the thesis. As the author of the diploma's thesis I furthermore declare that, concerning the creation of this diploma's thesis, I have not infringed any copyright. In particular, I have not unlawfully encroached on anyone's personal copyright and I am fully aware of the consequences in the case of breaking Regulation S 11 and the following of the Copyright Act No 121/2000 Vol., including the possible consequences of criminal law resulted from Regulation S 152 of Criminal Act No 140/1961 Vol.

Brno May 17th 2019 Jana Musilová

ACKNOWLEDGEMENT

I would like to thank my supervisor Mgr. Ing. Karel Sedlář Ph.D. for his expert advice, help and the time he devoted me during elaborating my thesis thorough the whole year.

Brno May 17th 2019 Jana Musilová

CONTENTS

INTRODUCTION

The production of biofuels as a renewable resource is becoming a popular topic, especially due to limited natural resources and efforts to protect the environment. The use of butanol-producing bacteria (i.e. solventogenic) seems to be an ideal solution, as they are undemanding and under certain conditions may be able to produce large amounts of gas. Solventogenic clostridia can be fed with waste materials and, with the successful genetic mutation, could be able to produce a very high amount of butanol.

The understanding and the possibility of subsequent mutation of the cell processes leading to butanol production requires a detailed knowledge of signaling pathway influenced to the production of this liquid, therefore the aim of the thesis is to describe the signaling pathway involved in butanol production in the promising solventogenic strain *C. beijerinckii* NRRL B-598.

The first three parts of the diploma thesis will be focused on the literary research of data acquisition and signaling pathways modeling. Section 1 will be aimed on biological networks, graph theory and systems biology. Part two will describe signaling pathways, mathematical models, tools, databases, data formats and data analysis using for modeling of signaling pathways. Data acquisition for signaling pathway modeling with the main focus on lab techniques for the detection of gene expression, such as RNA-Seq, microarray and blotting will be described in the third section as well as the detection of gene products and phenotype.

Last part of the thesis will give a preview of butanol-producing clostridia, especially signaling pathway for butanol production will be depicted: a comparison of signaling pathway of *C. acetobutylicum* and *C. beijerinckii*, five pathways of *C. acetobutylicum* and a genome-scale model of *C. beijerinckii*. General pathways description will be followed up by the main point of the thesis – signaling pathway for butanol production in *Clostridium beijerinckii* NRRL B-598, the creation of the dynamic model, its simulation, static and dynamic analysis as well as the comparison of the model with biological data and its statistical evaluation.

1 BIOLOGICAL NETWORKS

Living organisms, also labelled as biological systems, are complex and organized units made of molecules. Molecules are highly connected and therefore organisms reach immense complexity due to small set of molecules [1].

If we want to understand these systems and their connectivity, we need more than to study organisms in laboratory experiments. For this reason, virtual models of systems are made based on mathematical relationships, especially graph theory, which will be described in the subchapter below.

Models of biological systems are represented as networks [2] composed of nodes and edges (see Fig. 1). Nodes represent molecules, genes, proteins or other units and edges represent relation between nodes. Depending on the model part of the system, we can divide networks into several types – signaling pathways, interactions of proteins or drugs, metabolic networks, etc. [3].

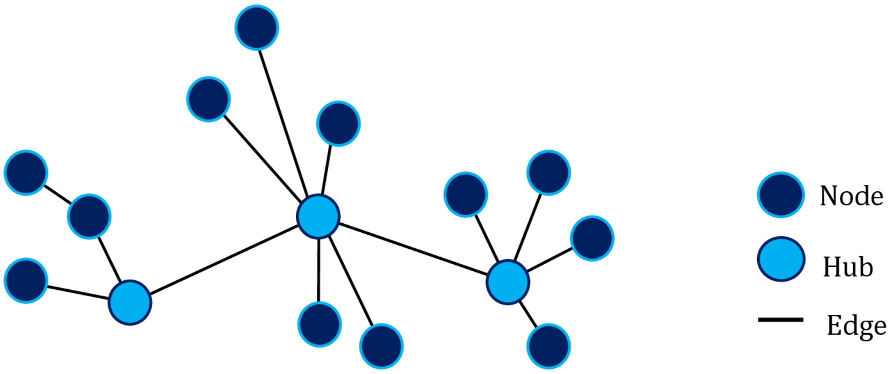


Fig. 1: Network example containing nodes, hubs and edges

Biological networks have certain properties that distinguish them from random networks. Particularly precise interaction and regulation thousands of molecules [2]. The number of possible interactions is given by combinatorial explosion, so we are not able to deduce the behaviour of the whole system from the individual parts. This property results from the non-linearity of the system [4]. Non-linear systems do not have the superposition principle, so the behaviour of the system is not given by the sum of the partial properties, which will appear as an emergent property.

The concept of emergence was first explained at 1843 by John Stuart Mill and says that whole system is more that only individual parts. For example, cardiac cells form the heart that pumps blood into the body, but the cell itself is unable to do this function [5].

Most networks, mainly constructed from biological data, share features described as scale-free behaviour [5] described in certain properties [1]: **(1)** Power law: most nodes have few connections with hubs (vital nodes) being highly interconnected (see Fig. 1). **(2)** self-similarity: an individual part of the network is similar to any other part. **(3)** small-world behaviour: two nodes can be connected via small number of other nodes and most of two nodes connected to the other node are also connected. **(4)** robustness: networks show a high degree of robustness – most nodes can be removed or disturbed with only local changes and any damage to network behaviour, but remove or damage of a hub can destroy the entire network [6, 7].

1.1 Graph theory

Graph theory is the study of graphs leaning on mathematic. Graphs are representing elements defined as an ordered pair *G = (V, E)* where *G* is graph, *V* is a set of vertices or nodes and *E* is a set of edges [8]. For given elements equation 1.1 applies [9]:

$$E \subseteq V \times V \; or \; E \subseteq \big\{\{u, v\}\big|u, v \in V, u \neq v\big\} \qquad (1.1)$$

Probably the first task in graph theory called Seven Bridges of Königsberg comes from Leonhard Euler. He proved in 1736 that it is not possible to find a way across all seven bridges so that everyone could only cross once because bridges do not form a Euler graph, which means that the path cannot be drawn in one stroke [8].

Depending on the type of edges, graphs are divided into several types [10]:

1. ***Undirected graphs:*** relation between *u* and *v* takes forms of disordered pair *e = {u, v}*. They represent symmetrical relations between nodes and express bilateral relations. An example of this graph is in the Fig. 2, L.
2. ***Directed graphs:*** relation *u* and *v* takes forms of ordered pair *e = (u, v)*. An edge *(u, v)* in directed graph begins in *u* node and ends in *v* node. The reverse edge *(v, u)* is different from *(u, v).* These graphs represent unidirectional, unsymmetrical relations between nodes and express unilateral relations. An example of this graph is in the Fig. 2, M.
3. ***Mixed graphs*** contain oriented and non-oriented edges. They are almost unused and replaced by directed graphs so that non-oriented edges are replaced by a pair of oppositely oriented edges (see Fig. 2, M). An example of this graph is in the Fig. 2, R.

Another division of graphs is based on the number of edges between nodes. Simple graphs have one edge connecting two nodes (see Fig. 2, L). Multigraphs have more than one edges between two nodes (see Fig. 2, M). Pseudographs contain a loop (a node is connected to itself).

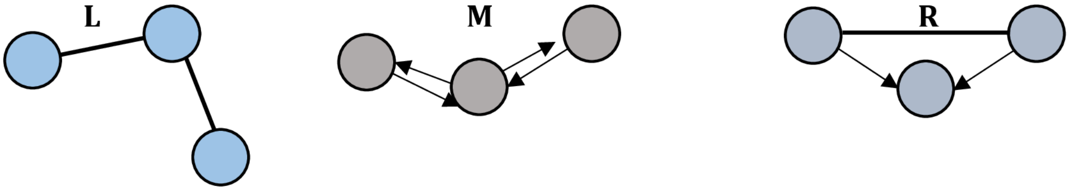


Fig. 2: Types of graphs according to type of edges.

L: undirected simple graph; M: directed multigraph; R: mixed simple graph

Subgraph *G′*, where *G′ = (V′, E′)* is a subset of graph *G*, where *G = (V, E)* if $V' \subseteq V$ and $E' \subseteq E$ [8, 9]. Statistically significant subgraphs are called motifs [11]. Motifs can be found in biological networks as a small subnet containing about three to five nodes [10].

Biological networks are usually presented as simple graphs. Nodes often contain loops and edges are weighted (indicate specific parameters related to the network). Directed graphs include networks: gene regulatory, metabolic, signaling pathway. Protein networks are usually undirected [2].

Isomorphism [9] is a sign for graphs that have the same number of edges and nodes. Graphs *G* and *H* are isomorphic if there is an isomorphism between them: $G \simeq H$. The isomorphism of graphs *G* and *H* is a bijective (i.e. mutually unambiguous) representation $f\colon V(G) \to V(H)$, for which is true that every pair of nodes $u, v \in V(G)$ is connected by an edge in *G* just when the pair *f(u), f(v)* is connected by an edge in *H* [8].

Graphs can be described by specific features such as sequence, trail, path and circle [8, 10]. The sequence denotes such a consecution of nodes where there is an edge between two adjacent peaks. The stroke is a sequence in which edges are not repeated. Specific cases are the Eulerian path (the path is a sequence in which nodes are not repeated) and the Hamiltonian path (the path going through all nodes). The circle is a path that begins and ends at the same node.

Node properties are described by neighborhood and degree [8, 9]. As neighborbood are named two nodes connected by an edge. Degree of the node is the size of its neighborhood (see Fig. 3). In case of directed graphs, we can divide input degree (number of edges entering the node) and output degree (number of edges exiting the node).

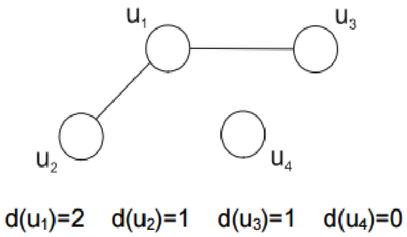


Fig. 3: Undirected graph depicting the degree of nodes [10]

Implementation of a graph into a computer is possible by two basic ways: adjacency matrix and adjacency list [8]. Adjacency matrix is a 2D array *g[][]* of size *V×V* where *V* is the number of nodes in graph. *g[i][j] = 1* indicates an edge between nodes *i* and *j*. Adjacency list is a 2D array *h[][]* and degree array *d[].* List of neighbors *i* is given by elements *h[i][0], h[i][1], ..., h[i][d[i] – 1]*.

Graph search is a graph analysis method used mainly in analysing all nodes. For example, finding the shortest path in robotics, route-planning or game-playing. There are two basic algorithms: Breadth First Search (BFS) and Depth First Search (DFS) [10]. Next algorithm of graph search is the Dijkstra's algorithm [8].

BFS uses the data structure FIFO (first in, first out). The basis is an initialization (inserting first element into a vector), next elements are sorted to the end of a vector. The first selected item is that one at the beginning of the vector and the shortest distance of all nodes from the currently selected node is calculated. The vector contains all nodes in graph in the beginning of searching and therefore has the same length as number of nodes in searched graph.

DFS algorithm passes all nodes sequentially, from the current to the neighboring element. Each node must be passed only once, for which auxiliary variables are used. It must be created a stack – a set of nodes that cannot be accessed immidiately. For example, if a node has two neighbors, one of them is stored in the stack and used later when the current node does not have a neighbor.

Dijkstra's algorithm is similar to BFS, but more complex and faster. It is often used to search for train or bus connections or in GPS navigation. Instead of BFS, Dijkstra's algorithm stores information about the shortest sequence length moreover. It chooses from the vector a node with the smallest distance because it is no better possible to reach this node. At the end of processing, these distance variables indicate the shortest distances from beginning to the other nodes.

A graph is an abstract concept. For illustration, there are several graph representation methods: diagram (see Fig. 3), definition (mathematical notation), matrices, lists and data structures [10]. The first two methods are computationally demanding, time-consuming and for large amounts of data are not effective, therefore they will not be discussed in detail.

Matrix representation is based on two basic relationships in the graph. The first one is the relationship between an edge and its end node called the relationship of incidence [10]. Incidence matrix I is a type $\{-1, 0, 1\}^{n \times m}$ if it is true that $/V/ = n$ and $/E/ = m$. The equation 1.2 applies [12]:

$$I_{v,e} = \begin{cases} -1, & e = (v,\cdot) \\ 1, & e = (\cdot,\mathrm{v}) \\ 0, & otherwise \end{cases} \tag{1.2}$$

The second one is the node adjacency describing adjacency matrix graph. Adjacency matrix A is a type $\{0, 1\}^{n \times m}$ if it is true that $/V/ = n$ and 1.3 applies [10]:

$$A_{u,v} = 1 \leftrightarrow \{u, v\} \in E \tag{1.3}$$

There are also distance matrices – edges are rated, instead 1 is given weight.

List representation is given by a list of neighbors [9]. This representation is advantageous mainly for sparse graphs for which applies: $/V/ = n$, $/E/ = m$ and $m \ll n^2$ [13]. Nodes are arranged into the array of size n and in the i-th element of this array is the pointer to linked list of nodes that are adjacent to the node I.

Graph representation as a data structure is used for the largest graphs [10]. The most used is representation using arrays. Structure of the graph is stored in two arrays. The first array has the same number of elements as the number of nodes in the graph. Each node corresponds to one array element. It holds an index value from which begins the list of nodes (neighbors of this node) in the second array.

1.2 Systems biology

Systems biology (SB) is a scientific field connecting several disciplines such as mathematics, biology, physics, engineering, informatics, medicine and chemistry. Subject of the SB studies are interactions in systems, not components description.

SB is based on holism, which means that it views the system as a complex and studies emergent properties. By contrast, bioinformatics is based on reductionism, which means that it studies individual components or individual interactions. Although these fields look at the system differently, they are usually focusing on similar themes as DNA, RNA, function of the organism, etc.

Fig. 4 shows two main approaches in SB. The first one – top-down approach studies system as a whole and decomposes it into smaller parts. It is usually used for understanding the system's behaviour, but does not describe the basic elements in detail. Bottom-up approach studies basic elements and compose it to larger units. It is mainly used to describe parts of the system.

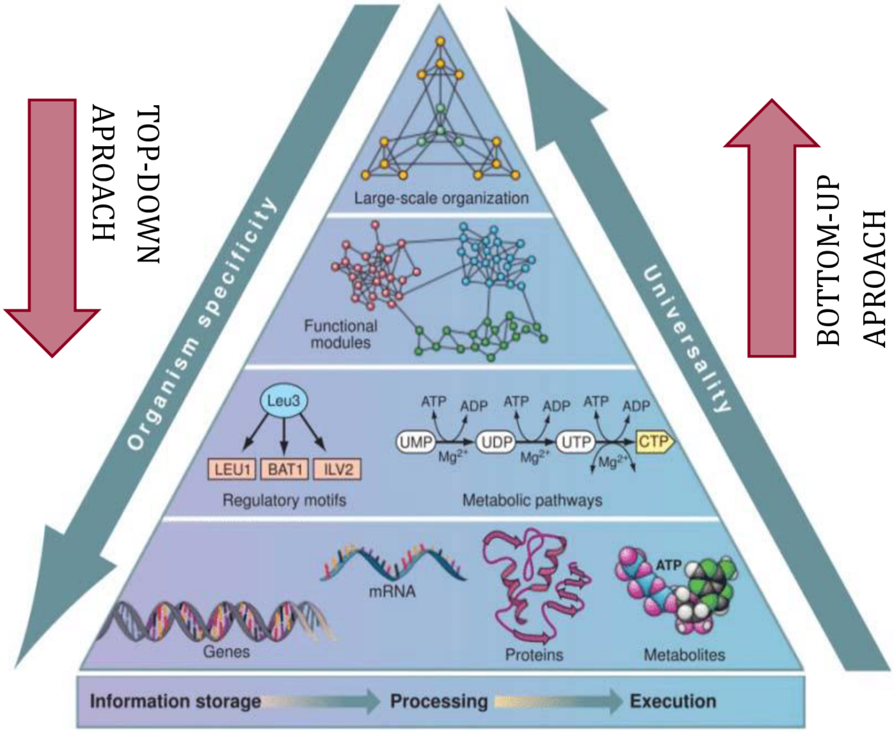


Fig. 4: Life's complexity pyramid. Top-down and bottom-up approach. [14]

We can divide several levels of view on the biological system [15]:

1. **Structure of the system:** basic organization of the system. The structure consists of elements of the network (genes, mRNA, proteins, metabolites, …), interaction between elements and associated parameters. Includes gene, metabolic, signal transduction networks and physical structures.
2. **Dynamics of the system:** behaviour of the system with known structure. Includes analysis as steady-state, flux balance (FBA), metabolic control; furthermore intracellular versus extracellular view, behavior in extreme conditions (temperature, pressure, starvation, …).
3. **Control of the system:** targeted changes to the structure or behaviour of the system. Includes drug design and genetic modification.
4. **Methods to design and modify the system:** e.g. "Signaling Pathway for Butanol Production in Solventogenic Clostridium Bacteria".

The historical development of the SB is illustrated in the Fig. 5. As will be mentioned, a boom in systems biology started about 2000 [16], mainly based on the development of powerful computers, genome sequencing and analysis.

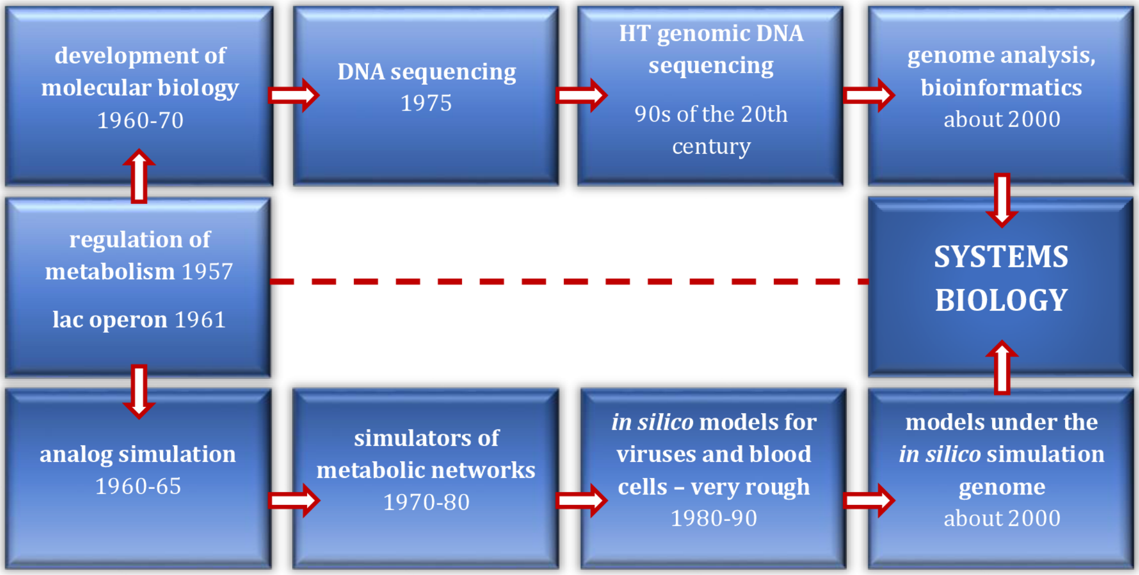

development of molecular biology 1960-70
DNA sequencing 1975
HT genomic DNA sequencing 90s of the 20th century
genome analysis, bioinformatics about 2000
regulation of metabolism 1957
lac operon 1961
SYSTEMS BIOLOGY
analog simulation 1960-65
simulators of metabolic networks 1970-80
in silico models for viruses and blood cells – very rough 1980-90
models under the *in silico* simulation genome about 2000

Fig. 5: Historical development of SB

Discipline was firstly proposed at 1998 in an article „Systems Biology: new opportunities arising from genomics, proteomics and beyond" [17] written by an American biologist Leroy Hood. He defined it as the science that studies all components and their interactions in biological systems. The greatest progress in the field of systems biology started from 2000, but interest in this science is still growing as the Fig. 6 shows.

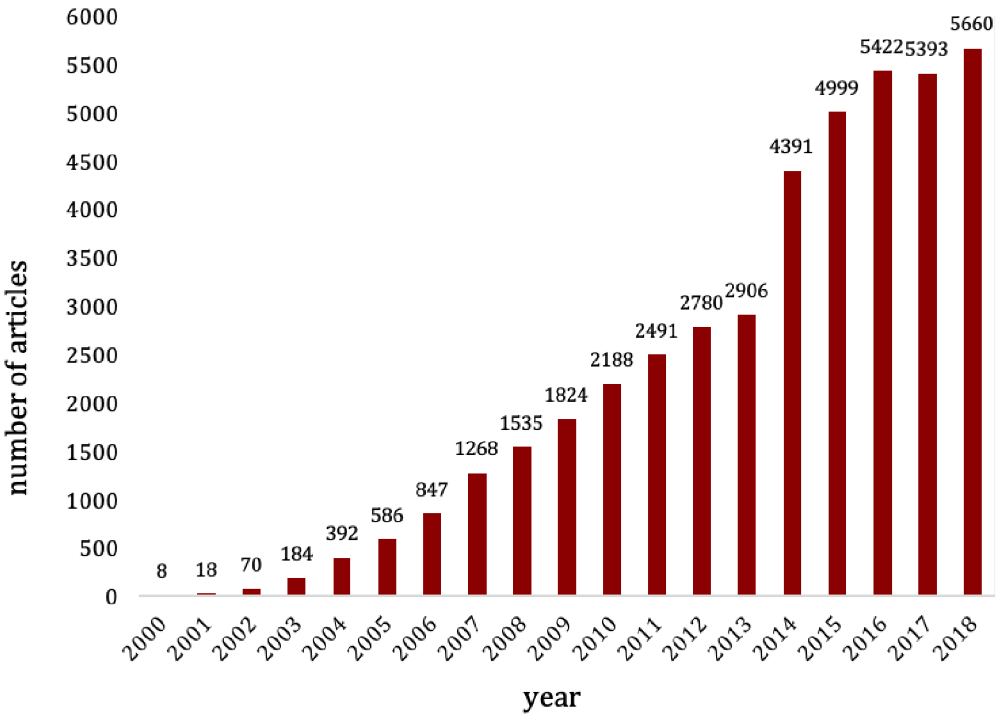


Fig. 6: Number of articles in PubMed containing phrase "systems biology"

2 SIGNALING PATHWAYS

Signaling pathways (SP) are made by a group of molecules that cooperate to control a specific cell function (examples are given in the next paragraph). The first receives one molecule signal in a pathway and passes the information to the next molecule. This process is repeated until the last molecule has this information and the specific function is carried out [18].

There are many types of signaling pathways. The most significant include [19]: the Akt SP playing a key role in the mediation of protein synthesis, metabolism, proliferation and cell cycle progression. The AMPK SP is used in the cellular response to low levels of available ATP (adenosine triphosphate). The apoptosis SP is a process for cell death. The MAPK SP – the mitogen-activated protein kinase pathways are important for the response to extracellular stimuli such as heat and stress. Signaling pathway for butanol production is a specific kind of the pathway that some organisms have (namely solventogenic bacteria) and will be resolved in the following text.

2.1 Mathematical models

Mathematical model of the signaling pathway is a simplified description of a real object designed for a better understanding of this object. It describes dynamics and structure of the system. Thanks to these models, it is possible to simulate situations that would be difficult or impossible to implement in real. They allow systemic description of process in a system and determination of regulatory process of a pathway of interest.

There are two approaches of modeling pathways: qualitative and quantitative methods [20]. Qualitative modeling is aimed at the structural organization and is mainly applied to the reconstruction of the pathways. Quantitative modeling should describe concentration and location of each component.

2.1.1 Qualitative methods

Qualitative models are created primarily for large networks that provide an overview of the dynamics. From these networks, we get information about the system state. S. Kaufmann and R. Thomas are considered as the pioneers of using logical models in biology. Their approach allows the fact that different signal intensities may exert different effects on target [20], which corresponds to basic approach of SB – non-linearity of the system.

There are several basic methodologies for qualitative modeling: interaction graphs, Boolean/logical networks, logic-based ordinary differential equations (ODE) and Petri nets. Interaction graphs allow the identification of important properties such as significant paths or feedback loops [21], but they are very simple, so they will not be further explored. Other methods are described below.

Boolean models are used mainly for large-scale signaling networks [22] to study the basic input-output behaviour of the system and to analyse its dynamic properties. Their components (usually nodes) have only two discrete states: 1 ("on" or "activated") and 0 ("off" or "inactivated"). This restriction to two states is a crude simplification of a system but regulatory interactions are often of sigmoidal shape: a regulator R has no or only little effect to a target T until R reaches a threshold θ. After θ is reached, B quickly rates its activation/synthesis rate. It means for B that A is inactive (or absent) when $A < \theta$ and A is active when $A > \theta$ [21]. Time is discretized and the state at time $t+1$ is a function of the component at time t as described in the equation 2.1 [22]:

$$\sigma_i(t+1) = sgn \sum_{j=1}^{k_{in}(i)} \left(J_{ji}\sigma_j(t) - \theta_i\right) \tag{2.1}$$

where i is the node represented by binary state σ_i, $i = 1, 2, \ldots, N$ in time t; $\sigma_1(t), \ldots, \sigma_N(t)$ is node activity pattern; $k_{in}(i)$ are other interactive nodes to the node i; J_{ji} is the interaction strength from j to i; $sgn(x)$ is the unitary step function; θ_i is the activation threshold of i [23]. In more general approaches of logic models, variables can have any number of discrete or continuous variables (fuzzy logic models).

Logic-based ODE [21] are the intermediate step between qualitative and quantitative models. They are transformed from Boolean networks and ODE models and allow studies on qualitative and dynamic features of a network, where time and states are continuous. As can be seen in the Fig. 7, logic-based ODE models are formatted based on the logical models (such as Boolean ones) and interaction graphs. Piecewise-linear differential equations or hybrid modeling is a method to form a continuous model and was firstly described in 1973 by Glass and Kauffman. It is a step function based on a sigmoidal shape of regulatory interactions. Logic-based ODE derived from multivariate polynomial interpolation transform logical models into systems of continuous differential equations that are derived from Boolean models without any more knowledge. Boolean variables and functions are replaced by continuous elements: Boolean $x_i \in \{0,1\}$ is replaced to $\bar{x}_i \in [0,1]$ which represents the normalized continuous variable of the i-th node. Discrete function B_i is replaced into continuous $\bar{B}_i^B$. Alternativelly to the piecewise-linear differential equation's sigmoid shape, polynomial interpolation uses the Hill function.

Hill function takes the form described in 2.2:

$$h(x) = \frac{x^n}{x^n + k^n} \tag{2.2}$$

where n is the Hill coefficient defining the steepness of the function; k is the activation level of a node x.

The relationship between modeling methods and the transition from the qualitative model to the quantitative is shown in the Fig. 7. Every Boolean/logical model is based on the interaction graph because it was created based on it and every logical-based ODE was also created from Boolean (logical) model. The most important thing is the retaining of systems and network properties when moving from simple to more complex model.

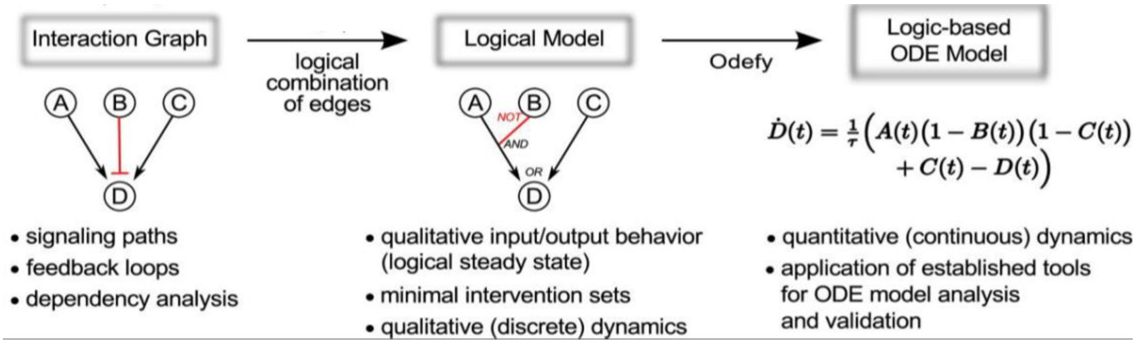

Interaction Graph
logical combination of edges
Logical Model
Odefy
Logic-based ODE Model
A
B
C
D
NOT
AND
OR
$\dot{D}(t) = \frac{1}{\tau}\Big(A(t)(1 - B(t))(1 - C(t)) + C(t) - D(t)\Big)$
• signaling paths
• feedback loops
• dependency analysis
• qualitative input/output behavior (logical steady state)
• minimal intervention sets
• qualitative (discrete) dynamics
• quantitative (continuous) dynamics
• application of established tools for ODE model analysis and validation

Fig. 7: Modeling: from qualitative information to quantitative model [21]

Petri nets [21] are an alternative to logical modeling. They are used mainly for large-scale networks. Petri nets are directed bipartite graphs made of nodes, places (usually nodes) and transitions (represent reactions or edges). Each transition has a certain number of input and output elements. The dynamic of the network is described by tokens: each place holds zero or a positive number of tokens. Transition can take place when all inputs of a transition carry certain number of tokens (or more). Input tokens are consumed and new tokens are generated in the output.

2.1.2 Quantitative methods

Quantitative models are exploited to large-scale networks behaviour prediction. They are often used as well for many analysis because they can show, for example, elements that are critically important and which ones are important less [23] or which edge is the most important to dissemination of information. Models are usually written in a form of mathematical equations.

The simplest models describe just the dependence of one element to another and they can be expressed by an algebraic equation described in 2.3:

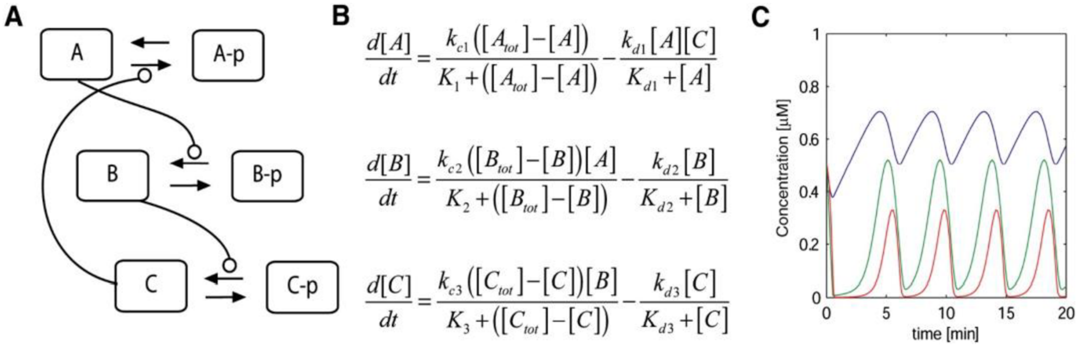
$$A_n + A_1 \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} A_{n+1} \qquad K = \frac{k_{off}}{k_{on}} \tag{2.3}$$

where A_1 and A_n denote the monomer and n-mer; K is the constant [24].

Deterministic temporal dynamics is used for more complex problems and it is described by ordinary differential equations (ODE) (see Fig. 8). ODE are the most common type of models used for cells signaling. An expansion of the ODE are partial differential equations (PDE) used mainly to model spatially heterogeneous dynamics [24]. Both models are derived from the Michaelis-Menten equation (see the equation 2.4):

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \overset{k2}{\rightarrow} E + P \tag{2.4}$$

where E is an enzyme; S is a substrate; k are rate constants; ES is an enzyme-substrate complex and P is a product of the enzymatic reaction [25].



$$\frac{d[A]}{dt} = \frac{k_{c1}([A_{tot}]-[A])}{K_1+([A_{tot}]-[A])} - \frac{k_{d1}[A][C]}{K_{d1}+[A]}$$

$$\frac{d[B]}{dt} = \frac{k_{c2}([B_{tot}]-[B])[A]}{K_2+([B_{tot}]-[B])} - \frac{k_{d2}[B]}{K_{d2}+[B]}$$

$$\frac{d[C]}{dt} = \frac{k_{c3}([C_{tot}]-[C])[B]}{K_3+([C_{tot}]-[C])} - \frac{k_{d3}[C]}{K_{d3}+[C]}$$

Fig. 8: ODE model of a simple network [24]

Left (A): a simple network; middle (B): ODE model with Michaelis-Menten expressions; right (C): simulation

2.2 Tools

Signaling pathways, as well as other large-scale networks, are complex and extensive models and it would be difficult or impossible to describe all parts of them without computer science. Therefore, many tools are created to facilitate network research. In addition, there are situations that cannot be implemented in real but can be made *in silico* (i.e. using computer model of living organism).

There are many tools for visualisation and analysis of networks that differ in availability, functions or graphic user interface. Here is an example of freely available tools suitable for visualization and analysing signaling pathways that will be described in more details: Cell Collective, Cytoscape, Gephi, GINsim, SQUAD, VisANT, PathVisio and Reactome.

Cell Collective [26] is a web-based platform created by T. Helikar's team at Omaha's University of Nebraska. The tool enables to create and use large-scale models based on the qualitative mathematical framework. The possibility to simulate and analyse models in real-time including loss/gain simulation of function; what-if testing is the next advantage. The Cell Collective is free for academic use tool accessible at cellcollective.org. The user registers first and then just logs-in from any computer because created networks will always be stored on a selected profile. In addition, platform allows to share networks with other users and therefore collaborate on projects from different locations. The user or collaborating group can, in addition to creating their own model, use already proposed models from a growing database and customize it.

Cytoscape [27] is a software for visualizing interaction networks and pathways and integrating them with gene expression and other data. The tool was originally made only for private use in biological research, but now it is an open source platform for complex networks available from cytoscape.org. Plugins are usually free to download at Cytoscape App Store where they are divided into categories such as data visualisation, graph analysis, pathway database, etc.

Gephi (The Open Graph Viz Platform) [28] is an open source software using a 3D render engine to display large-scale networks in real-time. Software is created for visualization and analysis the dynamics of networks, provides access to network data and allows for spatializing, filtering, navigating, manipulating and clustering. Gephi is available from gephi.org.

GINsim – Gene interaction Network simulation [29] is a tool made for genetic regulatory networks focused on qualitative models based on a discrete, logical formalism available at ginsim.org. The GINsim allows to model the net as asynchronous, multivalued logical functions and simulate/analyse the qualitative dynamical behaviour or explore an already created network from an extensive database.

SQUAD [30] was created at Swiss Instutite of Bioinformatic for the dynamic simulation of signaling networks (see omictools.com/squad-tool). It is based on the standardized qualitative dynamical systems. At first it converts the network into a discrete dynamical system, identify steady states by using of binary decision algorithm and at last the SQUAD creates a continuous dynamical system.

The tool allows to make a simulation on continuous systems and the network perturbation, which makes it possible to get closer to lab experiments (e.g. activating receptors or knocking out specific components).

VisANT [31] (available from visant.bu.edu) is a tool for integrating interaction data with multi-tiered architecture. A software offers an interface for a large range of published datasets and it is integrated with standard databases such as GenBank, KEGG and SwissProt. A Java-based tool is suitable for many applications, e.g. pathways study, gene regulation, systems biology, mining a visualizing data in context of sequence, pathway and structure. Data can be analysed, combined and overlaid using built-in functions. VisANT 4.0 (last version) provides functions to analyse networks of diseases, therapies, genes and drugs [32].

PathVisio [33] (available from www.pathvisio.org) is a pathway analysing and drawing tool that can be combined with other tools to computational augmentation of pathways, visual compilation of biological knowledge and interpretation of high-throughput expression datasets. A tool provides a basic set for pathway drawing, analysis and visualization, additional features (pathway building and analysis, data integration, etc.) are available in plugin repository [34].

Reactome [35] is a curated and peer-reviewed pathway database (available at reactome.org) allowing visualization, interpretation and analysis of biological pathways. A database is divided into four parts: pathway browser for visualization and interaction Reactome pathways; data analyser merges expression analysis and pathway identifier mapping; ReactomeFIViz is created to find pathways and network patterns related to diseases and documentation. Whereas pathway analysis is only one section of Reactome, which is primarily a database, Reactome is also described in the following subchapter Databases.

2.3 Databases

Specialist databases are used to store acquired data and share them with other professionals. Examples of large, freely available databases created and controlled by experts are: Reactome, KEGG, WikiPathways, UniProt, BioModels, RegulonDB.

Reactome [35] is a database of signaling pathways, metabolic molecules and their relations. It is designed as a graphical map of pathways and processes containing detailed information about components and relations by clicking through on the selected part of the network. Reactome pages cross-reference to over 100 different bioinformatics resources, including Ensembl, UniProt, ChEBI, KEGG and PubMed. Data are downloadable in data formats: Neo4j GraphDB, MySQL, BioPAX, SBML and PSI-MITAB.

KEGG (an encyclopedia of genes and genomes) [36] is still growing integrated database resource of sequenced genomes available at www.kegg.jp. The primary objective of the project is to assign functional meanings of genes and genomes at the molecular and higher levels. In this time, KEGG consists seventeen databases divided into four categories: systems, genomic, chemical and health. The KEGG Orthology (KO) database in the genomic category contains knowledge of molecular-level functions and it is organized with the concept of functional orthologs; each KO is defined as a functional ortholog of genes and proteins. Genome and Genes databases are also part of the genomic category and are derived from RefSeq, GenBank and NCBI Taxonomy databases.

WikiPathways [37] was established by biology community to contribute to pathway information (see wikipathways.org). It was developed as an open and collaborative platform. The WikiPathways is focused on genes, proteins and metabolic pathways. Data are encoded in GPML format, created with PathVisio tool and it is linked to other databases (Reactome, KEGG and Pathway Commons).

UniProt (the Universal Protein Resource) [38] is a large dataset of protein sequences and associated annotation (available at uniprot.org). The knowledgebase contains more than 60 million sequences. A database is divided into four main parts: UniProtKB – a database of proteins; UniRef – clustered sets of sequences; UniParc – non-redundant database of publicly available protein sequences; Proteomes – protein sets from fully sequenced genomes. In addition to these sections, there are another three (auxiliary) sections: annotation systems, supporting data and help. The section UniProtKB also offers tools for analysis and clarifying of biological data, including: BLAST – the basic local alignment search tool searches local similarities between sequences; Align – aligning protein sequences with the Clustal Omega program; Retrieve/ID mapping tool; Peptide search [39].

BioModels [40] is a dataset of mathematical models representing biological and biomedical systems and processes. Literature-based mechanistic models are stored in standard formats and in a high-quality. Models can be mooted in every modeling formats or approaches they are encoded, such as ODE, logical, agent-based, etc. It is also possible to load models of any formats such as SBML, CellML, MATEMATICA, etc.

RegulonDB [41] is a database of *Escherichia coli* K-12 gene regulation available at regulondb.ccg.unam.mx. *E. coli* is the best-characterized organism (including pathways, interactions, regulation, etc.) and so it is often used for studies in systems biology, whether the bacteria or other organisms with similar features. A database currently collects 232 interactions with RNAs affecting 192 genes, 189 GENSOR units (elementary genetic sensory-response units) and 304 transcription factors.

2.4 Data formats

Standardized data formats are indispensable for proposing, storaging and sharing of all signaling pathways *in silico* because they convert readable data into bits and thus allow them to be written to the computer. The most used formats in SB are XML, SBML, GML KGML and SIF.

XML (the eXtensible Markup Language) is a simple, flexible text format used for web-based applications in many domains. It is the basis for many other data formats such as SBML.

SBML (the Systems Biology Markup Language) is a standard XML-based format that allows the storage of arbitrarily complex stuctures of biological models. SBML code is divided into tags: body (information about the model), function and unit definitions, compartments and species types, parameters, rules, reactions, etc.

GML (the Graph Modeling Language) [42] is a text format supporting network data, used for example in Gephi, Cytoscape, etc. A GML file consists of a hierarchical key-value list.

KGML (KEGG Markup Language) [43] is a format of KEGG graphic objects, especially manually drawn and updated pathway maps. Pathways are represented as nodes and edges where nodes specify graph objects and edges represent relation (in protein networks) and reaction (for chemical networks) elements. It also allows facilities to computational analysis and protein/chemical networks modeling.

SIF (Simple Interaction File) [44] is a simple format that only specifies nodes and their interactions. SIF is especially used to import interactions when creating the network for the first time, because it is an easy to create this format in a text editor or spreadsheet. When is a network done, it is usually saved in another format (GML, XML, …).

2.5 Data analysis

Signaling pathway analysis is used to determine model properties. There are many parameters that can be monitored depending on the goal of the study. The basic division of data analysis is into static and dynamic, depending on whether network properties or model dynamicity are evaluated.

2.5.1 Static analysis

There are many tools for enumeration individual parameters (see chapter 2.2). Below is a description of the most significant elements of the static analysis.

All-pairs shortest path is a parameter specifying the shortest distance between every pair of nodes; as example can be finding the quickest way from one place to another one. Floyd Warshall's Algorithm (and its modifications) is usually used method to compute the shortest path in a weighted directed graph [45]. Result of the analysis is a matrix M containing $n \times n$ values, where n is number of nodes in the network.

Average shortest path (ASP) is an average value of all-pairs shortest path calculated according to equation 2.5:

$$ASP = \frac{2}{N(N+1)} \sum_{i \leq j; i,j \in G} d_{ij} \tag{2.5}$$

where N is number of nodes in the graph *G (V, E), i* and *j* are nodes in the graph *G*, d_{ij} is a distance between nodes *i, j* [46]. ASP can be similarly calculated for each node in the matrix *M*, which is referred to as average shortest path length (ASPL); the result is an average shortest path for every node – *n* values.

Network diameter (ND) is the longest length of all shortest paths and hence, a highest value in the matrix *M* (result of the all-pairs shortest path). The value also represents the linear size of a network: 1D lattice if ND ~ *n*; 2D lattice if ND ~ $n^{1/2}$; 3D lattice if ND ~ $n^{1/3}$ and random network if ND ~ ln(*n*) or if ND is smaller than ln (*n*) [47]. For example, ND of World Wide Web is 93.

Connectivity distribution (degree distribution) of the node is the number of edges attached to the node, or the number of nodes the node is connected to. Enumeration of the parameter is shown in green in the Fig. 9. This parameter is significant for searching hubs, i.e. single nodes with a high degree. Connectivity distribution is also used to evaluate whether a model matches a random graph (e.g. Bernoulli's random network: each node is connected with a certain probability, and connectivity has a binomial distribution) or a real network where most nodes have small degree, but a small number of nodes have high degree [6, 7].

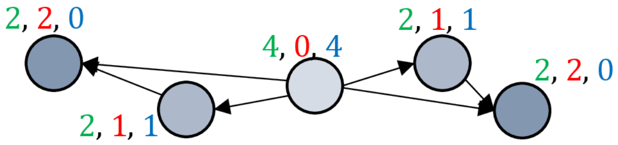


Fig. 9: Connectivity distribution (green), c. in degree (red), c. out degree (blue)

Connectivity in degree is a subset of connectivity distribution indicating number of edges entering the node (see Fig. 9, red).

Connectivity out degree is a subset of connectivity distribution, unlike in previous parameter, specifies number of edges exiting the node (see Fig. 9, blue).

Closeness centrality (CC) [48] is an average distance from one node to other nodes, describes how fast information is able to flow from a given node to other nodes. It is calculated as the normalized inverse of the sum of the topological inverse (see the equation 2.6) or simplified as the inverse the farness.

$$CC(i) = \frac{N - 1}{\sum_j d(i, j)} \tag{2.6}$$

where *i, j* are nodes, *i ≠ j*; *N* is number of nodes; d_{ij} is the shortest path between nodes *i* and *j*; $\sum_j d(i, j)$ is the farness.

Feedback loops (FL) are used in many branches such as software development, social sciences, biology, mechanical and electronic engineering. It is a path starts and ends at the same point; e.g. node *A* sends an information to the node *B*, node *B* sends an answer to the node *A*. FL can be divided into negative and positive feedback. Negative regulation in the example with nodes *A* and *B* causes a decrease in production of the node *A*, positive regulation increase in production of the node *A*.

Clustering coefficient [49] is used to determine the interconnection of the network. It can be divided into global (description of the entire graph) and local (enumeration for each node). Local clustering coefficient quantifies the degree to which nodes tend to associate and being cliques (i.e. complete subgraphs where every two nodes are adjacent). In other words, the coefficient indicates the probability that any two nodes that have a common neighbor are also connected.

Eccentricity [50] is a node centrality index indicating the maximum distance between node *u* and all other nodes. High eccentricity value of the node *u* assumes that all other nodes are in proximity and the node *u* could be easily influenced by other nodes or influence other nodes. Low values indicate that at least one node is remote from the individual node *u* and could mean a marginal role in the network.

Stress [50] is a parameter identifying node's relevance to hold communicating nodes together. It is measured as the number of shortest paths passing through node *u*. Usually high stress score indicates the significance of *u* to maintain the connection of the passing through nodes.

Betweenness centrality (B) [51] is a network property describing shortest path between pair of nodes passing through the node *u*. Applies the equation 2.7:

$$B_u = \frac{1}{(n - 1)(n - 2)} \sum_{s \neq u \neq t \in V} \frac{\mathrm{S}_{st}(u)}{S_{st}} \tag{2.7}$$

where *n* is total number of nodes, *u* is computed node, S_{st} is number of shortest paths from *s* to *t* and $S_{st}(u)$ is number of shortest paths from *s* to *t* passing through *u*. *B* is normalized by the number of node pairs.

2.5.2 Dynamic analysis

The testing evaluates model changes over time which is an important aid in examining the particular system, as it is not necessary to perform laboratory experiments that may be costly or impossible to realize. The method is also an ideal solution for model fitting. Comparing the course of the model (called simulation) with the behaviour of a living organism allows modifying model parameters to achieve the highest match.

Dynamic analysis studies model behaviour under certain conditions, not the model structure or setting of individual elements in the network. For example, it examines the response of the system to structural changes such as knockout or overexpression of a defined gene.

Using model simulation are studied parameters such as concentration, states of individual molecules, stages in processes, activity level, etc.; these variables can take place either in time (for continuous models) or in steps (for discrete models).

Continuous models are made up of quantitative methods, mainly using differential equations (see subchapter 2.1.2); every variable is defined at every moment and time changes are continuous [24]. Models are used mainly to simulate long time phenomenon, typically much longer than individual elements life or course [53], for example the reaction of the environment to the particular substance or the long-term evolution of the genetically modified species.

Discrete models describing step simulations, i.e. with abstracted time, are made up of qualitative methods – described by interaction graphs, Boolean/logical models, etc. (see subchapter 2.1.1); variables are described only at given steps or points, not at any moment. They are usually used to observe behaviour of individual cells, because define object as a set of points or states and the conditions under which they get into these states; conditions can be modified during the simulation due to behaviour of surrounding points or states [53].

3 DATA ACQUISITION FOR SIGNALING PATHWAYS MODELING

To construct a model of signaling pathways, we need to know all nodes involved in the process and the relationships between these nodes.

The first option how to get information of interest is to search the databases. There are many databases of genes (for example GenBank, GeneCards, KEGG Genes), interaction between genes (Biostars, BioGrid, etc.) and already created models.

The next way of data acquisition is the use of experimental methods. The organism or part of it is studied in the laboratory using molecular biology techniques. The most commonly used are techniques for the detection of gene expression to detect genes and determine their activity during the life cycle and the detection of gene products and the phenotype to evaluate the behaviour of the cell.

3.1 Detection of gene expression

Gene expression is the process of performing a product from the information saved in a gene. For example, expression of protein-coding gene causes the formation of a specific protein, which can cause further processes in the organism.

Lab techniques for gene expression study can be divided according to omics they use into: transcriptomics (RNA-Seq, microarray), proteomics (blotting) and metabolomics. All listed techniques belongs to hybridization methods (a target gene detection technique that uses pairing of singe-chain nucleid acids) [54, 55].

3.1.1 RNA-Seq

Whole transcriptome sequencing, RNA-Seq, is the most advanced technique for the detection of gene expression [55].

The sequencing process includes five major steps (see Fig. 10) [54, 55, 56]: **(1)** Total RNA isolation **(2)** Ribo-depletion **(3)** Reverse transcription into cDNA **(4)** Sequencing of the cDNA **(5)** Data analysis.

The last step, data analysis, is divided into four points: **(I)** Removal of sequencing adapters and residual contamination; **(II)** Mapping on the reference genome or *de novo* (synthesis of complex molecules from simple ones; translated: from the beginning) assembly of transcripts if the reference genome is not available; **(III)** Evaluation of the expression by creating the count table; **(IV)** Normalization of the count table.

Normalization of the count table can be done in several ways, depending on the type of data and the following use: negative binomial distribution, reads Per Kilobase per Million mapped reads (RPKM), fragments Per Kilobase per Million mapped reads (FPKM), Transcripts Per Kilobase per Million mapped reads (TPM).

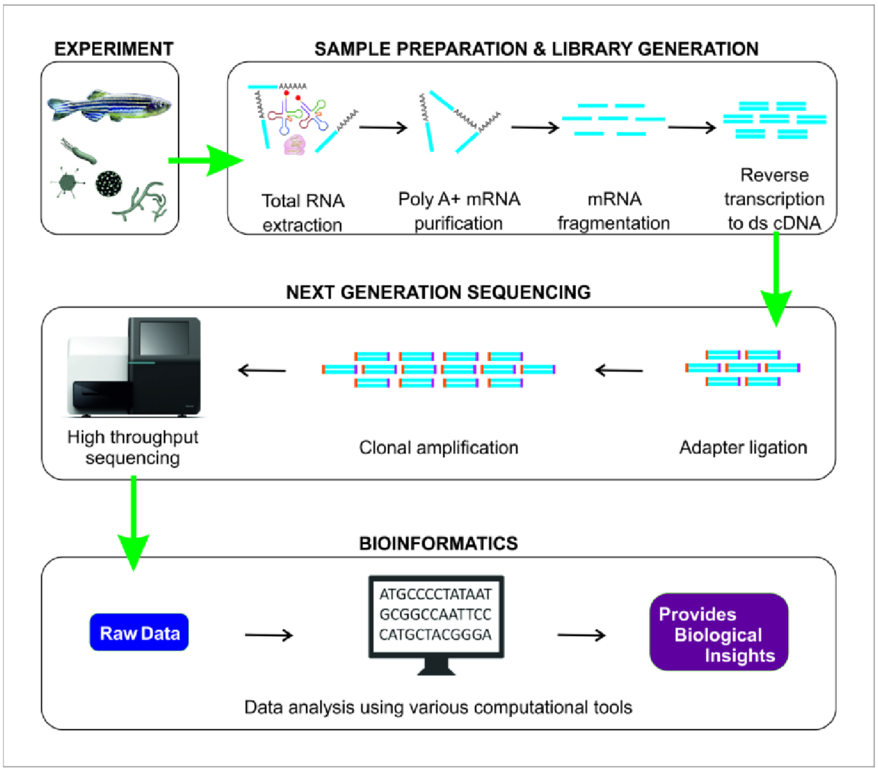

EXPERIMENT
SAMPLE PREPARATION & LIBRARY GENERATION
Total RNA extraction
Poly A+ mRNA purification
mRNA fragmentation
Reverse transcription to ds cDNA
NEXT GENERATION SEQUENCING
High throughput sequencing
Clonal amplification
Adapter ligation
BIOINFORMATICS
Raw Data
ATGCCCCTATAAT
GCGGCCAATTCC
CATGCTACGGGA
Provides Biological Insights
Data analysis using various computational tools

Fig. 10: Illustration of the RNA-Seq detection [56]

In a comparison to microarray (below), RNA-Seq has better results in detecting low abundance of transcripts, differentiating biologically critical isoforms and in the identification of genetic variants [55], which, besides, allows the analysis of non-model organisms whose genome is non-available [54]. Method detects a broader dynamic range and so the detection of more differentially expressed genes with higher fold-change [55]. Moreover, RNA-Seq is an easier method, as there is no need for processed such as cross-hybridization and non-specific hybridization. Disadvantage of RNA-Seq is the need for data processing. Next obstacle was the price, but the cost has now fallen to acceptable value [57].

3.1.2 Microarray

The technique is based on applying a sample to the plate – array. The array consists of thousands of features or spots to which is the sample attached. According to the type of sample that microarray analyses, it is divided into two types: DNA microarray or chip (features are short oligonucleotides) and protein microarray (features are immobilized antibodies for which are analysed proteins antigens).

DNA microarray was invented in 1990s for large-scale studies of gene expression [55]. It used to be the most used technique for the detection of gene expression, but for the reasons outlined above, nowadays it is more often replaced by RNA-Seq. Chips for the DNA detection of commonly analysed species are commercially produced, but for non-model organisms it is usually necessary to create own chip. The preparation of the microarray is possible by applying drops to the substrate or *in situ* (in the original place of the sample) synthesis of oligonucleotides. The next step of the analysis is to hybridize the sample to chip, wash and scan the chip. The result of the analysis is the qualitative information (which genes are expressed) and the quantitative information (the extent of gene expression).

3.1.3 Blotting

Blotting techniques are biochemical methods based on the transfer of the studied fragments from the electrophoresis gel to a nitrocellulose or nylon membrane. Studied fragments create blots, from which the method name is derived. Sir Edwin Southern invented the technique in 1975 for the analysis of DNA fragments. The method is called Southern blot [58]. Later, methods for RNA and protein analysis were developed and are named Northern blot (for RNA) and Western blot (for proteins) [59].

Blotting methods consist usually of four steps [59]: electrophoretic separation of DNA/RNA/protein fragments, transfer to and immobilisation on paper support, binding of analytical probe to the target molecule on paper and visualisation of bound probe.

3.2 Detection of gene products and phenotype

Gene products, RNA or proteins are the result of the individual gene activity – gene expression. Function of all products in the cell, together with the environmental effect, is referred to as phenotype, which indicates the resulting behaviour of the organism as a whole. Laboratory techniques such as HPLC and FC are used to evaluate these parameters and are described below.

3.2.1 HPLC

High-performance liquid chromatography (HPLC) [60] serves to separation, identification and quantification of the active compounds. A method sorts cells based on the molecular composition and structure.

As the first step of the chromatography, analysed sample is inserted into the column, where stationary and mobile phases are. The stationary phase, usually consisting of solid molecules, is fixed in the column. The mobile phase (gas or liquid) together with the analysed sample flow through the column and interacts with the stationary phase based on its size, chemical and physical properties. This slows down some molecules more than others, therefore the elements elute (arrive to the end of the column – to the detector) at a retention time (specific time typical for each group of molecules).

3.2.2 Flow cytometry

Flow cytometry (FC) is a high-throughput analysis method for sorting and discriminating individual cells or separating subpopulations. The multiparametric technique is based on passing the molecules one by one through laser beams, which measure scattered lights and fluorescence emissions. The FC process is shown in the Fig. 11.

Analysing cells are excited by laser light source, which they emit as a light of a certain wavelengths based on their size and shape. Forward scatter (FSC) measures cell size as well as distinguishes the living cells from the dead ones. Side scatter (SSC) detects cell's shape or granularity. Since the light intensity of the SSC is weak, a photomultiplier (PMT) is used to amplify the signal.

FC allows to analyse 32 parameters simultaneously and sort up to 100 thousand cells per second [62]. A technique is used in many fields, such as immunology, molecular biology, hematology, plants and bacterial research.

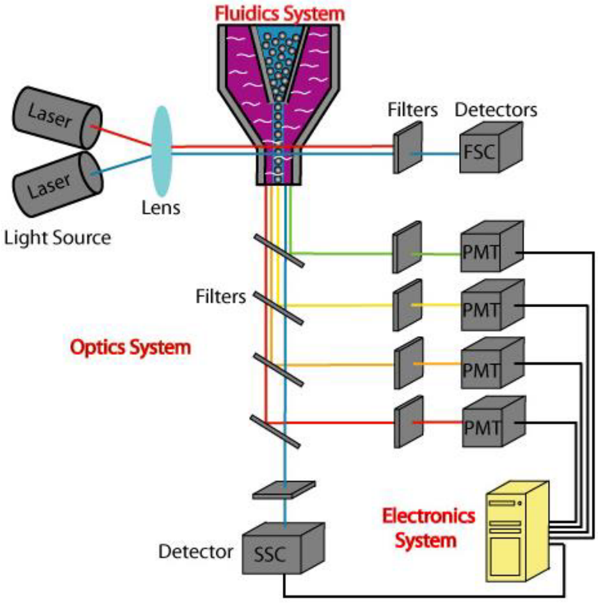


Fig. 11: Flow cytometer scheme [61]

4 SIGNALING PAHWAYS IN CLOSTRIDIA

4.1 Solventogenic clostridia

The clostridium genus is formed by anaerobic, sporadic bacteria occurring primarily in the soil. Morphologically, clostridium is rod-shape G+ (gram-positive) bacterium having a shape of bowling pin or a bottle in their endospore stage. More than 100 species of clostridia are described, many of which have effect on the human (or animal) organism. Positive effects have, for example, *C. leptum, C. coccoides*. Negative effects, as well as serious diseases, cause e.g. *C. botulinum, C. difficile, C. perfringens, C. tetani* [63]. Further, for this thesis the most important, there are species with acetone-butanol-ethanol (ABE) fermentation – solventogenic species. These are: *C. acetobutylicum, C. beijerinckii, C. saccharoperbutylacetonicum* [64]. Clostridia life cycle is shown in the Fig. 12.

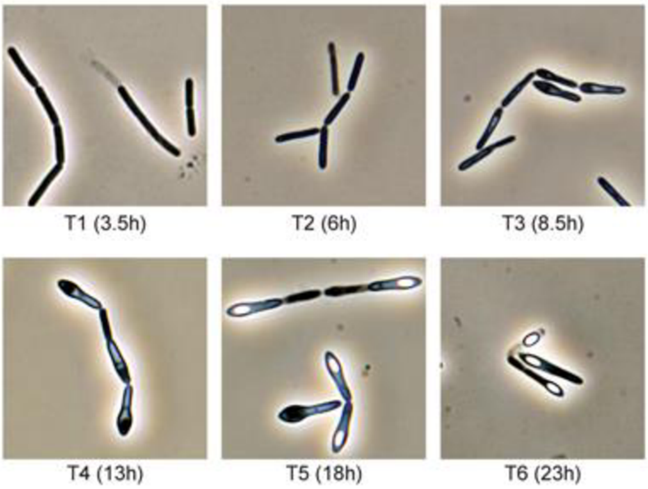


Fig. 12: Life cycle of solventogenic clostridia at six time points [65]

ABE fermentation is a biological process in which acetone, butanol and ethanol are produced. It was found in the early 20th century by Dr. Ch. Weizmann [64]. Butanol is a quarternary alcohol which can be used as a solvent, disinfectant, or as a biofuel. Its use as a biofuel or dopant in fuel has recently been often discussed, mainly in the attempt to use an alternative fuel from renewable sources and the reduction of oil reserves.

Clostridium acetobutylicum is a model organism of solventogenic clostridia containing 94 strains. It was firstly successfully isolated and used for the large-scale solvents production [66]. Genus research for butanol production has lasted over a hundred years [67]. *C. acetobutylicum* is sensitive to rifampicin, able to produce riboflavin and hydrolyse gelatine – unlike most other solventogenic species [66].

Clostridium beijerinckii is the genus utilizing a widest range of solvent production substrates and seems to be the most robust in terms of viability in a wide range of environmental conditions [65]. Therefore, *C. beijerinckii* appears to be the most suitable candidate for the large-scale butanol production in comparison to other solventogenic species [68] and its strain NRRL B-598 will be discussed in detail below.

Clostridium saccharoperbutylacetonicum [66] cells are in the form of straight, short and long rods with rounded ends and peritrichous flagella for movement. They usually occur singly or in pairs. During the life cycle, a species accumulate granule towards the end of growth and under adverse conditions, sporulation occurs. Endospores are oval, approximately the same size as the rod-shape.

4.2 Butanol production – general signaling pathways

Signaling pathways, as mentioned above, are cooperating molecules that control a specific cell function (e.g. butanol production), which can be described by mathematical models. Butanol production usually consist of two phases: acidogenesis and solventogenesis [65]; in solventogenic clostridia it is controlled especially by *Sol* operon. *Sol* operon is formed by a small open reading frame (ORF) and four genes [69]: *ctfA, ctfB, adc* (these genes are equal for all described clostridia) and *ald/bld/aad/adhE* (different for each species) – see the Fig. 11.

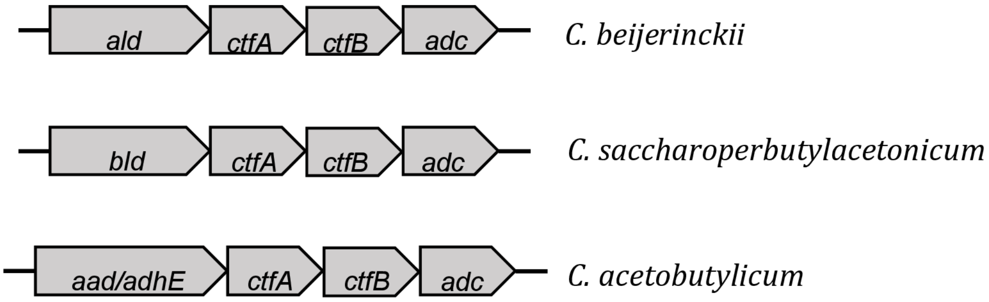
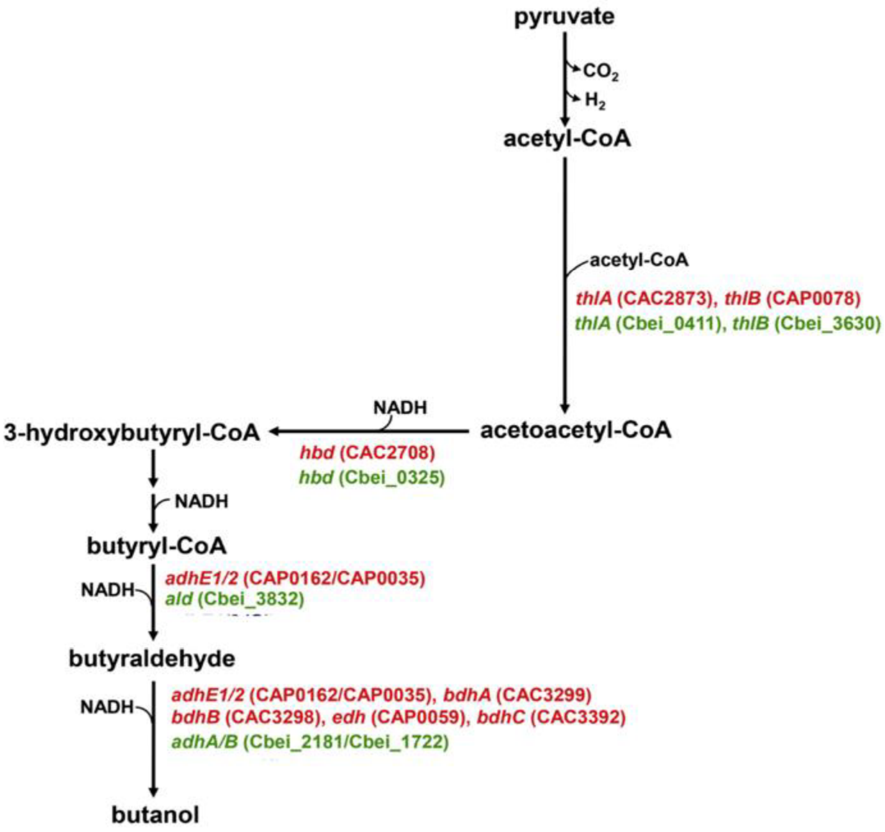


Fig. 13: *Sol* operon genes

Butanol production in bacteria, especially the effort to increase production, is a widely-studied area, so therefore exist many already created pathways. Most important models for this study are: a comparing model of two best described species, five already created pathways of the *C. acetobutylicum* and a genome-scale network of the *C. beijerinckii*. Pathways are described below.

A comparing model of *C. acetobutylicum* and *C. beijerinckii* [70] is shown in the Fig. 12 (its most important part for this study, the whole model is due to the large size available at [70] and in the attachment under the name *'comparison SP of beijerinckii and acetobutylicum'*). The picture shows the metabolic pathway of butanol production with contribution of acetyl-CoA (acetyl coenzyme A) and other metabolites, starting with pyruvate – a basic cell metabolite and final product of glycolysis. Genes that affect pro production of these metabolites are shown in red (*C. acetobutylicum*) and green (*C. beijerinckii*).



pyruvate
CO2
H2
acetyl-CoA
acetyl-CoA
thlA (CAC2873), thlB (CAP0078)
thlA (Cbei_0411), thlB (Cbei_3630)
acetoacetyl-CoA
NADH
hbd (CAC2708)
hbd (Cbei_0325)
3-hydroxybutyryl-CoA
NADH
butyryl-CoA
NADH
adhE1/2 (CAP0162/CAP0035)
ald (Cbei_3832)
butyraldehyde
NADH
adhE1/2 (CAP0162/CAP0035), bdhA (CAC3299)
bdhB (CAC3298), edh (CAP0059), bdhC (CAC3392)
adhA/B (Cbei_2181/Cbei_1722)
butanol

Fig. 14: SP of *C. acetobutylicum* (red) and *C. beijerinckii* (green) [70]

Already created cellular overview of *C. acetobutylicum* ATCC 824 pathways is stored in BioCyc Database Collection [71], shown in the Fig. 13 and available at https://biocyc.org/overviewsWeb/celOv.shtml?orgid=CACE272562&pnids=PWY-6594, where is it possible to work interactively with pathways – display only selected pathways, genes, metabolites, etc.

A pathway of butanol production called Superpathway of Clostridium acetobutylicum solventogenic fermentation, part of a cellular overview (coloured in green in the Fig. 13), is shown in the Fig. 14 and included in the attachment as *'C. acetobutylicum ATCC 824 - Superpathway of Clostridium acetobutylicum solventogenic fermentation'*.

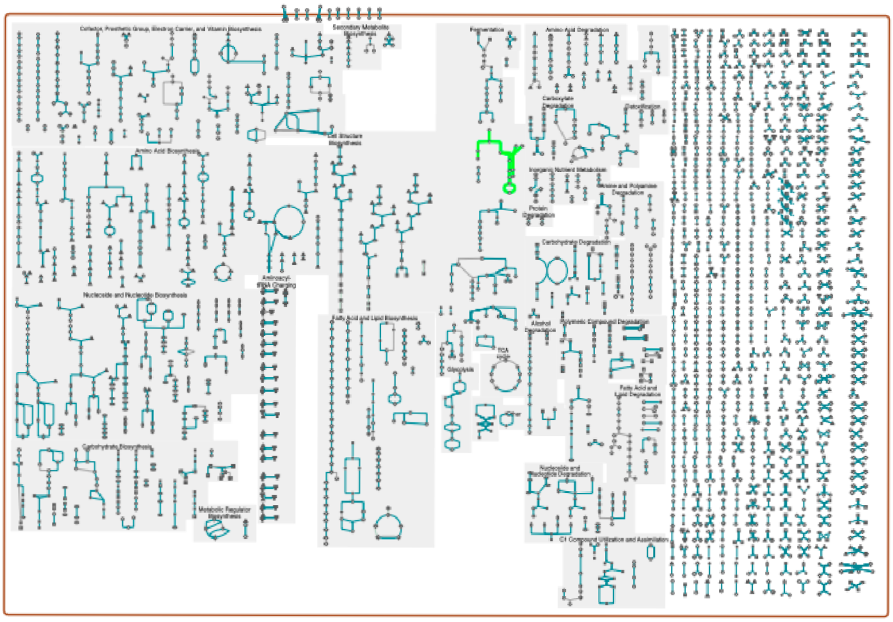

Cofactor, Prosthetic Group, Electron Carrier, and Vitamin Biosynthesis
Secondary Metabolite Biosynthesis
Fermentation
Amino Acid Degradation
Carboxylate Degradation
Detoxification
Amino Acid Biosynthesis
Cell Structure Biosynthesis
Inorganic Nutrient Metabolism
Amine and Polyamine Degradation
Protein Degradation
Carbohydrate Degradation
Nucleoside and Nucleotide Biosynthesis
Aminoacyl-tRNA Charging
Fatty Acid and Lipid Biosynthesis
Alcohol Degradation
Polymeric Compound Degradation
Glycolysis
TCA cycle
Fatty Acid and Lipid Degradation
Carbohydrate Biosynthesis
Metabolic Regulator Biosynthesis
Nucleoside and Nucleotide Degradation
C1 Compound Utilization and Assimilation

Fig. 15: *C. acetobutylicum* ATCC 824 pathways - an overview [71]

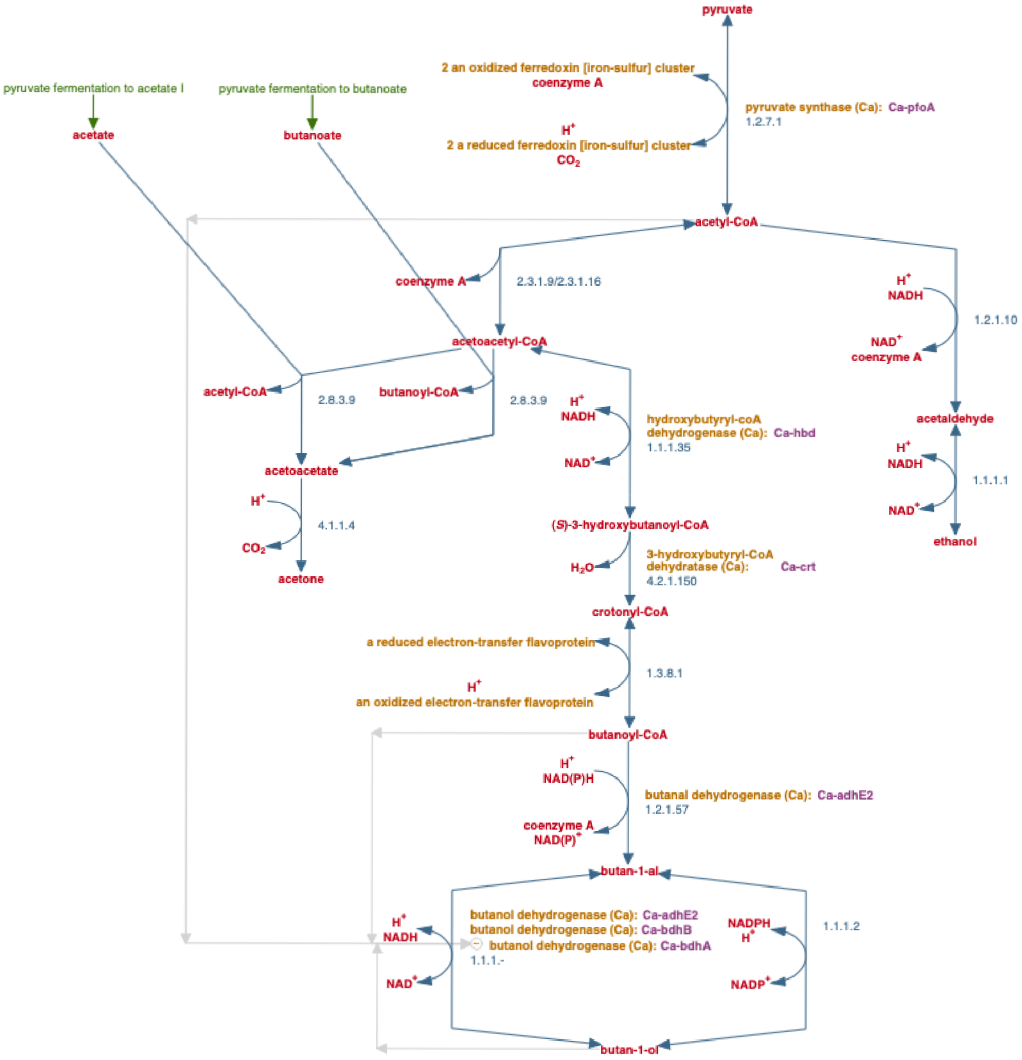

pyruvate
2 an oxidized ferredoxin [iron-sulfur] cluster
coenzyme A
pyruvate synthase (Ca): Ca-pfoA
1.2.7.1
H+
2 a reduced ferredoxin [iron-sulfur] cluster
CO2
pyruvate fermentation to acetate I
acetate
pyruvate fermentation to butanoate
butanoate
acetyl-CoA
coenzyme A
2.3.1.9/2.3.1.16
H+
NADH
1.2.1.10
NAD+
coenzyme A
acetoacetyl-CoA
acetyl-CoA
2.8.3.9
butanoyl-CoA
2.8.3.9
H+
NADH
hydroxybutyryl-coA dehydrogenase (Ca): Ca-hbd
1.1.1.35
NAD+
acetaldehyde
H+
NADH
1.1.1.1
NAD+
ethanol
acetoacetate
H+
4.1.1.4
CO2
acetone
(S)-3-hydroxybutanoyl-CoA
3-hydroxybutyryl-CoA dehydratase (Ca): Ca-crt
4.2.1.150
H2O
crotonyl-CoA
a reduced electron-transfer flavoprotein
1.3.8.1
H+
an oxidized electron-transfer flavoprotein
butanoyl-CoA
H+
NAD(P)H
butanal dehydrogenase (Ca): Ca-adhE2
1.2.1.57
coenzyme A
NAD(P)+
butan-1-al
butanol dehydrogenase (Ca): Ca-adhE2
butanol dehydrogenase (Ca): Ca-bdhB
butanol dehydrogenase (Ca): Ca-bdhA
1.1.1.-
H+
NADH
NAD+
NADPH
H+
NADP+
1.1.1.2
butan-1-ol

Fig. 16: *C. acetobutylicum* ATCC 824: solventogenic superpathway [71]

Butanoate metabolism pathway of *C. acetobutylicum* ATCC 824, shown in the Fig. 15 and included in the attachment as *'C. acetobutylicum ATCC 824 - Butanoate metabolism'*, is available also in the KEGG database [72]. Green elements are described in detail, white elements do not contain more information yet. Butanol and genes *ctfA* (2.8.3.8) and *adhE* (1.2.1.10) are highlighted in red in the picture; genes *ctfB* and *adc* are not described in detail now.

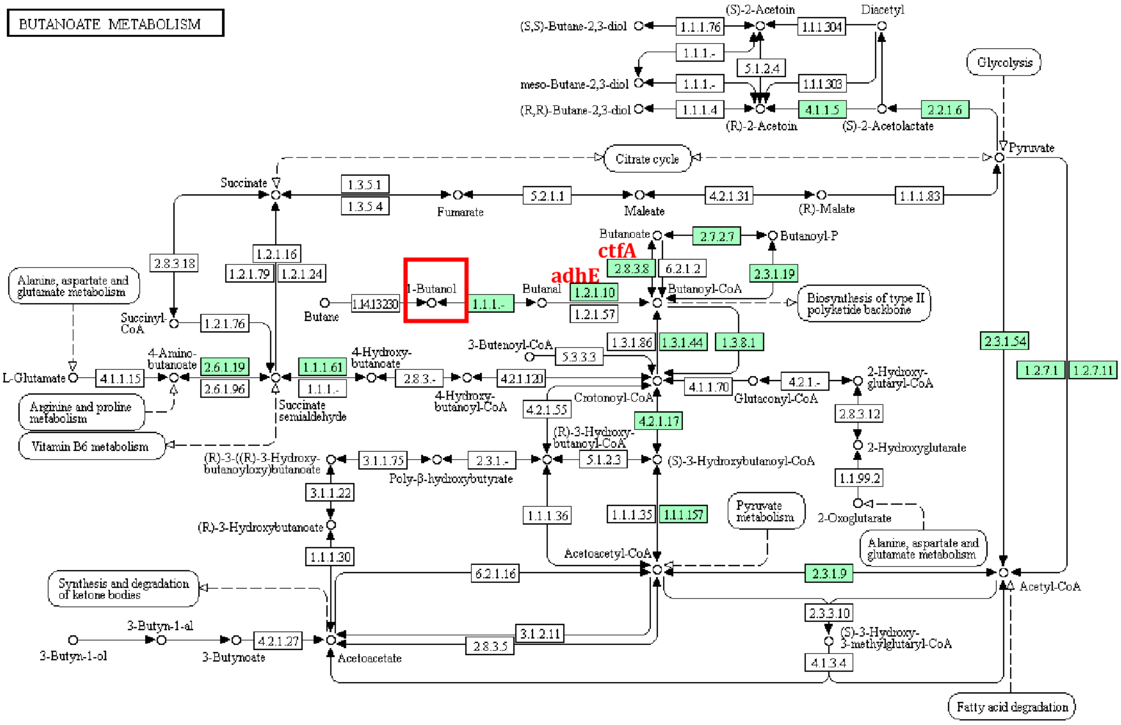


Fig. 17: Butanoate metabolism pathway of *C. acetobutylicum* ATCC 824 [72]

Two-component system of *C. acetobutylicum* ATCC 824 is a pathway available from KEGG database [73] and it is included in the attachment as *'C. acetobutylicum ATCC 824 – two-component system'*. Part of the pathway containing the *Sol* operon is in the database's section Other families and is showed in the Fig. 16.

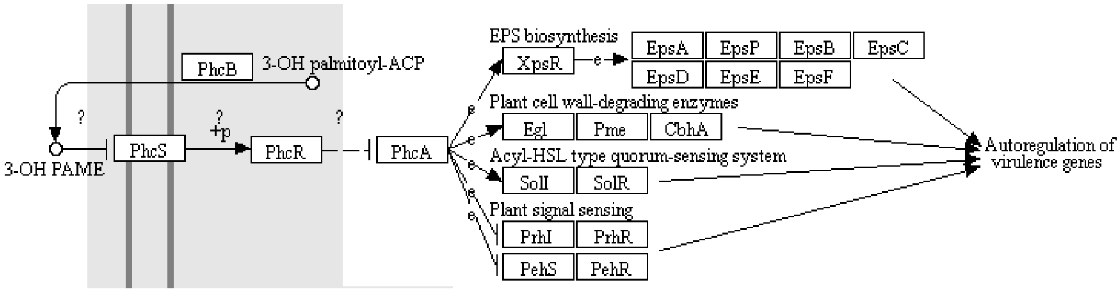


Fig. 18: Two-component system of *C. acetobutylicum* ATCC 824

section Other families, part containing *Sol* operon [73]

BioModels contains two genome-scale models of *C. acetobutylicum* ATCC 824 (iCac802 and iJL432) and a genome-scale network of *C. beijerinckii* NCIMB 8052 (iCB925). All of them are divided into three networks: all, base and kinetic.

iCac802 network was first published at 2008 by Senger and Papoutsakis [74], contains 4 148 nodes and 14 438 edges, 422 intracellular metabolites involved in 522 reaction and 80 membrane transport reactions. Semi-automated reverse engineering algorithm, thermodynamic analysis and systematic gene knock-out simulations were used to propose the net. iCac802 is available in the attachment as *'Genome-scale metabolic network of Clostridium acetobutylicum – Senger'.*

iLJ432 was reconstructed from annotated genomic sequence at 2008 by scientists Lee, Yun, Feist, Palsson and Lee [75]. It contains 1 605 nodes, 4 333 edges, 502 reactions and 479 metabolites. The network was used as the basis for an *in silico* model used to predict metabolic fluxes during the acidogenic phase. Single gene deletions were used to predict essential genes. Whole network is available in the attachment (*'Genome-scale metabolic network of Clostridium acetobutylicum – Lee').*

iCB925, the first metabolic genome-scale model of *C. beijerinckii* presented at 2011 by Milne et. al. [76] is shown in the Fig. 17 and available in the attachment as *'Genome-scale metabolic network of Clostridium beijerinckii'.* The network was built by semi-automated procedure using databases KEGG, BioCyc and The SEED. Containing 3 133 nodes, 9 087 edges, 925 genes, 938 reactions, 881 metabolites and 67 membrane transport reactions is iCB925 the largest genome-scale model for the clostridium species [76].

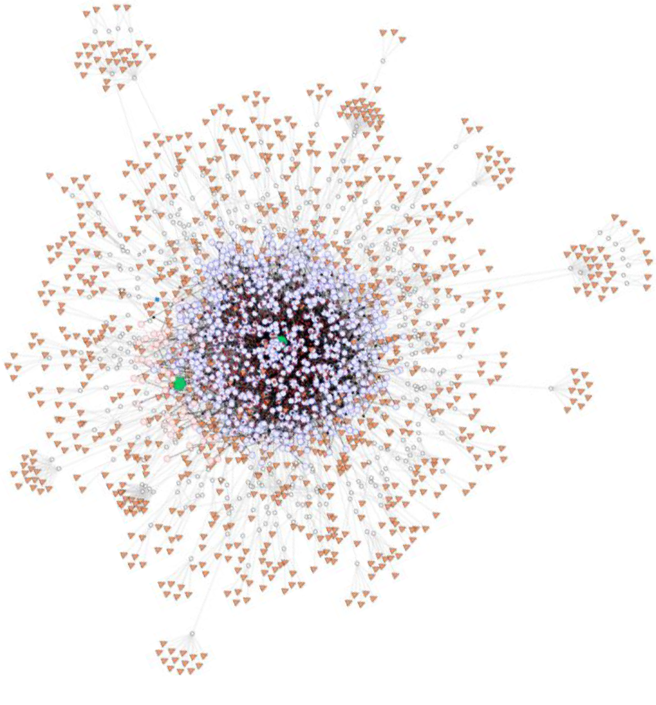
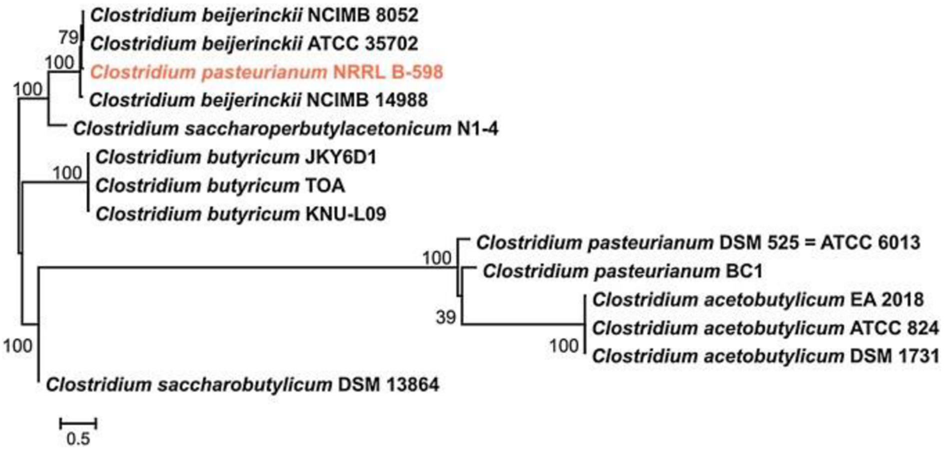


Fig. 19: Genome-scale network of *C. beijerinckii* NCIMB 8052 [76]

5 BUTANOL PRODUCTION IN THE STRAIN *C. BEIJERINCKII* NRRL B-598

C. beijerinckii NRRL B-598 was reidentified from the *C. pasteurianum* genus in 2017 based on the genome statistics similarity [77] (see phylogenetic tree in the Fig. 20). The complete genome was first sequenced in 2014 [78], then it was refined and version 3 is now available in NCBI Reference Sequence under the name NZ_CP011966.3 (https://www.ncbi.nlm.nih.gov/nuccore/CP011966.3/).



Clostridium beijerinckii NCIMB 8052
79
Clostridium beijerinckii ATCC 35702
100
Clostridium pasteurianum NRRL B-598
100
Clostridium beijerinckii NCIMB 14988
Clostridium saccharoperbutylacetonicum N1-4
Clostridium butyricum JKY6D1
100
Clostridium butyricum TOA
Clostridium butyricum KNU-L09
Clostridium pasteurianum DSM 525 = ATCC 6013
100
Clostridium pasteurianum BC1
Clostridium acetobutylicum EA 2018
39
Clostridium acetobutylicum ATCC 824
100
Clostridium acetobutylicum DSM 1731
100
Clostridium saccharobutylicum DSM 13864
0.5

Fig. 20: Phylogenetic position of *C. pasteurianum/beijerinckii* NRRL B-598 [77]

The non-type strain is able to produce butanol before the start of the sporulation up to maximum of 7.6 g/l, maximum production of acetone is about 3.9 g/l [57]. Ethanol production is negligible, we assume that due to the lack of the gene *aad* that ethanol-producing clostridia have.

5.1 Proposed signaling pathway

Data gain is based on the information obtained from laboratory using RNA-Seq, HPLC and FC as well as database and text mining. KEGG's tool BlastKOALA was used to compare gene identity and function. The pathway was proposed used the Cell Collective as it is an ideal tool for our purpose as it allows to propose a dynamic model and conditions between nodes, perform simulations and analysis, share and publish the model.

The model is shown in the Fig. 21 and available on the Cell Collective website under the name *'Signaling Pathway for Butanol Production in Clostridium beijerinckii NRRL B-598'*, version 1.1 (see bit.ly/2UWiQbJ).

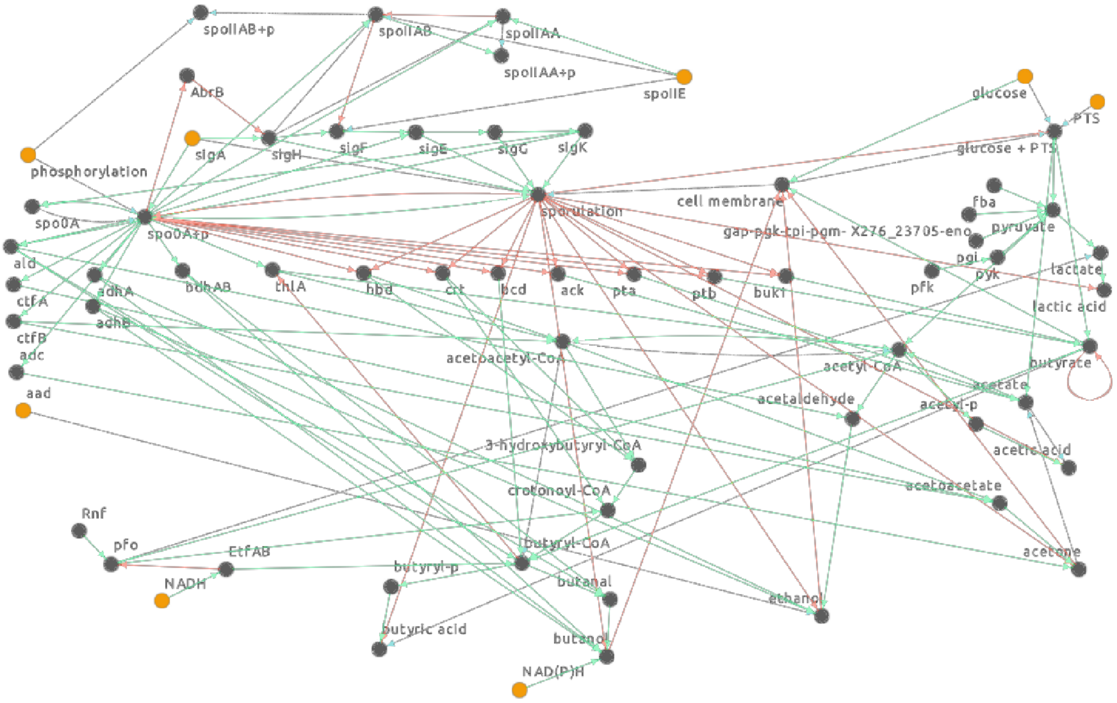


Fig. 21: Signaling pathway for butanol production in *C. beijerinckii* NRRL B-598

Signaling pathway is proposed as a dynamic model, thus allowing the approximation of cell's behaviour during butanol production. The pathway contains the dynamic part such as states and conditions and the static part.

Static part of the dynamic model consists of 66 nodes (elements – genes, proteins and metabolites) and 139 edges (connection between nodes).

Nodes are divided into internal (grey, e.g. butanal) and external (orange, e.g. *aad*). Internal ones contain both entering and exiting edges and they approximate "internal elements" of the cell (genes and proteins, eventually final metabolites) whose amount cannot be influenced directly, but only by their mutation or using indirect interactions with another substance. External ones contain only exiting edges and opposite to internal nodes can be changed in the activity level during the simulation. External components approximate usually elements whose amount is possible to change directly in a living organism, for example by injecting the element as a solution. External nodes are also made up of several elements that cannot be influenced directly, but the possibility of changing their activity level is necessary for the correctness of the model, such as the node *aad* – gene, which the strain does not contain, but is assumed to be crucial for ethanol production, since all other ethanol producing clostridia strains contain this gene.

Edges are divided into: activation (green, e.g. *sigH* → *sigF*) and inactivation (red, e.g. spo0A+p → *hbd*). Conditions are also illustrated as edges (grey; e.g. ethanol is activated only when *aad* is active). Activating edges increase the activity of the node that they target. If more of these edges target to the one node, it occurs to the increases proportionally to all inputs. Inactivating nodes, on the other hand, reduce the activity level of the targeting node. In case of both types of edges (activation and inactivation) targeting to one node, by default each entering edge has the same weight – activation and inactivation numbers are summed but it is possible to set the dominance of each inactivation edge; so that if the node from which the edge is active, the edge will inactivate the target node.

5.2 Simulation

Cell Collective tool allows model simulation in the Simulation panel. The first step is setting the properties such as simulation speed, sliding window and visibility of selected components for observing the course over time; initial state of internal components (active or inactive) and activity level of external components. The values used for the simulation shown in the Fig. 22 are described in the Tab. 1 and stored in the Cell Collective Simulation panel in the window External components: Environment under the name *'MyEnv'*. Due to the fact the dynamic model is discrete, i.e. time is abstracted, the simulation results are in steps (1 step = 1 model's change) and approximate the time in hours (1 step = 1 hour).

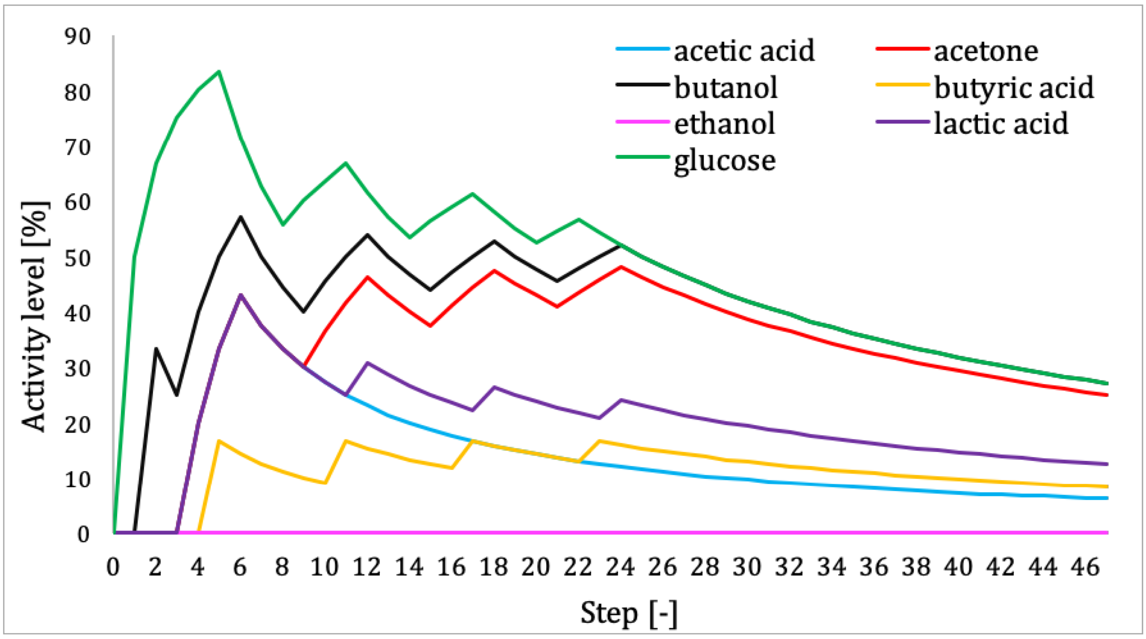


Fig. 22: Model simulation of seven metabolites

Tab. 1: Activity level of external components in individual steps

Activity [%] **Step [-]**	*aad*	glucose	NAD(P)H	NADH	phosphorylation	*PTS*	sigA	spoIIE
0 - 22	0	100	50	50	100	100	100	75
23 - 47	0	0	50	50	100	100	100	75

Since the aim of the work is to propose a dynamic model corresponding to biological data, I focused the simulation on the study of seven metabolites, the results of which obtained using laboratory measurement are available [57]. These metabolites can be considered as target products of the strain, so if the simulation process corresponds to biological data, we can assume the correctness of the model.

External components are nodes whose activity level can be changed. They approximate the biological elements the amount of which can be easily influenced, e.g. by adding them to the solution. External components activity levels (see Tab. 1) have been set based on the biological data matching (will be described in detail in the subchapter 5.4).

The activity level of the *aad* gene was set to 0 % throughout the simulation since the strain NRRL B-598 does not contain this gene, as mentioned above. The glucose value was reduced from 100 % to 0 % in the middle of the simulation, thereby approximating glucose consumption as an energy source during metabolic processes. NAD(P)H and NADH are secondary sources of energy, and their activity was half throughout the simulation, corresponding to their partial involvement in metabolic processes. Phosphorylation is an essential part of many processes in the cell as well as the butanol production process, as its known from laboratory experiments. Therefore, phosphorylation's activity was set to 100 % throughout the simulation. *PTS* is a set of genes responsible for glucose utilization to usable components, without which glucose would not be usable at all. For this reason, *PTS* activity level was set to maximum during the whole simulation. SigA is a sigma factor belonging to the sigma 70 family, which contributes significantly to both the activation of the transcription factor spo0A and the cascade of events leading to the sporulation, which are fundamental processes of the entire network. Therefore, SigA is considered as an essential part of the process and its value was set to 100 %. The SpoIIE sporulation factor is responsible for the activation or inactivation of other sporulation factors such as SigF and SpoIIAB. Since it causes many events depending on the environments, it is not possible to experimentally determine its activity at each case. The SpoIIE activity level was therefore set based on the comparison of simulation results with biological data to 75 %.

The simulation course can be seen in the Cell Collective's Simulation panel. The activity level of each node is shown in colour in the network. The red colour indicates inactivity, green represents activity and the transition over states shows the colour transition from green over yellow and orange to red. The exact activity level value in every step is written in the Internal components table and after the simulation is possible to download values for each node in each step in XLSX format.

5.3 Static analysis

Static analysis was performed using Cell Collective's Network analysis panel, Cytoscape and its plugin CentiScaPe. To verify the properties of the model, I calculated thirteen parameters: all-pairs shortest path, average shortest path, average shortest path length, network diameter, connectivity distribution, connectivity in degree, connectivity out degree, closeness centrality, feedback loops, clustering coefficient, eccentricity, stress and betweenness centrality.

All-pairs shortest path is a matrix M of $n \times n$ values (n = 66) containing the distances between all pairs of nodes that are in the pathway. Using the matrix, I have determined network parameters such as ASP, ASPL and ND. Matrix M is available in the attachment under the name *'allPairsShortestPath'.*

ASP – average value of the M is 4.5, which means that from one selected node to another information spreads over 4.5 nodes on average. The lower value, the faster information flow across the network and thus response to surrounding changes. The resulting ASP value is 14.6 % of all nodes, so responses to changes in environment are fast, although some delay occurs. This fact can be observed even during the simulation (see Fig. 22) – the response to glucose reduction occurred with a slight delay. The ASP suggests that cell's part responsible for controlling the butanol and other solvents production are well connected and capable of rapid communication.

ASPL specifies how much the node is connected or how fast the information will flow if we affect that node. Based on the ASPL results, I identified nodes that are best to modify to get a quick response. I searched for nodes with the lowest ASPL, except for zero values. The zero length indicates that the node has no output edge and cannot directly spread information. There are 4 zero nodes in the analysed network: acids acetic, butyric and lactic and proteins spoIIA+p and spoIIAB+p. These acids are the final products of the network, so they do not have an output edge. spoIIA+p and spoIIAB+p are auxiliary nodes in a transcriptional regulators subset that serve to create certain conditions and completeness of the subnet, but do not directly affect the butanol production. The lowest non-zero values are achieved by acetyl-p, butyryl-p and lactate followed by sporulation with the second highest value.

These nodes I classify as the most ideal for the direct influence in order to spread information as quickly as possible. The highest values are achieved by elements *ack, pta* and *pfk* which means that it is not appropriate to modify these elements directly if is required an immediate network response. ASP values for each node are stored in the attachment under the name *‚static analysis (Cytoscape)'*.

Network diameter – maximum value in the matrix *M* and the longest distance between two nodes of all shortest paths is 11 between *sigK* and *pfk*. This means that if one of these nodes (*sigK/pfk)* starts the flow of information, the last change occurs at the second node (*pfk/sigK)*. So, for studying whether the information has passed through the entire network, we will use this pair of nodes. If the information is reflected on the observed node, we know that all other nodes have already recorded the information. ND parameter also represents the linear size of a network: ND ~ $n^{1/2}$, so the network approximates a 2D lattice [47].

Connectivity distribution or degree distribution is shown in the Fig. 23. Most nodes (18) have connectivity distribution 3. The highest degree (25) reaches node spo0A+p, the second largest degree (22) has a sporulation. Whereas most nodes have low degree number and a small number of nodes have much more higher degree, we can conclude that the model approximates the real network [6, 7]. spo0A+p and sporulation are hubs (nodes with high degree) and thus they are very important for the network in means of topology - a network collision occurs when the hub is removed. Connectivity distribution together with connectivity in degree and connectivity out degree values are stored in the attachment under the name *'connectivityDistribution'.*

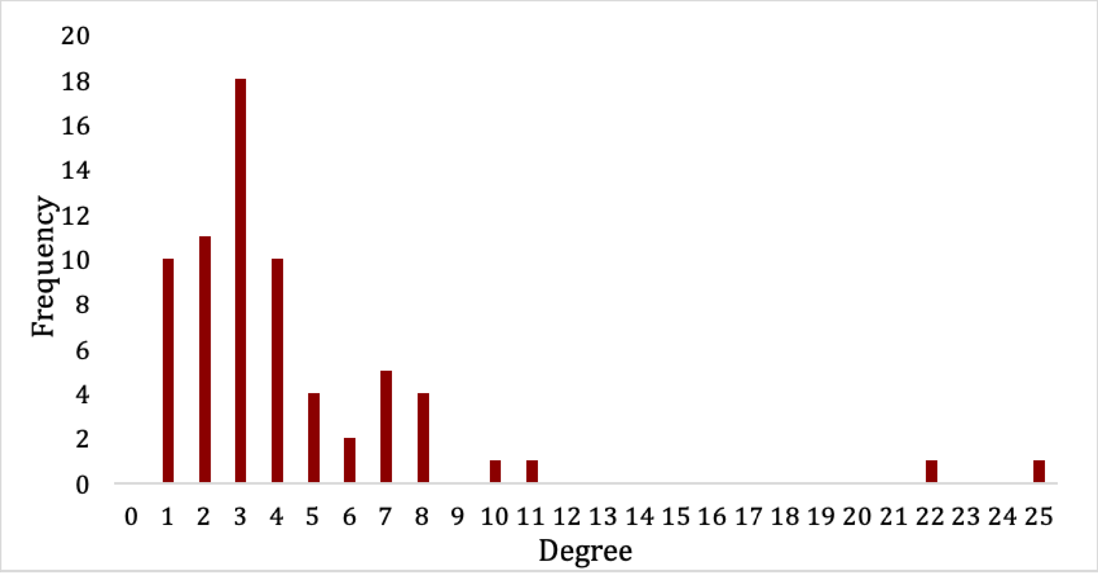


Fig. 23: Connectivity distribution

Closeness centrality values are displayed (sorted by size) in the Fig. 24 and in the attachment (*'closenessCentrality'*). The highest CC reach nodes: *AbrB*, acids acetic, lactic and butyric, spoIIAA+p, spoIIAB+p, butyril, acetyl and lactate. This implies that listed nodes are closest to all other nodes, or most associated with all nodes. On the contrary, the most remote from all other elements are the nodes with the lowest CC, which are NADH, *pfk*, 3-hydroxybutyryl-CoA, *crt*, *ack* and *pta*. It means that if we pass the information to one of the high CC nodes, the information will be spread to many other nodes or even to the entire network. On the other hand, if we insert the information to low CC nodes, they only pass the information to the small neighborhood. Therefore, if we want to influence as much network as possible in the signaling pathway for butanol production, we must primarily target to nodes *AbrB*, acetic acids, lactic and butyric, spoIIAA+p, spoIIAB+p, butyril, acetyl and/or lactate.

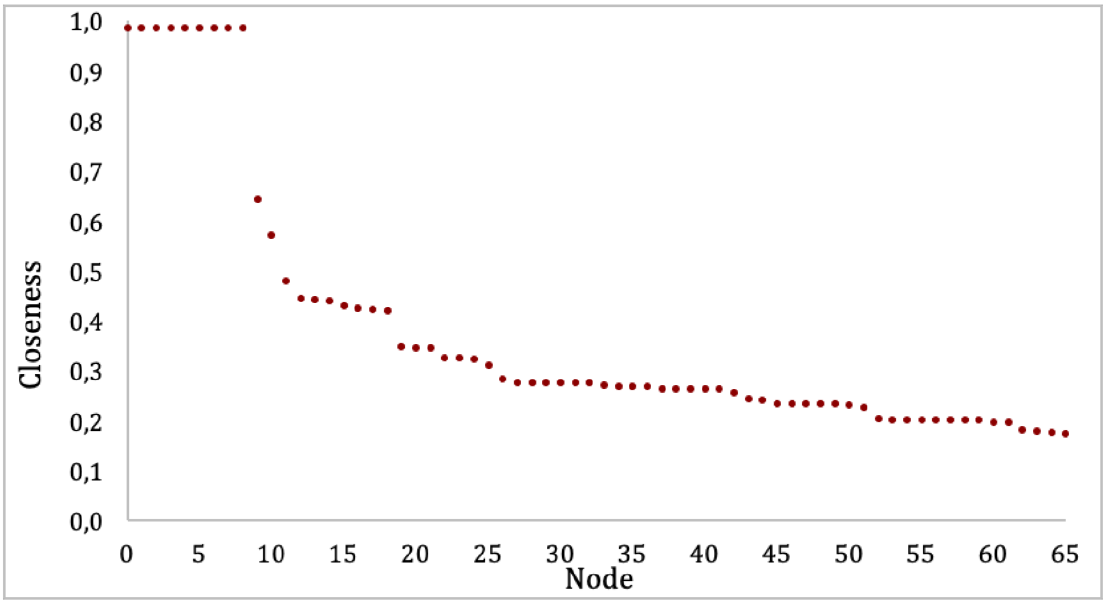


Fig. 24: Closeness centrality values of all nodes

Feedback loops are given as a path starting and ending at the same point. FL are very important as they serve to regulate processes in the cell and, in particular, to maintain homeostasis. The more FL network contains, the better response to environmental changes and adaptation. In this dynamic model is 825 FL, the shortest one contains one node (butyrate inactivate itself), longest loops consist of 14 nodes (e.g. spo0A+p → *ctfA* → acetoacetate → acetone → acetate → acetyl-CoA → acetoacetyl-CoA → 3-hydroxybutyryl-CoA → crotonoyl-CoA → butyryl-CoA → butanal → butanol → cell membrane → sporulation → spo0A+p). The network contains a large amount of FL, so it is capable of high-quality processes regulation and homeostasis maintenance. Based on it, we can assess that the network (and so living bacterium) can easily cope with changes in the environment without damage risk. All FL are stored in the attachment under the name *'feedbackLoops'.*

Next results of static analysis (clustering coefficient, eccentricity, stress and betweenness centrality) are stored in the attachment under the name *‚static analysis (Cytoscape)'* and will be described in detail in the following paragraphs.

Clustering coefficient describes the network interconnection and quantifies nodes tend to associate. The highest coefficient reach glucose, spoIIA+p and *sigA* with the value of 0.5. It means that average connection of listed node's neighbors is 50 % and these nodes are very highly interconnected and thus can easily influence the broad surroundings. A total of 22 nodes have a coefficient value higher than 0.09, so they are highly interconnected [79]. This implies that the network is medium interconnected - less than half of the nodes have a high clustering coefficient value, the remaining nodes medium or low.

Eccentricity interprets nodes influence to other elements or conversely other elements influence to the node. In contrast with clustering coefficient, which is calculated by number of edges, eccentricity is evaluated based on the distances between nodes. Highest eccentricity (11) reach genes *Rnf* and *pfk* followed by *ack, crt, fba*, ferredotoxin, 3-hydroxybutyril-CoA, *pta*, *pgi*, *pyk,* NADH and *pta,* with the value 10. These nodes can be easily functionally reached by other elements or they can reach other elements.

Stress is a parameter describing the relevance of nodes in meaning of holding communicating nodes together. The most relevant nodes (for communication flow) are spo0A+p, sporulation and cell membrane with stress values higher than one thousand. Removal or non-functionality of listed nodes would have a significant negative impact on the spread of the information across the network, or even information could not pass through the network to the target element. spo0A+p, sporulation and cell membrane are crucial for the cell's life processes, so it is appropriate that they are also of great importance in the transmission of information.

Betweenness centrality is generally in range 0-1; high result indicates the importance of the single node. Betweenness centrality values in this analysis are not higher than 0.5, which indicates that individual nodes are equal in the importance on the flow of information through them and removing of any node will have the same impact on the whole network. The highest values reach nodes spo0A+p, sporulation and cell membrane, same results as stress. Centralities stress and betweenness are similar, except that stress is enumerated as the absolute value of the shortest paths, and betweenness measures the fraction of the shortest paths passing through the node. This is confirmed by the previous statement that nodes spo0A+p, sporulation and cell membrane are very significant in the network.

5.4 Evaluating a model match with biological data

To evaluate the model approximation with real bacterium organism, I used biological data measured in the laboratory by HPLC, RNA-Seq and FC in previous studies of *C. beijerinckii* NRRL B-598 [57, 65] and compared them with model simulation and analysis results.

5.4.1 HLPC

Since the obtained laboratory data by HPLC was measured as the concentration of metabolites over time and model gives the result as an activity level of nodes over steps, I converted the measured data to a reaction rate (activity – production or consumption of each metabolite) over time using Matlab tool and the equation 5.1:

$$v_X = \frac{dc_X}{dt}\ [gl^{-1}h^{-1}] \tag{5.1}$$

where v_X is the reaction rate of metabolite *X*, c_X is the concetration of metabolite *X*, *t* is time. Result of the conversion is shown in the Fig. 25.

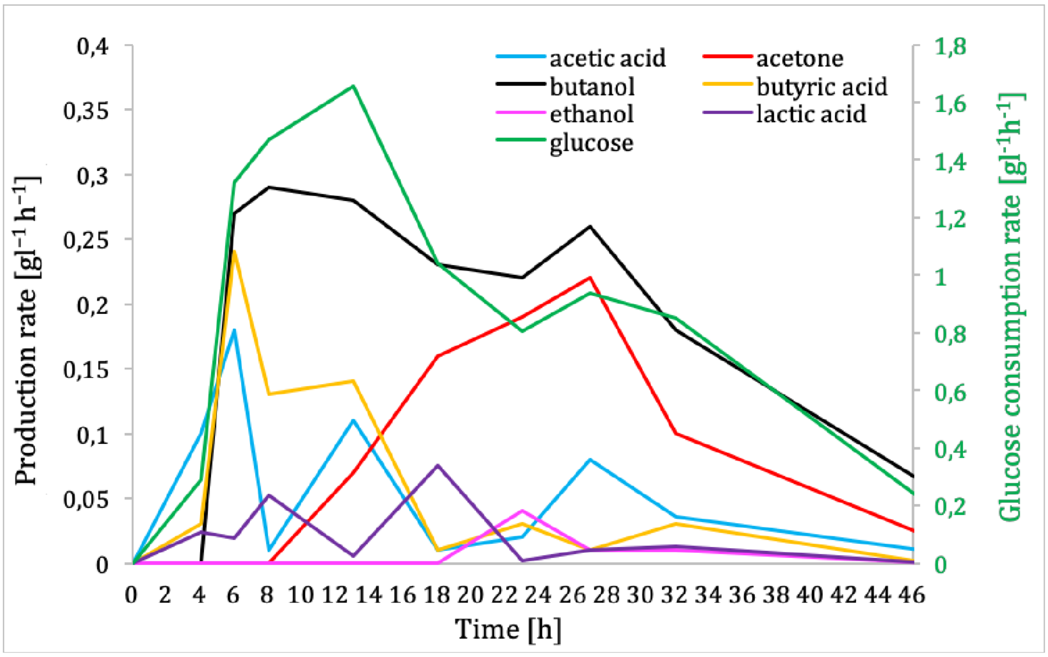


Fig. 25: Reaction rate of seven metabolites over time

To compare the simulation with measured data, I evaluated the similarity of a total seven metabolites at 10 time/step points. Part of the comparison values is in the Tab. 2, all values are in the attachment (*'statistics-simulationVShplc'*). I used the Spearman's correlation coefficient for the statistics since the model is discrete (gives the results as activity level over steps) and measured data are continuous. For the enumeration, I used the Matlab's function *corr (x, y, 'Type', 'Spearman').*

Tab. 2: Reaction rate and acivity level of selected metabolies; the statistics

Time [h] Step[-]	butanol		acetone		ethanol	
	production rate [gl^{-1}h^{-1}]	activity level [%]	production rate [gl^{-1}h^{-1}]	activity level [%]	production rate [gl^{-1}h^{-1}]	activity level [%]
0	0	0	0	0	0	0
4	0	25.0	0	0	0	0
6	0.27	50.0	0	33.3	0	0
8	0.29	50.0	0	37.5	0	0
13	0.28	53.8	0.07	46.2	0	0
18	0.23	50.0	0.16	44.4	0	0
23	0.22	47.8	0.19	43.5	0.04	0
27	0.26	48.1	0.22	44.4	0.01	0
32	0.18	40.6	0.10	37.5	0.01	0
47	0.06	27.7	0.02	25.5	0	0
Correlation	**0.766**					

I evaluated the model match with biological data with a correlation result 0.766 (see the Tab. 2), which is a strong correlation and satisfactory value. Although the biological material shows considerable variability in minor environmental changes and moreover I compared different variables, I can conclude that the model approximates real data very well, that confirms statistical evaluation and visual comparison of model simulation (Fig. 22) with biological data (Fig. 25).

The course of glucose in both graphs shows much higher values than the other metabolites and in both cases the element reaches its maximum value in a short period of time and then gradually decreases. Butanol reaches the highest values of all final products in both graphs, while maintaining a constant production/activity value with a slight decline after reaching its maximum. Ethanol shows the lowest result values of all metabolites; its activity level is zero at all and a minimal production rate appears in biological data. Acetone in both graphs reaches its maximum gradually and begins to decrease slightly in about half time/steps.

5.4.2 Flow cytometry

I used FC analysis to evaluate the match of sporulation level during butanol production. In the laboratory, the ratio of functional cells (live and active) and non-functional cells (dead and inactive) was measured. Since sporulation is an indicator of non-function or non-activity of cells, I compared the activity level of the node sporulation with percentage of non-functional cells labelled as D+I. Comparison values and the Spearman's correlation coefficient are shown in the Tab. 3.

Tab. 3: Percentage of D + I cells and sporulation activity level; the statistics

Time [h]/Step [-]	3.5	6	8.5	10	13	18	23	28	33	48
Activity level [%]	0	17	25	30	36	33	42	43	44	48
D+I [%]	35	39	38	19	27	43	43	71	86	96
Correlation	0.748									

The Spearman's correlation coefficient 0.748 shows a strong correlation, which means that the model approximates the sporulation process in the cell very well. The match is important as sporulation is one of the basic processes in the organism and is also undesirable in requiring an increase in butanol production. Based on the fact the model simulates the cellular sporulation process very well, as demonstrated by statistical results, it will be possible to use the model for sporulation process research as well as for simulations of sporulation avoidance experiments.

5.4.3 Heatmaps

In the next step, I compared heatmaps with activity level of each gene. Heatmap is the visualization of a Z score of average gene expression; Z score is usually used to evaluate transcriptional profiles of selected genes. Data gain for heatmaps creation is described in detail in [57]. An example of heatmaps (yellow-red colour; darker colour higher score) with comparing activity levels is shown in the Tab. 4, all values are in the attachment under the name *'statistics-activityLevelVSheatmaps'.* For the statistic evaluation, I used the Spearman's correlation coefficient and compared activity level of all genes (except of external components, because activity level of these genes is set manually and is unchangeable) with Z score rounded to 10^{-1}.

Tab. 4: Z score and activity level of selected genes; the statistics

	T1	T2	T3	T4	T5	T6
ack	🟥 75 %	🟥 83 %	🟧 60 %	🟨 50 %	🟨 44 %	🟨 54 %
adc	🟨 0 %	🟨 0 %	🟨 0 %	🟥 29 %	🟧 28 %	🟥 33 %
bdhAB	🟨 0 %	🟨 0 %	🟧 0 %	🟥 29 %	🟧 28 %	🟧 33 %
buk1	🟥 75 %	🟧 83 %	🟧 60 %	🟨 50 %	🟨 44 %	🟨 54 %
crt	🟥 75 %	🟧 83 %	🟧 60 %	🟧 50 %	🟨 44 %	🟨 54 %
hbd	🟥 75 %	🟧 83 %	🟧 60 %	🟧 50 %	🟨 44 %	🟨 54 %
pta	🟥 75 %	🟥 83 %	🟧 60 %	🟨 50 %	🟨 44 %	🟨 54 %
Correlation	**0.430**					

The correlation coefficient shows the medium correlation with the result 0.430 (see the Tab. 4). The outcome is relatively satisfactory with respect to comparison different values and moreover variable biological data. Improving the results would be achieved by involving all 5 000 genes that the cell contains; now the pathway consists of only genes involved in butanol production, but these genes are certainly influenced by other genes, proteins, metabolites etc. that are not directly involved in the solvent utilization.

5.5 Dynamic analysis

I conducted a series of simulations to determine network behaviour under different conditions with the main aiming to increase the butanol production. Based on changes in parameter settings such as different value of activity level of external components, I evaluated the significance of individual nodes and the resulting influence of changes on the butanol production.

The first, I examined changes in butanol production at different values of external components activity. Results of simulations are shown in the Fig. 26, individual experiments values are described in the Tab. 5.

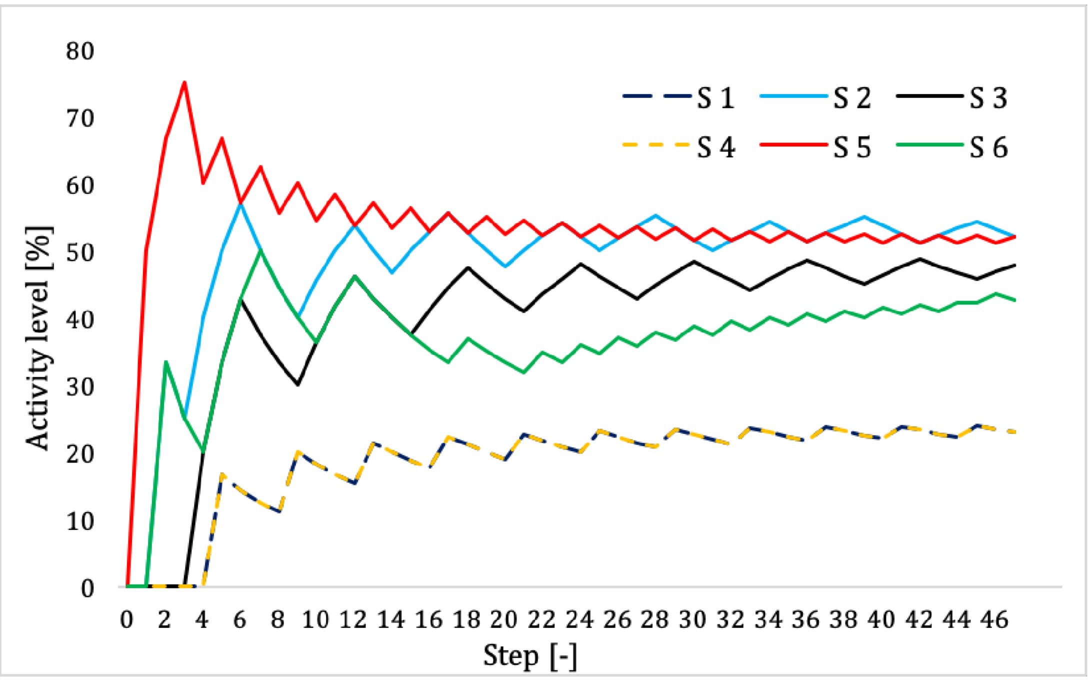


Fig. 26: Individual simulations with different external components activity level

Tab. 5: External components activity level during individual simulations

Activity [%] **Simulation**	*aad*	glucose	NAD(P)H	NADH	phosphorylation	*PTS*	sigA	spoIIE
S 1	0	0	50	50	100	100	100	75
S 2	0	100	50	50	0	100	100	75
S 3	0	100	0	0	100	100	100	75
S 4	0	0	0	0	0	0	100	75
S 5	0	100	100	100	100	100	100	100
S 6	0	50	50	50	100	100	100	75

To evaluate the similarity of butanol production during dynamic analysis and model simulation, I enumerated the relative error (RE) using the equation 4.1:

$$\delta_X = \frac{\Delta_X}{X} \times 100\ [\%], \Delta_X = X_M - X \qquad (4.1)$$

where δ_X is relative error, Δ_X is absolute error, X is conventionally true value (activity level of butanol during simulation), X_M is measured value (activity level of butanol during analysis). Finally, I averaged the RE values in each step to the total RE, results are in the Tab. 6.

Tab. 6: RE values of butanol production during simulation and dynamic analysis

Simulation	**S 1**	**S 2**	**S 3**	**S 4**	**S 5**	**S 6**
Error [%]	44.7	25.4	25.2	44.7	33.0	21.6

The highest butanol production shows the analysis S 5, where I tested an increase in solvent production at higher level of all external components (except of *aad*, a gene that does not contains the clostridium strain, but is needed in the network for proper ethanol function), the result follows from premises – butanol production has increased by nearly 20 % compared to simulation. On the contrary, the decrease in butanol production and the highest RE (i.e. lowest match) occurred in analyses S 1 and S 4, in the first case I inactivated glucose, in the second one all external components except *sigA* and *spoIIE* (genes that cannot be inactivated only by reducing the addition of a certain substance against other components) which suggests that glucose is the most important for butanol production, which was expected as it serves as an energy source for the cell. If glucose is not present, minimal butanol production occurs by conversion from acidogenic substances.

The best simulation-to-analysis match (the lowest RE) shows S 6, where glucose activity was 50 % instead of reducing from 100 to 0 during simulation. The second best match show S 2 and S 3, in which I firstly inactivated phosphorylation and then NAD(P)H with NADH. From these analysis follows that phosphorylation and NAD(P)H with NADH are not crucial for butanol utilization and the constant glucose value instead of its loss due to energy consumption has no major influence.

6 CONCLUSION

The diploma thesis Signaling Pathway for Butanol Production in Solventogenic Clostridia is focused on elaborating five points: literary research on the signaling pathways using systems biology methods, description of data gain for SP modeling with focus on lab techniques for the detection of gene expression, creation of a basic signaling pathway model involved in the production of butanol in solventogenic clostridia and the main parts – creation of the signaling pathway for butanol production in *C. beijerinckii* NRRL B-598, its static and dynamic analysis, comparing the model with biological data and results discussion.

The first three chapters are focused on the theoretical research. Chapter 1 describes basic information about biological networks, graph theory and systems biology. Graph theory is a mathematical discipline dealing with the properties of graphs. Systems biology is characterized by the study of a whole system as interconnected and cooperating elements. The second chapter focuses on signaling pathways, where mathematical models as well as tools for working with SP, databases and data formats have been described. Chapter 3 describes data acquisition for signaling pathways modeling with the main focus on lab techniques for the detection of gene expression, gene products and phosphorylation.

Chapter 4 gives a preview of clostridium bacteria with focus on butanol-producing species. The section is devoted to general signaling pathways involved in butanol production obtained from public databases. Specifically, a comparison of *C. acetobutylicum* and *C. beijerinckii* signaling pathways, five pathways of *C. acetobutylicum* and a genome-scale model of *C. beijerinckii* are included.

The chapter 5 contains the main points of the thesis – creation, simulation, static and dynamic analysis of the *C. beijerinckii's* NRRL B-598 signaling pathway involved in butanol production. Model is available in the Cell Collective tool under the name *'Signaling Pathway for Butanol Production in Clostridium beijerinckii NRRL B-598'*, version 1.1. The thesis also deals with the compassion of the model with biological data (HPLC, FC and heatmaps). Spearman's correlation coefficient shows strong similarity (HPLC and FC) demonstrating very good approximation of model with biological data, allowing future experiments can be replaced by computer simulations with the results of reducing cost and time as well as the ability to implement studies with real samples impossible. Middle correlation (heatmaps) shows satisfactory results with some deviations that could be eliminated by averaging more biological data and involving all 5 000 genes the strain contains.

BIBLIOGRAPHY

[1] Patrick, A., R. Russell. Taking the Mystery out of Biological Networks. *EMBO Reports*. 2004, 5(4): 349–350. DOI: 10.1038/sj.embor.7400129

[2] Network analysis in biology. *European Bioinformatics Intitute* [online]. [Accessed 24 September 2018]. Available from: https://www.ebi.ac.uk/training/online/course/network-analysis-protein-interaction-data-introduction/network-analysis-biology-0

[3] Zhu, X., M. Gerstein, M. Snyder. Getting connected: analysis and principles of biological networks. *Genes Dev.* 2007, 21:1010–24. DOI: 10.1101/gad.1528707

[4] Karl, S., T. Dandekar. Convergence behaviour and Control on Non-Linear Biological Networks. *Scientific Reports*. 2015, 5: 9746. DOI: 10.1038/srep09746

[5] An Explanation of Emergent Properties That Exist in Biology. *Biology Wise* [online]. [Accessed 25 September 2018]. Available from: https://biologywise.com/emergent-properties-explained-in-context-to-biology

[6] Barabasi, AL, Albert R. Emergence of scaling in random networks. *Science*. 1999, 286: 509–512. DOI: 10.1126/science.286.5439.509

[7] Reka, A., H. Jeong, A. Barabasi. Error and attack tolerance of complex networks. *Natur.* 2000, 406 (6794): 378–82. DOI: 10.1038/35019019

[8] Hliněný, P. Teorie Grafů (FI: MA010) [online]. [Accessed 2 October 2018]. Available from: http://is.muni.cz/el/1433/podzim2009/MA010/um/Grafy-text09.pdf

[9] Graphs. *Dr. B. S. Panda* [online]. [Accessed 13 February 2019]. Available from: http://web.iitd.ac.in/~bspanda/gtlecturenotes.pdf

[10] Večerka, A. Grafy a grafové algoritmy [online]. [Accessed 3 October 2018]. Available from: https://phoenix.inf.upol.cz/esf/ucebni/Grafy_a_grafove_algoritmy.pdf

[11] Ma, X., L. Gao. Biological Network analysis: Insights into Structure and Functions. *Briefings in Functional Genomics.* 2012, 11 (6): 434–42. DOI: 10.1093/bfgp/els045

[12] Rajagopalan, V. Computer-Aided Analysis of Power Electronic Systems. CRC Press, 1987.

[13] Masopust, T. Grafové algoritmy [online]. [Accessed 13 February 2019]. Available from: http://www.fit.vutbr.cz/~masopust/GAL/gal-text.pdf

[14] Oltvai Z., A. Barabási. Life's complexity pyramid. *Science*. 2002, 298: 763-764. DOI: 10.1126/science.1078563

[15] Kitano, H. Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology. *Curr Genet*. 2002, 41(1):1-10. DOI: 10.1007/s00294-002-0285-z

[16] Chuang, H. Y., M. Hofree, T. Ideker. A decade of systems biology. *Annual review of cell and developmental biology*. 2010, 26: 721-44. DOI: 10.1146/annurev-cellbio-100109-104122

[17] Hood, L. Systems Biology: new opportunities arising from genomics, proteomics and beyond. *Experimental Hematology*. 1998, 26: 681

[18] NCI Dictionary of Cancer Terms. *National Cancer Institute at the National Institutes of Health* [online]. [Accessed 5 October 2018]. Available from: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/signaling-pathway

[19] Signaling pathways. *Tocris Bioscience* [online]. [Accessed 5 October 2018]. Available from: https://www.tocris.com/signaling-pathways

[20] Lavrik, I. N., M. G. Samsonova. The Systems Biology of Signaling Pathways. *Biophysics*. 2016, 61(1): 78-84. DOI: 10.1134/S0006350916010127

[21] Samaga, R., S. Klamt. Modeling approaches for qualitative and semi-quantitative analysis of cellular signaling networks. *Cell Communication and Signaling*. 2013, 11(1): 11-43. DOI: 10.1186/1478-811X-11-43

[22] Wittmann, D.M, J. Krumsiek,J. Saez-Rodriguez, D. A. Lauffenburger, S. A. Klamt, F. J. Theis. Transforming Boolean models to continuous models: methodology and application to T-cell receptor signaling. *BMC Systems Biology*. 2009, 3(1): 3-98. DOI: 10.1186/1752-0509-3-98

[23] Fumiã, H. F., M. L. Martins, J. P. Brody. Boolean Network Model for Cancer Pathways: Predicting Carcinogenesis and Targeted Therapy Outcomes. *PLoS ONE*. 2013, 8(7). DOI: 10.1371/journal.pone.0069008

[24] Mogilner, A., R. Wollman, W. F Marshall. Quantitative Modeling in Cell Biology: What Is It Good for? *Developmental Cell*. 2006, 11(3): 279-287. DOI: 10.1016/j.devcel.2006.08.004

[25] Berg, J. M., J. L. Tymoczko, L. Stryer. *Biochemistry. 5th edition. New York: W. H. Freeman*. 2002. Section 8.4: The Michaelis-Menten Model Accounts for the Kinetic Properties of Many Enzymes. ISBN-10: 0-7167-3051-0

[26] Helikar, T., B. Kowal, S. Mcclenathan. The Cell Collective: Toward an open and collaborative approach to systems biology. *BMC Systems Biology*. 2012, 6(1): 96. DOI: 10.1186/1752-0509-6-96

[27] What is Cytoscape? *Cytoscape* [online]. [Accessed 16 October 2018]. Available from: http://cytoscape.org/what_is_cytoscape.html

[28] Bastian, M., S. Heymann, M. Jacomy. An Open Source Software for Exploring and Manipulating Networks. *International AAAI Conference on Weblogs and Social Media* [online]. [Accessed 15 October 2018]. Available from: https://gephi.org/publications/gephi-bastian-feb09.pdf

[29] Naldi, A., D. Berenguier, A. Fauré, F. Lopez, D. Thieffry, C. Chaouiya. Logical modelling of regulatory networks with GINsim 2.3. *Biosystems.* 2009, 97(2): 134-9. DOI: 10.1016/j.biosystems.2009.04.008

[30] Cara, A., A. Garg, G. Micheli, I. Xenarios, L. Mendoza. Dynamic simulation of regulatory networks using SQUAD. BMC *Bioinformatics*. 2007, 8: 462. DOI: 10.1186/1471-2105-8-462

[31] Hu, Z., J. Mellor, J. Wu, Ch. DeLisi. VisANT: an online visualization and analysis tool for biological interaction data. BMC *Bioinformatics.* 2004, 5:17. DOI: 10.1186/1471-2105-5-17

[32] Hu, Z., et al. VisANT 4.0: Integrative network platform to connect genes, drugs, diseases and therapies. *Nucleic Acids Research*. 2013, 41(1): 225-31. DOI: 10.1093/nar/gkt401

[33] Iersel, M., T. Kelder, R. Pico, K. Hanspers, S. Coort, B. Conklin, Ch. Evelo. Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics*. 2008, 9: 399. DOI: 10.1186/1471-2105-9-399

[34] What is PathVisio? *PathVisio* [online]. [Accessed 19 November 2018]. Available from: https://www.pathvisio.org

[35] What is Reactome? *Reactome* [online]. [Accessed 16 October 2018]. Available from: https://reactome.org/what-is-reactome

[36] Kanehisa, M., M. Furumichi, M. Tanabe, Y. Sato, K. Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. 2017, 45, D353-D361. DOI: 10.1093/nar/gkw1092

[37] Slenter, D. N., et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*. 2018, 46, D661–D667. DOI: 10.1093/nar/gkx1064

[38] UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. 2016, 45(D1): D158-D169. DOI: 10.1093/nar/gkw1099

[39] About UniProt. *UniProt* [online]. [Accessed 20 November 2018]. Available from: https://www.uniprot.org/help/about

[40] BioModels. *EMBL-EBI* [online]. [Accessed 20 November 2018]. Available from: https://www.ebi.ac.uk/biomodels/

[41] Gama-Castro, S., H. Salgado, A. Santos-Zavaleta, et al. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res*. 2015, 44(D1): D133-43. DOI: 10.1093/nar/gkv1156

[42] GML: A portable Graph File Format. *University of Passau* [online]. [Accessed 16 October 2018]. Available from: https://www.fim.uni-passau.de/fileadmin/files/lehrstuhl/brandenburg/projekte/gml/gml-technical-report.pdf

[43] KEGG Markup Language manual. *KEGG* [online]. [Accessed 20 November 2018]. Available from: https://www.kegg.jp/kegg/xml/docs/

[44] 7. Supported Network File Formats. *Cytoscape User Manual* [online]. [Accessed 20 November 2018]. Available from: http://manual.cytoscape.org/en/stable/Supported_Network_File_Formats.html

[45] Garg, H., P. Rawat. An Improved Algorithm for Finding All Pair Shortest Path. International *Journal of Computer Applications*. 2012, 47: 35-37. DOI: 10.5120/7539-0492.

[46] Zhao, X., ZP Liu. Analysis of Topological Parameters of Complex Disease Genes Reveals the Importance of Location in a Biomolecular Network. *Genes*. 2019, 10(2): 143. DOI:10.3390/genes10020143

[47] Barabási, A. *Network science.* Cambridge University Press, 2016. ISBN-10: 1107076269

[48] Closeness centrality. *EMBL-EBI* [online]. [Accessed 10 April 2019]. Available from: https://www.ebi.ac.uk/training/online/course/network-analysis-protein-interaction-data-introduction/building-and-analysing-ppins-1

[49] Chalancon, G., K. Kruse, M. Babu. Clustering Coefficient. *Springer, New York.* 2013, 422-424. DOI: https://doi.org/10.1007/978-1-4419-9863-7

[50] Scardoni, G., M. Petterlini, C. Laudanna. Analyzing Biological Network Parameters with CentiScaPe, *Bioinformatics.* 2009, 21: 2857–59. DOI:10.1093/bioinformatics/btp517

[51] Nacher, J. C., J. Schwartz. A Global View of Drug-Therapy Interactions. *BMC Pharmacology*. 2008, 1: 5. DOI: 10.1186/1471-2210-8-5

[52] Pachón, A., L. Sacerdote, S. Yang. Scale-free behavior of networks with the copresence of preferential and uniform attachment rules. *Physica D: Nonlinear Phenomena*. 2017, 371: 1-12. DOI: 10.1016/j.physd.2018.01.005.

[53] Savill, N., S. Turner, J. Sherratt, K. Painter. From a discrete to a continuous model of biological cell movement. *Physical review*. 2004, E 69. DOI: 10.1103/PhysRevE.69.021910

[54] Wang, Z., M. Gerstein, M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009, 10(1): 57-63. DOI: 10.1038/nrg2484

[55] Zhao, S., WP Fung-Leug, A. Bittner, K. Ngo, X. Liu. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS One*. 2014, 9(1): e78644. DOI: 10.1371/journal.pone.0078644

[56] Sudhagar, A., G. Kumar, M. El-Matbouli. Transcriptome Analysis Based on RNA-Seq in Understanding Pathogenic Mechanisms of Diseases and the Immune System of Fish: A Comprehensive Review. *Int. J. Mol. Sci.* 2018, 19(1), 245. DOI: 10.3390/ijms19010245

[57] Sedlar, K., P. Koscova, M. Vasylkivska, et al. Transcription profiling of butanol producer Clostridium beijerinckii NRRL B-598 using RNA-Seq. *BMC Genomics.* 2018, 19(1): 415. DOI: 10.1186/s12864-018-4805-8

[58] Southern, E. M. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology.* 1975, 98 (3): 503–517. DOI: 10.1016/S0022-2836(75)80083-0

[59] Hayes, P., R. Wolf, J. Hayes. Blotting techniques for the study of DNA, RNA, and proteins. *BMC Clinical Research*. 1989, 299(6705): 965-8. PMID: 2478239

[60] Malviya, R., V. Bansal, OP Pal, PK Sharma. High Performance Liquid Chromatography: A short review. *Global Pharma Tech*. 2010, 2(5): 22-26. DOI: 10.1234/jgpt.v2i5.208

[61] Forward Scatter vs. Side Scatter. *FlowJo* [online]. [Accessed 3 May 2019]. Available from: https://www.flowjo.com/learn/flowjo-university/flowjo/getting-started-with-flowjo/58

[62] Picot, J., CL Guerin, C. Le Van Kim, CM Boulanger. Flow cytometry: retrospective, fundamentals and recent instrumentation. *Cytotechnology*. 2012, 64(2): 109–130. DOI:10.1007/s10616-011-9415-0

[63] Al-Hinai MA, Jones SW, Papoutsakis ET. The Clostridium sporulation programs: diversity and preservation of endospore differentiation. *Microbiol Mol Biol Rev*. 2015, 79(1): 19-37. DOI: 10.1128/MMBR.00025-14

[64] Patáková, P., J. Kolek. Využití genového inženýrství pro zlepšení procesu fermentační výroby butanolu. *Chem. listy.* 2015, 109: 830-835.

[65] Patakova, P., B. Branska, K. Sedlar, et al. Acidogenesis, solventogenesis, metabolic stress response and life cycle changes in Clostridium beijerinckii NRRL B-598 at the transcriptomic level. *Sci Rep*. 2019, 9(1): 1371. DOI: 10.1038/s41598-018-37679-0

[66] Clostridium acetobutylicum. *Global Catalogue of Microorganisms* [online]. [Accessed 6 December 2018]. Available from: http://gcm.wfcc.info/speciesPage.jsp?strain_name=Clostridium%20acetobutylicum

[67] Keis, S., et.al. Emended descriptions of Clostridium acetobutylicum and Clostridium beijerinckii, and descriptions of Clostridium saccharoperbutylacetonicum sp. nov. and Clostridium saccharobutylicum sp. nov. *International Journal of Systematic Bacteriology*. 2001, 51: 2095-2103. DOI: 10.1099/00207713-51-6-2095

[68] Moon, H., Y. Jang, et al. One hundred years of clostridial butanol fermentation. *FEMS Microbiology Letters*. 2016, 363 (3). DOI: 10.1093/femsle/fnw001

[69] Berezina O. V., A. Brandt, S. Yarotsky, W. H. Schwarz, C. C Zverlov. Isolation of a new butanol-producing Clostridium strain: high level of hemicellulosic activity and structure of solventogenesis genes of a new Clostridium saccharobutylicum isolate. *Syst. Appl. Microbiol.* 2009, 32: 449–459. DOI: 10.1016/j.syapm.2009.07.005

[70] Yunpeng, Y., N. Xiaoqun, J. Yuqian, Y. Chen, G. Yang, J. Weihong. Metabolic regulation in solventogenic clostridia: regulators, mechanisms and engineering, *Biotechnology Advances*, 2018, 36(4): 905-914. DOI: 10.1016/j.biotechadv.2018.02.012

[71] Caspi, R., T. Altman, K. Dreher, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*, 2012, D742-53. DOI: 10.1093/nar/gkr1014

[72] Butanoate metabolism – Clostridium acetobutylicum ATCC 824. *KEGG* [online]. [Accessed 18 December 2018]. Available from: https://www.genome.jp/kegg-bin/show_pathway?cac00650

[73] Two-component system – Clostridium acetobutylicum ATCC 824. *KEGG* [online]. [Accessed 18 December 2018]. Available from: https://www.genome.jp/kegg-bin/show_pathway?cac02020

[74] Senger, R. S., E. T. Papoutsakis. Genome-Scale model for Clostridium acetobutylicum: Part I. Metabolic network resolution and analysis. *Biotechnol Bioeng*. 2008, 101(5):1036-52. DOI: 10.1002/bit.22010

[75] Lee, J., H. Yun, A. Feist, B. Palsson, S. Lee. Genome-scale reconstruction and *in silico* analysis of the *Clostridium acetobutylicum* ATCC 824 metabolic network. *Appl Microbiol Biotechnol*. 2008, 80: 849. DOI: 10.1007/s00253-008-1654-4

[76] Milne, C. B., J. A. Eddy, R. Raju, et al. Metabolic network reconstruction and genome-scale model of butanol-producing strain Clostridium beijerinckii NCIMB 8052. *BMC Syst Biol*. 2011, 5: 130. DOI: 10.1186/1752-0509-5-130

[77] Sedlar, K., J. Kolek, I. Provaznik, P. Patakova. Reclassification of non-type strain Clostridium pasteurianum NRRL B-598 as Clostridium beijerinckii NRRL B-598. *Journal of Biotechnology*. 2017, 244: 1-3. DOI: 10.1016/j.jbiotec.2017.01.003

[78] Kolek, J., K. Sedlar, I. Provaznik, P. Patakova. Draft Genome Sequence of Clostridium pasteurianum NRRL B-598, a Potential Butanol or Hydrogen Producer. *Genome Announc*. 2014, 2(2). DOI: 10.1128/genomeA.00192-14

[79] Newman, M. Networks: An introduction. *Oxford University Press*, 2010. ISBN: 978-0-19-920665-0

LIST OF FIGURES

LIST OF TABLES

LIST OF ABREVATIONS

ABE fermentation	acetone-butanol-ethanol fermentation
ASP(L)	average shortest path (length)
BFS	breath first search
CC	closeness centrality
DFT	depth first search
FBA	flux balance analysis
FC	flow cytometry
FIFO	first in, first out
FL	feedback loops
Gephi	The Open Graph Viz Platform
GML	Graph Modeling Language
HPLC	high-performance liquid chromatography
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG Orthology
ND	network diameter
ODE	ordinary differential equation
PDE	partial differential equation
RE	relative error
SB	systems biology
SBML	Systems Biology Markup Language
SP	signaling pathways
XML	eXtensible Markup Language

LIST OF ELECTRONIC ATTACHMENTS

‘*Signaling Pathway for Butanol Production in Clostridium beijerinckii NRRL B-598’:* the proposed signaling pathway in SBML format

‘static analysis (Cytoscape)’: a table of static analysis results (CSV format): average shortest path length, clustering coefficient, closeness centrality, eccentricity, stress, betweenness centrality

‘statisctics-sporulationVSfc’: a comparison table of sporulation activity levels and percentage of non-active cells in 10 points in CSV format

‘statistics-activityLevelVSheatmaps’: a comparison table of individual genes activity levels and heatmaps in CSV format

Subfolder ‘*acquired models’*

‘comparison SP of beijerinckii and acetobutylicum’: figure of a whole comparing SP

‘C. Acetobutylicum ATCC 824 - Butanoate metabolism’: a pathway in XML format

‘C. acetobutylicum ATCC 824 - Superpathway of Clostridium acetobutylicum solventogenic fermentation’: a pathway in XML format

‘C. Acetobutylicum ATCC 824 - two-component system’: a pathway in xml format

‘Genome-scale metabolic network of Clostridium acetobutylicum – Lee’: a genome-scale model in XML format labelled as iCac802

‘Genome-scale metabolic network of Clostridium acetobutylicum – Senger’: a genome-scale model in XML format labelled as iJL432

‘Genome-scale metabolic network of Clostridium beijerinckii’: a genome-scale model in XML format labelled as iCB925

Subfolder ‘*static analysis (Cell Collective)’*

‘allPairShortestPath’: a matrix *M* of all pair shortest path in the proposed signaling pathway in CSV format

‘connectivityDistribution’: tables of connectivity distribution, connectivity in degree and connectivity out degree in CSV format

‘feedbackLoops’: a table of all FL in the proposed signaling pathway in CSV format

‘closenessCentrality’: a table of closeness centralities of all nodes in the proposed signaling pathway in CSV format